JMLR: Workshop and Conference Proceedings 29:1–16, 2013 ACML 2013

Co-Training with Insufficient Views

Wei Wang WANGW@LAMDA.NJU.EDU.CN
Zhi-Hua Zhou ZHOUZH@LAMDA.NJU.EDU.CN
National Key Laboratory for Novel Software Technology
Nanjing University, Nanjing 210023, China

Editor: Cheng Soon Ong and Tu Bao Ho

Abstract

Co-training is a famous semi-supervised learning paradigm exploiting unlabeled data with two views. Most previous theoretical analyses on co-training are based on the assumption that each of the views is sufficient to correctly predict the label. However, this assumption can hardly be met in real applications due to feature corruption or various feature noise. In this paper, we present the theoretical analysis on co-training when neither view is sufficient. We define the diversity between the two views with respect to the confidence of prediction and prove that if the two views have large diversity, co-training is able to improve the learning performance by exploiting unlabeled data even with insufficient views. We also discuss the relationship between view insufficiency and diversity, and give some implications for understanding of the difference between co-training and co-regularization.

Keywords: Co-training, insufficient views

1. Introduction

During the past decade, many researchers have shown great interest in *semi-supervised learning* (Chapelle et al., 2006; Zhu, 2007), which deals with methods for automatically exploiting unlabeled data to improve learning performance, and developed a variety of semi-supervised learning algorithms, e.g., S3VMs, graph-based methods and disagreement-based methods. *Co-training* (Blum and Mitchell, 1998) is a representative paradigm of disagreement-based methods (Zhou and Li, 2010), which trains two classifiers separately on two views and allows them to label some unlabeled instances for each other. The two views are two disjoint sets of features. For example, the web page classification task has two views, i.e., the text appearing on the page and the anchor text attached to hyper-links pointing to this page (Blum and Mitchell, 1998). It has been found useful in many applications such as natural language processing (Hwa et al., 2003; Steedman et al., 2003) and image retrieval (Wang and Zhou, 2008; Zhou et al., 2006).

Several theoretical analyses on co-training have been studied. In the seminal work on co-training of Blum and Mitchell (1998), they proved that when the two sufficient and redundant views are *conditionally independent*, co-training can boost the performance of weak classifiers to arbitrarily high by using unlabeled data. Dasgupta et al. (2002) showed that when the conditional independence assumption holds the generalization error of co-training is upper-bounded by the disagreement between the two classifiers. To relax the conditional independence assumption, Abney (2002) found that *weak dependence* can also

lead to the success of co-training. After that, Balcan et al. (2005) proposed ϵ-*expansion* and proved that if the classifier in each view is never "confident but wrong", the ϵ-*expansion* assumption can guarantee the success of co-training. Wang and Zhou (2007) showed that if the two classifiers have *large diversity*, co-training style algorithms can also succeed in improving the learning performance. It is worthy mentioning that all above theoretical analyses on co-training are based on the assumption that each of the views can provide sufficient information to correctly predict the label, which means that the optimal classifier in each view can correctly classify all examples. However, in many real applications, due to feature corruption or various feature noise, neither view can provide sufficient information to correctly predict the label, i.e., there exist some examples $\big(\langle x_1, x_2\rangle, y\big)$, on which the posterior probability $P(y{=}{+}1|x_v)$ or $P(y{=}{-}1|x_v)$ $(v = 1, 2)$ is not equal to 1 due to the insufficient information provided by x_v for predicting the label[1]. For these examples, the optimal classifier may not correctly predict their labels because of the view insufficiency.

In semi-supervised learning, there is another method called *co-regularization* (Brefeld et al., 2006; Farquhar et al., 2006; Sindhwani et al., 2005), which exploits two-view unlabeled data, sometimes also known as *regularized co-training* (Balcan and Blum, 2010). Co-regularization directly minimizes the error rate on labeled data and the disagreement over unlabeled data with the intuition that the optimal classifiers in the two views are compatible with each other. It is worthy noting that co-training exploits unlabeled data with two views very differently from co-regularization. In detail, co-training uses initial labeled data to learn two weak hypotheses and allows them to label confident instances for each other; while co-regularization directly minimizes the error rate on labeled data and the disagreement over unlabeled data and no pseudo-labels are assigned to unlabeled instances. Balcan and Blum (2010) defined the compatibility of a pair of hypotheses $\langle f_1, f_2\rangle$ with distribution $\mathbb{D}_{\mathcal{X}}$ as $1{-}P_{\langle x_1,x_2\rangle\in\mathbb{D}_{\mathcal{X}}}(f_1(x_1){\neq}f_2(x_2))$ and provided a framework with the notion of compatibility to interpret co-regularization. Co-regularization allows for views with partial insufficiency, but it assumes that the two views provide almost the same information. Unfortunately, in real applications each view may be corrupted by different kind of noise, it is unreasonable to assume that the two views provide almost the same information. The two optimal classifiers in the two views may mistakenly classify different examples due to the different information sources, which causes the two optimal classifiers are no longer compatible with each other. When the two views are corrupted by different noise processes or provide diverse information, the performance of co-regularization will be adversatively influenced since it strongly encourages the agreement between the two views. Actually, Sridharan and Kakade (2008) presented an information theoretic framework for co-regularization which showed that the excess error between the output hypothesis of co-regularization and the optimal classifier is punished by the term $\sqrt{\epsilon_{info}}$, where $\epsilon_{info}{<}1$ measures the different information provided by the two views. This implies that it is hard for co-regularization to find the ϵ-approximation of the optimal classifier when the two views are insufficient and provide diverse information for predicting the label. In Section 2 we will show this theoretically.

In this paper, we present the theoretical analysis on co-training with insufficient views which is much more challenging but practical, especially when the two views provide diverse information. We give the definition on the insufficiency of the view and indicate that co-

1. Here x_1 and x_2 denote the two views of an example, see Section 2 for formal definition.

training might suffer from two limitations, i.e., *label noise* and *sampling bias*, when each view is insufficient. We also make definition on the diversity between two views with respect to the confidence of prediction and prove that if the two views have large diversity, co-training suffers little from the two limitations and could succeed in outputting the approximation of the optimal classifier by exploiting unlabeled data even with insufficient views. The rest of this paper is organized as follows. After stating some preliminaries in Section 2, we present our theoretical analysis in Section 3 and finish with conclusions in Section 4.

2. Insufficiency of the View

In two-view setting, an instance is described with two different disjoint sets of features and each set of features is called as one view. Suppose we have the instance space $\mathcal{X} = \mathcal{X}_1 \times \mathcal{X}_2$, where $\mathcal{X}_1$ and $\mathcal{X}_2$ correspond to the two different views of the instance space, respectively. Let $\mathcal{Y} = \{-1, 1\}$ denote the label space, $\mathbb{D}$ denote the unknown underlying distribution over $\mathcal{X} \times \mathcal{Y}$ and $\mathbb{D}_{\mathcal{X}}$ denote the unknown underlying marginal distribution over $\mathcal{X}$. In this paper, we focus on the situation where each of the views cannot provide sufficient information to correctly predict the label, i.e., there exist some examples $\big(\langle x_1, x_2\rangle, y\big)$ on which either $P(y = +1|x_v) \neq 1$ or $P(y = -1|x_v) \neq 1$ $(v = 1, 2)$, where $x_1 \in \mathcal{X}_1$, $x_2 \in \mathcal{X}_2$ and $y \in \mathcal{Y}$. Let $L \cup U$ denote the data set, where $L = \big\{\big(\langle x_1^1, x_2^1\rangle, y_1\big), \cdots, \big(\langle x_1^l, x_2^l\rangle, y^l\big)\big\} \subset \mathcal{X} \times \mathcal{Y}$ is the labeled data set i.i.d. sampled from $\mathbb{D}$ and $U = \Big\{\langle x_1^{l+1}, x_2^{l+1}\rangle, \cdots, \langle x_1^{l+u}, x_2^{l+u}\rangle\Big\} \subset \mathcal{X}$ is the unlabeled data set i.i.d. sampled from $\mathbb{D}_{\mathcal{X}}$. For an example $\big(\langle x_1, x_2\rangle, y\big)$, let $\varphi_v(x_v) = P(y = +1|x_v)$. If $\varphi_v(x_v) = \frac{1}{2}$, it implies that the features of x_v provide no helpful information to correctly predict its label y; while if $\varphi_v(x_v)$ is 1 or 0, it implies that the features of x_v provide sufficient information to correctly predict its label y. It is easy to understand that $|2\varphi_v(x_v) - 1|$ is a measurement of the information provided by x_v for predicting its label y. Now we give the definition on the insufficiency of the view.

Definition 1 ***(Insufficiency)*** *Let $\mathbb{D}$ denote the unknown underlying distribution over $\mathcal{X} \times \mathcal{Y}$. For $(x, y) \in \mathcal{X} \times \mathcal{Y}$, $\varphi(x) = P(y = +1|x)$. The insufficiency $\Upsilon(\mathcal{X}, \mathcal{Y}, \mathbb{D})$ on the learning task with respect to the example space $\mathcal{X} \times \mathcal{Y}$ and distribution $\mathbb{D}$ is defined as*

$$\Upsilon(\mathcal{X}, \mathcal{Y}, \mathbb{D}) = 1 - \int_{x \in \mathbb{D}_{\mathcal{X}}} |2\varphi(x) - 1| P(x) dx$$

$\Upsilon(\mathcal{X}, \mathcal{Y}, \mathbb{D}) \in [0, 1]$ measures the insufficiency of view $\mathcal{X}$ for correctly learning $\mathcal{Y}$ over distribution $\mathbb{D}$. When $|2\varphi(x) - 1| = 1$ for all examples, the insufficiency $\Upsilon(\mathcal{X}, \mathcal{Y}, \mathbb{D}) = 0$, i.e., view $\mathcal{X}$ provides sufficient information to correctly classify all examples; while $\varphi(x) = \frac{1}{2}$ for all examples, the insufficiency $\Upsilon(\mathcal{X}, \mathcal{Y}, \mathbb{D}) = 1$, i.e., view $\mathcal{X}$ provides no information to correctly classify any example. With Definition 1, we let $\Upsilon_v = \Upsilon(\mathcal{X}_v, \mathcal{Y}, \mathbb{D})$ denote the insufficiency of view $\mathcal{X}_v$.

Let $\mathcal{H}_v$: $\mathcal{X}_v \to [-1, +1]$ denote the hypothesis class for learning with view $\mathcal{X}_v$ $(v = 1, 2)$ and d_v denote the finite VC-dimension of hypothesis class $\mathcal{H}_v$. The classification rule induced by a hypothesis $h_v \in \mathcal{H}_v$ on an instance $x = \langle x_1, x_2\rangle$ is $sign(h_v(x_v))$. The error rate of a hypothesis h_v with distribution $\mathbb{D}$ is $R(h_v) = P_{(x_1, x_2, y) \in \mathbb{D}}(y \neq sign(h_v(x_v)))$ and let $R(\mathcal{S}_v) = \max_{h_v \in \mathcal{S}_v} R(h_v)$ for $\mathcal{S}_v \subseteq \mathcal{H}_v$. Let $c_v(x_v) = 2\varphi_v(x_v) - 1$, $sign(c_v(x_v)) = +1$ if $\varphi_v(x_v) \geq \frac{1}{2}$ and $sign(c_v(x_v)) = -1$ otherwise, $c_v(x_v) \in [-1, +1]$. Suppose c_v belongs to $\mathcal{H}_v$,

and it is well-known (Devroye et al., 1996) that c_1 and c_2 are the optimal Bayes classifiers in the two views, respectively. Let $\eta_v = R(c_v)$ denote the error rate of the optimal classifier c_v, we have the following Proposition 2.

Proposition 2 $\Upsilon_v = 2\eta_v$. $(v = 1, 2)$

Proof Given an example $(\langle x_1, x_2\rangle, y)$,

$$\begin{aligned}
&P\Big(sign\big(c_v(x_v)\big) \neq y|x_v\Big) \\
&= 1 - P\Big(sign\big(c_v(x_v)\big) = 1, y = 1|x_v\Big) - P\Big(sign\big(c_v(x_v)\big) = -1, y = -1|x_v\Big) \\
&= 1 - \mathbb{I}\Big\{sign\big(c_v(x_v)\big) = 1\Big\}P\big(y = 1|x_v\big) - \mathbb{I}\Big\{sign\big(c_v(x_v)\big) = -1\Big\}P\big(y = -1|x_v\big) \\
&= 1 - \mathbb{I}\big\{\varphi_v(x_v) > 1/2\big\}\varphi_v(x_v) - \mathbb{I}\big\{\varphi_v(x_v) \leq 1/2\big\}\big(1 - \varphi_v(x_v)\big)
\end{aligned}$$

So we get

$$\begin{aligned}
\eta_v &= \mathbb{E}\Big(1 - \mathbb{I}\big\{\varphi_v(x_v) > 1/2\big\}\varphi_v(x_v) - \mathbb{I}\big\{\varphi_v(x_v) \leq 1/2\big\}\big(1 - \varphi_v(x_v)\big)\Big) \\
&= \mathbb{E}\Big(1/2 - \big|\varphi_v(x_v) - 1/2\big|\Big) \\
&= \frac{1}{2}\Upsilon_v
\end{aligned}$$

■

Proposition 2 states that when the view is insufficient, the optimal classifier will mistakenly classify some examples. The larger the insufficiency, the worse the performance of the optimal classifier.

When the two views $\mathcal{X}_1$ and $\mathcal{X}_2$ are insufficient, they probably provide different information for predicting the label $\mathcal{Y}$ due to different information sources. Sridharan and Kakade (2008) used the conditional mutual information $I(A:B|C)$ to measure how much knowing A reduces the uncertainty of B conditioned on already knowing C, they assumed that $I(\mathcal{Y}:\mathcal{X}_v|\mathcal{X}_{3-v}) \leq \epsilon_{info}$ $(v = 1, 2)$ holds for some small $\epsilon_{info} > 0$, and provided an information theoretic framework for co-regularization which minimizes the following co-regularized loss for the pair (h_1, h_2) $(h_v \in \mathcal{H}_v)$.

$$Loss_{co}(h_1, h_2) = \frac{1}{2}\big(\widehat{R}_L(h_1) + \widehat{R}_L(h_2)\big) + \lambda_1\|h_1\| + \lambda_2\|h_2\| + \lambda_3\widehat{D}_U(h_1, h_2)$$

$\widehat{R}_L$ is the empirical risk with respect to the labeled data set L and $\widehat{D}_U$ is the empirical disagreement with respect to the unlabeled data set U. Note that $I(\mathcal{Y}:\mathcal{X}_v|\mathcal{X}_{3-v}) \leq \epsilon_{info}$ means that if we already knew view $\mathcal{X}_v$ then there is little more information that we could get from view $\mathcal{X}_{3-v}$ about $\mathcal{Y}$, i.e., the two views provide almost the same information. However, the two views generally provide diverse information due to different information sources, i.e., $I(\mathcal{Y}:\mathcal{X}_v|\mathcal{X}_{3-v}) > \epsilon_{info}$. We call the two views satisfying $I(\mathcal{Y}:\mathcal{X}_v|\mathcal{X}_{3-v}) \leq \epsilon_{info}$ as *similar views* and call the two views satisfying $I(\mathcal{Y}:\mathcal{X}_v|\mathcal{X}_{3-v}) > \epsilon_{info}$ as *diverse views*. For *diverse views*, there exist some instances $x = \langle x_1, x_2\rangle$ on which $P(y = +1|x_1)$ is different

from $P(y=+1|x_2)$, since x_1 and x_2 come from different sources to represent x. Thus, c_1 is not perfectly compatible with c_2 and let the following $D(c_1, c_2)$ denote the difference between c_1 and c_2.

$$D(c_1, c_2) = P_{x\in\mathbb{D}_{\mathcal{X}}}\big(sign(c_1(x_1)) \neq sign(c_2(x_2))\big)$$

Now we give the following Proposition 3 to show that co-regularization may never output the approximations of the optimal classifiers.

Proposition 3 *Suppose* $\|h_v\| = 1$ *for* $h_v \in \mathcal{H}_v$, $(v = 1, 2)$. *Let* $\mathcal{H}_v^L \subset \mathcal{H}_v$ *denote the hypotheses minimizing the empirical risk on the labeled data set* L *and* $(g_1, g_2) = \arg\min_{h_v\in\mathcal{H}_v^L} D(h_1, h_2)$. *If* $|U|$ *is sufficiently large,* $Loss_{co}(g_1, g_2)$ *is no larger than* $Loss_{co}(c_1, c_2)$.

Proof Considering that $\widehat{R}_L(g_v) = \widehat{R}_L(c_v)$ and that $\widehat{D}_U\big(g_1, g_2\big) \leq \widehat{D}_U\big(c_1, c_2\big)$ holds for sufficient large $|U|$, it is easy to get Proposition 3 proved. ■

Let us try to give an intuitive explanation to Proposition 3. Proposition 3 states that when the two views provide diverse information, co-regularization prefers to output the pair of hypotheses which simply minimizes the disagreement on the unlabeled data set rather than the optimal classifiers. Its performance will be adversatively influenced by the diverse information between the two views, especially when the unlabeled data set is very large while the labeled data set is small. In the rest part of the paper, we will study what co-training could do when it meets with insufficient views, especially when the two views provide diverse information.

3. Main Result

3.1. Limitations of Co-Training with Insufficient Views

Co-training trains two classifiers with initial labeled examples from the two views $\mathcal{X}_1$ and $\mathcal{X}_2$, respectively, and allows one of them to label some unlabeled instances to enlarge the training set for the other. The classifier in view $\mathcal{X}_v$ is updated with examples labeled by the other classifier in view $\mathcal{X}_{3-v}$ $(v = 1, 2)$. These examples with pseudo-labels may contain label noise and the following Proposition 4 shows that if the label noise is caused by some other classifier, we may not be able to achieve the optimal classifier.

Proposition 4 *Let* $\mathcal{Q}$ *be an instance set sampled i.i.d from* $\mathbb{D}_{\mathcal{X}}$, *for* $x = \langle x_1, x_2\rangle \in \mathcal{Q}$, *its pseudo-label* $\widehat{y}$ *is assigned according to some classifier* f *(*f *may not be in* $\mathcal{H}_1$ *or* $\mathcal{H}_2$*). Let* $h_v^{\mathcal{Q}} \in \mathcal{H}_v$ *be the hypothesis minimizing the empirical risk on* $\mathcal{Q}$ *with the pseudo-labels in view* $\mathcal{X}_v$ *and*

$$h_v^{\min} = \arg\min_{h_v\in\mathcal{H}_v} P_{x\in\mathbb{D}_{\mathcal{X}}}\big(sign(h_v(x_v)) \neq sign(f(x))\big).$$

For $\epsilon \in (0, \frac{1}{2})$ *and* $\delta \in (0, 1)$, *if* $|\mathcal{Q}| = O(\frac{d_v \ln\frac{1}{\delta}}{\epsilon^2})$, *the following bound holds with probability* $1 - \delta$.

$$P_{x\in\mathbb{D}_{\mathcal{X}}}\big(sign(h_v^{\mathcal{Q}}(x_v)) \neq sign(h_v^{\min}(x_v))\big) < \epsilon$$

Proof For $x=\langle x_1, x_2\rangle \in \mathcal{Q}$, its pseudo-label $\widehat{y}$ is assigned according to some classifier f and

$$h_v^{\min} = \arg\min_{h_v \in \mathcal{H}_v} P_{x\in\mathbb{D}_{\mathcal{X}}}\big(sign(h_v(x_v)) \neq sign(f(x))\big),$$

so $h_v^{\min}$ has the minimum expected empirical risk on $\mathcal{Q}$. According to standard PAC-theory, if $|\mathcal{Q}| = O\big(\frac{d_v \ln \frac{1}{\delta}}{\epsilon^2}\big)$, the following bound

$$P_{x\in\mathbb{D}_{\mathcal{X}}}\big(sign(h_v^{\mathcal{Q}}(x_v)) \neq sign(h_v^{\min}(x_v))\big) < \epsilon$$

holds with probability $1-\delta$. ■

Proposition 4 states the possibility that when the training data has label noise introduced by some classifier f, the output hypothesis minimizing the empirical risk might be drawn away from the optimal classifier c_v to $h_v^{\min}$. Considering that co-training allows one of the classifiers to label some instances for the other, so there exists the possibility that label noise prohibits co-training from outputting the optimal classifier. We call *label noise* as one of the limitations of co-training with insufficient views.

Sometimes, when the two views satisfy some strong assumption, label noise may not prohibit co-training from outputting the optimal classifier. We give the following Lemma 5 to show this.

Lemma 5 *Suppose the two views are conditionally independent given the class label and we can train two weak classifiers h_1^0 and h_2^0 whose error rates are less than $\eta < \frac{1}{2}$ with initial labeled examples in the two views, respectively. Let the classifier in view $\mathcal{X}_v$ $(v=1,2)$ only uses the examples labeled by the other classifier in view $\mathcal{X}_{3-v}$ as the retraining data, and excludes the examples labeled by itself and the initial labeled examples. Let $\mathcal{Q}_v$ denote the training data for view $\mathcal{X}_v$ and $h_v^{\mathcal{Q}_v}$ denote the hypothesis minimizing the empirical risk on $\mathcal{Q}_v$. For $\epsilon \in (0, \frac{1}{2})$ and $\delta \in (0,1)$, if $|\mathcal{Q}_v| = O\big(\frac{d_v \ln \frac{1}{\delta}}{\epsilon^2(1-2\eta)^2}\big)$, then $R\big(h_v^{\mathcal{Q}_v}\big) \leq R(c_v) + \epsilon$ holds with probability $1-\delta$.*

Proof For $\big(\langle x_1, x_2\rangle, \widehat{y}\big) \in \mathcal{Q}_1$, its pseudo-label is assigned by h_2^0. For $h_1 \in \mathcal{H}_1$, with the assumption that the two views are conditionally independent given the class label we get

$$\begin{aligned} P\big(sign(h_1(x_1)) \neq \widehat{y}\big) &= \big(1 - R(h_0^2)\big)R(h_1) + R(h_2^0)\big(1 - R(h_1)\big) \\ &= R(h_1)\big(1 - 2R(h_2^0)\big) + R(h_2^0). \end{aligned}$$

It is easy to find that c_1 has the minimum expected empirical risk on $\mathcal{Q}_1$. If $|\mathcal{Q}_1| = O\big(\frac{d_v \ln \frac{1}{\delta}}{\epsilon^2(1-2\eta)^2}\big)$, with Hoeffding's inequality we get $R\big(h_1^{\mathcal{Q}_1}\big) \leq R(c_1) + \epsilon$ with probability $1-\delta$. Similarly, we get $R\big(h_2^{\mathcal{Q}_2}\big) \leq R(c_2) + \epsilon$ with probability $1-\delta$. ■

However, the conditional independence assumption is overly strong to satisfy in real applications and Abney (2002) theoretically showed how unreasonably strong this assumption is. In order to reduce label noise in the examples with pseudo-labels, one usually lets one classifier label its confident instances for the other. Although this method can reduce label

noise, it would cause another problem that the training data in each view is not an i.i.d. sample from the marginal distribution $\mathbb{D}_{\mathcal{X}}$. If the training data is not an i.i.d. sample, we may not be able to achieve the optimal classifier even if there is no label noise. We give the following Proposition 6 to show this.

Proposition 6 *Let* $\mathcal{R}_v = \{(\langle x_1, x_2\rangle, y) \in \mathcal{X} \times \mathcal{Y} : sign(c_v(x_v)) \neq y\}$. *Obviously,* $\mathcal{R}_v$ *is a set of examples without label noise, and the optimal classifier* c_v *has the maximum empirical risk on* $\mathcal{R}_v$.

Proof For any example $(\langle x_1, x_2\rangle, y) \in \mathcal{R}_v$, y is its underground label, so $\mathcal{R}_v$ is a set of examples without label noise. Since $sign(c_v(x_v)) \neq y$, so the optimal classifier c_v mistakenly classifies x_v. Thus, c_v has the maximum empirical risk on $\mathcal{R}_v$. ■

Proposition 6 states the possibility that when there is sampling bias in the training data, we might not achieve the optimal classifier c_v even if there is no label noise in the training data. We call *sampling bias* as another limitation of co-training with insufficient views, which is caused by attempting to reduce *label noise*. It is worthy mentioning that when both views are sufficient, the optimal classifiers always have 0 empirical risk on any training data without label noise. Sampling bias will never make the optimal classifier perform worse than any non-optimal classifier on the noiseless training data, but might slow down the rate of convergence to the optimal classifier.

3.2. Learning Approximation of Optimal Classifier without Sampling Bias

Usually, co-training allows one of the classifiers to label its confident instances for the other and the process is described in Algorithm 1. When the confidence of the prediction on an instance is no less than some preset threshold, we use the predicted label as its pseudo-label and add it into the training set. However, if there is no prior knowledge about the relationship between hypothesis class and unlabeled data, it is hard to guarantee that selecting confident instances to label could reduce the label noise. In margin-based algorithms, margin could be used to measure the prediction quality. Intuitively, it is likely that similar hypotheses tend to have similar margin output, i.e., two hypotheses with small error difference should have small margin difference. With this intuition, we give the following Definition 7.

Definition 7 ***(Margin Lipschitz)*** *Let* $\mathcal{H}_v$ $(v = 1, 2)$ *denote the hypothesis class, for* $x = \langle x_1, x_2\rangle$ *and* $h_v \in \mathcal{H}_v$, *there exists some constant* $C_v^{\mathfrak{L}}$ *to satisfy* $|h_v(x_v) - c_v(x_v)| \leq C_v^{\mathfrak{L}}(R(h_v) - R(c_v))$.

Definition 7 states that the label predicted by weak classifiers with large margin is likely to be the same as the label predicted by the optimal classifier. Thus, the confident instances would help find the optimal classifier. To quantify the amount of the confident instances, we give the following Definition 8.

Definition 8 ***(Diversity)*** *Let* $\mathcal{F} \subseteq \mathcal{H}_1$ *and* $\mathcal{G} \subseteq \mathcal{H}_2$,

$$\mu(\gamma_1, \gamma_2, \mathcal{F}, \mathcal{G}) = P\{x = \langle x_1, x_2\rangle \in U : \exists f \in \mathcal{F}, g \in \mathcal{G} \text{ s.t. } |f(x_1)| \geq \gamma_1 \text{ or } |g(x_2)| \geq \gamma_2\}.$$

Algorithm 1 Margin-based co-training

Input: Labeled data set L, unlabeled data set U, $\sigma_0 = L$
for $i = 0, 1, 2, \ldots$ **do**
  Get $\mathcal{H}_v^i \subseteq \mathcal{H}_v$ by minimizing the empirical risk on σ_i with respect to view $\mathcal{X}_v$ and set $T_i = \emptyset$
  for $x = \langle x_1, x_2 \rangle \in U$ **do**
    for $h_1 \in \mathcal{H}_1^i$, $h_2 \in \mathcal{H}_2^i$ **do**
      if $|h_1(x_1)| \geq \gamma_1$ **then**
        $T_i = T_i \cup (x, sign(h_1(x_1)))$ and delete x from U
        break
      end if
      if $|h_2(x_2)| \geq \gamma_2$ **then**
        $T_i = T_i \cup (x, sign(h_2(x_2)))$ and delete x from U
        break
      end if
    end for
  end for
  if $T_i = \emptyset$ **then**
    return
  end if
  $\sigma_{i+1} = \sigma_i \cup T_i$
end for
Output: $\mathcal{H}_1^C = \mathcal{H}_1^i$ and $\mathcal{H}_2^C = \mathcal{H}_2^i$

We define $\mu(\gamma_1, \gamma_2, \mathcal{F}, \mathcal{G}) \in [0, 1]$ as the diversity between the two views with respect to margins γ_1 and γ_2, which quantifies the amount of confident instances. When μ is large, the two views could help each other strongly by providing many confident instances; while when μ is small, the two views only help each other little since there are few confident instances. For *similar views*, an instance which is labeled with small margin by one view may also be labeled with small margin by the other view, since the two views provide almost the same information for predicting the label; while for *diverse views*, an instance which is labeled with small margin by one view may be labeled with large margin by the other view, since the two views provide diverse information for predicting the label. Intuitively, *diverse views* would have large diversity. In the extreme case where the diversity μ is 1, we have the following Theorem 9.

Theorem 9 *Suppose the hypothesis class $\mathcal{H}_v$ $(v = 1, 2)$ satisfies Definition 7, let $\mathcal{H}_v^L \subseteq \mathcal{H}_v$ denote the hypotheses minimizing the empirical risk on initial labeled data set L, $R_v = \max_{h_v \in \mathcal{H}_v^L} R(h_v)$ and $\gamma_v = C_v^{\mathfrak{L}}(R_v - \eta_v)$. For $\epsilon \in (0, \frac{1}{2})$ and $\delta \in (0, 1)$, if $|U| = O(\frac{d_v \ln \frac{1}{\delta}}{\epsilon^2})$ and $\mu(\gamma_1, \gamma_2, \mathcal{H}_1^L, \mathcal{H}_2^L) = 1$, with probability $1 - \delta$ the output $\mathcal{H}_1^C$ and $\mathcal{H}_2^C$ in Algorithm 1 satisfy $R(\mathcal{H}_v^C) \leq \frac{\eta_1 + \eta_2 + D(c_1, c_2)}{2} + \epsilon$.*

Proof Since $\mu(\gamma_1, \gamma_2, \mathcal{H}_1^L, \mathcal{H}_2^L) = 1$, after 1 round all unlabeled instances in U are assigned with pseudo-labels and added into the data set σ_1. Then classifier set $\mathcal{H}_v^1$ is got by

minimizing the empirical risk on σ_1 with view $\mathcal{X}_v$. For $x = \langle x_1, x_2 \rangle$, $\widehat{y}$ denote its pseudo-label. If $|h_v(x_v)| \geq \gamma_v = C_v^{\mathfrak{L}}(R_v - \eta_v)$, with Definition 7 we know that h_v and c_v make the same prediction on x_v. So for any example $\big(\langle x_1, x_2 \rangle, \widehat{y}\big) \in \sigma_1$, either $\widehat{y} = sign\big(c_1(x_1)\big)$ or $\widehat{y} = sign\big(c_2(x_2)\big)$ holds. Here we consider the worst case that

$$\widehat{y} = \begin{cases} y & \text{if } sign\big(c_1(x_1)\big) = sign\big(c_2(x_2)\big) = y \\ -y & \text{otherwise} \end{cases}.$$

Let h_v^{com} denote the hypothesis that $sign\big(h_v^{com}(x_v)\big) = y$ if $sign\big(c_1(x_1)\big) = sign\big(c_2(x_2)\big) = y$, and $sign\big(h_v^{com}(x_v)\big) = -y$ otherwise. It is easy to find that h_v^{com} is consistent with the examples in σ_1 for the worst case and $R(h_v^{com}) = \frac{\eta_1 + \eta_2 + D(c_1, c_2)}{2}$. $R(h_v^{com})$ is larger than $R(c_v)$, so learning a classifier with error rate no larger than $R(h_v^{com}) + \epsilon$ is no harder than learning a classifier with error rate no larger than $R(c_v) + \epsilon$. Now we regard h_v^{com} as the optimal classifier in $\mathcal{H}_v$ and neglect the probability mass on the hypothesis whose error rate is less than $R(h_v^{com})$. Since the classifiers in $\mathcal{H}_v^C$ minimize the empirical risk on training data σ_1 which is an i.i.d sample with size of $|L| + |U|$ and $|U| = O\big(\frac{d_v \ln \frac{1}{\delta}}{\epsilon^2}\big)$, we get $\max_{h_v \in \mathcal{H}_v^C} R(h_v) \leq R(h_v^{com}) + \epsilon$ with probability $1 - \delta$. ■

Theorem 9 states that if the diversity with margins γ_1 and γ_2 between the two views is 1, i.e., every unlabeled instance in U could be labeled with large margin by one of the two views, co-training could output the near-good hypothesis set $\mathcal{H}_1^C$ and $\mathcal{H}_2^C$. Sometimes the pseudo-label which is the same as the prediction of the optimal classifier in view $\mathcal{X}_v$ is not good for achieving the optimal classifier in view $\mathcal{X}_{3-v}$, since there exists the difference $D(c_1, c_2)$ between the two optimal classifiers in the two views. Thus, the hypothesis in $\mathcal{H}_v^C$ is not very close to the optimal classifier c_v.

3.3. Learning ϵ-Approximation of Optimal Classifier without Sampling Bias

To achieve good approximations of the optimal classifier, some prior knowledge about the optimal classifier needs to be known, which is shown as follows.

Assumption 1 ***(Information Assumption)*** *For $\big(\langle x_1, x_2 \rangle, y\big) \in \mathcal{X} \times \mathcal{Y}$, if view $\mathcal{X}_v$ provides much information about it, i.e., $|P(y = +1|x_v) - \frac{1}{2}| \geq \gamma_v'/2$, then the optimal classifier c_v in view $\mathcal{X}_v$ classifies it correctly, i.e., $sign(c_v(x_v)) = y$.*

Assumption 1 states that for an example if one view can provide much information about it, it will be correctly classified by the optimal classifier in this view. Thus, we give the following Theorem 10.

Theorem 10 *Suppose the hypothesis class $\mathcal{H}_v$ $(v = 1, 2)$ satisfies Definition 7, let $\mathcal{H}_v^L \subseteq \mathcal{H}_v$ denote the hypotheses minimizing the empirical risk on initial labeled data L, $R_v = \max_{h_v \in \mathcal{H}_v^L} R(h_v)$ and $\gamma_v = C_v^{\mathfrak{L}}(R_v - \eta_v) + \gamma_v'$. For $\epsilon \in (0, \frac{1}{2})$ and $\delta \in (0, 1)$, if Assumption 1 holds, $|U| = O(\frac{d_v \ln \frac{1}{\delta}}{\epsilon^2})$ and $\mu(\gamma_1, \gamma_2, \mathcal{H}_1^L, \mathcal{H}_2^L) = 1$, with probability $1 - \delta$ the output $\mathcal{H}_1^C$ and $\mathcal{H}_2^C$ in Algorithm 1 satisfy $R(\mathcal{H}_v^C) \leq \eta_v + \epsilon$.*

Proof For $x = \langle x_1, x_2 \rangle$, $\widehat{y}$ denote its pseudo-label. If $|h_v(x_v)| \geq \gamma_v = C_v^{\mathfrak{L}}(R_v - \eta_v) + \gamma_v^{'}$, with Definition 7 we know that h_v and c_v make the same prediction on x_v and $|c_v(x_v)| \geq \gamma_v^{'}$. So we get $\left|P(y = +1|x_v) - \frac{1}{2}\right| \geq \gamma_v^{'}/2$. Then with Assumption 1 we know that $\widehat{y} = sign\big(h_v(x_v)\big) = sign\big(c_v(x_v)\big) = y$. So we get that the pseudo-label of any example in σ_1 is the same as its underground label. Since $\mu(\gamma_1, \gamma_2, \mathcal{H}_1^L, \mathcal{H}_2^L) = 1$, we know that all unlabeled instances in U are assigned with underground labels and added into σ_1. So σ_1 is an i.i.d sample with size of $|L| + |U|$. Considering that $|U| = O\big(\frac{d_v \ln \frac{1}{\delta}}{\epsilon^2}\big)$, we get $\max_{h_v \in \mathcal{H}_v^C} R(h_v) \leq R(c_v) + \epsilon$ with probability $1 - \delta$. ■

Theorem 10 states that if the diversity with larger margins γ_1 and γ_2 between $\mathcal{H}_1^L$ and $\mathcal{H}_2^L$ trained on initial labeled data L is 1, co-training could output the ϵ-approximation of the optimal classifier.

3.4. Learning ϵ-Approximation of Optimal Classifier with Sampling Bias

However, in real applications $\mu(\gamma_1, \gamma_2, \mathcal{H}_1^L, \mathcal{H}_2^L)$ may be smaller than 1, i.e., not all unlabeled instances could be classified with large margin by weak hypotheses trained on initial labeled data L. With Definition 7 we know that the threshold γ_v $(v = 1, 2)$ which guarantees the quality of confident instances is related with the error rate of weak hypotheses. An intuitive way to get more confident instances to augment the training data is updating the weak hypotheses with newly labeled confident instances and adaptively decreasing the threshold of margin, which is shown in Algorithm 2. When $\mu(\gamma_1, \gamma_2, \mathcal{H}_1^L, \mathcal{H}_2^L)$ is smaller than 1, it will make co-training suffer from the limitation of sampling bias, since the training data in each view might not be an i.i.d sample from the marginal distribution $\mathbb{D}_{\mathcal{X}}$. Now we give the following definition to approximately bound the difference between two training samples.

Definition 11 ***(Approximate KL Divergence)*** *Let Ω be a large example set i.i.d sampled from the unknown distribution $\mathbb{D}$ and $\Lambda \subseteq \Omega$ be a set of examples, define the following $D_{AKL}(\Lambda \parallel \Omega)$ as an approximate KL divergence from the distribution generating Λ to distribution $\mathbb{D}$.*

$$\begin{aligned} D_{AKL}(\Lambda \parallel \Omega) &= \sum_{x^j \in \Omega} P\big(\mathbb{I}\{x^j \in \Lambda\}\big) \ln \frac{P(\mathbb{I}\{x^j \in \Lambda\})}{P(\mathbb{I}\{x^j \in \Omega\})} \\ &= \sum_{x^j \in \Lambda} \frac{1}{|\Lambda|} \ln \frac{1/|\Lambda|}{1/|\Omega|} + 0 \\ &= \ln \frac{|\Omega|}{|\Lambda|} \end{aligned}$$

Let us interpret Definition 11 intuitively. Ω is a large example set i.i.d sampled from the unknown distribution $\mathbb{D}$, so we use the uniform distribution over Ω as an approximation of $\mathbb{D}$. In this way we use the uniform distribution over Λ as an approximation of the distribution generating Λ and define $D_{AKL}(\Lambda \parallel \Omega)$ as an approximate KL divergence from the distribution generating Λ to distribution $\mathbb{D}$. With Proposition 6 we know that sampling bias might make co-training suffer, we give the following assumption to bound the influence of sampling bias.

Algorithm 2 Adaptive margin-based co-training

Input: Labeled data L, unlabeled data U, $\sigma_0 = L$, $n = |L| + |U|$, $m_0 = |L|$ and $\gamma_v^0 = C_v^{\mathfrak{L}}(R_v - \eta_v) + \gamma_v^{'}$
for $i = 0, 1, 2, \ldots$ **do**
  Get $\mathcal{H}_v^i \subseteq \mathcal{H}_v$ by minimizing the empirical risk on σ_i with respect to view $\mathcal{X}_v$ and set $T_i = \emptyset$
  for $x = \langle x_1, x_2 \rangle \in U$ **do**
    for $h_1 \in \mathcal{H}_1^i$, $h_2 \in \mathcal{H}_2^i$ **do**
      if $|h_1(x_1)| \geq \gamma_1^i$ **then**
        $T_i = T_i \cup (x, sign(h_1(x_1)))$ and delete x from U
        break
      end if
      if $|h_2(x_2)| \geq \gamma_2^i$ **then**
        $T_i = T_i \cup (x, sign(h_2(x_2)))$ and delete x from U
        break
      end if
    end for
  end for
  if $i = 0$ and $|T_0| > \sqrt[3]{n^2 m_0} - m_0$ **then**
    $\gamma_v^1 = \gamma_v^0 - C_v^{\mathfrak{L}}(R_v - \eta_v)(1 - \frac{n\sqrt{m_0}}{(m_0 + |T_0|)^{3/2}})$, $\sigma_1 = \sigma_0 \cup T_0$, $m_1 = m_0 + |T_0|$
  end if
  if $|T_0| \leq \sqrt[3]{n^2 m_0} - m_0$ or $T_i = \emptyset$ **then**
    return
  end if
  if $i \geq 1$ **then**
    $\gamma_v^{i+1} = \gamma_v^0 - C_v^{\mathfrak{L}}(R_v - \eta_v)(1 - \frac{n\sqrt{m_0}}{(m_i + |T_i|)^{3/2}})$, $\sigma_{i+1} = \sigma_i \cup T_i$, $m_{i+1} = m_i + |T_i|$
  end if
end for
Output: $\mathcal{H}_1^C = \mathcal{H}_1^i$ and $\mathcal{H}_2^C = \mathcal{H}_2^i$

Assumption 2 ***(Sampling Bias Assumption)*** *Let Ω be a large example set i.i.d sampled from the unknown distribution $\mathbb{D}$ and $\Lambda \subseteq \Omega$ be a set of examples. Let f_Λ denote the hypothesis minimizing the empirical risk on Λ, R^* be the error rate of the optimal classifier and $R^{'}$ be the upper bound on the error rate of the hypothesis minimizing the empirical risk on an i.i.d. sample with size of $|\Lambda|$ from distribution $\mathbb{D}$, then $R(f_\Lambda) - R^* \leq (R^{'} - R^*) \cdot \exp\big(D_{AKL}(\Lambda \parallel \Omega)\big)$.*

Assumption 2 states that the error difference between the classifier trained with biased sample and the optimal classifier can be upper-bounded by that between the classifier trained with unbiased sample and the optimal classifier times an exponential function of the approximate KL divergence. If Λ is a large part of Ω, $D_{AKL}(\Lambda \| \Omega)$ is close to 0 and f_Λ suffers little from sampling bias.

Let Ω be an i.i.d sample size of m, it is well-known (Anthony and Bartlett, 1999) that there exists an universal constant C such that for $\delta \in (0, 1)$ we have $R(h_v) - R(c_v) \leq$

$\sqrt{\frac{C}{m}(d_v+\ln(\frac{1}{\delta}))}$ with probability $1-\delta$ for any h_v minimizing the empirical risk on sample Ω. Generally, there may exist more than one hypothesis which have the same empirical risk. Let H_v^{Ω} denote the set of hypotheses which have the same minimum empirical risk on sample Ω, it is reasonable to assume that $\max_{h_v\in\mathcal{H}_v^{\Omega}} R(h_v)-R(c_v)=\sqrt{\frac{C}{m}\big(d_v+\ln(\frac{1}{\delta})\big)}$, which means the PAC-bound is tight and the maximum error rate of the hypotheses which minimize the empirical risk on sample Ω is proportional to $\frac{1}{\sqrt{m}}$. We are now ready to give the theorem on co-training with insufficient views when there is sampling bias in the training data.

Theorem 12 *Suppose the hypothesis class $\mathcal{H}_v$ $(v=1,2)$ satisfies Definition 7, let σ_i denote the training data in the i-th round of Algorithm 2, $\mathcal{H}_v^i\subseteq\mathcal{H}_v$ denote the set of hypotheses minimizing the empirical risk on the training data σ_i, $n=|L|+|U|$ and $R_v=\max_{h_v^0\in\mathcal{H}_v^0} R(h_v^0)$. For $\epsilon\in(0,\frac{1}{2})$ and $\delta\in(0,1)$, if Assumptions 1 and 2 hold, $|U|=O(\frac{d_v\ln\frac{1}{\delta}}{\epsilon^2})$, $\mu(\gamma_1^0,\gamma_2^0,\mathcal{H}_1^0,\mathcal{H}_2^0)>\frac{\sqrt[3]{n^2|L|-|L|}}{n-|L|}$ and $|T_i|>0$ for $i\geq 1$ until $|\sigma_i|=n$, with probability $1-\delta$ the output $\mathcal{H}_1^C$ and $\mathcal{H}_2^C$ in Algorithm 2 satisfy $R(\mathcal{H}_v^C)\leq\eta_v+\epsilon$.*

Proof Since L is an i.i.d sample and $m_0=|L|$, for the hypothesis set $\mathcal{H}_v^{r_i}$ minimizing the empirical risk on an i.i.d sample with size of $m_i=|\sigma_i|$, with the assumption that the maximum error rate of the hypotheses minimizing the empirical risk on the i.i.d sample Ω is proportional to $\frac{1}{\sqrt{|\Omega|}}$ we have

$$\max_{h_v^{r_i}\in\mathcal{H}_v^{r_i}} R(h_v^{r_i})-R(c_v)=\frac{\sqrt{m_0}}{\sqrt{m_i}}\Big(\max_{h_v^0\in\mathcal{H}_v^0} R(h_v^0)-R(c_v)\Big)=\frac{\sqrt{m_0}}{\sqrt{m_i}}(R_v-\eta_v).$$

If $\gamma_v^0=C_v^{\mathfrak{L}}(R_v-\eta_v)+\gamma_v^{'}$, with the proof in Theorem 10 we know that the pseudo-label of any example in σ_1 is the same as the underground label. Since $L\cup U$ is a large i.i.d sample from the marginal distribution $\mathbb{D}_{\mathcal{X}}$, so with Assumption 2 we get

$$\max_{h_v^1\in\mathcal{H}_v^1} R(h_v^1)-R(c_v)\leq\frac{\sqrt{m_0}}{\sqrt{m_1}}(R_v-\eta_v)\cdot\exp\big(\ln\frac{n}{m_1}\big).$$

For $h_v^1\in\mathcal{H}_v^1$, if

$$|h_v^1(x_v)|\geq\gamma_v^1=\gamma_v^0-C_v^{\mathfrak{L}}(R_v-\eta_v)\big(1-\frac{n\sqrt{m_0}}{m_1\sqrt{m_1}}\big),$$

with Definition 7 we get $|c_v(x_v)|\geq\gamma_v^{'}$ and $sign\big(h_v^1(x_v)\big)=sign\big(c_v(x_v)\big)$. With Assumption 1 we know $sign\big(c_v(x_v)\big)=y$. Thus, the pseudo-label of any example in σ_2 is the same as the underground label. Similarly, for $h_v^i\in\mathcal{H}_v^i$, if

$$|h_v^i(x_v)|\geq\gamma_v^i=\gamma_v^0-C_v^{\mathfrak{L}}(R_v-\eta_v)\big(1-\frac{n\sqrt{m_0}}{m_i\sqrt{m_i}}\big),$$

we get $sign\big(h_v^i(x_v)\big)=y$. If $T_i\neq\emptyset$ until $|\sigma_i|=|L|+|U|$, all instances in U are labeled with underground labels. So σ_i is an i.i.d sample with size of $|L|+|U|$. Since $|U|=O\big(\frac{d_v\ln\frac{1}{\delta}}{\epsilon^2}\big)$, we

get $\max_{h_v \in \mathcal{H}_v^C} R(h_v) \leq \eta_v + \epsilon$ with probability $1-\delta$. If we want $\gamma_v^1 < \gamma_v^0$, $1 - \frac{n\sqrt{m_0}}{m_1\sqrt{m_1}}$ must be larger than 0, i.e., $m_1 > \sqrt[3]{n^2 m_0}$. It implies that $m_0 + \mu(\gamma_1^0, \gamma_2^0, \mathcal{H}_1^0, \mathcal{H}_2^0)|U| > \sqrt[3]{n^2 m_0}$, so we get $\mu(\gamma_1^0, \gamma_2^0, \mathcal{H}_1^0, \mathcal{H}_2^0) > \frac{\sqrt[3]{n^2|L|}-|L|}{n-|L|}$. ■

Theorem 12 states that if the diversity with margins γ_1^0 and γ_2^0 between $\mathcal{H}_1^0$ and $\mathcal{H}_2^0$ trained on initial labeled set L is large ($\mathcal{H}_1^0$=$\mathcal{H}_1^L$, $\mathcal{H}_2^0$=$\mathcal{H}_2^L$), i.e., $\mu(\gamma_1^0, \gamma_2^0, \mathcal{H}_1^0, \mathcal{H}_2^0) > \frac{\sqrt[3]{n^2|L|}-|L|}{n-|L|}$, co-training could improve the performance of weak hypotheses by exploiting unlabeled data until the diversity between the two views becomes 0. This result tells that the diversity between the two views plays an important role in co-training with insufficient views, which is consistent with the theoretical analysis on co-training with sufficient views in Wang and Zhou (2007).

3.5. Insufficiency vs. Diversity

In this section we study what influence the view insufficiency will bring to the learning process. Since we could not know the distribution and the posterior probability $\varphi_v(x_v)$ $(v = 1, 2)$ of the example space in advance, it is difficult to analyze the general case. We focus on the famous Tsybakov condition case (Tsybakov, 2004) that for some finite $C_v^0 > 0$, $k > 0$ and $0 < t \leq 1/2$,

$$P_{\langle x_1, x_2 \rangle \in \mathbb{D}_{\mathcal{X}}}\big(|\varphi_v(x_v) - 1/2| \leq t\big) \leq C_v^0 t^k,$$

where small k implies large view insufficiency Υ, and give a heuristic analysis to illuminate the relationship between view insufficiency and diversity. Considering the worst case of Tsybakov condition for the fixed parameter k, i.e., $P_{\langle x_1, x_2 \rangle \in \mathbb{D}_{\mathcal{X}}}\big(\big|\varphi_v(x_v) - 1/2\big| \leq t\big) = C_v^0 t^k$, we get $P_{\langle x_1, x_2 \rangle \in \mathbb{D}_{\mathcal{X}}}\big(|2\varphi_v(x_v) - 1| > \gamma\big) = 1 - C_v^0(\frac{\gamma}{2})^k$ for $0 < \gamma \leq 1$. $\big|2\varphi_v(x_v) - 1\big|$ is the output margin of the optimal classifier c_v, following Definition 7 with the intuition that similar hypotheses tend to have similar margin output, the magnitude of the instances with margin larger than γ in view $\mathcal{X}_v$ is probably $\alpha\big(1 - C_v^0(\frac{\gamma}{2})^k\big)$ for some parameter α. Define $\mu_1(\gamma_1, \mathcal{F})$ and $\mu_2(\gamma_2, \mathcal{G})$ as follows,

$$\mu_1(\gamma_1, \mathcal{F}) = P\big\{x = \langle x_1, x_2 \rangle \in U : \exists f \in \mathcal{F} \ \text{ s.t. } \ |f(x_1)| \geq \gamma_1\big\}$$
$$\mu_2(\gamma_2, \mathcal{G}) = P\big\{x = \langle x_1, x_2 \rangle \in U : \exists g \in \mathcal{G} \ \text{ s.t. } \ |g(x_2)| \geq \gamma_2\big\}$$

and let $\nu(\gamma_1, \gamma_2, \mathcal{F}, \mathcal{G})$ denote the probability mass on the instances which are labeled with large margin just by one view. $\mu_v \approx \alpha(1 - C_v^0(\frac{\gamma_v}{2})^k)$ quantifies the amount of instances labeled with large margin by view $\mathcal{X}_v$ and ν can be thought of as a measurement of the different information provided by the two views. It is not difficult to find that the diversity μ can be expressed as $\mu = \frac{\mu_1 + \mu_2 + \nu}{2}$. It implies that when ν is fixed, if the view insufficiency increases, the diversity between the two views decreases. For understanding the magnitude of the diversity between the two views better, we give the following example. There often have adequate unlabeled instances in real semi-supervised applications, suppose we have n=$|L|$+$|U|$=1000 and L=12, similar to the empirical study on co-training in the paper of Blum and Mitchell (1998), the diversity $\frac{\sqrt[3]{n^2|L|}-|L|}{n-|L|}$ at the first step in Theorem 12 should be 22%. With respect to μ=$(\mu_1$+μ_2+$\nu)/2$, for *diverse views* (ν is large), weak hypotheses

in each view predicting about 18% (even less) of the unlabeled instances with large margin might be enough to guarantee the 22% diversity, which is common in real applications.

3.6. Assumption Relaxation and Discussions

Our result is based on a little bit strong *Margin Lipschitz* assumption, which is caused by the fact that the learning task with insufficient views for semi-supervised learning is very difficult. As Ben-David et al. (2008) showed, without assumption on the relationship between hypothesis class and unlabeled data distribution, unlabeled data has limited usefulness. In this section we try to give a heuristic analysis for the case where the *Margin Lipschitz* assumption is relaxed. Instead, we give the following *Probabilistic Margin* assumption: for $\frac{1}{2} \leq \gamma_v \leq 1$ $(v = 1, 2)$,

$$P_{\langle x_1, x_2\rangle \in \mathbb{D}_{\mathcal{X}}}\left\{x_v : |h_v(x_v)| \geq \gamma_v \wedge sign(h_v(x_v)) \neq y\right\} \leq \phi(\gamma_v).$$

Here $\phi : [\frac{1}{2}, 1] \to [0, 1]$ is a monotonically decreasing function, e.g., $\phi(\gamma) = \beta \ln(\frac{1}{\gamma})$ for some parameter β. *Probabilistic Margin* assumption allows for small label noise in the examples labeled with large margin. Considering the worst case of the influence of label noise, i.e., the examples with noisy labels are completely inconsistent with the optimal classifiers, it can be found that when the two views have large diversity, co-training could output the hypotheses whose error rate are close to $\eta_v + \beta \ln(\frac{1}{\gamma_v})$, which could be smaller than the error rate of the classifier trained only on the small initial labeled data set L for some large γ_v. This shows that co-training could improve the learning performance by exploiting unlabeled data even with insufficient views.

In our result, the margin threshold $\gamma_v = C_v^{\mathfrak{L}}(R_v - \eta_v) + \gamma_v^{'}$ depends on several parameters. Generally, the optimal classifier would make mistakes only when the instances are close to the boundary, i.e., $P(y{=}+1|x)$ is close to $1/2$. So $\gamma_v^{'}$ is close to 0. $(R_v - \eta_v)$ depends on the number of initial labeled data L and is proportional to $\frac{1}{\sqrt{|L|}}$. So when $|L| \approx 4(C_v^{\mathfrak{L}})^2 C\big(d_v + \ln(\frac{1}{\delta})\big)$, $C_v^{\mathfrak{L}}(R_v - \eta_v)$ is close to $1/2$. Thus, γ_v can be close to $1/2$. In traditional co-training style algorithms, it often allows one classifier to label its most confident instance for the other in each round. Using this method can avoid setting the margin threshold, but might bring large sampling bias into the training data.

Ando and Zhang (2007) proposed a two-view model using unlabeled data to learn effective feature representations in the two views and then finding the optimum predictor as a linear combination of the features constructed from unlabeled data, which is different from co-training in the way of exploiting unlabeled data. Moreover, this two-view model is based on the assumption that the two views are conditionally independent given the class label, which is overly strong to satisfy in real applications.

4. Conclusions

We present the theoretical analysis on co-training with insufficient views in this paper, especially when the two views provide diverse information. We indicate that co-training might suffer from two limitations, i.e., *label noise* and *sampling bias*, when each view is insufficient. We also define the diversity between the two views with respect to the confidence of prediction and prove that when the two views have large diversity, co-training suffers little

from the two limitations and could succeed in outputting the approximation of the optimal classifier by exploiting unlabeled data even with insufficient views. Our result shows that the diversity between the two views is very important for co-training. It is possible to develop new algorithms based on the theoretical result.

This paper might contribute to understanding of the difference between co-regularization and co-training. For *similar views*, Sridharan and Kakade (2008) presented a framework for co-regularization; for *diverse views*, we show that co-regularization may fail while co-training which iteratively utilizes the confident information in one view to help the other is a good learning strategy. In this paper, we focus on the case where the information provided by each view is insufficient. Sometimes in real applications the views may be incomplete, i.e., the features of some examples are not available. It would be interesting to extend our work to co-training with incomplete views.

Acknowledgments

This research was partially supported by the National Fundamental Research Program of China (2010CB327903), the National Science Foundation of China (61305067, 61273301), and Huawei Fund (YBCB2012085).

References

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

R. K. Ando and T. Zhang. Two-view feature generation model for semi-supervised learning. In *Proceedings of the the 24th International Conference on Machine Learning*, pages 25–32, Corvallis, OR, 2007.

M. Anthony and P. Bartlett, editors. *Neural Network Learning: Theoretical Foundations*. Cambridge University Press, Cambridge, 1999.

M.-F. Balcan and A. Blum. A discriminative model for semi-supervised learning. *Journal of the ACM*, 57(3), 2010.

M.-F. Balcan, A. Blum, and K. Yang. Co-training and expansion: Towards bridging theory and practice. In *Advances in Neural Information Processing Systems 17*, pages 89–96. MIT Press, Cambridge, MA, 2005.

S. Ben-David, T. Lu, and D. Pal. Does unlabeled data provably help? worst-case analysis of the sample complexity of semi-supervised learning. In *Proceedings of the 21st Annual Conference on Learning Theory*, pages 33–44, Helsinki, Finland, 2008.

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.

U. Brefeld, T. Gärtner, T. Scheffer, and S. Wrobel. Efficient co-regularised least squares regression. In *Proceedings of the the 23th International Conference on Machine Learning*, pages 137–144, Pittsburgh, PA, 2006.

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

S. Dasgupta, M. Littman, and D. McAllester. PAC generalization bounds for co-training. In *Advances in Neural Information Processing Systems 14*, pages 375–382. MIT Press, Cambridge, MA, 2002.

L. Devroye, L. Györfi, and G. Lugosi, editors. *A Probabilistic Theory of Pattern Recognition*. Springer, New York, 1996.

J. Farquhar, D. R. Hardoon, H. Meng, J. Shawe-Taylor, and S. Szedmák. Two view learning: Svm-2k, theory and practice. In *Advances in Neural Information Processing Systems 18*. MIT Press, Cambridge, MA, 2006.

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.

V. Sindhwani, P. Niyogi, and M. Belkin. A co-regularization approach to semi-supervised learning with multiple views. In *Working Notes of the ICML'05 Workshop on Learning with Multiple Views*, Bonn, Germany, 2005.

K. Sridharan and S. M. Kakade. An information theoretic framework for multi-view learning. In *Proceedings of the 21st Annual Conference on Learning Theory*, pages 403–414, Helsinki, Finland, 2008.

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.

A. Tsybakov. Optimal aggregation of classifiers in statistical learning. *The Annals of Statistics*, 32(1):135–166, 2004.

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.

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.

Z.-H. Zhou and M. Li. Semi-supervised learning by disagreement. *Knowledge and Information Systems*, 24(3):415–439, 2010.

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.

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