Scientia Iranica D (2017) 24(3), 1303–1311

Sharif University of Technology
Scientia Iranica
Transactions D: Computer Science & Engineering and Electrical Engineering
www.scientiairanica.com

DiReT: An effective discriminative dimensionality reduction approach for multi-source transfer learning

J. Tahmoresnezhad[a,*] and S. Hashemi[b]

a. *Faculty of IT & Computer Engineering, Urmia University of Technology, Urmia, Iran.*
b. *School of Electrical and Computer Engineering, Shiraz University, Shiraz, Iran.*

Received 23 February 2015; received in revised form 16 April 2016; accepted 4 July 2016

KEYWORDS
Multi-source transfer learning;
Domain adaptation;
Discriminative dimensionality reduction;
Fisher discriminant analysis.

Abstract. Transfer learning is a well-known solution to the problem of domain shift in which source domain (training set) and target domain (test set) are drawn from different distributions. In the absence of domain shift, discriminative dimensionality reduction approaches could classify target data with acceptable accuracy. However, distribution difference across source and target domains degrades the performance of dimensionality reduction methods. In this paper, we propose a Discriminative Dimensionality Reduction approach for multi-source Transfer learning, DiReT, in which discrimination is exploited on transferred data. DiReT finds an embedded space, such that the distribution difference of the source and target domains is minimized. Moreover, DiReT employs multiple source domains and semi-supervised target domain to transfer knowledge from multiple resources, and it also bridges across source and target domains to find common knowledge in an embedded space. Empirical evidence of real and artificial datasets indicates that DiReT manages to improve substantially over dimensionality reduction approaches.

© 2017 Sharif University of Technology. All rights reserved.

1. Introduction

In machine learning and pattern recognition, dimensionality reduction is the process of reducing the number of features via obtaining a collection of principle variables [1-3]. Fisher Discriminant Analysis (FDA) [4] and Principal Components Analysis (PCA) [5] are pioneer approaches that concentrate on discovering a low-dimensional latent space. However, there are many real-world applications whose conditions for developing and using the models are different. In this case, the embedding for source and target domains is drawn from different distributions; therefore, the performance of model degrades dramatically.

Domain shift or data shift is a common challenge in real-world applications in which training and test sets have different distributions. This problem arises of a variety of applications, such as computer vision [6-9], multivariate time series [10], and sentiment analysis [11,12].

In this paper, an invariant latent space is extracted to tackle domain shift problem. DiReT, Discriminative Dimensionality Reduction approach for multi-source Transfer learning discovers a latent space, which is discriminative between different classes. DiReT employs Fisher discriminant analysis to find domain invariant features across source and target domains in a semi-supervised manner. Moreover, DiReT bridges source to target domain to transfer knowledge from labeled samples of target domain to the learned model.

In this work, we contribute to the solving of the domain shift problem and show:

i) How to formulate the problem of domain shift;

*. *Corresponding author. Tel.: +98 44 31980236;
Fax: +98 44 31980236
E-mail addresses: j.tahmores@it.uut.ac.ir (J. Tahmoresnezhad); s_hashemi@shirazu.ac.ir (S. Hashemi)*

ii) How to exploit Fisher discriminant analysis to find an embedding;

iii) How to bridge across source and target domains to transfer knowledge;

iv) How to employ multiple source domains to benefit from various resources;

v) How our proposed method outperforms other feature-based state-of-the-art transfer learning approaches.

The rest of this paper is organized as follows. We briefly review the related work in the next section. Section 3 introduces the proposed method and presents the main algorithm. We evaluate our method in variety of datasets with different number of features, instances, and distributions in Sections 4 and 5. This will be followed by conclusion and future work in the last section.

2. Related work

2.1. Transfer learning

In general, domain adaptation approaches are divided into two major categories: unsupervised and semi-supervised [13-15]. In unsupervised domain adaptation, no label is available in target domain. Blitzer et al. [11,16] proposed a structural correspondence learning approach that detects some pivotal features occurring frequently and behaving similarly in both domains. Ben-David et al. [17] presented a solution for feature representation functions that minimizes domain divergence and classification error. Wang and Mahadevan [18] faced with the unsupervised domain adaptation problem by manifold alignment.

In semi-supervised domain adaptation, there are a few labeled instances in the target domain. For the first time, natural language processing community paid attention to the semi-supervised domain adaptation problem. In this regard, Daume III and Marcu [19] detected the shared and individual components of the source and target domains by data distribution modeling. Kumar et al. [20] proposed a co-regularization approach that finds augmented feature space for modeling the source and target domains jointly. Pan et al. [21] presented Transfer Components Analysis (TCA) to project domains onto reproducing kernel Hilbert space. Saenko et al. [22] proposed a metric learning approach to predict domain drift using few labeled data of target domain. Kulis et al. [23] investigated the object detection problem in different vision domains. Tu et al. [24] proposed a Fisher-based discriminative dimensionality reduction approach for the problem of single source and target domains.

In many real-world applications, more than one source domain is available [25-27]. In this case, the focus is on transferring knowledge from multiple source domains or utilizing only the most informative ones. Yang et al. [27] presented a classifier related to target domain by adaptive SVM and multiple-source domains. Duan et al. [28] proposed a multi-kernel approach by different source domains on which optimal linear combination coefficients are learned simultaneously. Hoffman et al. [29] proposed a two-step probabilistic framework, so that, at first, source data are separated into latent clusters, and then a mixture-transform model is employed for adaptation.

2.2. Fisher discriminant analysis

Classical Fisher feature extraction is a famous method to look for linear combinations of variables which best explains the data. FDA explicitly attempts to model the difference between the classes of data by maximizing the between-class scatters and minimizing the within-class scatters in a low-dimensional space [30,31]. Assuming that the training and test samples are drawn from same distributions, classical Fisher maximizes the so-called Fisher criterion, J_F:

$$J_F(W) = (WS_WW^\intercal)^{-1}(WS_BW^\intercal), \tag{1}$$

where W is the output of the FDA that projects the original feature space, k dimension, to the latent feature space, d dimension. In fact, W is a linear mapping, a $d \times k$-matrix, where $d < k$. S_B and S_W are between-class and within-class scatter matrices, respectively. J_F is an optimization problem that is solved by an eigenvalue decomposition of $S_W^{-1}S_B$ [31,32]. In this way, the eigenvectors corresponding to the d largest eigenvalues are composed of the mapping matrix, W.

In this work, we are inspired by FDA to find an embedding in which the distribution difference of source and target domains is reduced. Moreover, we employ the labeled segment of target domain to bridge across domains. Also, multiple source domains are exploited to transfer knowledge from various related domains.

3. The proposed approach

The problem of dimensionality reduction is considered as follows. Suppose that there are N source domains, $\{X^1, X^2, \cdots, X^N\}$, and only one semi-supervised target domain, X^T. C denotes the number of classes in each domain, and X_i^u indicates the subset of instances of domain u that appertains to class i. μ_i^u and p_i^u are mean and class prior probability of subset X_i^u, respectively. The between-class and within-class scatter matrices can be calculated as $S_B = \sum_{i=1}^C p_i(\mu_i - \mu^S)(\mu_i - \mu^S)^\intercal$ and $S_W = \sum_{i=1}^C p_iS_i$, respectively. S_i is within-class covariance matrix, and p_i and μ_i are mean vector and prior of class i, respectively, i.e.:

$$\mu_i = 1/(\sum_{u=1}^{N} p_i^u) \sum_{u=1}^{N} p_i^u \mu_i^u$$

and:

$$p_i = 1/N \sum_{u=1}^{N} p_i^u.$$

The overall mean μ^S in X^S is calculated by:

$$\mu^S = 1/N \sum_{u=1}^{N} \sum_{i=1}^{C} p_i^u \mu_i^u.$$

3.1. Discriminative dimensionality reduction for transfer learning

Since the distribution of the source and target domains is different, the problem of dimensionality reduction in case of domain shift is considered as follows. The training set is composed of the labeled source domains and small labeled segments of target domain, $X_{tr} = \{X^1, X^2, \cdots, X^N, X_{tr}^T\}$. In this way, two types of criteria are considered to find an embedding in which the distribution difference of the source and target domains is minimized:

1. discrimination-based criteria;
2. Transferability-based criteria.

In the former, the main concentration is on the discrimination across various classes in the embedding. In the latter, the main concern is about transferability. In this way, the embedding will have the discrimination and transferability simultaneously. In the rest, the different criteria are investigated in detail.

3.1.1. Discrimination-based criteria

There are two discrimination-based criteria that are discriminated across various classes. Between-class scatter criteria maximize the various class means in the embedding; in parallel, within-class scatter criteria minimize the distance of each projected sample from its mean in the embedding, as well. The between-class scatter criterion, S'_B, is defined as follows:

$$S'_B = \sum_{i=1}^{C-1} \sum_{j=i+1}^{C} \sum_{r=1}^{N} \sum_{s=1}^{N} p_i^r p_j^s (\mu_i^r - \mu_j^s)(\mu_i^r - \mu_j^s)^\intercal, \quad (2)$$

where μ_i^r and μ_j^s specify the mean of i and j classes in r and s domains, respectively. In this way, the distance of various class means from different domains is maximized in the embedding. Moreover, within-class scatter criterion, S_W, is calculated as follows:

$$S_W = \sum_{i=1}^{C} \sum_{x \in X_{tr}^i} p_i (x - \mu_i)(x - \mu_i)^\intercal, \quad (3)$$

where X_{tr}^i denotes the samples of training set that belong to class i. In fact, the within-class scatter criteria cluster the same class samples of source and target domains in the embedding.

3.1.2. Transferability-based criteria

The transferability-based criteria are attempted to transfer knowledge from multiple source domains to a target domain. In this case, within-domain and within-dataset scatter criteria are minimized. In the former, the distance between mean of the same class samples and various domains is minimized in the embedding. In fact, the data points in the embedding with the same class label have minimum distance from each other. The latter minimizes the distance of all dataset domains from each other in the embedding. Within-domain scatter, S'_W, is defined as follows:

$$S'_W = \sum_{i=1}^{C} \sum_{u=1}^{N} \sum_{v=u+1}^{N+1} p_i p_j (\mu_i^u - \mu_i^v)(\mu_i^u - \mu_i^v)^\intercal, \quad (4)$$

where the value of $v = (N+1)$ indicates the labeled segment of target data. In fact, S'_W attempts to minimize the distance of the same classes from different domains against labeled segment of target domain. In this way, a bridge from source data to target data is made, with most similarity to unlabeled data. It is worth noting that the final goal is to predict the labels of unlabeled segment of target data. In this regard, within-dataset scatter criterion, S''_W, is formulated as follows:

$$S''_W = \sum_{u=1}^{N+1} (\mu^u - \mu)(\mu^u - \mu)^\intercal, \quad (5)$$

where $\mu^u = \sum_{i=1}^{C} p_i^u \mu_i^u$ is the mean of domain u, and μ denotes the mean of dataset.

3.2. DiReT

DiReT finds a linear mapping, W, that maximizes the following optimization problem:

$$J'_F(W) = \left(W \left(\gamma S_W + (1-\gamma) \left(\left(1 + \frac{n_l}{n} \right) S'_W + S''_W \right) \right) W^\intercal \right)^{-1} (W S'_B W^\intercal), \quad (6)$$

where n_l is the size of labeled segment of target data, and n denotes the number of samples in the target domain. Thus, we attach more importance to within-domain scatter criteria in which a bridge is made amongst labeled segment of target data and multiple source domains. γ is the control parameter that regularizes discrimination and transferability in the optimization problem. Small values of γ augment the transferability in the embedding, and large values of γ highlight the discrimination in the optimization problem. However, there is no exact value for γ, and it is determined by variety of experiments on different datasets. One straightforward way to estimate the value of γ is cross-validation whose best value is determined based on various experiments.

The main difference of $J'_F(W)$ and $J_F(W)$ belongs to the transferability-based criteria that handle shift problem. However, $J'_F(W)$ is solved using decomposition. The transformation matrix, W, is considered as d eigenvectors corresponding to the d largest eigenvalues of $((\gamma S_W + (1-\gamma)((1+\frac{n_l}{n})S'_W + S''_W))^{-1}S'_B)$.

The number of extracted features of FDA in the embedding is equal to $\min(C-1,k)$; however, DiReT extracts $\min((N+1)*C-1,k)$ features according to the rank of matrix S'_B in the optimization problem. Moreover, DiReT is invariant to the linear mapping of matrix W, similar to FDA.

4. Experimental setup

This section provides details on the experimental setup of the proposed approach. At first, we introduce the artificial and real-world benchmark datasets, and then DiReT is compared to other state-of-the-art transfer learning approaches.

4.1. Data description

Table 1 shows the list and details of artificial and real datasets. The number of samples in the source and target domains is considered the same. The artificial datasets are designed to evaluate DiReT in different conditions of shift. Also, real datasets are considered to show the performance of DiReT in facing with real applications. Subsequently, a short description of datasets is included.

4.1.1. Artificial data

The experiments are conducted on three artificial datasets. The number of source domains is considered two, and each domain contains two classes. Each domain is composed of variant and invariant features. The distribution of invariant features is similar across different source and target domains; however, variant features have different distributions. The number of invariant features is indicated by N, and the variant features are denoted by V. Moreover, V is considered from 1 through 40 where DiReT is evaluated in different conditions.

The dataset *Gau* contains 15 invariant features and 3000 samples. It is composed of two source domains and one target domain with dimension of $1000 \times (15+V)$. According to six different values for variant features, experiments are repeated six times with 16, 20, 25, 35, 45, and 55 features. The invariant features have the same mean and variance; however, variant features are drawn from different distributions with different mean and variance.

4.1.2. Lung dataset

The lung dataset is composed of 30 chest radiographs, obtained from publicly available JSRT dataset [33]. The number of features of each pixel is determined as 10 based on N-jets feature representation [34]. There are three available classes, i.e. lung, rib, and background. In each experiment, one radiograph is considered as source domain, and the rest of radiographs are supposed to be target data.

4.1.3. USPS handwritten digits

The USPS handwritten digits' dataset is considered as the next real dataset. USPS contains images of size 16×16 with pixel values going from 0 to 2. The numerous past works investigated the difficulties in classification between some digits, e.g. separating 4 from 7 and also 4 from 9 [35,36]. USPS contains 10 classes, which are equal to the number of digits.

4.2. Method evaluation

The performance of DiReT is compared with those of other four dimensionality reduction approaches. Since DiReT and other compared approaches are feature extraction methods, we employ SVM and 1-NN classifiers to compare the accuracy/error. PCA is a well-known feature extraction method that exploits covariance matrix to find principal components. Despite good performance of PCA in different applications, it shows

Table 1. Artificial and real-world benchmark datasets. In artificial datasets, the distribution property indicates the distribution of source and target data. For example, the distribution of source data of *UniPoi* dataset is *Uniform*, and the distribution of target data is the mixture of *Uniform* and *Poison*. Moreover, N denotes the number of invariant features across domains in which the number of variant features could be different.

Dataset	**Type**	**Distribution**	**Number of examples**	**Number of invariant features (N)**
Gau	Synthetic	Gaussian	1000	15
UniPoi	Synthetic	Uniform, Poison	1200	20
WeiGeo	Synthetic	Weibull, Geometric	800	25
lung	Real	—	30	—
USPS	Real	—	1470	—

low performance against FDA. In fact, PCA finds principal components in a fully unsupervised manner, i.e. without considering the label of samples.

TCA is a dimensionality reduction approach in which the shift problem is considered, as well. TCA has been inspired from PCA to tackle domain shift problem in dataset. The main drawback of TCA is similar to PCA, i.e. unsupervised transformation. f-MMD is another dimensionality reduction approach that employs feature selection instead of feature extraction. f-MMD assigns a weight to each feature according to its variation across source and target domains.

5. Experimental results and discussion

5.1. Artificial dataset

The experiments are conducted on three artificial datasets to compare the performance of DiReT with those of other dimensionality reduction approaches. Table 2 shows the performance of different approaches on various conditions, i.e. various number of variant features. As is clear from the results, transfer learning-based approaches (TCA, f-MMD, and DiReT) show a better performance against other dimensionality reduction methods. Also, DiReT preserves its accuracy even with increasing number of variant features.

PCA and FDA are high-performing approaches on non-shifted data; however, as is clear from results, they fail to face with the shifted data. Thus, the performance of PCA and FDA degrades with increasing the number of variant feature, i.e. random classifier. However, FDA shows a better performance than that of PCA, because it transforms data according to the discrimination criteria.

DiReT benefits from more number of features against other dimensionality reduction approaches such as FDA due to its optimization problem. Thus, DiReT reflects more properties and statistics from original dataset, and it preserves the geometric properties of source and target data. In addition, DiReT bridges across target and source domains using labeled segment of target data. In this case, distance of the samples with the same class label reduces from the target data. Tables 3 and 4 show the same results according to the above discussions.

Table 2. Error rates on *Gau* dataset using 1-NN classifier. The number of invariant features is 15, and the number of variant features is changing from 1 to 40.

Methods	$V = 1$	$V = 5$	$V = 10$	$V = 20$	$V = 30$	$V = 40$
PCA	28.1 ± 1.7	40.6 ± 2.2	45.7 ± 1.2	48.3 ± 0.9	47.1 ± 1.3	48.2 ± 2.7
TCA	23.9 ± 0.4	22.6 ± 1.4	24.1 ± 0.6	24.5 ± 1.2	23.1 ± 0.9	24.0 ± 0.4
f-MMD	19.8 ± 0.7	20.3 ± 0.9	25.4 ± 1.2	23.1 ± 0.5	21.5 ± 2.1	22.3 ± 0.7
FDA	17.5 ± 0.3	26.0 ± 0.5	35.7 ± 1.2	39.1 ± 1.9	42.3 ± 0.5	44.4 ± 0.7
DiReT	13.9 ± 0.3	14.1 ± 0.5	11.3 ± 1.0	15.9 ± 0.8	13.3 ± 0.2	13.6 ± 0.4

Table 3. Error rates on *UniPoi* dataset using 1-NN classifier. The number of invariant features is 20, and the number of variant features is changing from 1 to 40.

Methods	$V = 1$	$V = 5$	$V = 10$	$V = 20$	$V = 30$	$V = 40$
PCA	32.6 ± 1.2	41.7 ± 0.9	47.9 ± 1.7	46.1 ± 2.4	50.2 ± 2.7	51.3 ± 1.4
TCA	25.0 ± 0.4	27.2 ± 0.9	27.6 ± 1.4	26.3 ± 1.1	25.5 ± 0.6	29.2 ± 0.7
f-MMD	21.4 ± 0.7	19.2 ± 0.8	22.9 ± 1.1	23.5 ± 0.3	24.1 ± 1.9	27.2 ± 1.4
FDA	19.1 ± 0.3	22.3 ± 0.8	30.6 ± 1.9	41.1 ± 1.2	45.0 ± 0.8	43.2 ± 2.1
DiReT	12.5 ± 0.2	13.2 ± 0.5	15.0 ± 0.7	14.7 ± 0.6	12.3 ± 0.9	12.9 ± 0.3

Table 4. Error rates on *WeiGeo* dataset using 1-NN classifier. The number of invariant features is 25, and the number of variant features is changing from 1 to 40.

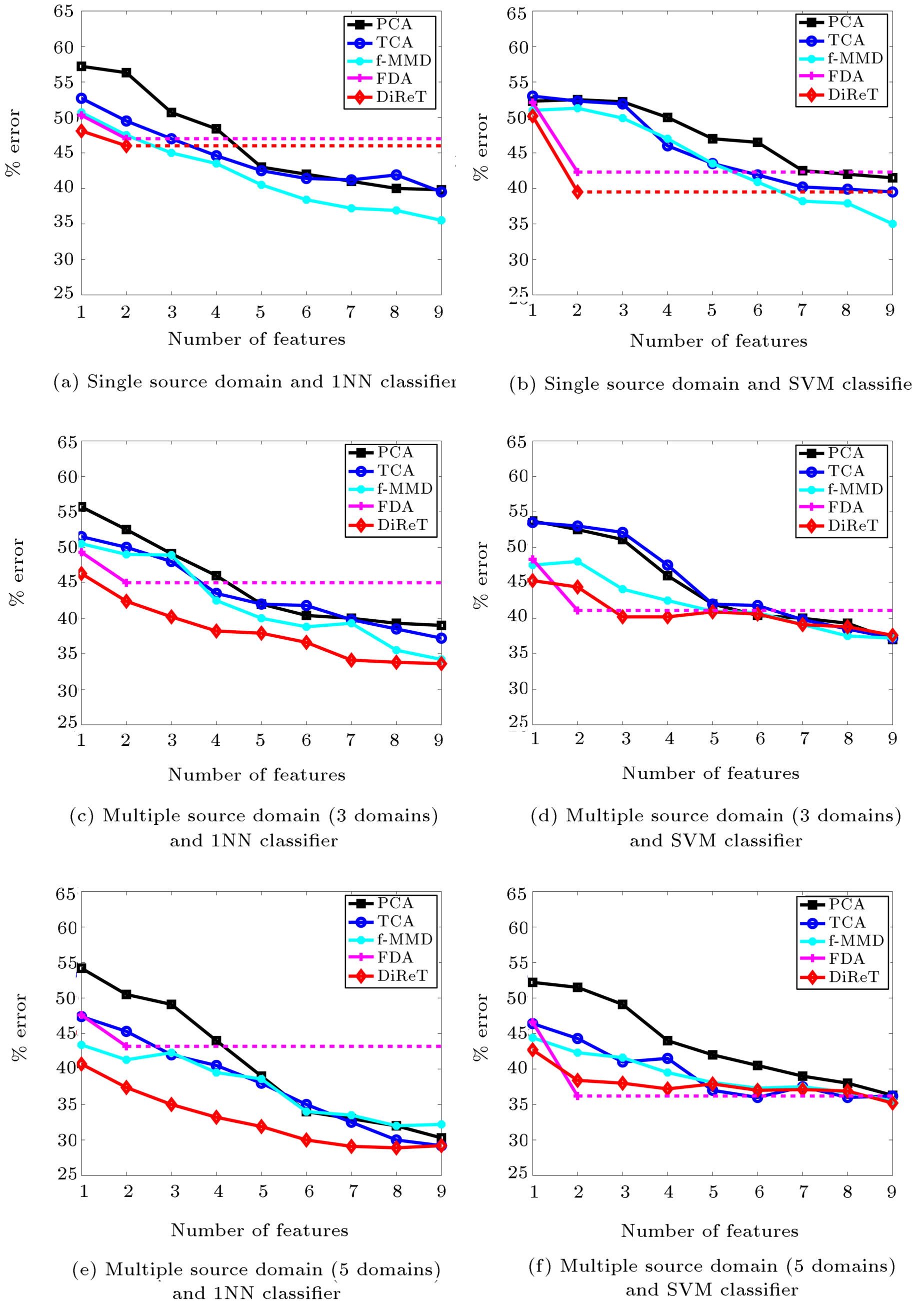
Methods	$V = 1$	$V = 5$	$V = 10$	$V = 20$	$V = 30$	$V = 40$
PCA	30.9 ± 2.2	33.7 ± 1.4	35.8 ± 2.7	45.1 ± 1.8	47.2 ± 0.9	50.1 ± 1.2
TCA	27.1 ± 0.8	27.5 ± 0.6	31.0 ± 1.2	26.5 ± 1.4	27.2 ± 0.9	26.3 ± 1.3
f-MMD	21.2 ± 1.0	21.5 ± 1.6	25.2 ± 0.7	25.8 ± 1.2	24.7 ± 1.5	28.3 ± 0.7
FDA	20.5 ± 1.2	25.1 ± 0.7	29.9 ± 1.5	38.2 ± 1.7	43.5 ± 2.6	41.4 ± 2.3
DiReT	11.2 ± 0.7	12.6 ± 0.3	11.2 ± 0.8	10.1 ± 0.4	10.7 ± 0.5	10.1 ± 1.1

5.2. Real data

The experiments on real-world datasets are conducted with three different scenarios. In the first scenario, a single-source domain is considered against a semi-supervised target domain. In this case, the proposed approach only benefits from one resource to transfer knowledge. In the second scenario, the knowledge is transferred from three-source domain to a semi-supervised target domain. Finally, in the last scenario, DiReT benefits from five-source domain, where it is expected to show a better performance.

In the multiple scenarios, the data are divided into different partitions; at each step, one partition is selected for training, and the rest are used for testing. The reported error is the average classification error over all partitions. In general, in multiple scenarios, we try to test the effect of presence of multiple information in the performance of our method. Instead, in the single scenario, only one set is used for training and the test is done on the target data.

Figures 1 and 2 depict the classification error rates on lung and USPS datasets, respectively. In



(a) Single source domain and 1NN classifier

(b) Single source domain and SVM classifie

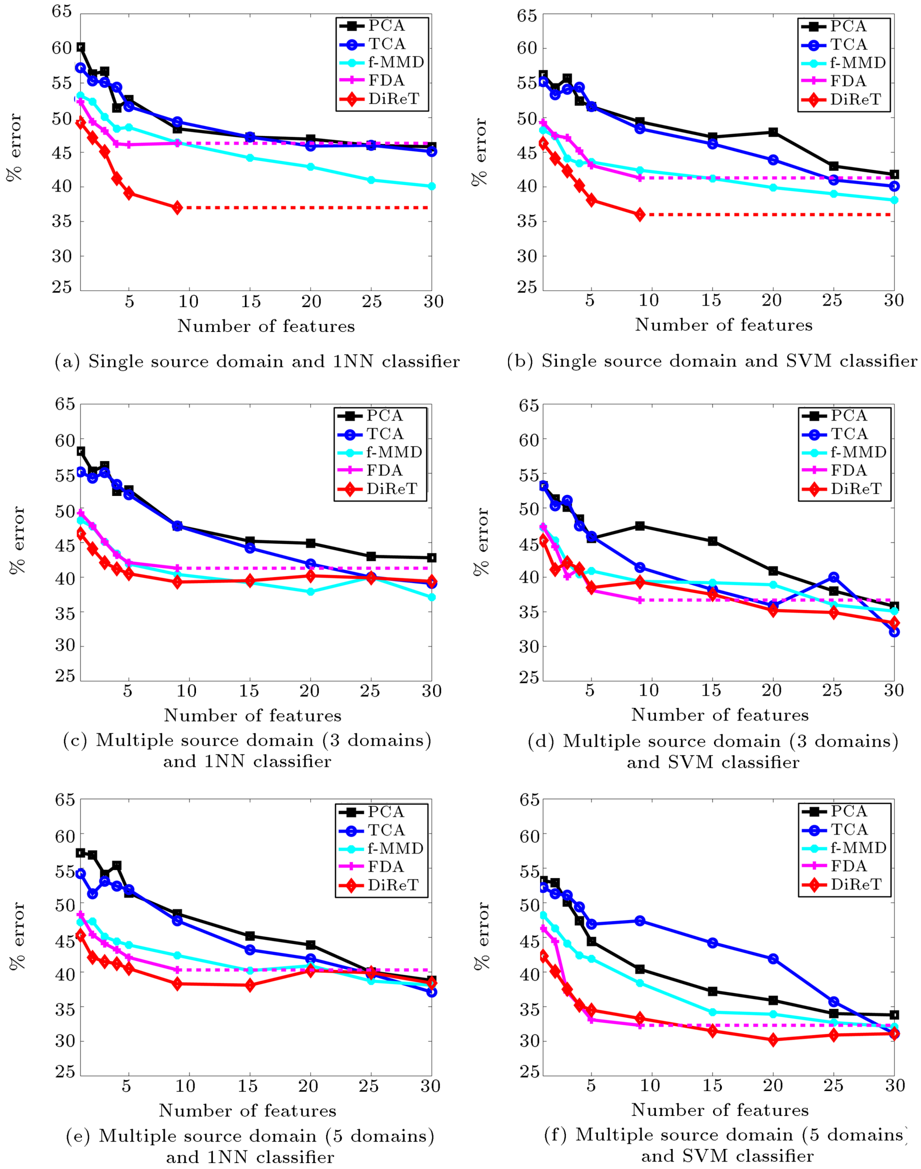
(c) Multiple source domain (3 domains) and 1NN classifier

(d) Multiple source domain (3 domains) and SVM classifier

(e) Multiple source domain (5 domains) and 1NN classifier

(f) Multiple source domain (5 domains) and SVM classifier

Figure 1. Error rates on different scenarios for lung dataset. The dotted lines in DiReT and FDA show that the number of extracted features is less than the number of depicted features on the horizontal axis.



(a) Single source domain and 1NN classifier

(b) Single source domain and SVM classifier

(c) Multiple source domain (3 domains) and 1NN classifier

(d) Multiple source domain (3 domains) and SVM classifier

(e) Multiple source domain (5 domains) and 1NN classifier

(f) Multiple source domain (5 domains) and SVM classifier

Figure 2. Error rates on different scenarios for USPS dataset. The dotted lines in DiReT and FDA show that the number of extracted features is less than the number of depicted features on the horizontal axis.

these experiments, the performance of feature extraction methods has been evaluated against the different number of extracted features. Each panel contains six figures, which shows the different scenarios and classification algorithms. The first row of the panel indicates single-domain scenario, and the second and third rows show multiple scenarios with three and five source domains, respectively. Also, the left column of panel shows the accuracy of the 1-NN classifier, and the right columns indicate the SVM classifiers' performance.

Since the number of classes in lung and USPS datasets is 3 and 10, respectively, FDA could only extract 2 and 9 features in the embedding according to different scenarios. However, DiReT extracts more features in multiple scenarios with regard to the rank of scatter matrices. In this case, DiReT has more knowledge to transfer across domains.

6. Conclusion and future work

In this paper, we have proposed a novel approach for tackling domain shift problem. Our approach is the generalization of Fisher discriminant analysis that copes with the transfer learning. In this work, we introduced DiReT, a Discriminative Dimensionality Reduction approach for multi-source Transfer learning, which maps source and target domains to embedding in a semi-supervised manner. DiReT maximizes the between-class scatters and concurrently minimizes the within-class and within-domains scatters. On benchmark tasks in artificial and real world, DiReT consistently outperforms other dimensionality reduction methods. For the future work, we plan to advance this direction further and extend DiReT to online transfer learning.

References

1. Gopalan, R., Li, R. and Chellappa, R. "Unsupervised adaptation across domain shifts by generating intermediate data representations", *IEEE T Pattern Anal*, **36**(11), pp. 2288-2302 (2014).
2. Xiong, C., McCloskey, S., Hsieh, S.-H. and Corso, J.J. "Latent domains modeling for visual domain adaptation", *23rd Conf. on Artif. Intel.*, Chicago, USA, pp. 2860-2866 (2008).
3. Tahmoresnezhad, J. and Hashemi, S. "Common feature extraction in multi-source domains for transfer learning", *7th IEEE Int. Conf. on Infor. and Know. Tech.*, Urmia, Iran, pp. 1-5 (2015).
4. Fisher, R.A. "The use of multiple measurements in taxonomic problems", *Ann Eugenic*, **7**(2), pp. 179-188 (1936).
5. Jolliffe, I., *Principal Component Analysis*, Wiley Online Library (2002).
6. Duan, L., Tsang, I.W., Xu, D. and Chua, T.-S. "Domain adaptation from multiple sources via auxiliary classifiers", *26th Int. Conf. on Mach. Learn.*, Montreal, Canada, pp. 289-296 (2009).
7. Pan, S.J., Tsang, I., Kwok, J.T. and Yang, Q. "Domain adaptation via transfer component analysis", *IEEE T Neural Networ*, **22**(2), pp. 199-210 (2011).
8. Gong, B., Shi, Y., Sha, F. and Grauman, K. "Geodesic flow kernel for unsupervised domain adaptation", *IEEE Int. Conf. on Comp. Vis.*, RI, USA, pp. 2066-2073 (2012).
9. Tahmoresnezhad, J. and Hashemi, S. "Visual domain adaptation via transfer feature learning", *Knowl Inf Syst.*, pp. 1-21 (2016).
10. Vidaurre, C., Kawanabe, M., Von Bunau, P., Blankertz, B. and Muller, K.R. "Toward unsupervised adaptation of LDA for brain-computer interfaces", *IEEE T Bio-Med Eng*, **58**(3), pp. 587-597 (2011).
11. Blitzer, J., Dredze, M. and Pereira, F. "Biographies, Bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification", *ACL*, **7**, pp. 440-447 (2007).
12. Calais Guerra, P.H., Veloso, A., Meira, Jr, W. and Almeida, V. "From bias to opinion: a transfer-learning approach to real-time sentiment analysis", *17th ACM SIGKDD Int. Conf. on Know. Disc. and Data Mining*, CA, USA, pp. 150-158 (2011).
13. Tahmoresnezhad, J. and Hashemi, S. "A generalized kernel-based random k-samplesets method for transfer learning", *Iran J Sci Technol Trans. of Elec. Eng.*, **39**, pp. 193-207 (2015).
14. Long, M., Wang, J., Sun, J. and Yu, P.S. "Domain invariant transfer kernel learning", *IEEE T Knowl Data En*, **27**(6), pp. 1519-1532 (2015).
15. Lu, J., Behbood, V., Hao, P., Zuo, H., Xue, S. and Zhang, G. "Transfer learning using computational intelligence: A survey", *Knowl-Based Syst*, **80**, pp. 14-23 (2015).
16. Blitzer, J., McDonald, R. and Pereira, F. "Domain adaptation with structural correspondence learning", *Conf. on Emp. Meth. in Natu. Lang. Proc.*, Sydney, Australia, pp. 120-128 (2006).
17. Ben-David, S., Blitzer, J., Crammer, K. and Pereira, F. "Analysis of representations for domain adaptation", *20th Conf. on ADV NEUR IN*, Vancouver, Canada, pp. 137-144 (2007).
18. Wang, C. and Mahadevan, S. "Manifold alignment without correspondence", *21st Int. Joint Conf. on Artif. Intel.*, CA, USA, pp. 1273-1278 (2009).
19. Daumé III, H. and Marcu, D. "Domain adaptation for statistical classifiers", *J. Artif. Intell. Res.*, **26**, pp. 101-126 (2006).
20. Kumar, A., Saha, A. and Daume, H. "Co-regularization based semi-supervised domain adaptation", *23rd Conf. on ADV NEUR IN*, Vancouver, Canada, pp. 478-486 (2010).
21. Pan, S.J., Tsang, I.W., Kwok, J.T. and Yang, Q. "Domain adaptation via transfer component analysis", *IEEE T Neural Networ*, **22**(2), pp. 199-210 (2011).
22. Saenko, K., Kulis, B., Fritz, M. and Darrell, T. "Adapting visual category models to new domains", *11th Eur. Conf. on Comp. Vis.*, Crete, Greece, pp. 213-226 (2010).
23. Kulis, B., Saenko, K. and Darrell, T. "What you saw is not what you get: Domain adaptation using asymmetric kernel transforms", *IEEE Int. Conf. on Comp. Vis. and Patt. Recog.*, Colorado Springs, USA, pp. 1785-1792 (2011).
24. Tu, W. and Sun, S. "Transferable discriminative dimensionality reduction", *23rd IEEE Int. Conf. on Tools with Artif. Intel.*, Florida, USA, pp. 865-868 (2011).
25. Mansour, Y., Mohri, M. and Rostamizadeh, A. "Domain adaptation with multiple sources", *22nd Conf. on ADV NEUR IN*, Vancouver, CANADA, pp. 1041-1048 (2009).
26. Dredze, M. and Crammer, K. "Online methods for multi-domain learning and adaptation", *Conf. on Emp. Meth. in Natu. Lang. Proc.*, Waikiki, Hawaii, pp. 689-697 (2008).
27. Yang, J., Yan, R. and Hauptmann, A.G. "Cross-domain video concept detection using adaptive svms", *15th Int. Conf. on Multimedia*, Augsburg, Germany, pp. 188-197 (2007).
28. Duan, L., Tsang, I.W. and Xu, D. "Domain transfer multiple kernel learning", *IEEE T. Pattern Anal.*, **34**(3), pp. 465-479 (2012).

29. Hoffman, J., Kulis, B., Darrell, T. and Saenko, K. "Discovering latent domains for multisource domain adaptation", *13th Eur. Conf. on Comp. Vis.*, Firenze, Italy, pp. 702–715 (2012).

30. Zhu, X. and Goldberg, A.B. "Introduction to semi-supervised learning", *Synthesis Lectures on Artificial Intelligence and Machine Learning*, **3**(1), pp. 1-130 (2009).

31. Tian, X., Tao, D. and Rui, Y. "Sparse transfer learning for interactive video search reranking", *Acm. T. Miltim. Comput.*, **8**(3), pp. 1-17 (2012).

32. Mesnil, G., Dauphin, Y., Glorot, X., et al. "Unsupervised and transfer learning challenge: A deep learning approach", *J. Mach. Learn. Res.*, **7**, pp. 97-110 (2012).

33. Shiraishi, J., Katsuragawa, S., Ikezoe, J., et al. "Development of a digital image database for chest radiographs with and without a lung nodule: receiver operating characteristic analysis of radiologists' detection of pulmonary nodules", *AM J. Roentgenol.*, **174**(1), pp. 71-74 (2000).

34. Koenderink, J.J. and van Doorn, A.J. "Representation of local geometry in the visual system", *Biol. Cybern.*, **55**(6), pp. 367-375 (1987).

35. Sugiyama, M., Id, T., Nakajima, S. and Sese, J. "Semi-supervised local Fisher discriminant analysis for dimensionality reduction", *Mach. Learn.*, **78**(1-2), pp. 35-61 (2010).

36. Zeng, H. and Cheung, Y.-M. "Feature selection for local learning based clustering", *13th Pacific-Asia Know. Dis. and Data Mining Conf.*, Pernambuco, Brazil, pp. 414-425 (2009).

Biographies

Jafar Tahmoresnezhad received his PhD degree in Computer Engineering from Shiraz University, Shiraz, Iran, in 2015. Following academic appointments in Urmia University of Technology, he is currently an Assistant Professor at Faculty of Information Technology and Computer Engineering, Urmia, Iran. His research interests include pattern recognition, transfer learning, deep learning, data mining, and computer security.

Sattar Hashemi received his PhD degree in Computer Science from Iran University of Science and Technology in conjunction with Monash University, Australia, in 2008. Following academic appointments in Shiraz University, he is currently an Associate Professor at Electrical and Computer Engineering School, Shiraz University, Shiraz, Iran. He is recognized for his contributions in the fields of machine learning and data mining. He has published many refereed papers and book chapters on data stream classification, game theory, social networks, database intrusion detection, and computer security.