When Semi-Supervised Learning Meets Ensemble Learning

Zhi-Hua Zhou

National Key Laboratory for Novel Software Technology
Nanjing University, Nanjing 210093, China
zhouzh@lamda.nju.edu.cn

Abstract. Semi-supervised learning and ensemble learning are two important learning paradigms. The former attempts to achieve strong generalization by exploiting unlabeled data; the latter attempts to achieve strong generalization by using multiple learners. In this paper we advocate generating stronger learning systems by leveraging unlabeled data and classifier combination.

1 Introduction

In many real applications it is difficult to get a large amount of labeled training examples although there may exist abundant unlabeled data, since labeling the unlabeled instances requires human effort and expertise. Exploiting unlabeled data to help improve the learning performance has become a very hot topic during the past decade. There are three major techniques for this purpose [28], i.e., *semi-supervised learning*, *transductive learning* and *active learning*.

Semi-supervised learning [6, 36] deals with methods for exploiting unlabeled data in addition to labeled data automatically to improve learning performance, where no human intervention is assumed. Transductive learning [25] also tries to exploit unlabeled data automatically, but it assumes that the unlabeled examples are exactly the test examples. Active learning deals with methods which assume that the learner has some control over the input space, and the goal is to minimize the number of queries from human experts on ground-truth labels for building a strong learner [22]. In this paper we will focus on semi-supervised learning.

From the perspective of generating strong learning systems, it is interesting to see that semi-supervised learning and *ensemble learning* are two important paradigms that were developed almost in parallel and with different philosophies. Semi-supervised learning tries to achieve strong generalization by exploiting unlabeled data, while ensemble learning tries to achieve strong generalization by using multiple learners. From the view of semi-supervised learning, it seems that using unlabeled data to boost the learning performance can be good enough, and so there is no need to involve multiple learners; while from the view of ensemble learning, it seems that using multiple learners can do all the things and therefore there is no need to consider unlabeled data. This partially explains why the MCS

community has not paid sufficient attention to semi-supervised ensemble methods [20]. Some successful studies have been reported [3, 7, 14, 15, 24, 32], while most are semi-supervised boosting methods [3, 7, 15, 24].

In this article we advocate combining the advantages of semi-supervised learning and ensemble learning. Using *disagreement-based semi-supervised learning* [34] as an example, we will discuss why it is good to leverage unlabeled data and classifier combination. After a brief introduction to disagreement-based methods in Section 2, we will discuss on why classifier combination can be helpful to semi-supervised learning in Section 3, discuss on why unlabeled data can be helpful to ensemble learning in Section 4, and finally conclude in Section 5.

2 Disagreement-Based Semi-Supervised Learning

Research on disagreement-based semi-supervised learning started from Blum and Mitchell's seminal work on co-training [5]. They considered the situation where data have two *sufficient and redundant* views (i.e., two attribute sets each of which contains sufficient information for constructing a strong learner and is conditionally independent to the other attribute set given the class label). The algorithm trains a learner from each view using the original labeled data. Each learner selects and labels some high-confident unlabeled examples for its peer. Then, each learner is refined using the newly labeled examples provided by its peer. The whole process repeats until no learner changes or a pre-set number of learning rounds is executed.

Blum and Mitchell [5] analyzed the effectiveness of co-training and disclosed that if the two views are conditionally independent, the predictive accuracy of an initial weak learner can be boosted to arbitrarily high using unlabeled data by employing the co-training algorithm. Dasgupta et al. [8] showed that when the two views are sufficient and conditionally independent, the generalization error of co-training is upper-bounded by the disagreement between the two classifiers. Later, Balcan et al. [2] indicated that if a PAC learner can be obtained on each view, the conditional independence assumption or even the weak independent assumption [1] is unnecessary, and a weaker assumption of "expansion" of the underlying data distribution is sufficient for iterative co-training to succeed.

Zhou et al. [35] showed that when there are two sufficient and redundant views, a single labeled training example is able to launch a successful co-training. Indeed, the existence of two sufficient and redundant views is a very luxury requirement. In most real-world tasks this condition does not hold since there is generally only a single attribute set. Thus, the applicability of the standard co-training is limited though Nigam and Ghani [18] showed that if there exist a lot of redundant attributes, co-training can be enabled through view split.

To deal with single view data, Goldman and Zhou [9] proposed a method which trains two learners by using different learning algorithms. The method requires each classifier be able to partition the instance space into equivalence classes, and uses cross validation to estimate the confidences of the two learners as well as the equivalence classes. Zhou and Li [32] proposed the *tri-training*

method, which requires neither two views nor special learning algorithms. This method uses three learners and avoids estimating the predictive confidence explicitly. It employs "majority teach minority" strategy in the semi-supervised learning process, that is, if two learners agree on an unlabeled instance but the third learner disagrees, the two learners will label this instance for the third learner. Moreover, classifier combination is exploited to improve generalization. Later, Li and Zhou [14] proposed the *co-forest* method by extending tri-training to include more learners. In co-forest, each learner is improved with unlabeled instances labeled by the ensemble consists of all the other learners, and the final prediction is made by the ensemble of all learners. Zhou and Li [31,33] proposed the first semi-supervised regression algorithm COREG which employs two kNN regressors facilitated with different distance metrics. This algorithm does not require two views either. Later it was extended to a semi-supervised ensemble method for time series prediction with missing data [17].

Previous theoretical studies [2, 5, 8] worked with two views, and could not explain why these single-view methods can work. Wang and Zhou [26] presented a theoretical analysis which discloses that the key for disagreement-based approaches to succeed is the existence of a large diversity between the learners, and it is unimportant whether the diversity is achieved by using two views, or two learning algorithms, or from other channels.

Disagreement-based semi-supervised learning approaches have been applied to many real-world tasks, such as natural language processing [10,19,21,23], image retrieval [28–30], document retrieval [13], spam detection [16], email answering [11], mammogram microcalcification detection [14], etc. In particular, a very effective method which combines disagreement-based semi-supervised learning with active learning for content-based image retrieval has been developed [29,30], and its theoretical analysis was presented recently [27].

3 The Helpfulness of Classifier Combination to Semi-Supervised Learning

Here we briefly introduce some of our theoretical results on the helpfulness of classifier combination to semi-supervised learning. Details can be found in a longer version of [26].

Let $\mathcal{H}$ denote a finite hypothesis space and $\mathcal{D}$ the data distribution generated by the ground-truth hypothesis $h^* \in \mathcal{H}$. Let $d(h^i, h^*) = \Pr_{x \in \mathcal{D}}[h^i(x) \neq h^*(x)]$ denote the difference between two classifiers h^i and h^*. Let h_1^i and h_2^i denote the two classifiers in the i-th round, respectively. We consider the following disagreement-based semi-supervised learning process:

Process: *First, we train two initial learners h_1^0 and h_2^0 using the labeled data set $\mathcal{L}$ which contains l labeled examples. Then, h_1^0 selects u number of unlabeled instances from the unlabeled data set $\mathcal{U}$ to label, and puts these newly labeled examples into the data set σ_2 which contains copies of all examples in $\mathcal{L}$; while h_2^0 selects u number of unlabeled instances from $\mathcal{U}$ to label and puts these newly*

labeled examples into the data set σ_1 which contains copies of all examples in $\mathcal{L}$. h_1^1 and h_2^1 are then trained from σ_1 and σ_2, respectively. After that, h_1^1 selects u number of unlabeled instances from $\mathcal{U}$ to label, and updates σ_2 with these newly labeled examples; while h_2^1 selects u number of unlabeled instances to from $\mathcal{U}$ label, and updates σ_1 with these newly labeled examples. The process is repeated for a pre-set number of learning rounds.

We can prove that even when the individual learners could not improve the performance any more, classifier combination is still possible to improve generalization further by using more unlabeled data.

Lemma 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*

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

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

Lemma 1 suggests that the individual classifier h_2^1 can be improved using unlabeled data when $d(h_1^0, h_2^1)$ is larger than a_0.

Considering a simple classifier combination strategy, that is, when two classifiers disagree on a test instance, the classifier which has a higher confidence is relied on. Let h_{com}^i denote the combination of h_1^i and h_2^i, S^i denote the set of examples on which $h_1^i(x) \neq h_2^i(x)$, and $\gamma = Pr_{x \in S^i}[h_{com}^i(x) \neq h^*(x)]$.

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

$$Pr[h_{com}^1(x) \neq h^*(x)] < Pr[h_{com}^0(x) \neq h^*(x)]. \tag{2}$$

Lemma 2 suggests that the classifier combination h_{com}^0 can be improved using unlabeled data when $d(h_1^1, h_2^1)$ is larger than $\frac{ua_0 + ub_0 + \left(l(1-2\gamma) - u\right) d(h_1^0, h_2^0)}{u + l(1-2\gamma)}$. By Lemmas 1 and 2, we have the following theorem.

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

Moreover, we can prove Theorem 2, which suggests that the classifier combination is possible to reach a good performance earlier than the individual classifiers.

Theorem 2. *Suppose* $a_0 > b_0$, *when* $\gamma < \frac{d(h_1^0,h_2^0)+b_0-a_0}{2d(h_1^0,h_2^0)}$, $Pr[h_{com}^0(x) \neq h^*(x)] < \min[a_0, b_0]$.

4 The Helpfulness of Unlabeled Data to Ensemble Learning

When there are very few labeled training examples, the necessity of exploiting unlabeled data is obvious, since it is impossible to build a strong ensemble otherwise. So, in this section we will only focus on situation where there are a lot of labeled training examples.

It is well-known that to construct a good ensemble, the base classifiers should be accurate and diverse; however, the diversity is difficult to measure and control [12]. We claim that when there are lots of labeled training examples, unlabeled instances are still helpful since they can help to increase the diversity among the base learners. We will briefly introduce a preliminary study below.

Let $\mathcal{X} = \mathcal{R}^d$ denote the d-dimensional input space and $\mathcal{Y} = \{-1, +1\}$ denote the binary label space. Given labeled training set $\mathcal{L} = \{(\boldsymbol{x}_1, y_1), \cdots, (\boldsymbol{x}_l, y_l)\}$ and unlabeled training set $\mathcal{U} = \{\boldsymbol{u}_1, \cdots, \boldsymbol{u}_n\}$, where $\boldsymbol{x}_i \in \mathcal{X}$, $\boldsymbol{u}_j \in \mathcal{X}$ and $y_i \in \mathcal{Y}$, let $\tilde{\mathcal{L}} = \{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_l\}$ denotes the set of unlabeled instances derived from $\mathcal{L}$. Assume that the classifier ensemble $\mathcal{E}$ consists of m linear classifiers $\{\boldsymbol{w}_1 \cdots, \boldsymbol{w}_m\}$, where $\boldsymbol{w}_k \in \mathcal{R}^d$ $(k = 1, \cdots, m)$ is the weight vector of the k-th classifier. Let $\boldsymbol{W} = [\boldsymbol{w}_1, \cdots, \boldsymbol{w}_m]$ be the matrix formed by concatenating all weight vectors. Then, we can generate an ensemble by minimizing the loss function

$$V(\mathcal{L},\mathcal{U},\boldsymbol{W}) = \frac{1}{2}\sum_{k=1}^{m} ||\boldsymbol{w}_k||_2^2 + C_1 \cdot V_{acc}(\mathcal{L},\boldsymbol{W}) + C_2 \cdot V_{div}(\mathcal{D},\boldsymbol{W}) \ , \qquad (3)$$

where the first term controls the model complexity, the second term corresponds to the loss of the ensemble in terms of *accuracy* on $\mathcal{L}$ (balanced by C_1), while the third term corresponds to the loss of the ensemble in terms of *diversity* on data set $\mathcal{D}$ (balanced by C_2). Here, we consider two ways to specify $\mathcal{D}$: (1) $\mathcal{D} = \tilde{\mathcal{L}}$, and (2) $\mathcal{D} = \tilde{\mathcal{L}} \bigcup \mathcal{U}$. The first way leads to the method LCD which does not consider unlabeled data, while the second way leads to the method LCDUD which considers both labeled and unlabeled data.

The second loss term in Eq. 3 can be calculated according to

$$V_{acc}(\mathcal{L},\boldsymbol{W}) = \sum_{k=1}^{m}\sum_{i=1}^{l} loss(\boldsymbol{w}_k, \boldsymbol{x}_i, y_i) \ , \qquad (4)$$

where $loss(\boldsymbol{w}_k, \boldsymbol{x}_i, y_i)$ measures the loss of the k-th base classifier, i.e., $\boldsymbol{w}_k$, on the i-th labeled training example, i.e., $(\boldsymbol{x}_i, y_i)$. Here we calculate it using the l_2 norm

$$loss(\boldsymbol{w}_k, \boldsymbol{x}_i, y_i) = \begin{cases} 0 & \text{if } y_i\langle \boldsymbol{w}_k, \boldsymbol{x}_i\rangle \geq 1 \\ (1 - y_i\langle \boldsymbol{w}_k, \boldsymbol{x}_i\rangle)^2 & \text{if } y_i\langle \boldsymbol{w}_k, \boldsymbol{x}_i\rangle < 1 \end{cases}$$

where $\langle \cdot, \cdot \rangle$ denotes the dot product between vectors.

We calculate the third term in Eq. 3 by considering the *prediction difference* between each pair of base classifiers, i.e.,

$$V_{div}(\mathcal{D}, \boldsymbol{W}) = \sum_{p=1}^{m-1} \sum_{q=p+1}^{m} d(\boldsymbol{w}_p, \boldsymbol{w}_q, \mathcal{D}) \ , \tag{5}$$

where

$$d(\boldsymbol{w}_p, \boldsymbol{w}_q, \mathcal{D}) = \begin{cases} 0 & \text{if } \mathcal{D} = \emptyset \\ \frac{\sum_{\boldsymbol{x} \in \mathcal{D}} \mathbf{sign}(\langle \boldsymbol{w}_p, \boldsymbol{x}\rangle) \cdot \mathbf{sign}(\langle \boldsymbol{w}_q, \boldsymbol{x}\rangle)}{|\mathcal{D}|} & \text{if } \mathcal{D} \neq \emptyset \end{cases} \ .$$

By putting Eqs. 4 and 5 into Eq. 3, and approximating $\mathbf{sign}(\cdot)$ by $\mathbf{tanh}(\cdot)$, the resulting loss function turns to be a continuous and differentiable function of the model parameters $\boldsymbol{W}$. Thus our goal becomes to find the optimal model $\boldsymbol{W}^*$ which minimizes

$$\boldsymbol{W}^* = \arg\min_{\boldsymbol{W}} V(\mathcal{L}, \mathcal{U}, \boldsymbol{W}) \ . \tag{6}$$

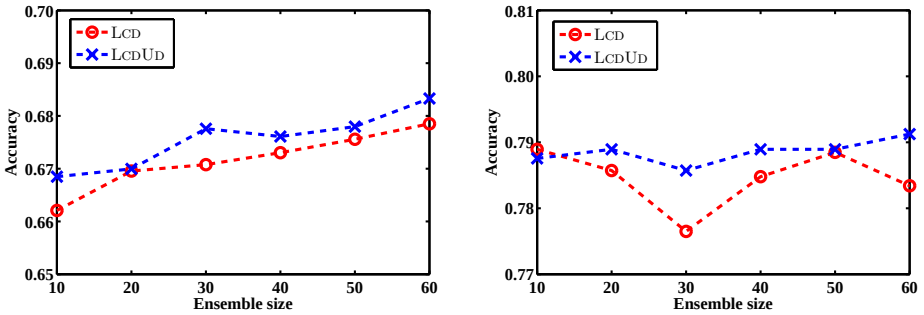
We initialize $\boldsymbol{W}$ by generating each classifier $\boldsymbol{w}_k$ from a bootstrap sample of $\mathcal{L}$, i.e., $\mathcal{L}_k = \{(\boldsymbol{x}_1^k, y_1^k), \cdots, (\boldsymbol{x}_l^k, y_l^k)\}$, by solving the SVM-style optimization problem

$$\min_{\boldsymbol{w}_k, \boldsymbol{\xi}} \frac{1}{2}||\boldsymbol{w}_k||_2^2 + C \sum_{i=1}^{l} \xi_i^k \quad \text{s.t. } \ y_i^k \langle \boldsymbol{w}_k, \boldsymbol{x}_i^k \rangle \geq 1 - \xi_i^k, \ \xi_i^k \geq 0 \ .$$

where $\boldsymbol{\xi} = [\xi_1^k, \xi_2^k, \cdots, \xi_l^k]$. The above problem falls into the category of quadratic programming (QP) and can be solved efficiently by a number of methods off-the-shelf. Then, we solve Eq. 6 by gradient descent.

Figure 1 shows some preliminary results on data sets `g241n` [1] and `vehicle` [4]. For each data set, a half data is randomly chosen to form the test set. Among the remaining data, 5% are used as labeled training examples while 95% are used as unlabeled instances. The experiments are repeated for ten times with random data splits. The parameters C_1 and C_2 are both set to 1. In Figure 1 the horizontal axis in each subfigure shows the size of the ensembles (from 10 to 60 with an interval of 10), and the vertical axis shows the average accuracy. The results show that LcdUd can outperform Lcd, while the only difference between LcdUd and Lcd is that the former considers the usefulness of unlabeled data.

[1] http://www.kyb.tuebingen.mpg.de/ssl-book/benchmarks.html



(a) `g241n` ($N = 1500,\ d = 241$) (b) `vehicle` ($N = 435,\ d = 26$)

Fig. 1. Comparing the performance of LCD and LCDUD. N is the number of instances; d is the dimensionality.

It is worth noting that the above method is far from an excellent one since it does not distinguish the priorities of the contribution from labeled data and unlabeled data. Ideally, the accuracy and diversity on labeled data should be considered at first to form a pool of comparable ensembles, and then from the pool an ensemble with high diversity on unlabeled data is selected. Powerful ensemble methods would be developed along this direction.

5 Conclusion

Semi-supervised learning and ensemble learning are two well-developed paradigms for improving generalization. Although there are some studies of semi-supervised ensemble methods, the MCS community has not devoted much effort to this line of research. In this article we argue that

- Classifier combination is helpful to semi-supervised learning. There are at least two reasons: 1) the performance of classifier combination can be improved further even though the individual learners could not be improved using unlabeled data; 2) the classifier combination can reach a good performance earlier than individual learners.
- Unlabeled data are helpful to ensemble learning. There are at least two reasons: 1) when there are very few labeled training examples, unlabeled data have to be exploited for constructing a strong ensemble; 2) unlabeled data can be used to help increase the diversity of base learners.

Our arguments were made on disagreement-based semi-supervised learning approaches, however, they are possible to generalize to other kinds of semi-supervised learning and ensemble learning approaches. We believe that semi-supervised ensemble methods are very worth studying. Moreover, we think it is possible to derive effective diversity controls for ensemble learning by considering the usefulness of unlabeled data.

Acknowledgments

The author wants to thank Wei Wang and Min-Ling Zhang for their help. This research was supported by the National Science Foundation of China (60635030, 60721002), the National High Technology Research and Development Program of China (2007AA01Z169), the Jiangsu Science Foundation (BK2008018) and the Jiangsu 333 High-Level Talent Cultivation Program.

References

1. S. Abney. Bootstrapping. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 360–367, Philadelphia, PA, 2002.
2. M.-F. Balcan, A. Blum, and K. Yang. Co-training and expansion: Towards bridging theory and practice. In L. K. Saul, Y. Weiss, and L. Bottou, editors, *Advances in Neural Information Processing Systems 17*, pages 89–96. MIT Press, Cambridge, MA, 2005.
3. K. Bennett, A. Demiriz, and R. Maclin. Exploiting unlabeled data in ensemble methods. In *Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 289–296, Edmonton, Canada, 2002.
4. C. Blake, E. Keogh, and C. J. Merz. UCI repository of machine learning databases. [http://www.ics.uci.edu/∼mlearn/MLRepository.html], Department of Information and Computer Science, University of California, Irvine, CA, 1998.
5. 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.
6. O. Chapelle, B. Schölkopf, and A. Zien, editors. *Semi-Supervised Learning*. MIT Press, Cambridge, MA, 2006.
7. F. d'Alché-Buc, Y. Grandvalet, and C. Ambroise. Semi-supervised MarginBoost. In T. G. Dietterich, S. Becker, and Z. Ghahramani, editors, *Advances in Neural Information Processing Systems 14*, pages 553–560. MIT Press, Cambridge, MA, 2002.
8. 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.
9. 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.
10. 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.
11. 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.
12. L. I. Kuncheva and C. J. Whitaker. Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. *Machine Learning*, 51(2):181–207, 2003.

13. M. Li, H. Li, and Z.-H. Zhou. Semi-supervised document retrieval. *Information Processing and Management*, 2009.
14. 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.
15. P. K. Mallapragada, R. Jin, A. K. Jain, and Y. Liu. SemiBoost: Boosting for semi-supervised learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2009.
16. 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.
17. T. A. Mohamed, N. El Gayar, and A. F. Atiya. A co-training approach for time series prediction with missing data. In *Proceedings of the 7th International Workshop on Multiple Classifier Systems*, pages 93–102, Prague, Czech, 2007.
18. 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.
19. 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.
20. F. Roli. Semi-supervised multiple classifier systems: Background and research directions. In *Proceedings of the 6th International Workshop on Multiple Classifier Systems*, pages 1–11, Seaside, CA, 2005.
21. 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.
22. B. Settles. Active learning literature survey. Technical Report 1648, Department of Computer Sciences, University of Wisconsin at Madison, Wisconsin, WI, 2009. http://pages.cs.wisc.edu/∼bsettles/pub/settles.activelearning.pdf.
23. 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.
24. H. Valizadegan, R. Jin, and A. K. Jain. Semi-supervised Boosting for multi-class classification. In *Proceedings of the 19th European Conference on Machine Learning*, pages 522–537, Antwerp, Belgium, 2008.
25. V. N. Vapnik. *Statistical Learning Theory*. Wiley, New York, 1998.
26. 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.
27. 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.
28. 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.
29. 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.

30. 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.
31. 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.
32. 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.
33. 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.
34. Z.-H. Zhou and M. Li. Semi-supervised learning by disagreement. *Knowledge and Information Systems*, in press.
35. 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.
36. 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.