Cross-Domain Sparse Coding

Jim Jing-Yan Wang[1,2]
[1] University at Buffalo, The State University of New York, Buffalo, NY 14203, USA
[2] National Key Laboratory for Novel Software Technology, Nanjing University, Nanjing 210023, China
jimjywang@gmail.com

ABSTRACT

Sparse coding has shown its power as an effective data representation method. However, up to now, all the sparse coding approaches are limited within the single domain learning problem. In this paper, we extend the sparse coding to cross domain learning problem, which tries to learn from a source domain to a target domain with significant different distribution. We impose the Maximum Mean Discrepancy (MMD) criterion to reduce the cross-domain distribution difference of sparse codes, and also regularize the sparse codes by the class labels of the samples from both domains to increase the discriminative ability. The encouraging experiment results of the proposed cross-domain sparse coding algorithm on two challenging tasks — image classification of photograph and oil painting domains, and multiple user spam detection — show the advantage of the proposed method over other cross-domain data representation methods.

Categories and Subject Descriptors

I.2.6 [**ARTIFICIAL INTELLIGENCE**]: Learning

Keywords

Cross-Domain Learning; Sparse Coding; Maximum Mean Discrepancy

1. INTRODUCTION

Traditional machine learning methods usually assume that there are sufficient training samples to train the classifier. However, in many real-world applications, the number of labeled samples are always limited, making the learned classifier not robust enough. Recently, cross-domain learning has been proposed to solve this problem [2], by borrowing labeled samples from a so called "source domain" for the learning problem of the "target domain" in hand. The samples from these two domains have different distributions but are related, and share the same class label and feature space. Two types of domain transfer learning methods have been studied: **classifier transfer** method which learns a classifier for the target domain by the target domain samples with help of the source domain samples, while **cross domain data representation** tries to map all the samples from both source and target domains to a data representation space with a common distribution across domains, which could be used to train a single domain classifier for the target domain [2, 5, 8, 9]. In this paper, we focus on the cross domain representation problem. Some works have been done in this field by various data representation methods. For example, Blitzer et al. [2] proposed the structural correspondence learning (SCL) algorithm to induce correspondence among features from the source and target domains, Daume III [5] proposed the feature replication (FR) method to augment features for cross-domain learning. Pan et al. [8] proposed transfer component analysis (TCA) which learns transfer components across domains via Maximum Mean Discrepancy (MMD) [3], and extended it to semi-supervised TCA (SSTCA) [9].

Recently, sparse coding has attracted many attention as an effective data representation method, which represent a data sample as the sparse linear combination of some codewords in a codebook [7]. Most of the sparse coding algorithms are unsupervised, due to the small number of labeled samples. Some semi-supervised sparse coding methods are proposed to utilize the labeled samples and significant performance improvement has been reported [4]. In this case, it would be very interesting to investigate the use of cross-domain representation to provide more available labeled samples from the source domain. To our knowledge, no work has been done using the sparse coding method to solve the cross-domain problem To fill in this gap, in this paper, we propose a novel cross-domain sparse coding method to combine the advantages of both sparse coding and cross-domain learning. To this end, we will try to learn a common codebook for the sparse coding of the samples from both the source and target domains. To utilize the class labels, a semi-supervised regularization will also be introduced to the sparse codes. Moreover, to reduce the mismatch between the distributions of the sparse codes of the source and target samples, we adapt the MMD rule to sparse codes.

The remaining of this paper is organized as follows: In Section 2, we will introduce the formulations of the proposed Cross-Domain Sparse coding (CroDomSc), and its implementations. Section 3 reports experimental results, and Section 4 concludes the paper.

2. CROSS-DOMAIN SPARSE CODING

In this section, we will introduce the proposed CroDomSc method.

2.1 Objective function

We denote the training dataset with N samples as $\mathcal{D} = \{\mathbf{x}_i\}_{i=1}^N \in \mathbb{R}^D$, where N is the number of data samples, $\mathbf{x}_i$ is the feature vector of the i-th sample, and D is the feature dimensionality. It is also organized as a matrix $X = [\mathbf{x}_1, \cdots, \mathbf{x}_N] \in \mathbb{R}^{D \times N}$. The training set is composed of the source domain set $\mathcal{D}^S$ and target domain set $\mathcal{D}^T$, i.e., $\mathcal{D} = \mathcal{D}^S \bigcup \mathcal{D}^T$. We also denote N^D and N^T as the number of samples in source and target domain set separatively. All the samples from the source domain set $\mathcal{D}^S$ are labeled, while only a few samples from the target domain $\mathcal{D}^T$ are labeled. For each labeled sample $\mathbf{x}_i$, we denote its class label as $y_i \in \mathcal{C}$, where $\mathcal{C}$ is the class label space. To construct the objective function, we consider the following three problems:

Sparse Coding Problem Given a sample $\mathbf{x}_i \in \mathcal{D}$ and a codebook matrix $U = [u_1, \cdots, u_K] \in \mathbb{R}^{D \times K}$, where the k-th column is the k-th codeword and K is the number of codewords in the codebook, sparse coding tries to reconstruct x by the linear reconstruction of the codewords in the codebook as $\mathbf{x}_i \approx \sum_{k=1}^K v_{ki}\mathbf{u}_k = U\mathbf{v}_i$, where $\mathbf{v}_i = [v_{1i}, \cdots, v_{Ki}]^\top \in \mathbb{R}^K$ is the reconstruction coefficient vector for $\mathbf{x}_i$, which should be as sparse as possible, thus $\mathbf{v}_i$ is also called sparse code. The problem of sparse coding can be formulated as follows:

$$\min_{U,V} \sum_{i:\mathbf{x}_i \in \mathcal{D}} \left(\|\mathbf{x}_i - U\mathbf{v}_i\|_2^2 + \alpha\|\mathbf{v}_i\|_1 \right)$$
$$= \|X - UV\|_2^2 + \alpha \sum_{i:\mathbf{x}_i \in \mathcal{D}} \|\mathbf{v}_i\|_1 \qquad (1)$$
$$s.t. \|\mathbf{u}_k\| \leq c, \; k = 1, \cdots, K$$

where $V = [\mathbf{v}_1, \cdots, \mathbf{v}_N] \in \mathbb{R}^{K \times N}$ is the sparse code matrix, with its i-th collum the sparse code of i-th sample.

Semi-Supervised Sparse Coding Regularization In the sparse code space, the intra-class variance should be minimized while the inter-class variance should be maximized for all the samples labeled, from both target and source domains. We first define the semi-supervised regularization matrix as $W = [W_{ij}] \in \{+1, -1, 0\}^{N \times N}$, where

$$W_{ij} = \begin{cases} +1, & if \; y_i = y_j, \\ -1, & if \; y_i \neq y_j, \\ 0, & if \; y_i \; or \; y_j \; is \; unkown. \end{cases} \qquad (2)$$

We define the degree of $\mathbf{x}_i$ as $d_i = \sum_{j:\mathbf{x}_j \in \mathcal{D}} W_{ij}$, $D = diag(d_1, \cdots, d_N)$, and $L = D - W$ as the is the Laplacian matrix. Then we formulate the semi-supervised regularization problem as the following problem:

$$\min_V \sum_{i,j:\mathbf{x}_i,\mathbf{x}_j \in \mathcal{D}} \|\mathbf{v}_i - \mathbf{v}_j\|_2^2 W_{ij}$$
$$= tr[V(D - W)V^\top] = tr(VLV^\top) \qquad (3)$$

In this way, the l_2 norm distance between sparse codes of intra-class pair ($W_{ij} = 1$) will be minimized, while inter-class pair ($W_{ij} = -1$) maximized.

Reducing Mismatch of Sparse Code Distribution To reduce the mismatch of the distributions of the source domain and target domain in the sparse code space, we adopt the MMD [3] as a criterion, which is based on the minimization of the distance between the means of codes from two domains. The problem of reducing the mismatch of the sparse code distribution between source and target domains could be formatted as follows,

$$\min_V \left\| \frac{1}{N_S} \sum_{i:\mathbf{x}_i \in \mathcal{D}^S} \mathbf{v}_i - \frac{1}{N_T} \sum_{j:\mathbf{x}_j \in \mathcal{D}^T} \mathbf{v}_j \right\|^2$$
$$= \|V\boldsymbol{\pi}\|_2^2 = Tr[V\boldsymbol{\pi}\boldsymbol{\pi}^\top V^\top] = Tr(V\Pi V^\top) \qquad (4)$$

where $\boldsymbol{\pi} = [\pi_1, \cdots, \pi_N]^\top \in \mathbb{R}^N$ with π_i the domain indicator of i-th sample defined as

$$\pi_i = \begin{cases} \frac{1}{N_S}, & \mathbf{x}_i \in \mathcal{D}^S, \\ -\frac{1}{N_T}, & \mathbf{x}_i \in \mathcal{D}^T. \end{cases} \qquad (5)$$

and $\Pi = \boldsymbol{\pi}\boldsymbol{\pi}^\top$.

By summarizing the formulations in (1), (3) and (4), the CroDomSc problem is modeled as the following optimization problem:

$$\min_{U,V} \|X - UV\|_2^2 + \beta Tr[VLV^\top] + \gamma Tr[V\Pi V^\top] + \alpha \sum_{i:\mathbf{x}_i \in \mathcal{D}} \|\mathbf{v}_i\|_1$$
$$= \|X - UV\|_2^2 + Tr[VEV^\top] + \alpha \sum_{i:\mathbf{x}_i \in \mathcal{D}} \|\mathbf{v}_i\|_1$$
$$s.t. \|\mathbf{u}_k\| \leq c, \; k = 1, \cdots, K$$
$$\qquad (6)$$

where $E = (\beta L + \gamma \Pi)$.

2.2 Optimization

Since direct optimization of (6) is difficult, an iterative, two-step strategy is used to optimize the codebook U and sparse codes V alternately while fixing the other one.

2.2.1 On optimizing V by fixing U

By fixing the codebook U, the optimization problem (6) is reduced to

$$\min_V \|X - UV\|_2^2 + Tr[VEV^\top] + \alpha \sum_{i:\mathbf{x}_i \in \mathcal{D}} \|\mathbf{v}_i\|_1 \qquad (7)$$

Since the reconstruction error term can be rewritten as $\|X - UV\|_2^2 = \sum_{i:\mathbf{x}_i \in \mathcal{D}} \|\mathbf{x}_i - U\mathbf{v}_i\|_2^2$, and the sparse code regularization term could be rewritten as $Tr[VEV^\top] = \sum_{i,j:\mathbf{x}_i,\mathbf{x}_j \in \mathcal{D}} E_{ij}\mathbf{v}_i^\top \mathbf{v}_j$, (7) could be rewritten as:

$$\min_V \sum_{i:\mathbf{x}_i \in \mathcal{D}} \|\mathbf{x}_i - U\mathbf{v}_i\|_2^2 + \sum_{i,j:\mathbf{x}_i,\mathbf{x}_j \in \mathcal{D}} E_{ij}\mathbf{v}_i^\top \mathbf{v}_j + \alpha \sum_{i:\mathbf{x}_i \in \mathcal{D}} \|\mathbf{v}_i\|_1 \qquad (8)$$

When updating $\mathbf{v}_i$ for any $\mathbf{x}_i \in \mathcal{D}$, the other codes $\mathbf{v}_j (j \neq i)$ for $\mathbf{x}_j \in \mathcal{D}, j \neq i$ are fixed. Thus, we get the following optimization problem:

$$\min_{\mathbf{v}_i} \|\mathbf{x}_i - U\mathbf{v}_i\|_2^2 + E_{ii}\mathbf{v}_i^\top \mathbf{v}_i + \mathbf{v}_i^\top \mathbf{f}_i + \alpha\|\mathbf{v}_i\|_1 \qquad (9)$$

with $\mathbf{f}_i = 2\sum_{j:\mathbf{x}_j \in \mathcal{D}, j \neq i} E_{ii}\mathbf{v}_j$. The objective function in (9) could be optimized efficiently by the modified feature-sign search algorithm proposed in [10].

2.2.2 On optimizing U by fixing V

By fixing the sparse codes V and removing irrelevant terms, the optimization problem (6) is reduced to

$$\min_{U}\|X - UV\|_2^2$$
$$s.t. \|\mathbf{u}_k\|_2^2 \le c, \ k = 1, \cdots, K \tag{10}$$

The problem is a least square problem with quadratic constraints, and it can be solved in the same way as [7].

2.3 Algorithm

The proposed **Cross Domain Sparse coding** algorithm, named as **CroDomSc**, is summarized in Algorithm 1. We have applied the original sparse coding methods to the samples from both the source and target domains for initialization.

Algorithm 1 CroDom-Ss Algorithm

INPUT: Training sample set from both source and target sets $\mathcal{D} = \mathcal{D}^S \bigcup \mathcal{D}^T$;
Initialize the codebooks U^0 and sparse codes V^0 for samples in $\mathcal{D}$ by using single domain sparse coding.
for $t = 1, \cdots, T$ **do**
  for $i : \mathbf{x}_i \in \mathcal{D}$ **do**
    Update the sparse code $\mathbf{v}_i^t$ for $\mathbf{x}_i$ by fixing U^{t-1} and other sparse codes $\mathbf{v}_j^{t-1}$ for $\mathbf{x}_j \in \mathcal{D}, j \ne i$ by solving (9).
  end for
  Update the codebook U^t by fixing the sparse code matrix V^t by solving (10).
end for
OUTPUT: U^T and V^T.

When a test sample from target domain comes, we simply solve problem (9) to obtain its sparse code.

3. EXPERIMENTS

In the experiments, we experimentally evaluate the proposed cross domain data representation method, CroDomSc.

3.1 Experiment I: Cross-Domain Image Classification

In the first experiment, we considered the problem of cross domain image classification of the photographs and the oil paintings, which are treated as two different domains.

3.1.1 Dataset and Setup

We collected an image database of both photographs and oil paintings. The database contains totally 2,000 images of 20 semantical classes. There are 100 images in each class, and 50 of them are photographs, and the remaining 50 ones are oil paintings. We extracted and concatenated the color, texture, shape and bag-of-words histogram features as visual feature vector from each image.

To conduct the experiment, we use photograph domain and oil painting domain as source domain and target domain in turns. For each target domain, we randomly split it into training subset (600 images) and test subset (400 images), while 200 images from the training subset are randomly selected as label samples and all the source domain samples are labeled. The random splits are repeated for 10 times. We first perform the CroDomSc to the training set and use

the sparse codes learned to train a semi-supervised SVM classifier [6]. Then the test samples will also be represented as sparse code and classified using the learned SVM.

3.1.2 Results

We compare our CroDomSc against several cross-Domain data representation methods: SSTCA [9], TCA [8], FR [5] and SCL [2]. The boxplots of the classification accuracies of the 10 splits using photograph and oil painting as target domains are reported in Figure 1. From Figure 1 we can see that the proposed CroDomSc outperforms the other four competing methods for both photograph and oil painting domains. It's also interesting to notice that the classification of the FR and SCI methods are poor, at around 0.7. SSTCA and TCA seems better than FR and SC but are still not competitive to CroDomSc.
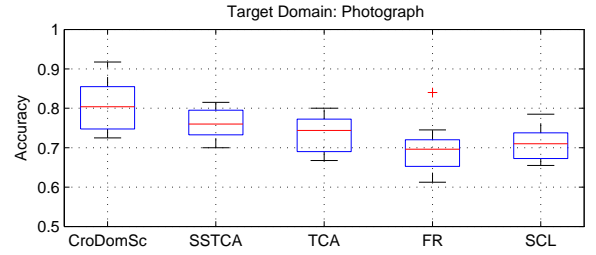
(a) Photograph as target domain
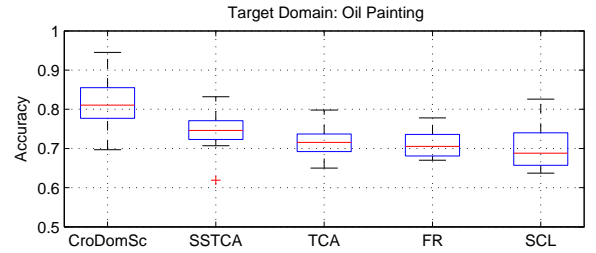
(b) Oil painting as target domain

Figure 1: The boxplot of classification accuracies of 10 splits of CroDomSc and compared methods.

3.2 Experiment II: Multiple User Spam Email Detection

In the second experiment, we will evaluate the proposed cross-domain data representation method for the multiple user based spam email detection.

3.2.1 Dataset and Setup

A email dataset with 15 inboxes from 15 different users is used in this experiment [1]. There are 400 email samples in each inbox, and half of them are spam and the other half non-spam. Due to the significant differences of the email source among different users, the email set of different users could be treated as different domains.

To conduct the experiment, we randomly select two users' inboxes as source and target domains. The target domain will further be split into test set (100 emails) and training set (300 emails, 100 of which labeled, and 200 unlabeled). The source domain emails are all labeled. The word occurrence frequency histogram is extracted from each email as original feature vector. The CroDomSc algorithm was performed to learn the sparse code of both source and target domain samples, which were used to train the semi-supervised classifier. The target domain test samples were also represented as sparse codes, which were classified using the learned classifier. This selection will be repeated for 40 times to reduce the bias of each selection.

3.2.2 Results

Figure 2 shows the boxplots of classification accuracies on the spam detection task. As we can observed from the figure, the proposed CroDomSc always outperforms its competitors. This is another solid evidence of the effectiveness of the sparse coding method for the cross-domain representation problem. Moreover, SSTCA, which is also a semi-supervised cross-domain representation method, seems to outperform other methods in some cases. However, the differences of its performances and other ones are not significant.
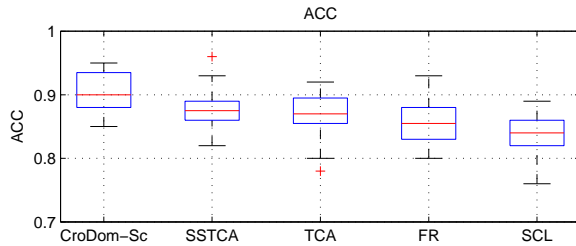

Figure 2: The boxplots of detection accuracies of 40 runs for spam detection task.

4. CONCLUSION

In this paper, we introduce the first sparse coding algorithm for cross-domain data representation problem. The sparse code distribution differences between source and target domains are reduced by regularizing sparse codes with MMD criterion. Moreover, the class labels of both source and target domain samples are utilized to encourage the discriminative ability. The developed cross-domain sparse coding algorithm is tested on two cross-domain learning tasks and the effectiveness was shown.

Acknowledgements

5. REFERENCES

[1] S. Bickel. ECML/PKDD Discovery Challenge 2006, September 2006.

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

[3] K. M. Borgwardt, A. Gretton, M. J. Rasch, H.-P. Kriegel, B. Schoelkopf, and A. J. Smola. Integrating structured biological data by Kernel Maximum Mean Discrepancy. *Bioinformatics*, 22(14):E49–E57, JUL 2006.

[4] L. Cao, R. Ji, Y. Gao, Y. Yang, and Q. Tian. Weakly Supervised Sparse Coding with Geometric Consistency Pooling. In *CVPR 2012*, pages 3578–3585. IEEE, 2012.

[5] H. Daume III. Frustratingly easy domain adaptation. In *ACL 2007 - Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*, pages 256 – 263, 2007.

[6] B. Geng, D. Tao, C. Xu, L. Yang, and X.-S. Hua. Ensemble Manifold Regularization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(6):1227–1233, JUN 2012.

[7] H. Lee, A. Battle, R. Raina, and A. Y. Ng. Efficient sparse coding algorithms. In *NIPS*, pages 801–808. NIPS, 2007.

[8] S. J. Pan, I. W. Tsang, J. T. Kwok, and Q. Yang. Domain Adaptation via Transfer Component Analysis. In *IJCAI 2009*, pages 1187–1192, 2009.

[9] S. J. Pan, I. W. Tsang, J. T. Kwok, and Q. Yang. Domain Adaptation via Transfer Component Analysis. *IEEE Transactions on Neural Networks*, 22(2):199–210, FEB 2011.

[10] M. Zheng, J. Bu, C. Chen, C. Wang, L. Zhang, G. Qiu, and D. Cai. Graph Regularized Sparse Coding for Image Representation. *IEEE Transactions on Image Processing*, 20(5):1327–1336, MAY 2011.