Domain Adaptation Transfer Learning by SVM Subject to a Maximum-Mean-Discrepancy-like Constraint

Xiaoyi Chen and Régis Lengellé
LM2S, Institut Charles Delaunay, UMR CNRS 6281, University of Technology of Troyes,
12 rue Marie Curie, CS 42060 - 10004, Troyes Cedex, France
{*xiaoyi.chen, regis.lengelle*}*@utt.fr*

Keywords: Transfer Learning, Kernel, SVM, Maximum Mean Discrepancy.

Abstract: This paper is a contribution to solving the domain adaptation problem where no labeled target data is available. A new SVM approach is proposed by imposing a zero-valued *Maximum Mean Discrepancy*-like constraint. This heuristic allows us to expect a good similarity between source and target data, after projection onto an *efficient* subspace of a *Reproducing Kernel Hilbert Space*. Accordingly, the classifier will perform well on source and target data. We show that this constraint does not modify the quadratic nature of the optimization problem encountered in classic SVM, so standard quadratic optimization tools can be used. Experimental results demonstrate the competitiveness and efficiency of our method.

1 INTRODUCTION

Recently, *Transfer Learning* has received much attention in the machine learning community. First formally defined in (Pan and Yang, 2010), the aim of Transfer Learning is to learn a good-performance classifier or regressor in a new domain with the help of previous knowledge issued from different but related domains; the new domain is designated as *target* while domains of previous knowledge are designated as *sources*. In this paper, we propose to solve the transfer learning problem where there is no labeled target data available. According to the taxonomy given in (Pan and Yang, 2010), our proposed method belongs to the transductive transfer learning where the source and the target share the same label space but differentiate from each other in the feature space. Marginal, conditional distributions and priors might differ. This problem is also known as domain adaptation.

There is a variety of methods for transfer learning. In this paper, we propose the use of a *Support Vector Machine (SVM)* subject to a zero valued *Maximum Mean Discrepancy (MMD)*-like constraint. The choice of a zero-valued MMD as the constraint is that MMD is a non-parametric measure of the distance between 2 distributions (Dudley, 2002) and it can be easily kernelized (Gretton et al., 2012). Therefore, the combination of MMD and SVM is promising. SVM is a widely known classification method used in binary classification. It is well known for its high generalization ability and the simplicity in dealing with non-linearly separable data set by using the kernel trick. Our method keeps these advantages while performing well in the transfer learning context. As shown in section 3, the optimization problem remains convex and can be directly implemented using standard quadratic optimization tools. Adding a MMD-like constraint is a heuristic that allows us to expect that source and target data will become similar in some selected subspace of the feature space. Therefore, the separating hyperplane found by SVM for source data can perform well for target data. The experimental results prove the effectiveness of our idea.

This paper is organized as follows: in section 2, we give a short summary of related work; then we present our method in section 3 together with the optimization solution to the problem (in section 4); we prove the effectiveness of the proposed method on synthetic and real data sets in section 5. Finally, we conclude this paper and suggest perspectives.

2 RELATED WORK

Because the aim of our work is to perform MMD-like SVM based transductive transfer learning, we first review the general transductive transfer learning problem, followed by a presentation of SVM based transfer learning and MMD based transfer learning. Inter-

Chen, X. and Lengellé, R.
Domain Adaptation Transfer Learning by SVM Subject to a Maximum-Mean-Discrepancy-like Constraint.
DOI: 10.5220/0006119900890095
In *Proceedings of the 6th International Conference on Pattern Recognition Applications and Methods (ICPRAM 2017)*, pages 89-95
ISBN: 978-989-758-222-6

ested readers are referred to (Pan and Yang, 2010) and (Jiang, 2008) for more general transfer learning and domain adaptation surveys. For a more recent survey on domain adaptation, readers are referred to (Patel et al., 2015)

Transductive transfer learning refers to a shared label space but different source and target feature spaces with different marginal and/or conditional distributions (Pan and Yang, 2010). To take full advantage of source information is the key issue to make the improvement in learning the target task. When target labels are not available, typical methods include instance weighting (Huang et al., 2006) with the necessary assumption of the same conditional distributions. Other authors propose structural corresponding learning for information retrieval (Blitzer et al., 2007).

SVM based transfer learning adapts the traditional SVM to the transfer learning context. To the best of our knowledge, there are five principal kinds of SVM based transfer learning methods:

- transferring common parameter ($w_{common} = w_{target} - w_{specific}$) (Zhang et al., 2009)
- iteratively using SVM to label target domain data (Bruzzone and Marconcini, 2010)
- reweighting the penalty term of SVM (Liang et al., 2014)
- adding extra regularization term to standard SVM (Huang et al., 2012), (Tan et al., 2012)
- SVM by integrating a transformed alignement constraint combining the knowledge of different natures (Li et al., 2011)

MMD based transfer learning combines the MMD, which will be presented later in this paper, with standard learning method to perform transfer. To the best of our knowledge, MMD is used as a regularization term of the objective function. The principal idea is to deal with the trade-off between the classification performance of source data and the similarity of source and target. The interested reader could refer to SVM-based transfer learning classification in (Quanz and Huan, 2009), multiple kernel learning in (Ren et al., 2010), multi-task clustering in (Zhang and Zhou, 2012), maximum margin classification in (Yang et al., 2012), feature extraction in (Pan et al., 2011) (Uguroglu and Carbonell, 2011), etc.

3 PRESENTATION OF THE MMD CONSTRAINED SVM METHOD

In this section, we present our MMD constrained SVM transfer learning method. We first briefly review the basic theoretical foundations of MMD and its kernelized version

3.1 Review of Basic Theoretical Foundations

3.1.1 Maximum Mean Discrepancy

Maximum Mean Discrepancy (MMD) is a non-parametric *distance* measure which can be used to evaluate the difference between two distributions. The definition of MMD is:

Definition 1 (Maximum Mean Discrepancy (Fortet and Mourier, 1953))**.**
Let $\mathcal{F}$ be a class of functions f: $\mathcal{X} \to \mathbb{R}$ and p, q two Borel probabilistic measures defined on $\mathcal{X}$. The *Maximum Mean Discrepancy (MMD)* between p and q is defined as:

$$MMD[\mathcal{F}, p, q] = sup_{f \in \mathcal{F}} (\mathbf{E}_p[f(x)] - \mathbf{E}_q[f(y)])$$

As a ”distance measure” between two distributions, MMD has the following property:

Theorem 1 (Dudley, 1984)**.**
Let ($\mathcal{X}$,d) be a metric space and p, q two Borel probabilistic measures defined on $\mathcal{X}$, $p = q$ iff $\mathbf{E}_p[f(x)] = \mathbf{E}_q[f(y)]$ for any function $f \in C(\mathcal{X})$, where $C(\mathcal{X})$ is the space of continuous bounded functions and x, y are random variables drawn from distribution p and q respectively.

Thanks to the works of Smola (Smola, 2006) and Gretton et al. (Gretton et al., 2012), distributions can be embedded in a Reproducing Kernel Hilbert Space (RKHS), where a distribution can be considered as some mean element of this RKHS ($\mathcal{H}$):

$$\mu[P_x] = E_x[k(x,.)]$$

(Smola et al., 2007). Accordingly, MMD can be evaluated as $MMD[\mathcal{F}, p, q] = \|\mu_p - \mu_q\|_{\mathcal{H}}$, where μ_r stands for $\mathbf{E}_r[k(x,.)]$ and $k(x,.)$ is the representation of x in the RKHS.

As a simple deduction, the squared MMD is:

$$MMD^2[\mathcal{F}, p, q] = \|\mu_p - \mu_q\|^2_{\mathcal{H}}$$
$$= \mathbf{E}_{p,p}[k(x,x')] - 2\mathbf{E}_{p,q}[k(x,y)] + \mathbf{E}_{q,q}[k(y,y')]$$

Here, x and x' are independent observations drawn from distribution p, y and y' are independent observations from distribution q, k designates a universal kernel function (which means that $k(x,.)$ is continuous for all x and the RKHS induced by k is dense in $C(\mathcal{X})$).

Theorem 2 (Steinwart (Steinwart, 2002) and Smola (Smola, 2006)**).**
$MMD[\mathcal{F},p,q]=0$ iff $p=q$ when $\mathcal{F}=\{f:\|f\|_{\mathcal{H}}\leq 1\}$ provided that $\mathcal{H}$ is universal.

An unbiased estimate of kernelized squared MMD is proposed in (Serfling, 2009):

$$\widehat{MMD}_u^2[\mathcal{F},X,Y]=\frac{1}{m(m-1)}\sum_{i=1}^{m}\sum_{j\neq i}^{m}k(x_i,x_j)$$
$$+\frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}^{n}k(y_i,y_j)-\frac{2}{nm}\sum_{i=1}^{m}\sum_{j=1}^{n}k(x_i,y_j)$$

where $x_i, i=1,\ldots,m$ and $y_i, i=1,\ldots,n$ are iid examples drawn from p and q respectively.

SVM aims to find the hyperplane that *maximally* separates two classes. The commonly used formulation is:

$$\min \frac{1}{2}||w||^2+C\sum_{i=1}^{n}\varepsilon_i$$
$$s.t.\ \varepsilon_i \geq 0$$
$$y_i(w^T\phi(x_i)+b)\geq 1-\varepsilon_i\ \forall i=1,\ldots,n$$

where, as usual, w is the hyperplane parameter, ε_i is the error term associated to observation i, C is the trade-off parameter between the margin term and the classification error, $\phi(x_i)$ is the kernel representation of x_i, y_i is the label of x_i and b is the bias.

3.2 MMD Constrained SVM Transfer Learning

We now propose a heuristic to constrain the hyperplane that maximizes the margin between the source classes (and minimizes the corresponding classification error) to lie in a subspace where source and target distributions are as similar as possible. Another assumption is that the conditional probability distributions of labels are also similar (hypothesis that cannot be verified because the target labels are supposed unknown). Accordingly, we can expect the classifier to perform well, both on source and target data. The heuristic used to *maximize* the similarity between source and target is to satisfy the proposed constraint:

$$<\mu_{X_s}-\mu_{X_t},w>_{\mathcal{H}}=0$$

where μ_{X_s} (μ_{X_t}) is the sample mean of source (target) data in $\mathcal{H}$ and can be estimated by $\mu_{X_s}=\frac{1}{n_s}\sum\phi(X_s)$ ($\mu_{X_t}=\frac{1}{n_t}\sum\phi(X_t)$).

By imposing $<\mu_{X_s}-\mu_{X_t},w>_{\mathcal{H}}=0$, we expect that source and target data will be similar in $\mathcal{H}$.

The SVM problem can now be formulated as follows:

$$\min \frac{1}{2}||w||^2+C\sum_{i=1}^{n}\varepsilon_i$$
$$s.t. <\mu_{X_s}-\mu_{X_t},w>_{\mathcal{H}}=0 \qquad (1)$$
$$\varepsilon_i\geq 0$$
$$y_i(w^T\phi(x_i)+b)\geq 1-\varepsilon_i\ \forall i=1,\ldots,n$$

Our approach of using a MMD-like constraint instead of a MMD-regularization-term is to guarantee the transfer ability. In (Quanz and Huan, 2009), Quanz and Huan suggest to solve the problem : $\min \frac{1}{2}||w||^2+C\sum_{i=1}^{n}\varepsilon_i+\lambda||<\mu_{X_s}-\mu_{X_t},w>_{\mathcal{H}}||^2$. In that case, depending on the finite value of the regularization parameter λ, we may sometimes sacrifice this similarity to achieve a high classification accuracy for source only. Furthermore, during the optimization process, their method requires the calculation of the inverse of a matrix which slows down the algorithm and causes inaccuracy, while this is avoided in our work.

4 DUAL FORM OF THE OPTIMIZATION PROBLEM

In order to solve the above primal problem, we use the *representer theorem* (Schölkopf et al., 2001). w, the optimum solution of Equation 1 in the above section , can be expressed as:

$$w=\sum_{k=1}^{n_s}\beta_k^s\phi(x_k^s)+\sum_{l=1}^{n_t}\beta_l^t\phi(x_l^t) \qquad (2)$$

where β_k^s and β_l^t are the unknowns. Incorporating this expression into the constraint, we obtain:

$$<\mu_{X_s}-\mu_{X_t},w>_{\mathcal{H}}$$
$$=<\frac{1}{n_s}\sum_{i=1}^{n_s}\phi(x_i)-\frac{1}{n_t}\sum_{j=1}^{n_t}\phi(x_j),$$
$$\sum_{k=1}^{n_s}\beta_k^s\phi(x_k^s)+\sum_{l=1}^{n_t}\beta_l^t\phi(x_l^t)>_{\mathcal{H}}$$
$$=\frac{1}{n_s}\sum_{k=1}^{n_s}\beta_k^s\sum_{i=1}^{n_s}<\phi(x_i),\phi(x_k)>_{\mathcal{H}}$$
$$-\frac{1}{n_t}\sum_{k=1}^{n_s}\beta_k^s\sum_{j=1}^{n_t}<\phi(x_j),\phi(x_k)>_{\mathcal{H}}$$
$$+\frac{1}{n_s}\sum_{l=1}^{n_t}\beta_l^t\sum_{i=1}^{n_s}<\phi(x_i),\phi(x_l)>_{\mathcal{H}}$$
$$-\frac{1}{n_t}\sum_{l=1}^{n_t}\beta_l^t\sum_{j=1}^{n_t}<\phi(x_j),\phi(x_l)>_{\mathcal{H}}$$
$$=(K\widetilde{1})^T\beta$$

where $K = \begin{bmatrix} K_{SS} & K_{TS} \\ K_{ST} & K_{TT} \end{bmatrix}, K_{SS} =< \phi(x_i), \phi(x_k) >_{\mathcal{H}}$, $K_{TS} =< \phi(x_j), \phi(x_k) >_{\mathcal{H}}, K_{ST} =< \phi(x_i), \phi(x_l) >_{\mathcal{H}}$, $K_{TT} =< \phi(x_j), \phi(x_l) >_{\mathcal{H}}$. Here $x_i, x_k \in X_s$ and $x_j, x_l \in X_t$; $\beta = [\beta^s, \beta^t]^T$ and $\widetilde{1} = [\underbrace{\frac{1}{n_s}, ..., \frac{1}{n_s}}_{n_s}, \underbrace{-\frac{1}{n_t}, ..., -\frac{1}{n_t}}_{n_t}]^T$.

Incorporating w (2) into $||w||^2$, we have : $||w||^2 = \beta^T K \beta$.

We now introduce the Lagrange parameters to solve this constrained problem:

$$L = \max_{\alpha,\mu,\eta} \min_{\beta,\varepsilon,b} \frac{1}{2} \beta^T K \beta + C \sum_{i=1}^{n_s} \varepsilon_i - \sum_{i=1}^{n_s} \alpha_i \varepsilon_i$$
$$- \sum_{i=1}^{n_s} \mu_i [y_i (\beta^T \phi(X) \phi(x_i) + b) - 1 + \varepsilon_i] - \eta (K \widetilde{1}^T \beta)$$

After some manipulations, we obtain the dual form:

$$\max_{\mu,\eta} \sum_{i=1}^{n_s} \mu_i - \frac{1}{2} (\sum_{i=1}^{n_s} \mu_i y_i K_{.i})^T K^{-1} (\sum_{j=1}^{n_s} \mu_j y_j K_{.j})$$
$$- \frac{1}{2} \eta^2 \widetilde{1}^T K^T \widetilde{1} - \eta (\sum_{i=1}^{n_s} \mu_i y_i K_{.i})^T \widetilde{1}$$
$$s.t. \ 0 \le \mu_i \le C \text{ and } \sum_{i=1}^{n_s} \mu_i y_i = 0$$

where $K_{.i} =< \phi(X), \phi(x_i) >_{\mathcal{H}}$ and X represents the ensemble of X_s and X_t; x_i is a single point either from X_s or X_t.

As there are two different kinds of Lagrange parameters μ and η, we eliminate one by first fixing the value of μ and maximizing only the two latter terms (related with η) of the Lagrange function. The optimal value of η can be expressed as a function of μ: $\eta = -\frac{(\sum_{i=1}^{ns} \mu_i y_i K_{.i})^T \widetilde{1}}{\widetilde{1}^T K^T \widetilde{1}}$. We now obtain the final dual form of the optimization problem:

$$\max_{\mu} \sum_{i=1}^{n_s} \mu_i - \frac{1}{2} (\sum_{i=1}^{n_s} \mu_i y_i K_{.i})^T (K^{-1} - \frac{\widetilde{1}\widetilde{1}^T}{\widetilde{1}^T K^T \widetilde{1}}) (\sum_{j=1}^{n_s} \mu_j y_j K_{.j})$$
$$s.t. \ 0 \le \mu_i \le C \text{ and } \sum_{i=1}^{n_s} \mu_i y_i = 0.$$

Let γ_i denote $\mu_i y_i$, the previous problem becomes:

$$\max_{\gamma} \gamma^T Y - \frac{1}{2} \gamma^T (K_{SS} - \frac{K_{S.} \widetilde{1}\widetilde{1}^T K_{S.}^T}{\widetilde{1}^T K^T \widetilde{1}}) \gamma$$
$$s.t. \ \sum_{i=1}^{n_s} \gamma_i = 0 \text{ and } min(0, Cy_i) \le \gamma_i \le max(0, Cy_i).$$

where $K_{S.} = \sum_{i=1}^{n_s} K_{i.}$. The matrix $K_{SS} - \frac{K_{S.} \widetilde{1}\widetilde{1}^T K_{S.}^T}{\widetilde{1}^T K^T \widetilde{1}}$ is the matrix of inner products (in the subspace orthogonal to w) of source data. As stated in (Paulsen, 2009), if $\mathcal{H}$ is a RKHS on X and $\mathcal{H}_0 \in \mathcal{H}$ is a closed subspace, then $\mathcal{H}_0$ is also a RKHS on X. Therefore, the matrix $K_{new} = K_{SS} - \frac{K_{S.} \widetilde{1}\widetilde{1}^T K_{S.}^T}{\widetilde{1}^T K^T \widetilde{1}}$ is the new Gram matrix corresponding to the projected kernel, K_{new} is positive semi-definite.

Considering the dual form of the optimization problem, we can solve it using standard quadratic programming tools. However, in order to shorten calculations, we used here an adaptation of the F-SVC decomposition algorithm proposed in (Tohmé and Lengellé, 2008). Adaptation and implementation are straightforward.

5 EXPERIMENTS

5.1 Data Sets

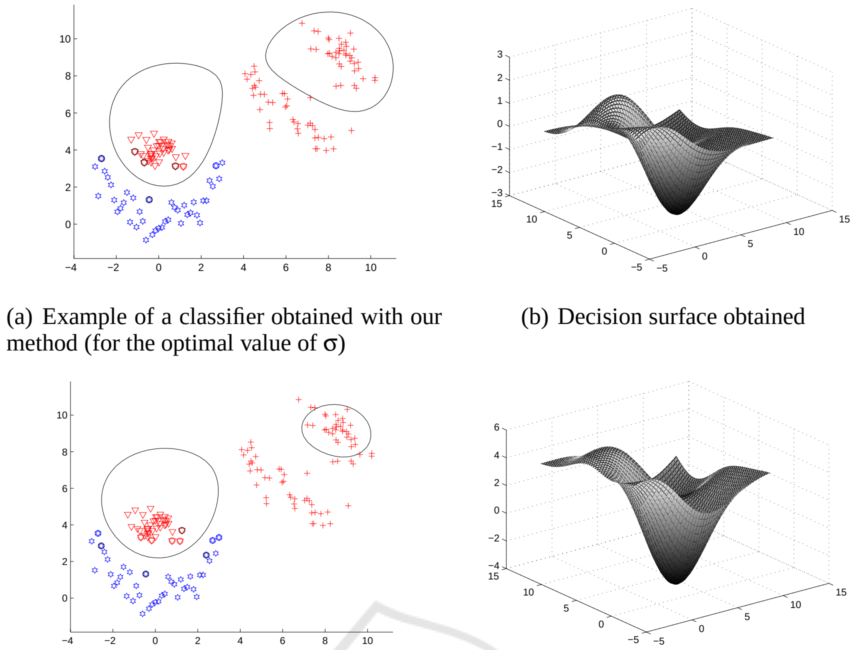
Our goal is to improve the classification performance on target data with the help of related but different source data.

To illustrate our method on a simple data set, we first consider some linearly separable data and we select the linear kernel (which is not universal so the heuristic should not lead to satisfactory results). We generate two almost linearly separable gaussian groups denoted as source-positive and source-negative. Then we do the same to generate the target data (there is no label provided for the target data). An example of this data set is shown in fig. 1.

A second, more complicated synthetic data set is the well-known banana-orange data set. We designate the banana as the source positive and the orange as the source negative. We also generate a target data set which is drawn from a translated and distorted version of the distribution of the source data. Here again, no label information is available for the target (see an example in fig. 3).

We now use the *USPS* data set, a famous handwritten digital number data set. The version used is composed of training and testing parts, both containing the image information (16 * 16 pixels) of 10 different numbers. As proposed in (Uguroglu and Carbonell, 2011), we choose to separate digits 4 and 7 as the source classification problem. All the source data is extracted from the training subset of *USPS* and is perfectly labeled. The target classification problem aims at separating digits 4 and 9 (without the use of the corresponding labels). All target data is extracted from the testing subset of the database *USPS*.

We compare the results we obtained with the method proposed in (Quanz and Huan, 2009) LM and also with standard SVM trained only on source data



(a) Example of a classifier obtained with our method (for the optimal value of σ)

(b) Decision surface obtained

(c) Example of a classifier obtained with LM (for the optimal value of σ)

(d) Decision surface (LM)

Figure 3: Results obtained on the banana-orange data set. In 3(a) and 3(c), circles and stars represent the labeled source data while ”plus” symbols are the unlabeled target data. In 3(b) and 3(d), the decision surfaces are plotted as functions of the input space coordinates. Thresholding these surfaces at 0 level gives the decision curves corresponding to the classifiers in 3(a) and 3(c), respectively.

(no transfer learning in this case). In (Quanz and Huan, 2009), LM has been proved superior to other transfer learning methods so we omit here the comparison to other transfer learning methods.

5.2 Experimental Results and Analysis

For a visual comprehension of our SVM-MMD method, we show in fig. 1 the results obtained on the first synthetic data set. Stars represent source-positive data, triangles are source-negative data, crosses are target data; the two circles are the means of source and target data, respectively. As can be seen, the normal to the obtained discriminant function is orthogonal to $\vec{m}_s - \vec{m}_t$, as expected (for this kernel, the mean of the original source (target) data coincides with μ_s (μ_t).)

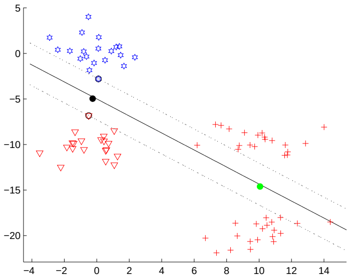


Figure 1: Linearly separable data set using the linear kernel (triangles and stars represent the labeled source data, while ”plus” symbols represent the unlabeled target data).

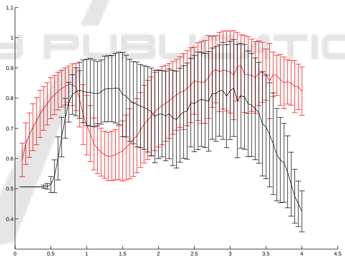


Figure 2: Average performance (good classification rate) ±1 s.d. as a function of the gaussian kernel parameter. Red line : our method. Black line : LM.

For the second synthetic data set (fig. 2), we show the classification result we obtained compared to those of LM. We do not compare with standard SVM on source target data, because obviously standard SVM will fail here (see fig. 3(a)). Example of classification results (data sets, discriminant functions obtained on source and target, decision surfaces) are shown in fig. 3.

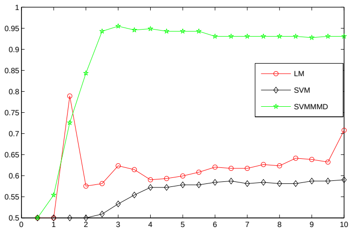


Figure 4: Results (good classification rates) obtained on the *USPS* data set as a function of the gaussian kernel parameter.

We independently generate 50 different banana-orange data sets and show the average performance (±1 standard deviation) in fig. 2. We conclude that most of the time our method achieves better results than LM for a wider range of the kernel parameter value.

We now show the results obtained on the *USPS* data set. As shown in fig. 4, our method provides higher performance for almost all the kernel parameter values considered.

6 CONCLUSION AND FUTURE DIRECTIONS

In this paper, we propose a new approach to solve the domain adaptation problem when no labeled target data is available. The idea is to perform a projection of source and target data onto a subspace of a RKHS where source and target data distributions are expected to be similar. To do so, we select the subspace which ensures nullity of a *Maximum Mean Discrepancy* based criterion. As source and target data become similar, the SVM classifier trained on source data performs well on target data. We have shown that this additional constraint on the primal optimization problem does not modify the nature of the dual problem so that standard quadratic programming tools can be used. We have applied our method on synthetic and real data sets and we have shown that our results compare favorably with Large Margin Transductive Transfer Learning.

As an important short term development, we must propose a method to automatically determine an adequate value of the gaussian kernel parameter used in our paper. We also have to consider multiple kernel learning. Finally, more complex real data sets are to be used to benchmark our transfer learning method.

REFERENCES

Blitzer, J., Dredze, M., Pereira, F., et al. (2007). Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *ACL*, volume 7, pages 440–447.

Bruzzone, L. and Marconcini, M. (2010). Domain adaptation problems: A dasvm classification technique and a circular validation strategy. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 32(5):770–787.

Dudley, R. M. (1984). A course on empirical processes. In *Ecole d'été de Probabilités de Saint-Flour XII-1982*, pages 1–142. Springer.

Dudley, R. M. (2002). *Real analysis and probability*, volume 74. Cambridge University Press.

Fortet, R. and Mourier, E. (1953). Convergence de la répartition empirique vers la réparation théorique. *Ann. Scient. École Norm. Sup.*, pages 266–285.

Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., and Smola, A. (2012). A kernel two-sample test. *J. Mach. Learn. Res.*, 13:723–773.

Huang, C.-H., Yeh, Y.-R., and Wang, Y.-C. F. (2012). Recognizing actions across cameras by exploring the correlated subspace. In *Computer Vision–ECCV 2012. Workshops and Demonstrations*, pages 342–351. Springer.

Huang, J., Gretton, A., Borgwardt, K. M., Schölkopf, B., and Smola, A. J. (2006). Correcting sample selection bias by unlabeled data. In *Advances in neural information processing systems*, pages 601–608.

Jiang, J. (2008). A literature survey on domain adaptation of statistical classifiers. *URL: http://sifaka. cs. uiuc. edu/jiang4/domainadaptation/survey*.

Li, L., Zhou, K., Xue, G.-R., Zha, H., and Yu, Y. (2011). Video summarization via transferrable structured learning. In *Proceedings of the 20th international conference on World wide web*, pages 287–296. ACM.

Liang, F., Tang, S., Zhang, Y., Xu, Z., and Li, J. (2014). Pedestrian detection based on sparse coding and transfer learning. *Machine Vision and Applications*, 25(7):1697–1709.

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

Pan, S. J. and Yang, Q. (2010). A survey on transfer learning. *Knowledge and Data Engineering, IEEE Transactions on*, 22(10):1345–1359.

Patel, V. M., Gopalan, R., Li, R., and Chellappa, R. (2015). Visual domain adaptation: A survey of recent advances. *IEEE signal processing magazine*, 32(3):53–69.

Paulsen, V. I. (2009). An introduction to the theory of reproducing kernel hilbert spaces. *Lecture Notes*.

Quanz, B. and Huan, J. (2009). Large margin transductive transfer learning. In *Proceedings of the 18th ACM conference on Information and knowledge management*, pages 1327–1336. ACM.

Ren, J., Liang, Z., and Hu, S. (2010). Multiple kernel learning improved by mmd. In *Advanced Data Mining and Applications*, pages 63–74. Springer.

Schölkopf, B., Herbrich, R., and Smola, A. J. (2001). A generalized representer theorem. In *Computational learning theory*, pages 416–426. Springer.

Serfling, R. J. (2009). *Approximation theorems of mathematical statistics*, volume 162. John Wiley & Sons.

Smola, A. (2006). Maximum mean discrepancy. In *13th International Conference, ICONIP 2006, Hong Kong, China, October 3-6, 2006: Proceedings*.

Smola, A., Gretton, A., Song, L., and Schölkopf, B. (2007). A hilbert space embedding for distributions. In *Algorithmic Learning Theory*, pages 13–31. Springer.

Steinwart, I. (2002). On the influence of the kernel on the consistency of support vector machines. *The Journal of Machine Learning Research*, 2:67–93.

Tan, Q., Deng, H., and Yang, P. (2012). Kernel mean matching with a large margin. In *Advanced Data Mining and Applications*, pages 223–234. Springer.

Tohmé, M. and Lengellé, R. (2008). F-svc: A simple and fast training algorithm soft margin support vector classification. In *Machine Learning for Signal Processing, 2008. MLSP 2008. IEEE Workshop on*, pages 339–344. IEEE.

Uguroglu, S. and Carbonell, J. (2011). Feature selection for transfer learning. In *Machine Learning and Knowledge Discovery in Databases*, pages 430–442. Springer.

Yang, S., Lin, M., Hou, C., Zhang, C., and Wu, Y. (2012). A general framework for transfer sparse subspace learning. *Neural Computing and Applications*, 21(7):1801–1817.

Zhang, P., Zhu, X., and Guo, L. (2009). Mining data streams with labeled and unlabeled training examples. In *Data Mining, 2009. ICDM'09. Ninth IEEE International Conference on*, pages 627–636. IEEE.

Zhang, Z. and Zhou, J. (2012). Multi-task clustering via domain adaptation. *Pattern Recognition*, 45(1):465–473.