Supervised Heterogeneous Domain Adaptation via Random Forests

Sanatan Sukhija[1], **Narayanan C Krishnan**[1], **Gurkanwal Singh**[2*]
[1]Department of Computer Science and Engineering, Indian Institute of Technology Ropar, Punjab, India sanatan@iitrpr.ac.in, ckn@iitrpr.ac.in
[2]Department of Computer Science and Engineering, PEC University of Technology, Chandigarh, India gurkanwal.singh7@gmail.com

Abstract

Heterogeneity of features and lack of correspondence between data points of different domains are the two primary challenges while performing feature transfer. In this paper, we present a novel supervised domain adaptation algorithm (SHDA-RF) that learns the mapping between heterogeneous features of different dimensions. Our algorithm uses the shared label distributions present across the domains as pivots for learning a sparse feature transformation. The shared label distributions and the relationship between the feature spaces and the label distributions are estimated in a supervised manner using random forests. We conduct extensive experiments on three diverse datasets of varying dimensions and sparsity to verify the superiority of the proposed approach over other baseline and state of the art transfer approaches.

1 Introduction

The key to success of many supervised learning algorithms is the availability of abundant labeled training data. However, for many real-world problems, collecting labeled data is often very expensive and cumbersome. Transfer learning algorithms help to overcome the scarcity of labeled data in a domain (often referred to as the target domain) by utilising information about the task, and data from single or multiple auxiliary domains (referred to as source domains). Transfer learning approaches have found success in many applications including activity recognition [Hu *et al.*, 2011; Cook *et al.*, 2013c], sentiment classification [Zhou *et al.*, 2014], document analysis and indoor localization [Pan and Yang, 2010].

A popular setting for performing transfer is when the source and the target domains are represented by the same set of features. The goal in this setting is to minimise the differences in the data distribution of the source and target domains. However, for applications such as sentiment analysis across different languages [Pan, 2010], and activity recognition across different domains [Cook *et al.*, 2013b], the source and target data are represented using heterogeneous features of different dimensions that may or may not overlap. Transfer learning for such heterogeneous domains can be performed by first bridging the gap between the features characterising the different domains. This is the principle behind feature-based transfer learning approaches.

The feature-based transfer approach proposed in this paper is motivated by the application of activity recognition in a smart home. Smart home based activity recognition deals with learning the daily activities of smart home resident(s), captured through a series of sensor observations. Transfer learning algorithms can be used to overcome the scarcity of labeled data of a new target smart home by utilising the labeled data of other source smart homes. However, different layouts and types of sensors deployed at different places lead to heterogeneous feature spaces [Hu and Yang, 2011] necessitating transfer methodologies. Figure 1 illustrates the layout and sensor locations for three smart homes from the CASAS datasets [Cook *et al.*, 2013c] used in this paper. Given only a few labeled instances in the target we leverage the common labels in the source and target domains to derive the relationship between the corresponding feature spaces. The key assumption of our algorithm is that features in both source and target domains that characterise data partitions with similar label distribution, must be related to each other. The shared label distributions across the two domains act as the pivot for learning the mapping between the feature spaces. The generated sparse mapping represents a target feature as a linear combination of source features. This mapping is estimated without assuming any correspondence between source and target data points.

1.1 Problem Definition

Let $\{X_S, Y_S\}_{i=1}^{m}$ and $\{X_T, Y_T\}_{j=1}^{n}$ represent the set of labeled instances in the source domain S and target domain T respectively, where $m \ggg n$. $x_S \in \mathbb{R}^{d_S}$ is a source data point with $y_S \in \mathcal{Y}$ the corresponding class label. Similarly, $x_T \in \mathbb{R}^{d_T}$ is a target data point and $y_T \in \mathcal{Y}$ is its associated label. The features that describe x_S and x_T are completely different and $d^S \neq d^T$. However, we assume that the source and target domains share a common label space. Let the number of shared labels be k. Our goal is to learn a mapping $f : \mathbb{R}^{d_S} \rightarrow \mathbb{R}^{d_T}$ such that the data from the source domain can be mapped to the target domain. This mapped source data

*The author contributed to this work during his internship at IIT Ropar.

Figure 1: Layouts of the three CASAS smart homes that differ in terms the layout, and count of the sensors deployed. The black squares represent the location of sensors.

can then be used in conjunction with the target data to learn the hypothesis $h : \mathbb{R}^{d_T} \rightarrow \mathcal{Y}$.

1.2 Contributions

The contributions of this paper can be summarised as follows:

1. The proposed algorithm yields a heterogeneous feature-space class-invariant mapping, assuming no correspondence between the data-points of the domains that share no overlapping features.
2. Our algorithm does not require the computation of an optimal code matrix for error correcting output code, which is a challenging task, that is a requirement for the current supervised state of the art feature transfer algorithm [Zhou *et al.*, 2014]. The proposed algorithm utilises naturally occurring label distributions at leaf nodes of a decision tree model as pivots to generate the mapping $P_S \in \mathbb{R}^{d_S \times d_T}$.
3. The experiments conducted on diverse datasets indicate the effectiveness of the algorithm even if very few labeled instances are available in the target domain.

2 Related Work

Bridging features across heterogeneous spaces for domain adaptation is a challenging problem. The approaches for heterogeneous domain adaptation can be broadly split into two categories based on the type of mapping learned, namely, *Feature Remapping* and *Latent Space Transformation*.

Feature Remapping approaches determine the transformation for converting source features to target features or vice-versa. It can in turn be of two types: one in which there is an explicit correspondence between the individual features of source and target domain such as the i^{th} source feature being mapped to the j^{th} target feature, and second in which a source or target feature is represented as a combination (often linear) of features from the other domain. Approaches for one-to-one source to target feature remapping have used genetic algorithms and other greedy methods to obtain an optimal mapping, using classification accuracy as the performance measure [Feuz and Cook, 2014]. Alternate approaches rely on domain independent features known as pivots that can be utilised to align the feature spaces [Blitzer *et al.*, 2006; He *et al.*, 2014]. These approaches partition the features of a domain into independent and dependent sets. The domain independent features are present across different domains, while the dependent features are specific to a domain. The goal is to learn the mapping between the dependent features by using the independent features. Spectral clustering algorithm can be used to obtain the feature-clusters from the co-aligned bipartite graph of domain independent and dependent features that acts as the common subspace [Pan, 2010]. In the absence of explicit domain independent features, statistical properties of domain specific features can be used to derive meta features to bridge the domains [Feuz, 2014]. A recent work on feature remapping for feature transfer constructs a class-invariant sparse transformation matrix by mapping the weight vectors of SVM classifier trained on labeled data from the domains [Zhou *et al.*, 2014]. Synthetically generated error correcting output codes (ECOC) are used to train the SVM model so as to estimate accurate transformations. We compare the performance of our algorithm against this supervised heterogeneous feature remapping approach (SHFR-ECOC).

Latent Space Transformation approaches determine transformations to project the data of different domains onto a common latent space. Specifically, these approaches compute two projection matrices P_S and P_T for source and target domains respectively, such that the difference between the projected source space $B_S : P_S \times X_S$ and projected target space $B_T : P_T \times X_T$ is minimised while trying to preserve the characteristics of the original feature spaces. Approaches in this category can be summarised as unsupervised Eigen transfer frameworks. Manifold alignment based approaches [Wang and Mahadevan, 2009] determine the transformed space by making the manifold assumption, where the mapping brings data distributions of the domains closer to each other while preserving the local geometric structure and maximising the alignment. However, these approaches only work for data that exhibit strong manifold property and therefore are not applicable to other scenarios. Heterogeneous spectral mapping (HeMap) [Shi and Yu, 2012] optimises the difference in the latent space in a general setting by learning two transformation matrices using spectral embedding without using any label information. The algorithm implicitly tries to discover the correspondence between the data points of source and target domain through an optimisation framework. Often, in situations where there is no explicit data correspondences, the recovered transformations are noisy. Since these approaches directly estimate the projected data, estimating the projection for out-of-sample data is a challenging problem. We compare the performance of our algorithm against the HeMap approaches that compute linear and non-linear transformations.

In the context of smart home activity recognition, heterogeneous layouts can be manually unified by defining a common meta-feature space that is shared by all smart apartments. These meta-features can be manually specified by a domain expert or can be derived from structural, temporal, spatial or

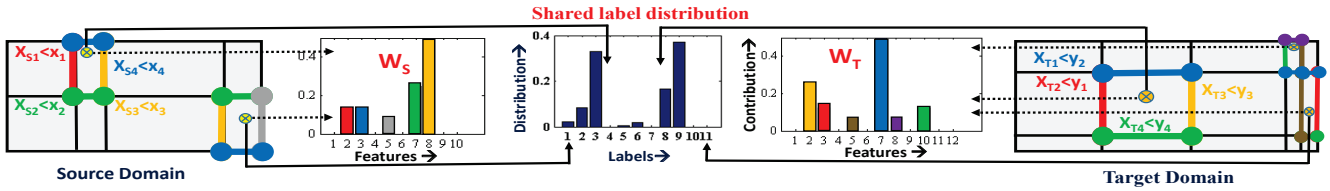

Figure 2: Illustration of the relationship between the features of the two domains based on a single pivot label distribution.

functional similarities of original features [Rashidi and Cook, 2010]. A manual mapping is not optimal and a poor mapping can drastically hinder the recognition performance of the model. The proposed algorithm intends to exploit label space distributions of data partitions to identify a mapping across the domains for addressing the problem of heterogeneous domain adaptation.

3 Proposed Methodology

The task of determining 'common' features between heterogeneous source and target feature spaces for knowledge transfer is a challenging problem [Li *et al.*, 2014]. Our novel solution to this problem leverages the common label information between the source and target domains as the pivot for knowledge transfer. The proposed algorithm determines the mapping P_S between source and target features based on the estimate of the contribution of the features towards creating data partitions having similar label distributions.

3.1 Estimating Pivots Across the Domains

The first step in our proposed approach is to derive the pivots that are used to construct the bridge across heterogeneous feature spaces. We define the pivots in terms of the shared labels between the source and target domains. In the simplest scenario each shared label is a pivot. When the number of shared labels between the domains is small, learning the feature mapping is a challenging problem [Zhou *et al.*, 2014]. The SHFR ECOC approach overcomes this limitation by using synthetically generated error-correcting output codes (ECOC) for representing each shared label. It is desirable that the class labels are independent as the relationships between the different labels are not effectively captured due to the randomness of the ECOC generation process [Rajan and Ghosh, 2004]. Thus selecting the optimal code matrix is a challenging problem for SHFR ECOC. Our approach overcomes this limitation by relying on naturally occurring label distributions in the complex data space.

To arrive at these label distributions, our approach looks at the leaf nodes of a decision tree modeled on the dataset. A decision tree follows a greedy strategy to recursively partition the data based on some feature value test. A leaf node in a decision tree holds a distribution L of class labels that is associated with a set of data partition. A path in a decision tree from the root to a leaf node contains a sequence of features chosen as split functions. The candidate split at a node involves a locally optimal partitioning based on some metric like Gini Impurity, Information Gain or Gain Ratio.

To ensure a sufficient number of pivotal label distributions for learning the mapping between the domains, we train a random forest, which also helps to reduce overfitting. Every tree in the forest is constructed using a random subset of features [Breiman, 2001]. Our algorithm first constructs n_s and n_t trees from S and T respectively. Every path in a decision tree leading to a partition of data is associated with a certain label distribution. Label distributions that appear both in the source and target random forest models are the pivots that are used for bridging the two domains.

3.2 Estimating Feature Relationships W_S and W_T

The next step in the algorithm computes the relationship matrices W_S and W_T between the domain dependent features and the shared pivots. Since our pivots are label distributions, we define this relationship as the contribution of the domain dependent features towards creating the pivot label distribution. This relationship can be easily extracted from the decision tree structure. The boundary of the data partition at the leaf node can be identified with the feature splits along the path to the leaf node from the root of the tree.

A simple approach to compute W_S and W_T would be to give equal importance to all the features that were used as split nodes in the path. Thus for a path, the i^{th} entry in the corresponding feature relationship vector would contain the frequency of the i^{th} feature getting selected as a split node. Another approach would be to give higher priority to a feature used at parent node compared to the features chosen as split nodes at its descendants. For every path, each entry in the feature contribution vector is given by $\sum_{i=1}^{c}(1/2)^{v(i)}$ where $v(i)$ denotes the decision tree depth at which the split was made and c represents the frequency of the feature being used as a candidate split in the path.

In practice, it is common to have duplicate label distributions at leaf nodes, i.e., different data partitions corresponding to the same label distribution. The feature contribution vectors for these data partitions are averaged. Thus at the end of this process, for each shared pivotal label distribution between S and T, we also have the domain dependent feature relationships to these pivots. Based on the similar source and target class label distributions, the estimated feature contribution matrices $W_S \in \mathbb{R}^{N_p \times d_S}$ and $W_T \in \mathbb{R}^{N_p \times d_T}$ are mapped to yield the source projection matrix P_S, where N_p is the number of pivots. This process is illustrated in Figure 2. The advantage with using a random forest model is that the pivots and the relationship between the domain dependent features and pivots across source and target can be estimated from a single model reducing the complexity of the transfer approach.

Algorithm 1 Supervised HDA via Random Forests (SHDA-RF)

Input: Source data: $S \in \mathbb{R}^{M \times d_S}$ and Target data: $T \in \mathbb{R}^{N \times d_T}$
Output: $P_S \in \mathbb{R}^{d_S \times d_T}$

1. Build a random forest with n_s trees from source features X_S.
2. For every path from the root to a leaf node in a tree, the contribution of a feature is estimated as $W(x_S(v)) = W(x_S(v)) + (1/2)^v$ where v denotes the level at which the feature $x \in X_S$ was selected as a candidate split. The corresponding label distribution L is acquired from the leaf node.
3. Similarly construct the target features contribution matrix W_T using n_t trees created from T.
4. Remove duplicates from L_S and L_T. For every duplicate entry in L_S and L_T, the corresponding feature vector entries in W_S and W_T are averaged.
5. Return the corresponding W_S and W_T for the identical class label distributions.
6. The mapping P_S can be obtained by running LASSO d_T times on obtained W_S and W_T from Step 5.

Table 1: Summary of CASAS-HH datasets.

Dataset	Feature count	Activity count
hh102	43	29
hh113	48	30
hh118	44	32

3.3 Deriving the Feature Transformation

The last step in our algorithm derives a sparse transformation P_S between the two domains. Our objective is to represent each target feature as a linear combination of a small set of source features. The Least Absolute Shrinkage and Selection Operator (**LASSO**) is used to learn P_S from W_S and W_T. It is defined as:

$$\min_{P_S} \quad \frac{1}{N_p} \sum_{i=1}^{N_p} \parallel W_T - W_S P_S \parallel_2^2 + \sum_{i}^{d_T} \lambda_i \parallel P_{S_i} \parallel_1,$$
$$s.t. \quad P_{S_i} \geq 0$$

The first part of the optimisation problem minimises the difference between the projected source feature contribution matrix $P_S \times W_S$ and target feature contribution matrix W_T. The second part is the L_1 regularisation term to obtain a sparse transformation matrix. The regularisation parameter λ controls the size of this subset. There are d_T minimisation problems that are solved using Least Angle Regression (**LARS**) [Hastie *et al.*, 2001].

The proposed methodology is summarised in Algorithm 1. Once the mapping $P_S \in \mathbb{R}^{d_S \times d_T}$ is obtained, the target model is retrained along with the projected source data ($S \times P_S$). The SHFR-ECOC approach does not retrain the model after finding the transformation. It uses the source model to predict the class labels of transformed target instances. In contrast, our approach utilizes the benefits of randomization and implicit feature selection of RF to retrain the model attuned for target domain.

4 Experiments

We compare the performance of the proposed algorithm against other baseline classifiers and approaches that perform transfer. Random forests (BRF) and SVM that uses ECOC (SVM ECOC) were chosen as the baseline classifiers. Transfer approaches include SHFR ECOC [Zhou *et al.*, 2014] and HeMAP (linear and non-linear) [Shi and Yu, 2012]. The number of trees in the random forest was set to 100. The number of bagged features for learning in a tree in the forest was set to $\sqrt{d} + 5$, where d is the total number of features.

The parameters for the SVM model with RBF kernel were fine-tuned using grid search. Based on cross validation experiments, the length of ECOC was set to 35, beyond which the performance plateaued. We choose three diverse datasets, varying in the size and sparsity of the features, for investigating the performance of the different algorithms.

The **CASAS dataset** [Cook *et al.*, 2013a] is a collection of smart home datasets that are widely used for investigating activity recognition algorithms. We use the horizon house (HH) datasets from this collection, which are records of sensor data from single resident smart homes. Sensor data from one smart home serves as the source and another acts as the target. A sliding window of 20 sensor events is used to build the feature vector that consists of counts of sensor events within the sliding window, along with temporal features such as time of the day and day of the week [Cook and Krishnan, 2015; Feuz, 2014]. The feature vector is annotated with the activity label associated with the last sensor event in the sliding window. The number of features and activity labels in each dataset are presented in Table 1. The feature values of the sensors in close vicinity appear to be mutually related. This motivates learning a sparse feature mapping instead of a dense mapping. The target training set consists of approximately 7000 samples that preserve the original class distribution. 16 such random subsets are used for evaluating the performance of the different algorithms.

The **20 Newsgroups** [Lang, 1995] text collection is a sparse dataset of approximately 19000 documents belonging to 20 classes that follow a label hierarchy. The transfer experiments were performed on two datasets each containing the subcategories falling under **rec and talk**, and **rec and sci** respectively. There are a total of 8 classes in each dataset with a vocabulary spanning over 26000 words. We considered only the first 10000 features that contributed the most towards the classification task. For each dataset, two transfer settings were created. In the first setting, the source and target consisted of random and mutually exclusive partition of 5000 features. Target training data is created by randomly selecting 10 samples per class. In the second setting, the roles of the source and target dataset were reversed. Since the baseline SVM ECOC model was unable to handle the high dimensional features, PCA was performed while preserving 75% variance on the TF-IDF feature values. Dimensionality reduction is not performed as a pre-processing step for the other approaches. The predefined test partitions of the dataset are used for testing the approaches.

The **Statlog (Landsat Satellite)** [Lichman, 2013] image dataset comprises of 6 classes and 36 real-valued features. It

Table 2: Performance comparison is depicted in terms of mean error(%). Statistically significant SHDA-RF results against BRF and SHFR-RF are highlighted in bold and indicated by $*$ respectively.

CASAS HH datasets								
	Baseline Results		**Transfer Results**					
S→T	**BRF**	**SVM-ECOC**	**SHFR-ECOC**	**HeMap Linear**	**HeMap Non-Linear**	**FA**	**SHFR-RF**	**SHDA-RF**
hh102→hh113	30.49±2.58	47.14±1.00	39.22±1.63	51.06±1.53	52.97±0.97	34.71±1.55	28.68±1.14	**27.93±2.54**
hh102→hh118	28.6±1.07	57.74±1.84	43.52±1.18	59.6±0.89	61.8±0.87	37.7±2.38	27.89±0.95	**26.97±1.15***
hh113→hh102	28.44±1.54	37.54±1.60	38.70±1.50	41.41±1.92	43.47±2.53	38.64±2.68	25.97±1.69	**25.97±1.00**
hh113→hh118	21.6±0.45	54.97±1.13	36.7±1.41	58.4±1.26	63±1.39	31±2.7	19.47±1.07	**18.38±1.29***
hh118→hh102	29.6±1.86	39.99±1.59	39.28±1.88	43±0.99	45.7±0.9	37.4±2.67	29.54±1.88	**27.83±2.64***
hh118→hh113	23.5±1.21	36.3±0.67	32.35±1.1	40.3±0.72	41.4±0.53	31±7.38	21.69±0.68	**21.54±1.47**
20 Newsgroups dataset								
	Baseline results		**Transfer Results**					
S → T	**BRF**	**SVM-ECOC (PCA)**	**SHFR-ECOC (PCA)**	**HeMap Linear**	**HeMap Non-Linear**		**SHFR-RF**	**SHDA-RF**
rec v/s sci								
F1:F5000→F5001:F10000	51.91±2.3	50.49±4.1	48.01±3.5	63.6±3.62	63.22±4.1		46.61±1.36	**40.06±2.9***
F5001:F10000→F1:F5000	68.41±3.6	67.09±4.0	60.23±6.6	73.1±3.9	72.8±4.6		58.12±2.13	**56.81±4.1***
rec v/s talk								
F1:F5000→F5001:F10000	55.79±1.1	56.12±1.6	51.55±2.5	66.2±3.9	66.0±3.55		49.99±0.12	**48.82±3.3***
F5001:F10000→F1:F5000	68.63±2.4	66.16±3.8	52.92±3.1	70.44±3.0	70.2± 6.11		44.67±0.23	**35.51±5.2***
Satellite Statlog dataset								
S→T	**BRF**	**SVM ECOC**	**SHFR-ECOC**	**HeMap-Linear**	**HeMap Non-Linear**		**SHFR-RF**	**SHDA-RF**
F1:F18→F19:F36	19.30±0.9	21.45±1.3	22.20±2.13	33.31±5.8	33.18±5.1		19.42±1.65	18.58±1.6
F19:F36→F1:F18	20.05±1.00	21.50±1.67	22.45±1.1	31.57±6.1	32.16±4.2		19.73±1.45	**18.66±0.78***

Table 3: Performance of SVM and SHFR ECOC models on original features of 20 Newsgroups dataset.

rec v/s sci		
S → T	**SVM ECOC**	**SHFR ECOC**
F1:F5000→F5001:F10000	79.48±3.39	87.09±4.1
F5001:F10000→F1:F5000	85.01±3.76	88.22±3.12
rec v/s talk		
F1:F5000→F5001:F10000	83.12±3.6	85.16±4.23
F5001:F10000→F1:F5000	89.55±4.1	86.92±3.75

Table 4: Performance of HeMap (Linear and Non-Linear) on 20 Newsgroups dataset without explicit correspondence between source and target data

rec v/s sci		
S → T	**HeMap Linear**	**HeMap Non-Linear**
F1:F5000→F5001:F10000	83.75±4.23	93.75± 4.79
F5001:F10000→F1:F5000	85.01±3.76	88.22±3.12
rec v/s talk		
F1:F5000→F5001:F10000	87.5±6.45	86.44±6.52
F5001:F10000→F1:F5000	83.75±4.79	85.0± 7.07

consists of 4435 examples in the training set and 2000 examples in the test set. The 36 features were split randomly into two equal groups for creating the source and target domains. To evaluate different algorithms, we used multiple sets of 10 labeled samples per class to create target training data.

5 Results and Discussion

The performance of different classifiers on the datasets is reported in Table 2. The common observation across all the datasets is the superior performance of baseline random forest (BRF) model over other baseline and some transfer learning approaches. This is another motivation behind adopting random forest model for performing transfer. The performance of the SHDA-RF algorithm on the CASAS-HH dataset is significantly better than all the other approaches by about 2-3% (p-value<0.05). Among the baseline classifiers, it is evident that the BRF models perform better than SVM ECOC. This can be explained by considering that the activity labels in the dataset are annotated by humans using rule based heuristics. It can be also noted that SHFR ECOC, a transfer strategy based on SVM ECOC, performs better than SVM ECOC significantly. This suggests that the possibility of knowledge transfer between the two domains, which is further reinforced by the performance improvement obtained by SHDA-RF over BRF model. Another common strategy that is used for performing transfer on activity recognition datasets is by defining a mapping that aggregates sensors to form layout independent functional areas (FA) [van Kasteren *et al.*, 2010] as an explicit meta-feature space. For example, individual sensors in the 'bedroom' are all clubbed together under a single feature. It can be observed from Table 2 that this approach performs worse than the BRF model. This suggests potential loss in information due to aggregation of different sensor events that is critical for differentiating activities happening in the same functional area. The FA approach, an unsupervised transfer approach, on the other hand performs significantly better than other unsupervised transfer approaches namely HeMAP (linear and non-linear).

On the high dimensional 20 Newsgroups dataset, SHDA-RF results in superior performance as compared to all the other approaches. The difference in the performance of

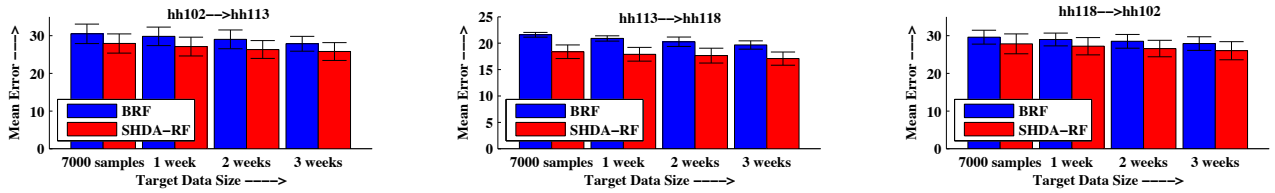

Figure 3: Availability of labeled target data helps to reduce mean error by learning a better mapping

SHDA-RF and the next best classifier SHFR-ECOC is on an average 7-8% (p-value$<$ 0.05). Handling high dimensional sparse data with only a few samples available per class necessitated the use of dimensionality reduction techniques for SVM ECOC and SHFR ECOC approaches. This can be observed by comparing the results of the SVM ECOC and SHFR ECOC models trained and tested on the original features (Table 3) and on transformed features (Table 2). However, the proposed approach does not require such a preprocessing step and is able to learn well in the original high dimensional space. The HeMAP approaches attempt to estimate a direct mapping between the source and target data. Learning this mapping in the presence of explicit correspondence between source and target data is easier than in the general case. As depicted in Table 4, the performance of HeMap suffers without explicit correspondence between source and target data. Even with explicit correspondence between the data points, the performance of the unsupervised transfer approaches are not at par with the other techniques. SHDA-RF

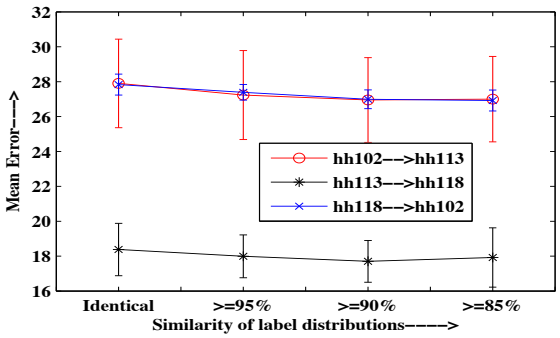

Figure 4: Effect of the number of pivots on the mean error

performs marginally better than BRF on the Statlog dataset. It is able to significantly outperform the BRF model only in one out of the two settings. Though the difference in the performances of SVM ECOC and SHFR ECOC is not significant, however if considered, it performs better than the SHFR ECOC transfer model. The smaller margin of improvement could be attributed to the property of the dataset, which is dense and real-valued.

Random Forest is used as the final model for comparing the different transfer mappings. The performance of the transfer mapping learned from SHFR-ECOC that uses random forest as the final model (SHFR-RF) is significantly poorer than SHDA-RF on the 20 Newsgroups dataset. On the CASAS-HH datasets, the results are only marginally poorer. Thus overall the results do seem to suggest that the transfer mapping learned through SHDA-RF is better than SHFR-ECOC.

Figure 3 presents the results for BRF and SHDA-RF models on a few CASAS-HH datasets with increasing target training data. It can be observed that the mean error for both the approaches reduces with increase in the number of labeled target domain data. However, the transfer approach performs marginally better than the baseline when number of target training examples is close to 50%.

The SHDA-RF algorithm uses only identical label distributions across the domains as pivots. We conducted experiments to study the effect of increasing the shared label distributions between the domains by relaxing the similarity between the distributions. We used Jensen-Shannon divergence [Lin, 1991] to determine the similarity between two label distributions. Figure 4 reports the mean error for models with increasing similarity relaxation on three CASAS-HH datasets. It can be observed that the mean error reduces only till about 90% relaxation beyond which the error increases marginally.

6 Summary and Future Work

In this paper we present a novel supervised heterogeneous domain adaptation technique that learns the mapping between heterogeneous feature spaces of different dimensions. Our algorithm uses the shared label distributions across the domains as the pivots for learning the feature transformation. We estimate the pivots using random forest models trained both on source and a small part of target labeled data. The experiments conducted on diverse datasets suggest the superiority of the proposed algorithm over other baseline and feature transfer approaches.

The SHDA-RF algorithm establishes the mapping between the feature spaces. A natural extension will be to consider instance transfer approaches to reduce the marginal and conditional probability distributions between the target and transformed source data. The proposed algorithm utilises a single source domain for knowledge transfer. Another direction that we would like to explore is how to effectively combine labeled data from multiple source domains to make an improved final prediction on the target. Finally, variants of random forests and sampling techniques can be used to improve upon on the random forest model that is the foundation of the SHDA-RF approach.

Acknowledgments

The authors are grateful to the anonymous reviewers for their valuable comments. This research is partially supported by the ISIRD grant from IIT Ropar.

References

[Blitzer *et al.*, 2006] John Blitzer, Ryan McDonald, and Fernando Pereira. Domain adaptation with structural correspondence learning. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 120–128, 2006.

[Breiman, 2001] Leo Breiman. Random forests. *Machine learning*, pages 5–32, 2001.

[Cook and Krishnan, 2015] Diane J. Cook and Narayanan C. Krishnan. *Activity Learning:Discovering, Recognizing, and Predicting Human Behavior from Sensor Data*. John Wiley and Sons Inc., 2015.

[Cook *et al.*, 2013a] Diane J. Cook, Aaron S. Crandall, Brian L. Thomas, and Narayanan C. Krishnan. Casas: A smart home in a box. *Computer*, 46(7):62–69, 2013.

[Cook *et al.*, 2013b] Diane J. Cook, Kyle Dillon Feuz, and Narayanan C. Krishnan. Transfer learning for activity recognition: a survey. *Journal of Knowledge and Information Systems*, pages 537–556, 2013.

[Cook *et al.*, 2013c] Diane J. Cook, Narayanan C. Krishnan, and Parisa Rashidi. Activity discovery and activity recognition: A new partnership. *IEEE Transactions on Systems Man and Cybernetics*, pages 820–828, 2013.

[Feuz and Cook, 2014] Kyle Dillon Feuz and Diane J. Cook. Heterogeneous transfer learning for activity recognition using heuristic search techniques. *International Journal of Pervasive Computing and Communications*, pages 393–418, 2014.

[Feuz, 2014] Kyle Dillon Feuz. *Preparing smart environments for life in the wild: Feature-space and Multi-view hetrogeneous learning*. PhD thesis, Washington State University, 2014.

[Hastie *et al.*, 2001] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Springer Series in Statistics. 2001.

[He *et al.*, 2014] Jingrui He, Yan Liu, and Qiang Yang. Linking heterogeneous input spaces with pivots for multi-task learning. In *Proceedings of the SIAM International Conference on Data Mining*, pages 181–189, 2014.

[Hu and Yang, 2011] Derek Hao Hu and Qiang Yang. Transfer learning for activity recognition via sensor mapping. In *Proceedings of the 22nd International Joint Conference on Artificial Intelligence*, pages 1962–1967, 2011.

[Hu *et al.*, 2011] Derek Hao Hu, Vincent Wenchen Zheng, and Qiang Yang. Cross-domain activity recognition via transfer learning. *Journal of Pervasive and Mobile Computing*, pages 344–358, 2011.

[Lang, 1995] Ken Lang. Newsweeder: Learning to filter netnews. In *Proceedings of the Twelfth International Conference on Machine Learning*, pages 331–339, 1995.

[Li *et al.*, 2014] Wen Li, Lixin Duan, Dong Xu, and Ivor W. Tsang. Learning with augmented features for supervised and semi-supervised heterogeneous domain adaptation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1134–1148, 2014.

[Lichman, 2013] M. Lichman. UCI machine learning repository, 2013.

[Lin, 1991] Jianhua Lin. Divergence measures based on the shannon entropy. *IEEE Transactions on Information Theory*, pages 145–151, 1991.

[Pan and Yang, 2010] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, pages 1345–1359, 2010.

[Pan, 2010] Jialin Pan. *Feature based transfer learning with real-world applications*. PhD thesis, Hong Kong University of Science and Technology, 2010.

[Rajan and Ghosh, 2004] Suju Rajan and Joydeep Ghosh. An empirical comparison of hierarchical vs. two-level approaches to multiclass problems. In *Multiple Classifier Systems*, Lecture Notes in Computer Science, pages 283–292. 2004.

[Rashidi and Cook, 2010] Parisa Rashidi and Diane J. Cook. Multi home transfer learning for resident activity discovery and recognition. In *Proceedings of the International Workshop on Knowledge Discovery from Sensor Data*, pages 56–63, 2010.

[Shi and Yu, 2012] Xiaoxiao Shi and Philip Yu. Dimensionality reduction on heterogeneous feature space. In *Proceedings of the 12th IEEE International Conference on Data Mining*, pages 635–644, 2012.

[van Kasteren *et al.*, 2010] T. L. M. van Kasteren, G. Englebienne, and B. J. A. Kröse. Transferring knowledge of activity recognition across sensor networks. In *Proceedings of the 8th International Conference on Pervasive Computing*, pages 283–300, 2010.

[Wang and Mahadevan, 2009] Chang Wang and Sridhar Mahadevan. Manifold alignment without correspondence. In *Proceedings of the 21st International Joint Conference on Artificial Intelligence*, pages 1273–1278, 2009.

[Zhou *et al.*, 2014] Joey Tianyi Zhou, Ivor W. Tsang, Sinno Jialin Pan, and Mingkui Tan. Heterogeneous domain adaptation for multiple classes. In *Proceedings of the 17th International Conference on Artificial Intelligence and Statistics*, pages 1095–1103, 2014.