

Universal Learning over Related Distributions and Adaptive Graph Transduction

Erheng Zhong¹, Wei Fan², Jing Peng³,
Olivier Verscheure², and Jiangtao Ren^{1,*}

¹ Sun Yat-Sen University, Guangzhou, China
{sw04zheh,issrjt}@mail2.sysu.edu.cn

² IBM T.J. Watson Research, USA
{weifan,ovl}@us.ibm.com

³ Montclair State University, USA
pengj@mail.montclair.edu

Abstract. The basis assumption that “training and test data drawn from the same distribution” is often violated in reality. In this paper, we propose one common solution to cover various scenarios of learning under “different but related distributions” in a single framework. Explicit examples include (a) sample selection bias between training and testing data, (b) transfer learning or no labeled data in target domain, and (c) noisy or uncertain training data. The main motivation is that one could ideally solve as many problems as possible with a single approach. The proposed solution extends graph transduction using the maximum margin principle over unlabeled data. The error of the proposed method is bounded under reasonable assumptions even when the training and testing distributions are different. Experiment results demonstrate that the proposed method improves the traditional graph transduction by as much as 15% in accuracy and AUC in all common situations of distribution difference. Most importantly, it outperforms, by up to 10% in accuracy, several state-of-art approaches proposed to solve specific category of distribution difference, i.e, BRSD [1] for sample selection bias, CDSC [2] for transfer learning, etc. The main claim is that the adaptive graph transduction is a general and competitive method to solve distribution differences implicitly without knowing and worrying about the exact type. These at least include sample selection bias, transfer learning, uncertainty mining, as well as those alike that are still not studied yet. The source code and datasets are available from the authors.

1 Introduction

One important assumption in many learning scenarios is that training and test data are drawn from the same distribution. However, this may not be true in many applications, and the following are some examples. First, suppose we wish

* The author is supported by the National Natural Science Foundation of China under Grant No. 60703110.

to generate a model in clinical trial of a new drug. Since people self-select, the training data is most likely having a different distribution from the general public. Second, if we want to use a topic model to classify articles from New York Times, but the only labeled data is from Reuters, we have to transfer knowledge across these two collections. Third, for data collected over sensor networks, the feature values are likely noisy or uncertain. Obviously, the distributions of training and test data in the above problems are different but related. In this paper, we formulate this situation within a “universal learning” framework.

Given a space of instances X and labels $Y = [-1, 1]$, let $p_{tr}(\mathbf{x}, y)$ denotes the joint distribution of training data $L \subseteq X \times Y$ and $p_{te}(\mathbf{x}, y)$ denotes the test data $U \subseteq X \times Y$. Using the standard decomposition, $p(\mathbf{x}, y)$ can be represented by $p(\mathbf{x}, y) = p(\mathbf{x})p(y|\mathbf{x})$ where $p(\mathbf{x})$ and $p(y|\mathbf{x})$ are the marginal and conditional distributions. Let $h \in X \rightarrow Y$ be a function from a fixed hypothesis space \mathcal{H} for X . Then, we define the distance between training and tests sets using a hypothesis class-specific distance measure. Let $A_{\mathcal{H}}$ be the subsets of X that are the support of some hypothesis in \mathcal{H} . In other words, for every hypothesis $h \in \mathcal{H}$, $\{\mathbf{x} : \mathbf{x} \in X, h(\mathbf{x}) = 1\} \in A_{\mathcal{H}}$. Then the distance between two distributions is:

$$d_{\mathcal{H}}(p_{tr}(\mathbf{x}), p_{te}(\mathbf{x})) = 2 * \sup_{A \in A_{\mathcal{H}}} |Pr_{p_{tr}}[A] - Pr_{p_{te}}[A]| \quad (1)$$

Using the conclusion from [3], we compute a finite-sample approximation to $d_{\mathcal{H}}$, where \mathcal{H} has finite VC dimensions. Thus, $d_{\mathcal{H}}$ can be treated as the indicator to measure the relationship between training and test set. We define the following framework as “universal learning over related but different distributions”:

Definition 1. *Given the training and test set, learning is universal iff $p_{tr}(\mathbf{x}, y) \neq p_{te}(\mathbf{x}, y)$ and $d_{\mathcal{H}}$ is small.*

With this definition, we study solutions for problems where $d_{\mathcal{H}}$ is reasonably small or training and testing data is related but different.

Note that this definition is unified in the sense that it covers many related problem formulations, such as sample selection bias, transfer learning, uncertainty mining and the alike that are not well studied and reported yet. For both sample selection bias and uncertainty mining, $p_{tr}(\mathbf{x}) \neq p_{te}(\mathbf{x})$ but $p_{tr}(y|\mathbf{x}) = p_{te}(y|\mathbf{x})$. For transfer learning, $p(\mathbf{x})$ and $p(y|\mathbf{x})$ of training and test data sets may both be different, but $p(y|\mathbf{x})$ are assumed to be related. As follows, we propose a generalized graph approach under this framework. One can solve as many different but similar problems as possible and avoid employing and remembering different approaches under different formulations.

1.1 Solution Based on Adaptive Graph Transduction

To solve universal learning problem, we propose an approach based on “maximum margin graph transduction”. Graph transduction explores the information from both training and test set. Its most important advantage for universal learning is that it does not explicitly assume the distributions of training and

test set to be the same, but only makes a weaker assumption that the decision boundary lies on the low density regions of the unlabeled data. However, the original graph transduction still suffers from the “unsmooth label problems” and these are common when learning different distributions. It may instead mislead the decision boundary to go through the high density regions, when labeled examples of training examples stand on the wrong location in the space of the testing data [4]. In margin-terms, unlabeled data with low margin are likely misclassified [5]. Clearly, if one employs graph transduction for universal learning, label information ought be regularized in order to maintain smoothness. Based on this motivation, we propose a maximal margin based graph transduction on the basis of “maximal unlabeled data margin principal.”

To solve this problem, we cast the graph transduction into a joint optimization over both the classification function and the labeled data. The optimization is solved iteratively. In each iteration, sample selection is performed on labeled set to select data which can maximize the unlabeled data margin. Based on the selected sample, graph transduction is invoked to predict the labels of unlabeled data. Those closest examples will be predicted with the same class label. Then, the prediction results from each iteration are averaged to form an ensemble, in order to remove the bias from any single graph [6] and further reduce the prediction error as shown in Section 3. By the analysis of Section 3, the risk of proposed method is bounded under reasonable terms even when training and testing distributions are different.

2 Graph Transduction over Related Distributions

We first summarize the traditional graph transduction using harmonic function [7], and then present details on the proposed algorithm MarginGraph that is generalized by the maximal margin principal over unlabeled data. The notations are summarized in Table 1.

2.1 Preliminaries: Graph Transduction Using Harmonic Functions

Suppose we have ℓ training examples $L = \{(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_\ell, y_\ell)\}$, and u test examples $U = \{(\mathbf{x}_{\ell+1}, y_{\ell+1}), \dots, (\mathbf{x}_{\ell+u}, y_{\ell+u})\}$. The labels of U are not known

Table 1. Definition of notation

Notation	Notation Description	Notation	Notation Description
X	Instance space	$maxIt$	Iteration times in algorithm
Y	Label space	G	Graph whose nodes are data points and the edges describe the similar between any nodes
\mathbf{x}_i	Instance(without label), $x_i \in X$	W	Weight matrix of the graph G , w_{ij} is the weight of the edge between node i and j
y_i	Label of instance \mathbf{x}_i , $y_i \in \{-1, 1\}$	D	Diagonal matrix, $D = diag(d_i)$, $d_i = \sum_j w_{ij}$
L	Training data set	$p(\mathbf{x})$	Margin distribution
ℓ	Number of instances in L	\mathcal{H}	A fixed hypothesis space
U	Test data set	$m_Q(U)$	Unlabeled data margin
u	Number of instances in U		
Δ	Laplacian matrix, $\Delta = D - W$		
$p(\mathbf{x}, y)$	Joint distribution		
$p(y \mathbf{x})$	Conditional distribution		
Q	A posterior distribution over \mathcal{H}		

apriori. In universal learning setting, L and U are drawn from the related but different distributions. We construct a connected graph $G = \{V, E\}$. Vertex V corresponds to both labeled and unlabeled data examples, and the edges $(i, j) \in E, i, j = 1, \dots, \ell + u$, are weighted according to the similarity between \mathbf{x}_i and \mathbf{x}_j . According to the results from [7], we set the weight $w_{ij} = \exp(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{\lambda^2})$, where λ is a bandwidth hyper-parameter. It is adaptive according to the data. Following the analysis in [7], we set $\lambda = d^0/3$, where d^0 is the minimal distance between class regions. Also, Let D be a diagonal matrix, $D = \text{diag}(d_i)$ where $d_i = \sum_j w_{ij}$. Based on the intuition that unlabeled examples that are close-by in the graph ought to have similar labels [7], a harmonic function is defined as $E(f) = \frac{1}{2} \sum_{i,j} w_{ij} (f(\mathbf{x}_i) - f(\mathbf{x}_j))^2 = f^T \Delta f$, where $f : V \rightarrow [-1, 1]$ is a real-valued mapping function to minimize the value of harmonic function and $\Delta = D - W$ is the Laplacian matrix. The harmonic solution $\Delta f = 0$ is given by:

$$f_U = \Delta_{UU}^{-1} \Delta_{UL} f_L \tag{2}$$

where f_U denotes the values on the unlabeled data, f_L represents the values on the labeled data (equal to the true label), Δ_{UU} is the sub-matrix of Δ relating the unlabeled data to unlabeled data, and Δ_{UL} relates unlabeled data to labeled data. For binary classification, we obtain a classifier based on the mapping function f and use it to predict the labels of unlabeled examples.

$$\hat{y}_i = I[f(\mathbf{x}_i)] = \begin{cases} 1, f(\mathbf{x}_i) > \theta \\ -1, \text{otherwise} \end{cases} \tag{3}$$

For balanced problem, θ is typically chosen to be 0. Based on the discussion in introduction, in universal learning over related but different distributions, the typical graph transduction suffers from label unsmooth problem that the initial labels stand on the wrong locations in the domain of testing data.

2.2 Adaptive Graph Transduction by Maximizing Unlabeled Data Margin

As follows, we focus on how to use sample selection to resolve the unsmooth problem based on maximal margin principal. Let Q be a posterior distribution over hypothesis space \mathcal{H} . Then the “ Q -weighted majority vote Bayes classifier B_Q ” is

$$B_Q(\mathbf{x}) = I[\mathbb{E}_{h \sim Q} h(\mathbf{x})] \tag{4}$$

where $I[x] = 1$ if $x > \theta$ and -1 otherwise, and h is a classifier in \mathcal{H} according to the distribution Q . Thus, the unlabeled data margin is:

$$\begin{aligned} m_Q(\mathbf{x}_i) &= |\mathbb{E}_{h \sim Q} [h(\mathbf{x}_i) = 1] - \mathbb{E}_{h \sim Q} [h(\mathbf{x}_i) = -1]| \\ &= |1 - 2\mathbb{E}_{h \sim Q} [h(\mathbf{x}_i) \neq y_i]| \end{aligned} \tag{5}$$

where $\mathbb{P}[\pi]$ denotes the probability that π is true. In graph transduction,

$$m_Q(\mathbf{x}_i) = ||f(\mathbf{x}_i)| - |1 - |f(\mathbf{x}_i)|| \tag{6}$$

And the unlabeled margin on whole unlabeled data set is $m_Q(U) = \sum_{i=\ell+1}^{\ell+u} m_Q(\mathbf{x}_i)$. From the intuition in introduction (formally analyzed in Section 3), graph transduction can be generalized by maximizing $m_Q(U)$ by performing sample selection in training set. This can be formulated as

$$S_L = \arg \max_{S'_L \subseteq L} m_Q(U)_{S'_L, h} \tag{7}$$

where S_L is the selected labeled subset used for building a maximal margin classifier h . To obtain S_L , we employ a greedy sequence searching procedure in the labeled data. Suppose the current mapping function is f . When we select one more labeled example $(\mathbf{x}_k, y_k) \in L$, we denote the new mapping function as f^{+k} . Then, in each iteration, we select one labeled example from L as follows:

$$k = \arg \max_{k'} \left(\sum_{i=\ell+1}^{\ell+u} ||f^{k'}(\mathbf{x}_i)| - |1 - |f^{k'}(\mathbf{x}_i)|| \right) \tag{8}$$

The above equation means that we select those data that can build a classifier to maximize the unlabeled data margin. This procedure is a joint optimization that simultaneously (1) maximizes the unlabeled data margin and (2) minimizes the harmonic function. Detailed description of MarginGraph can be found in Algorithm 1. In summary, it runs iteratively. In the beginning, the selected labeled data set S_L is empty. In each iteration, one labeled example is selected according to Eq(8), and added into the training set S_L in order to obtain the maximal unlabeled margin. Then the graph transduction is invoked on the new training set to calculate the harmonic function, as well as, to find a new mapping function f^{+k} . After the iteration, we combine all mapping functions f calculated during each iteration and obtain an averaging ensemble classifier which aims to remove the bias by any single graph and reduce the prediction error. In order to keep balanced prior class distribution, we select one positive and one negative labeled example alternatively during the iterative procedure. To practically guarantee the margin $m_Q(U)$ is not small, we propose a criterion to stop the iteration. If the following Eq(9) holds, we stop the iterative procedure.

$$m_Q(U)^* - m_Q(U)^k \geq \varepsilon m_Q(U)^* \tag{9}$$

where $m_Q(U)^k$ is the margin value after we select the k_{th} labeled data point, $m_Q(U)^*$ is the maximal margin value obtained in the procedure so far and ε is a real-value in $(0, 1]$ to control the iterations.

2.3 Implementation Details

At each iteration, we need to compute the mapping function f^{+k} after adding (\mathbf{x}_k, y_k) into the selected labeled data set. We discuss an efficient implementation to retrain and reduce the computational cost. We first “suppose” data from both L and U are all “unlabeled”. Thus the problem can be treated as labeling the selected data from L to maximize the margin on U . Denote the unlabeled data

```

Input:  $L, U, maxIt$ 
Output:  $B_Q$ : Averaging ensemble or Bayes classifier
1  $S_L = \{\}, f_0(\mathbf{x}_i) = 0, x_i \in U;$ 
2 for  $i = 1$  to  $maxIt$  do
3   Select one point  $(\mathbf{x}_k, y_k) \in L$  using Eq(8),  $S_L = S_L \cup (\mathbf{x}_k, y_k);$ 
4   Calculate  $f_i$  using Eq(2) based on  $S_L$  and new margin value
      $m_Q(U)^k;$ 
5   IF Eq(9) holds Break. ;
6   Otherwise, Update the maximal margin  $m_Q(U)^* = m_Q(U)^k;$ 
7 end
8 return  $B_Q = I[\frac{1}{maxIt} \sum_{i=1}^{maxIt} f_i];$ 
    
```

Fig. 1. MarginGraph

in U and the unselected labeled data in L as \mathcal{U} , and the selected data from L as \mathcal{L} with the labels $Y_{\mathcal{L}}$. Suppose we are labeling the k_{th} point from L , once we give the point \mathbf{x}_k label y_k , we obtain: $f_{\mathcal{U}}^{+k} = f_{\mathcal{U}} + (y_k - f_k) \frac{(\Delta_{\mathcal{U}\mathcal{U}}^{-1})_{.k}}{(\Delta_{\mathcal{U}\mathcal{U}}^{-1})_{kk}}$, where $(\Delta_{\mathcal{U}\mathcal{U}}^{-1})_{.k}$ is the k_{th} column of the inverse Laplacian on “unlabeled” data \mathcal{U} , and $(\Delta_{\mathcal{U}\mathcal{U}}^{-1})_{kk}$ is the k_{th} diagonal element of the same matrix. Importantly, this means updating the values of mapping function f is a linear computation. When one “unlabeled” example in L is labeled, the corresponding row and column are removed from $\Delta_{\mathcal{U}\mathcal{U}}^{-1}$, so $\Delta_{\mathcal{U}\mathcal{U}}^{-1}$ should be recomputed. Instead of naively taking the inverse, there are efficient algorithms to compute in linear time [8]. In summary, the proposed approach has computation complexity of $O(\ell * (\ell + u) * maxIt)$.

3 Formal Analysis

Theorem 1 shows that the error of a classifier across different distributions is bounded. Lemma 1 shows that this bound is reduced as margin being maximized. Finally, Lemma 2 demonstrates that the averaging ensemble achieves larger margin than any single classifiers, implying that the ensemble classifier has a lower error bound.

To measure the difference between training and test distributions, we define a distance based on Eq(1): $d_{\mathcal{H}\Phi\mathcal{H}}(p_{tr}, p_{te})$, where $\mathcal{H}\Phi\mathcal{H} = \{h(\mathbf{x}) \oplus h'(\mathbf{x}) : h, h' \in \mathcal{H}\}$ represents the symmetric difference hypothesis space, and \oplus denotes the XOR operator. A hypothesis $h \in \mathcal{H}\Phi\mathcal{H}$ assigns label +1 to \mathbf{x} when there is a pair of hypotheses in \mathcal{H} that disagree on \mathbf{x} . Thus, $A_{\mathcal{H}\Phi\mathcal{H}}$ is the subset of $A_{\mathcal{H}}$ for some $h, h' \in \mathcal{H}$, such that $A_{\mathcal{H}\Phi\mathcal{H}} = \{\mathbf{x} | \mathbf{x} \in X, h(\mathbf{x}) \neq h'(\mathbf{x})\}$. Let $\epsilon_{tr}(h, h')$ be the probability that a hypothesis h disagrees with another one h' according to the marginal distribution $p_{tr}(\mathbf{x})$. Then

$$\epsilon_{tr}(h, h') = \mathbb{E}_{\mathbf{x} \sim p_{tr}(\mathbf{x})} |h(\mathbf{x}) - h'(\mathbf{x})| \tag{10}$$

In particular, if $f^* : X \rightarrow Y$ denotes the (unknown) target (labeling) function, $\epsilon_{tr}(h) = \epsilon_{tr}(h, f_{tr}^*)$ represents the risk of the hypothesis h . Similarly, $\epsilon_{te}(h, h')$

and $\epsilon_{te}(h)$ are the corresponding definitions for the test distribution. It can be shown that for any hypotheses $h, h' \in \mathcal{H}$ [3],

$$|\epsilon_{tr}(h, h') - \epsilon_{te}(h, h')| \leq \frac{1}{2}d_{\mathcal{H}\Phi\mathcal{H}}(p_{tr}, p_{te}) \tag{11}$$

Since the margin evaluates the confidence of a classifier with regard to its decision [5], we define an ideal hypothesis as the one that maximizes the margin on unlabeled data.

$$h^* = \arg \max_{h \in \mathcal{H}}(m_Q(U)_h) \tag{12}$$

where $m_Q(U)_h$ is the margin obtained by classifier h . Following these definitions, the bound for the risk of any classifiers h can be established when the distributions of training and test data are related but different. The bound is adopted from Theorem 1 of [3], but with different terms.

Theorem 1. *Let \mathcal{H} be a hypothesis space of VC-dimension d_{VC} and U_{tr} and U_{te} be unlabeled samples of size m each, drawn according to $p_{tr}(\mathbf{x})$ and $p_{te}(\mathbf{x})$, respectively. Let $\hat{d}_{\mathcal{H}\Phi\mathcal{H}}$ be the empirical distance on U_{tr} and U_{te} . With probability of at least $1 - \delta$ (over the choice of the samples), for any classifiers $h \in \mathcal{H}$,*

$$\epsilon_{te}(h) \leq \epsilon_{tr}(h, h^*) + \epsilon_{te}(h^*) + \frac{1}{2}\hat{d}_{\mathcal{H}\Phi\mathcal{H}}(U_{tr}, U_{te}) + 4\sqrt{\frac{2d_{VC} \log(2m) + \log(\frac{4}{\delta})}{m}}$$

Proof. The proof uses the inequality Eq(11). Also, we assume the triangle inequality holds for classification error [3]. It implies for any functions f_1, f_2 and $f_3, \epsilon(f_1, f_2) \leq \epsilon(f_1, f_3) + \epsilon(f_2, f_3)$. Thus,

$$\begin{aligned} \epsilon_{te}(h) &\leq \epsilon_{te}(h, h^*) + \epsilon_{te}(h^*) \\ &\leq \epsilon_{tr}(h, h^*) + \epsilon_{te}(h^*) + |\epsilon_{tr}(h, h^*) - \epsilon_{te}(h, h^*)| \\ &\leq \epsilon_{tr}(h, h^*) + \epsilon_{te}(h^*) + \frac{1}{2}d_{\mathcal{H}\Phi\mathcal{H}}(p_{tr}, p_{te}) \\ &\leq \epsilon_{tr}(h, h^*) + \epsilon_{te}(h^*) + \frac{1}{2}\hat{d}_{\mathcal{H}\Phi\mathcal{H}}(U_{tr}, U_{te}) + 4\sqrt{\frac{2d \log(2m) + \log(\frac{4}{\delta})}{m}}. \quad \square \end{aligned}$$

Note that this bound is constructed by three terms. The first term, $\epsilon_{tr}(h, h^*)$ represents the training error in terms of approximating the ideal hypothesis h^* . The second is the risk of h^* . Recall the definition of Eq(12), h^* just relies on the test data set and is independent from any algorithms. When the unlabeled data are given, the risk of h^* is fixed. In addition, the third term is the distance between the training and test distributions, $d_{\mathcal{H}\Phi\mathcal{H}}(p_{tr}, p_{te})$. If the training and test distributions are related, $d_{\mathcal{H}\Phi\mathcal{H}}$ can be bounded. Therefore, the bound mostly relies on $\epsilon_{tr}(h, h^*)$.

The following lemma and analysis show when training and test distributions are related, if a classifier h achieves larger margin, $\epsilon_{tr}(h, h^*)$ becomes smaller. We assume that for a given instance \mathbf{x} , the misclassification probabilities of h and h^* are smaller than 50%.

Lemma 1. *Let $m_Q(U)_h$ denotes the unlabeled margin obtained by h , then $\epsilon_{tr}(h, h^*)$ is related to $|m_Q(U)_{h^*} - m_Q(U)_h|$.*

Proof. By the definition of Eq(6),

$$\begin{aligned}
 & |m_Q(U)_{h^*} - m_Q(U)_h| \\
 &= u * \mathbb{E}_{\mathbf{x}_i \sim p_{te}} \{ |1 - 2\llbracket h^*(\mathbf{x}_i) \neq y_i \rrbracket| - |1 - 2\llbracket h(\mathbf{x}_i) \neq y_i \rrbracket| \} \\
 &= u * \mathbb{E}_{\mathbf{x}_i \sim p_{te}} \{ (1 - 2\llbracket h^*(\mathbf{x}_i) \neq y_i \rrbracket) - (1 - 2\llbracket h(\mathbf{x}_i) \neq y_i \rrbracket) \} \\
 &= u * \mathbb{E}_{\mathbf{x}_i \sim p_{te}} 2 * (\llbracket h(\mathbf{x}_i) \neq y_i \rrbracket - \llbracket h^*(\mathbf{x}_i) \neq y_i \rrbracket) \\
 &= u * \mathbb{E}_{\mathbf{x}_i \sim p_{te}} (\llbracket h(\mathbf{x}_i) \neq y_i \rrbracket - \llbracket h^*(\mathbf{x}_i) \neq y_i \rrbracket) + (\llbracket h^*(\mathbf{x}_i) = y_i \rrbracket - \llbracket h(\mathbf{x}_i) = y_i \rrbracket)
 \end{aligned}$$

Obviously, when $|m_Q(U)_{h^*} - m_Q(U)_h|$ is small, h and h^* give similar classification probabilities. Thus, if the margins achieved by h^* and h are close, $\epsilon_{te}(h, h^*)$ is small. Recall the Eq(11), if the training and test distribution are related, smaller $\epsilon_{te}(h, h^*)$ induces smaller $\epsilon_{tr}(h, h^*)$. In summary, if one classifier h has larger unlabeled data margin, it will make the $\epsilon_{tr}(h, h^*)$ smaller. \square

More specifically, $\epsilon_{tr}(h, h^*)$ and $\epsilon_{te}(h, h^*)$ are equivalent when the training and test distributions are the same. Under this situation, the bound becomes $\epsilon_{te}(h) \leq \epsilon_{te}(h, h^*) + \epsilon_{te}(h^*)$. That implies, when the distributions of training and test data are the same, the error bound of the maximal margin classifier is the lowest.

As follows, we analyse the idea behind the averaging ensemble.

Lemma 2. *Unlabeled data margin achieved by averaging ensemble is not smaller than any single classifiers.*

Proof. Let $h_E = \mathbb{E}_{h \sim Q} h(\mathbf{x})$ denotes the averaging ensemble where Q is the posterior distribution of selecting h . And let $d_m(h, h^*)$ denotes the difference between the margins obtained by a classifier h and the ideal hypothesis h^*

$$\begin{aligned}
 d_m(h, h^*) &= \mathbb{E}_{h \sim Q, x_i \sim p_{te}} (m_Q(x_i)_h - m_Q(x_i))^2 \\
 &= \mathbb{E}_{h \sim Q, x_i \sim p_{te}} (m_Q(x_i)_h^2 - m_Q(x_i)_h * m_Q(x_i) + m_Q(x_i)^2)
 \end{aligned}$$

The margin difference between the ensemble h_E and the ideal hypothesis h^* is

$$\begin{aligned}
 d_m(h_E, h^*) &= \mathbb{E}_{x_i \sim p_{te}} (\mathbb{E}_{h \sim Q} m_Q(x_i)_h - m_Q(x_i))^2 \\
 &= \mathbb{E}_{x_i \sim p_{te}} ((\mathbb{E}_{h \sim Q} m_Q(x_i)_h)^2 - \mathbb{E}_{h \sim Q} m_Q(x_i)_h * m_Q(x_i) + m_Q(x_i)^2) \\
 &\leq d_m(h, h^*) \quad \text{as} \quad \mathbb{E}[f(x)]^2 \leq \mathbb{E}[f(x)^2] \quad \square
 \end{aligned}$$

Therefore, on average, margin distance between averaging ensemble and ideal hypothesis h^* is smaller than any single classifiers. In other words, the ensemble achieves larger margin. According to Lemma 1, the difference between the ensemble and ideal hypothesis h^* is smaller, and the bound is lower.

In summary, the error of a classifier can be bounded in universal learning and the proposed maximal margin classifier has a lower bound. Moreover, averaging ensemble further reduces the prediction error on the unlabeled data.

Table 2. Data Set Summary

Data Set	#Training	#Test	Description	Data Set	#Training	#Test	Description
Transfer learning							
O vs Pe	500	500	Documents from different sub categories	Sheep	61	65	Web pages with different contents
O vs Pl	500	500		Biomedical	61	131	
Pe vs Pl	500	500		Goats	61	70	
Sample Selection Bias Correction				Uncertainty Mining			
Ionosphere	34	317	Samples with feature bias	ColonTumor	30	35	Training and Test set
Diabetes	80	688		CNS	30	30	
Haberman	30	276		Leukemia	30	42	contain different Gaussian noises
Wdbc	30	539		ProstateCancer	30	106	

4 Experiment

MarginGraph is evaluated in three different scenarios of universal learning: transfer learning, sample selection bias correction and uncertainty mining. For each scenario, several frequently used data collections are selected. Results show that MarginGraph can reduce the domain adaptation risk significantly. Both maximal margin principal and ensemble play an important role in its performance.

4.1 Experiments Setting

The proposed approach is compared against different state-of-art algorithms specifically designed for each scenario (transfer learning, etc), as well as, the original graph transduction algorithm [7]. As a fair comparison, the original graph transduction is implemented in two different ways. The first uses the entire training data set, and the second one chooses a randomly selected sample whose size is equal to the number of examples chosen by MarginGraph. For naming convenience, the first one is called Graph, and the second as RandGraph. In transfer learning, CDSC [2] is selected as the comparative method. Its main idea is to find a mapping space which optimizes over consistency measure between the out-domain supervision and in-domain intrinsic structure. In sample selection bias correction, two approaches BRSD-BK and BRSD-DB [1] are adopted. Both methods correct the bias through structural discovery and re-balancing using unlabeled data. For both CDSC and BRSD, we use Graph as their base classifiers. For the proposed method, the number of iterations is chosen to be $n_t * 2$, where n_t is the number of labeled samples of the minority class. In addition, we set $\epsilon = 0.1$ for the stop criterion. The analysis of parameters can be found in Section 4.3. Both accuracy and AUC are reported as the evaluation metrics. Due to the randomness of obtained data set on sample selection bias and uncertainty mining tasks, the results below are averaged over 10 runs. The algorithm implementations are based on Weka [9].

4.2 Experiments Procedure

The descriptions, pre-processing procedures of different scenarios and the experiment results are presented below.

Table 3. Accuracy and AUC in Transfer Learning Data Set (%)

Methods	O vs Pe	O vs Pl	Pe vs Pl	Biomedical	Goats	Sheep
	Reuters-21578			SyskillWebert		
	Accuracy					
Graph	0.682	0.672	0.698	0.687	0.586	0.708
RandGraph	0.632	0.648	0.664	0.656	0.529	0.677
CDSC	0.704	0.720	0.736	0.610	0.586	0.677
MarginGraph	0.752	0.810	0.780	0.740	0.743	0.815
	AUC					
Graph	0.773	0.723	0.677	0.600	0.562	0.720
RandGraph	0.719	0.715	0.652	0.561	0.510	0.587
CDSC	0.783	0.799	0.682	0.582	0.523	0.536
MarginGraph	0.841	0.837	0.741	0.725	0.681	0.678

Transfer Learning. Two data collections from two different domains are employed. Among them, Reuters-21578 [10] is the primary benchmark of text categorization, and SyskillWebert [10] is the standard data set used to test web page ratings. Reuters-21578 collection is formed by different news with a hierarchical structure where it contains five top categories of news wire articles, and each main category contains several sub categories. Three top categories, “orgs”, “people” and “places” are selected in our study. All of the subcategories from each category are divided into two parts, one in-domain and one out-domain. They have different distributions and are approximately equal in size. Details are summarized in Table 2. The learning objective aims to classify articles into top categories. SyskillWebert collection is formed by the HTML source of web pages plus the a user rating (“hot” or “not hot”) on those web pages. It contains four separate subjects belonging to different topics. In the experiment, we randomly reserve “Bands-recording artists” as out-domain and the other three as in-domain data (Table 2). The learning task is to predict the user’s preferences for the given web pages.

Table 3 presents accuracy and AUC for each domain transfer data set, given by Graph, RandGraph, CDSC and the proposed algorithm MarginGraph. It is evident that MarginGraph achieves the best performance (accuracy and AUC) in 11 out of 12 runs. Due to “labeling unsmooth problem” caused by distribution difference between the training and test data, both Graph and RandGraph fail to make correct predictions most of the time. To be specific, they achieve accuracies just no more than 71% on the Reuters and SyskillWebert collections. By considering information from both domains, CDSC gets better performance in that it boosts the accuracy by 2% to 8%. However, the proposed approach, MarginGraph, has the highest accuracy in all data set, and highest AUC in 5 out of 6 data set. We notice that MarginGraph performs better than both Graph and RandGraph at least 7% in accuracy on most data set. Specifically, it achieves as high as 16% better than these baseline methods on the Goat data set. The better performance of MarginGraph than CDSC can be ascribed to both the maximal margin based sample selection and ensemble strategy. As analyzed, these criterions can give a low prediction risk. This, from the empirical perspective, provides justification to the analysis in Section 3.

Table 4. Accuracy and AUC in Sample Selection Bias Data Set (%)

Methods	Ionosphere	Diabetes	Haberman	Wdbc	Ionosphere	Diabetes	Haberman	Wdbc
	Accuracy				StDev			
Graph	0.642	0.602	0.649	0.892	0.028	0.034	0.074	0.001
RandGraph	0.590	0.601	0.586	0.890	0.070	0.068	0.002	0.001
BRSD-BK	0.699	0.643	0.631	0.890	0.038	0.043	0.004	0.002
BRSD-DB	0.649	0.624	0.627	0.887	0.010	0.018	0.059	0.887
MarginGraph	0.817	0.709	0.717	0.896	0.077	0.259	0.295	0.002
	AUC				StDev			
Graph	0.701	0.583	0.582	0.962	0.030	0.069	0.091	0.001
RandGraph	0.605	0.554	0.511	0.961	0.010	0.115	0.003	0.002
BRSD-BK	0.726	0.671	0.561	0.962	0.053	0.061	0.012	0.001
BRSD-DB	0.686	0.634	0.551	0.962	0.008	0.032	0.075	0.001
MarginGraph	0.654	0.650	0.592	0.963	0.259	0.063	0.276	0.001

Table 5. Accuracy and AUC in Uncertainty Mining Data Set (%)

Methods	CNS	ColonTumor	Leukemia	ProCancer	CNS	ColonTumor	Leukemia	ProCancer
	Accuracy				StDev			
Graph	0.647	0.813	0.928	0.762	0.078	0.032	0.042	0.040
RandGraph	0.566	0.838	0.916	0.721	0.461	0.85	0.88	0.741
MarginGraph	0.713	0.787	0.944	0.794	0.078	0.026	0.042	0.029
	AUC				StDev			
Graph	0.606	0.761	0.914	0.762	0.086	0.035	0.020	0.052
RandGraph	0.444	0.740	0.910	0.698	0.154	0.075	0.035	0.126
MarginGraph	0.640	0.792	0.930	0.782	0.036	0.063	0.015	0.031

Sample Selection Bias Correction. Four data sets from UCI Repository [10] are selected. “Haberman” aims to predict the survival of patients who had undergone surgery for breast cancer. “Ionosphere” is to detect which radar is “Good” or “Bad” based on the radar signals information. “Wdbc” contains digitized image with characteristics of the cell nuclei in a fine needle aspirate of breast masses. “Diabetes” records test information about diabetes of patients and the task is to figure out which patients have diabetes. To generate the sample selection bias data set, we first randomly select 50% of the features, and then we sort the data set according to each of the selected features (dictionary sort for categorical features and numerical sort for continuous). Then, we attain top instances from every sorted list as training set, with “#Training” instances, and use the remain examples as test set.

Table 4 summarizes the accuracy, AUC as well as their standard deviations of baselines: Graph, RandGraph, BRSD-BK, BRSD-DB, and the proposed algorithm MarginGraph on four biased data sets. Clearly, MarginGraph achieves higher accuracies (from 5% to 15%) for each data set than the corresponding baseline approaches. For example, on the Haberman data set, the accuracy has been improved from no more than 65% to 71%. In AUC, MarginGraph wins Graph at 3 rounds and just loses at 1 comparison. Importantly, MarginGraph outperforms BRSD-BK and BRSD-DB consistently, specifically designed for bias correction, in accuracy. Moreover, MarginGraph performs compatibly with them in AUC, 2 wins and 2 loses. These results demonstrate that MarginGraph also has good generalization in sample selection bias task. This is attributed to the

ensemble strategy, which makes the classifier more robust to bias [6], and the maximal margin strategy guarantees error bound of MarginGraph is low when the test and training distributions are related as shown in Section 3.

Uncertainty Mining. Four biomedical and gene expression data sets are selected from Kent Ridge Biomedical Repository [11] for this comparison. “Colon-Tumor” contains 62 samples collected from colon-cancer patients. “Central Nervous System(CNS)” aims to detect which patients are survivors or failures. “Leukemia” is the data set about classifying subtypes of pediatric acute lymphoblastic leukemia. “ProstateCancer” is about the tumor versus normal classification. Each of them is a typical example of high dimensional, low sample size (HDLSS) problem. To generate the uncertainty, we randomly partition the data set into training and test parts first. Then, we form two f_m -dimension Gaussian noises with different means and variances where f_m is the dimensions of the data set, and add them into two parts of data set separately. Thus, we obtain four uncertain data sets where the training set and test set contain different noises (Table 2).

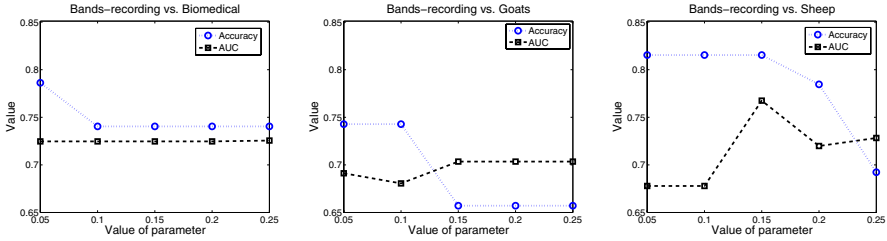
The accuracy and AUC of the proposed method, and original label propagation on uncertainty mining can be found in Table 5. Among the 4 tasks, MarginGraph outperforms the baseline by 3-1 in accuracy and 4-0 in AUC with smaller standard deviations. In particular, on the CNS data set, MarginGraph performs better than the baseline by as much as 6% in accuracy and 4% in AUC. The performance improvement results from the adaptation of maximal margin strategy that makes the decision boundary go through the low density region. In addition, the averaging ensemble strategy increases the resilience of the base classifier for feature noises [12].

4.3 Parameter Analysis

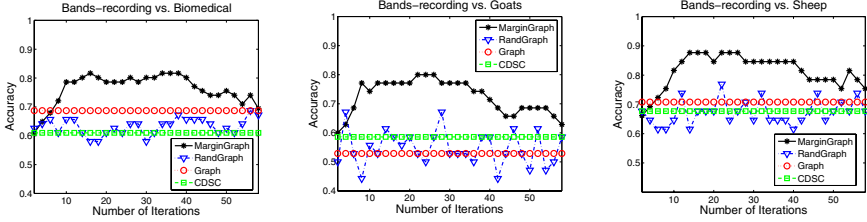
Three extended experiments were conducted on the SyskillWebert collection to test the parameter sensitivity and the relationship between the unlabeled data margin and prediction accuracy. As discussed in Section 2, there are two parameters to run MarginGraph, the parameter ϵ for the stopping criterion, as well as $maxIt$, the maximal number of iterations.

Figure 2(a) shows the different AUC and accuracy vs. different values of ϵ . We observe that AUC is insensitive to the value of ϵ , but the accuracy drops down when ϵ becomes large. The reason is that ϵ determines the threshold of average margin. Clearly, if the margin is too small, the prediction risk will increase and the accuracy will decrease, as analyzed in Section 3.

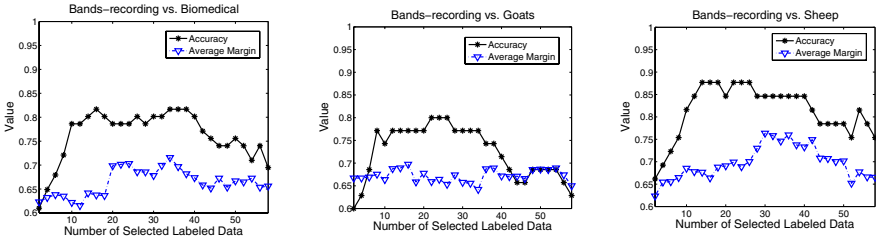
Figure 2(b) illustrates the relationship between the number of iterations and prediction accuracies. Because both Graph and CDSC use the entire training data, their accuracy results do not change. We observe that the accuracy of MarginGraph increases but then drops when the number of iterations is more than 40. That is because the number of labeled data is too few to build an effective classifier at the beginning. When useful labeled data are selected enough, adding more labeled data will reduce the unlabeled data margin and create unsmooth problems against the test data. This can be observed from Figure 2(c)



(a) Accuracy and AUC over ϵ



(b) Accuracy over Number of Iterations



(c) Accuracy over Average Margin

Fig. 2. Parameter Analysis

that the margin also drops down when we select more than 40 labeled points. However, we still see that MarginGraph achieves the best overall performance even with a large number of iterations.

Figure 2(c) shows that the average margin and accuracy have similar trend with a function of the number of selected examples. This implies that the same set of sampled training examples that reduce the unlabeled data margin can also reduce the prediction accuracy, and vice versa. Thus, the unlabeled data margin is a good criterion to select samples for graph transduction in universal learning.

4.4 Margin Analysis

As shown in Section 3, maximal margin and ensemble are two main contributing factors to guarantee a low error bound. As follows, we perform an additional experiment to study each factor. For comparison, we adopt three other approaches, similar with MarginGraph but with slight modifications. The first “MarginBase”

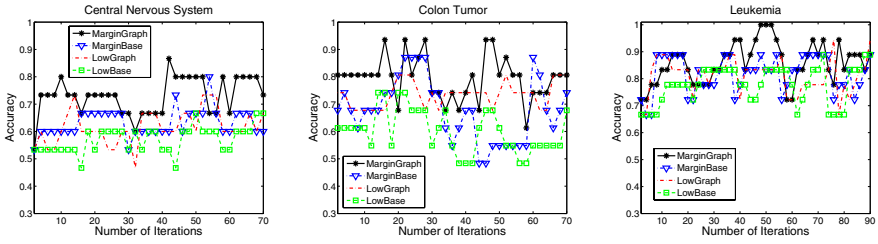


Fig. 3. Margin Analysis

is the base classifier of MarginGraph in each iteration. The second is a “minimal margin classifier” which selects samples for building a classifier with minimal unlabeled data margin, called “LowBase”. The third one is the averaging ensemble of LowBase, called “LowGraph”.

Figure 3 shows the relationship between the number of iterations and prediction accuracies on three uncertain data set. It is obvious that MarginGraph outperforms others in most cases. Especially, MarginGraph achieves the highest accuracy during iterations in all three data sets. In addition, LowGraph outperforms LowBase and MarginBase performs better than LowBase. That means maximal margin is better than minimal margin and ensemble is better than any single classifiers.

5 Related Works

Many solutions for transfer learning, sample selection bias correction and uncertainty mining have been proposed previously, such as but not limited to [2,13,1,14,15,16]. Among them, [2] designs a novel cost function from normalized cut that test data supervision is regularized by training data structural constraints. [13] learns a low-dimensional latent feature space where the distributions between the training data and the test data are the same or close to each other. [1] proposes to discover the natural structure of the test distribution, by which different types of sample selection biases can be evidently observed and then be reduced. [14] proposes a direct importance estimation method for sample selection bias that does not involve density estimation. [15] discusses a new method for handling error-prone and missing data with the use of density based approaches to data mining. [16] aims at minimizing the worst-case value of a loss function, over all possible realizations of the uncertainty data within given interval bounds. However, these methods are designed for each specific scenario, and are not generalized over universal learning where training and testing data are related but different. In our work, we do not distinguish transfer learning, sample selection bias, uncertainty mining and the alike. There are several significant extensions to graph transduction. For example, recently [7] introduces a semi-supervised learning framework based on Gaussian random fields and harmonic functions. Previously, [8] combines the graph transduction and active

learning, similar to the proposed work. However, the algorithm in this paper do not require any expert to label examples. Most recently, [4] introduces a new label propagation algorithm that can reliably minimize a cost function over both a function on the graph and a binary label matrix.

6 Conclusion

We have introduced the “universal learning” framework to cover different formulations where the training and test set are drawn from related but different distributions. Explicit scenarios include transfer learning, sample selection bias and uncertainty mining. We have proposed an adaptive graph transduction method using unlabeled data maximum margin principle to solve universal learning tasks. Unlike prior work, the proposed framework implicitly encompasses all three problem definitions and the alike. The same proposed solution can address different scenarios. The maximum margin graph transduction works as a joint optimization to maximize the unlabeled data margin and minimize the harmonic function over unlabeled data in the same time. It is an iterative strategy that removes the bias of any single graph. Formal analysis shows that the maximum margin based sample selection strategy has good generality over testing data with related but different distribution.

Empirical studies demonstrate that with different problem formulations in universal learning, the proposed approach significantly improves the original graph transduction. For transfer learning, it outperforms the original graph transduction by as much as 16% in accuracy and 12% in AUC. For sample selection bias correction, it achieves around 10% higher in accuracy in most cases. For uncertainty mining, its performance is the highest in 7 out of 8 comparisons. Most importantly, it consistently outperforms, by as much as much 10% in accuracy, than state-of-art approaches specifically designed for transfer learning and bias correction. These base line methods include CDSC [2] for transfer learning, and BRSD-BK and BRSD-DB [1] for sample selection bias correction. The main claims are that (1) universal learning framework provides a general formulation to cover and study various real-world application scenarios where training and testing data do not follow the same distribution, and (2) unlike previously proposed methods that cover only one scenario, the proposed adaptive graph transduction provides a more accurate solution to encompass all distribution differences under universal learning, and this provides utility and ease of use.

References

1. Ren, J., Shi, X., Fan, W., Yu, P.S.: Type-independent correction of sample selection bias via structural discovery and re-balancing. In: Proceedings of the Eighth SIAM International Conference on Data Mining, SDM 2008, pp. 565–576. SIAM, Philadelphia (2008)
2. Ling, X., Dai, W., Xue, G.R., Yang, Q., Yu, Y.: Spectral domain-transfer learning. In: KDD 2008: Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 488–496. ACM, New York (2008)

3. Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., Wortman, J.: Learning bounds for domain adaptation. In: Platt, J., Koller, D., Singer, Y., Roweis, S. (eds.) *Advances in Neural Information Processing Systems 20*, pp. 129–136. MIT Press, Cambridge (2008)
4. Wang, J., Jebara, T., Chang, S.F.: Graph transduction via alternating minimization. In: *ICML 2008: Proceedings of the 25th international conference on Machine learning*, pp. 1144–1151. ACM, New York (2008)
5. Amini, M., Laviolette, F., Usunier, N.: A transductive bound for the voted classifier with an application to semi-supervised learning. In: Koller, D., Schuurmans, D., Bengio, Y., Bottou, L. (eds.) *Advances in Neural Information Processing Systems 21* (2009)
6. Fan, W., Davidson, I.: On sample selection bias and its efficient correction via model averaging and unlabeled examples. In: *Proceedings of the Seventh SIAM International Conference on Data Mining, SDM 2007, Minneapolis, Minnesota*. SIAM, Philadelphia (2007)
7. Zhu, X.: *Semi-supervised learning with graphs*. PhD thesis, Carnegie Mellon University, Pittsburgh, PA, USA (2005)
8. Zhu, X., Lafferty, J.: Combining active learning and semi-supervised learning using gaussian fields and harmonic functions. In: *ICML 2003 workshop on The Continuum from Labeled to Unlabeled Data in Machine Learning and Data Mining*, pp. 58–65 (2003)
9. Witten, I.H., Frank, E.: *Data Mining: Practical Machine Learning Tools and Techniques*, 2nd edn. Morgan Kaufmann Series in Data Management Systems. Morgan Kaufmann, San Francisco (2005)
10. Asuncion, A., Newman, D.J.: UCI machine learning repository (2007), <http://www.ics.uci.edu/mllearn/MLRepository.html>
11. Liu, H., Li, J.: Kent ridge bio-medical data set repository (2005), <http://leo.ugr.es/elvira/DBCRepository/index.html>
12. Melville, P., Shah, N., Mihalkova, L., Mooney, R.J.: Experiments on ensembles with missing and noisy data. In: Roli, F., Kittler, J., Windeatt, T. (eds.) *MCS 2004*. LNCS, vol. 3077, pp. 293–302. Springer, Heidelberg (2004)
13. Pan, S.J., Kwok, J.T., Yang, Q.: Transfer learning via dimensionality reduction. In: *Proceedings of the Twenty-Third AAAI Conference on Artificial Intelligence, AAAI 2008, Chicago, Illinois, USA, July 13-17*, pp. 677–682 (2008)
14. Sugiyama, M., Nakajima, S., Kashima, H., Buenau, P.V., Kawanabe, M.: Direct importance estimation with model selection and its application to covariate shift adaptation. In: Platt, J., Koller, D., Singer, Y., Roweis, S. (eds.) *Advances in Neural Information Processing Systems 20*, pp. 1433–1440. MIT Press, Cambridge (2008)
15. Aggarwal, C.C.: On density based transforms for uncertain data mining. In: *Proceedings of the 23rd International Conference on Data Engineering, ICDE 2007, The Marmara Hotel, Istanbul, Turkey*, pp. 866–875. IEEE, Los Alamitos (2007)
16. El Ghaoui, L., Lanckriet, G.R.G., Natsoulis, G.: Robust classification with interval data. Technical Report UCB/CSD-03-1279, EECS Department, University of California, Berkeley (October 2003)