Modeling Sequential Domain Shift through Estimation of Optimal Sub-spaces for Categorization

Suranjana Samanta
ssamanta@cse.iitm.ac.in
Tirumarai A Selvan
tirumarai.selvan@gmail.com
Sukhendu Das
http://www.cse.iitm.ac.in/~sdas/

Visualization and Perception Lab
Dept. of CS&E
Indian Institute of Technology Madras
Chennai, India

Abstract

This paper describes a new method of unsupervised domain adaptation (DA) using the properties of the sub-spaces spanning the source and target domains, when projected along a path in the Grassmann manifold. Our proposed method uses both the geometrical and the statistical properties of the subspaces spanning the two domains to estimate a sequence of optimal intermediary subspaces. This creates a path of shortest length between the sub-spaces of source and target domains, where the distributions of the projected source and target domain data are identical when projected onto these intermediate sub-spaces (lying along the path). We extend our concept to the kernel space and perform non-linear projections on the subspaces using kernel trick. Projections of the source and target domains onto these intermediary sub-spaces are used to obtain the incremental (or gradual) change in the geometrical as well as the statistical properties of sub-spaces spanning the source and target domains. Results on object and event categorization using real-world datasets, show that our proposed optimal path in the Grassmann manifold produces better results for the problem of DA than the usual geodesic path.

1 Introduction

The amount of images and videos to be analyzed are increasing at an enormous rate due to availability of cheap hardware and memory chips. But it is difficult to annotate and create sufficient amount of labeled training samples from various datasets, to perform tasks like categorization, detection, recognition, retrieval etc. Domain adaptation (DA) is the process which uses labeled training samples available from one domain to improve the performance of statistical tasks on test samples drawn from a different domain.

The domain from which the training samples and the test samples are obtained are termed as the source domain and the target domain respectively. In order to estimate the distribution of target domain, few training samples are also necessary from the target domain. Using the training samples from both the domains, the method of DA improves classification performance on the test samples drawn from target domain. Based on the type of training samples

available from the target domain, there are mainly two types of DA: (i) unsupervised - a large number of unlabeled training samples are available from target domain and (ii) supervised - a few number of labeled training samples are available from target domain.

In this paper, we propose a new method of unsupervised DA where a set of domain invariant sub-spaces are estimated using the geometrical and statistical properties of the source and the target domains. This is a modification of the work done by Gopalan *et al*. [14], where the geodesic path from the principal components of the source to that of the target is considered in the Grassmann manifold, and the intermediary points are sampled to represent the incremental change in the geometric properties of the data in source and target domains. Instead of the geodesic path, we consider an alternate path of shortest length between the principal components of source and target, with the property that the intermediary sample points on the path form domain invariant sub-spaces. This is obtained by minimizing the difference in Reproducing Kernel Hilbert Space (RKHS), between means of the projected source and target domains on the intermediary subspaces, as the difference between the means of source and target data in the RKHS is a measure of the discrepancy of the distributions between the two domains [16]. Thus we model the change in the geometric properties of data in both the domains sequentially, in a manner that the distributions of projected data from both the domains always remain similar along the path. The entire formulation is done in the kernel space which makes it more robust to any non-linear transformations.

The work described in this paper successfully exploits two different concepts of DA, which are: (i) projecting onto a domain invariant sub-space where the distributions of source and target domains are similar and (ii) modeling the sequential (or gradual) change of the sub-spaces spanning the source and target domains. We also provide a formulation to calculate intermediary points on the desired path between the sub-spaces of source and target domains on the Grassmann manifold in the kernel space. Hence, non-linear transformation of data is also dealt implicitly. Results on real-world datasets show the effectiveness of our proposed method for the task of categorization, when compared with some state of the art methods of unsupervised DA.

The rest of the paper is as follows. Section 2 briefly describes the state of the art of the related works. Section 3 explains the proposed method of DA. Section 4 shows the experimental results and section 5 concludes the paper.

2 Literature Review

Domain adaptation has gained enormous attention in the recent past. The concept is similar to compensating the bias of a dataset [21] and covariate shift [27]. There are two notable categories of approaches to handle this problem of unsupervised DA. The first one is a statistical approach, where a domain invariant sub-space is estimated such that in the projected space disparity in the distributions of two domains is minimized. Several methods [1, 15, 23] have used this concept to build a suitable sub-space. Apart from the disparity in distributions, several other cost functions for locality preservation [23, 29], empirical error in target domain [30], divergence measure [24, 28] etc. have also been considered for estimating the optimal transformation matrix/sub-space. Most of these works perform the projection in the kernel space in order to handle non-linear transformation of data. The second well noted category of method for DA is to consider the incremental or sequential change in the geometrical properties of the source and target domains, rather than performing a one shot transformation [14], [13], [6]. The initial work was suggested by Gopalan *et al*. [14], where a geodesic

path between the principal components of source and target domain data is considered in the Grassmann manifold. The intermediate sampled points on the path gives an estimate of the continuous change of the properties of sub-spaces of source and target domain. This was later enhanced by Gong *et al*. [13], where an infinite number of intermediary sampled points are considered on the geodesic path, estimated by the geodesic flow kernel. In other notable works, Fernando et. al. [11] has calculated a sub-space using eigen-vectors of two domains such that the basis vectors of transformed source and target domains are aligned. Traditional SVM and other classifiers have been modified appropriately such that these classifiers when trained on source domain data reduces classification error on the target domain data [19, 22, 30]. Application of DA for improved results of object categorization and video classification have been discussed in [1, 7, 9, 11, 13, 14, 24, 25].

3 Proposed Method of Estimating Sequence of Domain Invariant Sub-spaces

A popular method of DA is to find a domain invariant sub-space [1, 23], where the distributions of the source and target domains are similar. It has also been shown in [13, 14], that using the concept of geodesic path between the source and the target domain sub-spaces on Grassmann manifold, boosts the performance of DA. The intermediary points sampled on the geodesic path, form a sequence or a set of sub-spaces. We thus formulate our proposed method of DA, where we estimate a set of intermediary sub-spaces, sampled from a path on Grassmann manifold between the two sub-spaces spanning the source and target domain respectively, such that the distributions of source and target domains are similar when projected onto the estimated sub-spaces. We perform this by considering an alternative path, instead of the geodesic path, between the sub-spaces spanning the two domains on the Grassmann manifold. To deal with non-linear transformations of data, we find the required path between the two aforesaid sub-spaces in a suitable kernel space.

The steps of the algorithm of the proposed method of unsupervised DA, are as follows: (i) Obtain the kernel Gram matrices for the source and target domain data, (ii) Obtain the eigen-vectors of the two Gram matrices, (iii) Find the appropriate start and end points of the path on the Grassmann manifold, (iv) Find the intermediary domain invariant sub-spaces on the Grassmann manifold, (v) Project the source and target domain data onto each of the intermediary sub-spaces along with the start and end points of the path and (vi) Concatenate the projections for each instance and apply PLS to obtain the final domain invariant features.

In the following sub-sections, we first briefly mention the underlying principles for the proposed method DA and then explain our formulation for exploiting these existing concepts to design a novel approach to obtain the domain invariant representation, which give better classification results on real-world datasets.

3.1 Underlying Principle

Maximum Mean Discrepancy using difference of means in two domains [16] - The main purpose of DA is to bridge the gap between the distributions of the source and target domain. Estimating the underlying distribution from the few available sample points is a cumbersome and error-sensitive process. One effective way to compare the two distributions is using the Maximum Mean Discrepancy (MMD), which says that if the distributions of two feature sets

are given by l and r, then the distributions are equal, i.e., $l = r \iff E_l(f(x_l)) = E_r(f(x_r))$, where x_l and x_r are the sets of sample points drawn from independently and identically distributed (i.i.d.) sets l and r respectively. Also, $f(.)$ is a continuous bounded function on x_l and x_r. Generally, $f(.)$ is taken as unit ball function. It is known to us that universal kernels (Gaussian and Laplacian kernels) are unit ball functions, though polynomial kernel function has also been used for calculating MMD.

Grassmann manifold [10] - The Grassmann manifold, $G_{d,p}$, consists of all sub-spaces of dimension $d \times p$. A point on the Grassmann manifold represents a sub-space of dimension $d \times p$ and hence these two terms are used inter-changeably in the rest of the paper. Since a sub-space can have multiple sets of basis vectors, only the orthogonality constraint for estimating the optimal sub-space is not enough. Optimization should be performed on Grassmann manifold, which has been well explained by Edelman *et al.* [10].

Distance between two sub-spaces [12] - Distance between two sub-spaces in $\mathbb{R}^{n\times p}$, which are represented by two points on the Grassmann manifold $(G_{d,p})$, can be measured by the set of principal angles $\theta_i, i = 1,2,\dots p$ between these sub-spaces. These principal angles measure the geometrical dissimilarity between two sub-spaces and the projection distance between two sub-spaces A_1 and A_2 is given as: $\delta_{proj}(A_1,A_2) = \left(\sum_{i=1}^{p} sin^2\theta_i\right)^{1/2}$. The span of sub-spaces A_1 and A_2 are given by $span(A_1) = A_1A_1^T$ and $span(A_2) = A_2A_2^T$ respectively. The sum of the squared cosine of the principal angles between two sub-spaces is given by the dot product of the spans of the sub-spaces: $\sum_{i=1}^{d} cos^2\theta_i = \langle span(A_1), span(A_2)\rangle$, where θ_i are the principal angles between two sub-spaces. If the trace of a matrix A be denoted as $tr(A)$, then the square of the projection distance (as in [18]) is expressed as:

$$\delta^2_{proj}(A_1,A_2) = p - \sum_{i=1}^{p} cos^2\theta_i = p - tr(A_1A_1^TA_2A_2^T) = p - tr(A_2^TA_1A_1^TA_2) \tag{1}$$

Geodesic flow in Grassmann manifold [10, 14] - It has been observed that modeling the intermediate incremental stages due to changes in the geometric properties of the source and target domains is more effective than a one shot transformation of the features of source domain. Gopalan *et al.* [14] have considered the geodesic path between the sub-spaces spanning source and target domains in the Grassmann manifold. A geodesic path between the source and target gives the shortest path of gradual change in geometric properties of sub-spaces between two domains. A finite number of points lying on the geodesic path are sampled and the features from both the domains are projected onto the intermediary subspaces. Projection of source and target domain data onto these intermediary points depicts the continuous change in the geometric properties of the sub-spaces spanned by the two domains.

If U_X and U_T be the principal vectors of source and target domains, then the geodesic flow from U_S to U_T is given as: $P^G(t) = Q[exp(tB)]J$, where $Q \in \mathbb{R}^{d\times d}$ is an orthogonal matrix such that $Q^TU_X = J = [I_p; 0_{(d-p)\times p}]$, $I_p \in \mathbb{R}^{p\times p}$ is an identity matrix and $B = \begin{bmatrix} 0 & A^T \\ -A & 0 \end{bmatrix}$ is a skew-symmetric matrix, where $A \in \mathbb{R}^{(d-p)\times p}$. A specifies the direction and speed of the geodesic flow between U_X and U_Y (refer to [14] for details); $P^G(0) = U_X$, $P^G(1) = U_Y$ and for any other value of t $(0 < t < 1)$, we obtain an intermediary sub-space G_i lying on P^G.

Notations - Let X and Y be the source and target domains having n_X and n_Y number of instances respectively. Let the i^{th} and j^{th} instances of X and Y be represented as x_i and y_j respectively. If $\Phi(.)$ is a universal kernel function, then in kernel space the source and

target domains are $\Phi(X) \in \mathbb{R}^{n_X \times d}$ and $\Phi(Y) \in \mathbb{R}^{n_Y \times d}$ respectively. Let K_{XX} and K_{YY} be the Gram matrices of $\Phi(X)$ and $\Phi(Y)$ respectively ($K_{XX} = \Phi(X)\Phi(X)^T$, $K_{XY} = \Phi(X)\Phi(Y)^T$ and $K_{YY} = \Phi(Y)\Phi(Y)^T$). Let $D = [X;Y]$ denote the combined source and target domain data, and the corresponding data in kernel space is $\Phi(D)$. The Gram matrix formed using D is given by $K = \begin{bmatrix} K_{XX} & K_{XY} \\ K_{XY}^T & K_{YY} \end{bmatrix}$. Let, the transpose of a matrix A be denoted as A^T. The formulation of the cost function required to find the optimal sub-spaces along the path (between the sub-space spanned by $\Phi(X)$ and $\Phi(Y)$) is explained in the following sub-sections.

3.2 Estimating distance between two means

The discrepancy of the distribution of two domains can be measured by the distance between the means of two domains in RKHS using a universal kernel function. Let $\Phi(\tilde{X})$ and $\Phi(\tilde{Y})$ represent the projections of $\Phi(X)$ and $\Phi(Y)$ respectively onto a subspace $W_i \in \mathbb{R}^{d \times p}$, which is a point on the Grassmann manifold $G_{d,p}$. Here, d is the dimension of the source and target domain in RKHS and p is the dimension of the optimal sub-spaces. The mean of $\Phi(\tilde{X})$ and $\Phi(\tilde{Y})$ are: $\frac{1}{n_X}\sum_{j=1}^{n_X} \Phi(x_j)W_i$ and $\frac{1}{n_Y}\sum_{j=1}^{n_Y} \Phi(y_j)W_i$ respectively. Then, the square of the distance between the means of two domains is given as:

$$\begin{aligned}
\delta_\mu^2(i) &= \left(\frac{1}{n_X}\sum_{j=1}^{n_X} \Phi(x_j)W_i - \frac{1}{n_Y}\sum_{j=1}^{n_Y} \Phi(y_j)W_i\right)\left(\frac{1}{n_X}\sum_{j=1}^{n_X} \Phi(x_j)W_i - \frac{1}{n_Y}\sum_{j=1}^{n_Y} \Phi(y_j)W_i\right)^T \\
&= tr(W_i^T\Phi(X)^T I_1 \Phi(X)W_i) - tr(W_i^T\Phi(X)^T I_2 \Phi(Y)W_i) \\
&\qquad - tr(W_i^T\Phi(Y)^T I_2 \Phi(X)W_i) + tr(W_i^T\Phi(Y)^T I_3 \Phi(Y)W_i) \\
&= tr\left(W_i^T\Phi(D)^T \begin{bmatrix} I_1 & -I_2 \\ -I_2 & I_3 \end{bmatrix} \Phi(D)W_i\right) = tr\left(Z_i^T \Gamma Z_i\right) \qquad (2)
\end{aligned}$$

where, $W_i = \Phi(D)^T Z_i$, $Z_i \in \mathbb{R}^{(n_X+n_Y)\times p}$, $\Gamma = \left(K \begin{bmatrix} I_1 & -I_2 \\ -I_2 & I_3 \end{bmatrix} K\right)$ and $[I_1]_{n_X \times n_X}$, $[I_2]_{n_Y \times n_X}$ and $[I_3]_{n_Y \times n_Y}$ are matrices containing all elements as $1/n_X^2$, $1/n_X n_Y$ and $1/n_Y^2$ respectively.

3.3 Estimating distance between two sub-spaces in kernel space

We consider a path of shortest length between the principal components of $\Phi(X)$ and $\Phi(Y)$ such that the projections of the two domains onto the intermediate sub-spaces along the path have identical distribution. Let $\{P\}$ be the set of all such paths with different lengths. Our aim is to find the shortest path $P^W \in \{P\}$. Using lemma 1 we can state that the principal components of $\Phi(X)$ and $\Phi(Y)U_Y^\Phi U_X^{\Phi T}$ are the same, where U_X^Φ and U_Y^Φ are the principal components of $\Phi(X)$ and $\Phi(Y)$ respectively.

Lemma 1. *If U_A and U_B are the principal components of two datasets A and B respectively, then the principal component of $BU_BU_A^T$ is U_A.*

Proof. Let Λ_A and Λ_B be two diagonal matrices whose diagonal elements are the eigenvalues of A and B respectively. The covariance matrices of A and B can be written as $U_A\Lambda_A U_A^T$ and $U_B\Lambda_B U_B^T$ respectively. Now, the covariance matrix of $\hat{B} = (BU_BU_A^T)$ is given as:

$$\hat{B}^T\hat{B} = U_AU_B^T B^T BU_BU_A^T = U_AU_B^T U_B\Lambda_B U_B^T U_BU_A^T = U_A\Lambda_B U_A^T$$

This shows that the principal components of $BU_BU_A^T$ is U_A. □

Hence, instead of considering U_X as the starting point of the path P^W, the principal components of $\Phi(D_s)$ is considered, where $\Phi(D_s) = [\Phi(X); \Phi(Y)U_Y^\Phi U_X^{\Phi^T}]$. Similarly, the end point of P^W can be obtained by the principal components of $\Phi(D_t) = [\Phi(X)U_X^\Phi U_Y^{\Phi^T}; \Phi(Y)]$. Let, U_s^Φ and U_t^Φ be the principal components of $\Phi(D_s)$ and $\Phi(D_t)$ respectively. Also, V_X^Φ and V_Y^Φ be the eigen-vectors of K_{XX} and K_{YY} respectively. Similarly, let V_s^Φ and V_t^Φ be the eigen-vectors of K_s and K_t respectively, where K_s and K_t are the Gram matrices built on $\Phi(D_s)$ and $\Phi(D_t)$ respectively. Then we can write [26],

$$U_X^\Phi = \Phi(X)^T V_X^\Phi \quad (3) \qquad U_s^\Phi = \Phi(D_s)^T V_s^\Phi \quad (5)$$

$$U_Y^\Phi = \Phi(Y)^T V_Y^\Phi \quad (4) \qquad U_t^\Phi = \Phi(D_t)^T V_t^\Phi \quad (6)$$

Let, G_i denote the i^{th} sampled point on the geodesic path P^G and the i^{th} sample point on P^W represent the sub-space W_i. As U_s^Φ and U_t^Φ change along the paths P^G and P^W, V_s^Φ and V_t^Φ also modify proportionally due to the linear relationship (Eqns. 5 & 6). The start and the end points of P^W are given by $W_1 = V_s^\Phi$ and $W_{N'} = V_t^\Phi$ respectively, while the intermediate points are denoted by W_i, $i = 1, \ldots N'-1$. Thus including U_s^Φ and U_t^Φ, N' number of sub-spaces or points on P^G and P^W are considered.

Now, P^W is the path of shortest length if the sampled points from P^W is closest to the corresponding sampled points from P^G, i.e. $d_{proj}(G_i, W_i)$ is minimum, $\forall i = 2, \ldots, (N'-1)$. The square of the distance between two sub-spaces, P_i^G and P_i^W in the kernel space, can be derived using lemma 1, as:

$$\begin{aligned} \delta^2_{proj}(W_i, G_i) &= p - tr(W_i^T U_i^\Phi U_i^{\Phi^T} W_i) && (7)\\ &= p - tr(Z_i^T \Phi(D)\Phi(\hat{D}_i)^T V_i^\Phi V_i^{\Phi^T} \Phi(\hat{D}_i)\Phi(D)^T Z_i) && (8)\\ &= p - tr(Z_i^T \hat{K}_i V_i^\Phi V_i^{\Phi^T} \hat{K}_i^T Z_i) = p - tr(Z_i^T \Pi_i Z_i) && (9) \end{aligned}$$

where, $\Pi_i = \hat{K}_i V_i^\Phi V_i^{\Phi^T} \hat{K}_i^T$. $\Phi(\hat{D}_i)$ is an appropriate projection of $\Phi(D)$ such that the principal component of $\Phi(\hat{D}_i)$ is U_i^Φ (using Lemma 1). V_i^Φ is the i^{th} intermediary point sampled on the geodesic path from V_s^Φ to V_t^Φ. Using lemma 1 and Eqn. 5, the Gram matrix $\hat{K}_i$ (for i^{th} sub-space in the sequence) is defined as:

$$\hat{K}_i = \Phi(D)\Phi(\hat{D}_i)^T = \Phi(D)U_i^\Phi U_s^{\Phi^T}\Phi(D)^T = K V_i^\Phi V_s^{\Phi^T} K \quad (10)$$

3.4 Estimating intermediary sub-spaces

We can now estimate the intermediary sub-spaces on P^W using Eqns. 2 and 9. For an optimal value of Z_i, δ^2_{mu} and $\delta^2_{proj}(G_i, W_i)$ should be minimum. The optimization framework for estimating Z_i ($\forall i = 2, \ldots, N'-1$), is:

$$\underset{Z_i}{\text{maximize}} \quad tr(Z_i^T \Pi_i \Gamma^{-1} Z_i) \quad (11)$$

$$\text{subject to} \quad Z_i^T Z_i = I \quad (12)$$

Since the orthogonality condition in Eqn. 12 is not enough to find the optimal point in Grassmann manifold, we solve the problem using Newton's method, as given by Edelman *et al.* [10]. We use the Manopt toolbox [3] for solving the optimization problem. After obtaining the set of optimal Z_is, the projections of the data onto W_is are given as:

$$\Phi(D)W_i = KZ_i, \quad \forall i = 2, \ldots, (N'-1) \quad (13)$$

The projection of the data points onto the first and last (or initial and final) points of the path P^W i.e. on U_s^Φ and U_t^Φ are:

$$\Phi(D)U_s^\Phi = \Phi(D)\Phi(D_s)^T V_s^\Phi = \begin{bmatrix} K_{XX} & K_{XX}V_X^\Phi V_Y^{\Phi T} K_{YY} \\ K_{XY}^T & K_{XY}^T V_X^\Phi V_Y^{\Phi T} K_{YY} \end{bmatrix} V_s^\Phi \tag{14}$$

$$\Phi(D)U_t^\Phi = \Phi(D)\Phi(D_t)^T V_t^\Phi = \begin{bmatrix} K_{XY}V_Y^\Phi V_X^{\Phi T} K_{XX} & K_{XY} \\ K_{YY}V_Y^\Phi V_X^{\Phi T} K_{XX} & K_{YY} \end{bmatrix} V_t^\Phi \tag{15}$$

For the final task of classification, only the projections of the source and target domain data onto the intermediary sub-spaces are required. Thus, once the projections on the intermediary sub-spaces are obtained, it is not required to find the actual path P^W. The path P^W can be approximated by a piece-wise geodesic path on the Grassmann manifold, which sequentially connects all the intermediary sampled points between the principal components of the source and target domains in the correct order. Hence, using the appropriate projections onto the N' sampled points on the path P^W (Eqns. 13, 14, 15), we model the sequential changes of both the domains in such a manner that their distributions become identical.

3.5 Calculating domain-invariant features for classification

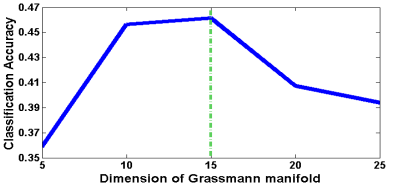
After obtaining the optimal sub-spaces, the projections of the source and target domains onto the intermediary sub-spaces are obtained and concatenated together, as done in [14]. If there are N' number of sub-spaces considered, including the principal components of source and target domain, we then get the source and target domain features as $n_X \times N'p$ and $n_Y \times N'p$ matrices. Finally, Partial Least Square (PLS) [17] is applied on the feature sets of source and target domains to obtain an m-dimensional $(m << N'p)$ feature vector for each instance, which are finally used for classification using KNN and SVM classifiers.

4 Experimental Results

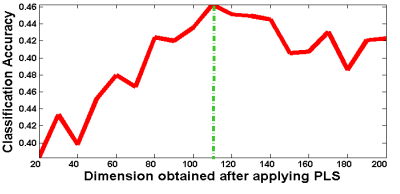
We evaluate our proposed method of unsupervised DA on real-world datasets for the tasks of object and event categorization. We have used the Gaussian kernel function to build the Gram matrices. Experiments performed using the different methods of DA, are discussed in the following.

Object Categorization in images - We evaluate the performance of the proposed method of DA for improving the results of object categorization using Office + Caltech datasets [13]. The dataset contains four domains: Amazon (A), Caltech (C), Dslr (D) and Webcam (W), with 10 classes of objects in each of the domains. Each image is resized to 300×300 dimension and SURF [2] features are extracted from the images to form a codebook of size 800. The results reported in this paper are obtained by using the features shared by the authors of [13]. We follow the same experimental protocols as described in [14]. For Amazon and Caltech domain, eight random samples per class and for Dslr and Webcam domain, three random samples per class have been chosen as training samples from target domain. We have considered nine intermediary sampled points on the path P^W, excluding the start and the end points, which gives the best average result. K nearest neighbor $(K = 1)$ has been used for the purpose of classification for all the results in this case.

It is necessary to compute optimal values of two parameters for experimentation. The first parameter is the dimension of the sub-spaces (and Grassmann manifold) which represents the points on the desired path and the second parameter is the dimension of the final domain invariant features which are obtained after performing PLS. We empirically obtain the pair of optimal parameter values in the range 5 to 25 and 20 to 200 respectively. The best average result across all the 12 source-target pairs, is obtained when the dimension of the Grassmann manifold is 15 and the dimension of the domain invariant features is 110. Plots shown in figure 1 (a) and (b) show the variation in the average classification accuracy, with respect to the change in the dimension of the Grassmann manifold and the dimension of the final domain invariant features obtained after PLS respectively.



(a)



(b)

Figure 1: Plots of average classification accuracies with change in dimensions of the: (a) Grassmann manifold and (b) final domain invariant features, after applying PLS.

Table 1 shows the classification accuracies for 12 different pairs of source and target domains, using a 25-fold cross validation. We compare our method with other unsupervised DA methods such as, Transfer Component Analysis [23], Geodesic Flow Subspace [14], Geodesic Flow Kernel [13] and Subspace alignment [11], while NA denotes 'No Adaptation', in which case only the source domain samples are used for training the classifier. In table 1, classification accuracies of the methods used for comparison are taken from [11]. From the above experimentations, we infer that the proposed method of unsupervised DA gives better result than other state of the art works in majority of the cases.

Table 1: Classification accuracies (in %-age) on Office+Caltech dataset [13] using different techniques of unsupervised domain adaptation (best results highlighted in bold).

Method	C→A	D→A	W→A	A→C	D→C	W→C
NA	21.5	26.9	20.8	22.8	24.8	16.4
TCA [23]	21.96	16.81	13.43	16.18	17.67	11.14
GFS [14]	36.9	32	27.5	35.3	29.4	21.7
GFK [13]	36.9	32.5	31.1	**35.6**	29.8	27.2
SA [11]	39.0	38.0	37.4	35.3	32.4	32.3
Proposed	**42.63**	**44.16**	**44.65**	34.40	**41.56**	**43.26**

Method	A→D	C→D	W→D	A→W	C→W	D→W	Average
NA	22.4	21.7	40.5	23.3	20.0	53.0	26.18
TCA [23]	16.69	22.8	32.31	23.60	22.03	44.69	21.61
GFS [14]	30.7	32.6	54.3	31.0	30.6	66.0	35.67
GFK [13]	35.2	35.2	70.6	34.4	33.7	74.9	39.76
SA [11]	37.6	39.6	80.3	38.60	36.80	**83.6**	44.24
Proposed	**38.82**	**43.64**	**80.57**	**39.31**	**42.27**	78.03	**47.76**

Event categorization in videos - We use three video datasets: Kodak [8, 9], YouTube [8, 9] and CCV dataset [20] as the 3 domains. We consider the YouTube data as the source domain and observed the classification accuracies on Kodak and CCV domains as done in [8]. We consider 6 common classes (events) between YouTube (906 videos) and Kodak (195 videos) as in [9], and 5 classes (events) between YouTube (821 videos) and CCV (2440 videos) as in [5]. For the first case, we use the distance matrices of Kodak and YouTube domains using SIFT and spatio-temporal (ST) features (HOG and HOF) [1]; and for the second case, we have obtained the codebook of size 2000 obtained using SIFT and ST features, as shared by the authors in [9]. Five samples per class have been randomly selected from the target domain for training the SVM classifier [4] with Gaussian kernel. Like the previous experiment, nine intermediate sub-spaces are considered apart from the start and the end points of the path on Grassmann manifold. The dimension of the Grassmann manifold is considered to be 10 and the dimension of the final domain invariant features obtained after applying PLS is 120. We have compared our proposed method with TCA [23]. Since, our experiments use the distance matrices as inputs, we are unable to obtain the performances of GFS [14], GFK [13] and SA [11] methods for this task. Figure 2 shows the Mean Average Precision (MAP) for the two cases of event categorization using both SIFT and ST features separately using a 25-fold cross validation. Results show that our proposed method of DA gives the best result in three out of four cases.

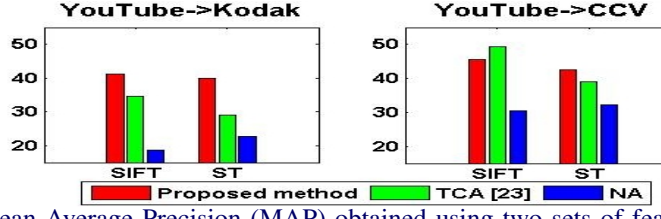


Figure 2: Mean Average Precision (MAP) obtained using two sets of features from three real-world datasets. Proposed method of domain adaptation (in red) performs better than TCA [23] (in green) and 'No Adaptation' (in blue) techniques.

5 Conclusion

In this paper, we have proposed a method of sequential domain adaptation by estimating a set of domain-invariant sub-spaces, along a path on Grassmann manifold. We achieve this by considering the shortest path between the sub-spaces spanning the source and target domains, where the intermediate points sampled on the optimal path represent domain-invariant sub-spaces with identical distributions of both the domains. The proposed method is able to handle non-linear transformation of data by using the representation in a higher dimensional kernel space. Experiments on real-world image and video datasets show that the proposed method performs better than most other relevant methods published on unsupervised DA. The method can be improved by automating the selection of the dimension of Grassmann manifold which has the potential to improve the classification performance.

Acknowledgment - This work has been partially supported by Tata Consultancy Service.

[1] http://vc.sce.ntu.edu.sg/index_files/VisualEventRecognition/VisualEventRecognition.html

References

[1] M. Baktashmotlagh, M.T. Harandi, B.C. Lovell, and M. Salzmann. Unsupervised domain adaptation by domain invariant projection. In *IEEE International Conference on Computer Vision*, pages 769–776, 2013.

[2] Herbert Bay, Andreas Ess, Tinne Tuytelaars, and Luc Van Gool. Speeded-up robust features (SURF). *Computer Vision Image Understanding*, 110(3):346–359, 2008.

[3] N. Boumal, B. Mishra, P.-A. Absil, and R. Sepulchre. Manopt: a Matlab toolbox for optimization on manifolds. *The Journal of Machine Learning Research*, 2014. Accepted for publication.

[4] Chih-Chung Chang and Chih-Jen Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

[5] Lin Chen, Lixin Duan, and Dong Xu. Event recognition in videos by learning from heterogeneous web sources. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2666–2673, 2013.

[6] Z. Cui, W. Li, D. Xu, S. Shan, X. Chen, and X. Li. Flowing on riemannian manifold: Domain adaptation by shifting covariance. *IEEE Transactions on Cybernetics*, in press, 2014.

[7] Jeff Donahue, Yangqing Jia, Oriol Vinyals, Judy Hoffman, Ning Zhang, Eric Tzeng, and Trevor Darrell. Decaf: A deep convolutional activation feature for generic visual recognition. In *International Conference on Machine Learning*, pages 647–655, 2014.

[8] Lixin Duan, Dong Xu, and Shih-Fu Chang. Exploiting web images for event recognition in consumer videos: A multiple source domain adaptation approach. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1338–1345, 2012.

[9] Lixin Duan, Dong Xu, Ivor W. Tsang, and Jiebo Luo. Visual event recognition in videos by learning from web data. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(9):1667–1680, 2012.

[10] Alan Edelman, Tomás A. Arias, and Steven T. Smith. The geometry of algorithms with orthogonality constraints. *SIAM Journal of Matrix Analysis and Applications*, 20(2): 303–353, 1999.

[11] Basura Fernando, Amaury Habrard, Marc Sebban, and Tinne Tuytelaars. Unsupervised visual domain adaptation using subspace alignment. In *IEEE International Conference in Computer Vision*, 2013.

[12] Gene H. Golub and Charles F. van Loan. *Matrix computations (3. ed.)*. Johns Hopkins University Press, 1996. ISBN 978-0-8018-5414-9.

[13] Boqing Gong, Yuan Shi, Fei Sha, and Kristen Grauman. Geodesic flow kernel for unsupervised domain adaptation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2066–2073, 2012.

[14] Raghuraman Gopalan, Ruonan Li, and Rama Chellappa. Domain adaptation for object recognition: An unsupervised approach. In *IEEE International Conference on Computer Vision*, pages 999–1006, 2011.

[15] Arthur Gretton, Alex Smola, Jiayuan Huang, Marcel Schmittfull, Karsten Borgwardt, and Bernhard Schölkopf. Covariate shift by kernel mean matching. *Dataset shift in machine learning, Chap. 8*, pages 131–160, 2009.

[16] Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning*, 13:723–773, 2012.

[17] Michael Haenlein and Andreas M. Kaplan. A Beginner's Guide to Partial Least Squares Analysis. *Understanding Statistics*, 3(4):283–297, 2004.

[18] Jihun Hamm and Daniel D. Lee. Grassmann discriminant analysis: A unifying view on subspace-based learning. In *International Conference on Machine Learning*, pages 376–383, 2008.

[19] Wei Jiang, Eric Zavesky, Shih fu Chang, and Alex Loui. Cross-domain learning methods for high-level visual concept classification. In *International Conference on Image Processing*, pages 161–164, 2008.

[20] Yu-Gang Jiang, Guangnan Ye, Shih-Fu Chang, Daniel Ellis, and Alexander C. Loui. Consumer video understanding: A benchmark database and an evaluation of human and machine performance. In *International Conference on Multimedia Retrieval*, pages 29:1–29:8, 2011.

[21] Aditya Khosla, Tinghui Zhou, Tomasz Malisiewicz, Alexei A. Efros, and Antonio Torralba. Undoing the damage of dataset bias. In *European Conference on Computer Vision*, pages 158–171, 2012.

[22] Fatemeh Mirrashed and Mohammad Rastegari. Domain adaptive classification. In *IEEE International Conference on Computer Vision*, 2013.

[23] Sinno Jialin Pan, I.W. Tsang, J.T. Kwok, and Qiang Yang. Domain adaptation via transfer component analysis. *IEEE Transactions on Neural Networks*, 22(2):199–210, 2011.

[24] Kate Saenko, Brian Kulis, Mario Fritz, and Trevor Darrell. Adapting visual category models to new domains. In *European Conference on Computer Vision*, pages 213–226, 2010.

[25] Suranjana Samanta and Sukhendu Das. Domain adaptation based on eigen-analysis and clustering, for object categorization. In *International Conference on Computer Analysis of Images and Patterns*, pages 245–253, 2013.

[26] John Shawe-Taylor and Nello Cristianini. *Kernel Methods for Pattern Analysis*. Cambridge University Press, New York, NY, USA, 2004. ISBN 0521813972.

[27] Hidetoshi Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of Statistical Planning and Inference*, 90(2): 227 – 244, 2000.

[28] Masashi Sugiyama, Shinichi Nakajima, Hisashi Kashima, Paul von Bünau, and Motoaki Kawanabe. Direct importance estimation with model selection and its application to covariate shift adaptation. In *Neural Information Processing Systems*, pages 1962–1965, 2007.

[29] Chang Wang and Sridhar Mahadevan. Heterogeneous domain adaptation using manifold alignment. In *International Joint Conferences on Artificial Intelligence*, pages 1541–1546. AAAI Press, 2011.

[30] Jun Yang, Rong Yan, and Alexander G. Hauptmann. Cross-domain video concept detection using adaptive svms. In *International conference on Multimedia*, pages 188–197, 2007.