Joint Transfer and Batch-mode Active Learning

Rita Chattopadhyay RCHATTOP@ASU.EDU
Arizona State University, 699 S. Mill Ave, Tempe, AZ 85281, USA

Wei Fan DAVID.FANWEI@HUAWEI.COM
Huawei Noah's Ark Lab, Hong Kong Science Park, Shatin, Hong Kong

Ian Davidson DAVID.FANWEI@HUAWEI.COM
Department of Computer Science, University of California, Davis, CA 95616

Sethuraman Panchanathan PANCH@ASU.EDU
Arizona State University, 699 S. Mill Ave, Tempe, AZ 85281, USA

Jieping Ye JIEPING.YE@ASU.EDU
Arizona State University, 699 S. Mill Ave, Tempe, AZ 85281, USA

Abstract

Active learning and transfer learning are two different methodologies that address the common problem of insufficient labels. Transfer learning addresses this problem by using the knowledge gained from a related and already labeled data source, whereas active learning focuses on selecting a small set of informative samples for manual annotation. Recently, there has been much interest in developing frameworks that combine both transfer and active learning methodologies. A few such frameworks reported in literature perform transfer and active learning in two separate stages. In this work, we present an integrated framework that performs transfer and active learning simultaneously by solving a single convex optimization problem. The proposed framework computes the weights of source domain data and selects the samples from the target domain data simultaneously, by minimizing a common objective of reducing distribution difference between the data set consisting of re-weighted source and the queried target domain data and the set of unlabeled target domain data. Comprehensive experiments on real data demonstrate the superior performance of the proposed approach.

Proceedings of the 30^{th} International Conference on Machine Learning, Atlanta, Georgia, USA, 2013. JMLR: W&CP volume 28.

1. Introduction

Traditional supervised machine learning methods require sufficient labeled examples in order to construct accurate models. These methods also assume that labeled examples belong to the same underlying distribution as the test data; in other words, both training and test data are drawn i.i.d. from the same distribution. However, for real world applications, as in the case of medical diagnosis, video concept detection, sentiment analysis, document classification etc, one may not have sufficient or any labeled data, belonging to the same underlying distribution as the test data. Two machine learning methods namely *transfer learning* (Pan & Yang, 2009) and *active learning* (Settles, 2009), address this problem in two different ways. Transfer learning methods try to solve this problem by utilizing labeled data from related domains, which may be available in plenty, e.g., labeled data from another lab or a machine, labeled video clips belonging to other TV channels or positive and negative reviews for another product category. As a different solution, active learning methods focus on selecting a small set of most informative samples, for which they acquire labels from the domain experts. Hence, under conditions where we have sufficient labeled data from a related domain and a budget to get a fix number of target samples labeled by an expert, a combination of transfer and active learning would provide an effective strategy. Indeed, availability of additional labeled data from a related source domain would increase the reliability of the classifier used in active learning; at

the same time, availability of informative labeled data from target domain would enable efficient transfer of knowledge from source domains.

However, to the best of our knowledge, not much work has been reported in the literature along this direction. Existing work (Shi et al., 2008; Rai et al., 2010) does not integrate the transfer and active learning methodologies into a single consolidated framework. Instead, transfer and active learning are performed in two stages, which may cause redundancy or information overlap between the instances selected from the source and target domain data (Figure 1). Besides, the transfer learning or domain adaptation is just performed initially once and is not dynamically updated at every iteration of active learning, as more informative samples are queried and labeled from target domain data.

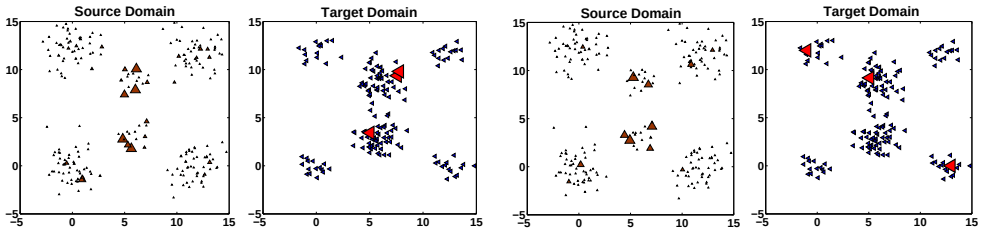
In this paper, we propose a novel transfer and active learning method that addresses the above mentioned issues. The proposed method re-weights source samples and selects a batch of query samples from the target domain, such that the marginal distribution represented by the data set consisting of re-weighted source samples, labeled target domain data (if any) and the selected query samples from the target domain, is closest to the distribution represented by the set of unlabeled target domain data, with the purpose of learning a classifier with low generalization error. This is achieved by solving a convex optimization problem, which minimizes the difference in a marginal probability measure between the two data sets. The optimization problem is found to minimize the similarities between the source data samples with large weights and the selected target samples, potentially avoiding information overlap between them. This process is repeated at every iteration to update transferred knowledge dynamically (Section 2.2).

To illustrate the problem of information overlap, we created source and target domain data with different marginal probability distributions, as shown in Figure 1. We created six dense regions of different densities for each of the domains. Figure 1 (a) shows the re-weighted source domain (size of the triangles is proportional to the weights) and the query set (batch size = 3) selected from target domain data by following a two step methodology, i.e., domain adaptation followed by active learning as in (Rai et al., 2010). Figure 1 (b) shows the re-weighted source domain data and the query set selected from target domain data by the proposed framework. We observe that the similarity in instances with large weights in the source domain and those selected from the target domain, in the two stage approach [Figure 1 (a)], is significantly higher compared to the case when the domain adaptation and active learning are done simultaneously [Figure 1 (b)], leading to considerable information overlap amongst the instances, in the former case.

To the best of our knowledge, this is the first work that performs transfer and batch-mode active learning simultaneously via a convex optimization problem. Batch-mode active learning selects a 'set' of most informative instances (Guo, 2010; Guo & Schuurmans, 2007; Hoi et al., 2006; Yu et al., 2006). Reducing the marginal distribution with the target domain data via re-weighting source instances has been previously used in the context of *transfer learning* applications (Huang et al., 2007; Pan et al., 2009; Shimodaira, 2000; Sugiyama et al., 2008; Bickel et al., 2009), however, performing active learning jointly on the basis of the same criterion, is a novel contribution of this work.

We measure the difference in the marginal probability distribution between the two sets of data using the Maximum Mean Discrepancy (MMD) proposed by Borgwardt et al. (Borgwardt et al., 2006; Gretton et al., 2007; Sriperumbudur et al., 2010). The subset selection problem is an NP-hard combinatorial integer programming problem. Specifically, the proposed formulation is an integer quadratic programming problem. We solve a continuous quadratic programming problem (by relaxing the integer constraint) on a convex function. The proposed formulation is also easily extendable to multi-source settings and is easily configurable for only transfer or active learning with corresponding parameter changes.

We tested the proposed method on two publicly available data sets, 20 Newsgroups and Sentiment Analysis and on two biomedical image data sets, Fly-FISH (Lecuyer et al., 2011) and BDGP (Tomancak et al., 2002), consisting of images representing 7 developmental stages in the life cycle of *Drosophila* embryo. Each developmental stage forms a class. The empirical results show that the combined approach of transfer and active learning performs significantly better than a framework performing transfer and active learning in two separate stages. We further extended the proposed method by incorporating uncertainty of the predicted labels of the unlabeled data, a commonly used criterion for active learning (Campbell et al., 2000; Schohn & Cohn, 2000; Tong & Koller, 2000; Joshi et al., 2009; Jing et al., 2004) and observed that the performance of the proposed method improved towards the later iterations, as the number of labeled data from the target domain increased.



(a) (b)

Figure 1. Source and target domains with different data distributions and the corresponding selected set of query data points from the target domain (red triangles) and weights of source instances shown by the size of the source data points based on (a) the two stage approach of domain adaptation and active learning and (b) the proposed single stage approach of domain adaptation and active learning (figures best viewed in color).

2. Proposed Framework

Given a parametric classification model, the learning algorithms often learn the parameters θ by maximizing the joint probability $P(X, Y|\theta) = P(X|\theta)P(Y|X, \theta)$ where X and Y are represented empirically by the training data $X_{tr} = \{x_1, x_2, \cdots, x_n\}$ and their corresponding labels $Y_{tr} = \{y_1, y_2, \cdots, y_n\}$ and $P(X)$ and $P(Y|X)$ denote the marginal and conditional probability distribution of X and Y respectively. Traditional machine learning algorithms are based on the assumption that the training data (X_{tr}, Y_{tr}) represents the true underlying distributions of X and Y and hence a model learned on this data works well on the test data (X_{tst}, Y_{tst}) which is also drawn i.i.d. from the same distribution. In this paper we propose a combined transfer and active learning method to perform simultaneous domain adaptation on the source data S and active sampling on the target domain data U so that the selected training data has similar probability distributions as the test data. Assuming that the labeling function or the conditional probability $P(Y|X)$ is the same for both source and target domain data, the problem reduces to performing domain adaptation on S, so as to obtain the domain adapted source data S_a, and selecting a subset of samples Q from U such that the marginal probability $P_{S_a \cup Q \cup L}(X)$ is similar to the marginal probability $P_{U \setminus Q}(X)$, where L denotes the existing labeled target domain data.

2.1. Proposed Joint Optimization Framework for Transfer and Batch-mode Active Learning

The proposed transfer and active sampling method, referred to as *Joint Optimization based Transfer and Active Learning* (JO-TAL) uses MMD (Borgwardt et al., 2006; Gretton et al., 2007; Sriperumbudur et al., 2010) to measure the marginal probability distribution difference between two sets of samples. Let us assume that we have n_s instances of source data or domain adapted source data S_a, n_u instances of unlabeled target domain data U and n_l instances of labeled target domain data L and we would like to select a batch Q of b instances such that the marginal distribution of $S_a \cup L \cup Q$ is similar to the marginal distribution of $U \setminus Q$. The MMD denoted as $\tilde{f}$ between these two sets can be expressed as follows:

$$\left\| \frac{1}{n_s + n_l + b} \sum_{j \in S_a \cup L \cup Q} \Phi(x_j) - \frac{1}{n_u - b} \sum_{i \in U \setminus Q} \Phi(x_i) \right\|_{\mathcal{H}}^2 . \tag{1}$$

where $\phi : \mathcal{X} \to \mathcal{H}$ is known as the feature space map from $\mathcal{X}$ to $\mathcal{H}$ (Borgwardt et al., 2006). Since we want to select a set Q that minimizes the mismatch between $S_a \cup L \cup Q$ and $U \setminus Q$, we propose to select a subset Q of U that minimizes $\tilde{f}$. Next, we define a binary vector α of size n_u where each entry α_i indicates whether the data $x_i \in U$ is selected or not. If a point is selected, the corresponding entry α_i is 1 else 0. The domain adaptation is performed by re-weighting source domain data, a common technique used in transfer learning (Huang et al., 2007; Shimodaira, 2000; Bickel et al., 2009; Cortes et al., 2010), which aims to match the marginal distributions between the source and target domain data. To this end we define another vector β of size n_s with each entry β_i indicating the weight of the data $x_i \in S$. Then, the problem reduces to finding α and β that minimize the following cost function:

$$\begin{aligned} \left\| \frac{1}{n_s + n_l + b} \left(\sum_{i \in S} \beta_i \Phi(x_i) + \sum_{j \in L} \Phi(x_j) + \sum_{i \in U} \alpha_i \Phi(x_i) \right) \right. \\ \left. - \frac{1}{n_u - b} \sum_{i \in U} (1 - \alpha_i) \Phi(x_i) \right\|_{\mathcal{H}}^2 , \\ \text{s.t.} \quad \alpha_i \in \{0, 1\}, \beta_i \in [0, 1], \alpha^T \mathbf{1} = b. \end{aligned} \tag{2}$$

where $\mathbf{1}$ is a vector of the same dimension as α with all entries being 1 and the symbol T is used to represent the matrix or vector *transpose* operation. The first term denotes the mean of the mapped features of re-weighted source data, already labeled target data and selected target data. Note that if a point x_i is

not selected in the current set then α_i will be 0 and this term would not get added in the summation. The second term is the mean of the mapped features of the unlabeled data set minus the selected query set. The first constraint ensures that each entry in α is either 0 or 1 and the third constraint ensures that exactly b entries of α are 1, meaning exactly b instances are selected from the unlabeled data set, where b is specified a priori by the user. The above formulation can be represented as:

$$\begin{aligned} \min \quad & \frac{1}{2}\alpha^T K_{u,u}\alpha + \frac{1}{2}\beta^T K_{s,s}\beta + \beta^T K_{s,u}\alpha \\ & - k_{u,u}^T\alpha - k_{s,u}^T\beta + k_{u,l}^T\alpha + k_{s,l}^T\beta + \text{const.} \\ \text{s.t.} \quad & \alpha_i \in \{0,1\}, \beta_i \in [0,1], \alpha^T\mathbf{1} = b. \end{aligned} \tag{3}$$

The various terms in the above expression are given as follows. We denote n_s, n_u and n_l as the number of source domain data S, unlabeled target domain data U and already labeled target domain data L respectively, G as the $(n_s+n_u+n_l)*(n_s+n_u+n_l)$ kernel Gram matrix over S, U and L, arranged in order, using a kernel function K such that $G(i,j) = K(x_i, x_j)$ and $c = \frac{n_l+n_s+n_u}{n_u-b}$. Then:

$$\begin{aligned} K_{s,s} &= \frac{1}{c^2}G(1:n_s, 1:n_s), \\ K_{u,u} &= G(n_s+1:n_s+1+n_u, n_s+1:n_s+1+n_u), \\ K_{s,u} &= \frac{1}{c}G(1:n_s, n_s+1:n_s+1+n_u), \\ k_{u,u}(i) &= \frac{n_l+n_s+b}{c^2(n_u-b)}\sum_{j=1}^{n_u} K_{u,u}(i,j), \\ k_{s,u}(i) &= \frac{n_l+n_s+b}{c^2(n_u-b)}\sum_{j=1}^{n_u} K_{s,u}(i,j), \\ k_{s,l}(i) &= \frac{1}{c^2}\sum_{j=1}^{n_l} G(i, n_s+n_u+j), \\ k_{u,l}(i) &= \frac{1}{c}\sum_{j=1}^{n_l} G(i+n_s, n_s+n_u+j). \end{aligned}$$

Based on the above expressions, we can draw the following observations regarding the properties of the re-weighted source data and the selected query set from target data:

- The first term ensures that the selected query set has minimum similarity within itself, avoiding *redundancy* in the selected set.
- The second term ensures that the re-weighted source instances have minimum similarity within themselves, again avoiding *redundancy* in the source set.
- The third term ensures that the selected query has minimum similarity with re-weighted source data thus avoiding *information overlap*.
- The fourth term enforces the selected query instances to be similar to the unselected ones, ensuring *representativeness*.
- The fifth term enforces the re-weighted data to be similar to the unselected ones, ensuring *representativeness* of the target domain data.
- The sixth term implies that the selected target data have less similarity with already labeled data, ensuring *diversity* in the selected set.
- The seventh term implies that the re-weighted source data have less similarity with already labeled target data, ensuring *diversity* in the re-weighted set.

Thus the proposed method selects examples which meet all the desirable properties for transfer and active learning, i.e., representativeness, diversity and minimum redundancy or information overlap.

2.2. Quadratic Programming (QP) Problem

The binary constraint on α_i makes the integer quadratic problem in (3) NP-hard. A common strategy is to relax the constraints to transform it into the following QP formulation:

$$\min_{X:X_i\in[0,1], X^T\mathbf{B}=b} \quad 0.5X^THX + f^TX \qquad \text{where}$$

$$X = \begin{pmatrix}\beta\\ \alpha\end{pmatrix},\ H = \begin{pmatrix}K_{s,s} & K_{s,u}\\ K_{s,u}^T & K_{u,u}\end{pmatrix},\ f = \begin{pmatrix}k_{s,l}-k_{s,u}\\ k_{u,l}-k_{u,u}\end{pmatrix},$$
$$B = \begin{pmatrix}O\\ I\end{pmatrix}, \qquad I = \mathbf{1}_{n_u\times 1}, \qquad O = \mathbf{0}_{n_s\times 1}. \tag{4}$$

The standard QP can be solved efficiently by applying many existing solvers. The key steps at each iteration are provided in Algorithm 1. The source weight vector, β_{new}, is initialized at the end of first iteration with the vector β. During subsequent iterations the source weights are added (step 7), thus reinforcing source weights, at every iteration.

Algorithm 1 JO-TAL

1: **Input:** S: source domain data; L: set of labeled target domain data; U: set of unlabeled target domain data; b: batch size; β_{new} : source weights (for iteration nos. > 1);
2: **Output:** β_{new}: source weights (updated), Q: target query set;
3: Compute H and f as explained in Section 2.1 and 2.2.
4: Compute β and α by solving (4).
5: $Q \rightarrow$ top b instances of U, sorted in descending order of α.
6: Update L, U : $L \leftarrow L \cup Q$, $U \leftarrow U \backslash Q$.
7: Update β_{new} : $\beta_{new} \leftarrow \beta_{new} + \beta$ (iteration nos. > 1).

The proposed formulation can be easily extended to include additional evaluation criteria (E_u) by adding a corresponding linear term $E_u^T\alpha$ in Equation (4), while still maintaining the quadratic form. In order to incorporate uncertainty of predictions by the existing

classifier, we added a term $(-E_u^T\alpha)$, where $E_u(x_i)$ is entropy of predicted labels computed as in (Guo & Schuurmans, 2007) for each unlabeled data x_i in target domain, with a weighting factor of n_l/n_u. Besides, the proposed formulation is easily configurable for only transfer learning or active learning with corresponding parameter changes, as discussed in Section 3.

3. Empirical Evaluations

Data sets. We evaluated the empirical performance of the proposed *JO-TAL* algorithm on three real-world data sets: the biological image data sets (Fly-FISH and BDGP), the 20 Newsgroups [1] and Sentiment Analysis [2] data sets, besides on a synthetic data set shown in Figure 1. The biological image data sets consist of images of 7 early developmental stages of *Drosophila* embryo. Each stage forms a class. Each image is represented by 3850 textural features that are extracted using Gabor filters (Liu & Wechsler, 2002). The 20 Newsgroups data set is a collection of approximately 20,000 newsgroup documents, partitioned (nearly) evenly across 20 different categories. We built two sets of source domain data vs. unlabeled data (target domain) as follows: (1) Sports: rec.sport.hockey vs. rec.sport.baseball and (2) Scientific: sci.med vs. sci.electronics. The positive class of each source and target domain data consists of 200 documents randomly sampled from the respective categories and the negative class consists of a random mixture of 200 samples from other categories as suggested in (Eaton & desJardins, 2009). We represented each document as a binary vector consisting of the 200 most discriminating words determined by Weka's info-gain filter (Witten & Frank, 2000), after removing stop words and using a document frequency of 5. The Sentiment Analysis data set contains positive and negative reviews on four product categories (or domains) including *kitched, book, dvd* and *electronics*. We processed the Sentiment Analysis data set to reduce the feature dimension to 200, from a union of features of all four categories, using a cutoff document frequency of 50. The source and target domain data consisted of 400 documents each. These consisted of randomly sampled 200 documents belonging to positive and negative classes each, from the respective set of 1000 documents.

Competing Methods. We compared the performance of *JO-TAL* with the following methods.

2-Stage based Transfer Active Learning (2S-TAL): In this method transfer and active learning are performed in two stages, as in (Rai et al., 2010). In the first stage, we perform domain adaptation on source data, using an instance re-weighting method (Huang et al., 2007) and in the second stage, we learn a classifier using the domain adapted source data and select a query set from unlabeled target data based on an existing batch-mode active learning method (Brinker, 2003).

Joint Optimization based Transfer Learning (JO-T-Rand): In this method the target domain data is selected randomly for labeling. However, the domain adaptation is performed by computing source weights β considering the randomly selected and labeled target domain data L. This is achieved via minimizing MMD between sets $S_a \cup L$ and U, by modifying the proposed QP formulation given in Equation (4), considering $b = 0$ and $\alpha = 0$, as follows:

$$\min_{X:X_i\in[0,1]} \quad 0.5X^THX + f^TX \quad \text{where}$$
$$X = \beta, \qquad H = K_{s,s}, \qquad f = k_{s,l} - k_{s,u}. \quad (5)$$

2-Stage based Transfer Learning (2S-T-Rand): In this method, we perform domain adaptation on source data (Huang et al., 2007) in the first stage, and we randomly select instances from the target domain, in the second stage. Unlike *JO-T-Rand* method, domain adaptation is performed without considering the selected and labeled target domain data.

Batch-mode Active Learning (AL): We also present the performance of a classifier learned only on target domain data, selected actively at every iteration based on reducing distribution difference between the queried (Q) and labeled target domain data (L) and the rest of the unlabeled target domain data ($U \backslash Q$). The batch-mode active learning formulation is obtained by modifying the proposed QP formulation given in Equation (4), considering $n_s = 0$ and $\beta = 0$, as follows:

$$\min_{X:X_i\in[0,1], X^T\mathbf{B}=b} \quad 0.5X^THX + f^TX \quad \text{where}$$
$$X = \alpha, \qquad H = K_{u,u}, \qquad f = k_{u,l} - k_{u,u},$$
$$B = I, \qquad I = \mathbf{1}_{n_u\times 1}. \quad (6)$$

We also extended the proposed *JO-TAL* to incorporate an entropy term as explained in Section 2.2 and referred it as *JO-TAL-Ent*.

Finally, for each of these methods, we learn a classifier using the re-weighted source and the selected target samples and compute the classification accuracy on an unseen fixed test set, from the target domain.

The above methods provide a basis for comparing the performance of the proposed method with the traditional method of performing active learning and trans-

[1] Available at http://www.ai.mit.edu/∽jrennie/
[2] Available at http://www.cs.jhu.edu/∽mdredze/

fer learning in two stages besides comparing with an active learning method.

Experimental Setup. We randomly divided each target domain data set into two sets. Batch selection based on active learning was performed on one set referred to as the unlabeled set (65%) and the effectiveness of the selection methodology was measured based on classification accuracy on the other unseen fixed set (35%) referred to as the test set. We start with *no* labeled instances from the unlabeled target domain data set. All the algorithms start with the same source domain data and same unlabeled set and test set from the target domain. The batch size for active learning was fixed at 10, except for the synthetic data set, where it was fixed at 3. Each algorithm performed transfer and active learning at each iteration, and evaluated the performance of a classifier learned using domain adapted source and randomly or actively selected target samples, on the fixed test set from the target domain. We used the Gaussian kernel function to compute the kernel Gram matrix in Equation (4) and Support Vector Machines with the Gaussian kernel as the classification model. The experiments were repeated 10 times with different sets of unlabeled test data and the average results were reported.

3.1. Comparative Studies- Synthetic data.

Figure 2 (a) shows the comparative performance of *JO-TAL*, on the synthetic data set shown in Figure 1. We observe that *JO-TAL* performed better than *2S-TAL*. This can be attributed to the efficient transfer and active learning, by selecting complementary samples from the target domain data as shown in Figure 1. Similarly, *JO-T-Rand* performed better than *2S-T-Rand*. It is however interesting to note that *JO-T-Rand* performed better than *2S-TAL* during initial iterations. However *2S-TAL* improved during the later iterations with more actively sampled data from the target domain. It is also interesting to note that the performance of *JO-TAL-Ent* improved at later iterations. This is due to the increased reliability of the classifier, when learned with more labeled data from the target domain.

Variation in MMD Vs Number of Selected Samples: We also investigated the change in the distribution difference (MMD) between the selected and target domain data, at every iteration. The MMD value is computed between the set consisting of re-weighted source data, queried and labeled target domain data and the set of unlabeled target domain data. Figure 2 (b) shows that our algorithm decreases the MMD value monotonically as more data samples are selected from the target domain.

3.2. Comparative Studies- Biomedical data.

The comparative performance of the proposed method *JO-TAL*, on Fly-FISH and BDGP data sets is shown in Figure 2; Figure 2 (c) shows the results with Fly-FISH as source and BDGP as target domain data and Figure 2 (d) shows the comparative performance with BDGP as source and Fly-FISH as target domain data. We observe that for both cases, *JO-TAL* and JO-TAL-Ent performed 9% to 10% better than *2S-TAL* (*when the number of labeled instances from the target domain is around 50.*). This can again be attributed to the efficient transfer and active learning, by selecting complementary samples from the source and target domain data sets, as shown in Figure 7, (provided in the supplementary material. We also note that *JO-T-Rand* performed comparably to *2S-TAL* and incorporating transfer learning improved the performance of the *AL* method by 11% to 9% for cases with BDGP and Fly-FISH as target domain data respectively. Thus the proposed joint optimization framework provides a viable solution to the problem of biological image annotation, by effectively using related image data sets to develop a classifier for a new data set.

Variation in MMD Vs Number of Selected Samples: As in the case of synthetic data, the MMD value between the set of re-weighted and the selected target samples and the set of unlabeled target data, reduced monotonically at every iteration (Figure 3). We also observe that the decrease in MMD value corresponded to the increase in classification accuracy on the test set as shown in Figures 2 (c) and 2 (d). The decrease in MMD value during the initial iterations is more than the decrease towards the later iterations, resulting in the higher increase in accuracy values during the initial iterations than later iterations.

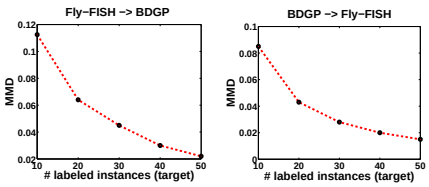


Figure 3. MMD vs. nos. of target data (biomedical data).

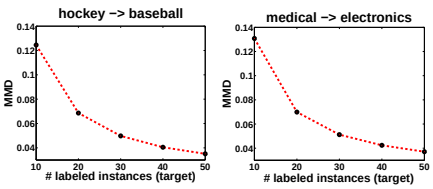


Figure 6. MMD vs. nos.of target data (20 Newsgroups).

3.3. Comparative Studies- 20 Newsgroups.

Figure 4 shows that *JO-TAL* performed better than *2S-TAL* by 5% and 8% for Sports and Scientific cate-

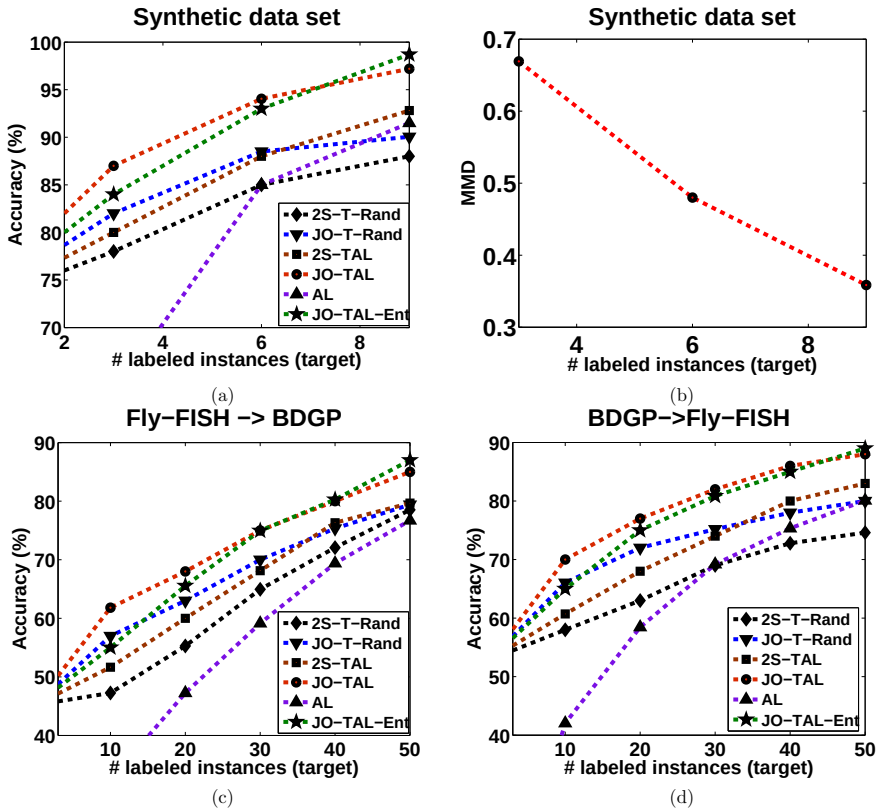


Figure 2. (a) Comparative performance on synthetic data shown in Figure 1 and (b) MMD vs. number of labeled target data. Comparative performance on biomedical data sets (c) Fly-FISH (source)- BDGP (target) and (d) BDGP (source) - Fly-FISH (target).

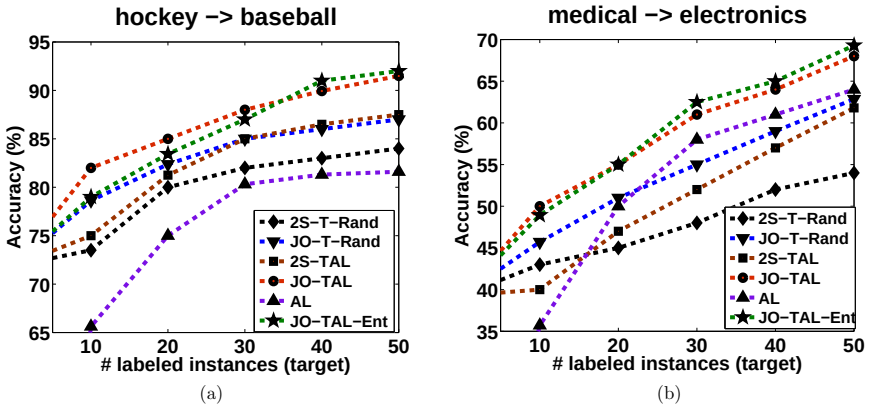


Figure 4. Comparative performance on 20 Newsgroups document categories: (a) Sports (b) Scientific.

gories respectively. Furthermore, for both categories, *JO-T-Rand* performed comparably to *2S-TAL*. This shows that performing transfer learning by taking into account the randomly selected samples from target domain can be equally or more effective than performing transfer and active learning in two stages, due to the selection of complementary samples from both domains as illustrated in Figure 1 and in Figure 7 of the supplementary material. Note that the performance of *JO-TAL-Ent* improved towards later iterations, for reasons explained in Section 3.1. We also observe that improvement in classification accuracies due to incorporation of transfer learning is more for Sports (10%) and moderate for Scientific category (5%). This can be attributed to the extent of difference in distribution between the source and target domain data, measured using MMD. The MMD value is 0.0121 and 0.0239 for Sports and Scientific categories respectively. Similar to previous results, the MMD value monotonically decreased for both of the 20 Newsgroups categories, as

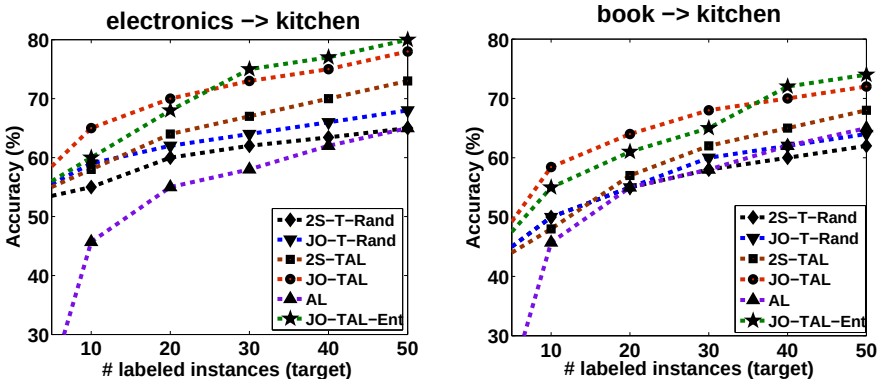


Figure 5. Comparative performance on Sentiment Analysis data set.

shown in Figure 6.

3.4. Comparative Studies- Sentiment Analysis. Figures 5 (a) and 5 (b) show the comparative performance of *JO-TAL* on the Sentiment Analysis data set. The first and second names in the title of the figures refer to the source and target domains respectively. We observe that *JO-TAL* performed better than *2S-TAL* by 7% and 5% for the cases with electronics and book data sets as source domains, while documents belonging to the category of kitchen forming the target domain, respectively. Similar to the 20 Newsgroups data set, the performance of *JO-TAL-Ent* improved towards later iterations. We also observe that incorporation of transfer learning has improved the classification accuracy on the kitchen data set by 13% and 9% with electronics and book as source data sets respectively. This can be explained by the differences in their MMD values, which are, 0.0145 and 0.0349 for electronics vs. kitchen and book vs. kitchen data sets respectively. More results are provided in the supplementary material.

4. Related work

There has not been much prior work towards combining transfer and active learning methodologies. A combination of transfer learning with active learning has been presented by Shi et al. (2008). In this method, a classifier is learned on the source domain data and another classifier is learned on an initial pool of labeled target domain data. The label for an unlabeled instance is predicted by both classifiers, and based on the confidence of predictions, the instance is queried for manual annotation. Another method, proposed by Rai et al. (2010) uses multiple classifiers to perform transfer and active learning. Both of these methods, perform transfer and active learning in multiple stages. A drawback of the first method is that the source data is used without any domain adaptation, besides the requirement of an initial pool of labeled target domain data for its operation. And the drawback of the second method is that the source domain adaptation is done once initially and is not refined as more target domain data gets queried. A combined transfer and active learning method was proposed by Chen et al. (2011), based on the assumption that the target domain may have unique features; a situation different from the one considered in this paper.

5. Conclusion

In this paper, we propose a novel convex optimization formulation for performing transfer and active learning simultaneously. The proposed formulation re-weights source instances and selects a set of query samples from the target domain, simultaneously based on minimizing the marginal probability differences between the set consisting of re-weighted source and selected target samples and the set of unlabeled target domain data. The motivation behind this approach is to ensure that a classifier learned on domain adapted source and labeled target domain data, has good generalization performance on the unlabeled target domain data and also on the unseen data coming from similar distribution. The proposed method is formulated as an integer quadratic programming problem and demonstrates sensible data selection properties. Our empirical studies on three real world data sets show that the proposed approach achieves superior performance compared to the existing approaches of performing transfer and active learning in two stages. In future work, we plan to study the generalization performance of the proposed formulation.

6. Acknowledgments

This research is sponsored by NSF IIS-0953662, CCF-1025177, NIH LM010730, and ONR N00014-11-1-0108.

References

Bickel, S., Brückner, M., and Scheffer, T. Discriminative learning under covariate shift. *JMLR*, 10:2137–2155, 2009.

Borgwardt, K.M., Gretton, A., Rasch, M.J., Kriegel, H.P., Schlkopf, B., and Smola, A.J. Integrating structured biological data by kernel maximum mean discrepancy. *Bioinformatics*, 22(14):49–57, 2006.

Brinker, K. Incorporating diversity in active learning with support vector machines. In *ICML*, 2003.

Campbell, C., Cristianini, N., and Smola, A.J. Query learning with large margin classifiers. In *ICML*, 2000.

Chen, M., Weinberger, K.Q., and Blitzer, J. Co-training for domain adaptation. In *NIPS*, 2011.

Cortes, C., Mansour, Y., and Mohri, M. Learning bounds for importance weighing. In *NIPS*, 2010.

Eaton, E. and desJardins, M. Set-based boosting for instance-level transfer. In *ICDM*, 2009.

Gretton, A., Borgwardt, K.M., Rasch, M., Schlkopf, B., and Smola, A.J. A kernel method for the two-sample-problem. In *NIPS*, 2007.

Guo, Y. Active instance sampling via matrix partition. In *NIPS*, 2010.

Guo, Y. and Schuurmans, D. Discriminative batch mode active learning. In *NIPS*, 2007.

Hoi, S.C.H., Jin, R., Zhu, J., and Lyu, M.R. Batch mode active learning and its application to medical image classification. In *ICML*, 2006.

Huang, J., Smola, A.J., Gretton, A., Borgwardt, K.M., and Scholkopf, B. Correcting sample selection bias by unlabeled data. In *NIPS*, 2007.

Jing, F., Li, M., Zhang, H., and Zhang, B. Entropy based active learning with support vector machines for content based image retrieval. In *ICME*, 2004.

Joshi, Ajay J., Porikli, F., and Papanikolopoulos, N. Multi-class active learning for image classification. In *CVPR*, 2009.

Lecuyer, E., Yoshida, H., Parthasarathy, N., Alm, C., and Babak, T. Global analysis of mrna localization reveals a prominent role in organizing cellular architecture and function. *Cell*, 2011.

Liu, C. and Wechsler, H. Gabor feature based classification using the enhanced fisher linear discriminant model for face recognition. *IEEE Trans. Image Processing*, 11:467–476, 2002.

Pan, S.J. and Yang, Q. A survey on transfer learning. *TKDE*, 2009.

Pan, S.J., Tsang, I.W., Kwok, J.T., and Yang, Q. Domain adaptation via transfer component analysis. In *IJCAI*, 2009.

Rai, P., Saha, A., H. Daumé III, and Venkatasubramanian, S. Domain adaptation meets active learning. In *NAACL-HLT Active Learning for NLP Workshop*, 2010.

Schohn, G. and Cohn, D. Less is more: Active learning with support vector machines. In *ICML*, 2000.

Settles, B. Active learning literature survey. In *Computer Sciences Technical Report 1648*. University of Wisconsin-Madison, 2009.

Shi, X., Fan, W., and Ren, J. Actively transfer domain knowledge. In *ECML/PKDD*. Antwerp, Belgium, 2008.

Shimodaira, H. Improving predictive inference under covariate shift by weighting the log-likelihood function. *JSPI*, 90:227–244, 2000.

Sriperumbudur, B., Gretton, A., Fukumizu, K., Schlkopf, B., and Lanckriet, G. Hilbert space embeddings and metrics on probability measures. *JMLR*, 11:1517–1561, 2010.

Sugiyama, M., Nakajima, S., Kashima, H., Buenau, P.V., and Kawanabe, M. Direct importance estimation with model selection and its application to covariate shift adaptation. In *NIPS*, 2008.

Tomancak, P., Beaton, A., Weiszmann, R., Kwan, E., and Shu, S. Systematic determination of patterns of gene expression during drosophila embryogenesis. *Genome Biology*, 3, 2002.

Tong, S. and Koller, D. Support vector machine active learning with applications to text classification. *JMLR*, 2:45–66, 2000.

Witten, I.H. and Frank, E. Data mining: Practical machine learning tools with java implementations. Morgan Kaufmann, 2000.

Yu, K., Bi, J., and Tresp, V. Active learning via transductive experimental design. In *ICML*, 2006.