Web-Induced Heterogeneous Transfer Learning with Sample Selection

Sanatan Sukhija, Narayanan C Krishnan

Indian Institute of Technology Ropar, Punjab, India PB-140001
sanatan@iitrpr.ac.in, ckn@iitrpr.ac.in

Abstract. Transfer learning algorithms utilize knowledge from a data-rich source domain to learn a model in the target domain where labeled data is scarce. This paper presents a novel solution for the challenging and interesting problem of Heterogeneous Transfer Learning (HTL) where the source and target task have heterogeneous feature and label spaces. Contrary to common space based HTL algorithms, the proposed HTL algorithm adapts source data for the target task. The correspondence required for aligning the heterogeneous features of the source and target domain is obtained through labels across two domains that are semantically aligned using web-induced knowledge. The experimental results suggest that the proposed algorithm performs significantly better than state-of-the-art transfer approaches on three diverse real-world transfer tasks.

Keywords: Heterogeneous Transfer Learning · Sample Selection

1 Introduction

Traditional supervised algorithms require sufficient labeled data to learn a computational model with a reasonable generalization to unseen examples. However, for many real-world problems, collecting labeled data is often very expensive and cumbersome. Transfer learning approaches utilize knowledge from an auxiliary domain with abundant labeled data (source domain) to perform tasks in domains with scarce labeled data (target domain). HTL [37] algorithms transfer knowledge from one domain to the other when the two domains have different features. Due to the heterogeneous feature spaces, the first task of any HTL algorithm is to decide a "common" space for adaptation. The second task is to bridge the gap between the data differences that arise when the data from both the domains is projected onto the common space. This is generally achieved by leveraging some pivotal information that is shared among the domains. These pivots could be in the form of instance correspondences [39], overlapping features [20], shared label space [29, 40, 28], common meta-features/latent space [38, 14, 12, 11] or any task specific/independent information [21, 41].

Latent Space Transformation (LST) approaches to HTL project the data from both the domains onto a shared subspace for adaptation, thus learning two transformations, one each for the source and target domain. On the other hand, Feature Space Remapping (FSR) approaches consider the common space as either of the two domains and determine a single transformation to transform data from the source domain to the target domain or vice-versa. The recent state-of-the-art HTL approaches leverage the common label space either to determine the cross-domain correspondences for learning the transformation(s) [29, 40, 28] or formulate a solution for obtaining the transformations as a minimization objective [13, 36, 33, 35, 17]. However, these approaches are not directly applicable for knowledge transfer between domains with heterogeneous label spaces.

We propose a novel FSR algorithm (refer [27, 26] for our preliminary work) that works even when there are no shared features and instance correspondences between the source and target domain. It utilizes the label space dependencies estimated through Normalized Google Distance to co-align the data from the two domains in the target space while preserving the original structure of the source data. Being a FSR framework, the proposed approach overcomes the need to determine an optimal shared subspace as compared to LST approaches and unlike [17, 36], does not suffer from out-of-sample extension problem [25]. The approach also utilizes source instances whose labels are absent in the target domain, by encoding the absent labels using the inter-label relationships to the target labels. Along with inter-label dependencies across the heterogeneous label spaces, the approach also utilizes intra-label relationships among target labels to learn a robust target model.

1.1 Problem Definition

Let $S \in \mathbb{R}^{n_S \times d_S}$ and $T \in \mathbb{R}^{n_T \times d_T}$ be the source and target domain data respectively where n_S and n_T represent the number of labeled data points in each domain respectively and $n_S \ggg n_T$. The number of features in the source and target domain are denoted by d_S and d_T respectively. The features in the two domains are different and $d_S \neq d_T$. x^S denotes a labeled source instance with y^S as the associated label. Similarly, x^T is a labeled target instance with y^T as its label. The source and the target label space may or may not be overlapping. However, we assume that there exists semantic relationships within and across the label spaces. Let the number of unique labels in the source and target domain be L_S and L_T respectively. The goal of the proposed approach is to learn relevant source data points $B_S \in \mathbb{R}^{n_S \times d_T}$ that adapt well to the target task. The set of relevant source data is used along with the limited target data $\{x_i^T, y_i^T\}_{i=1}^{n_T}$ to learn the model for the target task.

2 Related Work

There have been many approaches for transfer learning, some of which have been extended for heterogeneous transfer learning. Manifold alignment based LST approaches [30, 33, 32, 31] can be viewed as constrained dimensionality reduction frameworks that intend to find a low-dimensional embedding for multiple domains where the geometric structure of the original domains is preserved. These approaches assume that the heterogeneous source and target domain share a smooth low-dimensional manifold (subspace). However, such a strong manifold assumption may not hold good for real-world heterogeneous transfer tasks, especially for datasets with high-dimensional features [35]. Supervised Heterogeneous Feature Augmentation (HFA) [17] is a SVM-based LST optimization framework that uses a common augmented feature space for adaptation. Since HFA does not return the transformation matrices explicitly, it suffers from the out-of-sample extension problem. Subspace Co-Projection (SCP) [35] is a semi-supervised LST optimization framework that learns the model weights in the projected subspace simultaneously with the transformations. The closed form solution of SCP requires large matrix inversions for high-dimensional datasets. Co-regularized Heterogeneous Transfer Learning (Co-HTL) [34] is a supervised LST approach that jointly aligns the data from the domains in the shared subspace. A common limitation of these LST approaches is that they require determining the optimal subspace by performing a grid-search on the dimension of the shared subspace (d).

In contrast, the FSR approaches directly map the features across the domains. However, learning the direct transformation involves estimation of a larger set of parameters in comparison with LST approaches. Sparse Heterogeneous Feature Remapping (SHFR) [40] leverages the common labels across the domains encoded using error correcting output codes (ECOC) as pivots to generate cross-domain correspondences. Supervised Heterogeneous Domain Adaptation using Random Forests (SHDA-RF) [29] relies on common label distributions that are obtained from leaf nodes of decision trees trained on labeled source and target domain data as the pivots. Both SHFR and SHDA-RF rely on a common set of labels between source and target domain to estimate correspondences and hence, cannot be directly applied to bridge two domains with heterogeneous label spaces.

The proposed FSR optimization framework overcomes the limitations of the above-mentioned approaches. It bridges the domains with heterogeneous feature and label spaces in a generic setting without relying on instance or feature correspondences. Even if there exists very few labeled instances in the target domain, the proposed algorithm is effective for transferring knowledge as asserted by the experimental results.

3 Proposed Methodology

Unlike LST approaches that have an inherent limitation of determining the optimal subspace for transfer, the proposed HTL algorithm, Web-Induced Heterogeneous Transfer Learning with Sample Selection (WIHTLSS), is conceived as a FSR minimization objective with the goal of constructively utilizing data from the source and target domains to learn a robust target model. Given the source domain data S and target domain data T, the proposed objective (Equation 1) iteratively minimizes the overall loss incurred by jointly aligning the data of the source and target domain in the target space while learning the transformed source data B_S and the transformation $P \in \mathbb{R}^{d_S \times d_T}$ that links the heterogeneous features of the source and target domain.

$$\min_{B_S, P} L(S, B_S, P) + \beta D(B_S, T) + \kappa G(B_S, T, m) + \lambda R(P, B_S) \tag{1}$$

The first term, L, in Equation 1 preserves the original structure in the transformed source data, while the second and the third term, D and G, align the transformed source data closer to the target data distribution. These two terms together determine the extent of alignment between transformed source and target instances. Excluding D and G will not adapt the transformed source to the target, and excluding L will overly bias the transformed source instances towards the limited labeled target instances, thus not generalizing across the target. There is a trade-off between leveraging the relatedness to the source domain and the extent to which we want to adapt the transformed source data to the target task. This tradeoff is regulated by the variables β and κ. The last term, R, is the regularizer that prevents overfitting.

We define L in terms of the reconstruction error (Equation 2), which measures the extent to which the structure of the original source data is preserved in the target domain. The reconstruction error computes the loss incurred due to projection of source data, i.e., the difference between the original domain data and the transformed data being remapped to the original space.

$$L(S, B_S, P) = \mid\mid S - B_S P' \mid\mid^2 \tag{2}$$

Here, $P' \in \mathbb{R}^{d_T \times d_S}$ denotes the transpose of the transformation P. The reconstruction error takes advantage of the relatedness of the source and target domains to fill in the void in the target space with missing label data. An added benefit of using reconstruction error is that unlabeled data can be used to induce regularization which in turn helps to learn a more robust transformation.

The second and third term (D and G) that measure the mis-alignment between transformed source and target instances are defined in terms of weighted pairwise distances between the transformed source and target instances (presented in Equation 3).

$$\begin{aligned} D(B_S, T) &= \sum_{i=1}^{n_S} \sum_{j=1}^{n_T} \mid\mid B_{S_i} - x_j^T \mid\mid^2 W_{ij} \\ G(B_S, T, m) &= \sum_{i=1}^{n_S} \sum_{j=1}^{n_T} \max(0, m - \mid\mid B_{S_i} - x_j^T \mid\mid^2)(1 - W_{ij}) \end{aligned} \tag{3}$$

Inspired by the contrastive loss [10], the second term D constrains the transformed source instances to be closer to the target instances with the same or related labels whereas the third term G ensures that dissimilar transformed source instances are pushed apart and kept at a minimum distance m (the margin). The similarity between the labels of the i^{th} transformed source instance and the j^{th} labeled target instance (denoted as W_{ij}) is assigned as the weight for aligning them.

As the domains have heterogeneous labels, weighting the distance based on the relatedness of the labels allows the use of source data tagged with related labels, when the exact target label is absent in the source domain. We define the encoding of the source labels in terms of target labels using the Normalized Google Distance (NGD) [6]. Equation 4 defines the NGD between two search keywords y_1 and y_2, where $f()$ denotes the number of web page hits returned by the Google search engine and N is the number of pages indexed by Google multiplied by the average number of singleton search keywords on those pages.

$$NGD(y_1, y_2) = \frac{\max\{\log f(y_1), \log f(y_2)\} - \log f(y_1, y_2)}{\log N - \min\{\log f(y_1), \log f(y_2)\}} \tag{4}$$

Since NGD is a dissimilarity measure that returns the value between $[0, \infty)$, we standardize it to a similarity measure W that outputs a value between [0,1] (Equation 5).

$$W_{ij} = \left(1 - \frac{NGD_{ij}}{Z}\right) \tag{5}$$

where Z denotes the maximum NGD score obtained from the labels across the domains. Using the similarity matrix W, we induce semantic co-alignment in our framework by exploiting the inter-label space similarities.

While minimizing the inter-domain differences using the label information, there is a significant risk of over-fitting on the limited target training data. Hence, we adopt an explicit regularizer $R(.,.)$ in the objective function to penalize over-fitting as depicted in Equation 6.

$$R(P, B_S) = || B_S ||^2 + || P ||^2 \tag{6}$$

The overall objective $J()$ is given in Equation 7.

$$\begin{gathered} J(.) = \min_{B_S, P} \; || S - B_S P^{'} ||^2 \; + \; \beta(\sum_{i=1}^{n_S} \sum_{j=1}^{n_T} W_{ij} \; || B_{S_i} - x_j^T ||^2) + \\ \kappa(\sum_{i=1}^{n_S} \sum_{j=1}^{n_T} (1 - W_{ij}) \max(0, m - || B_{S_i} - x_j^T ||^2)) + \lambda(|| B_S ||^2 + || P ||^2) \end{gathered} \tag{7}$$

The proposed optimization problem is not jointly convex with respect to the two variables B_S and P. However, it is convex with respect to any one of them while the other has been fixed. Consequently, we utilize an alternating algorithm for solving the unconstrained optimization (similar to the E-M process), by iteratively fixing one variable to estimate the remaining one until convergence. The proposed unconstrained optimization problem can be easily solved by using methods that are gradient-based or hessian-based. In contrast to gradient-based methods, hessian-based methods such as quasi-newton have a faster rate of convergence but demand extensive memory $(O(d_S.d_T)^2)$ when dealing with high-dimensional datasets. The cross-lingual text transfer tasks involve high dimensional datasets where as the cross-domain activity recognition tasks have a significantly smaller dimension size. Hence, we use Conjugate Gradient Descent (CGD) method [4] (gradient-based) to solve the cross-lingual transfer tasks and quasi-newton method (hessian-based) for the cross-domain activity recognition tasks. The optimization routine comprises of two alternating steps.

Step 1: Fix B_S, and solve for P, using the gradient update for P.

$$\nabla_P(J) = 2(-S' B_S + P B_S' B_S + \lambda P) \tag{8}$$

After updating P, the next step involves solving for B_S.

Step 2: Use P from Step 1 and update the value of B_S.

$$\nabla_{B_S}(J) = 2(-SP - B_S P' P + \beta(W_S B_S - WT) - \kappa(W_D B_S - (1-W)T) + \lambda B_S) \quad (9)$$

Here, $(W_S)_{ii} = \sum_{j=1}^{n_T} W_{ij}$ and $(W_D)_{ii} = \sum_{j=1}^{n_T} (1 - W_{ij})$ where $W_S, W_D \in [0,1]^{n_S \times n_S}$. The two alternating steps of the proposed algorithm are repeated iteratively until convergence. These steps are summarized in Algorithm 1.

Algorithm 1 Web-Induced Heterogeneous Transfer Learning with Sample Selection (WIHTLSS)

```
Inputs: $S \in \mathbb{R}^{n_S \times d_S}$ and $T \in \mathbb{R}^{n_T \times d_T}$
Output: $B_S \in \mathbb{R}^{n_S \times d_T}$, $P \in \mathbb{R}^{d_S \times d_T}$
  (1) Randomly initialize $P$ and $B_S$.
  repeat
    (2) Fix $B_S$. Find optimal $P$ using quasi-newton or Conjugate Gradient Descent (CGD).
    (3) Use $P$ obtained from the previous step and then find optimal $B_S$ by using CGD or quasi-newton.
  until $P$ and $B_S$ are convergent
```

The existing HTL algorithms assume that the knowledge is transferred to a different but related target domain [23]. As the feature spaces are disparate, quantifying the extent of this relatedness apriori is a hard problem. In practice, the source and the target domain data can be very distant for real-world tasks and furthermore, only a handful of source data might be relevant to the target task. Since the proposed optimization objective tries to bring the similar instances closer while keeping the dissimilar ones away, we define a transformed source instance to be relevant if it is in the close proximity of at least one target instance. We select only the k-nearest transformed source instances for every target instance to train the final model. This helps to avoid negative transfer to a certain extent at the instance-level by selecting the relevant source instances for the target task. The optimal value of k is obtained through experimentation on a validation set. The extra computation cost of finding the nearest neighbors $O(n_S.n_T.d_T)$ is negligible as compared to the overall cost of the optimization routine.

3.1 Merging Heterogeneous Label Spaces

As the label spaces are heterogeneous, we have to link the label spaces first before using the relevant transformed source data and the target data T to train the final model. Given L_T unique labels in the target domain, we encode the label y of a training instance with a vector of size L_T. The i^{th} entry of the encoded vector is the semantic similarity, computed via NGD, between y and the i^{th} target label. This encoding encapsulates the inter-label space similarities of the source and target domain along with intra-label space similarities in the target domain. The associations between the labels within and across the domains enrich the transformation by providing additional discriminating information to learn a robust model. Replacing the label y with its encoding for every labeled instance $(x, y) \in (B_S \cup T)$ results in a multi-output regression problem. Independent single output regressors ignore relationships between the outputs. These are captured through stacked regressors [3]. Stacked regression is a 2-stage process that feeds the predictions made in the first stage as an augmented input for the second stage. For an unseen target instance, the target label encoding that is closest to the predicted values (computed via Euclidean distance) is chosen as the predicted label.

4 Experiments

The performance of WIHTLSS is compared against the following baseline and state-of-the-art transfer approaches:

- **Random Forest** (BRF) [5]: We train a random forest using 100 decision trees on only target training data where each tree is learned using $\sqrt{d_T}$ features [16]. Since the labeled target data is very limited, instance bagging is not used for tree construction.
- **Support Vector Machine with Error Correcting Output Codes** (SVM-ECOC): In our experiments, we use linear kernel and Error Correcting Output Codes (ECOC) [22] on the target data to get the SVM baseline results. We obtain the optimal value of the box-penalty parameter (C[1]) by experimentation on the validation set.
- **Sparse Heterogeneous Feature Remapping** (SHFR) [40]: SHFR-ECOC is a FSR approach based on SVM-ECOC that utilizes common labels as pivots to generate cross-domain correspondences in the form of SVM weight vectors. A linear and sparse mapping is learned by employing Least Absolute Shrinkage and Selection Operator (LASSO) [9] on these correspondences. We perform a grid search to get the optimal box constraint parameter and the optimal length of the Error Correcting Output Codes using a validation set (for every fold). Since SHFR-ECOC reuses the source SVM model to evaluate the transfer performance, to ensure a fair comparison against WIHTLSS, we modify the original approach to SHFR-RF, where we train a random forest model on the transformed source data along with the target data.
- **Supervised Heterogeneous Domain Adaptation using Random Forests** (SHDA-RF) [29]: Similar to SHFR, SHDA-RF is another supervised FSR transfer approach that leverages the shared label space to generate cross-domain correspondences. After estimating the cross-domain correspondences from the random forest models trained on the source and target data, we fine tune the regularizer λ and report the best results over 16-folds. Both SHDA and SHFR estimate corresponding data by leveraging the common labels across the domains. For tasks with heterogeneous outputs, we vary the label-similarity threshold ($W_{ij} \in [0.3, 1]$) to generate the correspondences.
- **Domain Adaptation using Manifold Alignment** (DAMA) [33]: DAMA is a LST approach that learns a low-dimensional embedding where the projected data from the source and target domain is aligned using labels while preserving the original

[1] `http://www.svms.org/parameters/`

structure. We vary the hyper-parameter β to capture the trade-off between preserving the original topology and label-induced alignment in the shared subspace to get the best results. NGD is used to calculate the similarity of the labels across the domains (instead of 0/1 value used in the paper). The optimal size of the shared subspace was determined through experimentation on the validation set.

- **Co-regularized Heterogeneous Transfer Learning** (Co-HTL) [34]: Similar to DAMA, Co-HTL is another LST approach that also leverages label relationships to co-align source and target data in a common space. Here too, NGD is used to calculate the similarity between the labels of the domains (instead of the using the divergence between the label-embeddings obtained from word2vec [19] model or the divergence between topic distribution vectors obtained via Latent Dirichlet Allocation [2]). The optimal size of the shared subspace was determined through experimentation on the validation set.

We consider three heterogeneous transfer tasks to compare the performance of WI-HTLSS against the aforementioned algorithms.

- **Cross-lingual text transfer**: We evaluate the cross-lingual transfer performance on two benchmark text datasets namely, 1) Amazon Cross-Lingual Sentiment (Amazon CLS) dataset[2] [18] and 2) Reuters Multi-Lingual dataset[3] [1]. We utilize the same experimental setting as mentioned by Zhou et al. [40] to test the performance of the proposed algorithm on cross-lingual sentiment and text classification tasks.
- **Cross-domain activity recognition**: We picked three single-resident CASAS[4] [7] datasets for this transfer task. We follow the same procedure as mentioned by Sukhija et al. [29] to construct the smart home datasets from raw sensor records. The labeled data from one smart home is treated as the source domain and limited labeled sensor data in another serves as the target domain. We employ the same experimental setting as mentioned by Sukhija et al. to evaluate six cross-domain activity recognition tasks.
- **Deep Representation Transfer**: In this experiment, the CIFAR-100[5] [15] image representations obtained from the last fully connected layer (4096 features) of the VGG-19 model [24] (pre-trained on the Imagenet dataset[6]) act as the target domain whereas the related Imagenet[7] [8] image representations acquired from the last fully connected layer (4096 features) of the pre-trained VGG-16 model[8] act as the source domain. We perform dimensionality reduction using Principal Component Analysis (PCA) [39] to preserve 60% variance on the obtained source and target image representations. In every fold, for a source-target pair, we pick 100

[2] `https://www.uni-weimar.de/en/media/chairs/computer-science-department/webis/data/corpus-webis-cls-10/`

[3] `https://archive.ics.uci.edu/ml/machine-learning-databases/00259/`

[4] `http://ailab.wsu.edu/casas/datasets/`

[5] `https://www.cs.toronto.edu/~kriz/cifar.html`

[6] `https://keras.io/applications/#vgg19`

[7] `http://www.image-net.org/`

[8] `https://keras.io/applications/#vgg16`

images per label to form the source domain data and 5 images per label as the target training data. The optimal value of the hyper-parameters is determined using a disjoint validation set (100 images per label). We report the classification error on the test set[9] consisting of 100 images per label. We repeat this process over 16 different folds and report the mean error and standard deviation on the following (Source→Target) transfer tasks:

1. Imagenet(321-326, 118-121, 125) → CIFAR-100(Butterfly-14, Crab-26)
2. Imagenet(52-68, 33-37) → CIFAR-100(Snake-78, Turtle-93)

The numbers within the brackets for the Imagenet(.) dataset are the label identities[10] corresponding to the human readable labels.

5 Results and Discussion

We report and analyze the performance of several state-of-the-art transfer algorithms against WIHTLSS in three HTL scenarios, namely 1) when the source and target domain are characterised by heterogeneous feature spaces but share the same label space, 2) with heterogeneous feature spaces but an overlapping label space and 3) when the labels across the domains are also heterogeneous.

The cross-lingual transfer experiments belong to the first scenario where the vocabulary differences in the source and target documents lead to heterogeneous feature spaces. However, the same set of labels are shared between the domains. The cross-domain activity recognition experiments correspond to the second HTL scenario due to an overlapping set of activity labels across the source and target smart homes. However, the daily activities of different smart home residents lead to semantically related label spaces. The deep representation transfer tasks belong to the third scenario. With respect to a binary target task (distinguishing 2 classes from the CIFAR-100 dataset), we pick images with distinct but related labels as the source (related images from the Imagenet) to analyze the impact of leveraging semantic label relationships determined through NGD to align heterogeneous representations obtained from deep models.

The reason for learning a linear transformation stems from the characteristics of these transfer tasks. As we are trying to map bag-of-words representation for cross-lingual transfer tasks, the relationship between words (that are synonymous) within a language and across different languages tends to be linear [40]. Similarly for cross-domain activity recognition tasks, the feature values of the sensors (in a functional area such as kitchen, bedroom etc.) across different smart homes appear to be linearly related.

The baseline and transfer results on all transfer tasks are shown in Table 1. Since the baseline random forest (BRF) performed significantly better than classical linear SVM on most transfer tasks, we kept random forest as the final model for all transfer approaches. For cross-lingual transfer tasks on Amazon CLS dataset and Reuters Multilingual dataset, the notable performance improvement of the transfer approaches over the baseline BRF indicates the feasibility of cross-lingual knowledge transfer. It

[9] https://www.cs.toronto.edu/~kriz/cifar-100-matlab.tar.gz

[10] https://gist.github.com/yrevar/942d3a0ac09ec9e5eb3a

Table 1: Performance comparison of state-of-the-art algorithms on cross-lingual tasks, cross-domain activity recognition tasks and deep representation transfer tasks is shown in terms of mean error and standard deviation (%) over 16 folds. The best performance has been highlighted in bold and statistically significant WIHTLSS results against the second best performing algorithm are indicated by * respectively.

Amazon CLS dataset							
S→T	**BRF**	**SVM-ECOC**	**DAMA**	**SHDA-RF**	**SHFR-RF**	**Co-HTL**	**WIHTLSS**
English→French	43.90±2.34	51.83±3.59	43.28±1.39	37.01±3.41	38.59±3.57	34.18±3.65	**32.05±2.67***
English→German	44.21±2.71	50.96±4.01	42.82±1.26	33.09±3.93	37.51±2.25	34.45±2.48	**30.55±2.27***
English→Japanese	49.0±3.47	52.91±4.60	47.75±1.49	34.58±4.89	37.79±3.7	35.08±2.71	**31.65±2.98***
Reuters Multilingual Dataset							
S→T	**BRF**	**SVM-ECOC**	**DAMA**	**SHDA-RF**	**SHFR-RF**	**Co-HTL**	**WIHTLSS**
English→Spanish	32.13±3.59	36.44±4.30	31.84±1.36	25.62±1.15	27.93±1.03	25.34±1.19	**22.96±1.10***
French→Spanish			30.48±1.48	24.61±0.85	26.52±0.90	25.20±1.0	**22.47±1.23***
German→Spanish			30.73±1.4	25.1±0.95	27.01±1.23	24.54±1.03	**22.32±1.08***
Italian→Spanish			31.59±1.82	23.57±1.35	24.86±1.47	22.82±1.28	**21.03±1.46***
CASAS horizon house (hh) datasets							
S→T	**BRF**	**SVM-ECOC**	**DAMA**	**SHDA-RF**	**SHFR-RF**	**Co-HTL**	**WIHTLSS**
hh102→hh113	29.67±2.51	34.58±2.20	31.98±2.93	26.37±2.23	27.98±2.51	27.75±2.55	**23.71±2.36***
hh102→hh118	36.41±2.42	43.51±3.01	37.72±2.52	31.50±2.01	32.01±2.09	30.54±2.76	**26.85±2.40***
hh113→hh102	36.70±1.95	41.23±2.93	36.21±2.28	29.88±1.76	32.81±2.38	31.98±2.44	**27.02±2.11***
hh113→hh118	32.35±2.56	39.41±1.89	34.49±2.42	28.0±2.02	30.15±2.44	29.33±2.01	**25.75±1.97***
hh118→hh102	38.95±2.57	41.80±2.21	39.55±2.59	33.02±2.13	35.40±2.35	36.0±2.45	**30.97±2.18***
hh118→hh113	31.01±1.80	34.73±3.39	31.81±2.09	27.73±1.08	30.46±2.95	30.81±2.39	**26.45±2.2***
Imagenet (VGG-16)→ CIFAR-100(VGG-19)							
Target Task	**BRF**	**SVM-ECOC**	**DAMA**	**SHDA-RF**	**SHFR-RF**	**Co-HTL**	**WIHTLSS**
Butterfly v/s Crab	31.16±5.61	27.78±6.61	30.05±2.95	24.71±3.40	25.20±3.28	26.09±3.84	**22.05±2.26***
Snake v/s Turtle	21.81±8.27	18.53±6.14	20.62±4.34	15.92±3.67	16.48±3.53	16.80±4.23	**13.24±2.15***

can be observed that the proposed framework WIHTLSS significantly outperforms (p-value$<$0.05) SHFR-RF by 3.5-7%, SHDA-RF by 2.5-3%, DAMA by 7-15% and Co-HTL by 1.5-3.5% in every cross-lingual transfer setting.

The experimental results on CASAS datasets are also depicted in Table 1. It can be seen that WIHTLSS outperforms all the other approaches on the CASAS datasets. Also, the performance of feature space remapping (FSR) approaches namely SHDA-RF and SHFR-RF is significantly better than DAMA. Among the FSR approaches, WIHTLSS outperforms the best performing algorithms, SHDA-RF and Co-HTL, by 3% on average (p-value$<$0.05). Even for the deep representation transfer tasks, the proposed algorithm significantly outperforms (p-value$<$0.05) all the baseline and transfer approaches.

We investigate the impact of having more labeled data in the target domain on the transfer performance of WIHTLSS. Due to lack of space, we show the results on a subset of the transfer tasks. The experimental results in Figure 1 illustrate that having more labeled data in the target domain yields better transfer performance for all approaches. It can be observed that the transfer improvement is significant even when there are just 10 or 50 samples per label in the target domain. However, the utility of transfer reduces beyond certain point at which there are sufficient labeled instances in the target domain to directly learn a model. Similar trend was observed for other tasks as well.

5.1 Impact of the hyper-parameters

We first perform a grid search on the hyper-parameter β (while keeping $\lambda = 0, \kappa = 0$) to assess the effectiveness of label-similarity induced alignment on the transfer per-

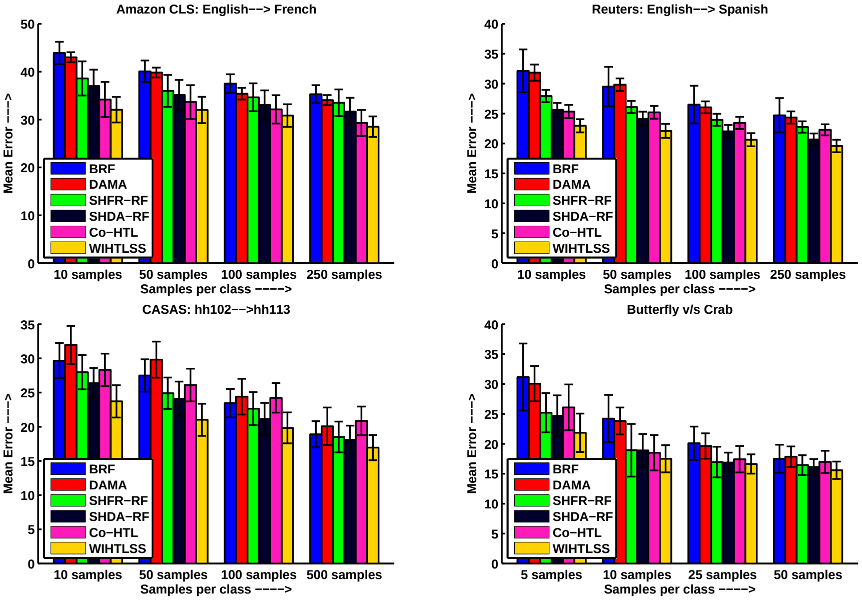
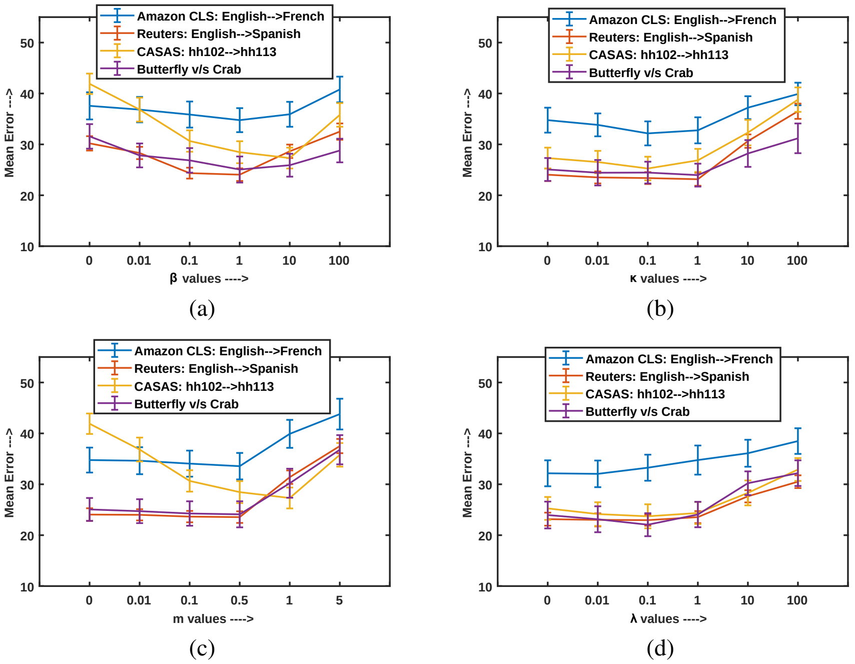


Fig. 1: Impact of the quantity of available labeled data in the target domain on the performance of baseline and transfer approaches.

formance of WIHTLSS. From Figure 2(a), it is evident that aligning data of the domains in the target space by leveraging semantic relationships between heterogeneous labels is beneficial. When $\beta = 0$, the performance is significantly poor in comparison to the baseline for all the transfer tasks. This suggests that only preserving the original structure of data in the target space does not guarantee a reasonable alignment where instances of related classes are closer than unrelated classes. We hypothesized to bring the instances with same or related labels closer to each other for learning a better transformation. It can be observed that there is substantial improvement in the transfer performance with increase in the importance of β. This indicates that leveraging the label similarities in the proposed formulation has a positive influence on the transfer performance of WIHTLSS. However, overemphasizing the importance of label-space induced alignment has a detrimental effect on the transfer performance as the learned transformation becomes biased towards the limited labeled target data (which does not necessarily represent the true target distribution). Further, replacing NGD based similarity measure with divergence from the word2vec embeddings[19] did not result in significant change in the performance.

After obtaining optimal β, we vary κ to investigate the importance of the 3rd term G. For this experiment, we fix the margin $m = 1$, keep $\lambda = 0$ and use the optimal β for D. It can be observed from Figure 2(b) that the transfer performance improves with increase in importance of G. This suggests that pushing dissimilar instances apart helps in learning a better transformation. However, over-regularization of this label-dissimilarity induced alignment worsens the transfer performance as it negatively affects the original



(a)

(b)

(c)

(d)

Fig. 2: (a) Impact of the hyper-parameters β (label-similarity induced semantic alignment), (b) κ (label-dissimilarity induced alignment), (c) λ (regularizer term) and (d) m (margin) on the performance of WIHTLSS.

structure. By fixing the optimal value of κ, we then determine the optimal margin (Refer Figure 2(c)). Our hypothesis here is that with low margin, dissimilar source instances will continue to remain close to the target instances resulting in poorer performance when compared to larger values of the margin. This can be observed in Figure 2(c). The decrease in the performance for smaller values of margin is low for the cross-lingual and cross-domain activity recognition tasks, but significant in the deep representation transfer task. Similarly, beyond the optimal value of the margin, pushing away the dissimilar instances affects the distribution of the instances in the target space significantly, resulting in performance degradation. Finally, we fine tuned the regularizer term by varying the hyper-parameter λ. It can be observed from Figure 2(d) that WIHTLSS performs the best, on average, when λ is set to 0.1.

Figure 3 depicts the impact of varying the number of nearest neighbors (k) on the transfer performance of WIHTLSS. It can be observed that the transfer performance improves with increase in the number of nearest neighbors. This result validates our hypothesis that the transformed source instances that were the nearest neighbors of any target instance are actually relevant for the target task. However, when we add more neighbors beyond a certain point, transformed source data points that are quite far also contribute towards learning the model. Apart from the deep transfer task (butterfly v/s crab), in the extreme case, when all the source instances are considered, it can be observed that the transfer performance suffers for all other tasks.

In our deep representation transfer experiment, for the binary target task (butterfly v/s crab), we used only related image representations as the source i.e. the source image representations only contain the encodings of different types of butterfly and crab. Hence, all the image representations of the source are actually relevant for the binary target task. This is in agreement with our observations as well (Refer Figure 3).

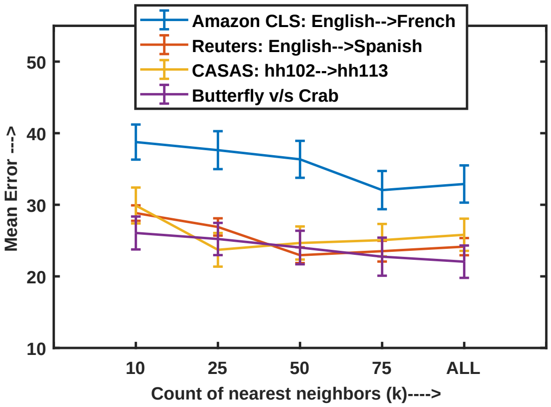


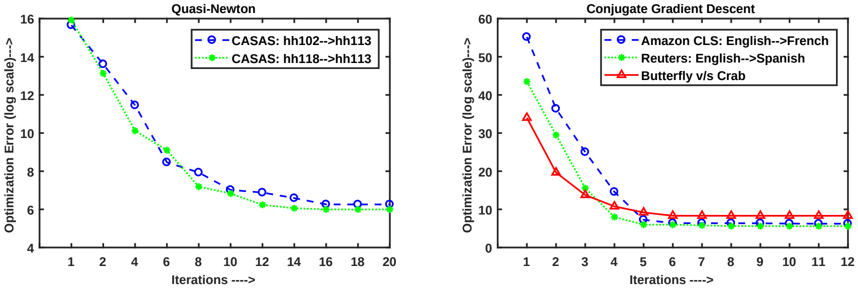
Fig. 3: Depicts the variation in the transfer performance (mean error and standard deviation) with increasing the nearest neighbors.

5.2 Convergence Analysis

The convergence results for the optimization routines are shown in Figure 4. With quasi-newton method for cross-domain activity recognition experiments, the optimization error decreases gradually over iterations and convergence is achieved within 15-20 iterations (on average). However, with CGD algorithm on cross-lingual transfer experiments, the error drops sharply in the first few iterations and convergence is reached inside 5-10 iterations.

6 Summary and Future Work

WIHTLSS is an integration of collective inference and FSR based transfer learning approach that bridges domains with heterogeneous feature and label spaces without relying on instance or feature correspondences. Assuming some semantic relationships within and across the label spaces, it utilizes web knowledge to align the source data in the target space while preserving the original structure. It can be viewed as a simple version of dictionary learning without orthogonality or sparsity constraints but using L2 norm constraints on the dictionary and the new representation. The label-induced



(a) (b)

Fig. 4: Convergence with (a) quasi-newton algorithm on two cross-domain activity recognition tasks and (b) CGD on two cross-lingual transfer tasks and a deep representation transfer task

alignment terms are the modified Laplacian regularization terms where the adjacency matrix is calculated based on the NGD. The experimental results on real-world HTL tasks with identical and different output spaces indicate the superiority of WIHTLSS over state-of-the-art supervised transfer approaches.

A limitation of the proposed approach is that it requires some amount of labeled data in the target domain. Consequently, if labeled data is absent in the target domain, annotating relatively small number of unlabeled data becomes an inescapable task. Besides leveraging the semantic label relationships to learn the transformation, one can investigate the possibility of leveraging unlabeled domain knowledge or the task associated characteristic properties to determine a quantifiable measure of the relatedness W. Further, the proposed algorithm utilizes a single source domain for knowledge transfer, as part of future work, we wish to explore how to effectively combine data from multiple related modalities simultaneously to make an improved final prediction on the target.

Acknowledgment

This research is supported by the Department of Science and Technology, India under grant YSS/2015/001206, and by the Indian Institute of Technology Ropar under the ISIRD grant.

References

1. Amini, M., Usunier, N., Goutte, C.: Learning from multiple partially observed views-an application to multilingual text categorization. In: Advances in Neural Information Processing Systems. pp. 28–36 (2009)
2. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. Journal of machine Learning research **3**(Jan), 993–1022 (2003)
3. Borchani, H., Varando, G., Bielza, C., Larrañaga, P.: A survey on multi-output regression. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery **5**(5), 216–233 (2015)

4. Branch, M.A., Coleman, T.F., Li, Y.: A subspace, interior, and conjugate gradient method for large-scale bound-constrained minimization problems. SIAM Journal on Scientific Computing **21**(1), 1–23 (1999)
5. Breiman, L.: Random forests. Machine learning pp. 5–32 (2001)
6. Cilibrasi, R.L., Vitanyi, P.M.B.: The google similarity distance. IEEE Transactions on Knowledge and Data Engineering **19**(3), 370–383 (2007)
7. Cook, D.J., Crandall, A.S., Thomas, B.L., Krishnan, N.C.: Casas: A smart home in a box. Computer **46**(7), 62–69 (2013)
8. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on. pp. 248–255. IEEE (2009)
9. Efron, B., Hastie, T., Johnstone, I., Tibshirani, R., et al.: Least angle regression. The Annals of statistics **32**(2), 407–499 (2004)
10. Hadsell, R., Chopra, S., LeCun, Y.: Dimensionality reduction by learning an invariant mapping. In: Computer vision and pattern recognition, 2006 IEEE computer society conference on. vol. 2, pp. 1735–1742. IEEE (2006)
11. He, J., Liu, Y., Yang, Q.: Linking heterogeneous input spaces with pivots for multi-task learning. In: Proceedings of the SIAM International Conference on Data Mining. pp. 181–189 (2014)
12. Hu, D.H., Yang, Q.: Transfer learning for activity recognition via sensor mapping. In: Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence. pp. 1962–1967. IJCAI'11, AAAI Press
13. Hubert Tsai, Y.H., Yeh, Y.R., Frank Wang, Y.C.: Learning cross-domain landmarks for heterogeneous domain adaptation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5081–5090 (2016)
14. van Kasteren, T.L.M., Englebienne, G., Kröse, B.J.A.: Transferring knowledge of activity recognition across sensor networks. In: Proceedings of the 8th International Conference on Pervasive Computing. pp. 283–300 (2010)
15. Krizhevsky, A., Hinton, G.: Learning multiple layers of features from tiny images (2009)
16. Kyrillidis, A., Zouzias, A.: Non-uniform feature sampling for decision tree ensembles. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing. pp. 4548–4552 (2014)
17. Li, W., Duan, L., Xu, D., Tsang, I.W.: Learning with augmented features for supervised and semi-supervised heterogeneous domain adaptation. IEEE Transactions on Pattern Analysis and Machine Intelligence pp. 1134–1148 (2014)
18. Lichman, M.: UCI machine learning repository (2013), `http://archive.ics.uci.edu/ml`
19. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Advances in neural information processing systems. pp. 3111–3119 (2013)
20. Pan, S.J., Yang, Q.: A survey on transfer learning. IEEE Transactions on Knowledge and Data Engineering **22**(10), 1345–1359 (2010)
21. Prettenhofer, P., Stein, B.: Cross-language text classification using structural correspondence learning. In: Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics. pp. 1118–1127. ACL '10 (2010)
22. Rajan, S., Ghosh, J.: An empirical comparison of hierarchical vs. two-level approaches to multiclass problems. In: Multiple Classifier Systems, pp. 283–292. Lecture Notes in Computer Science (2004)
23. Rosenstein, M.T., Marx, Z., Kaelbling, L.P., Dietterich, T.G.: To transfer or not to transfer. In: NIPS 2005 Workshop on Transfer Learning. vol. 898 (2005)

24. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
25. Strange, H., Zwiggelaar, R.: A generalised solution to the out-of-sample extension problem in manifold learning. In: AAAI. pp. 293–296 (2011)
26. Sukhija, S.: Label space driven feature space remapping. In: Proceedings of the ACM India Joint International Conference on Data Science and Management of Data. pp. 310–313. CoDS-COMAD '18 (2018)
27. Sukhija, S.: Label space driven heterogeneous transfer learning with web induced alignment. In: Proceedings of the 32nd AAAI Conference on Artificial Intelligence (2018)
28. Sukhija, S., Krishnan, N.C., Kumar, D.: Supervised heterogeneous transfer learning using random forests. In: Proceedings of the ACM India Joint International Conference on Data Science and Management of Data. pp. 157–166 (2018)
29. Sukhija, S., Krishnan, N.C., Singh, G.: Supervised heterogeneous domain adaptation via random forests. In: Proceedings of the International Joint Conference on Artificial Intelligence. pp. 193–200 (2016)
30. Wang, C., Mahadevan, S.: Manifold alignment using procrustes analysis. In: Proceedings of the International Conference on Machine learning. pp. 1120–1127 (2008)
31. Wang, C., Mahadevan, S.: A general framework for manifold alignment. In: Proceedings of the AAAI Fall Symposium: Manifold Learning and Its Applications. pp. 53–58 (2009)
32. Wang, C., Mahadevan, S.: Manifold alignment without correspondence. In: Proceedings of the International Joint Conference on Artificial Intelligence. pp. 1273–1278 (2009)
33. Wang, C., Mahadevan, S.: Heterogeneous domain adaptation using manifold alignment. In: Proceedings of the International Joint Conference on Artificial Intelligence. vol. 22, pp. 1541–1547 (2011)
34. Wei, Y., Zhu, Y., Leung, C.W.k., Song, Y., Yang, Q.: Instilling social to physical: Co-regularized heterogeneous transfer learning. In: Proceedings of the AAAI National Conference on Artificial Intelligence. pp. 1338–1344 (2016)
35. Xiao, M., Guo, Y.: Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2015, Porto, Portugal, September 7-11, 2015, Proceedings, Part II, chap. Semi-supervised Subspace Co-Projection for Multi-class Heterogeneous Domain Adaptation, pp. 525–540. Springer International Publishing (2015)
36. Yan, Y., Li, W., Ng, M., Tan, M., Wu, H., Min, H., Wu, Q.: Learning discriminative correlation subspace for heterogeneous domain adaptation. In: Proceedings of the 26th International Joint Conference on Artificial Intelligence. pp. 3252–3258. AAAI Press (2017)
37. Yang, Q., Chen, Y., Xue, G.R., Dai, W., Yu, Y.: Heterogeneous transfer learning for image clustering via the social web. In: Proceedings of the International Joint Conference on Natural Language Processing. pp. 1–9 (2009)
38. Zhou, G., He, T., Wu, W., Hu, X.T.: Linking heterogeneous input features with pivots for domain adaptation. In: Proceedings of the 24th International Conference on Artificial Intelligence. pp. 1419–1425. IJCAI'15, AAAI Press (2015)
39. Zhou, J.T., Pan, S.J., Tsang, I.W., Yan, Y.: Hybrid heterogeneous transfer learning through deep learning. In: Proceedings of the AAAI National Conference on Artificial Intelligence. pp. 2213–2219 (2014)
40. Zhou, J.T., Tsang, I.W., Pan, S.J., Tan, M.: Heterogeneous domain adaptation for multiple classes. In: Proceedings of the International Conference on Artificial Intelligence and Statistics. pp. 1095–1103 (2014)
41. Zhu, Y., Chen, Y., Lu, Z., Pan, S.J., Xue, G.R., Yu, Y., Yang, Q.: Heterogeneous transfer learning for image classification. In: Proceedings of the AAAI National Conference on Artificial Intelligence. pp. 1304–1310 (2011)