Multi-Similarity Based Multi-Source Transfer Learning and Its Applications

Zhen Liu, Jun-an Yang, Hui Liu, and Wei Wang
Electronic Engineering Institute, Hefei 230037, China
Key Laboratory of Electronic Restriction of Anhui Province, Hefei 230037, China
Email: [ahuiz, liuhui1983eei]@163.com; yangjunan@ustc.edu.cn; wwei009@mail.ustc.edu.cn

Abstract — In this paper, a novel multi-source transfer learning method based on multi-similarity ((MS)^2TL) is proposed. First, we measure the similarities between domains at two levels, i.e., “domain-domain” and “sample-domain”. With the multi-similarities, (MS)^2TL can explore more accurate relationship between the source domains and the target domain. Then, the knowledge of the source domains is transferred to the target based on the smoothness assumption, which enforces the requirement that the target classifier shares similar decision values with the relevant source classifiers on the unlabeled target samples. (MS)^2TL can increase the chance of finding the sources closely related to the target to reduce the “negative transfer” and also imports more knowledge from multiple sources for the target learning. Furthermore, (MS)^2TL only needs the pre-learned source classifiers when training the target classifier, which is suitable for large datasets. We also employ a sparsity-regularizer based on the ε-insensitive loss to enforce the sparsity of the target classifier with the support vectors only from the target domain such that the label prediction on any test sample is very fast. We also use the ε-insensitive loss function to enforce the sparsity of the decision function for fast label prediction. Validation of (MS)^2TL is performed with toy and real-life datasets. Experimental results demonstrate that (MS)^2TL can more effectively and stably enhance the learning performance. Finally, (MS)^2TL is also applied to the communication specific emitter identification task and the result is also satisfying.

Index Terms—Transfer learning, multiple source transfer, manifold assumption

I. INTRODUCTION

Transfer learning [1]-[2] can effectively exploit and transfer the knowledge from different but similar source domains for target domain learning. Recently, transfer learning has been applied to many real-world applications, such as text processing [3], computer vision [4]-[5], network identification [6], automatic control [7], etc.

For the single-source domain setting, much work has been developed [1]. In general, the effectiveness of the knowledge transfer from a source domain to the target domain depends on how well they are related. The stronger the relationship, the more usable will be the source knowledge. Often in practice, one may be offered more than one source domain for learning. If we only use one source domain for learning, it is wasteful and we also can’t ensure that the selected source domain is well related with the target domain. Brute force transferring in case of weak relationships may lead to performance deterioration of the target domain learning, i.e., “negative transfer”. In this paper, we propose a novel multi-source transfer learning method called (MS)^2TL (Multi-Similarity based Multi-Source Transfer Learning). (MS)^2TL explores the relationships between the source domains and the target domain by multi-similarity metric. Then, the knowledge of the source domains is transferred to the target based on the smoothness assumption, which enforces that the target classifier shares similar decision values with the relevant source classifiers on the unlabeled target samples.

We summarize the main contributions of this paper as follows: We propose a novel multi-source transfer learning method called (MS)^2TL, which can not only improve the ability to avoid the problem of “negative transfer” but also explore more knowledge from the source domains for the target domain learning. In (MS)^2TL, we measure the similarities between domains at two levels, i.e., “domain-domain” and “sample-domain”. With the multi-similarities, we then define a multi-source transfer manifold regularizer and add it into the optimal function of (MS)^2TL for knowledge transfer. We also use the ε-insensitive loss function to enforce the sparsity of the decision function for fast label prediction. Furthermore, (MS)^2TL only needs the pre-learned source classifiers when training the target classifier, which is suitable for large datasets. (MS)^2TL can be readily introduced to many kernel methods and extend these methods to the corresponding transfer learning methods [8]. In this paper, we give our method under the framework of least square SVM (LS-SVM) [9]. We evaluate our method in two multiple transfer learning related applications, i.e., target recognition and document retrieval. Experimental results demonstrate that (MS)^2TL can more effectively and stably enhance the learning performance. Finally, the proposed algorithm is applied to the communication specific emitter identification task and the result is also satisfying.

The rest of the paper is organized as follows: In Section II, we briefly review the related work; In Section III, the proposed method (MS)^2TL is introduced; In
Section IV, extensive experiments are performed; Some conclusions are given in Section V.

II. RELATED WORKS

Chattopadhyay et al. [10] proposed a weighting scheme which gives higher weights to those source domains with similar conditional probability distributions to the target. Based on [10], Sun et al. [11] proposed a two-stage transfer methodology in which the source samples are first weighted based on the marginal probability differences and then re-weighted by the weighting scheme in [10].

Ref. [10] and [11] use the source domain samples to train a target classifier whenever a new task is conducted. It is not efficient when the size of dataset is large. A more efficient way is to train a classifier in each source domain and is computed as follows

\[
\Omega(f^T) + \lambda_1 \Omega_s(f^T) + \lambda_2 \Omega_D(f^T)
\]

(1)
where \( \Omega(f^T) \) is a regularizer to control the complexity of the target domain classifier \( f^T \), \( \Omega_s(f^T) \) is a loss function of \( f^T \) on the labeled samples of target domain, \( \Omega_D(f^T) \) is the data-dependent regularizer defined on the unlabeled samples of target domain, and \( \lambda_1, \lambda_2 > 0 \) are the regularization parameters.

In DAM, the key of knowledge transfer is the data-dependent regularizer \( \Omega_D(f^T) \).

\[
\Omega_D(f^T) = \frac{1}{2} \sum_{i=1}^{N_T} \sum_{l=1}^{N_s} (f^T_i - f^s_i)^2
\]

(2)
where \( \gamma_s \) is the similarity weight of the \( s \)-th source \( D^s \) and is computed as follows

\[
\gamma_s = \exp(-\frac{\text{MMD}^2(D^s, D^T)}{\beta_i})
\]

(3)
where \( \text{MMD}^2(D^s, D^T) = \frac{1}{N_s} \sum_{i=1}^{N_s} \phi(x^s_i) - \frac{1}{N_T} \sum_{j=1}^{N_T} \phi(x^T_j) \) is the maximum mean discrepancy (MMD) [15] for measuring the data distributions between the \( s \)-th source domain and the target domain. MMD is an effective nonparametric distance metric for comparing data distributions in the reproducing kernel Hilbert space. \( \beta_i > 0 \) is the bandwidth parameter to control the spread of MMD and is usually fixed as the mean of MMD among domains.

DAM uses \( \gamma_s \) to measure the relationships between the source domains and the target domain. However, the \( \gamma_s \) only consider the relationships between domains as a whole. The similarity measurement is not enough detailed and accurate. As we know, it is important to find and measure the relationships between the source domains and the target domain for transfer learning. To explore the relationships between domains better, we give a multi-similarity measurement at two levels, i.e., “domain-domain” and “sample-domain”. With the defined multi-similarity weights, we modify the regularizer in (2) and then give our method (MS)TL under the framework of DAM.

A. Multi-Similarity

In Fig. 1, the triangles and the circles represent one class respectively. Fig. 1 shows that the classification model learned in the biased source domain is not reliable.
in the target domain. The key of transfer learning is to find and measure the relationships between the source domains and the target domain. Here we measure the similarities between domains at two levels, i.e., “domain-domain” and “sample-domain”.

Firstly, we concern the overall similarities between the source domains and the target domain, i.e., the similarity at the level of “domain-domain”. Here, we use the similarity weight \( \gamma_s \) in (3) as the measurement.

Since the samples in the target domain are different, their relevancies to a source domain are also different. To describe the relationship between the target domain and the source domains further in detail, we concern the similarities at the level of “sample-domain”. Here, two kinds of distance are first given: the average distance in the neighborhood (i.e., \( DN_i^s \)) and the minimum distance to the class center (i.e., \( DC_i^s \)).

\[
DN_i^s = \frac{1}{N_k} \sum_{k=1}^{N_k} d(x_i^T, x_k^s) \tag{4}
\]

where \( x_k^s \) is the \( k \)th neighbor of \( x_i^T \) in \( D^s \), \( N_k \) is the number of neighbors, \( d(\cdot) \) is a general distance metric. If \( DN_i^s \) is small, \( x_i^T \) is more likely to occur in \( D^s \) and thus more similar to \( D^s \).

\[
DC_i^s = \min_j d(x_i^T, c_j^s) \tag{5}
\]

where \( c_j^s \) is the mean of \( j \)th class samples in \( D^s \). If \( DC_i^s \) is small and \( x_i^T \) is most close to the \( j \)th class center of \( D^s \), \( x_i^T \) probably belongs to the \( j \)th class in \( D^s \).

The distances defined in (4)-(5) measure the similarity from the marginal and conditional distribution of the data respectively. Combining them together, we have the following similarity weight \( A_{ij} \) of \( x_i^T \) in the \( s \)th source domain at the level of “sample-domain”

\[
A_{ij} = \exp(-d_i^s \frac{2}{\beta_s}) \tag{6}
\]

where \( d_i^s = 0.5 \times (DN_i^s + DC_i^s) \), \( \beta_s > 0 \) is the bandwidth parameter to control the spread of \( d_i^s \) and is usually fixed as the mean of \( d_i^s \) in the whole unlabeled target set.

With (3) and (6), we defined the similarities between the source domains and the target domain at the two levels of “domain-domain” and “sample-domain”. The multi-similarities can measure the distribution relevance in more detail, which is in favor of the transfer learning.

B. Multi-Source Transfer Manifold Regularizer

Belkin et al. [16] proposed the manifold assumption which enforces the decision function to be smooth on the data manifold, namely, the two samples in a high-density region should share similar decision values. Motivated from the manifold assumption, we similarly assume that the target domain classifier \( f^T \) should have similar decision values on the unlabeled target samples with the pre-learned classifiers from the relevant source domains. Based on the similarities defined in section III.A, the multi-source transfer manifold regularizer \( \Omega_m(f^T) \) is given as follows

\[
\Omega_m(f^T) = \frac{1}{2} \sum_{i=1}^{N} \gamma_i \sum_{j=1}^{N_s} A_{ij}(f^T - f_i^j)^2 \tag{7}
\]

where \( \gamma_i \) and \( A_{ij} \) are the similarity weights defined in (3) and (6) respectively. As shown in Fig. 2, the regularizer \( \Omega_m(f^T) \) builds the connections between the sources and the target through the similarity weights \( \gamma_i \) and \( A_{ij} \). If \( \gamma_i \) and \( A_{ij} \) are large, the decision value of \( f^T \) and \( f_i^j \) on \( x_i^T \) will be similar. Thus, we can transfer the knowledge from the sources to the target under this assumption of “domain relevance-decision constraint”.

C. The Solution

The minimizer of the optimization problem in (1) admits a form of \( f^T(x) = w^T \phi(x) + b \) and then the regularizer \( \Omega(f^T) = \|w\|^2 / 2 \). In addition, \( \Omega_m(f^T) \) is modeled as the square error of the target domain classifier \( f^T \) on the labeled target samples, which is analogous to the LS-SVM [9]. Under the framework of DAM in (1), the optimal function of (MS)^2TL is then formulated as follows
\[
\min_{f^T, \mathbf{w}, b, \xi^+, \xi^-} \frac{1}{2} \| \mathbf{w} \|^2 + C \sum_{i=1}^{N_s} \xi_i^+ + \sum_{i=1}^{N_s} \xi_i^- + \frac{1}{2} \sum_{i=1}^{N_s} (f_i^R - y_i^T)^2 + \frac{1}{2} \sum_{i=1}^{N_s} \gamma_i \sum_{i=N_s+1}^{N} A_i (f_i^R - f_i^*)^2
\]

where \( \xi_i^+ = 0 \) or \( \xi_i^- = 0 \), otherwise \( \xi_i^+ = 0 \). C is the regularization parameter. Since \( \xi_i^+ \) is non-smooth, (9) is usually transformed as a constrained optimization problem

\[
\min_{f^T, \mathbf{w}, b, \xi^+, \xi^-} \frac{1}{2} \| \mathbf{w} \|^2 + C \sum_{i=1}^{N_s} \xi_i^+ + \sum_{i=1}^{N_s} \xi_i^- + \frac{1}{2} \sum_{i=1}^{N_s} (f_i^R - y_i^T)^2 + \frac{1}{2} \sum_{i=1}^{N_s} \gamma_i \sum_{i=N_s+1}^{N} A_i (f_i^R - f_i^*)^2
\]

\[
s.t. \quad w^T \phi(x_i^T) + b - f_i^T \leq \xi_i^+, \xi_i^- \geq 0
\]

\[
f_i^T - w^T \phi(x_i^T) - b \leq \xi_i^+, \xi_i^- \geq 0
\]

By introducing the Lagrange multipliers \( \alpha_i^+ \)'s and \( \eta_i^- \)'s (resp., \( \alpha_i^+ \)'s and \( \eta_i^- \)'s) for the constraints in (11) (resp., (12)), we obtain the following Lagrangian

\[
L = \frac{1}{2} \| \mathbf{w} \|^2 + C \sum_{i=1}^{N_s} \xi_i^+ + \sum_{i=1}^{N_s} \xi_i^- + \frac{1}{2} \sum_{i=1}^{N_s} (f_i^R - y_i^T)^2 + \frac{1}{2} \sum_{i=1}^{N_s} \gamma_i \sum_{i=N_s+1}^{N} A_i (f_i^R - f_i^*)^2
\]

\[
- \sum_{i=1}^{N_s} \alpha_i (\xi_i^+ + \xi_i^- + f_i^T - w^T \phi(x_i^T) - b) - \sum_{i=1}^{N_s} \eta_i \xi_i^-
\]

\[
- \sum_{i=1}^{N_s} \alpha_i (\xi_i^+ + \xi_i^- + f_i^T - w^T \phi(x_i^T) - b) - \sum_{i=1}^{N_s} \eta_i \xi_i^-
\]

Setting the derivatives of (13) w.r.t. the primal variables (\( \mathbf{w}, b, \xi^+, \xi^-, f_i^T \)) to zeros, respectively, we have

\[
\frac{\partial L}{\partial \mathbf{w}} = 0 \Rightarrow \mathbf{w} = \sum_{i=1}^{N_s} \phi(x_i^T) (\alpha_i^+ - \alpha_i^-) - \mathbf{\Phi} (\mathbf{a} - \mathbf{a}^*)
\]

\[
\frac{\partial L}{\partial b} = 0 \Rightarrow I_{N_s} \alpha = I_{N_s} \alpha^*
\]

\[
\frac{\partial L}{\partial \xi^+} = 0 \Rightarrow C = \alpha_i + \eta_i
\]

\[
\frac{\partial L}{\partial \xi^-} = 0 \Rightarrow C = \alpha_i + \eta_i
\]

\[
\frac{\partial L}{\partial f_i^T} = 0 \Rightarrow \tilde{y}_i = \hat{y}_i + q_i (\alpha_i - \alpha_i^-)
\]

where \( I_{N_s} \) is a column vector of all ones with size \( N_s \), \( \mathbf{a} = [\alpha_1, \ldots, \alpha_{N_s}]^T \), \( \mathbf{a}^* = [\alpha_1^+, \ldots, \alpha_{N_s}^+]^T \), \( \mathbf{\Phi} = [\phi(x_1^T), \ldots, \phi(x_{N_s}^T)] \). If \( i = 1, \ldots, N_s \), \( q_i = 1/\lambda_i \) and \( \tilde{y}_i = y_i^T \). If \( i = N_s + 1, \ldots, N \), \( q_i = \frac{1}{\lambda_i} \sum_{i=1}^{M} y_i A_i \) and \( \tilde{y}_i = \frac{1}{M} \sum_{i=1}^{M} y_i A_i f_i^T \).

Substituting them back into (13), we arrive at the following dual formulation

\[
\min_{a^+, a^-} \frac{1}{2} (a - a^*)^T \mathbf{K} (a - a^*) + \psi^T (a - a^*) + \varepsilon \mathbf{1}_{N_s} (a + a^*)
\]

s.t. \( \mathbf{1}_{N_s} a = \mathbf{1}_{N_s} a^*, 0_{N_s} \leq a, a^* \leq C \mathbf{1}_{N_s} \).

where \( \mathbf{K} = \mathbf{K} + \text{diag}(\psi), \mathbf{K} = \mathbf{\Phi}^T \mathbf{\Phi}, \psi = [q_1, \ldots, q_{N_s}]^T \).

Since the dual form of (15) is similar to the dual of \( \varepsilon \)-SVR [18], the objective function of (MS)TL in (9) can be solved efficiently by using state-of-the-art SVM solvers such as LIBSVM [19]. For any test sample \( x \), the decision value of the target classifier \( f^T(x) \) is

\[
f^T(x) = \mathbf{w}^T \phi(x) + b
\]

which is a linear combination of \( k(x^T, x) \)'s without involving any base classifiers. According to the Karush–Kuhn–Tucker (KKT) conditions in (14), if the target sample \( x^T \) has the value \( \| \mathbf{w}^T \phi(x) + b - f^T \| < \varepsilon \), then its corresponding coefficient \( (a_i - a_i^-) \) in (16) becomes zero. Therefore, the computation for the prediction can be greatly reduced by using the sparse representation in (16).

IV. EXPERIMENTS AND DISCUSSIONS

If we only concern the similarity at the level of “domain-domain”, namely, set all the \( A_i \) equal to 1, (MS)TL would change to be similar to the DAM algorithm [14]. Considering their relations, we compare our method (MS)TL with DAM in the experiments. To show the improvement by the transfer learning progress, we also compare our (MS)TL with a non-transfer learning classification strategy represented as "Base" in the experiments. The Base means that the source domain classifiers \( f^s \)'s are used to predict the unlabeled target samples directly and the average accuracy is the final result.

To demonstrate that (MS)TL can use different type of classifier as source classifiers \( f^s \), we also conduct experiments with three types of source classifiers respectively, i.e., LS-SVM, C4.5, and Naïve Bayes.

In the experiments, we need to fix some parameters empirically, i.e., the number of neighbors (i.e., \( N_s \) ) in (4),
and the regularization parameters $C$, $\lambda_L$, and $\lambda_D$ in (9). In the default setting, we set $\lambda_L=\lambda_D=1$, $C=1$, and $N_t=8$. The parameters in the source classifiers are set as the default values in Weka [20]. Gaussian kernel (i.e., $k(x_i, x_j) = \exp(-\|x_i-x_j\|^2/(2\sigma^2))$) is used as the default kernel in which the kernel parameter $\sigma$ is set as the mean distance between samples in the target domain. In the target domain, $n$ samples per class are randomly selected as the labeled target set for the experiments. $n$ is tuned in the range $[0, 2, 4, 6, 10, 15, 20]$. The experiments are repeated for 20 times with different source samples and target samples. The average classification accuracy is used as the evaluation measure.

In order to fully evaluate the algorithm performance, we evaluate (MS)$^2$TL for two multi-source transfer learning related applications: 1) target recognition, and 2) document retrieval. Finally, (MS)$^2$TL is also applied to the communication specific emitter identification task.

A. Experiments on Target Recognition

The experiments on target recognition use three datasets as the sources [21]: Amazon (images downloaded from online merchants), Webcam (low-resolution images by a web camera), and DSLR (high-resolution images by a digital SLR camera). We regard each dataset as a source domain. Caltech-256 [22] is used as the target domain. There are totally 10 classes which are common among all four datasets: Backpack, Touring-bike, Calculator, Head, Phones, Computer-keyboard, Laptop-101, Computer-monitor, Computer-mouse, Coffee-mug, and Video-projector. There are 8 to 151 samples per category per domain, and 2533 images in total. Fig. 3 highlights the differences among these domains with example images from the category of Computer-monitor.

![Fig. 3. Example images from the Computer-monitor category in different domains](image)

We extract the 4096 dimensional DeCAF$_6$ features [23] from the raw images. Then, these features from different domains are used to learn a classification model for the target domain. The classification accuracies of the proposed method compared with the Base and DAM are recorded in Table I. The highest accuracy among different methods is highlighted in bold.

| TABLE I: CLASSIFICATION ACCURACIES ON TARGET RECOGNITION EXPERIMENTS |
|-----------------|------------------|-------------------|-----------------|-----------------|-----------------|-----------------|-----------------|
| $n$             | C4.5             | Naive Bayes       | LS-SVM           | C4.5             | Naive Bayes       | LS-SVM           | C4.5             | Naive Bayes       | LS-SVM           |
|                 | Base             | (MS)/TL           | Base             | (MS)/TL           | Base             | (MS)/TL           | Base             | (MS)/TL           | Base             |
| 0               | 0.9202           | 0.9319            | 0.9427           | 0.6836           | 0.7254           | 0.8000           | 0.5125           | 0.5125           | 0.5482           |
| 2               | 0.9404           | 0.9625            | 0.9795           | 0.7023           | 0.7622           | 0.7708           | 0.4902           | 0.4902           | 0.5355           |
| 4               | 0.9559           | 0.9701            | 0.9876           | 0.6764           | 0.7239           | 0.7764           | 0.5465           | 0.5519           | 0.5677           |
| 6               | 0.9103           | 0.9439            | 0.9622           | 0.6534           | 0.7505           | 0.8528           | 0.5940           | 0.6974           | 0.7085           |
| 10              | 0.9342           | 0.9672            | 0.9744           | 0.6275           | 0.8577           | 0.9483           | 0.6124           | 0.7992           | 0.8602           |
| 15              | 0.9316           | 0.9747            | 0.9753           | 0.6488           | 0.9515           | 0.9555           | 0.5865           | 0.9121           | 0.9504           |
| 20              | 0.9371           | 0.9760            | 0.9772           | 0.7015           | 0.9708           | 0.9798           | 0.7573           | 0.9477           | 0.9657           |
| Ave.            | 0.9328           | 0.9609            | 0.9713           | 0.6705           | 0.8203           | 0.8691           | 0.5596           | 0.7015           | 0.7337           |

In Table I, (MS)$^2$TL can effectively improve the classification accuracy regardless of source classifiers and the number of the labeled target samples. The experimental results demonstrate that (MS)$^2$TL could better explore the relevant relationship between the source domains and the target domain, and transfer more knowledge from the source domains to promote the target domain learning. As the Base method uses the source classifiers directly without considering the difference between domains, its classification results are always bad. Besides, the average accuracies of (MS)$^2$TL are also higher than Base and DAM (last row in each table). In addition, it is can be found that the accuracies of (MS)$^2$TL and DAM generally increase along with the increasing of the number of the labeled target domain samples.

B. Experiments on Document Retrieval

In this section, the experiments are conducted for the application of document retrieval. The experimental dataset is the 20 Newsgroups dataset [24] which contains 18774 documents, and has a hierarchical structure with 6 main categories and 20 subcategories. To use the dataset for the purpose of multi-source transfer learning experiments, we regard the subcategories per main category as the samples of the common class in different domains. We choose the samples from three main categories with at least four subcategories and generate three settings to evaluate the algorithms (see Table II for the detailed settings). In every setting, we consider one main category as the positive class and use another one as the negative class, and employ all the samples from two subcategories to construct one domain.
In the 20 Newsgroups dataset, each document is represented by the 61188 dimensional word-frequency features. Since the feature dimension is very high, we only perform the experiments with one kind of source classifier (i.e. LS-SVM). The classification accuracies of the methods under different number of the labeled target domain samples are recorded in Table III. The highest accuracy among different methods is highlighted in bold.

Table III shows that our method (MS)\(^2\)TL outperforms other algorithms in most cases except that it performs slightly worse than DAM in two cases when setting \(n=0\), and 2 (see setting "comp vs. rec" and "comp vs. sci" in Table III). When the number of labeled samples per class (i.e., \(n\)) from the target domain increases, the performances of (MS)\(^2\)TL and DAM both improve. We observe that the Base method also achieves good results by only using the source domain classifiers directly, possibly because the source domains are highly relevant to the target domain. This conjecture is also supported by the similarities between the sources and the target according to section III.A.

### C. Parameter Analysis

For further studying the performance of the proposed (MS)\(^2\)TL, the influences of the parameters are considered. In this section, we evaluate the performance variations with respect to the regularization parameters \(\lambda_L\), \(\lambda_L\), and the number of neighbors \(N_k\) by using the datasets of target recognition described in Section IV.A. When evaluating the performance variations with respect to one parameter, we fix the other parameters as their default values (see the beginning of the Section IV).

First, we consider the performance variations w.r.t. the regularization parameter \(\lambda\). We choose the LS-SVM as the source domain classifier since it also has the parameter \(\lambda\). In the experiments, \(\lambda\) is tuned in the range \([10^{-3}, 10^{-2}, 10^{-1}, 1, 10, 10^2, 10^3}\). Here, the classification results of all methods with different number of the labeled target samples (i.e., \(n\)) are shown in Fig. 4. We observe that our method (MS)\(^2\)TL is better than other methods by using different \(\lambda\)’s in most cases. If there is no labeled samples in the target domain (i.e., \(n=0\)), DAM has no improvements compared with Base while (MS)\(^2\)TL still achieves the highest classification accuracy in most cases. In the case of labeled samples existing in the target domain (i.e., \(n=10\)), the performances of DAM and (MS)\(^2\)TL tend to saturate when \(\lambda\) becomes large while the classification results of Base are always not good. To sum up, fixing the value of \(\lambda\) at \([10^{-3}, 10^{-1}]\) is recommended.
The performance variations with respect to different $\lambda_L$ and $\lambda_D$ are shown in Fig. 5. Specifically, we set $\lambda_L$ and $\lambda_D$ for $(MS)^2$TL as 0.1, 1, 10, 10$^2$, 10$^3$, 3*10$^3$, 5*10$^3$, and 10$^4$ respectively. From Fig. 5, we observe that the performance of $(MS)^2$TL changes a little along with the variation of $\lambda_L$ while it changes dramatically along with the variation of $\lambda_D$. It demonstrates that the regularizer $\Omega_D(f^T)$ has a big influence on the performance of $(MS)^2$TL. Compared with the two settings in Fig. 5 (i.e., $n=0$ or 10), we also observe that $(MS)^2$TL achieves the highest accuracy at a larger value of $\lambda_D$ when there is no labeled target domain samples (i.e., $n=0$). It can be explained that $(MS)^2$TL depends more on the $\Omega_D(f^T)$ when no labeled target domain samples exist. We also observe that the performances become stable when setting $\lambda_D \geq 10$ in all cases.

$N_k$ is the number of neighbors for the calculation of $DN_k^L$ (see (4)). We show the performance of $(MS)^2$TL with different $N_k$ in Fig. 6, where $N_k$ is set as 2, 4, 6, 8, 10, 12, and 14. In both two settings of Fig. 5 (i.e., $n=0$ or 10), we can see that the performance of $(MS)^2$TL depends on the setting of parameter $N_k$. Especially, this dependence is evident when no labeled samples exist in the target domain. This may be because that $(MS)^2$TL will depend more on the knowledge from the sources if there are no labeled target domain samples, then $N_k$ will have a bigger influence on $(MS)^2$TL since $N_k$ is a key parameter for the knowledge transfer. In most cases, it is observed that the learning performance will be badly hurt if the value of $N_k$ is too high or too low. The reason can be concluded as: if the value of $N_k$ is set too small, the local scope can not cover all the affinitive examples; on the contrary, if the value is fixed beyond normal scope, the similarity measure may suffer interfere from the false distribution of the irrelevant data. Thus, fixing the value of $N_k$ at [6, 10] is recommended. In addition, the influence of $N_k$ is small when C4.5 is used as the source classifier.
D. Application to Communication Specific Emitter Identification

Communication specific emitter identification [25]-[26] is widely used in applications such as spectrum management, cognitive radio, network intrusion detection, intelligence gathering, etc. This system discerns wire-less radio emitters of interest only based on the external signal feature measurements. However, in the real-world application, the feature of the emitter signal always changes along with different operation modes, different times and other conditions. It is difficult to make sure that the training data collected previously is suitable for the current target task. Here, the proposed transfer learning algorithm (MS)\textsuperscript{2}TL is applied to this application.

Digital radios with the same type and same modulation mode are selected as the specific emitters, whose transmitting signal bandwidth is 25 KHz. The signal is sampled at the sampling frequency of 204.8 KHz under different conditions, e.g., different work frequencies (160MHz or 410MHz), different speakers (speaker 1, speaker 2 or speaker 3), and different receive distances (short distance with direct wave or long distance without direct wave). After the raw data of emitter signal are obtained, we extract the widely adopted emitter features such as envelope box dimension, envelope information dimension, Lempel-Ziv complexity, high-order spectrum, and Hilbert spectrum. Fig. 7 shows the instantaneous envelope and the extracted features of one radio emitter’s signal. To validate the performance of transfer learning, we select data sets under various conditions as the source domains and target domain. Information on these datasets is tabulated in Table IV.

**TABLE IV: EXPERIMENTAL RADIO Emitter DATA**

<table>
<thead>
<tr>
<th>Domains</th>
<th>Work frequency</th>
<th>Speaker</th>
<th>Receive distance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Target</td>
<td>160MHz</td>
<td>Speaker 1</td>
<td>long distance</td>
</tr>
<tr>
<td>Source 1</td>
<td>410MHz</td>
<td>Speaker 1</td>
<td>long distance</td>
</tr>
<tr>
<td>Source 2</td>
<td>160MHz</td>
<td>Speaker 2</td>
<td>long distance</td>
</tr>
<tr>
<td>Source 3</td>
<td>160MHz</td>
<td>Speaker 3</td>
<td>long distance</td>
</tr>
<tr>
<td>Source 4</td>
<td>160MHz</td>
<td>Speaker 1</td>
<td>short distance</td>
</tr>
</tbody>
</table>

100 samples of each emitter are randomly chosen from ‘Target’ in Table IV as the target domain. The other datasets in Table IV are used as the source domains. In the target domain, we also choose \( n \) samples per emitter as the labeled target samples for experiments. \( n \) is also tuned in the range [0, 2, 4, 6, 10, 15, 20]. For each source domain, we choose a certain number of samples per
emitter for experiments. The number is set as 10, 50, and 100. All the other parameters are set as the default values described at the beginning of Section IV. The experiments are also repeated for 20 times with different source samples and target samples. The average classification accuracies are recorded in Table V to Table VII.

It can be seen that the (MS)²TL has generally achieved higher classification accuracies compared with Base and DAM. The highest classification accuracy of (MS)²TL can be as high as 93.76%. All of these experimental results show that (MS)²TL is more suitable for communication specific emitter identification task.

<table>
<thead>
<tr>
<th>TABLE V: CLASSIFICATION ACCURACY WHEN THE SAMPLE NUMBER PER CLASS OF EACH SOURCE IS 10</th>
<th>C4.5</th>
<th>Naïve Bayes</th>
<th>LS-SVM</th>
</tr>
</thead>
<tbody>
<tr>
<td>n</td>
<td>Base</td>
<td>DAM</td>
<td>(MS)²TL</td>
</tr>
<tr>
<td>0</td>
<td>0.4769</td>
<td>0.7586</td>
<td>0.8879</td>
</tr>
<tr>
<td>2</td>
<td>0.4785</td>
<td>0.7719</td>
<td>0.8932</td>
</tr>
<tr>
<td>4</td>
<td>0.4724</td>
<td>0.8243</td>
<td>0.9035</td>
</tr>
<tr>
<td>6</td>
<td>0.4945</td>
<td>0.8477</td>
<td>0.9061</td>
</tr>
<tr>
<td>10</td>
<td>0.4673</td>
<td>0.8756</td>
<td>0.9126</td>
</tr>
<tr>
<td>15</td>
<td>0.4939</td>
<td>0.9248</td>
<td>0.9336</td>
</tr>
<tr>
<td>20</td>
<td>0.4986</td>
<td>0.9134</td>
<td>0.9194</td>
</tr>
<tr>
<td>Ave.</td>
<td>0.4832</td>
<td>0.8452</td>
<td>0.9081</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>TABLE VI: CLASSIFICATION ACCURACY WHEN THE SAMPLE NUMBER PER CLASS OF EACH SOURCE IS 50</th>
<th>C4.5</th>
<th>Naïve Bayes</th>
<th>LS-SVM</th>
</tr>
</thead>
<tbody>
<tr>
<td>n</td>
<td>Base</td>
<td>DAM</td>
<td>(MS)²TL</td>
</tr>
<tr>
<td>0</td>
<td>0.4830</td>
<td>0.7192</td>
<td>0.8072</td>
</tr>
<tr>
<td>2</td>
<td>0.4719</td>
<td>0.7983</td>
<td>0.8765</td>
</tr>
<tr>
<td>4</td>
<td>0.4821</td>
<td>0.7972</td>
<td>0.8756</td>
</tr>
<tr>
<td>6</td>
<td>0.4660</td>
<td>0.8583</td>
<td>0.8967</td>
</tr>
<tr>
<td>10</td>
<td>0.4899</td>
<td>0.8855</td>
<td>0.9099</td>
</tr>
<tr>
<td>15</td>
<td>0.4907</td>
<td>0.8984</td>
<td>0.9064</td>
</tr>
<tr>
<td>20</td>
<td>0.4781</td>
<td>0.9336</td>
<td>0.9376</td>
</tr>
<tr>
<td>Ave.</td>
<td>0.4802</td>
<td>0.8415</td>
<td>0.8871</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>TABLE VII: CLASSIFICATION ACCURACY WHEN THE SAMPLE NUMBER PER CLASS OF EACH SOURCE IS 100</th>
<th>C4.5</th>
<th>Naïve Bayes</th>
<th>LS-SVM</th>
</tr>
</thead>
<tbody>
<tr>
<td>n</td>
<td>Base</td>
<td>DAM</td>
<td>(MS)²TL</td>
</tr>
<tr>
<td>0</td>
<td>0.4668</td>
<td>0.7669</td>
<td>0.7755</td>
</tr>
<tr>
<td>2</td>
<td>0.4678</td>
<td>0.7812</td>
<td>0.8180</td>
</tr>
<tr>
<td>4</td>
<td>0.4682</td>
<td>0.8296</td>
<td>0.8379</td>
</tr>
<tr>
<td>6</td>
<td>0.4725</td>
<td>0.8579</td>
<td>0.8707</td>
</tr>
<tr>
<td>10</td>
<td>0.4973</td>
<td>0.8917</td>
<td>0.9032</td>
</tr>
<tr>
<td>15</td>
<td>0.4988</td>
<td>0.9101</td>
<td>0.9136</td>
</tr>
<tr>
<td>20</td>
<td>0.4915</td>
<td>0.9047</td>
<td>0.9068</td>
</tr>
<tr>
<td>Ave.</td>
<td>0.4804</td>
<td>0.8489</td>
<td>0.8608</td>
</tr>
</tbody>
</table>

V. CONCLUSIONS

In this paper, a novel multi-source transfer learning method called (MS)²TL is proposed. The method explores the relationships between the source domains and the target domain by multi-similarity metric. Then, the knowledge of the source domains is transferred to the target domain based on the smoothness assumption, which enforces that the target classifier shares similar decision values with the relevant source domain classifiers on the unlabeled target samples. The method can import more knowledge from multiple sources for the target learning and also increase the chance of finding the source domains closely related to the target domain to reduce the "negative transfer". Furthermore, the proposed method only needs the pre-learned source domain classifiers when training the target domain classifier, which is suitable for large datasets. We also use the ε-insensitive loss function to enforce the sparsity of the decision function for fast label prediction. Extensive experiments on the target recognition and document retrieval clearly demonstrate the effectiveness of our method. For further showing the practicality, (MS)²TL is also applied to the task of communication specific emitter identification and the result is also satisfying.

ACKNOWLEDGMENT

Both authors would like to acknowledge the support of Key Laboratory of Electronic Restriction of Anhui Province, National High-tech R&D Program (863 Program), and Anhui Provincial Natural Science Foundation (NO.1308085QF99, NO.1408085MKL46).

REFERENCES


Zhen Liu was born in Anhui Province, China, in 1989. He received the B.S. degree in electrical engineering & automation from Anhui University, Hefei, China, in 2010 and the M.S. degree in circuits & systems from Electronic Engineering Institute, Hefei, China, in 2013. He is currently pursuing the Ph.D. degree with the Department of Communications, Electronic Engineering Institute, Hefei, China. His research interests include communication specific emitter identification, intelligent computing, data mining and machine learning.

Jun-An Yang was born in Anhui Province, China, in 1965. He received the B.S. degree in radio technology from Southeast University, Nanjing, China in 1986 and the M.S. degree in communication & information systems from Electronic Engineering Institute, Hefei, China, in 1991. He received the Ph.D. degree in signal & information processing from University of Science and Technology of China (USTC), Hefei, China, in 2003. He is currently a professor in the Department of Communications at Electronic Engineering Institute, Hefei, China. His research interests include neural computing, large-scale machine and computer vision.
Hui Liu was born in Anhui Province, China, in 1983. He received the B.S. degree in communication engineering from Wuhan University, Wuhan, China, in 2005. He received the M.S. degree and the Ph.D. degree in communication & information systems from Electronic Engineering Institute, Hefei, China, in 2008 and 2011 respectively. He is currently a lecturer in the Department of Communications at Electronic Engineering Institute, Hefei, China. His research interests include intelligent information processing and cognitive communication.

Wei Wang was born in Anhui Province, China, in 1987. He received the B.S. degree in mechanism design, manufacturing & automatization from Northwestern Polytechnical University, Xi’an, China, in 2008. He received the M.S. degree and the Ph.D. degree in precision instrument and machinery from University of Science and Technology of China (USTC), Hefei, China, in 2011 and 2014 respectively. He is currently a lecturer in the Department of Communications at Electronic Engineering Institute, Hefei, China. His research interests include intelligent information processing and computer vision.