Proceedings of the Twenty-Third AAAI Conference on Artificial Intelligence (2008)

# Transfer Learning via Dimensionality Reduction

**Sinno Jialin Pan, James T. Kwok and Qiang Yang**
Department of Computer Science and Engineering
Hong Kong University of Science and Technology, Hong Kong
{sinnopan,jamesk,qyang}@cse.ust.hk

## Abstract

Transfer learning addresses the problem of how to utilize plenty of labeled data in a source domain to solve related but different problems in a target domain, even when the training and testing problems have different distributions or features. In this paper, we consider transfer learning via dimensionality reduction. To solve this problem, we learn a low-dimensional latent feature space where the distributions between the source domain data and the target domain data are the same or close to each other. Onto this latent feature space, we project the data in related domains where we can apply standard learning algorithms to train classification or regression models. Thus, the latent feature space can be treated as a bridge of transferring knowledge from the source domain to the target domain. The main contribution of our work is that we propose a new dimensionality reduction method to find a latent space, which minimizes the *distance* between distributions of the data in different domains in a latent space. The effectiveness of our approach to transfer learning is verified by experiments in two real world applications: indoor WiFi localization and binary text classification.

## Introduction

*Transfer learning* aims to solve the problem when the training data from a source domain and the test data from a target domain follow different distributions or are represented in different feature spaces (Caruana 1997). There are two main approaches to transfer learning in the past. The first approach can be referred to as instance-based approach (Dai et al. 2007; Huang et al. 2007; Sugiyama et al. 2008), where different weights are learned to rank training examples in a source domain for better learning in a target domain. Another approach can be referred to as feature-based approach (Ando and Zhang 2005; Argyriou, Evgeniou, and Pontil 2007; Blitzer, McDonald, and Pereira 2006; Raina et al. 2007), which tries to learn a common feature structure from different domains that can bridge the two domains for knowledge transfer. Several techniques have been developed for transfer learning, including *multi-task learning* (Ando and Zhang 2005; Argyriou, Evgeniou, and Pontil 2007), *multi-domain learning* (Blitzer, McDonald, and Pereira 2006) and *self-taught learning* (Raina et al. 2007). However, few previous feature-based methods have considered how to exploit a latent space as a bridge to facilitate knowledge transfer. As a result many of them may only have limited ability to transferring knowledge. In this paper, we focus on transfer learning in a latent feature space, so that even when the target domain have no labeled data, we can still learn a high performance classifier by making use of the training data from a source domain.

Our approach is intuitively appealing: if we can find a latent space where the marginal distributions of the data between different domains are close to each other, then this space can act as a bridge to propagate a classification model. More specifically, if two domains are related to each other, then there may exist several common *latent variables* that dominate the observed data. Some of them may cause the distributions of the observations to be different, while others may not. We can uncover these latent factors that do not cause change across domains, on which the source and target data distributions are found to be close to each other. Then, this is the lower-dimensional space we are looking for.

We illustrate our idea using a learning-based indoor localization problem as an example, where a client moving in a WiFi environment wishes to use the received signal strength (RSS) values to locate itself. In an indoor building, RSS values are affected by many hidden factors, such as temperature, human movement, building structure, properties of access points (APs), etc. Among these hidden factors, the temperature and the human movement may vary in time, resulting in changes in RSS values. However, the building structure and properties of APs are relatively stable. Thus, if we use the latter two factors to represent the RSS data, the distributions of the data collected in different time periods may be close to each other. Thus, this is the latent space where we can ensure a transferring of a learned localization model from one time period to another, or from one spatial area to another. Another example is learn to do text classification across domains. If two text-classification domains have different distributions, but are related to each other (e.g., news articles and blogs), there may be some *la-*

*tent topics* shared by these domains. Some of them may be relatively stable while others may not. If we use the stable latent topics to represent documents, the distance between the distributions of documents in related domains may be small. Then, in the latent space spanned by latent topics, we can transfer the text-classification knowledge.

In this paper, we propose a new dimensionality reduction algorithm designed to ensure effective transfer learning. This algorithm is driven by the objective to minimize the *distance* between distributions of the data in different domains in a low-dimensional latent space. In other words, we try to discover a latent space described by a feature transformation function $F$ such that the marginal distributions of $F(X_{src})$ and $F(X_{tar})$ are close to each other, where $F(X_{src})$ and $F(X_{tar})$ are new representations of patterns $X_{src}$ and $X_{tar}$ in the latent space. If the conditional probabilities $P(Y_{src}|F(X_{src}))$ and $P(Y_{tar}|F(X_{tar}))$ are similar, we can learn a model $f$ with $F(X_{src})$ and $Y_{src}$ and apply $f$ to predict labels of $F(X_{tar})$ directly.

In summary, our main contribution is a novel dimensionality reduction-based algorithm that aims to minimize the distance between distributions of different data sets in a latent space to enable effective transfer learning. We apply our new approach to two real world applications in a transfer learning setting to demonstrate its outstanding performance.

## Related Works and Preliminaries

### Transfer Learning

Feature-based methods have been widely used in many areas related to transfer learning. In multi-task learning, domain-specific information in related tasks is used to jointly train multiple classifiers in a way that they can benefit each other. A shared representation is exploited while the extra tasks can be used as an inductive bias during learning (Ando and Zhang 2005; Argyriou, Evgeniou, and Pontil 2007). In multi-domain learning, (Blitzer, McDonald, and Pereira 2006) described a heuristic method to construct new representations of the data for domain adaptation. In self-taught learning, (Raina et al. 2007) first learned high-level set of bases from a lot of unlabeled data for which may have different labels from the labeled data, and then projected the labeled data to these bases to get new representations for further classification problems.

The instance-based approach to transfer learning is an another way for solving the transfer learning problems (Dai et al. 2007; Huang et al. 2007; Sugiyama et al. 2008). Many instance-based methods make a common assumption that although the marginal probabilities $P(X_{src})$ and $P(X_{tar})$ are different, the conditional probabilities $P(Y_{src}|X_{src})$ and $P(Y_{tar}|X_{tar})$ are the same, where $X_{src}$ and $X_{tar}$ are patterns in a source domain and in a target domain, respectively. Here $Y_{src}$ and $Y_{tar}$ are the corresponding labels. In reality, however, this assumption may not hold. For example, in an indoor WiFi localization problem, we try to determine locations of a mobile device given its received signal strength (RSS) values sent from multiple transmitters or access points (APs). Some previous works have discovered that the distribution of RSS values $P(\mathbf{x})$, where **x** represents RSS values, may be non-Gaussian and can vary greatly due to dynamic environmental factors (Pan et al. 2007). Furthermore, the probability of locations given RSS values $P(y|\mathbf{x})$ estimated from one time period is not reliable for location estimation in another time period, where $y$ represents a location label. In this paper, we relax this assumption and only assume that there exists a latent space $F$ where $P(Y_{src}|F(X_{src}))$ and $P(Y_{tar}|F(X_{tar}))$ are similar.

### Dimensionality Reduction

Dimensionality reduction has been studied widely in machine learning community. (van der Maaten, Postma, and van den Herik 2007) gives a recent survey on various dimensionality reduction methods. Traditional dimensionality reduction methods try to project the original data to a low-dimensional latent space while preserving some properties of the original data. Since they cannot guarantee that the distributions between different domain data are similar in the reduced latent space, they cannot directly be used to solve transfer learning problems. Thus we need to develop a new dimensionality reduction algorithm for transfer learning.

A more recent dimensionality reduction technique is maximum variance unfolding (MVU) (Weinberger, Sha, and Saul 2004), which is motivated by designing kernels for kernel principal component analysis (KPCA) from the data itself. MVU extracts a low-dimensional representation of the data by maximizing the variance of the embedding while preserving the local distances between neighboring observations. MVU can be formulated in a *semidefinite programming* (SDP) (Lanckriet et al. 2004) optimization problem and solved by many optimization solvers. After estimating the kernel matrix $K$, MVU applies PCA to $K$ to choose a few eigenvectors as bases and projects the original data onto these bases to get low-dimensional representations.

### Maximum Mean Discrepancy

There are many criteria to estimate the distance between different distributions. A well-known example is *Kullback-Leibler* (K-L) *divergence*. Many criteria are parametric because they need an intermediate density estimate. To solve our problem, we wish to find a nonparametric estimate criterion of distance between distributions of data sets. *Maximum Mean Discrepancy* (MMD) is a relevant criterion for comparing distributions based on reproducing Kernel Hilbert Space (RKHS) (Borgwardt et al. 2006). Let $X = \{x_1, ..., x_{n_1}\}$ and $Y = \{y_1, ..., y_{n_2}\}$ be random variable sets with distributions $\mathcal{P}$ and $\mathcal{Q}$. The empirical estimate of distance between $\mathcal{P}$ and $\mathcal{Q}$ defined by MMD is as follows

$$Dist(\text{X,Y}) = \sup_{\|f\|_{\mathcal{H}} \leq 1} \left(\frac{1}{n_1}\sum_{i=1}^{n_1} f(x_i) - \frac{1}{n_2}\sum_{i=1}^{n_2} f(y_i)\right) \tag{1}$$

where $\mathcal{H}$ is a universal RKHS (Steinwart 2001). $Dist(X, Y)$ is non-negative, which vanishes if and only if $\mathcal{P} = \mathcal{Q}$, when $n_1, n_2 \rightarrow \infty$. By the fact that in a RKHS, function evaluation can be written as $f(x) = \langle\phi(x), f\rangle$, where $\phi(x) : \mathcal{X} \to \mathcal{H}$, the empirical estimate of MMD can be rewritten as follows:

$$Dist(\text{X,Y}) = \|\frac{1}{n_1}\sum_{i=1}^{n_1}\phi(x_i) - \frac{1}{n_2}\sum_{i=1}^{n_2}\phi(y_i)\|_{\mathcal{H}} \tag{2}$$

In summary, based on the MMD theory (Borgwardt et al. 2006), the distance between distributions of two samples is equivalent to the distance between the means of the two samples mapped into a RKHS.

## Dimensionality Reduction for Transfer Learning

### Problem Statement and Overall Approach

In a transfer learning setting, some labeled data $\mathcal{D}_{src}$ are available in a source domain, while only unlabeled data $\mathcal{D}_{tar}$ are available in the target domain. We denote the source domain data as $\mathcal{D}_{src} = \{(x_{src_1}, y_{src_1}), \ldots, (x_{src_{n_1}}, y_{src_{n_1}})\}$, where $x_{src_i} \in \mathbb{R}^m$ is the input and $y_{src_i}$ the corresponding label. Similarly, we denote the target domain data as $\mathcal{D}_{tar} = \{x_{tar_1}, \ldots, x_{tar_{n_2}}\}$, where, for simplicity, the input $x_{tar_i}$ is also assumed to be in $\mathbb{R}^m$. Let $\mathcal{P}(X_{src})$ and $\mathcal{Q}(X_{tar})$ (or $\mathcal{P}$ and $\mathcal{Q}$ in short) be the marginal distributions of $X_{src}$ and $X_{tar}$, respectively. In general, they can be different. Our task is then to predict the labels $y_{tar_i}$'s corresponding to the inputs $x_{tar_i}$'s in the target domain.

The proposed transfer learning approach is based on dimensionality reduction, and consists of two steps. First, we propose a new dimensionality reduction method (which will be called *Maximum Mean Discrepancy Embedding* (MMDE) in the sequel) to learn a low-dimensional latent space $F$ common to both domains. Let the projection map be $\psi$. We try to ensure that the distributions of the projected data, $\psi(X_{src})$ and $\psi(X_{tar})$, are close to each other. In the second step, we apply a traditional machine learning algorithm to train a model from $\psi(x_{src_i})$ in the latent space to $y_{src_i}$. The trained model can then be used for predicting the label of $x_{tar_i}$ in the target domain. In the sequel, we denote $\psi(X_{src})$ and $\psi(X_{tar})$ by $X'_{src} = \{x'_{src_i}\}$ and $X'_{tar} = \{x'_{tar_i}\}$, respectively.

### Step1: Maximum Mean Discrepancy Embedding

In this section, we address the problem of learning a common low-dimensional latent space $F$ such that the distributions of the source and target data ($X'_{src}$ and $X'_{tar}$) can be close to each other. On using (2), this is equivalent to minimizing

$$\text{dist}(X'_{src}, X'_{tar}) = \left\|\frac{1}{n_1}\sum_{i=1}^{n_1}\phi(x'_{src_i}) - \frac{1}{n_2}\sum_{i=1}^{n_2}\phi(x'_{tar_i})\right\|_{\mathcal{H}},$$

for some $\phi \in \mathcal{H}$. Thus,

$$\begin{aligned}&\text{dist}(X'_{src}, X'_{tar})\\ &= \text{dist}(\psi(X_{src}), \psi(X_{tar}))\\ &= \left\|\frac{1}{n_1}\sum_{i=1}^{n_1}\phi\circ\psi(x_{src_i}) - \frac{1}{n_2}\sum_{i=1}^{n_2}\phi\circ\psi(x_{tar_i})\right\|_{\mathcal{H}}\end{aligned} \tag{3}$$

Given that $\phi \in \mathcal{H}$, it is easy to show the following:

**Lemma 1** *Let $\phi$ be the feature map of an universal kernel. Then $\phi\circ\psi$ is also the feature map of an universal kernel for any arbitrary map $\psi$.*

Therefore, our goal becomes finding the feature map $\phi\circ\psi$ of some universal kernel such that (3) is minimized. Denote the corresponding universal kernel by $k$. Equation (3) can be written in terms of the kernel matrices defined by $k$, as:

$$\text{dist}(X'_{src}, X'_{tar}) = \text{trace}(KL), \tag{4}$$

where $K = \begin{bmatrix} K_{src,src} & K_{src,tar} \\ K^T_{tar,src} & K_{tar,tar}\end{bmatrix} \in \mathbb{R}^{(n_1+n_2)\times(n_1+n_2)}$ is a composite kernel matrix with $K_{src}$ and $K_{tar}$ being the kernel matrices defined by $k$ on the data in the source and target domains, respectively, and $L = [L_{ij}] \succeq 0$ with

$$L_{ij} = \begin{cases} \frac{1}{n_1^2} & x_i, x_j \in X_{src}, \\ \frac{1}{n_2^2} & x_i, x_j \in X_{tar}, \\ -\frac{1}{n_1 n_2} & \text{otherwise}. \end{cases}$$

In the transductive setting, we can learn this kernel matrix $K$ instead of learning the universal kernel $k$. However, we need to ensure that the learned kernel matrix does correspond to an universal kernel. To do this, we first recall the following property of universal kernels (Song 2007):

**Theorem 1** *A kernel is universal if for arbitrary sets of distinct points it induces strictly positive definite kernel matrices.*

While universal kernels induce strictly positive definite kernel matrices, the following proposition shows that certain strictly positive definite kernel matrices can also induce universal kernels.

**Proposition 1** *If a kernel matrix $K$ can be written as*

$$K = \widetilde{K} + \varepsilon I, \tag{5}$$

*where $\varepsilon > 0$, $\widetilde{K} \succeq 0$ and $I$ is the identity matrix, then the kernel function corresponding to $K$ is universal.*

Hence, as long as the learned kernel matrix is of the form in (5), we can be assured that the corresponding kernel is universal.

Besides minimizing the trace of $KL$ in (4), we also have the following constraints / objectives which are motivated from MVU:

1. The distance is preserved, i.e., $K_{ii} + K_{jj} - 2K_{ij} = d^2_{ij}$ for all $i, j$ such that $(i, j) \in \mathcal{N}$ [1];
2. The embedded data are centered;
3. The trace of $K$ is maximized.

[1] For all $i, j$, if $x_i$ and $x_j$ are k-nearest neighbors, we denote this by using $(i, j) \in \mathcal{N}$.

Thus, the embedding problem can be formulated as the following optimization problem:

$$\begin{aligned} \min_{K=\widetilde{K}+\varepsilon I} \quad & \operatorname{trace}(KL) - \lambda \operatorname{trace}(K) & (6) \\ \text{s.t.} \quad & K_{ii} + K_{jj} - 2K_{ij} = d_{ij}^2, \ \forall (i,j) \in \mathcal{N}, & \\ & K\mathbf{1} = \mathbf{0}, \ \widetilde{K} \succeq 0, & \end{aligned}$$

where $\varepsilon > 0$ and $\mathbf{1}$ and $\mathbf{0}$ are the vectors of ones and zeros, respectively. $\varepsilon$ is a small positive constant. The relative weightings of the two terms in the objective is controlled by the parameter $\lambda \geq 0$ [2]. This coefficient can be determined empirically.

We can further rewrite the above optimization problem as a semidefinite program (SDP):

$$\begin{aligned} \min_{\widetilde{K}\succeq 0} \quad & \operatorname{trace}(\widetilde{K}L) - \lambda \operatorname{trace}(\widetilde{K}) & (7) \\ \text{s.t.} \quad & \widetilde{K}_{ii} + \widetilde{K}_{jj} - 2\widetilde{K}_{ij} + 2\varepsilon = d_{ij}^2, \ \forall (i,j) \in \mathcal{N}, & \\ & \widetilde{K}\mathbf{1} = -\varepsilon\mathbf{1}. & \end{aligned}$$

This can be solved by standard SDP solvers. After obtaining $\widetilde{K}$, we can then apply PCA and select the leading eigenvectors to construct low-dimensional representations, $X'_{src}$ and $X'_{tar}$. In the sequel, we will call this approach *Maximum Mean Discrepancy Embedding* (MMDE). Note that, the optimization problem (7) is similar to a new supervised dimensionality reduction method, *colored MVU* (Song et al. 2008). However, there are two major differences between MMDE and colored MVU. First, the $L$ matrix in colored MVU is a kernel matrix that encodes label information of the data, while the $L$ in MMDE can be treated as a kernel matrix that encode distribution information of different data sets. Second, besides minimize the trace of KL, MMDE also aims to unfold the high dimensional data by maximize the trace of K.

### Step2: Training Models in the Latent Space

Using supervised or semi-supervised learning, we can train a model $f$ for the mapping between the estimated $X'_{src}$ and the class labels $Y_{src}$. This can then be used to obtain the predicted label $f(x'_{tar_i})$ of $x_{tar_i}$. Although we do not learn a function to explicitly project the original data $X_{tar}$ to $X'_{tar}$, we can still use the method of harmonic functions (Zhu, Ghahramani, and Lafferty 2003) to estimate the labels of new data in the target domain. The complete algorithm is shown in Algorithm 1.

## Experiments

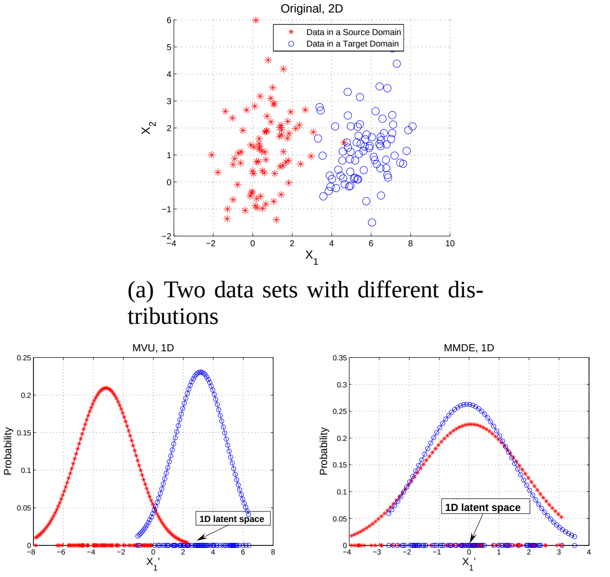
In this section, firstly, we use a synthetic data set to show explicitly why our method for transfer learning works. After that, we use two real world data sets to verify our method in a classification task and a regression task, respectively. In all experiments, to avoid over-fitting, we randomly select $60\%$ examples from $D_{src}$ as the training data and randomly select $60\%$ examples from $D_{tar}$ as the test data, repeat this $5$ times. The results published in all experiments are average results of these five individual results.

**Algorithm 1** Transfer Learning via Maximum Mean Discrepancy Embedding

```
Input: A labeled source domain data set $\mathcal{D}_{src} = \{(x_{src_i}, y_{src_i})\}$, a unlabeled target domain data set $\mathcal{D}_{tar} = \{x_{tar_i}\}$ and a positive $\lambda$.
Output: Labels $Y_{tar}$ of the unlabeled data $X_{tar}$ in the target domain.
1: Solve the SDP problem in (7) to obtain a kernel matrix $K$.
2: Apply PCA to the learned $K$ to get new representations $\{x'_{src_i}\}$ and $\{x'_{tar_i}\}$ of the original data $\{x_{src_i}\}$ and $\{x_{tar_i}\}$, respectively.
3: Learn a classifier or regressor $f : x'_{src_i} \rightarrow y_{src_i}$
4: Use the learned classifier or regressor to predict the labels of $\mathcal{D}_{tar}$, as: $y_{tar_i} = f(x'_{tar_i})$.
5: When new data $\mathcal{D}_{tar}^{new}$ arrive in the target domain, use harmonic functions with $\{x_{tar_i}, f(x'_{tar_i})\}$ to predict their labels.
```

### Synthetic Data Set

For the synthetic data, we generated two data sets: one represents the source domain (stars) and the other represents the target domain (circles), with different Gaussian distributions in a two-dimensional space (see Figure 1(a)). In Figures 1(b) and 1(c), the original 2D data are projected to a 1D latent space by applying MVU and MMDE, respectively. Gaussian distribution functions are used to fit the two data sets in the 1D latent space. We can see that, in the 1D latent space learned by MVU, which is a special case of KPCA,



(a) Two data sets with different distributions

(b) 1D projection by MVU.

(c) 1D projection by MMDE.

Figure 1: An Example of 2D Synthetic Data

[2] In particular, $\lambda$ contains a normalization term of trace($K$) and a tradeoff coefficient.

the distributions of the two data sets are still very different. However, in the latent space learned by MMDE, the distributions of the two data sets are close to each other. That is why MMDE can help transfer learning more effectively.

### Experimental Results on the WiFi Data Set

Our experimental data are collected in a WiFi area. To collect the WiFi experiment data, we carried an **IBM**© T60 laptop and walked in the floor of an office building, whose size is about $72 \times 37.5\ m^2$. The laptop is equipped with an **Intel**© Pro/3945ABG internal wireless card and installed a software to record WiFi signal strength every $0.5$ seconds. For obtaining the ground truth, we separated the environment into $135$ small grids, each of which is about $1.5 \times 1.5\ m^2$. We stopped at each grid for one or two seconds to collect the WiFi data. $500$ examples were collected in the midnight on one day as a source domain data set $D_{src}$ and $500$ examples were collected in the afternoon two days later as a target domain data set $D_{tar}$.

In a complex indoor environment, the distribution of WiFi signal strength at a certain location can change a lot due to dynamic environmental factors. Thus transfer learning becomes a necessary step to address indoor WiFi localization problems. To show that our proposed dimensionality reduction method for transfer learning works well for solving the WiFi localization problems, we compare the performance of various regressors trained in different feature space. Regression models used in our experiments are Regularized Least Square Regressor (RLSR), Support Vector Regressor (SVR) and Laplacian Regularized Least Square Regressor (LapRLSR) (Belkin, Niyogi, and Sindhwani 2006), respectively. Our goal is to verify that traditional regression models with help of MMDE can be applied to solve transfer learning problems. Thus, we use the default parameters of these regression models and do not change them in all experiments. Figure 2 shows the culmulative probabilities of these three regressors that are trained in the latent space learned by MMDE and in the original feature space, where culmulative probability means the estimation accuracy at different acceptable error distances. From this figure, we can see that regression models trained in the latent space, which are denoted by MMDE+RLSR, MMDE+SVR and MMDE+LapRLSR, get much higher performance than the ones trained in the original feature space.

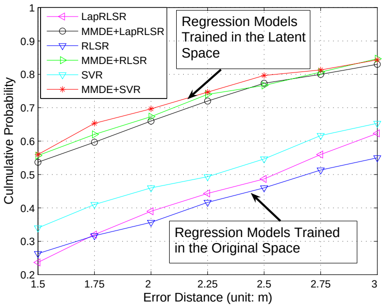


Figure 2: Comparison of Accuracy (The number of dimensions of the latent space is set to 10)

In Figure 3, we compare the performance of regression models trained in different latent space with different numbers of dimensions. From this figure, we can see that regression models built in the latent space either by MVU or by MMDE can improve the performance. However, the regression models based on MMDE achieve much higher performance. This is because that MMDE not only removes the noise from the original data but also reduces the *distance* of distributions between different data sets.

### Experimental Results on Text Data Sets

In text classification experiments, we used preprocessed data sets of Reuters-21578, which is in a transfer learning setting, to evaluate our proposed method. The basic idea is to utilize the hierarchy of the data sets. The binary classification task is defined as classifying top categories. Each top category is split into two disjoint parts with different subcategories, one for training and the other for test. In this case, distributions between the training and test data may be very different. Therefore, we have three data sets **orgs** vs **people**, **orgs** vs **places** and **people** vs **places** in transfer learning setting [3]. In this experiment, we use Support Vector Machines (SVMs) and Transductive Support Vector Machines (TSVMs) with linear kernel to verify the transferability of the MMDE algorithm. In Table 1, we can see that SVMs and TSVMs trained in the latent space that is learned by MMDE get much higher accuracy than those trained in the original space. From the table, we can find that the performance of traditional classifiers trained in the latent space learned by MMDE can be used in a transfer learning setting. In summary, MMDE based dimensionality reduction method can support various regression models and classification models for transfer learning.

## Conclusion and Future Work

In this paper, we have developed a novel transfer learning technique for learning in a latent space. We proposed a novel MMDE algorithm for transfer learning across two domains. Our experiments on two different applications demonstrated that our proposed solution can effectively improve the performance of many traditional machine learning algorithms for transfer learning. In the future, we plan to extend MMDE to nonnegative feature extraction, such that it can help transfer learning with other traditional classifiers, such as the Naive Bayes Classifier. Furthermore, we wish to find an efficient method to extend MMDE to handle large-scale transfer learning problems.

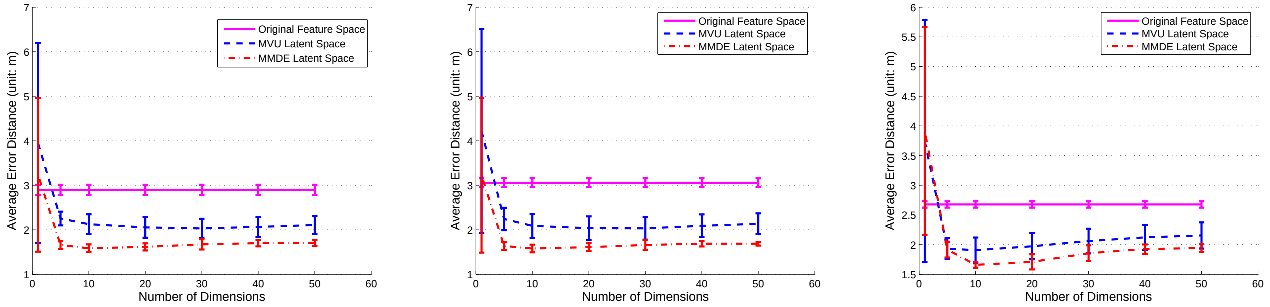
## Acknowledgment

Qiang Yang and Sinno Jialin Pan thank the support of Hong Kong CERG grant 621307.

## References

Ando, R. K., and Zhang, T. 2005. A framework for learning predictive structures from multiple tasks and unlabeled data. *Journal of Machine Learning Research* 6:1817–1853.

[3] Detailed description can be found in the following url: http://apex.sjtu.edu.cn/apex_wiki/dwyak



(a) LapRLSR in Different Feature Space.

(b) RLSR in Different Feature Space.

(c) SVR in Different Feature Space.

Figure 3: Comparison of Average Error Distance (unit: m) among Regression Models Trained in Different Feature Space.

| Data Set | Documents | | Words | SVM | | TSVM | |
|---|---|---|---|---|---|---|---|
| | $\lvert D_{scr}\rvert$ | $\lvert D_{tar}\rvert$ | | Original | MMDE | Original | MMDE |
| people vs places | 1079 | 1088 | 8000 | 0.519 (0.039) | **0.654** (0.021) | 0.553 (0.025) | **0.666** (0.036) |
| orgs vs people | 1239 | 1210 | 9729 | 0.661 (0.021) | **0.722** (0.034) | 0.694 (0.026) | **0.726** (0.033) |
| orgs vs places | 1016 | 1046 | 8568 | 0.670 (0.025) | **0.709** (0.021) | 0.704 (0.035) | **0.743** (0.036) |

Table 1: Comparison of Accuracy among Classification Models Trained in Different Feature Space (a value inside parentheses is the standard deviation of five round results).

Argyriou, A.; Evgeniou, T.; and Pontil, M. 2007. Multi-task feature learning. In *Proceedings of the 19th Annual Conference on Neural Information Processing Systems*. 41–48.

Belkin, M.; Niyogi, P.; and Sindhwani, V. 2006. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *Journal of Machine Learning Research* 7:2399–2434.

Blitzer, J.; McDonald, R.; and Pereira, F. 2006. Domain adaptation with structural correspondence learning. In *Proceedings of the Conference on Empirical Methods in Natural Language*, 120–128.

Borgwardt, K. M.; Gretton, A.; Rasch, M. J.; Kriegel, H.-P.; Schölkopf, B.; and Smola, A. J. 2006. Integrating structured biological data by kernel maximum mean discrepancy. In *Proceedings of the 14th International Conference on Intelligent Systems for Molecular Biology*, 49–57.

Caruana, R. 1997. Multitask learning. *Machine Learning* 28(1):41–75.

Dai, W.; Yang, Q.; Xue, G.; and Yu, Y. 2007. Boosting for transfer learning. In *Proceedings of the 24th International Conference on Machine Learning*, 193–200.

Huang, J.; Smola, A.; Gretton, A.; Borgwardt, K. M.; and Schölkopf, B. 2007. Correcting sample selection bias by unlabeled data. In *Proceedings of the 19th Annual Conference on Neural Information Processing Systems*.

Lanckriet, G. R. G.; Cristianini, N.; Bartlett, P. L.; Ghaoui, L. E.; and Jordan, M. I. 2004. Learning the kernel matrix with semidefinite programming. *Journal of Machine Learning Research* 5:27–72.

Pan, S. J.; Kwok, J. T.; Yang, Q.; and Pan, J. J. 2007. Adaptive localization in a dynamic WiFi environment through multi-view learning. In *Proceedings of the 22nd AAAI Conference on Artificial Intelligence*, 1108–1113.

Raina, R.; Battle, A.; Lee, H.; Packer, B.; and Ng, A. Y. 2007. Self-taught learning: Transfer learning from unlabeled data. In *Proceedings of the 24th International Conference on Machine Learning*, 759–766.

Song, L.; Smola, A.; Borgwardt, K.; and Gretton, A. 2008. Colored maximum variance unfolding. In *Proceedings of the 20th Annual Conference on Neural Information Processing Systems*.

Song, L. 2007. *Learning via Hilbert Space Embedding of Distributions*. Ph.D. Dissertation, The University of Sydney. Draft.

Steinwart, I. 2001. On the influence of the kernel on the consistency of support vector machines. *Journal of Machine Learning Research* 2:67–93.

Sugiyama, M.; Nakajima, S.; Kashima, H.; Buenau, P. V.; and Kawanabe, M. 2008. Direct importance estimation with model selection and its application to covariate shift adaptation. In *Proceedings of the 20th Annual Conference on Neural Information Processing Systems*.

van der Maaten, L. J. P.; Postma, E. O.; and van den Herik, H. J. 2007. Dimensionality reduction: A comparative review. Published online.

Weinberger, K. Q.; Sha, F.; and Saul, L. K. 2004. Learning a kernel matrix for nonlinear dimensionality reduction. In *Proceedings of the 21st International Conference on Machine Learning*.

Zhu, X.; Ghahramani, Z.; and Lafferty, J. D. 2003. Semi-supervised learning using gaussian fields and harmonic functions. In *Proceeding of The 22th International Conference on Machine Learning*, 912–919.