On Multi-View Active Learning and the Combination with Semi-Supervised Learning

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

Multi-view learning has become a hot topic during the past few years. In this paper, we first characterize the sample complexity of multi-view *active learning*. Under the α-*expansion* assumption, we get an exponential improvement in the sample complexity from usual $\widetilde{O}(\frac{1}{\epsilon})$ to $\widetilde{O}(\log\frac{1}{\epsilon})$, requiring neither strong assumption on data distribution such as the data is distributed uniformly over the unit sphere in $\mathbb{R}^d$ nor strong assumption on hypothesis class such as linear separators through the origin. We also give an upper bound of the error rate when the α-*expansion* assumption does not hold. Then, we analyze the combination of multi-view active learning and semi-supervised learning and get a further improvement in the sample complexity. Finally, we study the empirical behavior of the two paradigms, which verifies that the combination of multi-view active learning and semi-supervised learning is efficient.

1. Introduction

Learning from labeled data is well-established in machine learning, but labeling the training data is time consuming, sometimes may be very expensive since it may need human efforts. In many machine learning applications, unlabeled data can often be obtained abundantly and cheaply, so there has recently been substantive interest in using large amount of unlabeled data together with labeled data to achieve better learning performance.

There are two popular paradigms for using unlabeled data to complement labeled data. One is *semi-supervised learning*. Some approaches use a generative model for the classifier and employ EM to model the label estimation or parameter estimation process (Dempster et al., 1977; Miller & Uyar, 1997; Nigam et al., 2000); some approaches use the unlabeled data to regularize the learning process in various ways, e.g., defining a graph on the data set and then enforcing the label smoothness over the graph as a regularization term (Belkin et al., 2001; Zhu et al., 2003; Zhou et al., 2005); some approaches use the *multi-view* setting to train learners and then let the learners to label unlabeled examples (Blum & Mitchell, 1998; Goldman & Zhou, 2000; Zhou & Li, 2005). The multi-view setting is first formalized by Blum and Mitchell (1998), where there are several disjoint subsets of features (each subset is called as a *view*), each of which is sufficient for learning the target concept. For example, the web page classification task has two views, i.e., the text appearing on the page itself and the anchor text attached to hyper-links pointing to this page (Blum & Mitchell, 1998); the speech recognition task also has two views, i.e., sound and lip motion (de Sa & Ballard, 1998).

Another important paradigm for using unlabeled data to complement labeled data, which is the focus of this paper, is *active learning* (Cohn et al., 1994; Freund et al., 1997; Tong & Koller, 2001; Melville & Mooney, 2004). In active learning, the learners actively ask the user to label the *most informative* examples and hope to learn a good classifier with as few labeled examples as possible.

There have been many theoretical analyses on the sample complexity of *single-view* active learning. For some simple learning tasks the sample complexity of active learning can be $O(\log\frac{1}{\epsilon})$ which is exponentially improved in contrast to $O(\frac{1}{\epsilon})$ of *passive learning* taking into account the desired accuracy bound ϵ. Unfortunately, such an exponential improvement is not always achievable in active learning. Dasgupta (2006) illustrated that if the hypothesis class $\mathcal{H}$ is linear separators in $\mathbb{R}^2$ and if the data distribution is some density

Appearing in *Proceedings of the 25th International Conference on Machine Learning*, Helsinki, Finland, 2008.

supported on the perimeter of the unit circle, there are some target hypotheses in $\mathcal{H}$ for which $\Omega(\frac{1}{\epsilon})$ labels are needed to find a classifier with error rate less than ϵ, no matter what active learning approach is used. Under the strong assumptions that the hypothesis class is linear separators through the origin, that the data is distributed uniformly over the unit sphere in $\mathbb{R}^d$, and that the learning task is a *realizable* case (i.e., there exists a hypothesis perfectly separating the data), the sample complexity of active learning is $\widetilde{O}(d \log \frac{1}{\epsilon})$ taking into account the desired accuracy bound ϵ (Freund et al., 1997; Dasgupta et al., 2005) [1]. For some known data distribution $\mathbb{D}$ and specific hypothesis class, Dasgupta (2006) gave the coarse sample complexity bounds for realizable active learning. The study of sample complexity of active learning for realizable case without strong assumptions on the data distribution and the hypothesis class remains an open problem.

All the above results were obtained under the *single-view* setting. The first algorithm for active learning in multi-view setting is *co-testing* (Muslea et al., 2000; Muslea et al., 2006). It focuses on the set of *contention points* (i.e., unlabeled examples on which different views predict different labels) and asks the user to label some of them. This is somewhat related to Query-by-Committee (Freund et al., 1997) since *co-testing* also uses more than one learners to identify the most informative unlabeled examples to query, but the typical Query-by-Committee works under a single-view setting while *co-testing* exploits the multi-views explicitly. It was reported that *co-testing* outperforms existing active learners on a variety of real-world domains such as wrapper induction, Web page classification, advertisement removal and discourse tree parsing. To the best of our knowledge, however, there is no theoretical result on the sample complexity of multi-view active learning.

In this paper, we first theoretically analyze the sample complexity of multi-view active learning under the α-*expansion* assumption which is first mentioned by Balcan et al. (2005) and prove that the sample complexity of multi-view active learning can be exponentially improved to $\widetilde{O}(\log \frac{1}{\epsilon})$. A clear advantage is that we do not use strong assumptions which were employed in most previous studies, such as the hypothesis class is linear separators through the origin and the data is distributed uniformly over the unit sphere in $\mathbb{R}^d$. In case the α-*expansion* assumption does not hold, we give an upper bound of the error rate. Second, we analyze the combination of multi-view active learning and semi-supervised learning and get an further improvement in the sample complexity. Finally, we study the empirical behavior of the two paradigms, which verifies that the combination of multi-view active learning and semi-supervised learning is more efficient than pure multi-view active learning.

The rest of this paper is organized as follows. After introducing some preliminaries in Section 2, we analyze the sample complexity of multi-view active learning in Section 3. Then we analyze the sample complexity of the combination of multi-view active learning and semi-supervised learning in Section 4 and study the empirical behavior in Section 5. Finally we conclude the paper in Section 6.

2. Preliminaries

In the multi-view setting, an example x is described with several different disjoint sets of features. Without loss of generality, in this paper we only consider the *two-view* setting for the sake of simplicity. Suppose that the example space $X = X_1 \times X_2$ is with some unknown distribution $\mathbb{D}$, X_1 and X_2 are the two views, and $Y = \{-1, 1\}$ is the label space. Let $c = (c_1, c_2)$ be the underlying target concept, where c_1 and c_2 are the underlying target concepts in the two views, respectively. Suppose that the example space is consistent, that is, there is no such example $x = (x_1, x_2)$ that $c_1(x_1) \neq c_2(x_2)$ in X. Let $\mathcal{H}_1$ and $\mathcal{H}_2$ be the hypothesis class in each view, respectively. For any $h_j \in \mathcal{H}_j$ and $x = (x_1, x_2)$ we say $x_j \in h_j$ if and only if $h_j(x_j) = c_j(x_j)$ $(j = 1, 2)$. In this way any hypothesis in $\mathcal{H}_j$ can be thought of as a subset of X_j.

In each round of iterative multi-view active learning, the learners ask the user to label some unlabeled examples and add them into the labeled training data. These newly labeled examples provide more information about the data distribution. In this paper, we consider the *co-testing*-like Paradigm 1 described in Table 1. In Paradigm 1, the learners ask the user to label some contention points to refine the classifiers. If the confident set of each view is expanding by considering the other view together, Paradigm 1 may succeed. Intuitively, we can use the α-*expansion* assumption to analyze the process.

Suppose $S_1 \subseteq X_1$ and $S_2 \subseteq X_2$ denote the examples that are correctly classified in each view, respectively. Let $Pr(\mathbf{S}_1 \wedge \mathbf{S}_2)$ denote the probability mass on examples that are correctly classified in both views, while $Pr(\mathbf{S}_1 \oplus \mathbf{S}_2)$ denotes the probability mass on examples that are correctly classified only in one view (i.e., examples disagreed by the two classifiers). Now we give

[1] The $\widetilde{O}$ notation is used to hide factors $\log\log(\frac{1}{\epsilon})$, $\log(d)$ and $\log(\frac{1}{\delta})$

Input:

Unlabeled data set $\mathcal{U} = \{x^1, x^2, \cdots, \}$, where each example x^t is given as a pair (x_1^t, x_2^t)

Process:

Ask the user to label m_0 unlabeled examples drawn randomly from $\mathbb{D}$ to compose the labeled data set $\mathcal{L}$

Iterate $i = 0, 1, \cdots, s$

Train two classifiers h_1^i and h_2^i consistent with $\mathcal{L}$ in each view, respectively;

Apply h_1^i and h_2^i to the unlabeled data set $\mathcal{U}$ and find out the contention points set $\mathcal{Q}_i$;

Ask the user to label m_{i+1} unlabeled examples drawn randomly from $\mathcal{Q}_i$, then add them into $\mathcal{L}$ and delete them from $\mathcal{U}$.

Output:

$h_{final} = \text{combine}(h_1^s, h_2^s)$

Table 1. Paradigm 1: Multi-view active learning

our definition on *α-expansion*.

Definition 1 *$\mathbb{D}$ is α-expansion if for any $S_1 \subseteq X_1$, $S_2 \subseteq X_2$, we have*

$$Pr(\boldsymbol{S}_1 \oplus \boldsymbol{S}_2) \geq \alpha \min[Pr(\boldsymbol{S}_1 \wedge \boldsymbol{S}_2), Pr(\overline{\boldsymbol{S}_1} \wedge \overline{\boldsymbol{S}_2})].$$

We say that $\mathbb{D}$ is α-expanding with respect to hypothesis class $\mathcal{H}_1 \times \mathcal{H}_2$ if the above holds for all $S_1 \in \mathcal{H}_1 \cap X_1$, $S_2 \in \mathcal{H}_2 \cap X_2$ (here we denote by $\mathcal{H}_j \cap X_j$ the set $\{h \cap X_j : h \in \mathcal{H}_j\}$ for $j = 1, 2$).

Note that Definition 1 on *α-expansion* is almost the same as that in Balcan et al. (2005). To guarantee the success of iterative *co-training*, they made several assumptions such as that the learning algorithm used in each view is confident about being positive and is able to learn from positive examples only, and that the distribution $\mathbb{D}^+$ over positive examples is expanding. There are many concept classes, however, are not learnable from positive examples only. Apparently, all problems which satisfy the definition of Balcan et al. (2005) also satisfy our definition.

We will make use of the following lemma when deriving our sample complexity bound (Anthony & Bartlett, 1999).

Lemma 1 *Let $\mathcal{H}$ be a set of functions from X to $\{-1, 1\}$ with finite VC-dimension $V \geq 1$. Let $\mathbb{P}$ be an arbitrary, but fixed probability distribution over $X \times \{-1, 1\}$. For any ϵ, $\delta > 0$, if we draw a sample from $\mathbb{P}$ of size $N(\epsilon, \delta) = \frac{1}{\epsilon}(4V \log(\frac{1}{\epsilon}) + 2\log(\frac{2}{\delta}))$, then with probability $1 - \delta$, all hypotheses with error $\geq \epsilon$ are inconsistent with the data.*

3. Sample Complexity of Multi-View Active Learning

There are many strategies to combine the classifiers in Paradigm 1, for example, *weighted voting*, *majority voting* or *winner-take-all* (Muslea et al., 2006). In this paper, we use the following simple combination scheme for binary classification:

$$h_{com}^i(x) = \begin{cases} h_1^i(x_1) & \text{if } h_1^i(x_1) = h_2^i(x_2) \\ \textit{random guess} & \text{if } h_1^i(x_1) \neq h_2^i(x_2) \end{cases} \quad (1)$$

Assuming that the data distribution $\mathbb{D}$ is α-expanding with respect to hypothesis class $\mathcal{H}_1 \times \mathcal{H}_2$, we will analyze how many labels the user should label to achieve classifiers with error rate no larger than ϵ. We consider the iterative process and let $S_1^i \subseteq X_1$ and $S_2^i \subseteq X_2$ where S_1^i and S_2^i corresponds to the classifiers $h_1^i \in \mathcal{H}_1$ and $h_2^i \in \mathcal{H}_2$ in the i-th round, respectively. The initial m_0 unlabeled examples are randomly picked from $\mathbb{D}$ and labeled by the user according to the target concept c. Suppose m_0 is sufficient for learning two classifiers h_1^0 and h_2^0 whose error rates are at most $1/4$ (i.e., $Pr(\mathbf{S}_1^0) \geq 1 - 1/4$ and $Pr(\mathbf{S}_2^0) \geq 1 - 1/4$), and thus $Pr(\mathbf{S}_1^0 \wedge \mathbf{S}_2^0) \geq 1/2$. The *α-expansion* condition suggests

$$Pr(\mathbf{S}_1^0 \oplus \mathbf{S}_2^0) \geq \alpha Pr(\overline{\mathbf{S}_1^0} \wedge \overline{\mathbf{S}_2^0}).$$

In each round of Paradigm 1, the learners ask the user to label some unlabeled examples according to the target concept c and add them into the labeled data set. Then the two classifiers are refined. Some example x in X might be predicted with different labels between the i-th and $(i+1)$-th round. Intuitively, in order to get the classifiers improved in Paradigm 1, the reduced size of confident set should be no more than the size of contention set. Moreover, considering that there is no noise in the labeled data since all the labels are given by the user according to the target concept, and that the amount of labeled training examples are monotonically increasing, the asymptotic performance of PAC learners increase, we can assume that

$$Pr(\overline{\mathbf{S}_j^{i+1}} \mid \mathbf{S}_1^i \wedge \mathbf{S}_2^i) \leq \frac{\alpha Pr(\mathbf{S}_1^i \oplus \mathbf{S}_2^i)}{16 Pr(\mathbf{S}_1^i \wedge \mathbf{S}_2^i)} \quad (j \in \{1, 2\}) \quad (2)$$

Intuitively, by multiplying the denominator at the right-hand to the left-hand (16 is used for a faster convergence; it can be 2 for an easier understanding), Eq. 2 implies that the total reduced size of confident sets on both views after using the newly labeled contention points is no more than the size of contention set. Apparently, all problems that satisfy the assumption of Balcan et al. (2005) also satisfy Eq. 2. Now we give our main theorem.

Theorem 1 *For data distribution $\mathbb{D}$ α-expanding with respect to hypothesis class $\mathcal{H}_1 \times \mathcal{H}_2$, let ϵ and δ denote the final desired accuracy and confidence parameters. If $s = \lceil \frac{\log \frac{\alpha}{8\epsilon}}{\log \frac{1}{C}} \rceil$ and $m_i = \frac{16}{\alpha}(4V \log(\frac{16}{\alpha}) + 2\log(\frac{8(s+1)}{\delta}))$ $(i = 0, 1, \cdots, s)$, Paradigm 1 will generate a classifier with error rate no more than ϵ with probability $1 - \delta$.*

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

Proof. In Paradigm 1, we use Eq. 1 to combine the two classifiers, thus the error rate of the combined classifier h^i_{com} is

$$\begin{aligned} error_{h^i_{com}} &= Pr(\overline{\mathbf{S}^i_1} \wedge \overline{\mathbf{S}^i_2}) + \frac{1}{2} Pr(\mathbf{S}^i_1 \oplus \mathbf{S}^i_2) \\ &\leq Pr(\overline{\mathbf{S}^i_1} \wedge \overline{\mathbf{S}^i_2}) + Pr(\mathbf{S}^i_1 \oplus \mathbf{S}^i_2) \\ &= Pr(\overline{\mathbf{S}^i_1 \wedge \mathbf{S}^i_2}) \end{aligned}$$

With $m_0 = \frac{16}{\alpha}(4V \log(\frac{16}{\alpha}) + 2\log(\frac{8(s+1)}{\delta}))$, using Lemma 1 we have $Pr(\overline{\mathbf{S}^0_1}) \leq \frac{\alpha}{16}$ and $Pr(\overline{\mathbf{S}^0_2}) \leq \frac{\alpha}{16}$ with probability $1 - \frac{\delta}{4(s+1)}$. Generally, we have that an arbitrary S^i_j $(j = 1, 2)$ being consistent with the examples in $\mathcal{L}$ has an error rate at most $\frac{\alpha}{16}$ with probability $1 - \frac{\delta}{4(s+1)}$. So we have $Pr(\mathbf{S}^i_1 \wedge \mathbf{S}^i_2) \geq 1 - \frac{\alpha}{8}$ with probability $1 - \frac{\delta}{2(s+1)}$. Without loss of generality, consider $0 < \alpha \leq 1$ and therefore $1 - \frac{\alpha}{8} > \frac{1}{2}$. Thus the α-*expansion* condition suggests

$$Pr(\mathbf{S}^i_1 \oplus \mathbf{S}^i_2) \geq \alpha Pr(\overline{\mathbf{S}^i_1} \wedge \overline{\mathbf{S}^i_2}). \tag{3}$$

For $i \geq 1$, the learners ask the user to label m_i unlabeled examples drawn randomly from $S^{i-1}_1 \oplus S^{i-1}_2$ according to the target concept c and obtain two new classifiers S^i_1 and S^i_2. Similarly, if $m_i = \frac{16}{\alpha}(4V \log(\frac{16}{\alpha}) + 2\log(\frac{8(s+1)}{\delta}))$, using Lemma 1 we have

$$Pr(\overline{\mathbf{S}^i_j} \mid \mathbf{S}^{i-1}_1 \oplus \mathbf{S}^{i-1}_2) \leq \frac{\alpha}{16} \quad (j \in \{1, 2\})$$

with probability $1 - \frac{\delta}{4(s+1)}$. So we get that

$$Pr(\overline{\mathbf{S}^i_1 \wedge \mathbf{S}^i_2} \mid \mathbf{S}^{i-1}_1 \oplus \mathbf{S}^{i-1}_2) \leq \frac{\alpha}{8}$$

with probability $1 - \frac{\delta}{2(s+1)}$. Considering Eq. 2 we have

$$Pr(\overline{\mathbf{S}^i_1 \wedge \mathbf{S}^i_2} \mid \mathbf{S}^{i-1}_1 \wedge \mathbf{S}^{i-1}_2) \leq \frac{\alpha Pr(\mathbf{S}^{i-1}_1 \oplus \mathbf{S}^{i-1}_2)}{8 Pr(\mathbf{S}^{i-1}_1 \wedge \mathbf{S}^{i-1}_2)} .$$

Since

$$\begin{aligned} Pr(\overline{\mathbf{S}^i_1 \wedge \mathbf{S}^i_2}) = {} & Pr(\overline{\mathbf{S}^i_1 \wedge \mathbf{S}^i_2} \mid \overline{\mathbf{S}^{i-1}_1} \wedge \overline{\mathbf{S}^{i-1}_2}) \\ & \cdot Pr(\overline{\mathbf{S}^{i-1}_1} \wedge \overline{\mathbf{S}^{i-1}_2}) \\ & + Pr(\overline{\mathbf{S}^i_1 \wedge \mathbf{S}^i_2} \mid \mathbf{S}^{i-1}_1 \oplus \mathbf{S}^{i-1}_2) \\ & \cdot Pr(\mathbf{S}^{i-1}_1 \oplus \mathbf{S}^{i-1}_2) \\ & + Pr(\overline{\mathbf{S}^i_1 \wedge \mathbf{S}^i_2} \mid \mathbf{S}^{i-1}_1 \wedge \mathbf{S}^{i-1}_2) \\ & \cdot Pr(\mathbf{S}^{i-1}_1 \wedge \mathbf{S}^{i-1}_2), \end{aligned}$$

we have

$$Pr(\overline{\mathbf{S}^i_1 \wedge \mathbf{S}^i_2}) \leq \frac{\alpha}{4} Pr(\mathbf{S}^{i-1}_1 \oplus \mathbf{S}^{i-1}_2) + Pr(\overline{\mathbf{S}^{i-1}_1} \wedge \overline{\mathbf{S}^{i-1}_2}) .$$

From Eq. 3 we can get that

$$Pr(\overline{\mathbf{S}^{i-1}_1} \wedge \overline{\mathbf{S}^{i-1}_2}) \leq Pr(\mathbf{S}^{i-1}_1 \oplus \mathbf{S}^{i-1}_2)/\alpha .$$

Thus, considering

$$Pr(\overline{\mathbf{S}^{i-1}_1 \wedge \mathbf{S}^{i-1}_2}) = Pr(\mathbf{S}^{i-1}_1 \oplus \mathbf{S}^{i-1}_2) + Pr(\overline{\mathbf{S}^{i-1}_1} \wedge \overline{\mathbf{S}^{i-1}_2}),$$

we have

$$\begin{aligned} & \frac{Pr(\overline{\mathbf{S}^i_1 \wedge \mathbf{S}^i_2})}{Pr(\overline{\mathbf{S}^{i-1}_1 \wedge \mathbf{S}^{i-1}_2})} \\ \leq {} & \frac{\frac{\alpha}{4} Pr(\mathbf{S}^{i-1}_1 \oplus \mathbf{S}^{i-1}_2) + Pr(\overline{\mathbf{S}^{i-1}_1} \wedge \overline{\mathbf{S}^{i-1}_2})}{Pr(\mathbf{S}^{i-1}_1 \oplus \mathbf{S}^{i-1}_2) + Pr(\overline{\mathbf{S}^{i-1}_1} \wedge \overline{\mathbf{S}^{i-1}_2})} \\ \leq {} & \frac{\frac{\alpha}{4} Pr(\mathbf{S}^{i-1}_1 \oplus \mathbf{S}^{i-1}_2) + Pr(\mathbf{S}^{i-1}_1 \oplus \mathbf{S}^{i-1}_2)/\alpha}{Pr(\mathbf{S}^{i-1}_1 \oplus \mathbf{S}^{i-1}_2) + Pr(\mathbf{S}^{i-1}_1 \oplus \mathbf{S}^{i-1}_2)/\alpha} \\ = {} & \frac{\alpha/4 + 1/\alpha}{1 + 1/\alpha} . \end{aligned}$$

Now we get

$$\begin{aligned} Pr(\overline{\mathbf{S}^s_1 \wedge \mathbf{S}^s_2}) &\leq (\frac{\alpha/4 + 1/\alpha}{1 + 1/\alpha})^s Pr(\overline{\mathbf{S}^0_1 \wedge \mathbf{S}^0_2}) \\ &\leq \frac{\alpha}{8} (\frac{\alpha/4 + 1/\alpha}{1 + 1/\alpha})^s . \end{aligned}$$

So when $s = \lceil \frac{\log \frac{\alpha}{8\epsilon}}{\log \frac{1}{C}} \rceil$ where C is a constant and $\frac{\alpha/4+1/\alpha}{1+1/\alpha} < 1$, we have $Pr(\overline{\mathbf{S}^s_1 \wedge \mathbf{S}^s_2}) \leq \epsilon$. In other words, we get a classifier h^s_{com} whose error rate is no more than ϵ with probability $1 - \delta$. □

From Theorem 1 we know that we only need to label $\sum_{i=0}^{s} m_i = O(\log \frac{1}{\epsilon} \log(\log \frac{1}{\epsilon}))$ examples to get a classifier with error rate no more than ϵ with probability

$1-\delta$. Thus, we achieve an exponential improvement in sample complexity from $\widetilde{O}(\frac{1}{\epsilon})$ to $\widetilde{O}(\log\frac{1}{\epsilon})$ as in Dasgupta et al. (2005) and Balcan et al. (2007). Note that we have not assumed a specific data distribution and a specific hypothesis class which were assumed in the studies of Dasgupta et al. (2005) and Balcan et al. (2007). From the proof of Theorem 1 we can also know that the proportion $\frac{\alpha}{16}$ in Eq. 2 can be relaxed to close to $\frac{\alpha}{2}$. Such relaxation will not affect the exponential improvement, but will reduce the convergence speed.

Further, considering that not every data distribution $\mathbb{D}$ is α-expanding with respect to hypothesis class $\mathcal{H}_1 \times \mathcal{H}_2$, we will give a coarse upper bound of the generalization error for Paradigm 1 for cases when the α-*expansion* assumption does not hold.

Let $Pr(\mathbf{S}_1^i \oplus \mathbf{S}_2^i) = \alpha_i Pr(\overline{\mathbf{S}_1^i} \wedge \overline{\mathbf{S}_2^i})$ $(i = 0, 1, \cdots)$. If the α-*expansion* assumption does not hold in Paradigm 1, for any $\epsilon > 0$ and any integer $N > 0$, the size of the set $\{\alpha_i\colon\ i > N \wedge \alpha_i < \epsilon\}$ is infinite. We set a parameter $\epsilon_c > 0$ as the stop condition. When $Pr(\mathbf{S}_1^i \oplus \mathbf{S}_2^i)$ is less than ϵ_c, we terminate the iteration in Paradigm 1. Now we make the definition on *expanded region with respect to* ϵ_c.

Definition 2 *Let γ_{ϵ_c} denote the expanded region with respect to ϵ_c in Paradigm 1,*

$$\gamma_{\epsilon_c} = Pr(\overline{\boldsymbol{S}_1^0} \wedge \overline{\boldsymbol{S}_2^0}) - Pr(\overline{\boldsymbol{S}_1^i} \wedge \overline{\boldsymbol{S}_2^i}),$$

where $i = \min\{i : Pr(\boldsymbol{S}_1^i \oplus \boldsymbol{S}_2^i) < \epsilon_c \wedge i \geq 1\}$.

After i rounds the region in which both classifiers wrongly predict becomes smaller and smaller, from $Pr(\overline{\mathbf{S}_1^0} \wedge \overline{\mathbf{S}_2^0})$ to $Pr(\overline{\mathbf{S}_1^i} \wedge \overline{\mathbf{S}_2^i})$. This *expanded region* can be thought of as an approximation of $\Sigma_{k=1}^{i} Pr(\mathbf{S}_1^k \oplus \mathbf{S}_2^k)$.

Theorem 2 *When the α-expansion assumption does not hold, set $\epsilon_c > 0$ to terminate Paradigm 1. The error rate of h_{com}^i can be smaller than h_{com}^0 for $\gamma_{\epsilon_c} + \frac{1}{2}(Pr(\boldsymbol{S}_1^0 \oplus \boldsymbol{S}_2^0) - \epsilon_c)$.*

Proof. Considering $error_{h_{com}^i} = Pr(\overline{\mathbf{S}_1^i} \wedge \overline{\mathbf{S}_2^i}) + \frac{1}{2}Pr(\mathbf{S}_1^i \oplus \mathbf{S}_2^i)$ and $Pr(\mathbf{S}_1^i \oplus \mathbf{S}_2^i) < \epsilon_c$, we have that $error_{h_{com}^0} - error_{h_{com}^i}$ is larger than $\gamma_{\epsilon_c} + \frac{1}{2}(Pr(\mathbf{S}_1^0 \oplus \mathbf{S}_2^0) - \epsilon_c)$. □

Theorem 2 implies that Paradigm 1 could not boost the performance to arbitrarily high and gives a coarse upper bound of the error rate, when the α-*expansion* assumption does not hold. The improvement depends on the *expanded region* γ and the disagreement between the initial two classifiers. The larger the *expanded region* γ, the better the improvement of Paradigm 1. Theorem 2 can also be applied to one-shot *co-training* (Balcan et al., 2005).

4. Sample Complexity of Combination of Multi-View Active Learning and Semi-Supervised Learning

We can try to reduce the sample complexity further by combining multi-view active learning with semi-supervised learning. Previously this has been tried in some applications and led to good results (Zhou et al., 2006), yet to the best of our knowledge, there is no theoretical analysis which supports such argument. For computational simplicity, we consider the following case in this section. Suppose that the hypothesis class $\mathcal{H}_j$ is the subset of mappings from X_j to $[-1, 1]$ and $y = sign(c(x))$, $c = (c_1, c_2)$ is the underlying target concept, where c_1 and c_2 is the underlying target concept in each view, respectively. Let $d(f, g)$ denote the probability that the two classifiers $f \in \mathcal{H}_j$ and $g \in \mathcal{H}_j$ predict different labels on an example x_j drawn randomly from X_j, then

$$d(f, g) = Pr_{x_j \in X_j}\Big(sign\big(f(x_j)\big) \neq sign\big(g(x_j)\big)\Big).$$

Suppose that for any $f, g \in \mathcal{H}_j$, there exists some constant $L_1 > 0$ to hold that $|f(x_j) - g(x_j)| \leq L_1 \cdot d(f, g) \cdot \|x_j\|_2$, where $\|x_j\|_2$ denotes the 2-norm of x_j. Without loss of generality, suppose that there exists some constant $L_2 > 0$ to hold that $\|x_j\|_2 \leq L_2$ for $x_j \in X_j$ $(j = 1, 2)$. Now we have the following theorem for Paradigm 2 which combines multi-view active learning with semi-supervised learning.

Theorem 3 *For data distribution $\mathbb{D}$ α-expanding with respect to hypothesis class $\mathcal{H}_1 \times \mathcal{H}_2$, let ϵ and δ denote the final desired accuracy and confidence parameters. If $s = \lceil \frac{\log \frac{\alpha}{8\epsilon}}{\log \frac{1}{C}} \rceil$, $m_0 = \frac{1}{L}(4V\log(\frac{1}{L}) + 2\log(\frac{8(s+1)}{\delta}))$ and $m_i = \frac{16}{\alpha}(4V\log(\frac{16}{\alpha}) + 2\log(\frac{8(s+1)}{\delta}))$ $(i = 1, 2, \cdots,)$, Paradigm 2 will generate a classifier with error rate no more than ϵ with probability $1 - \delta$.*

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

Proof. In Paradigm 2, we also use Eq. 1 to combine the two classifiers. With $m_0 = \frac{1}{L}(4V\log(\frac{1}{L}) + 2\log(\frac{8(s+1)}{\delta}))$ where constant $L = \min[\frac{\alpha}{16}, \frac{1}{16L_1L_2}]$, using Lemma 1 we have $Pr(\overline{\mathbf{S}_1^0}) \leq \frac{1}{L}$ and $Pr(\overline{\mathbf{S}_2^0}) \leq \frac{1}{L}$ with probability $1 - \frac{\delta}{4(s+1)}$. Generally, we have that an

Input:

Unlabeled data set $\mathcal{U} = \{x^1, x^2, \cdots, \}$, where each example x^t is given as a pair (x_1^t, x_2^t)

Threshold thr

Process:

Ask the user to label m_0 unlabeled examples drawn randomly from $\mathbb{D}$ to compose the labeled data set $\mathcal{L}$

Iterate $i = 0, 1, \cdots, s$

Set counter n_1^{i+1} to 0. If $\mathbb{D}$ is expanding, set counter n_2^{i+1} to $+\infty$; Otherwise, set counter n_2^{i+1} to 0;

Train two classifiers h_1^i and h_2^i consistent with $\mathcal{L}$ in each view, respectively;

Apply h_1^i and h_2^i to the unlabeled data set $\mathcal{U}$ and find out the contention points set $\mathcal{Q}_i$;

for $k = 1, \cdots, m_{i+1}$

Draw an example $x^k = (x_1^k, x_2^k)$ randomly from $\mathcal{Q}_i$;

if $|h_1^i(x_1^k)| > thr$ **then** $y^k = sign(h_1^i(x_1^k))$;

else if $|h_2^i(x_2^k)| > thr$ **then** $y^k = sign(h_2^i(x_2^k))$;

else ask the user to label x^k and $n_1^{i+1} = n_1^{i+1} + 1$;

Add (x^k, y^k) into $\mathcal{L}$ and delete it from $\mathcal{U}$ and $\mathcal{Q}_i$.

end for

for $w = 1, 2, \cdots$

if $n_2^{i+1} \geq m_{i+1} - n_1^{i+1}$ **break**;

Draw an example $x^w = (x_1^w, x_2^w)$ randomly from $\mathcal{U} - \mathcal{Q}_i$;

if $|h_1^i(x_1^w)| > thr$ **then** $y^w = sign(h_1^i(x_1^w))$;

else if $|h_2^i(x_2^w)| > thr$ **then** $y^w = sign(h_2^i(x_2^w))$;

else ask the user to label x^w and $n_2^{i+1} = n_2^{i+1} + 1$;

Add (x^w, y^w) into $\mathcal{L}$ and delete it from $\mathcal{U}$.

end for

Output:

$h_{final} = \text{combine}(h_1^s, h_2^s)$

Table 2. Paradigm 2: Combination of multi-view active learning and semi-supervised learning

arbitrary S_j^i $(j = 1, 2)$ being consistent with the examples in $\mathcal{L}$ has an error rate at most $\frac{1}{L}$ with probability $1 - \frac{\delta}{4(s+1)}$. So, for any example $x = (x_1, x_2)$,

$$|h_j^i(x_j) - c_j(x_j)| \leq L_1 \cdot L_2 \cdot d(h_j^i, c_j) \leq \frac{1}{16} .$$

We can set the threshold thr in Paradigm 2 to $\frac{1}{16}$. If $|h_j^i(x_j)| > \frac{1}{16}$, h_j^i and c_j make the same prediction on x_j. When $s = \lceil \frac{\log \frac{\alpha}{8\epsilon}}{\log \frac{1}{C}} \rceil$, from the proof of Theorem 1 we have $Pr(\overline{\mathbf{S}_1^s \wedge \mathbf{S}_2^s}) \leq \epsilon$. Thus we get a classifier h_{com}^s whose error rate is no more than ϵ with probability $1 - \delta$ using Paradigm 2. □

The sample complexity of Paradigm 2 is $m_0 + \sum_{i=1}^{s} n_1^i$, which is much smaller than that of Paradigm 1. From Theorem 3 we know that the sample complexity can be further reduced by combining multi-view active learning with semi-supervised learning, however, it needs a stronger assumption on the hypothesis class $\mathcal{H}_1 \times \mathcal{H}_2$. If this assumption holds, in contrast to Paradigm 1, when α-*expansion* does not hold, we can query $\sum_{i=1}^{s}(m_i - n_1^i)$ more examples on which both classifiers have small margin, which can help to reduce the size of the region $\overline{S_1} \wedge \overline{S_2}$.

5. Empirical Study

In this section we empirically study the performance of the Paradigms 1 and 2 on a real-world data set, i.e., the *course* data set (Blum & Mitchell, 1998). This data set has two views (*pages* view and *links* view) and contains 1,051 examples each corresponds to a web page, and the task is to predict whether an unseen web page is a course page or not. There are 230 positive examples (roughly 22%). We randomly use 25% data as the test set and use the remaining 75% data as the unlabeled set $\mathcal{U}$ in Tables 1 and 2. Then, we randomly draw 10 positive and 30 negative examples from $\mathcal{U}$ to generate the initial m_0 labeled examples.

In practice, the thr in Paradigm 2 can be determined by cross validation on labeled examples. Here in our experiments, for the ease of comparison, we do not set thr and instead, we fix the number of examples to be queried in both Paradigms. Thus, we can study their performance under the same number of queries. In detail, in the i-th round, Paradigm 1 picks out two contention points randomly to query; while Paradigm

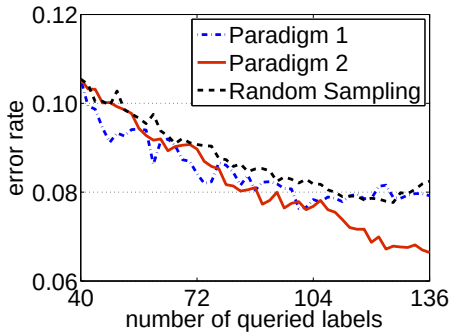

Figure 1. Comparison of the performances

2 picks out the example with the smallest absolute sum of the two classifiers' outputs from $\mathcal{Q}_i$ and $\mathcal{U}-\mathcal{Q}_i$ respectively to query, and picks out the example with the largest absolute sum of the two classifiers' outputs from $\mathcal{Q}_i$ and $\mathcal{U}-\mathcal{Q}_i$ respectively to label as $sign\big(h_1^i(x_1)+h_2^i(x_2)\big)$. That is, the two examples to be queried in Paradigm 2 are $\arg\min_{x\in\mathcal{Q}_i}\big(|h_1^i(x_1)+h_2^i(x_2)|\big)$ and $\arg\min_{x\in\mathcal{U}-\mathcal{Q}_i}\big(|h_1^i(x_1)+h_2^i(x_2)|\big)$, while the two examples Paradigm 2 labels for itself by semi-supervised learning are $\arg\max_{x\in\mathcal{Q}_i}\big(|h_1^i(x_1)+h_2^i(x_2)|\big)$ and $\arg\max_{x\in\mathcal{U}-\mathcal{Q}_i}\big(|h_1^i(x_1)+h_2^i(x_2)|\big)$. We use Random Sampling as the baseline and implement the classifiers with SMO in WEKA (Witten & Frank, 2005). The experiments are repeated for 20 runs and Figure 1 plots the average error rates of the three methods against the number of examples that have been queried.

It can be found from Figure 1 that with the same number of queried examples, although there are some fluctuation, the performance of Paradigm 1 is generally better than that of Random Sampling, while the performance of Paradigm 2 is better than that of the others. In particular, the advantage of Paradigm 2 becomes more prominent as the number of queries increases. This is not difficult to understand since with more labeled data the learners become stronger and thus the labels obtained from the semi-supervised learning process become more helpful.

Overall, the empirical study verifies that comparing with pure active learning, the combination of multi-view active learning and semi-supervised learning can reduce the sample complexity.

6. Conclusion

In this paper, we first characterize the sample complexity of multi-view active learning and get an exponential improvement in the sample complexity from $\widetilde{O}(\frac{1}{\epsilon})$ to $\widetilde{O}(\log\frac{1}{\epsilon})$. The α-*expansion* assumption we employed is weaker than assumptions taken by previous theoretical studies on active learning, such as that the data is distributed uniformly over the unit sphere in $\mathbb{R}^d$ and that the hypothesis class is linear separators through the origin. We also give an upper bound of the error rate for cases where the α-*expansion* assumption does not hold. Then, we analyze the combination of multi-view active learning with semi-supervised learning and get that such a combination can reduce the sample complexity further, which is verified by an empirical study. This provides an explanation to that why the method described in (Zhou et al., 2006) can lead to good results.

Our work is the first theoretical analysis on the sample complexity of realizable multi-view active learning. Recently, *non-realizable* active learning, where there does not exist a hypothesis perfectly separating the data, starts to attract attention (Balcan et al., 2006; Balcan et al., 2007; Dasgupta et al., 2008). Extending our work to non-realizable multi-view active learning is a future work.

Acknowledgments

This research was supported by the National Science Foundation of China (60635030, 60721002), the Foundation for the Author of National Excellent Doctoral Dissertation of China (200343) and the National High Technology Research and Development Program of China (2007AA01Z169).

References

Anthony, M., & Bartlett, P. L. (Eds.). (1999). *Neural network learning: Theoretical foundations*. Cambridge, UK: Cambridge University Press.

Balcan, M.-F., Beygelzimer, A., & Langford, J. (2006). Agnostic active learning. *Proceedings of the 23rd International Conference on Machine Learning* (pp. 65–72). Pittsburgh, PA.

Balcan, M.-F., Blum, A., & Yang, K. (2005). Co-training and expansion: Towards bridging theory and practice. In L. K. Saul, Y. Weiss and L. Bottou (Eds.), *Advances in neural information processing systems 17*, 89–96. Cambridge, MA: MIT Press.

Balcan, M.-F., Broder, A. Z., & Zhang, T. (2007). Margin based active learning. *Proceedings of the 20th Annual Conference on Learning Theory* (pp. 35–50). San Diego, CA.

Belkin, M., Matveeva, I., & Niyogi, P. (2001). Reg-

ularization and semi-supervised learning on large graphs. *Proceedings of the 17th Annual Conference on Learning Theory* (pp. 624–638). Banff, Canada.

Blum, A., & Mitchell, T. (1998). Combining labeled and unlabeled data with co-training. *Proceedings of the 11th Annual Conference on Computational Learning Theory* (pp. 92–100). Madison, WI.

Cohn, D. A., Atlas, L. E., & Ladner, R. E. (1994). Improving generalization with active learning. *Machine Learning*, *15*, 201–221.

Dasgupta, S. (2006). Coarse sample complexity bounds for active learning. In Y. Weiss, B. Schölkopf and J. Platt (Eds.), *Advances in neural information processing systems 18*, 235–242. Cambridge, MA: MIT Press.

Dasgupta, S., Hsu, D., & Monteleoni, C. (2008). A general agnostic active learning algorithm. In J. Platt, D. Koller, Y. Singer and S. Roweis (Eds.), *Advances in neural information processing systems 20*, 353–360. Cambridge, MA: MIT Press.

Dasgupta, S., Kalai, A. T., & Monteleoni, C. (2005). Analysis of perceptron-based active learning. *Proceedings of the 18th Annual Conference on Learning Theory* (pp. 249–263). Bertinoro, Italy.

de Sa, V. R., & Ballard, D. H. (1998). Category learning through multi-modality sensing. *Neural Computation*, *10*, 1097–1117.

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

Freund, Y., Seung, H. S., Shamir, E., & Tishby, N. (1997). Selective sampling using the query by committee algorithm. *Machine Learning*, *28*, 133–168.

Goldman, S., & Zhou, Y. (2000). Enhancing supervised learning with unlabeled data. *Proceedings of the 17th International Conference on Machine Learning* (pp. 327–334). San Francisco, CA.

Melville, P., & Mooney, R. J. (2004). Diverse ensembles for active learning. *Proceedings of the 21st International Conference on Machine Learning* (pp. 584–591). Banff, Canada.

Miller, D. J., & Uyar, H. S. (1997). A mixture of experts classifier with learning based on both labelled and unlabelled data. In M. Mozer, M. I. Jordan and T. Petsche (Eds.), *Advances in neural information processing systems 9*, 571–577. Cambridge, MA: MIT Press.

Muslea, I., Minton, S., & Knoblock, C. A. (2000). Selective sampling with redundant views. *Proceedings of the 17th National Conference on Artificial Intelligence* (pp. 621–626). Austin, TX.

Muslea, I., Minton, S., & Knoblock, C. A. (2006). Active learning with multiple views. *Journal of Artificial Intelligence Research*, *27*, 203–233.

Nigam, K., McCallum, A. K., Thrun, S., & Mitchell, T. (2000). Text classification from labeled and unlabeled documents using EM. *Machine Learning*, *39*, 103–134.

Tong, S., & Koller, D. (2001). Support vector machine active learning with applications to text classification. *Journal of Machine Learning Research*, *2*, 45–66.

Witten, I. H., & Frank, E. (2005). *Data Mining: Practical Machine Learning Tools and Techniques*. San Francisco, CA: Morgan Kaufmann. 2nd edition.

Zhou, D., Schölkopf, B., & Hofmann, T. (2005). Semi-supervised learning on directed graphs. In L. K. Saul, Y. Weiss and L. Bottou (Eds.), *Advances in neural information processing systems 17*, 1633–1640. Cambridge, MA: MIT Press.

Zhou, Z.-H., Chen, K.-J., & Dai, H.-B. (2006). Enhancing relevance feedback in image retrieval using unlabeled data. *ACM Transactions on Information Systems*, *24*, 219–244.

Zhou, Z.-H., & Li, M. (2005). Semi-supervised learning with co-training. *Proceedings of the 19th International Joint Conference on Artificial Intelligence* (pp. 908–913). Edinburgh, Scotland.

Zhu, X., Ghahramani, Z., & Lafferty, J. (2003). Semi-supervised learning using Gaussian fields and harmonic functions. *Proceedings of the 20th International Conference on Machine Learning* (pp. 912–919). Washington, DC.