Scientia Iranica D (2016) 23(3), 1239–1250

SCIENTIA
IRANICA

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

Transductive transfer learning via maximum margin criterion

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 8 April 2015; received in revised form 5 January 2016; accepted 13 February 2016

KEYWORDS
Transductive transfer learning;
Unsupervised domain adaptation;
Maximum margin criterion;
Maximum mean discrepancy.

Abstract. In this paper, we propose a transductive transfer learning framework, referred to as Transfer Maximum Margin Criterion (T-MMC). This framework is suitable to transfer the knowledge acquired in one domain, the source domain, to another domain, the target domain, where no labeled examples are available in the target domain. We introduce an effective feature weighting approach, which proceeds to reduce the domain difference between the source and target domains. Moreover, we exploit maximum margin criterion to well discriminate various classes in the reduced domains. We simultaneously transfer knowledge from the source domain to target domain and also discriminate various classes in the reduced domains. Comprehensive experiments on the synthetic and real datasets demonstrate that T-MMC outperforms existing transfer learning methods.

© 2016 Sharif University of Technology. All rights reserved.

1. Introduction

In some machine learning problems, in contrast to finding a general prediction rule, the predictions are done only on some constant number of known test points, which is referred to as transductive setting [1-3]. In this case, the learning problem is changed to a particular type of semi-supervised learning where the learning problem is allowed to exploit the location of test points. Transductive Support Vector Machines (TSVMs) exploit the test points in computation of the margin based on the idea of transductive learning [4]. The obtained margin on the test points provides prior knowledge for learning. Despite high performance of TSVM, it suffers from some issues as follows [5]:

1. Increase in the number of labeled sets has not significant improvement for the performance of model accordingly;
2. With decrease in the size of extremely small unlabeled samples in target domain, the time complexity goes correspondingly high;
3. TSVM shows some instability in some cases so that the results of different runs are not the same.

The high dimensionality of input data causes serious challenges to most of learning tasks, specifically the curse of dimensionality. A prevalent method to tackle this issue is dimensionality reduction, which has been exploited in machine learning and data mining since past decades. The dimensionality reduction approaches are mainly categorized into two distinct ways, which are feature extraction and feature selection. In the former, new features are extracted through algebraic transformation. In the latter, subsets of features are selected from original feature space.

Conventional feature selection algorithms for machine learning and data mining have good performance under the assumption that the training and testing sets are from the same distribution, i.e. independent and identically distributed (i.i.d). However, for most of the real world applications, this assumption is violated

*. *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)

through domain difference of training and test sets, and it seriously reduces the performance of conventional approaches [6,7]. Transductive transfer learning is an effective solution for the problem of domain shift where the training data from source domain and the testing data from the target domain follow different distributions [8,9]. In fact, transductive transfer learning looks for some common structures and knowledge across source and target domains to utilize them as a bridge for transference [10,11].

In general, transfer learning approaches are divided into three settings, i.e. inductive transfer learning, unsupervised transfer learning, and transductive transfer learning [12]. In the inductive transfer learning setting, the source domain labels could be available or unavailable, whereas target domain labels are certainly available. In the unsupervised transfer learning setting, no label is available in the source and target domains. Our proposed approach belongs to the transductive transfer learning setting on which source domain labels are available and there is no label in the target domain.

In this paper, we address the challenging setting in which the source and target data have different distributions. We therefore propose a novel approach, referred to as Transfer Maximum Margin Criterion (T-MMC), to jointly perform transfer learning and discrimination across source and target domains. Specifically, we implement feature weighting by minimizing nonparametric Maximum Mean Discrepancy (MMD) [13] in a Reproducing Kernel Hilbert Space (RKHS) and maximizing discrimination parameter, i.e. Maximum Margin Criterion (MMC) [14]. Therefore, input feature space is categorized into the variant and invariant feature sets. The former is the set of variant features that are violated across the domain. The latter is a set of features that are invariant between the source and target domains.

Contributions: The main contributions of this paper include the following:

1. We successfully extend the traditional machine learning and data mining algorithms, such as MMD and MMC, to solve the transfer learning problems;
2. To tackle the significant distribution difference between the source and target domains, T-MMC minimizes the distribution distance via an important criterion that is MMD;
3. We formulate an optimization problem that concurrently minimizes the domain difference and maximizes margin criterion.

Organization of the paper: The rest of the paper is organized as follows. In Section 2, the previous related works are discussed and the preliminaries, including MMD and MMC, are introduced. We present our proposed approach for transductive transfer learning and corresponding solutions in Section 3. The experimental results of both synthetic and real world benchmark datasets are discussed in Sections 4 and 5. Finally, we draw a conclusion and present the future work.

2. Previous works and preliminaries

2.1. Transfer learning

According to the literature survey [12], existing transfer learning approaches can roughly be organized into four categories that are instance-based transfer learning, parameter-based transfer learning, relational-knowledge transfer learning, and feature-based transfer learning.

The instance-based transfer learning methods reweight or select some data in source domain to reduce the distance between the source and target domains. TrAdaBoost [15], TrAdaBoost.R2 [16], KLIEP [17], and TransferBoost [18] are the representative instance-based transfer learning methods.

In parameter-based transfer learning approaches, it is assumed that hyper-parameters of the model are transferred from the source to the target task. Thus, source and target domains share common knowledge and priors to bridge across domains. In this category, MI-IVM [19], GPDRTL [20], and TLVM [21] are the most representative algorithms.

In relational-knowledge transfer learning approaches, it is assumed that source and target data have some similar relationship. MLN [22] is a statistical relational learning method that is representative in this area.

In feature-based transfer learning, a shared feature space is discovered, on which the distribution difference between the source and target domains is reduced. The shared feature space can be created in the projected latent space [23-25] or original feature space [26,27].

Our work belongs to the category of feature-based transfer learning. Thus, we focus on some previous known algorithms in this category as follows. Blitzer et al. [27] proposed Structured Correspondence Learning (SCL), which finds pivot features, high frequency, and similar meaning features in source and target domains. Pivot features are employed to map non-pivot features onto each other from the unlabeled data of the source and target domains. In a follow-up work, Blitzer et al. [26] exploited Mutual Information (MI) to find the pivot features against employing heuristic criteria. In this way, the dependency between labels and source domain samples is considered in finding pivot features. Pan et al. [28] proposed Maximum Mean Discrepancy Embedding (MMDE), which is a transductive transfer learning method. MMDE is a dimensionality reduction approach that reduces the difference of distributions between source and target domains; however, the computational complexity of MMDE is high. In a follow-up work, Pan et al. [23] proposed Transfer Component Analysis (TCA) to resolve the drawback of MMDE.

From another view [29] and in line with answering the fundamental question of transfer learning and how to transfer the knowledge, transfer learning approaches are categorized into three different styles: adaptive knowledge transfer, collective knowledge transfer, and integrative knowledge transfer. The aim of adaptive knowledge transfer is to adapt auxiliary domain knowledge for the target domain. Collective knowledge transfer jointly learns the shared knowledge and unshared effect of source and target data, simultaneously. Integrative knowledge transfer is the incorporation of raw knowledge of source domains into the learning task of target domain as the known knowledge. Thus, it is noteworthy that including raw data of source domain instead of extracted knowledge of it makes this method different from the adaptive knowledge transfer style. Our proposed approach belongs to the integrative knowledge transfer category in which the knowledge is transferred in the original space and does not project domains into a latent space.

2.2. Maximum Mean Discrepancy (MMD)

In this work, we intend to measure dissimilarity between two probability distributions of the source domain s and the target domain t. There are dozens of methods to measure the distance between two distributions, e.g. *Kullback-LeiblerDivergence (KLD)* which is a widely used method to measure the difference between domains. KLD suffers from the expensive density calculation and non-symmetry. MMD is a non-parametric criterion that compares the distributions of two domains by mapping the data on a rich Reproducing Kernel Hilbert Space. Given two distributions s and t, MMD is defined as:

$$\mathrm{MMD}(X_s, X_t, F) = \sup(E[f(x_s)] - E[f(x_t)]), \quad (1)$$

where X_s and X_t are the source and target datasets, respectively, and $E[f(x_s)]$ and $E[f(x_t)]$ are expectations under distributions of s and t, in turn. F is defined as a rich class of functions, e.g. unit ball in the universal RKHS. $\mathrm{MMD}(X_s, X_t, F)$ tends towards zero if and only if $s = t$. The main idea is that if the feature means of domains under RKHS are close to each other, the distributions of domains will be close in the original space [30]. $X_s = \{x_s^1, x_s^2, \cdots, x_s^n\}$ and $X_t = \{x_t^1, x_t^2, \cdots, x_t^m\}$ are defined as the observations drawn i.i.d. from s and t, respectively. An empirical estimate of M M D can be calculated as:

$$D(X_s, X_t) = \left\| \frac{1}{n}\sum_{i=1}^{n} \Phi(x_s^i) - \frac{1}{m}\sum_{j=1}^{m} \Phi(x_t^j) \right\|_H, \quad (2)$$

where n and m are the numbers of source and target samples, respectively, and $\Phi(x)$ is the feature map defined as $\Phi(x) : X \to H$, where H is a universal RKHS [23]. If the universal kernel associated with this mapping is defined as $k(z_i, z_j^T) = \Phi(z_i)\Phi(z_j^T)$ according to Baktashmotlagh et al., 2013, the distance can be rewritten as:

$$D(X_s, X_t) = \left(\sum_{i=1}^{n}\sum_{j=1}^{n} \frac{k(x_s^i, x_s^j)}{n^2} + \sum_{i=1}^{m}\sum_{j=1}^{m} \frac{k(x_t^i, x_t^j)}{m^2} - 2\sum_{i=1}^{n}\sum_{j=1}^{m} \frac{kt(x_s^i, x_t^j)}{nm} \right)^{\frac{1}{2}}. \quad (3)$$

In a nutshell, MMD between the distributions of two sets of observations is equivalent to the distance between the sample means in a high-dimensional feature space.

2.3. Maximum Margin Criterion (MMC)

MMC [14] maximizes the average margin between various classes in the reduced domain. Therefore, the criterion for feature reduction is defined as follows.

$$J = \frac{1}{2}\sum_{i=1}^{C}\sum_{j=1}^{C} p_i p_j d(C_i, C_j), \quad (4)$$

where p_i and p_j are the prior probabilities of classes i and j; C is the number of source and target classes; and d is the interclass margin and is defined as:

$$d(C_i, C_j) = d(m_i, m_j) - S(C_i) - S(C_j)S(C_i), \quad (5)$$

where $S(C_j) = tr(S_j)$, and m_i and m_j are the mean vectors of the classes C_i and C_j, respectively. Also, S_i and S_j are the covariance matrices of the classes C_i and C_j. Following [14], $d(C_i, C_j) = tr(S_i)$, and Eq. (4) could be simplified to the following formula:

$$J = tr(S_b - S_w), \quad (6)$$

where S_b is the between-class scatter matrix, and S_w is the within-class scatter matrix, which are defined as follows:

$$S_b = \sum_{i=1}^{C} n_i(m_i - m)(m_i - m)^T,$$

$$S_w = \sum_{i=1}^{C} (X_i - m_i)(X_i - m_i)^T, \quad (7)$$

where m is the mean vector of all data and n_i indicates the number samples in class C_i. Then, MMC is achieved as the following optimization problem:

$$\arg\max_{W} F(W) = \arg\max_{W} tr(W^T(S_b - S_w)W). \quad (8)$$

$F \in \mathbb{R}^{n\times c}$ is defined as the indicator matrix, where c is the number of predefined clusters and F is calculated as follows:

$$F_{ij} = \begin{cases} \frac{1}{\sqrt{l_j}} & \text{if } x_i \text{ belong to } j\text{th cluster} \\ 0 & \text{otherwise} \end{cases}$$

where l_j indicates the number of samples in the jth cluster. Following [31], S_b and S_w are achieved as follows:

$$S_w = X(I - FF^T)X^T, \qquad S_b = XFF^TX^T. \quad (9)$$

In a nutshell, MMC maximizes the average margin between domain classes and discriminates various classes from each other in the reduced domain.

3. The proposed approach

In this section, Transfer Maximum Margin Criterion (T-MMC) approach for effectively tackling the problem of domain shift is presented in detail.

3.1. Motivation

Unlike most existing methods that project input data into latent space, our methodology to address the problem is inspired from joint domain reduction and maximum margin discrimination (that minimizes the distribution difference between domains and discriminates various classes according to invariant feature representation). T-MMC tries to find domain invariant features that reduce source and target domains difference and also construct maximum margin across classes. Figure 1 represents the main idea of our proposed method. In short, in search of the new representation, we:

1. Obtain an optimization problem that assigns a weight to each feature of the domain;
2. Reduce the number of domain features according to the assigned weights; and
3. Classify target data using a trained standard machine learning classification method on source data.

3.2. Transfer Maximum Margin Criterion (T-MMC)

In this paper, we propose a joint transfer learning and discrimination methodology that finds a low-dimensional reduced representation so that it concurrently minimizes domain difference between the source and target data and discriminates various classes based on maximum margin criterion. Since some of the source and target data features have distribution difference with one another (variant features), we aim to find invariant common features which have unique distribution across domains. In the rest, we will discuss kernel-based feature weighting and maximum discrimination in detail.

Feature weighting. Given source data as $X_s \in \mathbb{R}^{n\times d}$, and target data as $X_t \in \mathbb{R}^{m\times d}$, where n is the number of source instances, m is the number of target instances and d is the number of features. We aim to predict unknown target label, $\{y_{t_i}\}$. Let $W \in \mathbb{R}^{d\times d}$ be a diagonal weight matrix that assigns a weight to each feature. The major issue is to minimize the distribution difference between domains by assigning optimal weight to each feature. Therefore, optimal weight, W^*, is obtained based on the domain distance, D, minimization between the source and target domains.

$$W^* = \arg\min_{W} D^2(X_s, X_t),$$
$$\text{s.t.} \quad \|\text{diag}(W)\| = 1 \quad \text{and} \quad W > 0, \quad (10)$$

where diag(W) is the diagonal of the weight matrix. The constraints control the range of W; the first

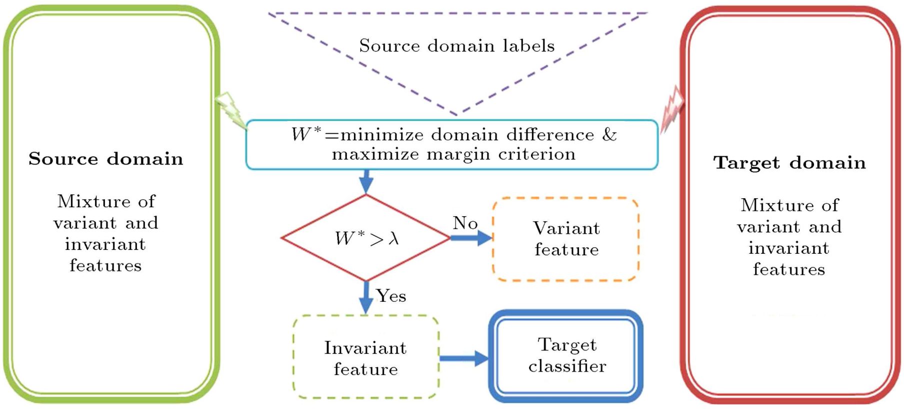

Source domain labels
Source domain
Mixture of variant and invariant features
W^*=minimize domain difference & maximize margin criterion
$W^* > \lambda$
No
Yes
Variant feature
Invariant feature
Target classifier
Target domain
Mixture of variant and invariant features

Figure 1. The flowchart of T-MMC. Each feature obtains a weight based on its discrepancy and aligned margin.

constraint restricts the size of weights and the second one lets W have only positive values. According to Eq. (3), the objective function can be written as a Gaussian kernel function due to MMD expression in terms of a kernel function:

$$\begin{aligned} D^2(X_s, X_t) =& \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} e^{-\frac{(x_s^i - x_s^j)^T (x_s^i - x_s^j)}{\delta}} \\ &+ \frac{1}{m^2}\sum_{i=1}^{m}\sum_{j=1}^{m} e^{-\frac{(x_t^i - x_t^j)^T (x_t^i - x_t^j)}{\delta}} \\ &- \frac{2}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m} e^{-\frac{(x_s^i - x_t^j)^T (x_s^i - x_t^j)}{\delta}}. \end{aligned} \tag{11}$$

Following Pan et al., 2011, the above equation can be rewritten in the form below using the kernel trick, i.e. $k(z_i, z_j^T) = \Phi(z_i)\Phi(z_j^T)$ where k is a positive definite kernel:

$$D^2(X_s, X_t) = tr(KL), \tag{12}$$

where:

$$K = \begin{bmatrix} K_{s,s} & K_{s,t} \\ K_{t,s} & K_{t,t} \end{bmatrix}, \tag{13}$$

and $K \in \mathbb{R}^{(n+m)\times(n+m)}$ is a composite kernel matrix. $K_{s,s}$, $K_{t,t}$ and $K_{s,t}$ are kernel matrices that have been defined by k on the source, target, and cross domains, respectively. $L \in \mathbb{R}^{(n+m)\times(n+m)}$ is the coefficient matrix with $L_{s,s} = \frac{1}{n^2}$, $L_{t,t} = \frac{1}{m^2}$ and $L_{s,t} = \frac{-1}{nm}$. Each element in K is computed using the kernel function; thus, they depend on W. The polynomial kernel $K_{s,s}$ with the degree p is calculated by $K_{s,s} = (xWx + 1)^p$.

Maximum discrimination. In this paper, an efficient and stable method is proposed to find the most discriminant features based on maximum margin criterion. Our proposed approach shows better separability in reduced domain. Moreover, the selected features based on MMC maximize the between-class scatter in the input space instead of null space of S_w.

Here, we want to predict the label of target data. For this purpose, some measures need to be exploited to evaluate the similarity or dissimilarity of samples. We need to preserve similarity or dissimilarity information in the reduced domain as much as possible. We hope that a sample would be close to those in the same class but far from those in other classes. Therefore, we need to maximize the distance between classes after reduction. In this way, we define our optimization criterion as:

$$J = tr(S_b - S_w). \tag{14}$$

A large J means that patterns are close to each other if they are from the same class and small J indicates that patterns are far from each other if they are from different classes. Therefore, our method preserves more discriminative information in the reduced domain.

When performing transfer learning, we want to find a feature space such that J is maximized after reduction. In this section, we discuss how to find an optimal weight matrix that transfers knowledge from source to target domain and simultaneously discriminates various classes from each other. By merging Eqs. (12) and (14), we achieve an optimization problem that bridges across source and target domains and has maximum separability.

$$\begin{aligned} &W^* = \arg\min_W tr(KL) + \beta \arg\max_W (W^T(S_b - S_w)W), \\ &\text{s.t.} \quad \|\mathrm{diag}(W)\| = 1 \quad \text{and} \quad W > 0, \end{aligned} \tag{15}$$

where β is the regularization parameter to guarantee the optimization problem to be well-defined. Eq. (15) is converted to the minimization optimization problem as follows:

$$\begin{aligned} &W^* = \arg\min_W tr(KL) - \beta \arg\min_W (W^T(S_w - S_b)W), \\ &\text{s.t.} \quad \|\mathrm{diag}(W)\| = 1 \quad \text{and} \quad W > 0. \end{aligned} \tag{16}$$

Thus, W is achieved in a way that the samples from the same class have lower distance from class mean. This can be obtained through minimizing the distance between the samples and the mean of each class. In this way, every instance falls into a predefined compact cluster, hence, considerably increasing the classification performance.

Since Eq. (16) is a Quadratically Constrained Quadratic Program (QCQP), it should be solved using a QCQP solver such as CVX (abbreviation for ConVeX). CVX is a strong tool for convex function optimization. Algorithm 1 shows T-MMC, where W^* contains the optimized weights. Because the weight values in matrix W are very small, W is normalized before feature discrimination. The weight of each feature classifies it as either variant or invariant.

4. Experimental setup

In this section, we present the setup of our experiments on various datasets for our proposed approach.

4.1. Data description

We evaluate T-MMC on two types of real world datasets that are benchmark in domain shift problem and four synthetic datasets. Table 1 presents a short view of synthetic datasets.

```
**Input:** Source data $X_s \in \mathbb{R}^{n\times d}$; target data $X_t \in \mathbb{R}^{m\times d}$;
        number of classes $C$; regularization parameter $\beta$;
        polynomial kernel degree $p$
**Output:** Optimal weight matrix $W^*$
*cvx_begin*
*variable* $W(d,d)$ *diagonal*
$K_{ij} = \begin{cases} (1+x_sWx_s)^p & \text{if} \quad x_i, x_j \in X_s \\ (1+x_tWx_t)^p & \text{if} \quad x_i, x_j \in X_t \\ (1+x_sWx_t)^p & \text{if} \quad x_i \in X_s, x_j \in X_t \\ (1+x_tWx_s)^p & \text{if} \quad x_i \in X_t, x_j \in X_s \end{cases}$
$L_{ij} = \begin{cases} \frac{1}{nm} & \text{if} \quad x_i, x_j \in X_s \\ \frac{1}{mm} & \text{if} \quad x_i, x_j \in X_t \\ \frac{-1}{nm} & \text{otherwise} \end{cases}$
$S_b = X_s(I - FF^T)X_s^T;$
$S_w = X_sFF^TX_s^T;$
$W^* = \min(\text{trace}(KL) - \beta * \text{trace}(W'(S_w - S_b)W));$
*subject to*
$W > 0$
$\|\text{diag}(W)\| = 1$
*cvx_end*
```

Algorithm 1. The optimization problem of T-MMC.

Table 1. The list and properties of the synthetic datasets. N is the number of invariant features and V denotes the number of variant features.

Dataset	N	V	Number of instances	Distribution of source domain	Distributions of target domain
GAU-GAU	10	40	300	Gaussian	Gaussian-Gaussian
GAU-EXP	20	30	300	Gaussian	Gaussian-exponential
WEI-GEO	30	20	600	Weibull	Weibull-geometric
POI-UNI	40	10	600	Poison	Poison-uniform

4.1.1. Synthetic datasets

Synthetic datasets are randomly generated to evaluate the performance of T-MMC in different difficulty situations. Each dataset is composed of the source and target domains, which contains N invariant and V variant features. According to Table 1, GAU-GAU is a shifted dataset, which has been composed of 50 features. For both source and target domains, N invariant features are sampled from N randomly picked distributions with zero mean and unit variance. For the source domain, V variant features are sampled from V randomly picked distributions with zero mean and unit variance. For the target domain, V variant features are sampled from V randomly picked distributions with shifted mean and unit variance.

GAU-EXP dataset is generated in the same way, except that V variant features generated in the target domain are sampled from randomly picked *exponential* distributions. Datasets are sampled from six standard distributions: *Gaussian, exponential, Weibull, geometric, uniform*, and *poison.*

The labels for synthetic datasets are generated using standard *sign* function. At first, number of r features are randomly selected from the total number of features, d. Next, vector $g \in \mathbb{R}^{d\times 1}$ is drawn from uniform distribution. Then, the elements of vector g are set to zero if they are not included in the feature set r. Finally, the label set l is generated via $l = \text{sign}(g*x)$, where x is the input data.

4.1.2. Real datasets

Indoor WiFi localization is a benchmark dataset in domain shift problem. For the first time, WiFi dataset was published in 2007 IEEE ICDM Contest for transfer learning. WiFi contains the labels of 247 locations and Received Signal Strength (RSS) by different access points. Since the value of RSS is a function of time, device, and space, the collected data in time period A (source domain) and B (target domain) has significant distribution difference. WiFi dataset contains 621 labeled source samples and 3,128 unlabeled target samples. In our experiments, we choose 1020 WiFi samples and compose WiFi1 and WiFi2 datasets with 510 items as source and target domains. Next, we switch source and target data to form another experiment.

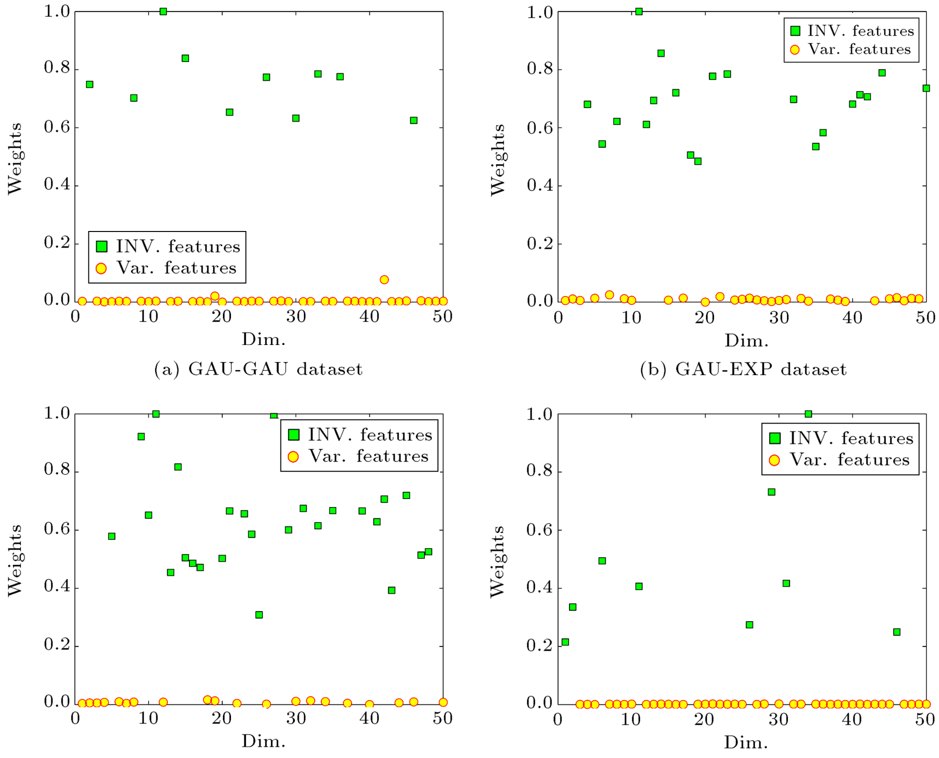
Digit dataset contains USPS and MNIST handwritten digits. USPS dataset refers to the handwritten digits scanned from envelops of U.S. Postal Service. The numbers of training and test samples are 7291 and 2007, respectively, and the objects have been normalized in 16 × 16 gray-scale images. MNIST is a large dataset of handwritten digits that was taken from mixed American Census Bureau employees and American high school students. The images were normalized to fit into 20×20 pixel bounding box with gray-scale level. MNIST has a training set of 60,000 samples and a test set with 10,000 samples. USPS and MNIST share 10 classes of digits, but they have been drawn from very different distributions. We uniformly rescale all images in USPS and MNIST to size 16×16, and fit them into gray-scale pixel feature vectors. Thus, the feature space of the source and target data is unified. Also, we randomly select 500 images from USPS as source data and 500 images from MNIST as target data. Next, we switch source and target data to form another dataset.

4.2. Method evaluation

We compare T-MMC with two other state-of-the-art transfer learning approaches (TCA [23], f-MMD [32]). Since ROWA and other domain adaptation methods are dimensionality reduction approaches, Nearest Neighbor (NN) is used for classification and regression on the labeled source data and unlabeled target data. We compared our approach with f-MMD and TCA keeping the dimension size equal for all three methods. All methods are evaluated by their reported best results. The parameters of TCA and f-MMD are adjusted to 1 and 0.1, respectively, and they are fixed during the tests.

5. Empirical results and discussion

In this section, the performance of T-MMC on synthetic and real world benchmark datasets is evaluated. Figure 2 shows the weight of each feature assigned by T-MMC. The horizontal axis shows the domain dimensions and the vertical axis indicates the weight of each feature. As is clear from the plots, variant and invariant features have been separated from each other and T-MMC has assigned higher weights to invariant features. Therefore, the feature spaces of the variant and invariant features are distinguishable using a margin which has various widths according to the type of domain distributions. Moreover, we are able



(a) GAU-GAU dataset

(b) GAU-EXP dataset

(c) WEI-GEO dataset

(d) POI-UNI dataset

Figure 2. Weight assignment of the proposed approach. T-MMC assigns higher weights to invariant features and weights close to zero to variant features (best viewed in color).

to determine a threshold weight that could be general across different datasets. In the next section, we will discuss how to determine the threshold value.

5.1. Synthetic datasets

In this section, we discuss the performance of T-MMC on synthetic datasets. Since T-MMC discovers and removes the variant features from the feature space, it only contains the invariant features of the domain. Moreover, T-MMC preserves the original properties of the input data and incorporates all samples for target domain label prediction.

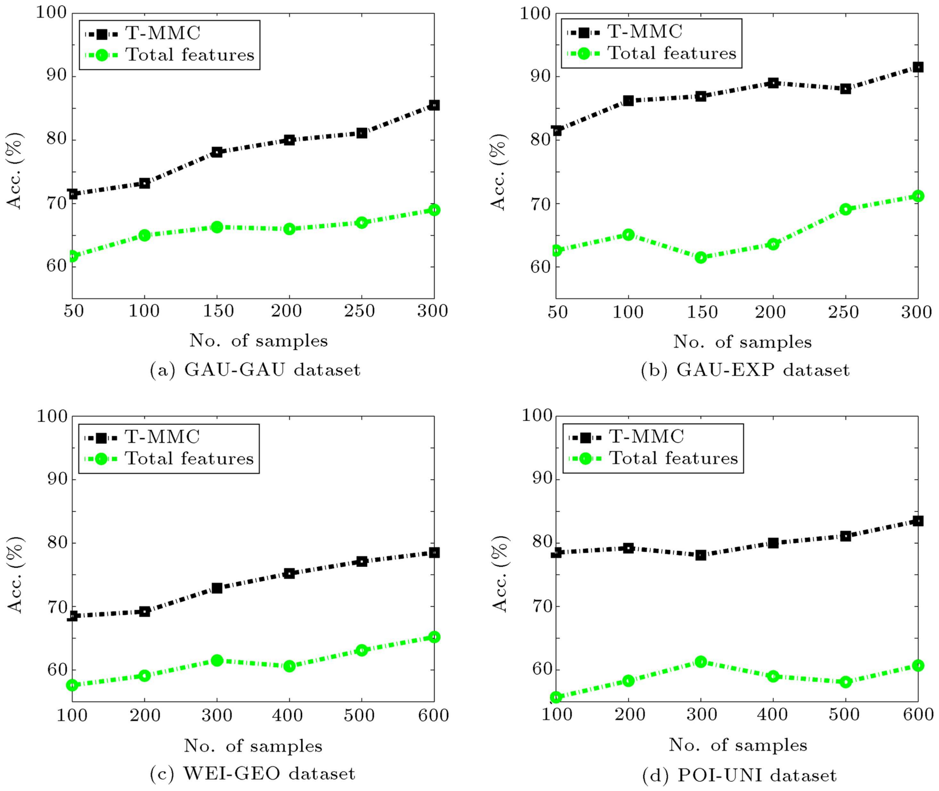
Figure 3 shows the performances of T-MMC against NN on four synthetic datasets. In fact, we examine the performance of T-MMC with only invariant features with a case that all features are available in learning process. The horizontal axis denotes the number of samples that varies from 50 to 300 for *GAU-GAU* and *GAU-EXP* datasets, and it varies from 100 to 600 for WEI-GEO and POI-UNI datasets. The vertical axis is the classification accuracy of T-MMC and NN.

Figure 3(a) shows the accuracy of T-MMC against NN on GAU-GAU dataset. According to the reported results, T-MMC shows better performance in all cases. Moreover, with increasing the number of samples, T-MMC could generate accurate margins between various classes. In this way, according to Eq. (7), the within-class scatter is determined by employing a large number of samples and the width of margin increases based on class differences.

The amounts of improvement in Figure 3(a), (b), (c), and (d) are different. In fact, some factors determine the performance of classification algorithms, including number of features, number of samples, distribution of domain, and so on. It is worth noting that increasing the number of samples in most cases enhances the accuracy of T-MMC in reduced domain.

As the second synthetic experiment, we design a dataset with different numbers of instances and 30 features, where the numbers of variant and invariant features are equal. Also, half of the features are randomly selected to generate the class label using the procedure previously mentioned. The distributions of variant and invariant features are Gaussian with different means and the same variance.

Figure 4 shows the performance of T-MMC where the number of samples raises from 100 to 350. T-MMC outperforms other feature-based transfer learning approaches, i.e. TCA and f-MMD, in all cases. With increasing the number of samples, the accuracy of f-MMD and T-MMC increases while TCA shows instability in results in some situations. TCA and f-MMD are fully unsupervised and could not exploit the labels of samples to find the sufficient results; thus, their performances unexpectedly degrade.



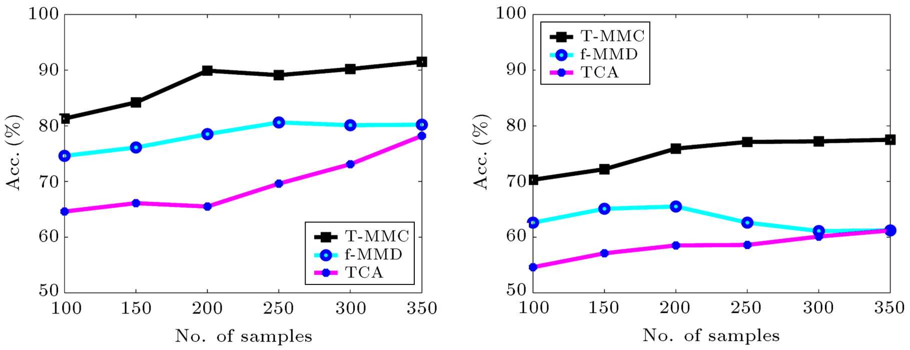
(a) GAU-GAU dataset

(b) GAU-EXP dataset

(c) WEI-GEO dataset

(d) POI-UNI dataset

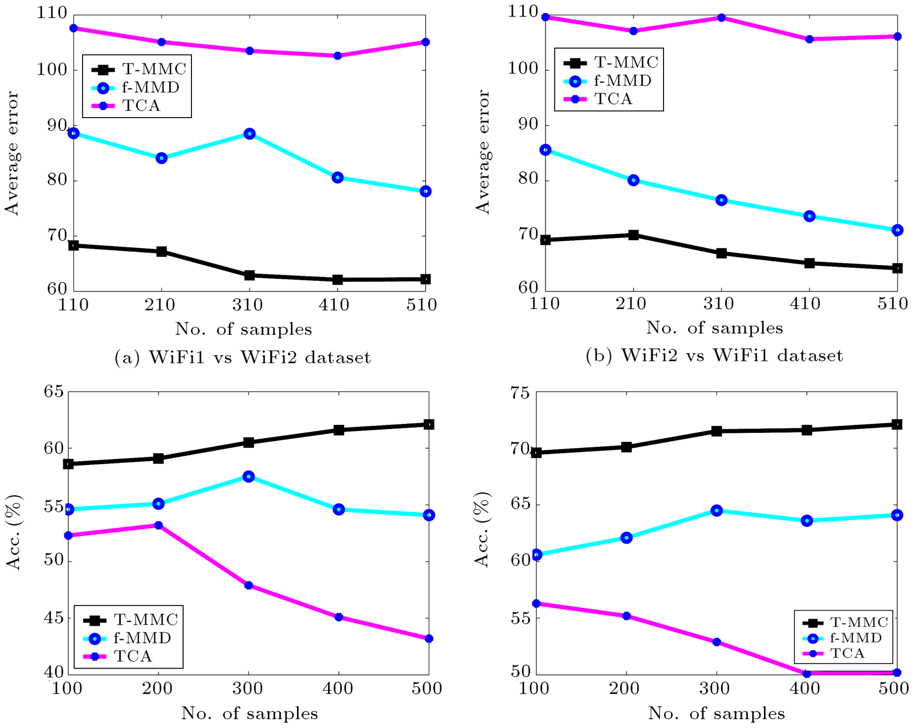
Figure 3. Performance evaluation. T-MMC is compared with NN with total features. Employing only invariant features by T-MMC enhances its performance.



(a) $\mu = 1$

(b) $\mu = 5$

Figure 4. Performance evaluation. Comparison between the performance of T-MMC and that of other transfer learning methods on synthetic datasets. μ indicates the mean parameter of Gaussian distribution.



(a) WiFi1 vs WiFi2 dataset

(b) WiFi2 vs WiFi1 dataset

(c) USPS vs MNIST dataset

(d) MNIST vs USPS dataset

Figure 5. Performance evaluation. Comparison between T-MMC and other transfer learning methods on real datasets.

The performance of T-MMC has almost a growing trend; but in some cases, its accuracy degrades with respect to the number of samples. For example, in Figure 4(a), when the number of instances increases from 200 to 250, accuracy decreases to 89.1; because in some cases, class labels have strong relations to variant features and removing them decreases the performance of the model.

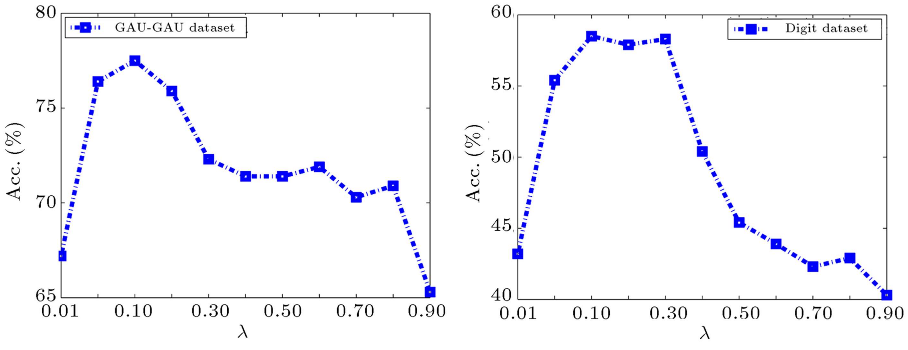
5.2. Real datasets

The real world transfer learning datasets naturally have domain shift, e.g. in indoor WiFi localization dataset the WiFi signal strength may be a function of device, space, time, or other dynamic factors. Thereupon, we need to adapt the shifted data of source and target domains. Figure 5 shows the performance of T-MMC compared to those of other feature-based transfer learning methods.

Figure 5(a) illustrates the experiments on the indoor WiFi localization dataset. The task is to identify the labels of the WiFi data collected during time period B according to the data collected during time period A. Each experiment is repeated 10 times and the Average Error Distance (AED) is calculated according to the following relation:

$$\text{AED} = \frac{\sum_{(x_i, y_i) \in D} |f(x_i) - y_i|}{N},$$

where x_i, $f(x_i)$, and y_i are vectors of RSS values,



(a) GAU-GAU dataset

(b) Digit dataset

Figure 6. Impact of parameter setting. Parameter evaluation with respect to the classification accuracy based on threshold parameter λ.

predicted location, and corresponding true location, respectively. The number of samples is increased from 110 to 510 and the performance of methods is evaluated based on the AED.

In Figure 5(a), WiFi1 is considered as the source domain and WiFi2 denotes the target domain. Figure 5(b) shows the contrary situation on which WiFi2 is considered as the source domain and WiFi1 denotes the target domain. As is clear from the plots, T-MMC generalizes more samples in the target domain. In fact, T-MMC has a decreasing error trend on datasets with more samples to predict target label.

TCA has the worst performance, because it projects the data into the latent space without considering the relation between the features and class labels. In fact, TCA only determines the transfer components based on the variance of data. However, f-MMD has better performance than TCA where it transfers the knowledge in the original space and does not project the domains into a latent space.

Figure 5(c) and (d) show the performance of T-MMC and other transfer learning approaches on digit dataset. We first set USPS as the source and MNIST as the target domain (Figure 5(c)), and next repeat our experiment by exchanging the source and target domains. T-MMC in all cases outperforms TCA and f-MMD as it distinguishes the target digits based on source samples with high accuracy. In fact, T-MMC only preserves the features that are common across domains and also could discriminate between classes via achieving a trustworthy margin. On the other hand, T-MMC could bridge across domains and it transfers knowledge from labeled source domain to unlabeled target domain.

Since TCA projects input data into a shared latent space in a fully unsupervised manner, it fails to classify target domain samples. This is reflected obviously in Figure 5(c) and (d). With increasing the number of samples, negative transfer happens and the performance of TCA dramatically degrades. f-MMD has reasonable performance, but it is also unsupervised and does not exploit source domain labels to transfer knowledge from source to target domain.

5.3. Impact of parameter settings

T-MMC is evaluated with respect to different values of parameter to analyze its performance in various conditions. In general, we should tune the threshold parameter, λ, for T-MMC on different datasets. Due to the page limitation, we only report the results of T-MMC on *GAU-GAU* and *Digit* datasets.

Figure 6 illustrates the experiments on *GAU-GAU* and *Digit datasets*. We run T-MMC with varying values of λ. We report the classification accuracy of T-MMC with $\lambda \in [0.010.9]$ on both datasets. The value of λ determines the margin between the variant and invariant features. The plot indicates that in most cases, increasing the value of λ decreases the performance of T-MMC while the accuracy has negative slope. Indeed, T-MMC shows better performance with low values of λ. In this way, $\lambda \in [0.05\ 0.25]$ is chosen as the acceptable interval. We select $\lambda = 0.1$ for our experiments.

6. Conclusion and future work

In this paper, we presented a Transfer Maximum Margin Criterion (T-MMC) approach for cross domain classification. T-MMC exploits transfer learning strategies to cope with domain shift problem. Moreover, T-MMC employs domain invariant clustering to enhance the adaptation performance in the reduced subspace. The reduced subspaces for source and target domains are most effective and robust for cross domain problems. Performance of T-MMC is evaluated from different perspectives and its yields are compared with other state-of-the-art baseline methods. Our comprehensive experiments on a variety of synthetic and real datasets with different difficulties show that T-MMC significantly outperforms other adaptation methods. For future work, we plan to advance in this direction, i.e. proposing T-MMC for multi domain setting.

References

1. Chen, Y., Wang, G. and Dong, S. "Learning with progressive transductive support vector machine", *Pattern Recogn. Lett.*, **24**(12), pp. 1845-1855 (2003).
2. 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).
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. Joachims, T. "Transductive inference for text classification using support vector machines", *16th Int. Conf. on Mach. Learn.*, Bled, Slovenia, pp. 200-209 (1999).
5. Helmi, H., Garibaldi, J.M. and Aickelin, U. "Examining the classification accuracy of TSVMs with feature selection in comparison with the GLAD algorithm", arXiv preprint arXiv:1307.1387 (2013).
6. 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).
7. Zhao, P., Hoi, S.C., Wang, J. and Li, B. "Online transfer learning", *Artif. Intell.*, **216**, pp. 76-102 (2014).
8. Gheisari, M. and Baghshah, M.S. "Unsupervised domain adaptation via representation learning and adaptive classifier learning", *Neurocomputing*, **165**, pp. 300-311 (2015).
9. 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).
10. Gong, B., Grauman, K. and Sha, F. "Learning kernels for unsupervised domain adaptation with applications to visual object recognition", *Int. J. Comput. Vision*, **109**(1-2), pp. 3-27 (2014).
11. Long, M., Wang, J., Ding, G., Pan, S.J. and Yu, P.S. "Adaptation regularization: A general framework for transfer learning", *IEEE T Knowl Data En*, **26**(5), pp. 1076-1089 (2014).
12. Pan, S.J. and Yang, Q. "A survey on transfer learning", *IEEE T Knowl. Data En.*, **22**(10), pp. 1345-1359 (2010).
13. Borgwardt, K.M., Gretton, A., Rasch, M.J., Kriegel, H.P., Schölkopf, B. and Smola, A.J. "Integrating structured biological data by kernel maximum mean discrepancy", *Bioinformatics*, **22**(14), pp. e49-e57 (2006).
14. Li, H., Jiang, T. and Zhang, K. "Efficient and robust feature extraction by maximum margin criterion", *IEEE T Neural Networ*, **17**(1), pp. 157-165 (2006).
15. Dai, W., Yang, Q., Xue, G.R. and Yu, Y. "Boosting for transfer learning", *24th Int. Conf. on Mach. Learn.*, Oregon, USA, pp. 193-200 (2007).
16. Deng, Z., Choi, K.S., Jiang, Y. and Wang, S. "Generalized hidden-mapping ridge regression, knowledge-leveraged inductive transfer learning for neural networks, fuzzy systems and kernel methods", *IEEE T Cybern*, **44**(12), pp. 2585-2599 (2014).
17. Sugiyama, M., Nakajima, S., Kashima, H., Buenau, P.V. and Kawanabe, M. "Direct importance estimation with model selection and its application to covariate shift adaptation", *Adv. Neur. In.*, Vancouver, Canada, pp. 1433-1440 (2008).
18. Eaton, E. and DesJardins, M. "Set-based boosting for instance-level transfer", *IEEE Int. Conf. on Data Mining Workshops*, Florida, USA, pp. 422-428 (2009).
19. Lawrence, N.D. and Platt, J.C. "Learning to learn with the informative vector machine", *Proceedings of the Twenty-First International Conference on Machine Learning, ACM,* (2004).
20. Tong, B., Gao, J., Nguyen, T.H. and Suzuki, E. "Gaussian process for dimensionality reduction in transfer learning", *Int. Conf. on Data Mining*, Arizona, USA, pp. 783-794 (2011).
21. Gao, X., Wang, X., Li, X. and Tao, D. "Transfer latent variable model based on divergence analysis", *Pattern Recogn*, **44**(10), pp. 2358-2366 (2011).
22. Davis, J. and Domingos, P. "Deep transfer via second-order Markov logic", *26th Int. Conf. on Mach. Learn.*, Montreal, Canada, pp. 217-224 (2009).
23. 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).
24. Dinh, C.V., Duin, R.P., Piqueras-Salazar, I. and Loog, M. "FIDOS: A generalized Fisher based feature extraction method for domain shift", *Pattern Recogn.*, **46**(9), pp. 2510-2518 (2013).
25. Fernando, B., Habrard, A., Sebban, M. and Tuytelaars, T. "Unsupervised visual domain adaptation using subspace alignment", *IEEE Int. Conf. on Comp. Vis.*, Sydney, Australia, pp. 2960-2967 (2013).
26. Blitzer, J., Dredze, M. and Pereira, F. "Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification", *ACL*, **7**, pp. 440-447 (2007).
27. 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).
28. Pan, S.J., Kwok, J.T. and Yang, Q. "Transfer learning via dimensionality reduction", *23d Conf. on Artif. Intel.*, Chicago, USA, pp. 677-682 (2008).
29. Pan, W. "A survey of transfer learning for collaborative recommendation with auxiliary data", *Neurocomputing*, **177**, pp. 447-453 (2016).
30. Gretton, A., Borgwardt, K.M., Rasch, M., Schlkopf, B. and Smola, A.J. "A kernel method for the two-sample-problem", *20th Conf. of Adv. Neur. In.*, pp. 513-520 (2006).

31. Yang, S., Lin, M., Hou, C., Zhang, C. and Wu, Y. "A general framework for transfer sparse subspace learning", *Neural Comput. Appl.*, **21**(7), pp. 1801-1817 (2012).

32. Uguroglu, S. and Carbonell, J. "Feature selection for transfer learning", *Lect. Notes Artif. Int.*, pp. 430-442 (2011).

Biographies

Jafar Tahmoresnezhad received the PhD degree in Computer Engineering from Shiraz University, Shiraz, Iran, in 2015. Following academic appointments at Urmia University of Technology, he is currently an Assistant Professor in the 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 the PhD degree in Computer Science from Iran University of Science and Technology in conjunction with Monash University, Australia, in 2008. Following academic appointments at Shiraz University, he is currently an Associate Professor in Electrical and Computer Engineering School, Shiraz University, Shiraz, Iran. He is recognized for 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.