

# Semisupervised learning from different information sources

Tao Li, Mitsunori Ogihara

Department of Computer Science, University of Rochester, Rochester, NY, USA

**Abstract.** This paper studies the use of a semisupervised learning algorithm from different information sources. We first offer a theoretical explanation as to why minimising the disagreement between individual models could lead to the performance improvement. Based on the observation, this paper proposes a semisupervised learning approach that attempts to minimise this disagreement by employing a co-updating method and making use of both labeled and unlabeled data. Three experiments to test the effectiveness of the approach are presented in this paper: (i) webpage classification from both content and hyperlinks; (ii) functional classification of gene using gene expression data and phylogenetic data and (iii) machine self-maintaining from both sensory and image data. The results show the effectiveness and efficiency of our approach and suggest its application potentials.

**Keywords:** Decision tree; Minimise disagreement; Semisupervised; Support vector machines; Unlabelled data

---

## 1. Introduction

Many real-world problems can be cast as problems of learning from different information sources. One such example is self-maintenance of xerographic machines based on both sensory data of the faulty machines and on image samples of the hard copies. Another one is gene function discovery based on microarray expression data and phylogenetic profiles.

The problem of learning from multiple information sources has been extensively studied in machine learning and in computer vision, where the problem is referred to as multimodal learning.<sup>1</sup> Generally, there are two types of multimodal learning: Feature-level integration and semantic integration (Wu et al. 1999). The feature in-

---

*Received 18 August 2003*  
*Revised 15 November 2003*  
*Accepted 19 December 2003*  
*Published online 21 May 2004*

<sup>1</sup> In this paper, we interchangeably use modal, component and information source.

tegration combines the information at the feature level and performs learning in the joint feature space. The correlation structure between different sources can be discovered via learning. The semantic integration, on the other hand, first builds individual models based on separate information sources and then combines these models via some processes, say, mutual information maximization (Becker 1996).

In many applications, the ability to use both labelled and unlabelled data is very useful because labelling samples is usually expensive and, in most cases, only a limited number of labelled samples and a lot of unlabelled samples are available. Having data from different information sources could help us take advantage of both labelled and unlabelled data. In our recent work, we have provided a theoretical explanation as to why minimising the disagreement between two individual models could lead to the improvement of the classification accuracy of individual models. In light of the observation, we proposed a co-updating method (as an improvement of the algorithm presented in one of our papers) for minimising the disagreement between the individual models taking both labelled and unlabelled data into consideration. To demonstrate the effectiveness and efficiency of our approach, we conducted in this paper three sets of experiments: (i) webpage classification from both content and hyperlinks, (ii) functional classification of genes using gene expression data and phylogenetic data and (iii) machine self-maintaining from both sensory and image data.

The co-updating approach can be thought of as a kind of semantic integration. We prefer semantic integration over feature-level integration for the following four reasons: First, although the structure in the joint feature space is often more informative than that available to each of the individual sources, feature integration tends to generalize poorly. The model complexity, computation intensity and training difficulty typically are other problems associated with the feature-integration approach (Wu et al. 1999). Second, learning in the joint space is not able to marginalize over the missing sources and requires future patterns for classification containing all feature dimensions (De Sa and Ballard 1998). While we intend to learn from a joint feature space, we still need to be able to analyse and act on the information from a single source. For example, we sometimes would like to predict the copy/printer failures solely based on the image information or predict the functional classifications solely based on the DNA microarray information. Feature integration is thus not suitable. Third, we would like to utilize both labelled and unlabelled data because there are scenarios in which some gene category information is available while the other is unavailable. Finally, the semantic integration appears to have biological and physical plausibility. It is well known that the cerebral cortex competently classifies unimodal stimuli while keeping the different modalities largely separate. Also, McGurk showed that, although the information from different sensory modalities is combined in determining human's perception, the combination is often not subject to conscious control (McGurk and MacDonald 1976).

A preliminary version of this method was published in Bio-informatics track in 2003 ACM Symposium on Applied Computing (SAC2003) (Li et al. 2003), where the focus is gene functional discovery from heterogeneous data types. We provide detailed theoretical analysis and present extensive experimental results with three case studies in this paper. The paper is organized as follows. Section 2 shows the theoretical reason as to why minimisation of disagreement is useful and Sect. 3 presents the co-updating algorithm. Section 4 reviews related work on using labelled and unlabelled data and discusses the connections and differences between them. Section 5 presents the results of the experiments. Finally, Sect. 6 concludes and discusses future research directions.

## 2. Minimising disagreement

### 2.1. Theoretical results

In this section, we show that, theoretically, minimising the disagreement between two individual models could lead to the improvement of the classification accuracy of individual models. In this paper, we focus on binary classification problems and we use 0, 1 to label the two classes respectively<sup>2</sup>. Suppose we have an instance space  $X = (X_1, X_2)$ , where  $X_1$  and  $X_2$  are from different observations. Let  $D$  be the distribution over  $X$ . If  $f$  is the target function over  $D$ , then for any example  $x = (x_1, x_2)$ , we would have  $f(x_1, x_2) = f_1(x_1) = f_2(x_2)$ , where  $f_1$  and  $f_2$  are the target functions over  $X_1$  and  $X_2$ , respectively. We also use  $Y$  to denote the target label.

**Definition 1.** We say that  $f$  is a nontrivial classifier if  $\Pr(f = u|Y = u) > \Pr(f = \bar{u}|Y = u)$ , where  $u \in \{0, 1\}$  and  $\bar{u}$  is the complement of  $u$ .  $\square$

*Remark 1.* The nontrivial conditions can be restated as  $\Pr(f = u|Y = u) > 1/2$  or  $\Pr(f \neq Y) \leq \Pr(f = u)$ ,  $u \in \{0, 1\}$ .  $\square$

In Blum and Mitchell (1998), it is assumed that  $x_1$  and  $x_2$  are conditionally independent given the labels, i.e.

$$\Pr(x_1 = x'_1|x_2 = x'_2) = \Pr(x_1 = x'_1|f_2(x_2) = f_2(x'_2)).$$

The independence assumption is rather strong but it is used by many successful applications. Suppose we build hypotheses  $f'_1$  on  $X_1$  and  $f'_2$  on  $X_2$ . Thus, if  $x_1$  and  $x_2$  are conditional independent given the labels, then  $f'_1$  and  $f'_2$  are also conditional independent. The conditional independence of  $f'_1$  and  $f'_2$  can be interpreted as follows:

$$\Pr(f'_1(x_1) = u|f'_2(x_2) = v, Y = y) = \Pr(f'_1(x_1) = u|Y = y), \quad (1)$$

where  $u, v, y \in \{0, 1\}$ . The following theorem holds.

**Theorem 1.** Under the conditional independence assumption, the disagreement upper bounds the misclassification error for the nontrivial classifier.

*Proof.* The misclassification error of  $f'_1$  is

$$\begin{aligned} \Pr(f'_1 \neq Y) &= \Pr(f'_1 = 0, Y = 1) + \Pr(f'_1 = 1, Y = 0) \\ &= \Pr(f'_1 = 0, Y = 1, f'_2 = 0) + \Pr(f'_1 = 0, Y = 1, f'_2 = 1) \\ &\quad + \Pr(f'_1 = 1, Y = 0, f'_2 = 0) + \Pr(f'_1 = 1, Y = 0, f'_2 = 1). \end{aligned}$$

The disagreement between two models is

$$\begin{aligned} \Pr(f'_1 \neq f'_2) &= \Pr(f'_1 = 0, f'_2 = 1) + \Pr(f'_1 = 1, f'_2 = 0) \\ &= \Pr(f'_1 = 0, f'_2 = 1, Y = 1) + \Pr(f'_1 = 0, f'_2 = 1, Y = 0) \\ &\quad + \Pr(f'_1 = 1, f'_2 = 0, Y = 0) + \Pr(f'_1 = 1, f'_2 = 0, Y = 1). \end{aligned}$$

---

<sup>2</sup> For a multiclass classification problem, several approaches can be used to reduce it to binary ones (Allwein et al. 2000).

To show that  $\Pr(f'_1 \neq Y) \leq \Pr(f'_1 \neq f'_2)$ , it is enough to verify that

$$\begin{aligned} & \Pr(f'_1 = 0, Y = 1, f'_2 = 0) + \Pr(f'_1 = 1, Y = 0, f'_2 = 1) \\ & \leq \Pr(f'_1 = 0, f'_2 = 1, Y = 0) + \Pr(f'_1 = 1, f'_2 = 0, Y = 1). \end{aligned}$$

Note that

$$\begin{aligned} & \Pr(f'_1 = 0, Y = 1, f'_2 = 0) + \Pr(f'_1 = 1, Y = 0, f'_2 = 1) \\ & = \Pr(f'_2 = 0, Y = 1) \Pr(f'_1 = 0 | f'_2 = 0, Y = 1) \\ & \quad + \Pr(f'_2 = 1, Y = 0) \Pr(f'_1 = 1 | f'_2 = 1, Y = 0) \\ & = \Pr(f'_2 = 0, Y = 1) \Pr(f'_1 = 0 | Y = 1) \\ & \quad + \Pr(f'_2 = 1, Y = 0) \Pr(f'_1 = 1 | Y = 0) \text{ by Eq. (1)}. \end{aligned}$$

$$\begin{aligned} & \Pr(f'_1 = 0, f'_2 = 1, Y = 0) + \Pr(f'_1 = 1, f'_2 = 0, Y = 1) \\ & = \Pr(f'_2 = 1, Y = 0) \Pr(f'_1 = 0 | f'_2 = 1, Y = 0) \\ & \quad + \Pr(f'_2 = 0, Y = 1) \Pr(f'_1 = 1 | f'_2 = 0, Y = 1) \\ & = \Pr(f'_2 = 1, Y = 0) \Pr(f'_1 = 0 | Y = 0) \\ & \quad + \Pr(f'_2 = 0, Y = 1) \Pr(f'_1 = 1 | Y = 1) \text{ by Eq. (1)}. \end{aligned}$$

Hence, it reduces to check

$$\begin{aligned} & \Pr(f'_1 = 0 | Y = 1) \leq \Pr(f'_1 = 1 | Y = 1) \\ & \Pr(f'_1 = 1 | Y = 0) \leq \Pr(f'_1 = 0 | Y = 0). \end{aligned}$$

This amounts to requiring that

$$\Pr(f'_1 = 0, Y = 1) \leq \Pr(f'_1 = 1, Y = 1) \quad (2)$$

$$\Pr(f'_1 = 1, Y = 0) \leq \Pr(f'_1 = 0, Y = 0). \quad (3)$$

Observe that

$$\begin{aligned} & \Pr(f'_1 = 0, Y = 1) \leq \Pr(f'_1 = 1, Y = 1) \\ & \Leftrightarrow \Pr(f'_1 = 0, Y = 1) + \Pr(f'_1 = 1, Y = 0) \\ & \quad \leq \Pr(f'_1 = 1, Y = 1) + \Pr(f'_1 = 1, Y = 0) \\ & \Leftrightarrow \Pr(f'_1 \neq Y) \leq \Pr(f'_1 = 1). \end{aligned}$$

Similarly,

$$\begin{aligned} & \Pr(f'_1 = 1, Y = 0) \leq \Pr(f'_1 = 0, Y = 0) \\ & \Leftrightarrow \Pr(f'_1 = 1, Y = 0) + \Pr(f'_1 = 0, Y = 1) \\ & \quad \leq \Pr(f'_1 = 0, Y = 0) + \Pr(f'_1 = 0, Y = 1) \\ & \Leftrightarrow \Pr(f'_1 \neq Y) \leq \Pr(f'_1 = 0). \end{aligned}$$

In other words, if  $\Pr(f'_1 \neq Y) \leq \Pr(f'_1 = u)$ ,  $u \in \{0, 1\}$ , then the disagreement upper bounds the misclassification error:

$$\Pr(f'_1 \neq Y) \leq \Pr(f'_1 \neq f'_2). \quad (4)$$

□

*Remark 2.* When the conditional independence condition (e.g., Eq. (1)) doesn't hold, to guarantee that disagreement upper bounds the misclassification error, we need

$$\begin{aligned}\Pr(f'_1 = 0 | f'_2 = 0, Y = 1) &\leq \Pr(f'_1 = 1 | f'_2 = 0, Y = 1) \\ \Pr(f'_1 = 1 | f'_2 = 1, Y = 0) &\leq \Pr(f'_1 = 0 | f'_2 = 1, Y = 0).\end{aligned}$$

In other words, if

$$\Pr(f'_1 \neq Y | f'_2 \neq Y) \leq \Pr(f'_1 = Y | f'_2 \neq Y),$$

then the disagreement still upper bounds the misclassification error without the conditional independence condition.  $\square$

In essence, this shows that, under certain conditions, the disagreement upper bounds the misclassification error. Thus, minimising the disagreement would decrease the upper bound on the misclassification error and could bootstrap the learning algorithm.

## 2.2. Bayes perspective

Theorem 1 can also be derived from Bayes perspective. Let  $x = (x_1, x_2)$  be an observation vector, the Bayes decision rule for the first modal is

$$\Pr(Y = 1 | x_1) \stackrel{\leq_0}{\leq} \Pr(Y = 0 | x_1),$$

which indicates that, if the *posteriori* probability of class 1 given  $x_1$  is larger than the probability of class 0,  $x_1$  is assigned to class 1 and vice versa. Using Bayes theorem and eliminating the common term  $\Pr(x_1)$ , we get

$$\Pr(Y = 1) \Pr(x_1 | Y = 1) \stackrel{\leq_0}{\leq} \Pr(Y = 0) \Pr(x_1 | Y = 0).$$

The Bayes error can be computed as<sup>3</sup>

$$\begin{aligned}\epsilon &= \int \min\{\Pr(Y = 1) \Pr(x_1 | 1), \Pr(Y = 0) \Pr(x_1 | 0)\} dx_1 \\ &= \Pr(Y = 1) \int_{L_2^1} \Pr(x_1 | 1) dx_1 + \Pr(Y = 0) \int_{L_1^1} \Pr(x_1 | 0) dx_1,\end{aligned}$$

where  $L_1^1$  is the region where

$$\Pr(Y = 1) \Pr(x_1 | Y = 1) > \Pr(Y = 0) \Pr(x_1 | Y = 0)$$

and  $L_2^1$  is the region where

$$\Pr(Y = 1) \Pr(x_1 | Y = 1) < \Pr(Y = 0) \Pr(x_1 | Y = 0).$$

In other words, if an observation  $x_1 \in L_1^1$ , it will be classified as in class 1 and if  $x_1 \in L_2^1$ , it will be classified as in class 0.

<sup>3</sup> We use  $\Pr(x_i | j)$  to denote  $\Pr(x_i | Y = j)$ , where  $i = 1, 2$  and  $j = 0, 1$ .

Under the conditional independence assumption, the disagreement between two components can be computed as

$$\begin{aligned}
 E(x_1, x_2) &= \Pr\{\Pr(Y = 1|x_1) > \Pr(Y = 0|x_1) \& \Pr(Y = 1|x_2) < \Pr(Y = 0|x_2)\} \\
 &\quad + \Pr\{\Pr(Y = 1|x_1) < \Pr(Y = 0|x_1) \& \Pr(Y = 1|x_2) > \Pr(Y = 0|x_2)\} \\
 &= \int_{L_1^1} \int_{L_2^2} (\Pr(Y = 1) \Pr(x_1|Y = 1) \Pr(x_2|Y = 1)) dx_1 dx_2 \\
 &\quad + \int_{L_1^1} \int_{L_2^2} (\Pr(Y = 0) \Pr(x_1|Y = 0) \Pr(x_2|Y = 0)) dx_1 dx_2 \\
 &\quad + \int_{L_2^1} \int_{L_1^2} (\Pr(Y = 1) \Pr(x_1|Y = 1) \Pr(x_2|Y = 1)) dx_1 dx_2 \\
 &\quad + \int_{L_2^1} \int_{L_1^2} (\Pr(Y = 0) \Pr(x_1|Y = 0) \Pr(x_2|Y = 0)) dx_1 dx_2,
 \end{aligned}$$

where  $L_1^2$  is the region where

$$\Pr(Y = 1) \Pr(x_2|Y = 1) > \Pr(Y = 0) \Pr(x_2|Y = 0)$$

and  $L_2^2$  is the region where

$$\Pr(Y = 1) \Pr(x_2|Y = 1) < \Pr(Y = 0) \Pr(x_2|Y = 0).$$

Similarly, if an observation  $x_2 \in L_1^2$ , it will be classified as in class 1 and if  $x_2 \in L_2^2$ , it will be classified as in class 0.

Observe that

$$\begin{aligned}
 \epsilon &= \Pr(Y = 1) \int_{L_2^1} \Pr(x_1|Y = 1) dx_1 \\
 &\quad + \Pr(Y = 0) \int_{L_1^1} \Pr(x_1|Y = 0) dx_1 \\
 &= \Pr(Y = 1) \int_{L_2^1} \Pr(x_1|Y = 1) \left( \int \Pr(x_2|Y = 1) dx_2 \right) dx_1 \\
 &\quad + \Pr(Y = 0) \int_{L_1^1} \Pr(x_1|Y = 0) \left( \int \Pr(x_2|Y = 1) dx_2 \right) dx_1 \\
 &= \int_{L_2^1} \int_{L_2^2} \Pr(Y = 1) \Pr(x_1|Y = 1) \Pr(x_2|Y = 1) dx_1 dx_2 \\
 &\quad + \int_{L_2^1} \int_{L_1^2} \Pr(Y = 1) \Pr(x_1|Y = 1) \Pr(x_2|Y = 1) dx_1 dx_2 \\
 &\quad + \int_{L_1^1} \int_{L_2^2} \Pr(Y = 0) \Pr(x_1|Y = 0) \Pr(x_2|Y = 0) dx_1 dx_2 \\
 &\quad + \int_{L_1^1} \int_{L_1^2} \Pr(Y = 0) \Pr(x_1|Y = 0) \Pr(x_2|Y = 0) dx_1 dx_2.
 \end{aligned}$$

So to ensure that  $\epsilon \leq E(x_1, x_2)$ , it is sufficient that

$$\begin{aligned} & \int_{L_2^1} \int_{L_2^2} \Pr(Y = 1) \Pr(x_1|Y = 1) \Pr(x_2|Y = 1) dx_1 dx_2 \\ & < \int_{L_1^1} \int_{L_2^2} \Pr(Y = 1) \Pr(x_1|Y = 1) \Pr(x_2|Y = 1) dx_1 dx_2 \end{aligned}$$

and

$$\begin{aligned} & \int_{L_1^1} \int_{L_1^2} \Pr(Y = 0) \Pr(x_1|Y = 0) \Pr(x_2|Y = 0) dx_1 dx_2 \\ & < \int_{L_2^1} \int_{L_1^2} \Pr(Y = 0) \Pr(x_1|Y = 0) \Pr(x_2|Y = 0) dx_1 dx_2. \end{aligned}$$

The above formula can be reduced to

$$\Pr(x_1 \in L_2^1 | Y = 1) < \Pr(x_1 \in L_1^1 | Y = 1) \quad (5)$$

$$\Pr(x_1 \in L_1^1 | Y = 0) < \Pr(x_1 \in L_2^1 | Y = 0). \quad (6)$$

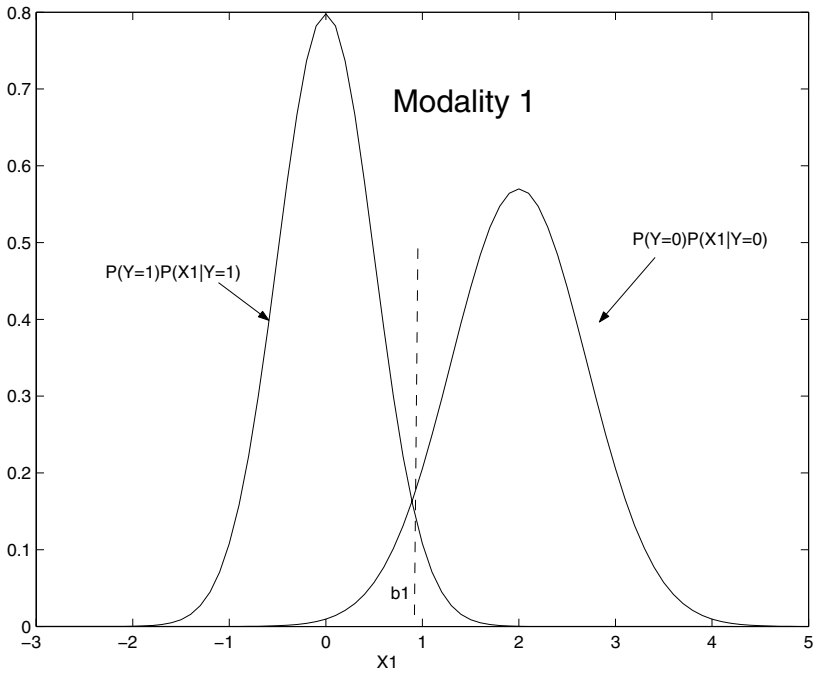
The formula in Eqs. (5) and (6) are essentially the same as those in Eqs. (2) and (3). Hence, this quantity of disagreement is an upper bound on the Bayes error.

*Remark 3.* Figure 1 gives an explicit example to illustrate the mathematical derivations above. The example is a one-dimensional classification problem with two different modalities. In this example,  $L_1^1 = (-\infty, b_1)$ ,  $L_2^1 = (b_1, \infty)$ ,  $L_1^2 = (-\infty, b_2)$ ,  $L_2^2 = (b_2, \infty)$ , the Bayes error  $\epsilon$  and the disagreement  $E(x_1, x_2)$  are calculated as follows:

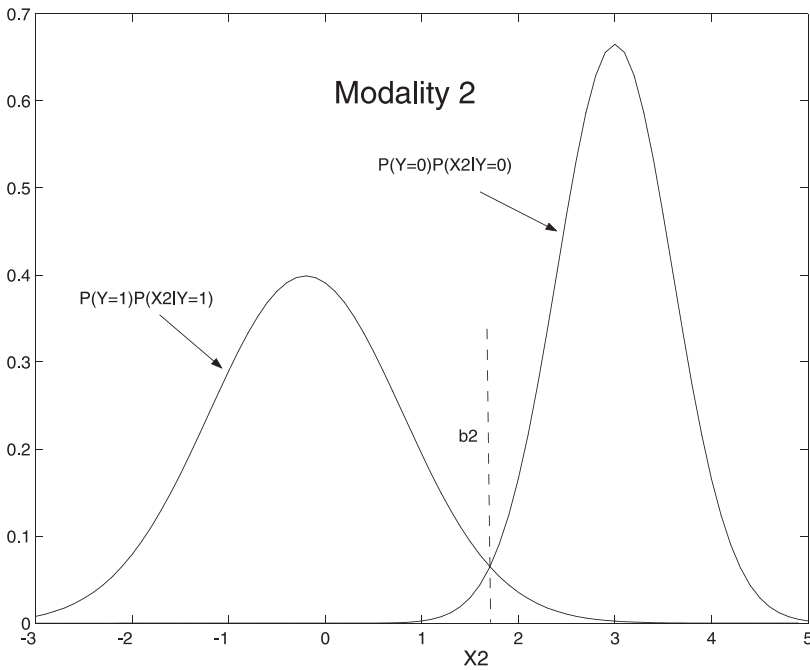
$$\begin{aligned} \epsilon &= \int_{b_1}^{\infty} \int_{b_2}^{\infty} \Pr(Y = 1) \Pr(x_1|Y = 1) \Pr(x_2|Y = 1) dx_1 dx_2 \\ &+ \int_{b_1}^{\infty} \int_{-\infty}^{b_2} \Pr(Y = 1) \Pr(x_1|Y = 1) \Pr(x_2|Y = 1) dx_1 dx_2 \\ &+ \int_{-\infty}^{b_1} \int_{b_2}^{\infty} \Pr(Y = 0) \Pr(x_1|Y = 0) \Pr(x_2|Y = 0) dx_1 dx_2 \\ &+ \int_{-\infty}^{b_1} \int_{-\infty}^{b_2} \Pr(Y = 0) \Pr(x_1|Y = 0) \Pr(x_2|Y = 0) dx_1 dx_2 \end{aligned}$$

$$\begin{aligned} E(x_1, x_2) &= \int_{-\infty}^{b_1} \int_{b_2}^{\infty} (\Pr(Y = 1) \Pr(x_1|Y = 1) \Pr(x_2|Y = 1) dx_1 dx_2 \\ &+ \int_{-\infty}^{b_1} \int_{b_2}^{\infty} \Pr(Y = 0) \Pr(x_1|Y = 0) \Pr(x_2|Y = 0)) dx_1 dx_2 \\ &= \int_{b_1}^{\infty} \int_{-\infty}^{b_2} (\Pr(Y = 1) \Pr(x_1|Y = 1) \Pr(x_2|Y = 1) dx_1 dx_2 \\ &+ \int_{b_1}^{\infty} \int_{-\infty}^{b_2} \Pr(Y = 0) \Pr(x_1|Y = 0) \Pr(x_2|Y = 0)) dx_1 dx_2. \end{aligned}$$

□



(a) Modality 1



(b) Modality 2

**Fig. 1.** A one-dimensional example with two different modalities.  $b_1$  and  $b_2$  denote the Bayes decision boundaries of the two modalities

### Algorithm co-updating

Input: A collection of labelled and unlabelled data

$\alpha$  — default 0.15

$T$  — default 30

Output: Two classifiers that predict class labels for new instances based on different information sources

- 1: Build classifier  $f_1^0$  using the first component of labelled samples
- 2: Build classifier  $f_2^0$  using the second one
- 3: Loop for  $T$  times:
  - 3.1: **E-step:** Using  $f_1^{i-1}$  get the labels of all the unlabelled samples based on their first component; using  $f_2^{i-1}$  on their second component
  - 3.2: **M-step:** With probability  $\alpha$ , select the unlabelled samples on which the two classifiers have the same predictions. Rebuild the classifiers  $f_1^i$  and  $f_2^i$  using the labelled samples and selected unlabelled samples
- 4: Output  $f_1^T, f_2^T$

Fig. 2. The algorithm description of co-updating

## 3. Algorithm description

Based on Theorem 1, we have developed a co-updating approach to learn from both labelled and unlabelled data that aims at minimising the disagreement on unlabelled data. The co-updating approach is an iterative expectation-maximization (EM)-type procedure. Its basic idea is as follows: The labelled samples are first used to get weak classifiers  $f_1^0$  on  $X_1$  and  $f_2^0$  on  $X_2$ . Then for each iteration, the expectation step uses current classifiers to predict the labels of unlabelled data, the maximization step rebuilds the classifiers using the labelled samples and a random collection of unlabelled samples on which the classifiers agree (i.e. they have the same predictions). This process is then repeated until some termination criterion is met. The detailed description of the algorithm is given in Fig. 2.

The intuition behind the co-updating approach is that we stochastically select the unlabelled samples on which the two component classifiers agree and then use them, along with the labelled samples, to train/update the classifiers. Co-updating iteratively updates classifier models by using current models to infer (a probability distribution on) labels for unlabelled data and then adjusting the models to fit the (distribution on) filled-in labels. When the model defines a joint probability distribution over observable data and unobservable labels, each iteration of the EM algorithm can be shown to increase the probability of the observable data given the model parameters (Dasgupta et al. 2001). Intuitively, the added agreed unlabelled samples help the classifiers to achieve more agreement and hence minimise disagreement. If the initial classifiers are nontrivial and the agreement of the predictions from the two modal is conditional independent, then, on average, the co-updating approach, trying to minimise the disagreement, should progress toward higher performance.

In the above algorithm,  $T$  is the number of iterations and is used as the stopping criterion.  $\alpha$  defines the fraction of agreed unlabelled samples we want to include for updating the model at each iteration.<sup>4</sup> The co-updating approach is a variant of the EM algorithm, which is a general method for parameter estimates in the presence of missing data. In the generalized formulation (Neal and Hinton 1998), a current parameter estimate and a distribution  $Q$  over the missing data is maintained, the

<sup>4</sup> Providing theoretical heuristics and justification for the choice of parameters is one of our future works. Currently, their choice is mainly based on experience.

E step updates the distribution  $Q$  and the M-step modifies the estimate. In the co-updating approach, we use current classifiers to predict the unlabelled data, which amounts to computing the distribution  $Q$  over the missing data. In the M-step, we use the selected unlabelled samples along with labelled samples to retrain the classifiers, which is equivalent to modifying the parameter estimates.

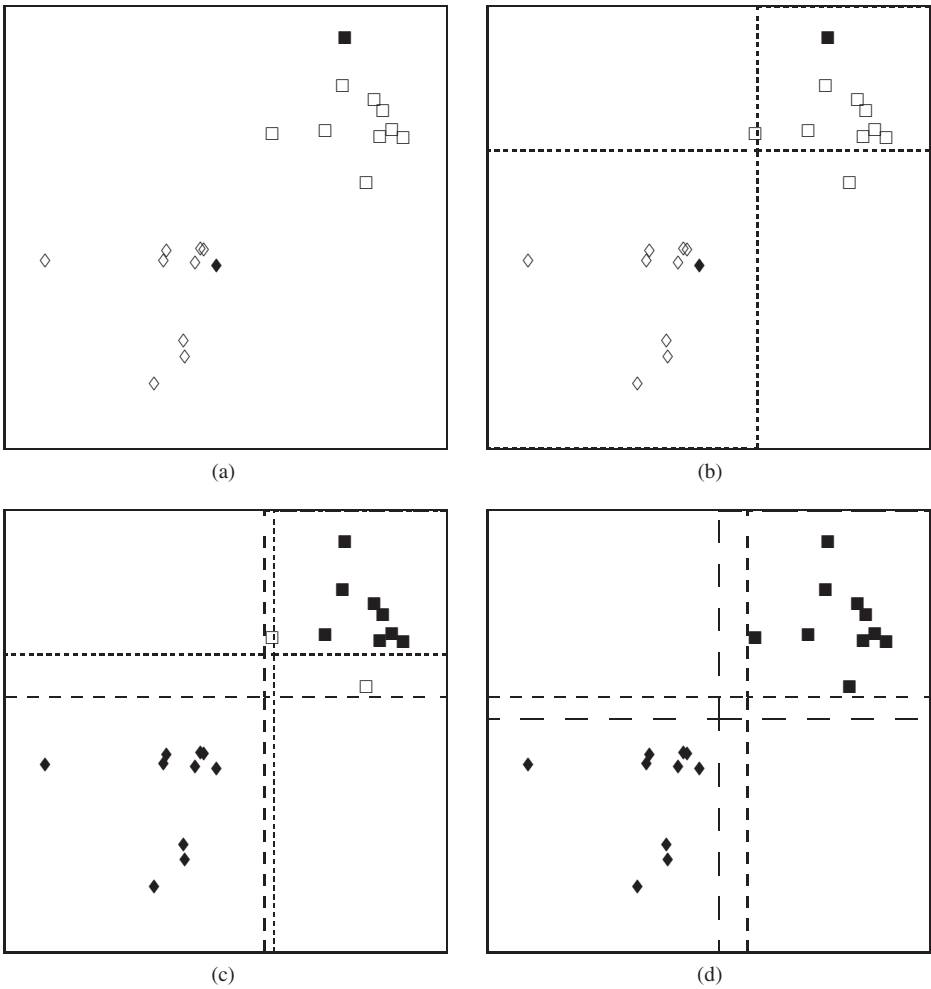
### 3.1. Example illustration

Figure 3 illustrates the co-updating procedure through an example. The simple example helps us understand the procedures of the co-updating approach. One thing to note is that, in the example, all the agreed unlabelled samples along with labelled samples are used to retrain the classifiers. In our algorithm, to avoid the phenomenon that the model may be overwhelmed by the unlabelled samples, especially when the number of labelled samples is small, a parameter  $\alpha$  is used to define the fraction of agreed unlabelled samples we want to include for updating the model at each iteration.

## 4. Literature review

### 4.1. Related work

There has been much recent interest in the problem of learning with both labelled and unlabelled data. From a theoretical perspective, it has been shown that labelled data are exponentially more useful than unlabelled data in risk reduction under certain assumptions (Castelli and Cover 1996). Discriminative and generative learning architectures have been shown to take advantage of unlabelled data in various probabilistic frameworks (Zhang and Oles 2000). Generally, the approaches for handling both labelled and unlabelled data can be roughly divided into four categories: probabilistic, cotraining, transductive interference and others. Probabilistic approaches include using expectation-maximization (Dempster et al. 1977; Ghahramani and Jordan 1994) to maximize the estimates of *posteriori* parameters and using generative models to perform classification (Nigam et al. 2000). Unfortunately, when the data do not match the generative assumptions, the information from unlabelled data may overwhelm that of labelled data and the algorithm goes astray. The ideas of cotraining first appeared on unsupervised learning. Becker (1996) proposed the approach of learning coherence structure in data by maximizing the mutual information between the outputs of two groups of units, which receive inputs physically separated in modality. De Sa and Ballard (1998) described a unsupervised network algorithm learning from co-occurring patterns of lip motion and sound signals from a human speaker to minimise the fraction of training samples on which the two patterns disagree. The cotraining paradigms combining both labelled and unlabelled data were first introduced by Blum and Mitchell (1998), where the features in the problem domain are naturally divided into two disjoint sets (or in other words, is two-modal) to exploit the *compatibility* between different views of the samples. The *compatibility* of the instance distribution means that the target functions over each feature set predict the same label. Blum and Mitchell (1998) showed that, under certain assumptions, PAC-style guarantee on learning with labelled and unlabelled data holds. Roth and Zelenko (2000) developed a theory for learning scenarios where multiple learners coexist but there are mutual compatibility constraints on their outcomes. Nigam and



**Fig. 3.** An example on the co-updating approach. The two different components of the data are represented by the two different dimensions. Initially, only a small portion of the data is labelled, ■ and ◇, in (a). The labelled data is used to train two classifiers separately, denoted by the two lines, in (b). In (c), all the unlabelled data, □ and ◇, are given pseudo-labels by the classifiers and the agreed unlabelled along with the original labelled samples are used to update the classifiers, and two new lines are obtained. Another iteration of the co-updating approach is exerted in (d)

Ghani (2000) showed that, when an independent and redundant feature split exists, cotraining algorithms outperform other algorithms using unlabelled data. A co-EM algorithm, which runs EM in each view and interchanges the probabilistic labels before new EM iterations, was also proposed in the paper. Abney (2002) refined the cotraining analysis and evaluated a greedy agreement algorithm by constructing rules on unlabelled samples. Goldman and Zhou (2000) presented a new cotraining strategy that does not assume there are two redundant views both of which are sufficient for perfect classification. When we have two different information sources for self-maintaining and gene functional classification, we generally fall into the cotraining setting.

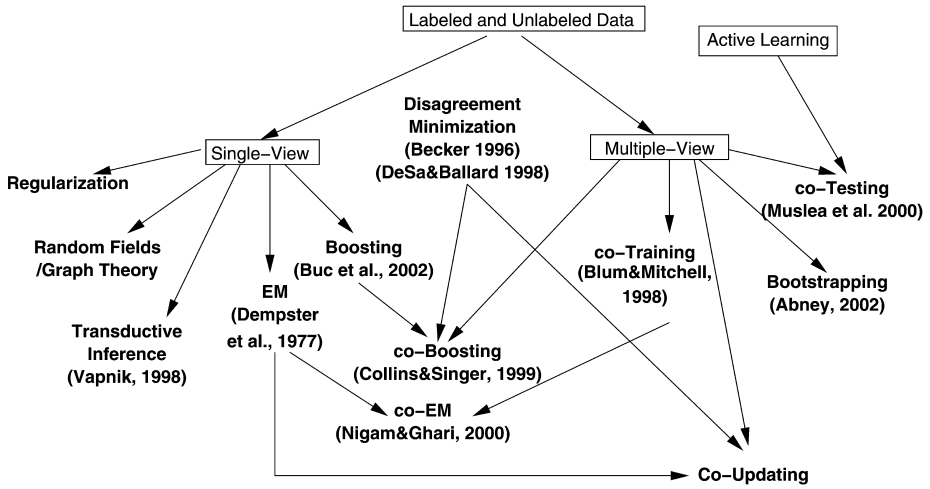


Fig. 4. The connections between various semisupervised algorithms

Transductive inference, first introduced by Vapnik (1998), aims to estimate the test labels directly from the labelled and unlabelled data without first building the latent function models. There are some other approaches that could handle both labelled and unlabelled data, such as using graph mincuts (Blum and Chawla 2001), ensemble methods (Bennett et al. 2002), boosting (Buc et al. 2002) and a combination of clustering and support vector machines (Fung and Mangasarian 2001). A detailed survey on labelled and unlabelled data is given in Seeger (2000). In brief, the related work with co-updating can be summarized in Fig. 4.

### 4.2. Interpretations and connections

In this section, we discuss the connections and differences between them and our co-updating approach as presented in Fig. 4. Cotesting (Muslea et al. 2000) is a family of active-learning algorithms that searches for the most informative unlabelled samples and asks the user to label them.

The major difference between co-updating and cotraining lies in two aspects. First, the cotraining procedure does not directly seek to find classifiers that agree on the unlabelled data and the proof given in Blum and Mitchell (1998) did not justify the procedure of the algorithm. Co-updating aims to minimise disagreement via iterative expectation-maximization (EM)-type procedures and we have presented theoretical analysis on minimising the disagreement, which would decrease the misclassification error. Second, co-updating does not commit to a label for the unlabelled samples; instead, it uses probabilistic labels that may change from iteration to iteration. By contrast, cotraining incrementally incorporates the unlabelled data into learning: the most confident unlabelled samples assigned to each class are chosen at each iteration and those labels become permanent. Hence, during the early iterations when the models have less prediction power, cotraining’s commitment to high-confidence prediction may add a large number of mislabelled samples into training. There are also some connections between them because both methods are trying to use the knowledge in both modalities to utilize unlabelled samples. In addition,

if the agreement between two views and the accuracy of the individual classifier is positively correlated, the agreed unlabelled samples usually have high confidence to be correctly labelled.

The co-updating also differs from the co-EM algorithm (Nigam and Ghani 2000) in that it directly seeks the classifiers that agree on unlabelled samples. Co-boosting (Collins and Singer 1999) explicitly expresses the minimising disagreement in the objective function and adaptively updates the distribution of the samples to build the ensemble classifier. Co-updating differs from co-boosting in that, at each iteration, only the distribution on the unlabelled samples is updated and, in particular, only the agreed unlabelled samples are considered because the algorithm only chooses a portion of the agreed unlabelled samples for training. In some sense, the ideas of coboosting and co-updating are complementary. The cotesting (Muslea et al. 2000) is an active learning strategy by selecting the disagreed samples for human labelling because these disagreed samples are most informative in the sense that at least one of them is correct. The ideas of co-updating and cotesting are orthogonal and they can be easily combined: co-updating uses the agreed unlabelled samples to train the classifier and cotesting selects the disagreed unlabelled samples for human labelling. The combination is expected to further improve the performance.

## 5. Experimental results

We performed three sets of experiments to investigate the behaviour of co-updating.

### 5.1. Experiments on the webpage dataset

We used the same dataset of web pages as in Blum and Mitchell (1998). The dataset<sup>5</sup> consists of 1,051 web pages collected from Computer Science department web sites at four universities: Cornell, University of Washington, University of Wisconsin and University of Texas. The task is to identify the web pages that are course home-pages (about 22% pages fall into the category). The two sources for each web page are the words that occur on the web page and the words occurring in the anchor text of hyperlinks pointing to the page. Classifiers that were trained separately for the individual source are referred as the page-based and hyperlink-based classifiers, respectively.

Experiments were conducted to determine whether our co-updating algorithm could successfully use the unlabelled data to outperform standard supervised training of naive Bayes classifiers. The experiment setups are the same as that in Blum and Mitchell (1998). In each experiment, 263 (25%) of the 1,051 web pages were randomly selected as a test set. The remaining data were used to generate a labelled data set containing three positive and nine negative examples drawn at random. Five trials of the experiments were conducted using different training/test splits. The results of supervised training were obtained using only the 12 labelled training samples. The combined classifiers were constructed, with the naive Bayes assumption of conditional independence, by multiplying the probability outputs of the page-based and hyperlink-based classifiers. The results are summarized in Table 1.

Numbers shown here are the test set error rates averaged over the five trials. The first row of the table shows the test error rates for the three classifiers formed by

---

<sup>5</sup> <http://www-2.cs.cmu.edu/afs/cs.cmu.edu/project/theo-51/www/co-training/data/>.

**Table 1.** Experimental results on webpages

Classifier	Page based	Hyperlink based	Combined
Supervised	15.11	12.98	11.12
Co-updating	4.50	11.52	3.86
Co-training	6.2	11.6	5.0

supervised learning; the second row shows error rates for the classifiers formed by co-updating. The third row presents the error rates of cotraining cited from Blum and Mitchell (1998).<sup>6</sup> Note that, for this data, the default hypothesis that always predicts negative achieves an error rate of 22%. Figure 5(a) gives a plot of error versus number of iterations and Fig. 5(b) gives a plot of error versus the different choice of  $\alpha$ . For all three types of classifiers (hyperlink based, page based, and combined), the classifier obtained by co-updating outperforms the classifier formed by supervised learning. On page-based and combined classifiers, the co-updating approach achieves better performance than cotraining. The hyperlink-based classifier is helped less by both cotraining and co-updating. This may be due to the fact that the hyperlinks contain fewer words and are less capable of expressing an accurate approximation to the target function. These results do indicate that the co-updating can provide a useful way of taking advantage of unlabelled data. From Fig. 5(a), we observe that the error rate of page-based classifier drops sharply as the number of iterations increases from 5 to 25 and the variations become small thereafter. As shown in Fig. 5(b), we observe that, when the selection probability  $\alpha$  becomes large, the performance of classifiers may degrade, as more unlabelled data may dominate the training. Currently, we choose  $\alpha$  based on experience. A good future research direction is to explore more the effects of  $\alpha$ .

## 5.2. Gene functional classification from heterogeneous data

### 5.2.1. Background

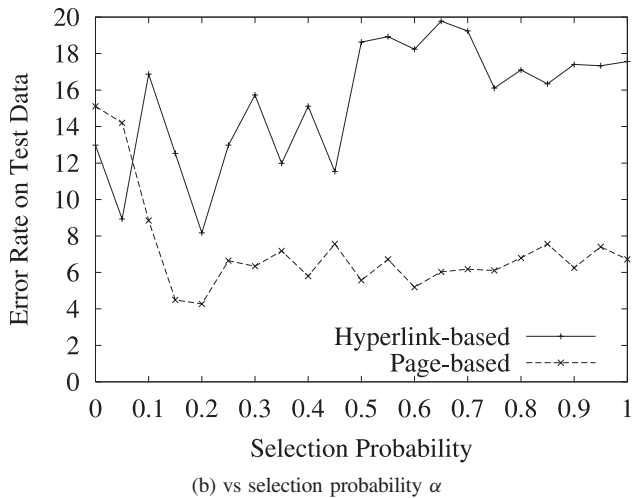
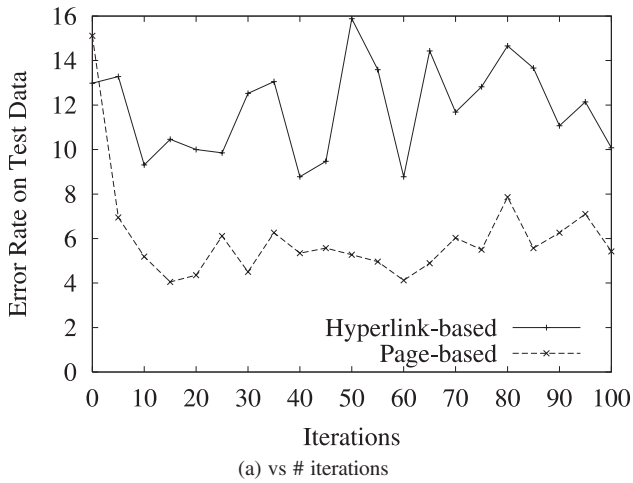
Discovery of gene functions is a fundamental problem in genetics. The microarray technology enables monitoring of gene expression of tens of thousands of genes in a single experiment and thereby makes it possible to infer functions of genes by studying correlations of expression among genes. In Pavlidis et al. (2001), gene function classifiers are built by means of feature-level integration of gene expression data and phylogenetic profiles (the existence of homologs of each gene in 24 other genomes). We used the same data set as Pavlidis et al. (2001) and applied the co-updating method.

### 5.2.2. Data description

The first set of data is gene expression of 2,465 yeast genes on 79 samples. The second set of data is the result of BLAST homolog search of these genes on 24 other genomes. Gene classification was based on the Munich Information Center for

---

<sup>6</sup> Although the results may not be directly comparable due to the randomness of the data selection, they should be able to provide enough insights on the behaviour of the co-updating approach.



**Fig. 5.** Error rate on test data

Protein Sequences Comprehensive Yeast Genome Database (CYGD)<sup>7</sup>. The database contains several hundred functional classes, the Definitions of which come from biochemical and genetic studies of gene function. Our experiments used the five most learnable classes of Pavlidis et al. (2001). The five classes, which will be referred to by I–V, are given in Table 2.

To build the classifiers, we used support vector machines (SVMs). SVMs are systems that try to classify data points in the input space by mapping them into a higher dimension feature space (using kernel function) and then finding the separating hyperplane in the feature space with the largest margin. More details on SVMs can be found in Vapnik (1998). In our experiments, as in Brown et al. (2000) and Pavlidis

<sup>7</sup> <http://mips.gsf.de/proj/yeast/CYGD/db/index.html>.

**Table 2.** The five gene functional classes

Number	Class name	Size
I	Amino acid transporters	22
II	Ribosomal proteins	173
III	Sugar & carbohydrate transporters	32
IV	Deoxyribonucleotide metabolism	9
V	Mitochondrial organization	296

**Table 3.** The distributions of the samples used for gene functional classification. TL = total labelled samples, LP = labelled positive samples, LN = labelled negative samples, UL = unlabelled samples, TS = test samples, TSP = positive test samples and TSN = negative test samples

Class	TL	LP	LN	UL	TS	TSP	TSN
I	246	5	241	1603	616	10	606
II	246	20	226	1603	616	43	573
III	246	6	240	1603	616	15	601
IV	246	2	224	1603	616	3	613
V	246	40	206	1603	616	74	542

et al. (2001), we used the kernel

$$K(X, Y) = \left( \frac{X \cdot Y}{\|X\| \|Y\|} + 1 \right)^3, \quad (7)$$

where  $X \cdot Y$  is the inner product and  $\|\cdot\|$  is the L2 norm. The co-updating parameters were set to their default values.

### 5.2.3. Results on gene functional classification

Experiments were then conducted using co-updating. For each class, about 25% genes were chosen as the test set, 10% genes were selected as labelled samples and all the remaining genes were used as unlabelled samples. The detailed distributions of the samples for each class are shown in Table 3. The experimental results are presented in Table 4. E and P show, respectively, the accuracy of expression-based classifiers and of phylogeny-based classifiers. Column A shows the results when the classifiers are individually trained using the labelled samples only. Column B shows co-updating use and Column C shows when the unlabelled samples are included in the training set with their correct labels. The quadruples are the number of true positives, the number of true negatives, the number of false positives and the number of false negatives. Such quadruples are useful in measuring accuracy when the class distribution is highly skewed (Weiss and Provost 2001). The results seem to indicate that the co-updating approach utilizes the unlabelled samples to improve the learning performance on most counts. Co-updating, however, does not help when the class data are too sparse (class IV and the expression-based classifier for classes I and III). In those cases, the initial supervised classifiers, obtained by training only with the labelled samples, have no true positives on the test sets due to the sparse nature

**Table 4.** The results on gene functional classification

Class I			
	A	B	C
E	0.9870130 (0 608 8 0)	0.9870130 (0 608 8 0)	0.9870130 (0 608 8 0)
P	0.9464286 (6 577 2 31)	0.9853896 (6 601 2 7)	0.9918831 (3 608 5 0)
Class II			
	A	B	C
E	0.9092382 (1 560 42 13)	0.9238250 (21 548 23 24)	0.9821718 (43 562 1 10)
P	0.8298217 (7 505 37 67)	0.8363047 (12 504 32 68)	0.9141005 (23 541 21 31)
Class III			
	A	B	C
E	0.9724026 (0 599 15 2)	0.9724026 (0 599 15 2)	0.9772727 (1 601 14 0)
P	0.9204545 (10 557 5 44)	0.9415584 (0 580 15 21)	0.9870130 (10 598 5 3)
Class IV			
	A	B	C
E	0.9951299 (0 613 3 0)	0.9951299 (0 613 3 0)	0.9951299 (0 613 3 0)
P	0.9935065 (0 612 3 1)	0.9935065 (0 612 3 1)	0.9967532 (1 613 2 0)
Class V			
	A	B	C
E	0.8620130 (44 487 29 56)	0.9074675 (26 536 47 7)	0.9042208 (36 521 37 22)
P	0.7678571 (20 453 53 90)	0.7889610 (21 465 52 78)	0.8474026 (31 491 42 52)

of the data. In all the other cases, the co-updating approach improves the learning performance of both classifiers.

### 5.3. Self-maintaining from different sources

#### 5.3.1. Background

Self-maintaining, which includes machine self-diagnosis, machine self-repair and customer self-help, is attracting more and more attention in many industrial systems. It brings intelligence to the industrial system to enable its robustness, high availability and cost effectiveness.

Machine failure prediction is an important step in self-maintaining and most failure prediction problems could be abstracted as the following classification prob-

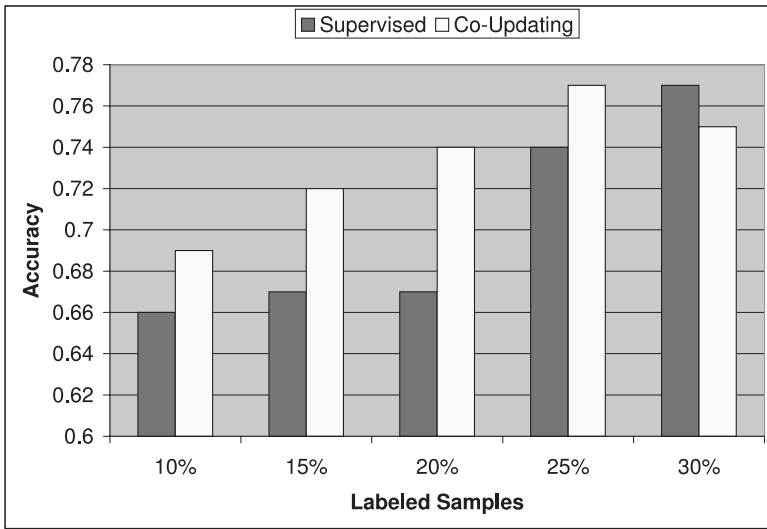
lem of predicting whether there are any failures based on current system status. Usually, when a failure occurs in an industrial system, there are several different kinds of information available for analysis. For example, when an error occurs in a copier/printer, the relevant machine data of the faulty machine, such as sensory readings and job control data, are available for analysis. In addition, image samples can also be acquired based on the hard-copy samples obtained using the faulty machine. Either information could be sufficient to predict certain kinds of failures. We have successfully used the machine data to predict cleaner failure and the image-defect analysis for ground failure of a xerographic engine. However, for some failures, methods are required to combine both information to improve the prediction accuracy. Moreover, collecting samples for machine failure prediction is usually expensive and, in most cases, we might have only limited number of labelled samples and a lot of unlabelled samples. In this section, we present the case study of using semisupervised learning for machine self-maintaining from different information sources.

### 5.3.2. Results on self-maintaining

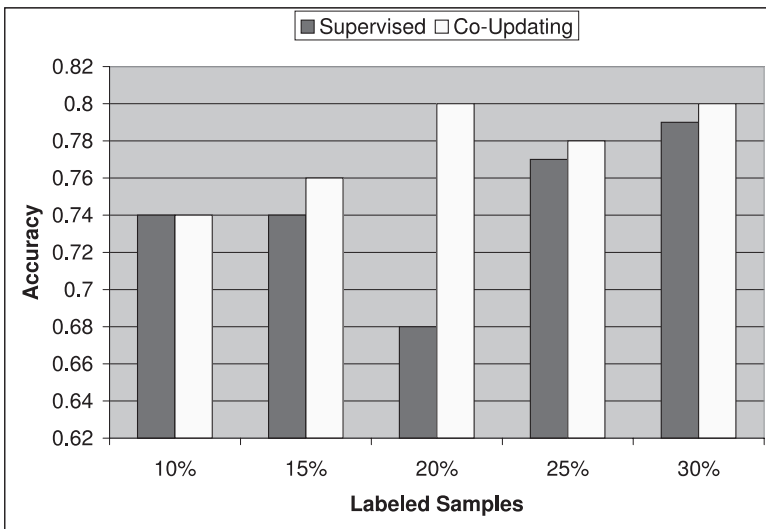
We used a dataset of cleaner failure data containing both machine and image information for a xerographic machine. The image data for each sample are a 32-dimension vector and machine data are a 46-dimension vector. One hundred samples were selected as a test set. We chose decision tree C4.5 (Quinlan 1993) as base classifiers. Decision tree produces interpretable results and has been widely used in many industrial diagnostic applications. The decision tree is a classification model in a tree structure and it is built up based on training samples. C4.5 uses the information gain of the attribute, which is the expected reduction in entropy caused by partitioning samples according to the attribute, as the measure to recursively select the splitting attribute and build the tree. A postpruning process is then carried out to prevent overfitting. A more detailed description of the algorithm can be found in Mitchell (1997). Experiments were then conducted to determine whether this co-updating algorithm could successfully use the unlabelled data to outperform standard supervised algorithms. Classifiers that were trained separately for different sources are referred as image-based classifiers and machine-based classifier. Figures 6(a) and 6(b) present, respectively, the performance comparisons of supervised learning and co-updating on the image-based classifier and machine-based classifiers with different sizes of labelled samples. The results of supervised learning were obtained by using the classifiers built from the labelled data. The co-updating approach outperforms the supervised learning on the machine-based classifier. However, as shown in Fig. 6(a), co-updating degrades the performance of the image-based classifier in some cases. This suggests that unlabelled data sometimes may hurt. The fact that using unlabelled data in addition to a set of labelled data occasionally hurts instead of being beneficial has been previously reported (Zhang and Oles 2000; Nigam 2001; Cozman and Cohen 2002). The reason may be that the consistency constraints in the model is violated.

## 6. Conclusion and discussion

In this paper, we propose a co-updating approach for semisupervised learning from multiple information sources. The co-updating approach tries to minimise the dis-



(a) Image-based classifier



(b) Machine-based classifier

Fig. 6. Performance comparison for self-maintaining

agreement between the individual models and makes use of both labelled and unlabelled data. We have conducted three sets of experiments on different datasets investigating the behaviour of the co-updating approach.

There are several natural avenues for future research. First, as mentioned in Sect. 5.1, a good future research direction is to investigate the effects of the selection probability  $\alpha$ . Second, the co-updating approach can be applied to multimodal learning tasks, such as word learning and object recognition, because spatiotemporal and cross-modal coherence is a powerful constraint in sensory data of the physical world. We could apply the co-updating approach to utilize the approximately co-

incident information of different modalities in these tasks. Third, another obvious research direction is to extend the co-updating approach for multiple data sources and to handle multiclass classification. Finally, many feature selection techniques could be incorporated into the co-updating approach.

## References

- Abney S (2002) Bootstrapping. In: Proceedings of the 40th annual meeting of the Association for Computational Linguistics (ACL 2002). Morgan Kaufmann, San Francisco, CA, pp 360–367
- Allwein EL, Schapire RE, Singer Y (2000) Reducing multiclass to binary: a unifying approach for margin classifiers. In: Langley P (ed) Proceedings of the 17th international conference on machine learning (ICML 2000). Morgan Kaufmann, San Francisco, CA, pp 9–16
- Becker S (1996) Mutual information maximization: models of cortical self-organization. *Netw Comput Neural Syst* 7:7–31
- Bennett KP, Demiriz A, Maclin R (2002) Exploiting unlabeled data in ensemble methods. In: Proceedings of the eighth ACM SIGKDD international conference on knowledge discovery and data mining (SIGKDD 2002). ACM Press, New York, NY, pp 289–296
- Blum A, Chawla S (2001) Learning from labeled and unlabeled data using graph mincuts. In: Brodley CE, Danyluk AP (eds) Proceedings of the eighteenth international conference on machine learning (ICML 2001). Morgan Kaufmann, San Francisco, CA, pp 19–26
- Blum A, Mitchell T (1998) Combining labeled and unlabeled data with co-training. In: Proceedings of the eleventh annual conference on computational learning theory (COLT 1998). ACM Press, New York, NY, pp 92–100
- Brown MPS, Grundy WN, Lin D, Cristianini N, Sugnet C, Furey TS, Manuel Ares J, Haussler D (2000) Knowledge-based analysis of microarray gene expression data using support vector machines. *Proceedings of the National Academy of Sciences* 97(1):262–267
- Buc F, Grandvalet Y, Ambroise C (2002) Semi-supervised marginboost. In: Dietterich TG, Becker S, Ghahramani Z (eds) Advances in neural information processing systems 14 (NIPS 2001). The MIT Press, Cambridge, MA, pp 553–560
- Castelli V, Cover T (1996) The relative value of labeled and unlabeled samples in pattern recognition with an unknown mixing parameter. *IEEE Trans Inf Theory* 42:2102–2117
- Collins M, Singer Y (1999) Unsupervised models for named entity classification. In: Proceedings of the 1999 joint SIGDAT conference on empirical methods in natural language processing and very large corpora. ACM Press, New York, NY, pp 100–110
- Cozman FG, Cohen I (2002) Unlabeled data can degrade classification performance of generative classifiers. In: Proceedings of the fifteenth international Florida artificial intelligence society conference (FLAIRS 2002), pp 327–331
- Dasgupta S, Littman ML, McAllester D (2001) PAC generalization bounds for co-training. In: Dietterich TG, Becker S, Ghahramani Z (eds) Advances in neural information processing systems 14 (NIPS 2001). The MIT Press, Cambridge, MA, pp 375–382
- De Sa VR, Ballard D (1998) Category learning through multimodality sensing. *Neural Comput* 10:1097–1117
- Dempster A, Laird N, Rubin D (1977) Maximum likelihood from incomplete data via the EM algorithm. *J Roy Stat Soc B* 39(1):1–38
- Fung G, Mangasarian O (2001) Semi-supervised support vector machines for unlabeled data classification. *Optim Methods Softw* 15:29–44
- Ghahramani Z, Jordan MI (1994) Supervised learning from incomplete data via an EM approach. In: Cowan JD, Tesauro G, Alspector J (eds) Advances in neural information processing systems 6 (NIPS 1993). Morgan Kaufmann, San Francisco, CA, pp 120–127
- Goldman S, Zhou Y (2000) Enhancing supervised learning with unlabeled data. In: Langley P (ed) Proceedings of the 17th international conference on machine learning (ICML 2000). Morgan Kaufmann, San Francisco, CA, pp 327–334
- Li T, Zhu S, Li Q, Ogihara M (2003) Gene functional classification by semi-supervised learning from heterogeneous data. In: Proceedings of the 18th annual ACM symposium on applied computing (SAC'03). ACM Press, New York, NY, pp 78–82
- McGurk H, MacDonald J (1976) Hearing lips and seeing voices. *Nature* 264:746–748
- Mitchell TM (1997) Machine learning. McGraw-Hill
- Muslea I, Minton S, Knoblock CA (2000) Selective sampling with redundant views. In: Proceedings of the seventeenth national conference on artificial intelligence and twelfth conference on innovative applications of artificial intelligence (AAAI/IAAI 2000). The MIT Press, Cambridge, MA, pp 621–626

- Neal R, Hinton G (1998) A view of the EM algorithm that justifies incremental, sparse, and other variants. In: Jordan MI (ed) Learning in graphical models. The MIT Press, Cambridge, MA, pp 355–368
- Nigam K (2001) Using unlabeled data to improve text classification. Doctoral Dissertation, Computer Science Department, Carnegie Mellon University, Pittsburgh, PA
- Nigam K, Ghani R (2000) Analyzing the effectiveness and applicability of co-training. In: Proceedings of the ninth international conference on information and knowledge management (CIKM 2000). ACM Press, New York, NY, pp 86–93
- Nigam K, McCallum AK, Thrun S, Mitchell TM (2000) Text classification from labeled and unlabeled documents using EM. *Mach Learn* 39:103–134
- Pavlidis P, Weston J, Cai J, Grundy WN (2001) Gene functional classification from heterogeneous data. In: Proceedings of fifth annual international conference on computational molecular biology (RECOMB 2001). ACM Press, New York, NY, pp 249–255
- Quinlan J (1993) C4.5: Programs for machine learning. Morgan Kaufmann, San Francisco, CA
- Roth D, Zelenko D (2000) Toward a theory of learning coherent concepts. In: Proceedings of the seventeenth national conference on artificial intelligence and twelfth conference on innovative applications of artificial intelligence (AAAI/IAAI 2000). The MIT Press, Cambridge, MA, pp 639–644
- Seeger M (2000) Learning with labeled and unlabeled data. Technical report, Institute for Adaptive and Neural Computation, University of Edinburgh, Edinburgh, UK
- Vapnik VN (1998) Statistical learning theory. Wiley, New York, NY
- Weiss G, Provost F (2001) The effect of class distribution on classifier learning: an empirical study. Technical report ML-TR 44, Department of Computer Science, Rutgers University, New Brunswick, NJ
- Wu L, Oviatt SL, Cohen PR (1999) Multimodal integration—a statistical view. *IEEE Trans Multimedia* 1:334–341
- Zhang T, Oles F (2000) A probability analysis on the value of unlabeled data for classification problems. In: Langley P (ed) Proceedings of the 17th international conference on machine learning (ICML 2000). Morgan Kaufmann, San Francisco, CA, pp 1191–1198

## Author biographies



**Tao Li** received his B.S. degree in computer science from Fuzhou University, China, and M.S. degree in computer science from Chinese Academy of Science. He also got a M.S. degree in mathematics from Oklahoma State University. He is currently a doctoral candidate in the computer science department at the University of Rochester. His primary research interests are data mining, machine learning and music information retrieval.



**Mitsunori Ogihara** received a Ph.D. in information sciences at Tokyo Institute of Technology in 1993. He is currently professor and chair of the Department of Computer Science at the University of Rochester. His primary research interests are data mining, computational complexity and molecular computation.