Supervised Heterogeneous Domain Adaptation via Random Forests

Sanatan Sukhija SANATAN@IITRPR.AC.IN
Narayanan C Krishnan CKN@IITRPR.AC.IN
Department of Computer Science and Engineering, IIT Ropar, Rupnagar, Punjab, India

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 using random forests. Our algorithm uses the shared label distributions present across the domains as pivots for learning a sparse feature transformation. 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

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 different but related, single or multiple auxiliary domains (referred to as source domains). For most transfer applications such as cross-lingual sentiment analysis and cross-domain activity recognition (Pan, 2010), the source and target data are represented using heterogeneous features of different dimensions. As the domains have heterogeneous feature spaces, the goal is to discover a common space or a mapping that bridges the domains.

The proposed algorithm yields a heterogeneous feature space class-invariant mapping $P_S \in d_S \times d_T$ by bridging the two domains using the common label space. The generated mapping returns a target feature as a linear combination of source features, assuming no correspondence between the data-points of the domains that share no overlapping features.

Appearing in Proceedings of the 2^{nd} Indian Workshop on Machine Learning, IIT Kanpur, India, 2016.

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

2. Related Work

A domain independent feature space remapping is the focal point of heterogeneous domain adaptation. The binding task at hand is to find a translator that reduces the differences between the domains in the common space. Based on how this common space is determined, the transfer approaches can be broadly split into two categories, namely, Feature Remapping and Latent Space Transformation. *Feature Remapping* strives to reduce the differences of the domains by mapping the features of one domain to the other i.e. $g : X_S \rightarrow X_T$ or $g : X_T \rightarrow X_S$. The simplest strategy entails greedily mapping the features across the domains based on some fitness criterion (Feuz, 2014). Alternate approaches rely on domain independent features known as pivots that can be utilised to align the feature spaces. 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. 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. *Latent Space Transformation* intends to discover a non-trivial common subspace while trying to preserve certain characteristics of the original feature spaces. Heterogeneous spectral mapping (HeMap) (Shi & 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. 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.

3. Proposed Methodology

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.

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. 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. Every leaf node is associated with a data partition that follows a specific label distribution. Similar label distributions from the source and target are the pivots that are used for bridging the two domains. 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.

3.2. Estimating Feature Relationships

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. One simple way to compute feature importance towards creating a data partition is to give equal importance to all the features that were used as split nodes along 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. 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.

3.3. Deriving the Feature Transformation

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:

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

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. 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.

Table 1. 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-L**	**HeMap-NL**	**SHFR-RF**	**SHDA-RF**
hh102→hh118	28.6±1.07	57.74±1.84	43.52±1.18	59.6±0.89	61.8±0.87	27.89±0.95	**26.97±1.15***
hh113→hh118	21.6±0.45	54.97±1.13	36.7±1.41	58.4±1.26	63±1.39	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	29.54±1.88	**27.83±2.64***
20 Newsgroups dataset							
	Baseline results		**Transfer Results**				
S → T	**BRF**	**SVM-ECOC**	**SHFR-ECOC**	**HeMap-L**	**HeMap-NL**	**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***
Amazon CLS dataset							
S→T	**BRF**	**SVM-ECOC**	**SHFR-ECOC**	**HeMap-L**	**HeMap-NL**	**SHFR-RF**	**SHDA-RF**
English→French	44.46±2.73	52.23±3.64	39.01±2.54	56.69±4.24	55.35±4.35	38.85±3.51	**36.66±3.38***
English→German	45.43±2.92	51.36±3.92	38.33±3.18	57.76±3.1	57.32±3.71	37.62±2.31	**33.29±4.12***
English→Japanese	49.2±3.28	53.28±4.69	39.84±2.63	59.99±4.55	59.89±4.47	38.22±3.59	**34.87±4.93***

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 (L) and non-linear (NL)) (Shi & Yu, 2012). The hyper-parameters associated with random forest (Breiman, 2001) were set using cross-validation experiments. 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 & Krishnan, 2015) 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. The feature vector is annotated with the activity label associated with the last sensor event in the sliding window. 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** (Lichman, 2013) 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. The predefined test partitions of the dataset are used for testing the approaches. The **Amazon Cross Lingual Sentiment (CLS) dataset** (Lichman, 2013) consists of product reviews written in English, French, German and Japanese for three different product categories, namely, books, music and dvds. The English language reviews act as the Source domain and the reviews written in the other languages are treated as the target domain. To handle high dimensional features, PCA was performed while preserving 75% variance on the TF-IDF feature values. The target domain was constructed with 10 instances per class and the remaining instances were used for testing the performance of model.

5. Results and Discussion

The performance of different classifiers on the datasets is reported in Table 1. The superior performance of baseline random forest (BRF) model was one of the motivations 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. On the high dimensional 20 Newsgroups dataset, SHDA-RF results in superior performance as compared to all the other approaches. 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. However, the proposed approach does not require such a pre-processing 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. It was observed that even with explicit correspondence between the data points, the performance of the unsupervised transfer approaches are not at par with the other techniques. Even on the the Amazon CLS dataset, SHDA-RF performs significantly better than all the baseline and transfer approaches.

To compare different transfer mappings, random forest was used as the final model. The results suggest that the transfer mapping learned through SHDA-RF is better than all the transfer approaches under consideration. It was also observed that with increase in availability of labeled target data, the mean error reduces by learning a better mapping. 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 to determine the similarity between two label distributions. It was observed that the mean error reduces only till about 90% relaxation beyond which the error increases marginally.

6. Summary

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.

Dual Submission

The submission has been accepted for presentation at the 25th International Joint Conference on Artificial Intelligence (IJCAI) http://ijcai-16.org/, July 9-15, 2016, New York.

References

Breiman, Leo. Random forests. *Machine learning*, pp. 5–32, 2001.

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

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

Lichman, M. UCI machine learning repository, 2013. URL http://archive.ics.uci.edu/ml.

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

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

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