A New Analysis of Co-Training

Wei Wang WANGW@LAMDA.NJU.EDU.CN
Zhi-Hua Zhou ZHOUZH@LAMDA.NJU.EDU.CN
National Key Laboratory for Novel Software Technology, Nanjing University, China

Abstract

In this paper, we present a new analysis on co-training, a representative paradigm of disagreement-based semi-supervised learning methods. In our analysis the co-training process is viewed as a combinative label propagation over two views; this provides a possibility to bring the graph-based and disagreement-based semi-supervised methods into a unified framework. With the analysis we get some insight that has not been disclosed by previous theoretical studies. In particular, we provide the *sufficient and necessary* condition for co-training to succeed. We also discuss the relationship to previous theoretical results and give some other interesting implications of our results, such as combination of weight matrices and view split.

1. Introduction

Semi-supervised learning (Chapelle et al., 2006; Zhu, 2007) deals with methods for automatically exploiting unlabeled data in addition to labeled data to improve learning performance. During the past decade, many semi-supervised learning algorithms have been developed, e.g., S3VMs, graph-based methods and disagreement-based methods. Co-training (Blum & Mitchell, 1998) is a representative paradigm of disagreement-based methods (Zhou & Li, in press). In its initial form, co-training trains two classifiers separately on two sufficient and redundant views and lets the two classifiers label some unlabeled instances for each other. It has been found useful in many applications such as statistical parsing and noun phrase identification (Hwa et al., 2003; Steedman et al., 2003).

All machine learning methods work with specific assumptions, so do semi-supervised learning methods. S3VMs and graph-based methods generally work with the cluster assumption or the manifold assumption. The cluster assumption concerns on classification, while the manifold assumption can also be applied to tasks other than classification. Without these assumptions concerning the relationship between the labels and unlabeled data distribution, semi-supervised learning has limited usefulness (Ben-David et al., 2008).

Like other semi-supervised methods, co-training also needs some assumptions to guarantee its success. When co-training was proposed, Blum & Mitchell (1998) proved that if the two sufficient and redundant views are conditionally independent to the other given the class label, co-training can be successful. Yu et al. (2008) proposed a graphical model for co-training based on the conditional independence assumption as well. Abney (2002) showed that *weak dependence* can also guarantee successful co-training. After that, a weaker assumption called ϵ-*expansion* was proved sufficient for iterative co-training to succeed (Balcan et al., 2005). The above studies give theoretical support to co-training working with two views. For tasks with only a single view, some effective variants have been developed (Goldman & Zhou, 2000; Zhou & Li, 2005). Wang & Zhou (2007) proved that if the two classifiers are with large diversity, co-training style algorithms can succeed. This gives theoretical support to the success of single-view co-training variants, and also contributes to a further understanding of co-training with two-views.

To the best of our knowledge, all previous analyses studied the *sufficient condition* for the success of co-training, yet the *sufficient and necessary* condition is untouched. In this paper, we provide a new analysis of co-training, where the learner in each view is viewed as label propagation and thus the co-training process can be viewed as the combinative label propagation over the two views. Based on this new analysis, we get some insights that have not been discovered by

Appearing in *Proceedings of the 27th International Conference on Machine Learning*, Haifa, Israel, 2010.

previous theoretical studies. In particular, we provide the sufficient and necessary condition for co-training to succeed. We also discuss the relationship to previous theoretical results and give some other interesting implications of our results. Finally we verify our theoretical findings empirically.

The rest of this paper is organized as follows. After introducing some preliminaries we provide our new analysis of co-training in Section 3 and discuss the relationship between our results and previous theoretical results in Section 4. Then we verify our theoretical results in Section 5 and study the implication of our theoretical results on view split in Section 6. Finally we conclude the paper in Section 7.

2. Preliminaries

Suppose we have example space $X = X^1 \times X^2$, where X^1 and X^2 correspond to the two different views of the example space, respectively. Let $\mathbb{D}$ denote the distribution over X, C^1 and C^2 denote the concept classes over X^1 and X^2, respectively. Let $Y = \{-1, 1\}$ denote the label space and $c = (c^1, c^2)$, where $c^1 \in C^1$ and $c^2 \in C^2$, denote the underlying target concept. In co-training, we assume that the labels on examples are consistent with the target concepts c^1 and c^2, i.e., there is no such example $x = (x^1, x^2)$ that $c^1(x^1) \neq c^2(x^2)$ in X. Suppose that the data set is $L \cup U$, where $L = \{\langle (x_1^1, x_1^2), y_1\rangle, \cdots, \langle (x_l^1, x_l^2), y_l\rangle\} \subset X \times Y$ is the labeled data set and $U = \{(x_{l+1}^1, x_{l+1}^2), \cdots, (x_{l+u}^1, x_{l+u}^2)\} \subset X$ is the unlabeled data set.

3. Graph View of Co-training

Co-training (Blum & Mitchell, 1998) trains two learners respectively from two different views and lets the learners label the most confident unlabeled instances to enlarge the training set of the other learner. Such a process can be repeated until some stopping condition is met. In this section, we will show how co-training can be viewed as a combinative label propagation over the two views and then give the sufficient and necessary condition for co-training to succeed.

Generally, assigning a label to an unlabeled instance x_t^v $(v = 1, 2)$ based on a labeled example x_s^v can be viewed as estimating the conditional probability $P\big(y(x_t^v) = y(x_s^v)|x_t^v, x_s^v\big)$. For controlling the confidence of the estimation of the learner, we can set a threshold $\eta^v > 0$ (generally $\eta^v = 1/2$). If $P\big(y(x_t^v) = y(x_s^v)|x_t^v, x_s^v\big) < \eta^v$, we set $P\big(y(x_t^v) = y(x_s^v)|x_t^v, x_s^v\big) = 0$. Note that $P\big(y(x_t^v) = y(x_s^v)|x_t^v, x_s^v\big) = 0$ does not mean that x_s^v and x_t^v must have different labels. In this way, we can assign label to x_t^v according to $P\big(y(x_t^v) = y(x_s^v)|x_t^v, x_s^v\big)$ and the label of x_s^v. For two labeled examples x_w^v and x_q^v, if they have the same label, we set $P\big(y(x_w^v) = y(x_q^v)|x_w^v, x_q^v\big) = 1$ and otherwise $P\big(y(x_w^v) = y(x_q^v)|x_w^v, x_q^v\big) = 0$. Let each entry P_{ij}^v of the matrix P^v correspond to $P\big(y(x_i^v) = y(x_j^v)|x_i^v, x_j^v\big)$ $(1 \leq i, j \leq n = l + u)$ and $f^v = \begin{bmatrix} f_L^v \\ f_U^v \end{bmatrix} = \begin{bmatrix} Y_L \\ 0 \end{bmatrix}$. Without loss of generality, P^v can be normalized to a probabilistic transition matrix according to Eq. 1.

$$P_{ij}^v \leftarrow P_{ij}^v / \sum_{m=1}^{n} P_{im}^v \tag{1}$$

In this way, the labels can be propagated from labeled examples to unlabeled instances according to the process (Zhu, 2005): 1) Propagate $f^v = P^v f^v$; 2) Clamp the labeled data $f_L^v = Y_L$; 3) Repeat from step 1 until f^v converges. The labels of unlabeled instances in U can be assigned according to $sign(f_U^v)$. For some unlabeled instance x_t^v if $f_t^v = 0$, it means that label propagation on graph P^v has no idea on x_t^v. Thus, in each view the learner can be viewed as label propagation from labeled examples to unlabeled instances on graph P^v and we focus on label propagation in this paper. The error $err(f_U^v)$, the accuracy $acc(f_U^v)$ and the uncertainty $\perp(f_U^v)$ of this graph-based method can be counted on U as $acc(f_U^v) = P_{x_t^v \in U}[f_U^v(x_t^v) \cdot c^v(x_t^v) > 0]$, $err(f_U^v) = P_{x_t^v \in U}[f_U^v(x_t^v) \cdot c^v(x_t^v) < 0]$, and $\perp(f_U^v) = P_{x_t^v \in U}[f_U^v(x_t^v) = 0]$. In one view, the labels can be propagated from initial labeled examples to some unlabeled instances in U and these newly labeled examples can be added into the other view. Then the other view can propagate the labels of initial labeled examples and these newly labeled examples to the remaining unlabeled instances in U. This process can be repeated until the stopping condition is met. Thus, the

Table 1. Re-description of co-training

Input: Labeled examples L, unlabeled instances U and probabilistic transition matrix P^v $(v = 1, 2)$.
Process: Perform label propagation from labeled examples L to unlabeled instances U on graph P^v and get the labeled examples set S_0^v.
iterate $k = 0, 1, 2, \cdots$
  if $S_k^1 \oplus S_k^2 = \emptyset$
    break;
  end if
  Perform label propagation from labeled examples $S_k^{3-v} \cap (U - S_k^v)$ to unlabeled instances $U - S_k^v$ on graph P^v and get the labeled examples set T_k^v;
  $S_{k+1}^v = S_k^v \cup T_k^v$.
end iterate
Output: f_U^v corresponding to S_k^v

co-training algorithm can be re-described as the combinative label propagation process over the two views in Table 1, where $S_k^1 \oplus S_k^2 = (S_k^1 - S_k^2) \cup (S_k^2 - S_k^1)$.

3.1. Co-training with Perfect Graphs

Label propagation needs a graph which is represented by the matrix P. In this paper, we focus on the co-training process with two graphs P^1 and P^2 constructed from the two views. How to construct a graph is an important issue studied in graph-based methods (Jebara et al., 2009; Maier et al., 2009) and is beyond the scope of this paper.

First, we assume that P^v $(v = 1, 2)$ is *perfect graph* in this section, i.e., if $P\big(y(x_t^v) = y(x_s^v) | x_t^v, x_s^v\big) > 0$, x_s^v and x_t^v must have the same label. It means that the learner is either "confident of labeling" or "having no idea". Before showing the sufficient and necessary condition for co-training with perfect graphs to succeed, we need Lemma 1 to indicate the relationship between label propagation and connectivity.

Lemma 1 *Suppose that P is perfect graph. Unlabeled instance x_{t_0} can be labeled by label propagation on graph P if and only if it can be connected with some labeled example x_{t_r} in graph P through a path R in the form of $V_R = \{t_0, t_1, \cdots, t_r\}$, where $P_{t_\rho t_{\rho+1}} > 0$ $(\rho = 0, \cdots, r-1)$.*

Proof. It is well known (Zhu, 2005) that the label propagation process has the following closed form solution for each connected component in graph P.

$$f_{U_\theta} = (I - P_{U_\theta U_\theta})^{-1} P_{U_\theta L_\theta} Y_{L_\theta}. \tag{2}$$

Here $U_\theta \cup L_\theta$ is a connected component π_θ in graph P, where $U_\theta \subseteq U$ and $L_\theta \subseteq L$.

If an unlabeled instance x_t cannot be connected with any labeled example, with respect to Eq. 2, we know that $f_t = 0$. If x_{t_0} can be connected with some labeled example x_{t_r} through a path R in the form of $V_R = \{t_0, t_1, \cdots, t_r\}$, considering that P is a perfect graph we get $|f_{t_0}| \geq \prod_{\rho=0}^{r-1} P_{t_\rho t_{\rho+1}} |y_{t_r}|$. Thus, x_{t_0} can be labeled with label $sign(f_{t_0})$ by label propagation. □

From Lemma 1 we know that when every unlabeled instance can be connected with some labeled example through a path in perfect graph P, label propagation on graph P is successful. Now we give Theorem 1.

Theorem 1 *Suppose P^v $(v = 1, 2)$ is perfect graph. $f_U^v(x_t^v) \cdot c^v(x_t^v) > 0$ for all unlabeled instance $x_t \in U$ $(t = l+1, \cdots, l+u)$ if and only if $S_k^1 \oplus S_k^2$ is not $\emptyset$ in Table 1 until $S_k^v = L \cup U$.*

Proof. Here we give a proof by contradiction. Suppose that for any unlabeled instance $x_t \in U$ $(t = l+1, \cdots, l+u)$, $f_U^v(x_t^v) \cdot c^v(x_t^v) > 0$. From Lemma 1 and the process in Table 1 we know that for any unlabeled instance $x_{t_0} \in U$, x_{t_0} can be connected with some labeled example $x_{t_r} \in L$ through a path R in the form of $V_R = \{t_0, t_1, \cdots, t_r\}$, where $P^1_{t_\rho t_{\rho+1}} > 0$ or $P^2_{t_\rho t_{\rho+1}} > 0$ $(\rho = 0, \cdots, r-1)$. If $S_k^1 \oplus S_k^2 = \emptyset$ while $S_k^v \neq L \cup U$, there must exist some unlabeled instances in $U - S_k^v$. Considering that S_k^v are obtained by label propagation on graph P^v, so from Lemma 1 we know that for any unlabeled instance $x_h \in U - S_k^v$, there is no path between x_h and any labeled example $x_d \in S_k^v$ in graph P^v, i.e., $P_{hd}^v = 0$. It is in contradiction with that any unlabeled instance in U can be connected with some labeled example in L through a path R. Therefore, if $f_U^v(x_t^v) \cdot c^v(x_t^v) > 0$ for all unlabeled instance x_t, $S_k^1 \oplus S_k^2$ is not $\emptyset$ until $S_k^v = L \cup U$.

Suppose the graph P^v contains λ_v connected components. If one example in some connected component is labeled, from Lemma 1 we know that all unlabeled instances in this connected component can be labeled by label propagation. If $S_k^1 \oplus S_k^2$ is not $\emptyset$ until $S_k^v = L \cup U$, in the k-th iteration of Table 1, the unlabeled instances in at least one connected component of either P^1 or P^2 will be labeled by label propagation. Thus, after at most $\lambda_1 + \lambda_2$ iterations all unlabeled instances in U can be assigned with labels by the process in Table 1. Considering that P^v in each view is perfect graph, we get that for any unlabeled instance $x_t \in U$, $f_U^v(x_t^v) \cdot c^v(x_t^v) > 0$. □

Theorem 1 provides the sufficient and necessary condition for co-training with perfect graphs to succeed. With this theorem, for tasks with two views, if two perfect graphs can be constructed from the two views, we can decide whether co-training will be successful.

3.2. Co-training with Non-perfect Graphs

In many real applications, it is generally hard to construct a perfect graph. We will discuss the case when the perfect graph assumption is waived in this section.

In label propagation on *non-perfect graph*, an unlabeled instance may be connected with labeled examples belonging to different classes. As discussed in the proof of Lemma 1, the label propagation for each connected component π_θ in graph P has the closed form of $f_{U_\theta} = (I - P_{U_\theta U_\theta})^{-1} P_{U_\theta L_\theta} Y_{L_\theta}$. Let $A = (I - P_{U_\theta U_\theta})^{-1}$, we can get Eq. 3 from Eq. 2.

$$f_t = \sum\nolimits_{s \in L_\theta} \sum\nolimits_{j \in U_\theta} A_{tj} P_{js} Y_s \quad (t \in U_\theta) \tag{3}$$

From Eq. 3 we know that in each connected component

the contribution of the labeled example x_s ($s \in L_\theta$) to the unlabeled instance x_t ($t \in U_\theta$) is $\sum_{j\in U_\theta} A_{tj}P_{js}$. Now we define the *positive contribution* and *negative contribution* for an unlabeled instance.

Definition 1 *Let L_θ denote the labeled examples and U_θ denote the unlabeled instances belonging to the connected component π_θ in graph P. For an unlabeled instance x_t ($t \in U_\theta$), the positive contribution to x_t is*

$$\sum\nolimits_{Y_s=y_t} \sum\nolimits_{j\in U_\theta} A_{tj}P_{js}|Y_s| \tag{4}$$

and the negative contribution to x_t is

$$\sum\nolimits_{Y_s\neq y_t} \sum\nolimits_{j\in U_\theta} A_{tj}P_{js}|Y_s|. \tag{5}$$

If the positive contribution is larger than the negative contribution, the unlabeled instance x_t will be labeled correctly by label propagation.[1] Now we give Theorem 2 for co-training with non-perfect graphs.

Theorem 2 *Suppose P^v ($v = 1, 2$) is non-perfect graph. $f_U^v(x_t^v) \cdot c^v(x_t^v) > 0$ for all unlabeled instance $x_t \in U$ ($t = l+1, \cdots, l+u$) if and only if both (1) and (2) hold in Table 1: (1) $S_k^1 \oplus S_k^2$ is not $\emptyset$ until $S_k^v = L \cup U$; (2) For any unlabeled instance in the connected component $\pi_{\theta_k}^v$, where $\pi_{\theta_k}^v \subseteq (U - S_k^v)$ and $\pi_{\theta_k}^v \cap S_k^{3-v} \neq \emptyset$, its positive contribution is larger than its negative contribution.*

Proof. Here we give a proof by contradiction. Suppose for any unlabeled instance $x_t \in U$, $f_U^v(x_t^v) \cdot c^v(x_t^v) > 0$. If $S_k^1 \oplus S_k^2$ is equal to $\emptyset$ while $S_k^v \neq L \cup U$, for any unlabeled instance $x = (x^1, x^2)$ in $U - S_k^v$, $f_U^v(x^v) = 0$. It is in contradiction with $f_U^v(x^v) \cdot c^v(x^v) > 0$. If for some unlabeled instance x in the connected component $\pi_{\theta_k}^v$, where $\pi_{\theta_k}^v \subseteq (U - S_k^v)$ and $\pi_{\theta_k}^v \cap S_k^{3-v} \neq \emptyset$, its positive contribution is no larger than negative contribution, $f_U^v(x^v) \cdot c^v(x^v) \leq 0$. It is also in contradiction with $f_U^v(x^v) \cdot c^v(x^v) > 0$.

If conditions (1) and (2) hold, with Definition 1 it is easy to get that for any unlabeled instance $x_t \in U$, $f_U^v(x_t^v) \cdot c^v(x_t^v) > 0$. □

Theorem 2 provides the sufficient and necessary condition for co-training with non-perfect graphs to succeed. Note that in both Theorem 1 and Theorem 2, $S_k^1 \oplus S_k^2$ is not $\emptyset$ until $S_k^v = L \cup U$ ($v = 1, 2$) is a necessary condition. In the following part of this section we will further study what this necessary condition means and how to verify it before co-training.

First, we introduce the combinative graph P^c in Eq. 6 which aggregates graphs P^1 and P^2.

$$P_{ij}^c = \max[P_{ij}^1, P_{ij}^2] \tag{6}$$

Then we give Theorem 3 which indicates that each unlabeled instance can be connected with some labeled example in graph P^c is the necessary condition.

Theorem 3 *$S_k^1 \oplus S_k^2$ is not $\emptyset$ in Table 1 until $S_k^v = L \cup U$ ($v = 1, 2$) if and only if each unlabeled instance $x_{t_0} \in U$ can be connected with some labeled example $x_{t_r} \in L$ in graph P^c through a path R^c in the form of $V_{R^c} = \{t_0, t_1, \cdots, t_r\}$, where $P_{t_\rho t_{\rho+1}}^c > 0$ ($\rho = 0, \cdots, r-1$).*

Proof. If we neglect the probability mass on the instances for which the non-zero positive contribution is equal to the non-zero negative contribution in this paper, similarly as the proof of Lemma 1 we get that: Unlabeled instance can be labeled by label propagation on graph P if and only if it can be connected with some labeled example in graph P through a path.

If $S_k^1 \oplus S_k^2$ is not $\emptyset$ until $S_k^v = L \cup U$, any unlabeled instance $x_t \in U$ can be labeled by the process in Table 1. So x_t must belong to one of S_0^1, S_0^2, T_k^1 or T_k^2 for some $k \geq 0$. Considering Eq. 6, the above discussion and the fact that S_0^1, S_0^2, T_k^1 and T_k^2 have been obtained in previous iteration by label propagation and will be used as labeled examples in next iteration, we can get that x_{t_0} can be connected with some labeled example $x_{t_r} \in L$ in graph P^c through a path R^c.

If each unlabeled instance $x_{t_0} \in U$ can be connected with some labeled example $x_{t_r} \in L$ through a path R^c, with respect to Eq. 6, we can get that either $P_{t_\rho t_{\rho+1}}^1$ or $P_{t_\rho t_{\rho+1}}^2$ is larger than 0 for $\rho = 0, \cdots, r-1$. Because x_{t_r} is a labeled example, with above discussion and the process in Table 1 we know that $x_{t_{r-1}}, \cdots, x_{t_0}$ can be labeled by label propagation on either P^1 or P^2. Therefore, finally $S_k^v = L \cup U$. □

3.3. Co-training with ϵ-Good Graphs

It is somehow overly optimistic to expect to learn the target concept perfectly using co-training with non-perfect graphs. While learning the approximately correct concept using co-training with approximately perfect graphs is more reasonable. In perfect graph, all edges between the examples are reliable; while in non-perfect graph, it is hard to know which and how many edges are reliable. Restricting the reliability and allowing an ϵ-fraction exception is more feasible in real

[1] We neglect the probability mass on the instances for which the non-zero positive contribution is equal to the non-zero negative contribution in this paper.

applications. In this section, we focus on the approximately perfect graph and provide sufficient condition for co-training the approximately correct concept.

Let $\pi_1^v, \cdots, \pi_{\lambda_v}^v$ $(v = 1, 2)$ denote the λ_v connected components in graph P^v, the definitions of *purity* and *ϵ-good graph* are given as follows.

Definition 2 *Let $pur(\pi_\theta^v)$ denote the purity of the connected component π_θ^v in graph P^v, then*

$$pur(\pi_\theta^v) = \max\Big[|\{x^v : x^v \in \pi_\theta^v \wedge c^v(x^v) = 1\}|/|\pi_\theta^v|, |\{x^v : x^v \in \pi_\theta^v \wedge c^v(x^v) = -1\}|/|\pi_\theta^v|\Big] \quad (7)$$

If $pur(\pi_\theta^v) \geq 1 - \epsilon$ for all $1 \leq \theta \leq \lambda_v$, we say that P^v is an ϵ-good graph.

The purity of the connected components reflects the reliability of the graph. The higher the purity, the more reliable the graph. With the purity, we can define the label of π_θ^v as $c^v(\pi_\theta^v)$.

$$c^v(\pi_\theta^v) = \begin{cases} 1 & \text{if } |\{x^v : x^v \in \pi_\theta^v \wedge c^v(x^v) = 1\}| \geq \\ & |\{x^v : x^v \in \pi_\theta^v \wedge c^v(x^v) = -1\}| \\ -1 & \text{otherwise} \end{cases}$$

With ϵ-good graph, predicting the labels of all π_θ^v correctly is sufficient to get a learner whose error rate is less than ϵ. From Definition 1 we know that the *contribution* is related to the number of labeled examples in the connected component. In a connected component, if the labeled examples with label y ($y \in \{-1, 1\}$) is much more than the labeled examples with label $-y$, the unlabeled instances belong to this connected component may be labeled with y. Based on this, we assume graph P^v satisfies the following condition: in the connected component $\pi_{\theta_k}^v$ of graph P^v where $\pi_{\theta_k}^v \subseteq (U - S_k^v)$ and $\pi_{\theta_k}^v \cap S_k^{3-v} \neq \emptyset$, let f_k^v denote the learner corresponding to S_k^v, if $|\{x_t : x_t \in \pi_{\theta_k}^v \cap S_k^{3-v} \wedge f_k^{3-v}(x_t) \cdot y > 0\}|/|\pi_{\theta_k}^v| > |\{x_t : x_t \in \pi_{\theta_k}^v \cap S_k^{3-v} \wedge f_k^{3-v}(x_t) \cdot y < 0\}|/|\pi_{\theta_k}^v| + \gamma$, the unlabeled instances belonging to $\pi_{\theta_k}^v$ can be labeled with y by label propagation on graph P^v. Here $\gamma \in [0, 1)$ can be thought of as a form of *margin* which controls the confidence in label propagation. With this assumption, we get Theorem 4 which provides a margin-like sufficient condition for co-training the approximately correct concept with ϵ-good graphs.

Theorem 4 *Suppose P^v $(v = 1, 2)$ is ϵ-good graph. $acc(f_U^v) \geq 1 - \epsilon$ if both (1) and (2) hold in Table 1: (1) $S_k^1 \oplus S_k^2$ is not $\emptyset$ until $S_k^v = L \cup U$; (2) In the connected component $\pi_{\theta_k}^v$, where $\pi_{\theta_k}^v \subseteq (U - S_k^v)$ and $\pi_{\theta_k}^v \cap S_k^{3-v} \neq \emptyset$, $|\{x_t : x_t \in \pi_{\theta_k}^v \cap S_k^{3-v} \wedge f_k^{3-v}(x_t) \cdot c^v(\pi_{\theta_k}^v) > 0\}|/|\pi_{\theta_k}^v| > |\{x_t : x_t \in \pi_{\theta_k}^v \cap S_k^{3-v} \wedge f_k^{3-v}(x_t) \cdot c^v(\pi_{\theta_k}^v) < 0\}|/|\pi_{\theta_k}^v| + \gamma$.*

4. Relationship to Previous Results

There are several theoretical analyses on co-training indicating that co-training can succeed if some condition holds, i.e., *conditional independence*, *weak dependence*, *α-expansion* and *large diversity*. In this section we will discuss the relationship between our results and the previous results at first, then we will discuss some other interesting implications of our results.

4.1. Conditional Independence

Blum & Mitchell (1998) proved that when the two sufficient views are conditionally independent given the class label, co-training can be successful. The *conditional independence* means that for the connected components $\pi_{\theta_i}^1$ of P^1 and $\pi_{\theta_j}^2$ of P^2, $P(\pi_{\theta_i}^1 \cap \pi_{\theta_j}^2) = P(\pi_{\theta_i}^1)P(\pi_{\theta_j}^2)$. Since S_k^v $(v = 1, 2)$ is the union of some connected components of P^v, we have $P(S_k^1 \cap S_k^2) = P(S_k^1)P(S_k^2)$. It means that $P(S_k^1 \oplus S_k^2) = P(S_k^1)(1 - P(S_k^2)) + P(S_k^2)(1 - P(S_k^1))$, which implies that condition (1) in Theorem 4 holds. In addition, Eqs. 8 and 9 can be obtained for ϵ-good graphs.

$$\begin{aligned} & P\big(\pi_{\theta_k}^v \cap S_k^{3-v} \wedge f_k^{3-v}(x_t) \cdot c^v(\pi_{\theta_k}^v) > 0\big) \\ \geq\ & P(\pi_{\theta_k}^v)P(S_k^{3-v})(1 - \epsilon) \quad (8) \\ & P\big(\pi_{\theta_k}^v \cap S_k^{3-v} \wedge f_k^{3-v}(x_t) \cdot c^v(\pi_{\theta_k}^v) < 0\big) \\ <\ & P(\pi_{\theta_k}^v)P(S_k^{3-v})\epsilon \quad (9) \end{aligned}$$

Thus, we get that condition (2) in Theorem 4 holds with $\gamma = P(S_k^{3-v})(1 - 2\epsilon)$. However, in real applications the conditional independence assumption is overly strong to satisfy for the two views.

4.2. Weak Dependence

Abney (2002) found that *weak dependence* can lead to successful co-training. The *weak dependence* means that for the connected components $\pi_{\theta_i}^1$ of P^1 and $\pi_{\theta_j}^2$ of P^2, $P(\pi_{\theta_i}^1 \cap \pi_{\theta_j}^2) \leq \tau P(\pi_{\theta_i}^1)P(\pi_{\theta_j}^2)$ for some $\tau > 0$. It implies that the number of examples in $S_k^1 \oplus S_k^2$ is not very small. So condition (1) in Theorem 4 holds. For ϵ-good graphs, without loss of generality, assume that $P(\pi_{\theta_k}^v \cap S_k^{3-v}) = \tau_1 P(\pi_{\theta_k}^v)P(S_k^{3-v})$ and that

$$\begin{aligned} & P\big(\pi_{\theta_k}^v \cap S_k^{3-v} \wedge f_k^{3-v}(x_t) \cdot c^v(\pi_{\theta_k}^v) < 0\big) \\ \leq\ & \tau_2 P(\pi_{\theta_k}^v)P(S_k^{3-v})\epsilon \quad (10) \end{aligned}$$

for some $\tau_1 > 0$ and $\tau_2 > 0$, we can have

$$P\big(\pi_{\theta_k}^v \cap S_k^{3-v} \wedge f_k^{3-v}(x_t) \cdot c^v(\pi_{\theta_k}^v) > 0\big)$$

$$\geq P\left(\pi^v_{\theta_k} \cap S^{3-v}_k\right) - \tau_2 P(\pi^v_{\theta_k}) P(S^{3-v}_k)\epsilon$$
$$= P(\pi^v_{\theta_k}) P(S^{3-v}_k)(\tau_1 - \tau_2\epsilon). \quad (11)$$

Thus, we get that condition (2) in Theorem 4 holds with $\gamma = P(S^{3-v}_k)(\tau_1 - 2\tau_2\epsilon)$.

4.3. α-Expansion

Balcan et al. (2005) proposed α-expansion and proved that it can guarantee the success of co-training. They assumed that the learner in each view is never "confident but wrong", which corresponds to the case with perfect graphs in Theorem 1. The α-*expansion* means that S^1_k and S^2_k satisfy the condition that $Pr(S^1_k \oplus S^2_k) \geq \alpha \min[Pr(S^1_k \cap S^2_k), Pr(\overline{S^1_k} \cap \overline{S^2_k})]$. When α-expansion is met, it is easy to know that the condition in Theorem 1 holds. Note that $S^1_k \oplus S^2_k \neq \emptyset$ is weaker than α-*expansion*, since $Pr(S^1_k \oplus S^2_k)$ does not need to have a lower bound with respect to some positive α.

4.4. Large Diversity

Wang & Zhou (2007) showed that when the diversity between the two learners is larger than their errors, the performance of the learner can be improved by co-training style algorithms. Since the learners have both error and uncertainty with non-perfect graphs in Table 1, it is very complicated to define the diversity between them. Therefore, we only discuss co-training with perfect graphs here. For perfect graphs, the learners in Table 1 are "confident of labeling", so the error is 0. Thus, that the diversity between the two learners is larger than their errors means $Pr(S^1_k \oplus S^2_k) > 0$, which implies that the condition in Theorem 1 holds.

4.5. Other Implications

From above discussions it can be found if any of the previous condition holds, our condition also holds; this means that our results are more general and tighter. Our results also have other interesting implications.

Firstly, there were some works which combine the weight matrices or the Laplacians for each graph and then classify unlabeled instances according to the combination (Sindhwani et al., 2005; Argyriou et al., 2006; Zhang et al., 2006; Zhou & Burges, 2007), yet the underlying principle is not clear. To some extent, Theorem 3 can provide some theoretical support to these methods, i.e., these methods are developed to satisfy the necessary condition for co-training with graphs to succeed as much as possible.

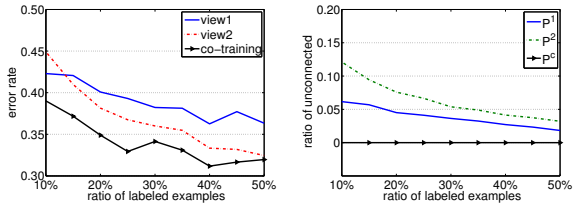
Secondly, in tasks where there does not exist two views, several single-view variants of co-training have been developed (Goldman & Zhou, 2000; Zhou & Li, 2005); to apply the standard two-view co-training directly, *view split* is a possible solution. This has been explored in Nigam & Ghani (2000) and Brefeld et al. (2005). Their studies show that when there are a lot of features and the features have much redundancy, a random split of the features is able to generate two views that enable co-training to outperform several other single-view learning algorithms. However, it is evident that a random split would not be effective in most cases and how to judge a method for view split is also an open problem. From Theorem 3 we know that each unlabeled instance can be connected with some labeled example in the combinative graph P^c is the necessary condition for co-training to succeed. Actually, it implies a possible view split method, i.e., to select the view split which makes more unlabeled instances become connected with labeled examples in graph P^c. In Section 6, we will report on some preliminary experimental results on this method.

5. Verification of Theoretical Results

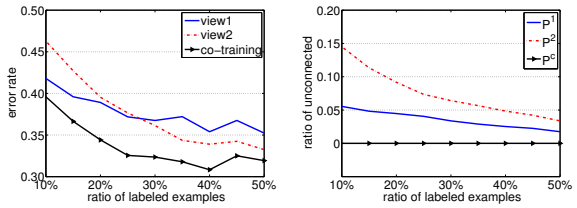
We use the *artificial* data set (Muslea et al., 2002) and the *course* data set (Blum & Mitchell, 1998) in the experiments. The *artificial* data set has two artificial views which are created by randomly pairing two examples from the same class and contains 800 examples. For controlling the connectivity between the two views, the number of clusters per class can be set as a parameter. Here we use 2 clusters and 4 clusters, respectively. The *course* data set has two natural views: *pages* view (i.e., the text appearing on the page) and *links* view (i.e., the anchor text attached to hyper-links pointing to the page) and contains 1,051 examples. We use *1*-NN in each view to approximate the matrix P^v ($v = 1, 2$), i.e., if example x^v_s is the nearest neighbor of x^v_t, $P^v_{st} = P^v_{ts} = 1$ and otherwise $P^v_{st} = P^v_{ts} = 0$. The combinative graph P^c is constructed according to Eq. 6. P^1, P^2 and P^c are normalized according to Eq. 1. We randomly select some data to be used as the labeled data set L and use the remaining data to generate the unlabeled data set U. The error rate is calculated over U. To study the performance with different amount of labeled examples, we run experiments with different sizes of L, from 10% to 50% with interval 5%. The experiments are repeated for 20 runs and the average results are shown in Figure 1.

From Figure 1(a), 1(c) and 1(e) we can see that on both data sets the performance of co-training is much better than the learner in each view. This can be successfully explained by our graph view explanation. As Figure 1(b), 1(d) and 1(f) show, the amount of unlabeled instances that are not connected with any la-

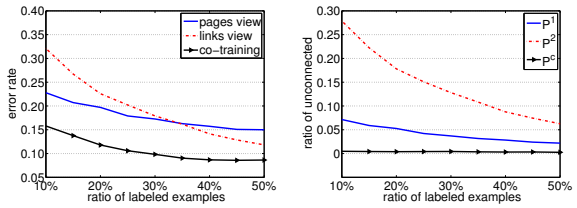
beled example in the graph P^c is much smaller than that in the graphs P^1 and P^2. So, co-training can label not only the unlabeled instances that can be labeled by a single view, but also the unlabeled instances that cannot be labeled by either a single view.



(a) *artificial with 2 clusters* (b) *artificial with 2 clusters*



(c) *artificial with 4 clusters* (d) *artificial with 4 clusters*



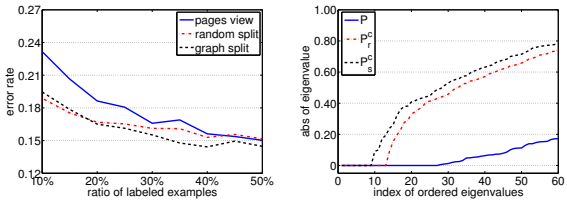
(e) *course* (f) *course*

Figure 1. Results on the *artificial* data set and the *course* data set. The x-axis shows the amount of labeled examples (in the ratio of L); the y-axis in (a), (c) and (e) shows the error rate of co-training and learners using a single view; the y-axis in (b), (d) and (f) shows the amount (in ratio) of unlabeled data that are not connected with any labeled example in the graphs P^1, P^2 and P^c.

6. View Split for Single-view Data

As mentioned in the end of Section 4, our theoretical results imply a method which enables co-training to work on single-view data, i.e., to select the view split which makes more unlabeled instances become connected with labeled examples in graph P^c, and then generate two views for co-training to work on.

We use the *course* data set with only the *pages* view here. Thus, the experimental data is with a single view. We split the features of the *pages* view into two parts randomly ten times and use *1*-NN to approximate the matrices. The combinative graph P^c is constructed according to Eq. 6. Let P denote the graph on the *pages* view, P and P^c are normalized according to Eq. 1. L and U are generated in the way as similar as that described in Section 5. Among the ten view splits, we select the one which leads to the largest amount of unlabeled instances connected with labeled examples in the graph P^c. The results are shown in Figure 2(a), where we also present the performances of using the original *pages* view and co-training with random view split. From Figure 2(a) we can see that the performance of co-training with selected view split is always better than using the single view and is always superior to co-training with random view split except in very few cases where they are comparable.



(a) (b)

Figure 2. Result on the view split method. (a) The x-axis shows the amount of labeled examples (in the ratio of L), and the y-axis shows the error rate; (b) the x-axis shows the index of the ascendingly ordered eigenvalues, and the y-axis shows the absolute value of the eigenvalues.

To study the result further, we calculate the eigenvalues of the Laplacian matrices related to the graphs P, P_r^c and P_s^c (P_r^c corresponds to the combinative graph with random view split and P_s^c corresponds to the combinative graph with selected view split) and sort the absolute value of these eigenvalues in ascending order, respectively. The first 60 ones are plotted in Figure 2(b). By setting $\Delta = 10^{-10}$, we get that the Laplacian matrix related to the graphs P and P_r^c have 27 and 13 eigenvalues whose absolute value is smaller than Δ, respectively. While the Laplacian matrix related to the graph P_s^c has only 9 eigenvalues whose absolute value is smaller than Δ. This implies that, the graph P has 27 connected components and the graph P_r^c has 13 connected components, while the graph P_s^c has only 9 connected components, with an apparent improvement. In other words, through the selected view split, more unlabeled instances become connected with labeled examples in the graph P_s^c. This validates the usefulness of the simple view split method derived from our theoretical results.

7. Conclusion

In this paper, we provide a new analysis of co-training, based on which we get the sufficient and necessary condition for co-training to succeed. Although previously

there were many theoretical studies on co-training, to the best of our knowledge, this is the first result on the sufficient and necessary condition for co-training. We also discuss the relationship between our results and previous theoretical results. Moreover, our results have some other interesting implications, such as combination of weight matrices and view split.

Our results can be extended to multi-view cases. Similar sufficient and necessary condition for multi-view learning can be obtained. Note that such an extension only cares the multiple matrices rather than where these matrices come from, and therefore it is also suited for applications where there is only one view in the data set but multiple conditional probability matrices can be obtained in different concept spaces.

It is noteworthy that in previous semi-supervised learning studies, the disagreement-based and graph-based methods were developed separately, in two parallel threads. While our analysis provides a possibility to bring them into a unified framework, which will be explored further in the future.

Acknowledgments

Supported by NSFC (60635030, 60721002), 973 Program (2010CB327903) and JiangsuSF (BK2008018).

References

Abney, S. Bootstrapping. In *ACL*, pp. 360–367, 2002.

Argyriou, A., Herbster, M., and Pontil, M. Combining graph laplacians for semi-supervised learning. In *NIPS 18*, pp. 67–74. 2006.

Balcan, M.-F., Blum, A., and Yang, K. Co-training and expansion: Towards bridging theory and practice. In *NIPS 17*, pp. 89–96. 2005.

Ben-David, S., Lu, T., and Pal, D. Does unlabeled data provably help? worst-case analysis of the sample complexity of semi-supervised learning. In *COLT*, pp. 33–44, 2008.

Blum, A. and Mitchell, T. Combining labeled and unlabeled data with co-training. In *COLT*, pp. 92–100, 1998.

Brefeld, U., Büscher, C., and Scheffer, T. Multi-view discriminative sequential learning. In *ECML*, pp. 60–71, 2005.

Chapelle, O., Schölkopf, B., and Zien, A. (eds.). *Semi-Supervised Learning*. MIT Press, 2006.

Goldman, S. and Zhou, Y. Enhancing supervised learning with unlabeled data. In *ICML*, pp. 327–334, 2000.

Hwa, R., Osborne, M., Sarkar, A., and Steedman, M. Corrected cotraining for statistical parsers. In *ICML Workshop*, 2003.

Jebara, T., Wang, J., and Chang, S.-F. Graph construction and b-matching for semi-supervised learning. In *ICML*, pp. 441–448, 2009.

Maier, M., von Luxburg, U., and Hein, M. Influence of graph construction on graph-based clustering measures. In *NIPS 21*, pp. 1025–1032. 2009.

Muslea, I., Minton, S., and Knoblock, C. A. Active + semi-supervised learning = robust multi-view learning. In *ICML*, pp. 435–442, 2002.

Nigam, K. and Ghani, R. Analyzing the effectiveness and applicability of co-training. In *CIKM*, pp. 86–93, 2000.

Sindhwani, V., Niyogi, P., and Belkin, M. A co-regularization approach to semi-supervised learning with multiple views. In *ICML Workshop*, 2005.

Steedman, M., Osborne, M., Sarkar, A., Clark, S., Hwa, R., Hockenmaier, J., Ruhlen, P., Baker, S., and Crim, J. Bootstrapping statistical parsers from small data sets. In *EACL*, pp. 331–338, 2003.

Wang, W. and Zhou, Z.-H. Analyzing co-training style algorithms. In *ECML*, pp. 454–465, 2007.

Yu, S., Krishnapuram, B., Rosales, R., Steck, H., and Rao, R. B. Bayesian co-training. In *NIPS 20*, pp. 1665–1672. 2008.

Zhang, T., Popescul, A., and Dom, B. Linear prediction models with graph regularization for web-page categorization. In *SIGKDD*, pp. 821–826, 2006.

Zhou, D. and Burges, C. Spectral clustering and transductive learning with multiple views. In *ICML*, pp. 1159–1166, 2007.

Zhou, Z.-H. and Li, M. Tri-training: Exploiting unlabeled data using three classifiers. *IEEE TKDE*, 17 (11):1529–1541, 2005.

Zhou, Z.-H. and Li, M. Semi-supervised learning by disagreement. *KAIS*, in press.

Zhu, X. *Semi-supervised learning with graphs*. PhD thesis, CS School, CMU, 2005.

Zhu, X. Semi-supervised learning literature survey. Technical Report 1530, Department of Computer Sciences, University of Wisconsin at Madison, 2007.