Under consideration for publication in Knowledge and Information Systems

Semi-Supervised Learning by Disagreement

Zhi-Hua Zhou and Ming Li

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

Abstract. In many real-world tasks there are abundant unlabeled examples but the number of labeled training examples is limited, because labeling the examples requires human efforts and expertise. So, semi-supervised learning which tries to exploit unlabeled examples to improve learning performance has become a hot topic. *Disagreement-based semi-supervised learning* is an interesting paradigm, where multiple learners are trained for the task and the disagreements among the learners are exploited during the semi-supervised learning process. This survey article provides an introduction to research advances in this paradigm.

Keywords: Machine Learning; Data Mining; Semi-Supervised Learning; Disagreement-Based Semi-Supervised Learning

1. Introduction

In traditional supervised learning, hypotheses are learned from a large number of training examples. Each training example has a *label* which indicates the desired output of the event described by the example. In classification, the label indicates the *category* into which the corresponding example falls into; in regression, the label is a real-valued output such as temperature, height, price, etc.

Advances in data collection and storage technology enable the easy accumulation of a large amount of training instances without labels in many real-world applications. Assigning labels to those unlabeled examples is expensive because the labeling process requires human efforts and expertise. For example, in computer-aided medical diagnosis, a large number of X-ray images can be obtained from routine examination, yet it is difficult to ask physicians to mark

Received October 16, 2008
Revised March 16, 2009
Accepted April 03, 2009

all focuses in all images. If we use traditional supervised learning techniques to build a diagnosis system, then only a small portion of training data, on which the focuses have been marked, are useful. Due to the limited amount of labeled training examples, it may be difficult to get a strong diagnosis system. Then, a question arises: Can we leverage the abundant unlabeled training examples with a few labeled training examples to generate a strong hypothesis? Roughly speaking, there are three major techniques for this purpose [82], i.e., *semi-supervised learning*, *transductive learning* and *active learning*.

Semi-supervised learning [21, 92] deals with methods for automatically exploiting unlabeled data in addition to labeled data to improve learning performance, where no human intervention is assumed. Transductive learning is a cousin of semi-supervised learning, which also tries to exploit unlabeled data automatically. The main difference between them lies in the different assumptions on the test data. Transductive learning takes a "close-world" assumption, i.e., the test data set is known in advance and the goal of learning is to optimize the generalization ability on this test data set, while the unlabeled examples are exactly the test examples. Semi-supervised learning takes an "open-world" assumption, i.e., the test data set is not known and the unlabeled examples are not necessary test examples. In fact, the idea of transductive learning originated from statistical learning theory [69]. Vapnik [69] believed that one often wants to make predictions on test examples at hand instead of on all potential examples, while inductive learning that seeks the best hypothesis over the whole distribution is a problem more difficult than what is actually needed; we should not try to solve a problem by solving a more difficult intermediate problem, and so, transductive learning is more appropriate than inductive learning. Up to now there is still a debate in the machine learning community on this learning philosophy. Nevertheless, it is well recognized that transductive learning provides an important insight into the exploitation of unlabeled data.

Active learning deals with methods that assume that the learner has some control over the input space. In exploiting unlabeled data, it requires an oracle, such as a human expert, from which the ground-truth labels of instances can be queried. The goal of active learning is to minimize the number of queries for building a strong learner. Here, the key is to select those unlabeled examples where the labeling will convey the most helpful information to the learner. There are two major schemes, i.e., uncertainty sampling and committee-based sampling. Approaches of the former train a single learner and then query the unlabeled example on which the learner is least confident [45]; approaches of the latter generate multiple learners and then query the unlabeled example on which the learners disagree to the most [1, 63].

In this survey article, we will introduce an interesting and important semi-supervised learning paradigm, i.e., *disagreement-based semi-supervised learning*. This line of research started from Blum & Mitchell's seminal paper on co-training [13] [1]. Different relevant approaches have been developed with different names, and recently the name *disagreement-based semi-supervised learning* was coined [83] to reflect the fact that they are actually in the same family, and the key for the learning process to proceed is to maintain a large disagreement between base learners. Although transductive learning or active learning may be involved in some place, we will not talk more on them. In the following we will start by a

[1] This seminal paper won the "ten years best paper award" at ICML'08.

brief introduction to semi-supervised learning, and then we will go to the main theme to introduce representative disagreement-based semi-supervised learning approaches, theoretical foundations, and some applications to real-world tasks.

2. Semi-Supervised Learning

In semi-supervised learning, a labeled training data set $L = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), ..., (\boldsymbol{x}_{|L|}, y_{|L|})\}$ and an unlabeled training data set $U = \{\boldsymbol{x}'_1, \boldsymbol{x}'_2, ..., \boldsymbol{x}'_{|U|}\}$ are presented to the learning algorithm to construct a function $f : X \mapsto Y$ for predicting the labels of unseen instances, where X and Y are respectively the input space and output space, $\boldsymbol{x}_i, \boldsymbol{x}'_j \in X$ $(i = 1, 2, ..., |L|, j = 1, 2, ..., |U|)$ are d-dimensional feature vectors drawn from X, and $y_i \in Y$ is the label of $\boldsymbol{x}_i$; usually $|L| \ll |U|$.

It is well-known that semi-supervised learning originated from [64]. In fact, some straightforward use of unlabeled examples appeared even earlier [40, 50, 52, 53, 57]. Due to the difficulties in incorporating unlabeled data directly into conventional supervised learning methods (e.g., BP neural networks) and the lack of a clear understanding of the value of unlabeled data in the learning process, the study of semi-supervised learning attracted attention only after the middle of 1990s. As the demand for automatic exploitation of unlabeled data increases and the value of unlabeled data was disclosed by some early analyses [54, 78], semi-supervised learning has become a hot topic.

Most early studies did not provide insight or explanation to the reason why unlabeled data can be beneficial. Miller and Uyar [54] provided possibly the first explanation to the usefulness of unlabeled data from the perspective of data distribution estimation. They assumed that the data came from a Gaussian mixture model with L mixture components, i.e.,

$$f(\boldsymbol{x}|\theta) = \sum_{l=1}^{L} \alpha_l f(\boldsymbol{x}|\theta_l) \ , \tag{1}$$

where α_l is the mixture coefficient satisfying $\sum_{l=1}^{L} \alpha_l = 1$, while $\theta = \{\theta_l\}$ are the model parameters. In this case, label c_i can be considered a random variable C whose distribution $P(c_i|\boldsymbol{x}_i, m_i)$ is determined by the mixture component m_i and the feature vector $\boldsymbol{x}_i$. The optimal classification rule for this model is the MAP (maximum *a posterior*) criterion, that is,

$$h(\boldsymbol{x}) = \arg\max_k \sum_j P\left(c_i = k|m_i = j, \boldsymbol{x}_i\right) P\left(m_i = j|\boldsymbol{x}_i\right) \ , \tag{2}$$

where

$$P\left(m_i = j|\boldsymbol{x}_i\right) = \frac{\alpha_j f(\boldsymbol{x}_i|\theta_j)}{\sum\limits_{l=1}^{L} \alpha_l f(\boldsymbol{x}_i|\theta_l)} \ . \tag{3}$$

Thus, the objective of learning is accomplished by estimating the terms $P\left(c_i = k|m_i = j, \boldsymbol{x}_i\right)$ and $P\left(m_i = j|\boldsymbol{x}_i\right)$ from the training data. It can be found that only the estimate of the first probability involves the class label. So, unlabeled examples can be used to improve the estimate of the second probability, and hence improve the performance of the learned hypothesis. Later, Zhang and Oles

[78] analyzed the value of unlabeled data for parametric models. They suggested that if a parametric model can be decomposed as $P(\boldsymbol{x}, y|\theta) = P(y|\boldsymbol{x}, \theta)P(\boldsymbol{x}|\theta)$, the use of unlabeled examples can help to reach a better estimate of the model parameters.

There are two basic assumptions in semi-supervised learning, that is, the *cluster assumption* and the *manifold assumption*. The former assumes that data with similar inputs should have similar class labels; the latter assumes that data with similar inputs should have similar outputs. The cluster assumption concerns classification, while the manifold assumption can also be applied to tasks other than classification. In some sense, the manifold assumption is a generalization of the cluster assumption. These assumptions are closely related to the idea of *low density separation*, which has been taken by many semi-supervised learning algorithms. No matter which assumption is taken, the common underlying belief is that the unlabeled data provide some helpful information on the ground-truth data distribution. So, a key of semi-supervised learning is to exploit the distributional information disclosed by unlabeled examples.

Many semi-supervised learning algorithms have been developed. Roughly speaking, they can be categorized into four categories, i.e., generative methods [54,56,64], S3VMs (Semi-Supervised Support Vector Machines) [22,37,42,44], graph-based methods [7–9, 80, 93], and disagreement-based methods [13, 16, 36, 48, 55, 85, 88, 89, 91].

In generative approaches, both labeled and unlabeled examples are assumed to be generated by the same parametric model. Thus, the model parameters directly link unlabeled examples and the learning objective. Methods in this category usually treat the labels of the unlabeled data as missing values of model parameters, and employ the EM (expectation-maximization) algorithm [29] to conduct maximum likelihood estimation of the model parameters. The methods differ from each other by the generative models used to fit the data, for example, mixture of Gaussian [64], mixture of experts [54], Naïve Bayes [56], etc. The generative methods are simple and easy to implement, and may achieve better performance than discriminative models when learning with a very small number of labeled examples. However, methods in this category suffer from a serious deficiency. That is, when the model assumption is incorrect, fitting the model using a large number of unlabeled data will result in performance degradation [23, 26]. Thus, in order to make it effective in real-world applications, one needs to determine the correct generative model to use based on domain knowledge. There are also attempts of combining advantages of generative and discriminative approaches [4, 33].

S3VMs try to use unlabeled data to adjust the decision boundary learned from the small number of labeled examples, such that it goes through the less dense region while keeping the labeled data being correctly classified. Joachims [42] proposed TSVM (Transductive Support Vector Machine). This algorithm firstly initiates an SVM using labeled examples and assigns potential labels to unlabeled data. Then, it iteratively maximizes the margin over both labeled and unlabeled data with their potential labels by flipping the labels of the unlabeled examples on different sides of the decision boundary. An optimal solution is reached when the decision boundary not only classifies the labeled data as accurate as possible but also avoids going through the high density region. Chapelle and Zien [22] derived a special graph kernel using the low density separation criterion, and employed gradient descent to solve the SVM optimization problem. The non-convexity of the loss function of TSVM leads to the fact that there are

many local optima. Many studies tried to reduce the negative influence caused by the non-convexity. Typical methods include: employing a continuation approach, which begins by minimizing an easy convex objective function and sequentially deforms it to the non-convex loss function of TSVM [20]; employing a deterministic annealing approach, which decomposes the original optimization problem into a series of convex optimization problems, from easy to hard, and solves them sequentially [65,66]; employing the convex-concave procedure (CCCP) [77] to directly optimize the non-convex loss function [25], etc.

The first graph-based semi-supervised learning method is possibly [11]. Blum and Chawla [11] constructed a graph whose nodes are the training examples (both labeled and unlabeled) and the edges between nodes reflect certain relation, such as similarity, between the corresponding examples. Based on the graph, the semi-supervised learning problem can be addressed by seeking the minimum cut of the graph such that nodes in each connected component have the same label. Later, Blum et al. [12] disturbed the graph with some randomness and produced a "soft" minimum cut using majority voting. Note that the predictive function in [11] and [12] is discrete, i.e., the prediction on unlabeled examples should be one of the possible labels. Zhu et al. [93] extended the discrete prediction function to continuous case. They modelled the distribution of the prediction function over the graph with Gaussian random fields and analytically showed that the prediction function with the lowest energy should have the harmonic property. They designed a label propagation strategy over the graph using such a harmonic property. Zhou et al. [80] defined a quadratic loss of the prediction function over both the labeled and unlabeled data, and used a normalized graph Laplacian as the regularizer. They provided an iterative label propagation method yielding the same solution of the regularized loss function. Belkin and Niyogi [7] assumed that the data are distributed on a Riemannian manifold, and used the discrete spectrum and its eigenfunction of a nearest neighbor graph to reform the learning problem to interpolate over the data points in Hillbert space. Then, Belkin et al. [8,9] further extended the idea of manifold learning in semi-supervised learning scenario, and proposed manifold regularization framework in Reproducing Kernel Hillbert Space (RKHS). This framework directly exploits the local smoothness assumption to regularize the loss function defined over the labeled training examples such that the learned prediction function is biased to give similar output to the examples in a local region. Sindhwani et al. [67] embedded the manifold regularization into a semi-supervised kernel defined over the overall input space. They modified the original RKHS by changing the norm while keeping the same function space. This leads to a new RKHS, in which learning supervised kernel machines with only the labeled data is equivalent to a certain manifold regularization over both labeled and unlabeled data in original input space.

Most of the previous studies on graph-based semi-supervised learning usually focus on how to conduct semi-supervised learning over a given graph. It is noteworthy that how to construct a graph which reflects the essential relationship between examples is a key that will seriously affect the learning performance. Although the graph construction might favor certain domain knowledge, some researchers have attempted to construct graphs of high quality using some domain-knowledge-independent properties. Carreira-Perpinan and Zemel [19] generated multiple minimum spanning trees based on perturbation to construct a robust graph. Wang and Zhang [70] used the idea of LLE [60] that instances can be reconstructed by their neighbors to obtain weights over the edges in the graph.

Zhang and Lee [79] selected a better RBF bandwidth to minimize the predictive error on labeled data using cross validation. Hein and Maier [39] attempted to remove noisy data and hence obtained a better graph. Note that, although graph-based semi-supervised learning approaches have been used in many applications, they suffer seriously from poor scalability. This deficiency has been noticed and some efforts have been devoted to this topic [34, 76, 94]. Recently, Goldberg et al. [35] proposed an online manifold regularization framework as well as efficient solutions, which improves the applicability of manifold regularization to large-scale and real-time problems.

The name *disagreement-based semi-supervised learning* was coined recently by Zhou [83], but this line of research started from Blum & Mitchell's seminal work [13]. In those approaches, multiple learners are trained for the same task and the disagreements among the learners are exploited during the learning process. Here, unlabeled data serve as a kind of "platform" for information exchange. If one learner is much more confident on a disagreed unlabeled example than other learner(s), then this learner will teach other(s) with this example; if all learners are comparably confident on a disagreed unlabeled example, then this example may be selected for query. Since methods in this category do not suffer from the model assumption violation, nor the non-convexity of the loss function, nor the poor scalability of the learning algorithms, disagreement-based semi-supervised learning has become an important learning paradigm. In the following sections, we will review studies of this paradigm in more detail.

3. Disagreement-Based Semi-Supervised Learning

A key of disagreement-based semi-supervised learning is to generate multiple learners, let them collaborate to exploit unlabeled examples, and maintain a large disagreement between the base learners. In this section, we roughly classify existing disagreement-based semi-supervised learning techniques into three categories, that is, learning with multiple views, learning with single view multiple classifiers, and learning with single view multiple regressors.

3.1. Learning with Multiple Views

In some applications, the data set has several disjoint subsets of attributes (each subset is called as a *view*). For example, the web page classification task has two views, i.e., the texts appearing on the web page itself and the anchor text attached to hyper-links pointing to this page [13]. Naturally, we can generate multiple learners with these multiple views and then use the multiple learners to start disagreement-based semi-supervised learning. Note that there were abundant research on multi-view learning, yet a lot of work was irrelevant to semi-supervised learning, and so they are not mentioned in this section.

The first algorithm of this paradigm is the *co-training* algorithm proposed by Blum and Mitchell [13]. They assumed that the data has two *sufficient and redundant* views (i.e., attribute sets), where each view is sufficient for training a strong learner and the views are conditionally independent to each other given the class label.

The co-training procedure, which is illustrated in Fig. 1, is rather simple. In co-training, each learner is generated using the original labeled data. Then,

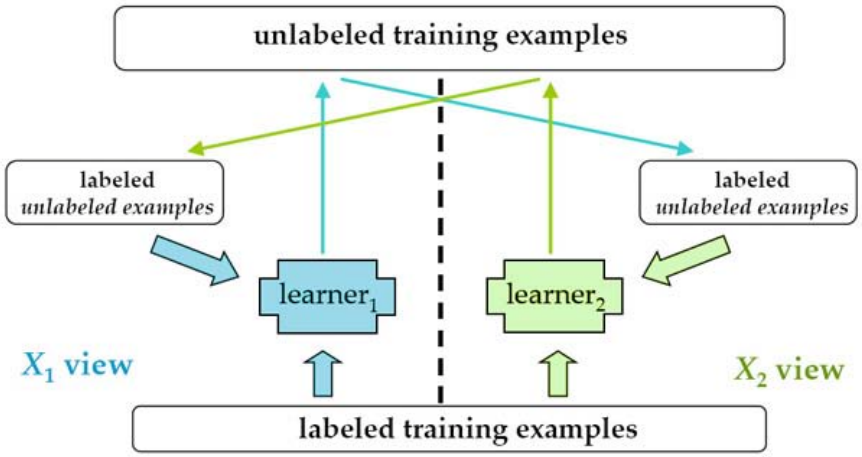

Fig. 1. An illustration of the co-training procedure

each learner will select and label some high-confident unlabeled examples for its peer learner. Later, the learners will be refined using the newly labeled examples provided by its peer. With such a process, when two learners disagree on an unlabeled example, the learner which misclassifies this example will be taught by its peer. The whole process will repeat until no learner changes or a pre-set number of learning rounds has been executed. Blum and Mitchell [13] analyzed the effectiveness of the co-training algorithm, and showed that co-training can effectively exploit unlabeled data to improve the generalization ability, given that the training data are described by sufficient and redundant views which are conditionally independent to each other given the class label.

Another famous multi-view semi-supervised learning algorithm, co-EM [55], combines multi-view learning with the probabilistic EM approach. This algorithm requires the base learners be capable of estimating class probabilities, and so naïve Bayes classifiers are generally used. Through casting linear classifiers into a probabilistic framework, Brefeld and Scheffer [15] replaced the naïve Bayes classifiers by support vector machines. The co-EM algorithm has also been applied to unsupervised clustering [10].

Brefeld et al. [14] tried to construct a hidden markov perceptron [3] in each of the two views, where the two hidden markov perceptrons were updated according to the heuristic that, if the two perceptrons disagree on an unlabled example then each perceptron is moved towards that of its peer view. Brefeld et al. [14] did not mention how to extend this method to more than two views, but it could be able to move the perceptrons towards their median peer view when they disagree, according to the essential of their heuristics. However, the convergence of the process has not been proved even for two-view case. Brefeld and Scheffer [16] extended SVM-2K [32], a supervised co-SVM that minimizes the training error as well as the disagreement between the two views, to semi-supervised learning and applied it to several tasks involving structured output variables such as multi-class classification, label sequence learning and natural language parsing.

In real-world applications, when the data has two views, it is rarely that the two views are conditionally independent given the class label. Even a weak

conditional independence [2] is difficult to be met in practice. In fact, the assumption of sufficient and redundant views, which are conditionally independent to each other given the class label, is so strong that when it holds, a single labeled training example is able to launch a successful semi-supervised learning [91].

Zhou et al. [91] effectively exploited the "compatibility" of the two views to turn some unlabeled examples into labeled ones. Specifically, given two sufficient and redundant views v_1 and v_2 (in this case, an instance is represented by $\boldsymbol{x} = (\boldsymbol{x}^{(v_1)}, \boldsymbol{x}^{(v_2)})$), a prediction function f_{v_i} is learned from each view respectively. Since the two views are sufficient, the learned prediction functions satisfy $f_{v_1}(\boldsymbol{x}^{(v_1)}) = f_{v_2}(\boldsymbol{x}^{(v_2)}) = y$, where y is the ground-truth label of $\boldsymbol{x}$. Intuitively, some projections in these two views should have strong correlation with the ground-truth. For either view, there should exist at least one projection which is correlated strongly with the ground-truth, since otherwise this view could not be sufficient. Since the two sufficient views are conditionally independent given the class label, the most strongly correlated pair of projections should be in accordance with the ground-truth. Thus, if such highly correlated projections of these two views can be identified, they can help induce the labels of some unlabeled examples. With those additional labeled examples, two learners can be generated, and then they can be improved using the standard co-training routine, i.e., if the learners disagree on an unlabeled example, the learner which misclassifies this example will be taught by its peer. To identify the correlated projections, Zhou et al. [91] employed kernel canonical component analysis (KCCA) [38] to find two sets of basis vectors in the feature space, one for each view, such that after projecting the two views onto the corresponding sets of basis vectors, the correlation between the projected views is maximized. Here the correlation strength (λ) of the projections are also given by KCCA. Instead of considering only the highest correlated projections, they used top m projections with the m-highest correlation strength. Finally, by linearly combining similarity in each projection, they computed confidence ρ_i of each unlabeled example $\boldsymbol{x}_i$ as being with the same label as that of the single labeled positive example $\boldsymbol{x}_0$, as shown in Eq. 4,

$$\rho_i = \sum_{j=1}^{m} \lambda_j sim_{i,j} \ , \tag{4}$$

where

$$sim_{i,j} = \exp\left(-d^2\left(P_j(\boldsymbol{x}_i^{(v_1)}), P_j(\boldsymbol{x}_0^{(v_1)})\right)\right) + \exp\left(-d^2\left(P_j(\boldsymbol{x}_i^{(v_2)}), P_j(\boldsymbol{x}_0^{(v_2)})\right)\right) \tag{5}$$

and $d(a, b)$ measures the Euclidian distance between a and b.

Thus, several unlabeled examples with the highest and lowest confidence values can be picked out, respectively, and used as extra positive and negative examples. Based on this augmented labeled training set, standard co-training can be employed for semi-supervised learning. Again, intuitively, when the two learners disagree on an unlabeled example, the learner which misclassifies this example will be taught by its peer. Such kind of method has been applied to content-based image retrieval [91] where there is only one example image in the first round of query.

3.2. Learning with Single View Multiple Classifiers

In most real-world applications the data sets have only one attribute set rather than two. So the effectiveness and usefulness of the standard co-training is limited. To take advantage of the interaction between learners when exploiting unlabeled data, methods that do not rely on the existence of two views have been developed.

A straightforward way to tackle this problem is to partition the attribute sets into two disjoint sets, and conduct standard co-training based on the manually generated views. Nigam and Ghani [55] empirically studied the performance of standard co-training algorithm in this case. The experimental results suggested that when the attribute set is sufficiently large, randomly splitting the attributes and then conducting standard co-training may lead to a good performance. However, many applications are not described by a large number of attribute, and co-training on randomly partitioned views is not always effective. Thus, a better way is to design single-view methods that can exploit the interaction between multiple learners rather than tailoring the data sets for standard two-view co-training.

Goldman and Zhou [36] proposed a method that does not rely on two views. They employed different learning algorithms to train the two classifiers, respectively. It is required that each classifier is able to partition the instance space into a number of equivalence classes. In order to identify which unlabeled example to label, and to decide how to make the prediction when two classifiers disagree, ten-fold cross validations are executed such that the confidences of the two classifiers as well as the confidences of the equivalence classes that contain the concerned instance can be compared. Later, this idea was extended to involving more learning algorithms [81]. Note that although [36] does not rely on the existence of two views, it requires special learning algorithms to construct classifiers. This prevents its application to other kinds of learning algorithms.

Zhou and Li [88] proposed the *tri-training* method, which requires neither the existence of two views nor special learning algorithms, thus it can be applied to more real-world problems. In contrast to the previous studies [13, 36, 55], tri-training attempts to exploit unlabeled data using three classifiers. Such a setting tackles the problem of determining how to efficiently select most confidently predicted unlabeled examples to label and produces final hypothesis. Note that the essential of tri-training is extensible for more than three classifiers, which will be introduced later. The use of more classifiers also provides a chance to employ ensemble learning techniques [84] to improve the performance of semi-supervise learning.

Generally, tri-training works in the following way. First, three classifiers are initially trained from the original labeled data. Unlike [36], tri-training uses the same learning algorithm (e.g., C4.5 decision tree) to generate the three classifiers. In order to make the three classifiers diverse, the original labeled example set is bootstrap sampled [31] to produce three perturbed training sets, on each of which a classifier is then generated. The generation of the initial classifiers is similar to training an ensemble from the labeled example set using Bagging [17]. Then, intuitively, in each tri-training round, if two classifiers agree on the labeling of an unlabeled example while the third one disagrees, then these two classifiers will teach the third classifier on this example. Finally, three classifiers are combined by majority voting. Note that the "majority teaches minority" strategy serves as an *implicit* confidence measurement, which avoids the use of complicated time-

consuming approaches to explicitly measure the predictive confidence, and hence the training process is efficient. Such an implicit measurement, however, might not be as accurate as an explicit estimation, since sometimes "minority holds the truth". Thus, some additional control is needed to reduce the negative influence of incorrectly labeled examples. Zhou and Li [88] analytically showed that the negative influence can be compensated if the amount of newly labeled examples is sufficient under certain conditions.

Inspired by [36], Zhou and Li [88] derived the criterion based on theoretical results of learning from noisy examples [5]. In detail, if a sequence σ of m samples is drawn, where the sample size m satisfies Eq. 6,

$$m \geq \frac{2}{\epsilon^2 \left(1-2\eta\right)^2} \ln\left(\frac{2N}{\delta}\right) , \tag{6}$$

where ϵ is the hypothesis worst-case classification error rate, η (< 0.5) is an upper bound on the classification noise rate, N is the number of hypotheses, and δ is the confidence, then a hypothesis H_i that minimizes disagreement with σ will have the PAC property, i.e.,

$$\Pr\left[d(H_i, H^*) \geq \epsilon\right] \leq \delta , \tag{7}$$

where $d(,)$ is the sum over the probability of elements from the symmetric difference between the two hypothesis sets H_i and H^* (the ground-truth). Let $c = 2\mu \ln\left(\frac{2N}{\delta}\right)$ where μ makes Eq. 6 hold equality. After some reforming, Eq. 6 becomes Eq. 8.

$$u = \frac{c}{\epsilon^2} = m\left(1-2\eta\right)^2 \tag{8}$$

For each classifier, in order to keep improving the performance in the training process, the u value of the current round should be greater than that in its previous round. Let L^t and L^{t-1} denote the newly labeled data set of a classifier in the t-th round and $(t-1)$-th round, respectively. Then the training sets for this classifier in the t-th round and $(t-1)$-th round are $L \cup L^t$ of the size of $|L \cup L^t|$ and $L \cup L^{t-1}$ of the size of $|L \cup L^{t-1}|$, respectively. Let $\check{e}^t$ and $\check{e}^{t-1}$ denote the upper bound of the classification error rate of the hypothesis derived from the combination of the other two classifiers in the t-th round and $(t-1)$-th round, respectively. By comparing Eq. 8 in the subsequent rounds, the condition that a classifier's performance can be improved through the refinement in the t-th round is shown as

$$0 < \frac{\check{e}^t}{\check{e}^{t-1}} < \frac{|L^{t-1}|}{|L^t|} < 1 . \tag{9}$$

Such a condition is used as the stopping criterion for tri-training algorithm. If none of the three classifiers satisfies the condition shown in Eq. 9 in the t-th round, tri-training stops and outputs the learned classifiers. Note that Eq. 9 sometimes could not be satisfied, due to the fact that $|L^t|$ may be far bigger than $|L^{t-1}|$ instead of $\check{e}^t$ being higher than $\check{e}^{t-1}$. When this happens, in order not to stop training before the error rate of the classifier becomes low, L^t are randomly subsampled to size s according to Eq. 10 to make Eq. 9 hold again,

$$s = \left\lceil \frac{\check{e}^{t-1}|L^{t-1}|}{\check{e}^t} - 1 \right\rceil , \tag{10}$$

Table 1. Pseudo-code describing the tri-training algorithm [88]

tri-training(L, U, $Learn$)
Input: L: Original labeled example set
U: Unlabeled example set
$Learn$: Learning algorithm
for $i \in \{1..3\}$ **do**
$S_i \leftarrow BootstrapSample(L)$
$h_i \leftarrow Learn(S_i)$
$e_i^{'} \leftarrow .5$; $l_i^{'} \leftarrow 0$
end of for
repeat until none of $h_i\,(i \in \{1..3\})$ changes
for $i \in \{1..3\}$ **do**
$L_i \leftarrow \emptyset$; $update_i \leftarrow FALSE$
$e_i \leftarrow MeasureError(h_j \& h_k)\ (j, k \neq i)$
if $(e_i < e_i^{'})$ % otherwise Eq. 9 is violated
then for every $x \in U$ **do**
if $h_j(x) = h_k(x)\ (j, k \neq i)$
then $L_i \leftarrow L_i \cup \{(x, h_j(x))\}$
end of for
if $(l_i^{'} = 0)$ % h_i has not been updated before
then $l_i^{'} \leftarrow \frac{e_i}{e_i^{'} - e_i} + 1$ % refer Eq. 11
if $(l_i^{'} < |L_i|)$ % otherwise Eq. 9 is violated
then if $(e_i|L_i| < e_i^{'} l_i^{'})$ % otherwise Eq. 9 is violated
then $update_i \leftarrow TRUE$
else if $l_i^{'} > \frac{e_i}{e_i^{'} - e_i}$ % refer Eq. 11
then $L_i \leftarrow Subsample(L_i,\ \frac{e_i^{'} l_i^{'}}{e_i} - 1\ )$
% refer Eq. 10
$update_i \leftarrow TRUE$
end of for
for $i \in \{1..3\}$ **do**
if $update_i = TRUE$
then $h_i \leftarrow Learn(L \cup L_i)$; $e_i^{'} \leftarrow e_i$; $l_i^{'} \leftarrow |L_i|$
end of for
end of repeat
Output: $h(x) \leftarrow \arg\max_{y \in label} \sum_{i:\, h_i(x) = y} 1$

where L^{t-1} should satisfy Eq. 11 such that the size of L^t after subsampling, i.e. s, is still bigger than $|L^{t-1}|$.

$$|L^{t-1}| > \frac{\check{e}^t}{\check{e}^{t-1} - \check{e}^t} \tag{11}$$

The pseudo-code of tri-training algorithm is shown in Table 1. The function $MeasureError(h_j \& h_k)$ estimates the classification error rate of the hypothesis derived from the combination of h_j and h_k. The function $Subsample(L^t, s)$ randomly removes $|L^t| - s$ examples from L^t where s is computed according to Eq. 10.

As mentioned before, the essential of tri-training is extensible for much more

classifiers. Li and Zhou [48] proposed the *Co-Forest* algorithm, which extended tri-training to the collaboration of many classifiers in training process. By using an ensemble of classifiers, several immediate benefits can be achieved. First, for each classifier h_i, its concomitant ensemble H_i, i.e., ensemble of all the other classifies, is used to label several unlabeled examples for this classifier. As an ensemble of classifiers usually achieves a better generalization than a single classifier [73, 84], the labeling of unlabeled data becomes more reliable, and thus, classifier refinement using these high quality newly labeled data can lead to better performances. Second, without employing sophisticated and time-consuming methods to estimate the predictive confidence, the use of multiple classifiers enables an efficient estimation of the predictive confidence. Such an explicit estimate can be further exploited to guide the use of the corresponding unlabeled example in the training stage.

Although the advantages of using multiple classifiers seems promising, a straightforward extension suffers from a problem, that is, the "majority teaches minority" process hurts the generalization ability of an ensemble of classifiers. It is known that the diversity of the learners is a key of a good ensemble [84]. During the "majority teaches minority" process, the behaviors of the learners will become more and more similar, and thus the diversity of the learners decreases rapidly.

To address this problem, Li and Zhou [48] proposed to inject certain randomness into the semi-supervised learning process. They employed two strategies. First, randomness is injected into the classifier learning process, such that any two classifiers in the ensemble can be diverse even when their training data are similar. For implementation convenience, they used Random Forest [18] to construct the ensemble. Second, randomness is injected into the unlabeled data selection process. Instead of directly selecting the highly confident unlabeled examples, some candidate examples for labeling are randomly subsampled from the original unlabeled training set to meet a condition similar as Eq. 9, and highly confident examples in the candidate pool are then selected and labeled. Thus, the learners will encounter different training sets in each round. Such a strategy is helpful not only for the diversity, but also for reducing the chance of being trapped into local minima, just like a similar strategy adopted in [13].

3.3. Learning With Single View Multiple Regressors

Previous studies on semi-supervised learning mainly focus on classification tasks. Although regression is almost as important as classification, semi-supervised regression has rarely been studied. One reason is that for real valued labels the cluster assumption is not applicable. Although methods based on manifold assumption can be extended to regression, as pointed out by [92], these methods are essentially transductive instead of really semi-supervised since they assume that the unlabeled examples are exactly test examples.

Zhou and Li [87] first proposed a disagreement-based semi-supervised regression approach COREG, which employs two kNN regressors [27] to conduct the data labeling as well as the predictive confidence estimation. The use of kNN regressors as base learners enables efficient refinement of a regressor based on the newly labeled data from its peer regressor since this lazy learning approach does not hold a separate training phase when updating the current regressor.

Moreover, kNN regressor can be easily coupled with the predictive confidence estimation method.

In order to choose appropriate unlabeled examples for labeling in semi-supervised regression, the labeling confidence should be estimated such that the most confidently labeled example can be identified. In classification this is relatively straightforward because when making classifications, many classifiers (e.g. a Naïve Bayes classifier) can also provide an estimated probability (or an approximation) for the classification. Therefore, the predictive confidence can be estimated through consulting the probabilities of the unlabeled examples being labeled to different classes. Unfortunately, in regression there is no such estimated probability that can be used directly. This is because in contrast to classification where the number of labels to be predicted is finite, the possible predictions in regression are infinite. Therefore, Zhou and Li [87] proposed a predictive confidence estimation criterion for disagreement-based semi-supervised learning method.

Intuitively, the most confidently labeled example of a regressor should decrease most the error of the regressor on the labeled example set, if the the most confidently labeled example is utilized. In other words, the most confidently labeled example should be the one which makes the regressor most *consistent* with the labeled example set. Thus, the mean squared error (MSE) of the regressor on the labeled example set can be evaluated first. Then, the MSE of the regressor utilizing the information provided by a newly labeled example $(\boldsymbol{x}_u, \hat{y}_u)$ can be evaluated on the labeled example set, where the real-valued label $\hat{y}_u$ of the unlabeled instance $\boldsymbol{x}_u$ is generated by the regressor. Let Δ_u denote the result of subtracting the latter MSE from the former MSE. Note that the number of Δ_u to be estimated equals to the number of unlabeled examples. Finally, $(\boldsymbol{x}_u, \hat{y}_u)$ associated with the biggest positive Δ_u can be regarded as the most confidently labeled example.

To avoid repeatedly measuring the MSE of the kNN regressor on the whole labeled training set in each iteration, approximation is employed to compute the MSE based on only the k-nearest labeled examples of an unlabeled instance. Let Ω_u denote the set of its k-nearest labeled examples of $\boldsymbol{x}_u$, then the most confidently labeled example $\tilde{\boldsymbol{x}}$ is identified through maximizing the value of $\Delta_{\boldsymbol{x}_u}$ in Eq. 12,

$$\Delta_{\boldsymbol{x}_u} = \sum_{\boldsymbol{x}_i \in \Omega_u} \left(\left(y_i - h\left(\boldsymbol{x}_i\right)\right)^2 - \left(y_i - h'\left(\boldsymbol{x}_i\right)\right)^2 \right), \tag{12}$$

where h denotes the original regressor while h' denotes the refined regressor which has utilized the information provided by $(\boldsymbol{x}_u, \hat{y}_u)$, $\hat{y}_u = h(\boldsymbol{x}_u)$.

Another important aspect in COREG is the diversity between two regressors. Note that the labeling of an unlabeled example is obtained by averaging the real-valued labels of its k-nearest neighbors in the labeled training set. As only a few examples are labeled at early stage, the labeling of unlabeled data can be noisy. Zhou and Li [87, 89] showed that using diverse regressors can help to reduce the negative influence of the noisy newly labeled data. Since kNN regressor is used as the base learner, a natural way to make the kNN regressors different is to enable them to identify different vicinities, which can be achieved by manipulating the parameter settings of the kNN regressors. In [87], Minkowsky distances of different orders were used to generate two diverse kNN regressors. This strategy was extended to a more general case, i.e., achieving diversity by using different distance metrics and/or different number of neighbors identified for a given

Table 2. Pseudo-code describing the COREG algorithm [89]

ALGORITHM: COREG

INPUT: labeled example set L, unlabeled example set U,
maximum number of learning iterations T,
number of nearest neighbors k_1, k_2
distance metrics D_1, D_2

PROCESS:

$L_1 \leftarrow L$; $L_2 \leftarrow L$
Create pool U' of size s by randomly picking examples from U
$h_1 \leftarrow kNN(L_1, k_1, D_1)$; $h_2 \leftarrow kNN(L_2, k_2, D_2)$
Repeat for T rounds:
  for $j \in \{1, 2\}$ **do**
    for each $\boldsymbol{x}_u \in U'$ **do**
      $\Omega_u \leftarrow Neighbors\left(\boldsymbol{x}_u, L_j, k_j, D_j\right)$
      $\hat{y}_u \leftarrow h_j(\mathbf{x}_u)$
      $h'_j \leftarrow kNN(L_j \cup \{(\boldsymbol{x}_u, \hat{y}_u)\}, k_j, D_j)$
      $\Delta_{\boldsymbol{x}_u} \leftarrow \sum_{\boldsymbol{x}_i \in \Omega_u} \left(\mathbf{y}_i - h_j\left(\boldsymbol{x}_i\right)\right)^2 - \left(y_i - h'_j\left(\boldsymbol{x}_i\right)\right)^2$
    end of for
    if there exists an $\Delta_{\boldsymbol{x}_u} > 0$
    then $\tilde{\boldsymbol{x}}_j \leftarrow \arg\max_{\boldsymbol{x}_u \in U'} \Delta_{\boldsymbol{x}_u}$; $\tilde{y}_j \leftarrow h_j(\tilde{\boldsymbol{x}}_j)$
      $\pi_j \leftarrow \{(\tilde{\boldsymbol{x}}_j, \tilde{y}_j)\}$; $U' \leftarrow U' - \tilde{\boldsymbol{x}}_j$
    else $\pi_j \leftarrow \emptyset$
  end of for
  $L_1 \leftarrow L_1 \cup \pi_2$; $L_2 \leftarrow L_2 \cup \pi_1$
  if neither of L_1 and L_2 changes **then exit**
  else
    $h_1 \leftarrow kNN(L_1, k_1, D_1)$; $h_2 \leftarrow kNN(L_2, k_2, D_2)$
    Replenish U' to size s by randomly picking examples from U
end of Repeat

OUTPUT: regressor $h^*(\boldsymbol{x}) \leftarrow \frac{1}{2}\left(h_1(\boldsymbol{x}) + h_2(\boldsymbol{x})\right)$

example [89]. Additionally, such a setting also brings another benefit, that is, since it is usually difficult to decide the appropriate parameter settings for kNN regressor for a specific task, combining the regressors with different parameter settings can obtain somewhat complementary effect.

The pseudo code of COREG is shown in Table 2, where $kNN(L_j, k_j, D_j)$ is a function that returns a kNN regressor on the labeled training set L_j, whose k value is k_j and distance metric is D_j. The learning process stops when the maximum number of learning iterations, i.e. T, is reached, or there is no unlabeled example which is capable of reducing the MSE of any of the regressors on the labeled example set. A pool of unlabeled examples smaller than U is used, as what was used in [13]. Note that in each iteration the unlabeled example chosen by h_1 won't be chosen by h_2, which is an extra mechanism for encouraging the diversity of the regressors. Thus, even when h_1 and h_2 are similar, the examples they labeled for each other will still be different.

It is evident that the method introduced in this section is closely related to those introduced in Section 3.2. The key is to generate multiple diverse learners and then try to exploit their disagreements on unlabeled data to implement the

performance boost. Actually, "learning with multiple views" (Section 3.1) is a special case which uses multiple views to help generate multiple diverse learners.

3.4. The Combination with Active Learning

In disagreement-based semi-supervised learning approaches, the unlabeled examples that are labeled for a learner are examples on which most other learners agree but the concerned learner disagrees. If all learners disagree on the labeling of an unlabeled example, this example is simply neglected. However, it is highly probable that such an example is not able to be learned well by the learning system itself. As mentioned in Section 1, active learning is another major technique of learning with labeled and unlabeled data. It actively selects some informative unlabeled examples and queries their labels from an *oracle* independent to the learning system. It is evident that the unlabeled example on which all learners disagree is a good candidate to query.

Zhou et al. [85, 86] proposed a disagreement-based active semi-supervised learning method SSAIR for content-based image retrieval. After obtaining a small number of labeled images from *relevance feedback*, they constructed two learners using the labeled images. Each learner attempts to assign a rank to all images in the imagebase. The smaller the rank, the higher the chance that the concerned image is relevant to the user query. The two most confident irrelevant images of each learner are passed to the other learner as negative examples. Such a process is repeated and the two learners are refined. In previous relevance feedback methods, the user randomly picks some images from the retrieval result to give feedback. Zhou et al. [85, 86] thought that letting the user to give feedback on images that have been learned well is not helpful to improve the performance. So, instead of passively waiting user feedback, they actively prepared a pool of images for user to give feedback. The pool contains images on which the two learners are with contradict predictions but similar confidences, and images on which the two learners are both with low confidences. Thus, in each round of relevance feedback, both semi-supervised learning and active learning are executed to exploit the images existing in imagebase to the most. It is evident that although the combination of the disagreement-based semi-supervised learning and active learning is simple, it provides a good support to the interesting *active semi-supervised relevance feedback* scheme which is useful in information retrieval tasks.

4. Theoretical Foundations for Disagreement-Based Semi-Supervised Learning

Early theoretical analyses of the disagreement-based semi-supervised learning approaches mainly focus on the case where there exists two views.

Blum and Mitchell [13] analyzed the effectiveness of co-training. Let X_1 and X_2 denote the two sufficient and redundant views of the input space X, and hence an instance can be represented by $\boldsymbol{x} = (\boldsymbol{x}_1, \boldsymbol{x}_2) \in X_1 \times X_2$. Assume that $f = (f_1, f_2) \in C_1 \times C_2$ is a target function defined over X and C_1, C_2 are concept classes defined over X_1, X_2 respectively, and then $f(\boldsymbol{x}) = f_1(\boldsymbol{x}_1) = f_2(\boldsymbol{x}_2) = y$ should be held due to the sufficiency of the two views, where y is the ground-truth label of $\boldsymbol{x}$. Therefore, they defined *compatibility* between target

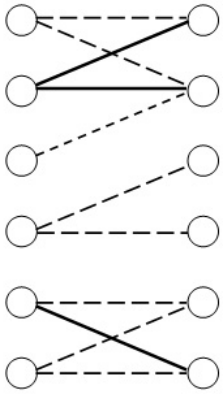

Fig. 2. The bipartite graph for instance distribution. Plot based on a similar figure in [13].

function $f = (f_1, f_2)$ and the unknown data distribution $\mathcal{D}$, based on which co-training is analyzed. Here, $f = (f_1, f_2)$ is compatible with $\mathcal{D}$ means that $\mathcal{D}$ assigns probability zero to any instance $(\boldsymbol{x}_1, \boldsymbol{x}_2)$ such that $f_1(\boldsymbol{x}_1) \neq f_2(\boldsymbol{x}_2)$.

If the "compatibility" is satisfied, as [13] pointed out, even if the concept classes C_1 and C_2 are large and complex (i.e., in high VC-dimension), the set of target concepts that is compatible with $\mathcal{D}$ could be smaller and simpler. Therefore, unlabeled data can be used to verify which are the compatible target concepts, and hence lead to the reduction of the number of labeled data needed in learning. They illustrated this idea by an bipartite graph shown in Fig. 2. In this graph, vertices in left hand side and right hand side denote the instances in X_1 and X_2, respectively. For any pair of vertices on each side of the graph, there exists an edge on between if and only if the corresponding instance $(\boldsymbol{x}_1, \boldsymbol{x}_2)$ can be drawn with non-zero probability under the distribution $\mathcal{D}$. The solid edges denote the instances observed in the finite training set. Obviously, under this representation, the concepts that are compatible with $\mathcal{D}$ correspond to the graph partitions without any cross-edges. The instances in the same connected component share the same label, and only a labeled example is required to determine the labeling of this component. The value of unlabeled data is their assistance for identifying the connected components of the graph, and in fact identifying the distribution $\mathcal{D}$.

Based on this bi-partite representation, Blum and Mitchell [13] analytically showed that "If C_2 is learnable in the PAC model with classification noise, and if the conditional independence assumption is satisfied, then (C_1, C_2) is learnable in Co-training model from unlabeled data only, given an initial weak-useful predictor $h(\boldsymbol{x}_1)$". This is a very strong conclusion which implies that if the two views are conditionally independent, the predictive accuracy of an initial weak learner can be boosted to arbitrarily high with unlabeled data using co-training.

Later, Dasgupta et al. [28] analyzed the generalization bound for standard co-training. Let S be a set of i.i.d samples. For any statement $\Phi[s]$, let $S(\Phi)$ be a subset of S that satisfies Φ. For two statements Φ and Ψ, the empirical estimate $\hat{P}(\Phi|\Psi) = |S(\Phi \wedge \Psi)|/|S(\Psi)|$. They assumed that the data is drawn from some distribution over triples $\langle \boldsymbol{x}_1, y, \boldsymbol{x}_2 \rangle$ with $\boldsymbol{x}_1 \in \mathcal{X}_1$ and $\boldsymbol{x}_2 \in \mathcal{X}_2$, and

$P(\boldsymbol{x}_1|y,\boldsymbol{x}_2) = P(\boldsymbol{x}_1|y)$ and $P(\boldsymbol{x}_2|y,\boldsymbol{x}_1) = P(\boldsymbol{x}_2|y)$; in other words, the data has two views that are independent given the class label. Assume that there are k different classes, and if a learner h fails to classify $\boldsymbol{x}$ into the k classes, then $h(\boldsymbol{x}) = \perp$. Let $|h|$ denote a complex measurement of h, and h_1 and h_2 denote the learner constructed in $\mathcal{X}_1$ and $\mathcal{X}_2$, respectively. Then, they showed that with probability at least $1-\delta$ over the choice of S, for all pairs of learners h_1 and h_2 such that $\gamma_i(h_1,h_2,\delta/2) > 0$ and $b_i(h_1,h_2,\delta/2) \leq (k-1)/k$, the following inequity holds,

$$error(h_1) \leq \left(\hat{P}(h_1 \neq \perp) - \epsilon(|h_1|,\delta/2)\right) \max_j b_j(h_1,h_2,\delta/2) + \frac{k-1}{k}\left(\hat{P}(h_1 = \perp) + \epsilon(|h_1|,\delta/2)\right) \tag{13}$$

where

$$\epsilon(k,\delta) = \sqrt{\frac{k\ln 2 + \ln 2/\delta}{2|S|}} \tag{14}$$

$$b_i(h_1,h_2,\delta) = \frac{1}{\gamma_i(h_1,h_2,\delta)}\left(\hat{P}(h_1 \neq i | h_2 = i, h_1 \neq \perp) + \epsilon_i(h_1,h_2,\delta)\right) \tag{15}$$

$$\epsilon_i(h_1,h_2,\delta) = \sqrt{\frac{(\ln 2)(|h_1|+|h_2|) + \ln(2k/\delta)}{2|S(h_2 = i, h_1 \neq \perp)|}} \tag{16}$$

$$\gamma_i(h_1,h_2,\delta) = \hat{P}(h_1 = i | h_2 = i, h_1 \neq \perp) - \hat{P}(h_1 \neq i | h_2 = i, h_1 \neq \perp) - 2\epsilon_i(h_1,h_2,\delta) \tag{17}$$

The result of Dasgupta et al. [28] shows that when there are two sufficient and redundant views which are conditionally independent given the class label, the generalization error of co-training is upper-bounded by the disagreement between the two classifiers. This suggests that a better learning performance can be obtained if the disagreement can be exploited in a better way.

Note that in the analyses in [13] and [28], it was assumed that there exist two sufficient and redundant views that are conditionally independent given the class label. Since such a strong requirement is not often satisfied, analyses under more realistic assumptions are desired. Balcan et al. [6] pointed out that if a PAC learner can be obtained on each view, the conditional independence assumption or even weak independent assumption [2] is unnecessary; a weaker assumption of "expansion" of the underlying data distribution is sufficient for iterative co-training to succeed. They consider that the learning algorithm used in each view is confident about being positive and is able to learn from positive examples only, and the "expansion" is defined as follows: Let X^+ denote the positive region and $\mathcal{D}^+$ denote the distribution over X^+. For $S_1 \subseteq X_1$ and $S_2 \subseteq X_2$, let $\mathbf{S}_i$ be the event that an instance $(\boldsymbol{x}_1,\boldsymbol{x}_2)$ has $\boldsymbol{x}_i \in S_i$ $(i = 1,2)$. Let $P(\mathbf{S}_1 \wedge \mathbf{S}_2)$ denote the probability on examples for being confident on both views, and $P(\mathbf{S}_1 \oplus \mathbf{S}_2)$ denote the probability on examples for being confident on only one view. Let

$H_i \cap X_i^+ = \{h \cap X_i^+ : h \in H_i\}$, where H_i $(i = 1, 2)$ is hypothesis class. If Eq. 18 holds for any $S_1 \subseteq X_1$ and $S_2 \subseteq X_2$, then $\mathcal{D}^+$ is ϵ-expanding; if Eq. 18 holds for any $S_1 \subseteq H_1 \cap X_1$ and $S_2 \subseteq H_2 \cap X_2$, then $\mathcal{D}^+$ is ϵ-expanding with respect to hypothesis class $H_1 \times H_2$.

$$P(\mathbf{S}_1 \oplus \mathbf{S}_2) \geq \epsilon \min\left(P(\mathbf{S}_1 \wedge \mathbf{S}_2), P(\bar{\mathbf{S}}_1 \wedge \bar{\mathbf{S}}_2)\right) \tag{18}$$

If some data distribution satisfies the expansion assumption, with a small confidence set S_j of the hypothesis of view j $(j = 1, 2)$, the iterative co-training can succeed to achieve a classifier whose error rate is smaller than ϵ with a large probability.

All the previous theoretical studies investigated the standard two-view co-training. Theoretical foundation of other disagreement-based semi-supervised learning approaches, in particular, those work on a single view, has not been established, although the effectiveness of those approaches have been empirically verified. Wang and Zhou [71] presented a theoretical study for those approaches. Let $\mathcal{H}$ denote the hypothesis space and $\mathcal{D}$ is the data distribution generated by the ground-truth hypothesis $h^* \in \mathcal{H}$. Let $d(h^i, h^*)$ denote the difference between the two classifier h^i and h^*, which can be measured by $\Pr_{x \in \mathcal{D}}[h^i(x) \neq h^*(x)]$. Let h_1^i and h_2^i denote the classifiers in i-th round of the iterative co-training process. Then, their main result is summarized in Theorem 1 shown as follows.

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 classifiers h_1^0 and h_2^0 whose upper bound of the generalization error is $a_0 < 0.5$ and $b_0 < 0.5$ with high probability (more than $1-\delta$) in the PAC model, respectively, 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*

$$\Pr[d(h_2^1, h^*) \geq b_1] \leq \delta. \tag{19}$$

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

Such a theorem suggests that the key for the disagreement-based approaches to succeed is the large difference between the learners, which explains the reason why the disagreement-based approaches still work well even when there are no two views. Note that in contrast to all previous studies which assumed that data is drawn from some distribution over triples $\langle \boldsymbol{x}_1, y, \boldsymbol{x}_2 \rangle$ (that is, the data has two views), the above theorem does not assume that data is drawn from distribution over two views. Actually, from Theorem 1 we can know that the existence of two views is a sufficient condition instead of necessary condition for disagreement-based approaches. This is because when there are two sufficient and redundant views, the learners trained from the two views respectively are of course diverse, and so the disagreement-based learning process can succeed. When there are no two views, it is also possible to get two diverse learners, and thus disagreement-based approaches are also able to succeed. It is worth mentioning that all previous studies, either theoretical or algorithmic, tried to maximize the consensus among the learners; in other words, they always tried to minimize the error for labeled examples and maximize the agreement for

unlabeled examples, but never revealed that keeping a large disagreement among the learners is a necessary condition for co-training to proceed.

Moreover, Wang and Zhou [71] analyzed the reason why the performance of the disagreement-based approaches could not be improved further after a number of training rounds. Such a problem is frequently encountered in many practical applications of the disagreement-based approaches but could not be explained by previous theoretical results. Based on Theorem 1, Wang and Zhou [71] showed that as the learning process continues, the learners will become more and more similar, and therefore, the required diversity could not be met and the learners could not improve each other further. Based on this recognition, a preliminary method for roughly estimating the appropriate iteration to terminate the learning process was proposed.

Section 3.4 introduced that the combination of disagreement-based semi-supervised learning with active learning can lead to good performance. Recently, Wang and Zhou [72] analyzed this situation and got the result in Theorem 2.

Theorem 2 *For data distribution $\mathcal{D}$ α-expanding with respect to hypothesis class $\mathcal{H}_1 \times \mathcal{H}_2$, let ϵ and δ denote the final desired accuracy and confidence parameters. If $s = \lceil \frac{\log \frac{\alpha}{8\epsilon}}{\log \frac{1}{C}} \rceil$, $m_0 = \frac{1}{L}(4V \log(\frac{1}{L}) + 2\log(\frac{8(s+1)}{\delta}))$ and $m_i = \frac{16}{\alpha}(4V \log(\frac{16}{\alpha}) + 2\log(\frac{8(s+1)}{\delta}))$ $(i = 1, 2, \cdots,)$, a classifier will be generated with error rate no more than ϵ with probability $1 - \delta$, according to a similar approach in [85].*

Here, $V = \max[VC(\mathcal{H}_1), VC(\mathcal{H}_2)]$ where $VC(\mathcal{H})$ denotes the VC-dimension of the hypothesis class $\mathcal{H}$, constant $C = \frac{\alpha/4+1/\alpha}{1+1/\alpha}$ and constant $L = \min[\frac{\alpha}{16}, \frac{1}{16L_1L_2}]$.

This theorem suggests that under assumption of α-expansion on the hypothesis class $\mathcal{H}_1 \times \mathcal{H}_2$, the sample complexity can be exponentially reduced by combining disagreement-based semi-supervised learning with active learning in contrast to pure disagreement-based semi-supervised learning. This is the first theoretical analysis on the combination of semi-supervised learning with active learning, which also contains the first analysis on multi-view active learning.

Both Theorem 1 and Theorem 2 provide theoretical explanations to the effectiveness of general disagreement-based semi-supervised learning approaches whose effectiveness has been empirically verified in practice. Although some strong assumptions are still required in the analyses, the results serve as an important step towards the establishment of the whole theoretical foundation for the disagreement-based learning framework. However, note that all the current theoretical analyses are on the use of two learners, while theoretical analysis on disagreement-based semi-supervised learning with more than two learners remains an open problem.

5. Applications to Real-World Tasks

Disagreement-based semi-supervised learning paradigm has been successfully applied to many real-world tasks, particularly in natural language processing. In fact, in the middle of 1990s, it was accepted that constructing prediction models based on the different attribute sets of the problem may help to achieve a better result. Yarowsky [75] constructed a word sense classifier using the local context of the word and a classifier based on the senses of other occurrences of that word

in the same document for word sense disambiguation. Riloff and Jones [59] considered both the noun phrase itself and the linguistic context in which the noun phrase appears for classifying noun phrase for geographic locations. Collins and Singer [24] utilized both the spelling of the entity and the context in which the entity appears for named entity classification.

Pierce and Cardie [58] applied standard co-training to conduct named entity identification. They treated the current word and k immediate words before it as the first view, and similarly, the current word and k immediate words after it as the second view. Based on these two views, standard co-training were directly applied with some necessary adaptations to multi-class classification. By utilizing unlabeled data with co-training, the identification error rate reduced by 36% compared to the identification using only the labeled data. Sarkar [62] decomposed statistical parser into two sequentially related probabilistic models. The first model, which is called tagging probability model, is responsible to select the most likely trees for each word by examining the local context, while the second model, which is called parsing probability model, is responsible for attaching the selected trees together to provide a consistent bracketing of the sentences. In the learning process, these two models employ a disagreement-based approach to exploit the unlabeled examples, where each model uses its most confident information about the prediction to help the other model to reduce the uncertainty in statistical parsing, and hence achieves a better performance in terms of both precision and recall. Later, Steedman et al. [68] solved this problem from a different perspective. Unlike [62], they used the two different statistical parsers for co-training. In the training process, each parser assigns scores to unlabeled sentences that have been parsed by itself, using a scoring function to indicate the confidence of the parse results. Then, the parser passes the parsed sentences with the top scores to the other parser. They empirically showed that such a method could also improve the performance of statistical parsing. Hwa et al. [41] combined disagreement-based semi-supervised learning with active learning in statistical parsing, where each learner teaches the other learner with its most confidently parsed sentences, while its peer learner queries the parse result for its least confidently parsed sentences from the user and feeds them to this learner. By applying such a method, the number of manually labeling can be greatly reduced.

In addition to natural language processing, disagreement-based semi-supervised learning paradigm has been applied to content-based image retrieval. Given a query image, a CBIR system is required to return the images in the imagebase that are relevant to the query image. Due to the semantic gap between the high-level image semantics and the low-level image features, relevance feedback [61] is usually employed to bridge the gap. Since it is usually infeasible for the user to provide many rounds of feedback, the number of images with the relevance judgement is insufficient to achieve a good performance. Thus, unlabeled images in the imagebase can be further exploited to improve the performance of CBIR based on semi-supervised learning, while CBIR itself becomes a good application of learning with labeled and unlabeled data [82]. In fact, semi-supervised learning in CBIR scenario has been studied in [30, 74].

Zhou et al. [85, 86] first applied a disagreement-based semi-supervised learning method to exploit unlabeled images in the imagebase of a CBIR system. The method actually combines semi-supervised learning with active learning (see Section 3.4 for details). This research also leads to a new user interface design in CBIR. As shown in Fig. 3, the region above the dark line displays the retrieved

Fig. 3. User interface of a prototype system [85]

images while the region below the dark line displays the pooled images for relevance feedback. This is quite different from common interface which provides retrieval results only. In common interfaces, which provide retrieval results only, user may label images that have already been learned well by the system. In contrast, in the new interface, the images for user to give feedback are the selected ones that will give the most helpful information to the system, and thus the retrieval effectiveness will be improved much more effectively and efficiently.

Exploiting unlabeled examples is more difficult in the first round of retrieval as only one labeled image, i.e., the user query, can be used. Such an extreme setting has not been studied before in the area of learning with unlabeled data. A recent work [91] (see Section 3.1 for details) showed that when the images are with textual annotations, improving retrieval performance using unlabeled images is still feasible, even in the initial retrieval. Such a goal is achieved by exploiting the correlations between visual features and textual annotations.

Disagreement-based approaches have also been applied to other real-world problems. For example, Kockelkorn et al. [43] applied several algorithms including transductive SVM and co-training to email answering, i.e., predicting which of several frequently used answers a user will choose to respond to an email, and found that the benefit of both transduction and co-training is greatest when only few labeled data are available. Li and Ogihara [49] applied a disagreement-based approach to machine failure prediction, where the failure data contain both machine and image information of a xerographic machine. Mavroeidis et al. [51] applied tri-training to email spam detection. They joined the ECML-PKDD 2006 Discovery Challenge and achieved a top five rank. Li and Zhou [48] applied Co-Forest to detect microcalcification clusters in mammograms for breast cancer diagnosis, and significantly reduced the false negative rate without increasing the false positive rate, after exploiting the undiagnosed samples. Li et al. [46] developed the SSRank algorithm for document retrieval, by using a traditional document retrieval method BM25 as one base learner and RankNet as another, and achieved good performance on both benchmark document retrieval data and real web search data.

6. Conclusion

During the past decade, many disagreement-based approaches have been proposed, many theoretical supports have been discovered, and many successful real-world applications have been reported. All of these make disagreement-based semi-supervised learning become an important paradigm for semi-supervised learning. This article provides a review on this topic.

Note that even when a learner is with very high confidence in labeling unlabeled examples for the other learners, it may still give incorrect labels. For standard co-training with sufficient and redundant views, such a classification noise can be regarded as random noise due to the fact that the two views are conditionally independent. Thus, the performance of co-training would not be affected much if the learners could adapt to random noise. For other approaches, especially those using single view multiple learners, the learners are correlated and thus, the noise in the newly labeled examples cannot be considered as random noise. The accumulation of such noise might seriously mislead the learned hypotheses. Li and Zhou [47] tried to identify and remove some potentially mislabeled examples using *data editing* before these newly labeled examples are used for learner refinement. Such an idea works well for self-training [55]; this suggests that this seems a promising way to tackle the noise accumulation problem for disagreement-based semi-supervised learning approaches. The combination with active learning may also be helpful for addressing this problem.

Given the sufficient and redundant views, the minimum number of labeled examples required for triggering a successful semi-supervised learning has been reduced to one [91]. There is no such study on the minimum number of the required labeled examples for other disagreement-based approaches. This problem is interesting because requiring a smaller number of labeled examples implies requiring fewer user intervene, which is important for many online applications.

Current semi-supervised learning approaches, including disagreement-based approaches, are not "safe". In other words, sometimes the exploitation of unlabeled data may lead to performance degeneration. Designing "safe" semi-supervised learning approaches is the holy grail of this field. Previous studies on semi-supervised learning almost neglect the fact that although there exist abundant or even unlimited unlabeled data, the computational and storage resource that can be used is generally not unlimited. *Budget semi-supervised learning* [90] is worth noting, where effective algorithms should be able to adjust behaviors considering the given resource budget.

Acknowledgments

The authors want to thank Wei Wang, Sheng-Jun Huang and Ju-Hua Hu for proof reading the article, and the anonymous reviewers for helpful comments. The authors were partially supported by the National Science Foundation of China (60635030, 60721002), the Jiangsu Science Foundation (BK2008018) and the Jiangsu 333 High-Level Talent Cultivation Program.

References

[1] N. Abe and H. Mamitsuka. Query learning strategies using boosting and bagging. In *Proceedings of the 15th International Conference on Machine Learning*, pages 1–9, Madison, WI, 1998.

[2] S. Abney. Bootstrapping. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 360–367, Philadelphia, PA, 2002.

[3] Y. Altun, I. Tsochantaridis, and T. Hofmann. Hidden markov support vector machines. In *Proceedings of the 20th International Conference on Machine Learning*, pages 3–10, Washington, DC, 2003.

[4] M. R. Amini and P. Gallinari. Semi-supervised learning with an imperfect supervisor. *Knowledge and Information Systems*, 8(4):385–413, 2005.

[5] D. Angluin and P. Laird. Learning from noisy examples. *Machine Learning*, 2(4):343–370, 1988.

[6] 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.

[7] M. Belkin and P. Niyogi. Semi-supervised learning on Riemannian manifolds. *Machine Learning*, 56(1-3):209–239, 2004.

[8] 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.

[9] 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.

[10] S. Bickel and T. Scheffer. Estimation of mixture models using co-EM. In *Proceedings of the 16th European Conference on Machine Learning*, pages 35–46, Porto, Portugal, 2005.

[11] A. Blum and S. Chawla. Learning from labeled and unlabeled data using graph mincuts. In *Proceedings of the 18th International Conference on Machine Learning*, pages 19–26, Williamston, MA, 2001.

[12] A. Blum, J. Lafferty, M. Rwebangira, and R. Reddy. Semi-supervised learning using randomized mincuts. In *Proceedings of the 21st International Conference on Machine Learning*, pages 13–20, Banff, Canada, 2004.

[13] 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.

[14] U. Brefeld, C. Büscher, and T. Scheffer. Multi-view hidden markov perceptrons. In *Proceedings of the GI Workshops*, pages 134–138, Saarbrücken, Germany, 2005.

[15] U. Brefeld and T. Scheffer. Co-EM support vector learning. In *Proceedings of the 21st International Conference on Machine Learning*, Banff, Canada, 2004.

[16] U. Brefeld and T. Scheffer. Semi-supervised learning for structured output variables. In *Proceedings of the 23rd International Conference on Machine Learning*, pages 145–152, Pittsburgh, PA, 2006.

[17] L. Breiman. Bagging predictors. *Machine Learning*, 24(2):123–140, 1996.

[18] L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

[19] M. A. Carreira-Perpinan and R. S. Zemel. Proximity graphs for clustering and manifold learning. In L. K. Saul, Y. Weiss, and L. Bottou, editors, *Advances in Neural Information Processing Systems 17*. MIT Press, Cambridge, MA, 2005.

[20] O. Chapelle, M. Chi, and A. Zien. A continuation method for semi-supervised SVMs. In *Proceedings of the 23rd International Conference on Machine Learning*, pages 185–192, Pittsburgh, PA, 2006.

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

[22] 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.

[23] I. Cohen, F. G. Cozman, N. Sebe, M. C. Cirelo, and T. S. Huang. Semisupervised learning of classifiers: Theory, algorithm, and their application to human-computer interaction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(12):1553–1567, 2004.

[24] M. Collins and Y. Singer. Unsupervised models for named entity classifications. In *Pro-*

ceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora, pages 100–110, College Park, MD, 1999.

[25]R. Collobert, F. Sinz, J. Weston, and L. Bottou. Trading convexity for scalability. In *Proceedings of the 23rd International Conference on Machine Learning*, pages 201–208, Pittsburgh, PA, 2006.

[26]F. G. Cozman and I. Cohen. Unlabeled data can degrade classification performance of generative classifiers. In *Proceedings of the 15th International Conference of the Florida Artificial Intelligence Research Society*, pages 327–331, Pensacola, FL, 2002.

[27]B. V. Dasarathy. *Nearest Neighbor Norms: NN Pattern Classification Techniques*. IEEE Computer Society Press, Los Alamitos, CA, 1991.

[28]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.

[29]A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, 39(1):1–38, 1977.

[30]A. Dong and B. Bhanu. A new semi-supervised EM algorithm for image retrieval. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, pages 662–667, Madison, WI, 2003.

[31]B. Efron and R. Tibshirani. *An Introduction to the Bootstrap*. Chapman & Hall, New York, 1993.

[32]J. D. R. Farquhar, D. Hardoon, H. Meng, J. Shawe-Taylor, and S. Szedmak. Two view learning: SVM-2K, theory and practice. In Y. Weiss, B. Schölkopf, and J. Platt, editors, *Advances in Neural Information Processing Systems 18*, pages 355–362. MIT Press, Cambridge, MA, 2006.

[33]A. Fujino, N. Ueda, and K. Saito. A hybrid generative/discriminative approach to semi-supervised classifier design. In *Proceedings of the 20th National Conference on Artificial Intelligence*, pages 764–769, Pittsburgh, PA, 2005.

[34]J. Garcke and M. Griebel. Semi-supervised learning with sparse grids. In *Working Notes of the ICML'05 Workshop on Learning with Partially Classified Training Data*, Bonn, Germany, 2005.

[35]A. B. Goldberg, M. Li, and X. Zhu. Online manifold regularization: A new learning setting and empirical study. In *Proceedings of the 19th European Conference on Machine Learning*, pages 393–407, Antwerp, Belgium, 2008.

[36]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.

[37]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.

[38]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.

[39]M. Hein and M. Maier. Manifold denoising. In B. Schölkopf, J. C. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems 19*, pages 561–568. MIT Press, Cambridge, MA, 2007.

[40]W. Hosmer. A comparison of iterative maximum likelihood estimates of the parameters of a mixture of two normal distributions under three different types of sample. *Biometrics*, 29(4):761–770, 1973.

[41]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.

[42]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.

[43]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.

[44]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.

[45]D. Lewis and W. Gale. A sequential algorithm for training text classifiers. In *Proceedings*

of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 3–12, Dublin, Ireland, 1994.

[46]M. Li, H. Li, and Z.-H. Zhou. Semi-supervised document retrieval. *Information Processing and Management*, 45(3):341–355, 2009.

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

[48]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.

[49]T. Li and M. Ogihara. Semisupervised learning from different information sources. *Knowledge and Information Systems*, 7(3):289–309, 2005.

[50]R. P. Lippmann. Pattern classification using neural networks. *IEEE Communications*, 27(11):47–64, 1989.

[51]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.

[52]J. McLachlan. Iterative reclassification procedure for constructing an asymptotically optimal rule of allocation in discriminant analysis. *Journal of the American Statistical Association*, 70(350):365–369, 1977.

[53]J. McLachlan and S. Ganesalingam. Updating a discriminant function on the basis of unclassified data. *Communications in Statistics: Simulation and Computation*, 11(6):753–767, 1982.

[54]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.

[55]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.

[56]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.

[57]T. O'Neill. Normal discrimination with unclassified observations. *Journal of the American Statistical Association*, 73(364):821–826, 1978.

[58]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.

[59]E. Riloff and R. Jones. Learning dictionaries for information extraction by multi-level bootstrapping. In *Proceedings of the 16th National Conference on Artificial Intelligence*, pages 474–479, Orlando, FL, 1999.

[60]S. T. Roweis and L. K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500):2323–2326, 2000.

[61]Y. Rui, T. S. Huang, M. Ortega, and S. Mehrotra. Relevance feedback: a power tool for interactive content-based image retrieval. *IEEE Transactions on Circuits and Systems for Video Technology*, 8(5):644–655, 1998.

[62]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.

[63]H. Seung, M. Opper, and H. Sompolinsky. Query by committee. In *Proceedings of the 5th ACM Workshop on Computational Learning Theory*, pages 287–294, Pittsburgh, PA, 1992.

[64]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.

[65]V. Sindhwani and S. S. Keerthi. Large scale semi-supervised linear SVMs. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 477–484, Seattle, WA, 2006.

[66]V. Sindhwani, S. S. Keerthi, and O. Chapelle. Deterministic annealing for semi-supervised kernel machines. In *Proceedings of the 23rd International Conference on Machine Learning*, pages 123–130, Pittsburgh, PA, 2006.

[67]V. Sindhwani, P. Niyogi, and M. Belkin. Beyond the point cloud: From transductive to semi-supervised learning. In *Proceedings of the 22nd International Conference on Machine Learning*, pages 824–831, Bonn, Germany, 2005.

[68]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.

[69]V. N. Vapnik. *Statistical Learning Theory*. Wiley, New York, 1998.

[70]F. Wang and C. Zhang. Label propagation through linear neighborhoods. In *Proceedings of the 23rd International Conference on Machine Learning*, pages 985–992, Pittsburgh, PA, 2006.

[71]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.

[72]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.

[73]X. Wu, V. Kumar, J. R. Quinlan, J. Ghosh, Q. Yang, H. Motoda, G. J. McLachlan, A. Ng, B. Liu, P. S. Yu, Z.-H. Zhou, M. Steinbach, D. J. Hand, and D. Steinberg. Top 10 algorithms in data mining. *Knowledge and Information Systems*, 14(1):1–37, 2008.

[74]Y. Wu, Q. Tian, and T. S. Huang. Discriminant-EM algorithm with application to image retrieval. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, pages 222–227, Hilton Head, SC, 2000.

[75]D. Yarowsky. Unsupervised word sense disambiguation rivaling supervised methods. In *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics*, pages 189–196, Cambridge, MA, 1995.

[76]K. Yu, S. Yu, and V. Tresp. Blockwise supervised inference on large graphs. In *Working Notes of the ICML'05 Workshop on Learning with Partially Classified Training Data*, Bonn, Germany, 2005.

[77]A. L. Yuille and A. Rangarajan. The concave-convex procedure (CCCP). In T. G. Dietterich, S. Becker, and Z. Ghahramani, editors, *Advances in Neural Information Processing Systems 14*, pages 1033–1040. MIT Press, Cambridge, MA, 2002.

[78]T. Zhang and F. J. Oles. A probability analysis on the value of unlabeled data for classification problems. In *Proceedings of 17th International Conference on Machine Learning*, pages 1191–1198, Stanford, CA, 2000.

[79]X. Zhang and W. S. Lee. Hyperparameter learning for graph based semi-supervised learning algorithms. In B. Schölkopf, J. Platt, and T. Hofmann, editors, *Advances in Neural Information Processing Systems 19*, pages 1585–1592. MIT Press, Cambridge, MA, 2007.

[80]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.

[81]Y. Zhou and S. Goldman. Democratic co-learning. In *Proceedings of the 16th IEEE International Conference on Tools with Artificial Intelligence*, pages 594–602, Boca Raton, FL, 2004.

[82]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.

[83]Z.-H. Zhou. Semi-supervised learning by disagreement. In *Proceedings of the 4th IEEE International Conference on Granular Computing*, Hangzhou, China, 2008.

[84]Z.-H. Zhou. Ensemble learning. In S. Z. Li, editor, *Encyclopedia of Biometrics*. Springer, Berlin, 2009.

[85]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.

[86]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.

[87]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.

[88]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.

[89] 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.
[90] Z.-H. Zhou, M. Ng, Q.-Q. She, and Y. Jiang. Budget semi-supervised learning. In *Proceedings of the 13th Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 588–595, Bangkok, Thailand, 2009.
[91] 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.
[92] 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.
[93] 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.
[94] X. Zhu and J. Lafferty. Harmonic mixtures: Combining mixture models and graph-based methods for inductive and scalable semi-supervised leanring. In *Proceedings of the 22nd International Conference on Machine Learning*, pages 1052–1059, Bonn, Germany, 2005.

Author Biographies

Zhi-Hua Zhou is currently Professor in the Department of Computer Science & Technology and director of the LAMDA group at Nanjing University. His main research interests include machine learning, data mining, pattern recognition and information retrieval. He is associate editor-in-chief of *Chinese Science Bulletin*, associate editor of *IEEE Transactions on Knowledge and Data Engineering*, and on the editorial boards of *Artificial Intelligence in Medicine*, *Intelligent Data Analysis*, *Journal of Computer Science & Technology*, *Science in China*, etc. He was an associate editor of *Knowledge and Information Systems* (2003-2008). He is a steering committee member of PAKDD and PRICAI, and has served as program chair/co-chair of PAKDD'07 and PRICAI'08, and vice chair or area chair of ICDM'06, ICDM'08, SDM'09, CIKM'09, etc.

Ming Li received his B.Sc. and Ph.d. degrees in computer science from Nanjing University, China, in 2003 and 2008, respectively. Currently he is an assistant professor in the Department of Computer Science & Technology at Nanjing University, and is a member of the LAMDA Group. His main research interests include machine learning and data mining, especially in learning with labeled and unlabeled examples.

Correspondence and offprint requests to: Zhi-Hua Zhou, National Key Laboratory for Novel Software Technology, Nanjing University, Nanjing 210093, China. Email: zhouzh@nju.edu.cn