Chinese Science Bulletin

© 2009 SCIENCE IN CHINA PRESS

Springer

Inductive transfer learning for unlabeled target-domain via hybrid regularization

ZHUANG FuZhen[1,3†], LUO Ping[2], HE Qing[1] & SHI ZhongZhi[1]

[1] Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China;
[2] Hewlett Packard Labs China, Beijing 100084, China;
[3] Graduate University of Chinese Academy of Sciences, Beijing 100190, China.

Recent years have witnessed an increasing interest in transfer learning. This paper deals with the classification problem that the *target-domain* with a different distribution from the *source-domain* is totally unlabeled, and aims to build an inductive model for unseen data. Firstly, we analyze the problem of class ratio drift in the previous work of transductive transfer learning, and propose to use a normalization method to move towards the desired class ratio. Furthermore, we develop a hybrid regularization framework for inductive transfer learning. It considers three factors, including the distribution geometry of the target-domain by *manifold regularization*, the entropy value of prediction probability by *entropy regularization*, and the class prior by *expectation regularization*. This framework is used to adapt the inductive model learnt from the source-domain to the target-domain. Finally, the experiments on the real-world text data show the effectiveness of our inductive method of transfer learning. Meanwhile, it can handle unseen test points.

Transfer Learning, Inductive Learning, Transductive Learning, Hybrid Regularization

Classification plays a key role in intelligent information processing, including the analysis of Web pages, images, videos and so on. Traditional classification has the assumption that the labeled training and unlabeled test data are drawn from the same distribution, and thus might fail to deal with the situation when the new unlabeled data are obtained from fast evolving, related but different information sources. Under this situation, the labeled data $D_s = \{\mathrm{x}_i^{(s)}, y_i^{(s)}\}|_{i=1}^{n_s}$ are under one distribution in one domain known as source-domain, while the unlabeled data $D_t = \{\mathrm{x}_i^{(t)}\}|_{i=1}^{n_t}$ are under a related but different domain called target-domain. This distribution mismatch between the source-domain and target-domain attracts the research of transfer learning[1–11].

So far most of the efforts in transfer learning[1,2,5–7] have been invested in the problems that there are some labeled data from the target-domain. In this paradigm, the labeled target-domain data are used to change the weight of the labeled source-domain data or adapt the model trained from the source-domain. Several advances were recently achieved, like Migratory-Logit[5], boosting based transfer learning[1], and cross-domain adaptive SVM[6]. Wu et al.[7] presented a method that explores a second, auxiliary data drawn from distribution mismatch training data to improve SVM performance on test data.

On the other hand, to further remove the human efforts of collecting labeled target-domain examples, transfer learning is required to deal with the totally unlabeled target-domain. For this problem, Dai et al.[3] proposed a co-clustering based classification method (CoCC), in which the class labels are transferred through

Received July 08, 2008; accepted November 03, 2008
doi: 10.1007/s11434-009-0171-x

† Corresponding author (email: zhuangfz@ics.ict.ac.cn)

Supported by the National Science Foundation of China (Grant Nos. 60435010, 60675010), National High Technology Research and Development of China (Grant Nos. 2006AA01Z128, 2007AA01Z132), National Basic Research Priorities Programme (Grant No. 2007CB311004) and National Science and Technology Support Plan (Grant No. 2006BAC08B06)

Citation: Zhuang F Z, Luo P, He Q, et al. Inductive transfer learning for unlabeled target-domain via hybrid regularization. Chinese Sci Bull, 2009, 54: 2470−2478, doi: 10.1007/s11434-009-0171-x

the bridge of co-clustering. Xing et al.[4] introduced a bridged-refinement method for transfer learning, which corrects the labels predicted by the model of the source-domain via bridged refinement process and obtains the labels with high accuracy. It is clear that bridged-refinement method performs in a transductive setting that predicts only for observed instances, rather than generating a model for new test data. A brute force approach is to incorporate the new test points and restart the bridged refinement process. However, this is very inefficient.

In this paper, we first analyze the bridged-refinement method of transductive transfer learning, and find that the class ratio changes during the iterating process. This problem of class ratio drift greatly affects the performance of transductive transfer learning. However, in some applications the desired class ratio (the number of samples in each class divides the total number in data set) can be obtained by domain knowledge[16]. For example, one might estimate that about 20% of the pages crawled on the Internet belong to news. Thus, we can leverage this class ratio prior to avoid the drift of class ratio.

Furthermore, we propose an inductive algorithm of transfer learning for totally unlabeled target-domain. The characteristic which makes our algorithm unique is that it can not only deal with totally unlabeled data from the target-domain, but also output the classifier for further unseen instances. The proposed algorithm is composed of two phases. In the first phase, we learn a classifier h_s in the source-domain, which represents the learnt source-domain knowledge. In the second phase, a hybrid regularization framework adapts h_s to the target-domain by leveraging the inherent structure of the unlabeled data. Specifically, the hybrid regularization framework considers three principles, including manifold regularization[14], entropy regularization[15], and expectation regularization[16] (These three regularization criterions will be detailed later). Our algorithm framework uses the classifier h_s as the initial point for optimization, and will converge to a local optimum point by this combination framework of regularization. We implement this regularization framework by Logistic Regression[19], and the experimental results on text classification show that our inductive method outperforms the previous transductive algorithms for totally unlabeled target-domain.

1 Transductive transfer learning

1.1 Bridged refinement

This section briefly describes the bridged refinement[4] method of transductive transfer learning. The key component of this two-phase bridged refinement process is an iterating algorithm.

Let $T \in i_+^{n\times|c|}$ (where i_+ denotes the set of nonnegative real numbers, |*c*| denotes the number of data classes, |*n*| denotes the number of samples) be a probability matrix, and T_{ij} be the probability that the instance *i* belongs to the class *j*. Let *M* be the adjacence matrix of the samples,

$$M_{ij} = \begin{cases} 0 & \text{if } x_j \text{ is not the } k\text{-nearest neighbor of } x_i, \\ \dfrac{1}{K} & \text{if } x_j \text{ is the } k\text{-nearest neighbor of } x_i. \end{cases} \tag{1}$$

The similarity degree between x_i and x_j can be measured by

$$cos(x_i, x_j) = \frac{x_i^T x_j}{\| x_i \| \cdot \| x_j \|}. \tag{2}$$

In each round of the iteration, each sample x absorbs a fraction of label information from its neighborhood, and retains some label information of its initial state. The iterated equation is

$$T_{i\cdot}^{m+1} = \alpha \sum_{j:x_j \in N_i} \frac{T_{j\cdot}^{m}}{K} + (1-\alpha) T_{j\cdot}^{0}, \tag{3}$$

where $T_{i\cdot}^{m+1}$ denotes the *i*-th row of the matrix T in the (*m*+1)-th iteration, $N_i$is the set of *k*-nearest neighbors of the *i*-th instance, and $0<\alpha<1$ is the trade-off factor. The above equation can also be writte *n* in the following form of matrix computation,

$$T^{m+1} = \alpha M T^{m} + (1-\alpha) T^{0}. \tag{4}$$

It can be proved that the matrix T will converge to

$$T^{*} = (1-\alpha)(1-\alpha M)^{-1} T^{0}. \tag{5}$$

The details of this theoretical analysis can be found in Wang et al.[17].

Actually, this iteration algorithm contains two inputs, *M* and T^0. For the first phase of bridged refinement, *M* is the adjacent matrix of the data in both the source- domain and target-domain, and T^0contains true class labels of the source-domain data and the prediction results of the target-domain data by the classifier learnt from the source-domain. For the second phase, *M* is the adjacent matrix of the target-domain data only, and $T^0$is set to the prediction result from the first phase on the target domain data. Therefore, the main idea of the bridged refinement method is that it leverages the manifold structure firstly in the data $D = D_s \cup D_t$ and then in the target domain D_t.

1.2 The Enhancement of bridged refinement

In this paper, we investigate the property of T during the refinement process. It can be proved that the sum of each row of T maintains constant. That is $\sum_{j=1}^{|c|} T_{ij} = 1$. However, the value $s = \sum_{i=1}^{n} T_{ij}$, indicating the number of samples belonging to class c_j, is changing during the iteration. However, we always expect s reaches the actual number of samples belonging to the corresponding class.

Considering two-class classification problem,

$$r^p = \frac{\sum_{i=1}^{n} T_{i1}}{\| T \|} \tag{6}$$

and

$$r^n = \frac{\sum_{i=1}^{n} T_{i2}}{\| T \|} \tag{7}$$

stand for the ratios of samples in positive and negative class respectively, where

$$\| T \| = \sum_{j=1}^{2} \sum_{i=1}^{n} T_{ij} \tag{8}$$

Appendix A and n is the number of samples. As shown in Figure 1, the value of r^p ($r^n = 1 - r^p$) changes during the iteration process of Problem 10 (The description of this problem is given in). Eventually the value of r^p converges to 0.571, which is greatly different from the actual value 0.748. Thus, it might lead to a poor performance.

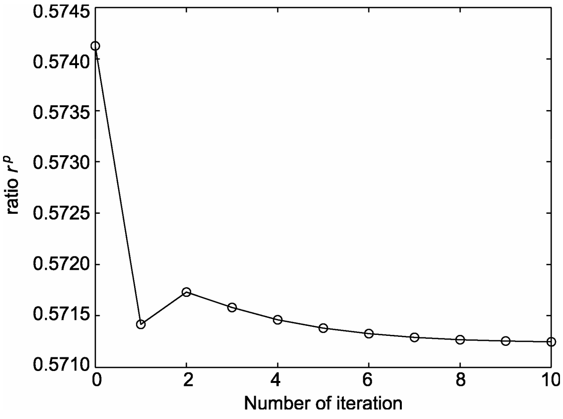

Figure 1 Change of class ratio for Problem 10 during the iterations.

Actually, Xing et al.[4] normalized this class ratio to 1:1, which can perform well on the balanced data. However, it is not proper especially for the imbalanced data. Considering the true class ratio provided, we propose the algorithm of bridged refinement with class ratio normalization in Algorithm 1 (In Appendix B). In each round of this algorithm we normalize the class ratio to the true class ratio provided. Through this technique the accuracy performance of this problem increases 11.5% from 80.47%, which will be further shown in the experimental section.

2 Inductive transfer learning via hybrid regularization

The transductive transfer learning in Section 1 cannot generate a classifier for new coming test data. However, in many applications we need a classifier to directly predict unseen data. In this section we propose a two-phase method for inductive transfer learning via hybrid regularization. In the following we only consider two-class classification problem, but the algorithm framework can be extended to multi-class setting naturally. Since we implement this regularization framework by Logistic Regression[19] we first give a brief description of this classification model.

2.1 Logistic regression

Logistic regression[19] is an approach to learn functions of $P(Y/X)$ in the case where Y is discrete-valued, and X is any vector containing discrete or continuous random variables. Logistic regression assumes a parametric form for the distribution $P(Y/X)$, then directly estimates its parameters from the training data. The parametric model assumed by logistic regression in the case where Y is Boolean is

$$P(y = \pm 1 \mid x; w) = \sigma(y w^T x) = \frac{1}{1 + \exp(-y w^T x)}, \tag{9}$$

where w is the parameter of the model. Under the principle of Maximum A-Posteriori (MAP), w is estimated under the Laplacian prior. Given a data set $D = \{\mathrm{x}_i, y_i\}|_{i=1}^{N}$, we want to find the parameter w which maximizes:

$$\sum_{i=1}^{N} \log \frac{1}{1 + \exp(-y_i w^T x_i)} - \frac{\lambda}{2} w^T w. \tag{10}$$

This criterion is a concave function of w, so that the global solution can be obtained by methods of the nonlinear numerical optimization. After w is estimated, eq.

(9) can be used to compute the probabilities of an instance belonging to the positive and negative class.

2.2 Regularization principles

The proposed hybrid framework includes three regularization principles: manifold regularization[14], entropy regularization[15], and expectation regularization[16]. In previous work these principles were proposed to exploit the inherent structure in labeled and unlabeled data for semi-supervised learning. However, in this paper we utilize them in transfer learning, which means that these principles effect on the target domain $D_t = \{\mathrm{x}_i^{(t)}\}|_{i=1}^{n_t}$, where n_t is the number of samples. Also, we adopt the function σ in Eq. (9) to express the conditional probability $P\,(\,y=1|x)$, provided the model is w.

Manifold regularization[14]. Belkin et al. proposed the manifold regularization that efficiently exploits manifold structure in labeled and unlabeled data for semi-supervised learning. It requires that the label of an instance be similar to the labels of its neighbors. When applying it to the target domain for transfer learning, this regularization is to minimize the following formula:

$$g_m(w) = \frac{1}{n_t}\sum_{i=1}^{n_t}\left[\frac{1}{K}\sum_{k=1}^{K}\sigma(w^T x_{i_k}) - \sigma(w^T x_i)\right]^2, \qquad (11)$$

where K is the number of sample x_i's neighbors, x_{ik} is the k-th ($1 \leqslant k \leqslant K$) neighbor of sample x_i. Any distance metric can be used to compute the neighboring relationship among the instances.

Entropy regularization[15]. Grandvalet et al. proposed the entropy regularization to minimize the entropy of the probability vector $p_i = (p_{i1}, l, p_{i|c|})$ on any instance x_i, where p_{ij} is the probability that x_i belongs to class c_j and $|c|$ is the number of classes. This regularization is based on the fact that any instance belongs to only one class, which results in the entropy minimum on its true probability vector. For binary classification, the entropy regularization is equivalent to minimize the following formula:

$$g_c(w) = -\frac{1}{n_t}\sum_{i=1}^{n_t}\left[\sigma(w^T x_i) - \frac{1}{2}\right]^2. \qquad (12)$$

Expectation regularization[16]. Mann et al. proposed the expectation regularization to force the prediction results to match some prior knowledge (such as class ratio), which might be approximately estimated in advance. Specifically, it requires that the class ratio in prediction results be similar to the true class ratio provided. It can be expressed in minimizing the following formula:

$$g_e(w) = \frac{1}{n_t}\Big[\sum_{i=1}^{n_t}\sigma(w^T x_i) - r\cdot n_t\Big]^2, \qquad (13)$$

where r is the true ratio of positive samples.

2.3 Two-phase method for inductive transfer learning

Our method of inductive transfer learning includes two phases—training initial classifier and classifier refinement.

Phase one trains the initial classifier h_s. Let $D_s = \{x_i^{(s)}, y_i^{(s)}\}|_{i=1}^{n_s}$ be source-domain with fully labeled data, and then we use supervised-learning algorithm—Logistic Regression[19] to learn initial model h_s on D_s.

Phase two refines the initial classifier h_s by the hybrid regularization framework. Given the target- domain data D_t, we refine the model w by minimizing the following objective function f:

$$f(w) = w^T w + \alpha\cdot g_m + \beta\cdot g_c + \gamma\cdot g_e, \qquad (14)$$

where α, β, γ are trade-off parameters among these regularization principles, and g_m, g_c and g_e are expressed by eq. (11)—(13), respectively.

The difference between this regularization framework and that for semi-supervised learning is that we do not combine the factor of the log-likelihood of the labeled source-domain data into the regularization framework. The reason is that the labeled and unlabeled data in semi-supervised learning are from the same distribution, which makes them naturally share the same model. However, in transfer learning the labeled source-domain data and unlabeled target-domain data come from different distributions, thus we might not obtain a model that performs well on both source-domain and target-omain in one optimization phase. Therefore, we firstly consider the model trained in the labeled source-domain data, and then refine it only utilizing the unlabeled target-domain data.

To solve the optimization problem, we list the partial differential of g_m, g_c, g_e and f as follows:

$$\nabla_w g_m = \frac{2}{n_t}\cdot\sum_{i=1}^{n_t}\left(\frac{1}{K}\sum_{k=1}^{K}\sigma(w^T x_{i_k}) - \sigma(w^T x_i)\right)\cdot$$
$$\left(\frac{1}{K}\sum_{k=1}^{K}\sigma(w^T x_{i_k})(1-\sigma(w^T x_{i_k}))x_{i_k}\right.$$
$$\left. - \sigma(w^T x_i)(1-\sigma(w^T x_i))x_i\right). \qquad (15)$$

$$\nabla_w g_c = -\frac{2}{n_t} \cdot \sum_{i=1}^{n_t} \left(\sigma(w^T x_i) - \frac{1}{2} \right)$$

$$\sigma(w^T x_i)(1 - \sigma(w^T x_i)) x_i \,. \qquad (16)$$

$$\nabla_w g_e = \frac{2}{n_t} \left(\sum_{i=1}^{n_t} \sigma(w^T x_i) - r \cdot n_t \right)$$

$$\left(\sum_{i=1}^{n_t} \sigma(w^T x_i)(1 - \sigma(w^T x_i)) x_i \right). \qquad (17)$$

So the partial differential of objective function f is

$$\nabla_w f = 2 \cdot w + \alpha \cdot \nabla_w g_m + \beta \cdot \nabla_w g_c + \gamma \cdot \nabla_w g_e. \qquad (18)$$

Since the objective function f is neither convex nor concave, it is hard to obtain the global optimum solution for this problem. However, using the non-linear numerical optimization technique we can obtain the local optimum, starting from the initial model h_s. In this paper, we use conjugate gradient method to solve the optimization problem. The detail process of conjugate gradient method is shown in Algorithm 2 (In Appendix C). We use the function fminunc in Matlab to solve the single-variable optimization problem in Step 3 of this algorithm.

3 Experimental results

The experiments designed in this section are to evaluate the validity of the proposed algorithm.

3.1 Data preparation

We restructure the data collection 20 newsgroups[1)] to make it fit our problem setting. This data set has two level hierarchical structures. Suppose A and B are two top categories in this data set, and A_1, A_2 and B_1, B_2 are sub-level categories of A and B respectively. Now we form the source-domain and target-domain data in this way. Let $A \cdot A_1$, $B \cdot B_1$ be the positive and negative examples in the source-domain data respectively. Let $A \cdot A_2$, $B \cdot B_2$ be the positive and negative examples in the target-domain data respectively. Additionally, we intentionally put different numbers of positive and negative instances into the target-domain. Finally, we get 12 problems described in Table 1 (In Appendix C). This table also lists the true ratios r^p of positive instances in target-domain D_t. We use the model of $tf \cdot idf$ to express each document in the data set, and the threshold of Document Frequency with the value of 5 is used to select the features.

3.2 Performance comparison

The baseline algorithms for performance comparison with our method of Inductive transfer learning via Hybrid Regularization (IHR) include:

(1) Traditional classification algorithms: SVM[18] and Logistic Regression (LR)[19];

(2) Transductive transfer learning methods: Bridged Refinement[4](BR) and Bridged Refinement with class ratio Prior (PBR, Section 1.2 in this paper). Since these methods are based on the prediction input of a classifier, BR^{LR}, BR^{SVM}, PBR^{LR} and PBR^{SVM} denotes the methods of BR and PBR when the inputs from LR and SVM are adopted, respectively. In the experiments, we set the number of neighbors K = 70 for these transductive methods, and SVM with a linear kernel and all the other parameters in both SVM and Logistic Regression set by default[2)].

(3) Inductive transfer learning method: CoCC[3] (Co-Clustering based Classification algorithm)

(4) Semi-supervised learning methods: TSVM[20] (Transductive Support Vector Machines) and SGT[21] (Spectral Graph Transducer).

Comparison of IHR, LR, SVM, BR^{LR}, PBR^{LR}, BR^{SVM} and PBR^{SVM}. We compare the performance of these 7 algorithms on the 12 problems. The performances of all these algorithms on each problem are shown in Figure 2, and the average performances of these algorithms on the 12 problems are shown in Table 1. We also conduct t-test with 95% confidence to check whether the superiority of an algorithm over another one is statistically significant. From these results we get the following

Table 1 Average accuracy (%) of IHR, LR, SVM, BR^{LR}, PBR^{LR}, BR^{SVM} and PBR^{SVM} on all problems instances (α = 0.4, β = 15, γ = 0.12)

$\overline{\mathrm{LR}}$	$\overline{\mathrm{SVM}}$	$\overline{\mathrm{BR}^{\mathrm{LR}}}$	$\overline{\mathrm{BR}^{\mathrm{SVM}}}$	$\overline{\mathrm{PBR}^{\mathrm{LR}}}$	$\overline{\mathrm{PBR}^{\mathrm{SVM}}}$	$\overline{\mathrm{IHR}}$
77.92	79.81	91.46	90.59	94.57	94.27	96.31

1) http://people.csail.mit.edu/jrennie/20Newsgroups/

2) We also conduct the experiments to show that the parameter tuning of the base classifier (SVM or Logistic Regression) does not improve the performances of BR and PBR, and the default parameters of SVM (adopted in this paper) always achieve the best accuracy of BR and PBR. We omit the detailed experimental data due to the space limitation.

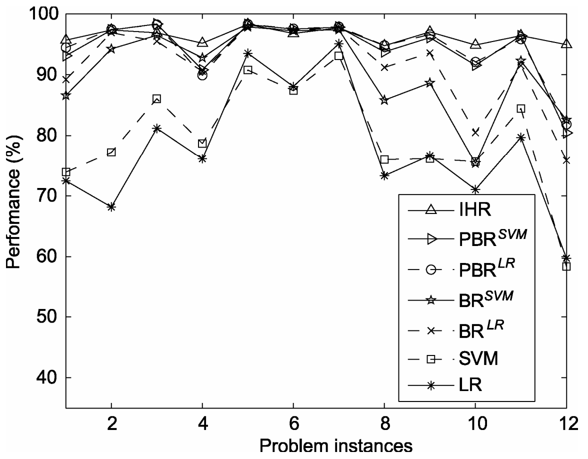

Figure 2 The Performance Comparison among IHR, LR, SVM, BR^{LR}, PBR^{LR}, BR^{SVM} and PBR^{SVM} on the 12 Problem Instances (α = 0.4, β = 15, γ = 0.12).

findings: 1) IHR significantly outperforms LR, SVM, BR^{LR}, and BR^{SVM}; 2) the performance differences among IHR, PBR^{LR} and PBR^{SVM} are not significantly clear. However, on average IHR outperforms PBR^{LR} and PBR^{SVM}, as shown in Table 1.

Comparison of IHR, CoCC, TSVM, SGT. We use the same data sets in Dai et al.[3] (the detailed data description can be found in Appendix A of Dai et al.[3]) to conduct this comparison, and the results are shown in Table 2. We can find that IHR outperforms CoCC, TSVM and SGT on every data set, which again shows the effectiveness of our algorithm.

Table 2 The Performance comparison (%) among TSVM, SGT, CoCC and IHR (α = 0.4, β = 15, γ = 0.12)

Data set	TSVM	SGT	CoCC	IHR
Res vs. talk	96	90.9	96.4	98.71
Rec vs. sci	93.8	93.8	94.5	97.15
Comp vs. talk	90.3	97.2	98	98.28
Comp vs. sci	81.7	72.1	87	96.65
Comp vs. rec	90.2	95.3	95.8	97.04
Sci vs. talk	89.2	91.7	94.6	96.15

3.3 Effects of different parameters on IHR

In our experiments, we also analyze the performances of IHR under different settings of the three parameters α, β and γ on the six problems. In these experiments we record the performances of IHR along the increase of a parameter while the other two parameters are fixed. The results are shown in Figure 3. The findings and analysis on these results are as follows:

(1) From Figure 3(a) we find that the performance of IHR is not sensitive to the parameter α in the range of [0, 1]. It also indicates that on some problems (e.g. Problems 5 and 6) the manifold regularization is useless. We further check the data characteristics of Problems 5 and 6 to analyze why this happens. We find that prediction results from the initial classifier h_s on these two problems satisfy with a high degree that the label of an instance is similar to the labels of its neighbors. Then, the

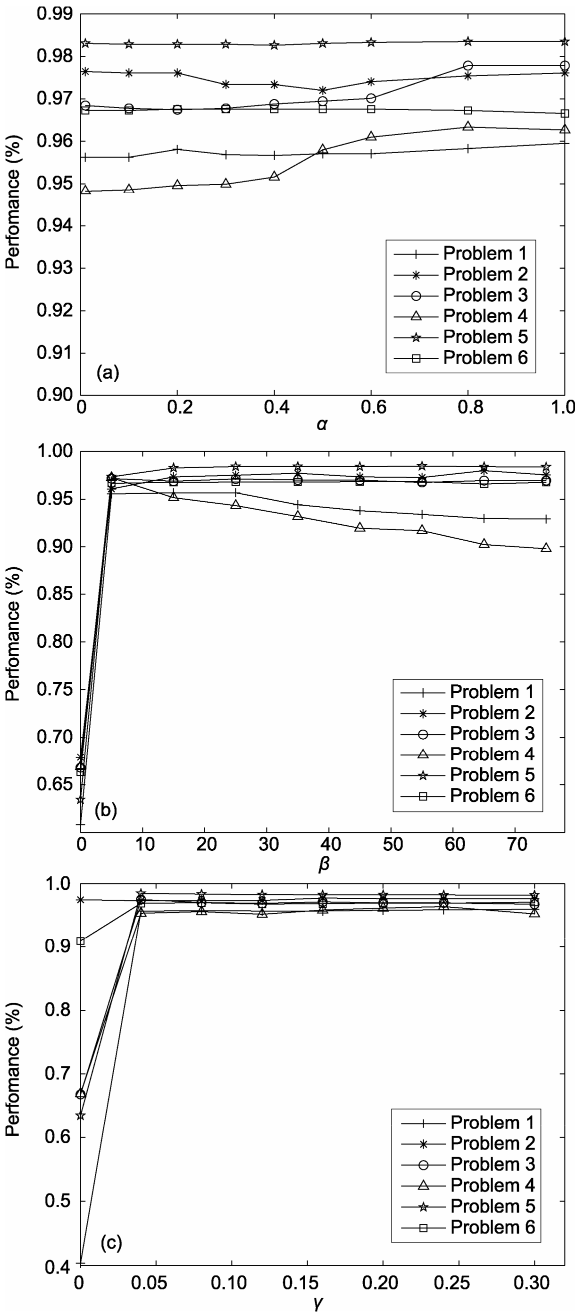

Figure 3 The effects of different parameters on IHR. (a) α vs. the Performance of IHR (β = 15, γ = 9.12); (b) β vs. the Performance of IHR (α = 0.4, γ = 0.12); (c) γ vs. the Performance of IHR (α = 0.4, β= 15).

manifold regularization does not affect the performance of IHR under this situation. Therefore, the effects of different parameters on IHR are affected by not only the hybrid regularization framework itself but also the data characteristics of the problems.

(2) From Figure 3(b) and Figure 3(c) we find that the parameters of β and γ greatly affect the performance of IHR. And the performance of IHR is stable when γ is bigger than 0.05 as shown in Figure 3(c).

To show the robustness of our algorithm IHR for the parameters, we relax the bounds of the parameters α, β and γ, rather than the specific values. In the experiments, we set the bounds of parameters $\alpha\in(0,1)$, $\beta\in(0,30)$ and $\gamma\in(0,5)$ after some preliminary experiments, and evaluate them on the 12 problems. We randomly sample m (here $m = 15$) combinations of parameters, and average the performances of each parameter setting on 12 problems. The results are shown in Table 3. We can find that the average performance is almost the same as the results shown in Section 3.2. Also we can find the performances are always good when the parameters are in the bounds, which further validates the effectiveness of algorithm IHR.

3.4 Inductive setting of algorithm IHR

To further validate that our algorithm IHR is an inductive learning algorithm, we also evaluate the inductive setting of our algorithm IHR on 12 problems in our paper. In details, we randomly sample (without replacement) ratio p of the data in the target-domain D_t to form a new data set D_t^1, and the left data in D_t form the other data set D_t^2. Then, the unlabeled data in D_t^1 is used for the model refinement in the training process, and the unlabeled data in D_t^2 is used to test the generalization ability of the model output by the training process. We also test the accuracy of this model on D_t^1. Additionally, we test the performance of the trained model under different values of p for each problem. The whole experimental results are recorded in Figure 4. From these results we have the findings:

(1) The accuracy performances of the trained model on both D_t^1 and D_t^2 are almost the same.

(2) The increase of the unlabeled data in D_t^1 used for training improves the generalization ability of the resultant model. When $p\geqslant 0.6$ the accuracy of the model on D_t^2 is greater than 90% for all the problems.

These results show that our algorithm IHR has good generalization ability on the unseen data when more than 60% of the unlabeled target-domain data are used in the training.

4 Related Work

In this section we will discuss some related work in the field of transfer learning. Transfer learning deals with the problem of distribution mismatch between training and test data. In general, previous work in this area can

Table 3 Parameter affection for performance (%) of Algorithm IHR

Sample ID	α	β	γ	Problem ID												Average
				1	2	3	4	5	6	7	8	9	10	11	12	
1	0.550	8.897	0.254	95.8	97.0	97.8	97.6	97.8	97.1	97.3	95.2	96.5	93.9	95.5	95.3	96.4
2	0.014	17.411	0.432	95.7	97.4	96.8	94.8	98.2	96.7	97.6	94.1	97.1	95.4	96.4	94.5	96.2
3	0.088	8.604	0.177	95.6	96.8	97.6	97.4	97.7	97.0	97.6	94.7	96.4	94.6	95.4	95.0	96.3
4	0.745	20.812	0.310	96.0	97.5	97.1	95.3	98.3	97.0	97.7	93.9	97.2	95.7	96.6	94.1	96.4
5	0.626	17.952	0.128	95.8	97.3	97.0	94.9	98.4	96.6	97.5	94.9	96.7	95.0	96.4	94.3	96.2
6	0.008	16.624	0.346	95.8	97.5	96.7	94.9	98.2	96.8	97.6	94.2	97.0	95.3	96.4	94.2	96.2
7	0.732	27.048	0.146	95.3	97.6	97.3	93.9	98.4	97.0	97.6	94.1	97.2	95.2	96.6	93.3	96.1
8	0.416	24.654	0.359	95.5	97.5	97.2	94.6	98.4	97.0	97.8	94.2	97.3	94.7	96.5	93.5	96.2
9	0.128	1.819	0.234	93.0	92.2	93.8	94.1	95.1	94.9	95.6	87.7	92.2	86.7	87.9	90.5	92.0
10	0.182	9.711	0.142	95.6	96.9	97.6	96.1	97.9	96.8	97.5	95.4	96.3	92.6	95.8	94.9	96.1
11	0.842	22.251	0.257	95.9	97.5	97.1	95.3	98.4	97.0	97.9	94.6	97.3	95.5	96.8	94.1	96.4
12	0.196	25.486	0.082	95.3	97.4	97.2	93.8	98.5	96.8	97.6	93.8	96.6	95.4	96.3	93.2	96.0
13	0.981	13.580	0.427	95.9	97.5	97.8	97.0	98.1	97.2	97.7	96.3	97.2	95.2	96.2	95.2	96.8
14	0.303	29.857	0.405	94.8	97.8	97.0	93.7	98.4	96.9	97.7	93.8	97.0	92.6	95.8	92.9	95.7
15	0.793	6.347	0.295	95.8	96.6	97.5	98.0	97.5	96.9	97.1	95.7	96.2	93.5	95.0	96.2	96.3
	0.4	15	12	95.7	97.3	96.9	95.2	98.3	96.8	97.7	94.7	97.0	94.9	96.4	94.9	96.3

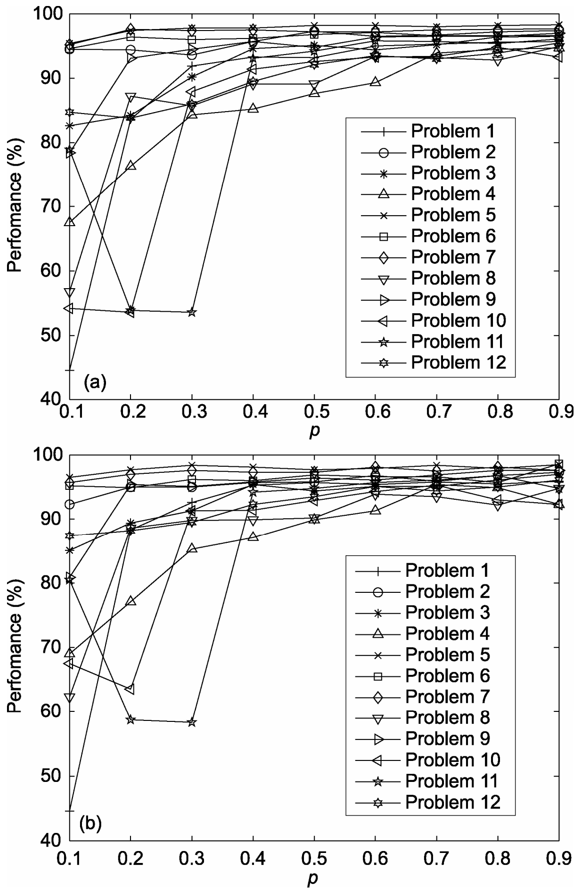

Figure 4 The Performance (%) of Algorithm IHR (α = 0.4, β = 15, γ = 0.12) on both D_t^1 and D_t^2. (a) Sampling ratio p vs. Performance on D_t^1; (a) Sampling ratio p vs. Performance on D_t^2.

be grouped into two categories. The first category of studies is under the assumption that there are some labeled data from the target domain. For instance, Liao et al.[5] estimated the degree of mismatch of each instance in the source domain with the whole target domain, and incorporated this information into logistic regression. Also, Dai et al.[1] extended boosting-based learning algorithms to transfer learning, in which the source-domain data with very different distribution were less weighted for data sampling. They also validated the algorithm by theoretical analysis using the Probability approximately correct (PAC) theory.

In the second category of transfer learning, the data from the target-domain are totally unlabeled. For this problem Ben-David et al.[9] analyzed the representations of domain data, and explored a promising model that not only minimized the generalization error on training data, but also minimized the difference between the source and target domains. Ling et al.[12] developed a new spectral classification algorithm that optimized an objective function to seek for the maximal consistency between the supervised information from the source domain and the intrinsic structure of the target domain. Mahmud et al.[13] studied transfer learning from the perspective of algorithmic information theory. They measured the relatedness between tasks, and then decided how much information to transfer and how to transfer. Xing et al.[4] proposed a transductive learning algorithm for transfer learning. This method is based on exploiting the geometry of marginal distribution on the data from the source-domain and the target-domain. However, this method cannot output a classifier for future unlabeled data.

5 Conclusions

In this paper we address the problem of inductive transfer learning, in which the target-domain is totally unlabeled. First, we analyze the transductive transfer learning method of bridged refinement, expose the problem of class ratio drift in this method, and find its limitation for imbalanced data. Then, the normalization method is proposed to incorporate the provided information of class ratio prior into this transductive method. Second, a new method of inductive transfer learning via hybrid regularization is proposed, which combines three regularization principles, including manifold regularization, entropy regularization, and expectation regularization. Compared with other transductive methods, this inductive algorithm can output a classifier for unseen test data. Additionally, the experiments on real-world data validate the effectiveness of this method. In the future, we will investigate when and how these regularization criterions work.

1 Dai W Y, Yang Q, Xue G R, et al. Boosting for Transfer Learning. In: Ghahramani Z B, eds. Proceeding of 24th International Conference on Machine Learning, 2007 Jun 20—24, Corvalis, Oregon. ACM, 2007. 193—200

2 Raina R, Battle A, Lee H, et al. Self-taught Learning: Transfer Learning from Unlabeled Data. In: Ghahramani Z B, eds. Proceeding of 24th International Conference on Machine Learning, 2007 Jun 20—24, Corvalis, Oregon. ACM, 2007. 759—766

3 Dai W Y, Xue G R, Yang Q, et al. Co-clustering based Classification for Out-of-domain Documents. In: Berkhin P, Caruana R, Wu X D, et al, eds. Proceeding of 13th ACM International Conference on Knowledge Discovery and Data Mining, 2007, Aug 12—15, San Jose, California. ACM, 2007. 210—219

4 Xing D K, Dai W Y, Xue G R, et al. Bridged Refinement for Transfer

Learning. In: Joost N K, Jacek K, Ramon L, et al, eds. Proceeding of 11th European Conference on Practice of Knowledge Discovery in Databases, 2007 Sep 17－21, Warsaw Poland. Springer, 2007. 324－335

5 Liao X J, Xue Y, Carin L, et al. Logistic Regression with an Auxiliary Data Source. In: Raedt L D, Wrobel S, eds. Proceeding of 22th International Conference on Machine Learning, 2007 Aug 7－11, Bonn, Germany. ACM, 2005. 505－512

6 Yang J, Yan R, Hauptmann A G, et al. Cross-domain Video Concept Detection Using Adaptive SVMs. In Rainer L, Anand R P, Alan H, et al, eds. Proceeding of 15th International Conference on Multimedia, 2007 Sep 24－29, Augsburg Germany. ACM, 2007. 188－197

7 Wu P C, Dietterich T G, et al. Improving SVM Accuracy by Training on Auxiliary Data Sources. In: Brodley C E, eds. Proceeding of 21th International Conference on Machine Learning, 2004 Jul 4－8, Banff, Alberta, Canada. ACM, 2004. 871－878

8 Mahmud M M H, Ray S, et al. Transfer Learning Using Kolmogorov Complexity: Basic Theory and Empirical Evaluations. Technical Report, UIUC-DCS-R-2007-2875, Department of Computer Science, University of Illinois at Urbana-Champaign. 2007

9 Ben-David S, Blitzer J, Crammer K, Pereira F, et al. Analysis of Representations for Domain Adaptation. In: Koller D, Singer Y, Platt J, et al, eds. Proceeding of Advances in Neural Information Processing Systems 20, MIT Press, Cambridge, MA, 2007(20): 137－144

10 Dai W Y, Xue G R, Yang Q, et al. Transferring Naive Bayes Classifiers for Text Classification. In: John C, Peyman F, Simon P, et al, eds. Proceeding of 22nd Conference on Artificial Intelligence, 2007 Jul 22－26, Vancouver, British Columbia. AAAI Press, 2007. 540－545

11 Samarth S, Sylvian R, et al. Cross Domain Knowledge Transfer Using Structured Representations. In: Proceeding of 21nd Conference on Artificial Intelligence, 2006 Jul 16－22, Boston, Massachusetts. AAAI Press, 2006

12 Ling X, Dai W Y, Xue G R, et al. Spectral Domain-Transfer Learning. In: Li Y, Liu B, Sunita S, et al, eds. Proceeding of 14th ACM International Conference on Knowledge Discovery and Data Mining, 2008, Aug 24－27, Las Vegas, Nevada. ACM, 2008. 488－496

13 Mahmud, M M H. On Universal Transfer Learning. In: Rocco M H, Servedio R A, Takimoto E, et al, eds. Proceeding of 18th International Conference on Algorithmic Learning Theory, 2007, Oct 1－4, Sendai, Japan. LNCS, 2007. 135－149

14 Belkin M, Niyogi P, Sindhwani V, et al. Manifold Regularization: A Geometric Framework for Learning from Labeled and Unlabeled Examples. Journal of Machine Learning Research, 2006: 2399－2434

15 Grandvalet Y, Bengio Y, et al. Semi-supervised Learning by Entropy Minimization. Proceeding of 19th Conference on Neural Information Processing Systems, 2005 Dec 5－8, Vancouver, British Columbia. MIT Press, 2005. 529－536

16 Mann G S, McCallum A, et al. Simple, Robust, Scalable Semi-supervised Learning via Expectation Regularization. In: Ghahramani Z B, eds. Proceeding of 24th International Conference on Machine Learning, 2007 Jun 20－24, Corvalis, Oregon. ACM, 2007. 593－600

17 Wang F, Zhang C S, et al. Label Propagation through Linear Neighborhoods. IEEE Transations on Knowledge and Data Engineering, 2008: 55－67

18 Joachims T. Making Large-scale SVM Learning Practical. In: Schölkopf B, Burges C, Smola A, et al, eds. Proceeding of Advances in Kernel Methods, 1999. MIT Press, Cambridge, 1999. 169－184

19 Davie H, Stanley L, et al. Applied Logistic Regression. Wiley, New York, 2000

20 Joachims T. Transductive Inference for Text Classification Using Support Vector Machines. In: Bratko I, Dzeroski S, eds. Proceeding of 16th International Conference on Machine Learning, 1999 Jun 27－30, Bled, Slovenia. ACM, 1999. 200－209

21 Joachims T. Transductive Learning via Spectral Graph Partitioning. In: Tom F, Nina M, eds. Proceeding of 20th International Conference on Machine Learning, 2003 Aug 21－24, Washington DC. ACM, 2003. 290－297