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 $\epsilon$-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
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 $D$ 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 examine the label space and $\cup_{c \in C^1}$ over $X$ (generally $\cup_{c \in C}$). Suppose we have example space, respectively. Let $X$ and $\cup_{c \in C^1}$, 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 $L = \{(x^1_1, x^2_1), y_1\}, \ldots, (x^1_l, x^2_l), y_l\} \subset X \times Y$ is the labeled data set and $U = \{ (x^1_{l+1}, x^2_{l+1}), \ldots, (x^1_{l+u}, x^2_{l+u}) \} \subset X$ 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 $x^u_i$ ($v = 1, 2$) based on a labeled example $x^v_j$ can be viewed as estimating the conditional probability $P(y(x^v_j)|x^v_j, x^u_i)$ for controlling the confidence of the estimation of the learner, we can set a threshold $\eta^v > 0$ (generally $\eta^v = 1/2$). If $P(y(x^v_j)|x^v_j, x^u_i) < \eta^v$, we set $P(y(x^v_j)|x^v_j, x^u_i) = 0$. Note that $P(y(x^v_j)|x^v_j, x^u_i) = 0$ does not mean that $x^v_j$ and $x^u_i$ must have different labels. In this way, we can assign label to $x^u_i$ according to $P(y(x^v_j)|x^v_j, x^u_i)$ and the label of $x^v_j$. For two labeled examples $x^v_j$ and $x^v_j$ if they have the same label, we set $P(y(x^v_j)|x^v_j, x^v_j) = 1$ and otherwise $P(y(x^v_j)|x^v_j, x^v_j) = 0$. Let each entry $P_{ij}^v$ of the matrix $P^v$ correspond to $P(y(x^v_i)|x^v_j|x^v_j)$ $(1 \leq i, j \leq n)$ and $f^v = \left[ f_{ij}^L \ f_{ij}^U \right] = \left[ \begin{array}{c} Y_L \\ 0 \end{array} \right]$. Without loss of generality, $P^v$ can be normalized to a probabilistic transition matrix according to Eq. 1.

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

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 $f_{ij}^L = Y_{ij}$; 3) Repeat from step 1 until $f^v$ converges. The labels of unlabeled instances in $U$ can be assigned according to $sign(f_{ij}^v)$. For some unlabeled instance $x^u_i$ if $f_{ij}^L = 0$, it means that label propagation on graph $P^v$ has no idea on $x^u_i$. 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 $err(f_{ij}^v)$, the accuracy $acc(f_{ij}^v)$ and the uncertainty $\perp(f_{ij}^v)$ of this graph-based method can be counted on $U$ as $acc(f_{ij}^v) = P_{x^u_i \in U} \left[ f_{ij}^v(x^u_i) \cdot c^v(x^u_i) > 0 \right]$, $err(f_{ij}^u) = P_{x^u_i \in U} \left[ f_{ij}^v(x^u_i) \cdot c^v(x^u_i) < 0 \right]$, and $\perp(f_{ij}^v) = P_{x^u_i \in U} \left[ f_{ij}^v(x^u_i) = 0 \right]$.

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

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 $P(y(x^v_0) = y(x^v_s)(x^v_s, x^v_0) > 0$, $x^v_s$ and $x^v_0$ 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 $x_{t_e}$ in graph $P$ through a path $R$ in the form of $V_R = \{t_0, t_1, \cdots , t_r\}$, where $P_{t_t, t_{t+1}} > 0$ ($\rho = 0, \cdots , r - 1$).

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

$$f_{U\theta} = (I - P_{U\theta}U\theta)^{-1}P_{U\theta}L\theta Y_{L\theta}. \quad (2)$$

Here $U\theta \cup L\theta$ is a connected component $\pi_\theta$ 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 $x_{t_e}$ 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 $|f_{t_0}| \geq \prod_{\rho=0}^{r-1} P_{t_\rho, t_{\rho+1}} |y_{t_\rho}|$. Thus, $x_{t_0}$ can be labeled with label $\text{sign}(f_{t_0})$ by label propagation. \hfill \Box

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. $f^v_{U^v}(x^v_t) \cdot c^v(x^v_t) > 0$ for all unlabeled instance $x_t \in U$ ($t = l + 1, \cdots , l + u$) if and only if $S_k^1 \oplus S_k^2_0$ 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 $x_t \in U$ ($t = l + 1, \cdots , l + u$), $f^v_{U^v}(x^v_t) \cdot c^v(x^v_t) > 0$. 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 $x_{t_e} \in L$ through a path $R$ in the form of $V_R = \{t_0, t_1, \cdots , t_r\}$, where $P_{t_\rho, t_{\rho+1}} > 0$ ($\rho = 0, \cdots , r - 1$). If $S_k^1 \oplus S_k^2_0 = \emptyset$ while $S_k^v \neq L \cup U$, there must exist some unlabeled instances in $U - S_k^v$. Considering that $S_k^1$ 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., $P^v_{hd} = 0$. 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 $f^v_{U^v}(x^v_t) \cdot c^v(x^v_t) > 0$ for all unlabeled instance $x_t$, $S_k^1 \oplus S_k^2_0$ is not $\emptyset$ until $S_k^v = L \cup U$.

Suppose the graph $P^v$ contains $\lambda_0$ 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 $S_k^1 \oplus S_k^2_0$ 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 $x_t \in U$, $f^v_{U^v}(x^v_t) \cdot c^v(x^v_t) > 0$. \hfill \Box

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 $\pi_\theta$ 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.
provides the sufficient and necessary condition for co-training with non-perfect graphs. Now we define the positive contribution and negative contribution for an unlabeled instance.

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

$$f_t = \sum_{s \in L_\emptyset} \sum_{j \in U_\emptyset} A_{ij} P_{js} Y_s \quad (t \in U_\emptyset)$$

(3)

From Eq. 3 we know that in each connected component the contribution of the labeled example $x_s$ ($s \in L_\emptyset$) to the unlabeled instance $x_t$ ($t \in U_\emptyset$) is $\sum_{j \in U_\emptyset} A_{ij} P_{js}$. Now we define the positive contribution and negative contribution for an unlabeled instance.

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. $f^v_U(x^v_t) \cdot c^v(x^v_t) > 0$ for all unlabeled instance $x_t \in U$ ($t = l + 1, \ldots, l + u$) if and only if both (1) and (2) hold in Table 1: (1) $S^1_k \oplus S^2_k$ is not $\emptyset$ until $S^v_k = L \cup U$; (2) For any unlabeled instance in the connected component $\pi^v_\emptyset$, where $\pi^v_\emptyset \subseteq (U - S^v_k)$ and $\pi^v_\emptyset \cap S^3-v \neq \emptyset$, its positive contribution is larger than its negative contribution.

**Proof.** Here we give a proof by contradiction. Suppose for any unlabeled instance $x_t \in U$, $f^v_U(x^v_t) \cdot c^v(x^v_t) > 0$. If $S^1_k \oplus S^2_k$ is equal to $\emptyset$ while $S^v_k \neq L \cup U$, for any unlabeled instance $x = (x^1, x^2)$ in $U - S^v_k$, $f^v_U(x^v) = 0$. In contradiction with $f^v_U(x^v) \cdot c^v(x^v) > 0$. If for some unlabeled instance $x$ in the connected component $\pi^v_\emptyset$, where $\pi^v_\emptyset \subseteq (U - S^v_k)$ and $\pi^v_\emptyset \cap S^3-v \neq \emptyset$, its positive contribution is no larger than negative contribution, $f^v_U(x^v) \cdot c^v(x^v) \leq 0$. It is also in contradiction with $f^v_U(x^v) \cdot c^v(x^v) > 0$.

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

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, $S^1_k \oplus S^2_k$ is not $\emptyset$ until $S^v_k = L \cup U$ ($v = 1, 2$) 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 $P^2$.

$$P^c_{ij} = \max[P^1_{ij}, P^2_{ij}]$$

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

**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 $S^1_k \oplus S^2_k$ is not $\emptyset$ until $S^v_k = L \cup U$, any unlabeled instance $x_t \in U$ can be labeled by the process in Table 1. So $x_t$ must belong to one of $S^1_0, S^2_0, T^1_k$ or $T^2_k$ for some $k > 0$. Considering Eq. 6, the above discussion and the fact that $S^1_0, S^2_0, T^1_k$ and $T^2_k$ 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 $x_{t_\rho} \in L$ 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 $x_{t_\rho} \in L$ through a path $R^c$, with respect to Eq. 6, we can get that either $P^1_{t_{\rho-1}, t_\rho}$ or $P^2_{t_{\rho-1}, t_\rho}$ is larger than 0 for $\rho = 0, \ldots, r - 1$. Because $x_{t_1}$ is a labeled example, with above discussion and the process in Table 1 we know that $x_{t_{r-1}}, \ldots, x_{t_0}$ can be labeled by label propagation on either $P^1$ or $P^2$. Therefore, finally $S^v_k = L \cup U$.\(\Box\)

### 3.3. Co-training with $\epsilon$-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 $\epsilon$-fraction exception is more feasible in real applications. In this section, we focus on the approximately perfect graph and provide sufficient condition for co-training the approximately perfect concept.

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

**Definition 2** Let $\text{pur}(\pi_v^\theta)$ denote the purity of the connected component $\pi_v^\theta$ in graph $P^v$, then

$$
\text{pur}(\pi_v^\theta) = \max \left\{ \left[ \{ x^v : x^v \in \pi_v^\theta \land c^v(x^v) = 1 \} \right] / |\pi_v^\theta|, \right.
\left. \{ x^v : x^v \in \pi_v^\theta \land c^v(x^v) = -1 \} / |\pi_v^\theta| \right\}
$$

(7)

If $\text{pur}(\pi_v^\theta) \geq 1 - \epsilon$ for all $1 \leq \theta \leq \lambda_v$, we say that $P^v$ is an $\epsilon$-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 $\pi_v^\theta$ as $c^v(\pi_v^\theta)$.

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

With $\epsilon$-good graph, predicting the labels of all $\pi_v^\theta$ correctly is sufficient to get a learner whose error rate is less than $\epsilon$. 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 $f_{\theta_k}^v$ denote the learner corresponding to $S_k^v$, if $\{ x_t : x_t \in \pi_{\theta_k}^v \cap S_k^{3-v} \land f_{\theta_k}^{3-v}(x_t) \cdot y > 0 \} / |\pi_{\theta_k}^v| > \{ x_t : x_t \in \pi_{\theta_k}^v \cap S_k^{3-v} \land f_{\theta_k}^{3-v}(x_t) \cdot y < 0 \} / |\pi_{\theta_k}^v| + \gamma$, 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 $\epsilon$-good graphs.

**Theorem 4** Suppose $P^v$ ($v = 1, 2$) is $\epsilon$-good graph. $\text{acc}(f_{\theta_k}^v) \geq 1 - \epsilon$ 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 $\pi_{\theta_k}^v \cap S_k^{3-v} \neq \emptyset$, $\{ x_t : x_t \in \pi_{\theta_k}^v \cap S_k^{3-v} \land f_{\theta_k}^{3-v}(x_t) \cdot c^v(\pi_{\theta_k}^v) > 0 \} / |\pi_{\theta_k}^v| > \{ x_t : x_t \in \pi_{\theta_k}^v \cap S_k^{3-v} \land f_{\theta_k}^{3-v}(x_t) \cdot c^v(\pi_{\theta_k}^v) < 0 \} / |\pi_{\theta_k}^v| + \gamma$.

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, $\alpha$-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 $\pi_{\theta_1}^1$ of $P^1$ and $\pi_{\theta_2}^2$ of $P^2$, $P(\pi_{\theta_1}^1 \cap \pi_{\theta_2}^2) = P(\pi_{\theta_1}^1)P(\pi_{\theta_2}^2)$. 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 $P(S_k^1 \cup S_k^2) = P(S_k^1)(1 - P(S_k^2)) + P(S_k^2)(1 - P(S_k^1))$, which implies that condition (1) in Theorem 4 holds. In addition, Eqs. 8 and 9 can be obtained for $\epsilon$-good graphs.

$$
P(\pi_{\theta_k}^v \cap S_k^{3-v} \land f_{\theta_k}^{3-v}(x_t) \cdot c^v(\pi_{\theta_k}^v) > 0) \geq P(\pi_{\theta_k}^v)P(S_k^{3-v})(1 - \epsilon) \quad (8)
$$

$$
P(\pi_{\theta_k}^v \cap S_k^{3-v} \land f_{\theta_k}^{3-v}(x_t) \cdot c^v(\pi_{\theta_k}^v) < 0) < P(\pi_{\theta_k}^v)P(S_k^{3-v})\epsilon \quad (9)
$$

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 $\pi_{\theta_1}^1$ of $P^1$ and $\pi_{\theta_2}^2$ of $P^2$, $P(\pi_{\theta_1}^1 \cap \pi_{\theta_2}^2) \leq \tau P(\pi_{\theta_1}^1)P(\pi_{\theta_2}^2)$ 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 $\epsilon$-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

$$
P(\pi_{\theta_k}^v \cap S_k^{3-v} \land f_{\theta_k}^{3-v}(x_t) \cdot c^v(\pi_{\theta_k}^v) < 0) < \tau_2 P(\pi_{\theta_k}^v)P(S_k^{3-v})\epsilon \quad (10)
$$
for some $\tau_1 > 0$ and $\tau_2 > 0$, we can have
\[
P(\pi_k^u \cap S_k^{3-v} \cap f_k^{3-v}(x_t) \cdot c^v(\pi_k^u) > 0) \\
\geq P(\pi_k^u \cap S_k^{3-v}) - \tau_2 P(\pi_k^e) P(S_k^{3-v}) \epsilon \\
= P(\pi_k^u) P(S_k^{3-v}) (\tau_1 - \tau_2 \epsilon).
\]
Thus, we get that condition (2) in Theorem 4 holds with $\gamma = P(S_k^{3-v}) (\tau_1 - 2 \tau_2 \epsilon)$.

4.3. $\alpha$-Expansion

Balcan et al. (2005) proposed $\alpha$-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 $\alpha$-expansion means that $S_k^1$ and $S_k^2$ satisfy the condition that $Pr(S_k^1 \oplus S_k^2) \geq \alpha \min\{Pr(S_k^1 \cap S_k^2), Pr(S_k^2 \cap S_k^1)\}$. When $\alpha$-expansion is met, it is easy to know that the condition in Theorem 1 holds. Note that $S_k^1 \oplus S_k^2 \neq \emptyset$ is weaker than $\alpha$-expansion, since $Pr(S_k^1 \oplus S_k^2)$ does not need to have a lower bound with respect to some positive $\alpha$.

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 $Pr(S_k^1 \oplus S_k^2) > 0$, 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 Ngiam & 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 $x^u$ is the nearest neighbor of $x^v$, $P^v_{ux} = P^v_{tx} = 1$ and otherwise $P^v_{ux} = P^v_{tx} = 0$. 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
We use the generate two views for co-training to work on.

connected with labeled examples in graph which makes more unlabeled instances become con-

work on single-view data, i.e., to select the view split

results imply a method which enables co-training to

As mentioned in the end of Section 4, our theoretical

6. View Split for Single-view Data

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
element in the graphs $P^1$, $P^2$ and $P^c$.

Figure 1(b), 1(d) and 1(f) show, the amount of un-
labeled instances that are not connected with any la-
abeled 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.

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.

structured according to Eq. 6. Let $P$ denote the graph
on the 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. To study the result further, we calculate the eigenval-
ues of the Laplacian matrices related to the graphs
$P$, $P^c$ and $P^c$ ($P^c$ corresponds to the combinative
graph with random view split and $P^c$ 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 $P^c$ have 27 and 13 eigenvalues whose absolute value is smaller than $\Delta$, respectively. While the Laplacian matrix related to the graph $P^c$ has only 9 eigenvalues whose absolute value is smaller than $\Delta$. This implies that, the graph $P$ has 27 connected components and the graph $P^c$ has 13 connected components, while the graph $P^c$ 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 $P^c$. 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 con-
A New Analysis of Co-Training

dition 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


