SCIENCE CHINA
Information Sciences

November 2020, Vol. 63 210203:1–210203:14
https://doi.org/10.1007/s11432-020-3080-3

• **RESEARCH PAPER** •
Special Focus on Multi-source Information Fusion

Evidential combination of augmented multi-source of information based on domain adaptation

Linqing HUANG[1], Zhunga LIU[1*], Quan PAN[1] & Jean DEZERT[2]

[1]*School of Automation, Northwestern Polytechnical University, Xi'an* 710072*, China;*
[2]*ONERA-The French Aerospace Lab, Palaiseau* F-91761*, France*

Received 16 June 2020/Revised 19 August 2020/Accepted 6 October 2020/Published online 22 October 2020

Abstract In the applications of domain adaptation (DA), there may exist multiple source domains, and each source domain usually provides some auxiliary information for object classification. The combination of such complementary knowledge from different source domains is helpful for improving the accuracy. We propose an evidential combination of augmented multi-source of information (ECAMI) method. The information sources are augmented at first by merging several randomly selected source domains to generate extra auxiliary information. We can obtain one piece of classification result with the assistance of each information source based on DA. Then these multiple classification results are combined by belief functions theory, which is expert at dealing with the uncertain information. Nevertheless, the classification results derived from different information sources may have different weights. The optimal weights are calculated by minimizing an given error criteria defined by the distance between the combination result and the ground truth using some training data. For each object, the augmented information sources will produce multiple classification results that will be discounted by the learnt weights under the belief functions framework. Then the combination of these discounted results is employed to make the final class decision. The effectiveness of ECAMI is evaluated with respect to some related methods based on several real data sets, and the experimental results show that ECAMI can significantly improve the classification accuracy.

Keywords information fusion, domain adaptation, evidence theory, belief functions, pattern classification

Citation Huang L Q, Liu Z G, Pan Q, et al. Evidential combination of augmented multi-source of information based on domain adaptation. Sci China Inf Sci, 2020, 63(11): 210203, https://doi.org/10.1007/s11432-020-3080-3

1 Introduction

In pattern classification, most of the machine learning methods work with a common assumption that the training and test patterns have the same distribution and feature space. If the distribution or feature space changes, the classification model should be rebuilt from scratch using newly collected training data. However, the labeled patterns may be expensive or time-consuming to obtain in some cases, and the standard machine learning methods cannot work well with few or no labeled training data. It would be helpful if we can transfer the knowledge from some related domains (called source domains) into the new domain (called target domain) for building a reliable classification model. Domain adaptation as one special setting of transfer learning to solve classification problem without many labeled patterns has been successfully employed in applications [1]. It utilizes the labeled patterns (knowledge) in the source domain for classifying unseen objects in the target domain when the source and target domain data are in the same feature space but drawn from different distributions.

* Corresponding author (email: liuzhunga@nwpu.edu.cn)

The major issue of domain adaptation is how to effectively reduce the distribution discrepancy between domains and preserve the original information in the source and target domains as much as possible. Existing studies for solving such problem can be summarized into two categories. (1) Patterns re-weighting [2], which reuses the patterns in the source domains with some weighting techniques for building classification models. For instance, Dai et al. [2] extended the AdaBoost method to Transfer AdaBoost (TrAdaBoost) method for solving domain adaptation problem. It adds a mechanism that increases the weights of wrongly predicted patterns in the target domain and decreases the weights of incorrectly classified patterns in the source domain at each iteration. (2) Feature matching (also called distribution alignment) [3], which discovers a new feature representation to make distributions close to each other. Recently, many approaches [3–8] have been proposed to learn the new feature representation. Pan et al. [3] proposed a transfer component analysis (TCA) method to acquire a new representation by minimizing the marginal distribution discrepancy between domains. Joint distribution adaptation (JDA) [4] considers both the marginal and conditional distributions in a dimensionality reduction procedure. The pseudo labels of target data which are predicted by the classifier in the source domain are used to approximately estimate the conditional distribution. Balanced distribution adaptation (BDA) [7] adaptively leveraged the importance of the marginal and conditional distribution discrepancies to discover more robust domain-invariant feature. Deng et al. [8] proposed an explicit map-based feature selection (EMFS) method which uses explicit feature map and feature selection to reveal the high-order invariant features. Above mentioned methods focus on learning shallow features by minimizing domain discrepancy between the source and target domains. In recent studies [9–11], the deep networks can learn much more transferable features for domain adaptation. Wen et al. [9] proposed a new deep transfer learning (DTL) method using three-layer sparse auto-encoder to extract the features of data for fault diagnosis by minimizing the discrepancy between domains. Long et al. [10] proposed a deep adaptation networks (DAN) method to embed the deep features of all task-specific layers into reproducing kernel Hilbert spaces and match the distributions. The deep features are made more transferable by exploiting low-density separation of target domain data in deep architectures. Tzeng et al. [11] proposed an adversarial discriminative domain adaptation (ADDA) method based on adversarial learning which combines discriminative modeling, untied weight sharing and a GAN loss to handle domain shifts.

In applications, multiple source domains are often available. The multi-source domain adaptation problem is still an open challenging question for object classification [12]. The classification accuracy is expected to be improved using the complementary information in different source domains, and many methods [13–18] have been proposed in recent years. Liu et al. [13] proposed a structure-preserved multi-source domain adaptation (SPMDA) method. In this method, the source and target data are put together for clustering to explore the structures of the source and target domains for improving the classification accuracy in the target domain. Duan et al. [14] proposed to learn a robust target classifier by utilizing all task functions in multiple source domains into a unified loss function which contains domain-dependent regularizer and data-dependent regularizer. Ding et al. [15] proposed an incomplete multi-source domain adaptation method for sharing knowledge from two directions, i.e., cross-domain transfer and cross-source transfer. Similarly, the deep feature can provide more transferable information in the multi-source domains setting. Multiple feature space adaptation network (MFSAN) [16] aligns not only the domain-specific distribution of each source and the target domains by learning multiple domain-invariant representations but also the outputs of classifiers from multiple sources. Xu et al. [17] proposed a deep cocktail network (DCTN) to battle the domain shifts and category shifts where classes from different sources are non-consistent among multiple sources. Multi-source domain adversarial network (MDAN) [18] aims to learn the feature representations that are invariant under multiple domain shifts while at the same time being discriminative for learning task.

These methods for solving multi-source domain adaptation issue can be regarded as data-level combination techniques. They lose some specific information among source domains, and may not effectively extract the complementary knowledge. We want to take full advantage of the information contained in different source domains from the decision-level standpoint, which can preserve all the specific knowledge. In fusion system, the source domain data sets are regarded as information sources to provide knowledge for classifying objects. The classification results yielded by the auxiliary of different information sources can be combined by many decision-level fusion methods. If the classification results are crisp outputs (single crisp labels), the simple majority vote (MV) method is often employed. Fuzzy rules [19] and evidence theory [20] can be used to combine soft classification results which can provide more information than single crisp labels. Belief functions offer an interesting mathematical framework to model uncertainty, and to fuse uncertain sources of evidence. They have been successfully applied in multi-source information fusion [21], so it is employed here to improve accuracy by combining multiple soft classification results.

For unseen objects, the combination of classification results yielded by multiple information sources usually gets a good classification performance because the knowledge provided by different information sources is complementary to each other. The fusion operation makes the combination result be closer to the ground truth compared with only using knowledge in individual source domain. In our previous work [22], we proposed a transfer classification method with multiple source domains. Each source domain produced one piece of classification result, and the different classification results corresponding to different source domains were combined (with weighting factors) for making the final class decision of the object under concern. Nevertheless, the weighting factors in the combination were determined by the distribution distances between the source and target domains, and they were not adaptively learnt by training data. Because the proper augmentation [23, 24] of information sources (i.e., source domains with different distributions) can generate more complementary knowledge to further improve the classification accuracy, we propose to use it in our new evidential combination of augmented multi-source of information (ECAMI) method presented in this paper. For this, we use the augmentation operation to obtain more information sources to produce extra classification for combination. The distribution of the union of several existing source domain data sets is different from that of the singleton source domain data set. Its discernment information keeps after distribution alignment is diverse from that contained in the individual source domains; i.e., the union of several existing source domain data sets can also provide some useful knowledge. Thus, we regard the unions of several source domain data sets as new information sources to yield extra classification results.

In general, high-quality information sources usually have positive influence on the fusion result. The classification results obtained by low-quality information sources having low reliabilities/weights yield poor combination results. Thus, we select some high-quality information sources, i.e., singleton source domain data set and some unions of several source domain data sets, that provide useful knowledge for classifying the objects. The combination of multiple classification results obtained by high-quality information sources will help to improve the accuracy. It is worth noting that the classification results acquired by multiple information sources often have different weights because the domain-consistency between diverse source and target domains is different. The source domain which is very consistent to the target domain can provide quite large useful information for object classification, and the obtained classification results will be quite reliable. The yielded classification results will not be reliable if the source domain is not close to the target domain. In other words, the reliabilities/weights of information sources are different in general. The augmented information sources are coming from the singleton source domain data sets, and the original information in different source domains overlaps to some extent. Thus, the information sources may correlate to each other. We have to discount the classification results obtained by the information sources to reduce the negative influence of reliability and correlation among information sources. The weighting factors are learnt by an optimization operation using labeled patterns in the source domain. The discounted results with corresponding weights can be combined by some fusion rules (here we use Dempster's rule) to make the fusion result be close to the ground truth.

The remainder of this paper is organized as follows. Section 2 briefly introduces transfer learning. The evidential combination of augmented multi-source of information method is presented in Section 3, and the experiments to validate ECAMI are reported in Section 4. Section 5 concludes the paper.

2 Basics of transfer learning

Transfer learning is an appealing paradigm to handle classification problem with few or no labeled training patterns [1]. It has achieved great success in many fields [25], such as cross-domain classification [26], clustering [27], and WiFi localization [28]. It has two important concepts: domain and task. Domain $\mathcal{D}$ consists of two elements: a feature space $\mathcal{X}$ and a marginal probability $P(\boldsymbol{X})$, where $\boldsymbol{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n] \in \mathcal{X}$ is the set of patterns in the feature space $\mathcal{X}$. Task $\mathcal{T}$ also has two components: a label space $\mathcal{Y}$ and a prediction function $f(\cdot)$ which is used to predict the label of unseen objects.

In transfer learning, the domain and task are respectively denoted by $\mathcal{D} = \{\mathcal{X}, P(\boldsymbol{X})\}$ and $\mathcal{T} = \{\mathcal{Y}, f(\cdot)\}$. The source domain, source task and source domain data set are described as $\mathcal{D}_S = \{\mathcal{X}_S, P(\boldsymbol{X}_S)\}$, $\mathcal{T}_S = \{\mathcal{Y}_S, f_S(\cdot)\}$ and $D_S = \{(\boldsymbol{x}_1^S, y_1^S), (\boldsymbol{x}_2^S, y_2^S), \ldots, (\boldsymbol{x}_{n_S}^S, y_{n_S}^S)\}$. Similarly, the target domain, target task and target domain data set are denoted by $\mathcal{D}_T = \{\mathcal{X}_T, P(\boldsymbol{X}_T)\}$, $\mathcal{T}_T = \{\mathcal{Y}_T, f_T(\cdot)\}$ and $D_T = \{\boldsymbol{x}_1^{\mathrm{T}}, \boldsymbol{x}_2^{\mathrm{T}}, \ldots, \boldsymbol{x}_{n_T}^{\mathrm{T}}\}$. A unified definition of transfer learning based on above concepts and notations is as follows.

Definition 1 ([1])**.** Given a source domain $\mathcal{D}_S$ and source task $\mathcal{T}_S$, a target domain $\mathcal{D}_T$ and target task $\mathcal{T}_T$, transfer learning aims to help improve the learning of target prediction function $f_T(\cdot)$ in $\mathcal{D}_T$ using the knowledge (i.e., labeled patterns) in $\mathcal{D}_S$ when $\mathcal{D}_S \neq \mathcal{D}_T$ or $\mathcal{T}_S \neq \mathcal{T}_T$.

Obviously, $\mathcal{D}_S \neq \mathcal{D}_T$ represents either $\mathcal{X}_S \neq \mathcal{X}_T$ or $P(X_S) \neq P(X_T)$. When the feature spaces of the source and target domains are the same but the distributions are quite different, i.e., $\mathcal{X}_S = \mathcal{X}_T$ and $P(X_S) \neq P(X_T)$, this case is called domain adaptation or homogeneous transfer learning. If $\mathcal{X}_S \neq \mathcal{X}_T$, it is called heterogeneous transfer learning. In this paper, we consider only the domain adaptation case.

3 Combining augmented transfer classification with muti-source domains

We consider n ($n \geqslant 2$) source domains $\{\mathcal{D}_{S_i}\}_{i=1}^n$ and one target domain $\mathcal{D}_T$ with n source domain data sets $D_{S_i} = \{(\boldsymbol{x}_p^{S_i}, y_p^{S_i})\}_{p=1}^{N_i}, i = 1, \ldots, n$ and one target domain data set $D_T = \{\boldsymbol{x}_q^{\mathrm{T}}\}_{q=1}^{N_T}$, where N_i and N_T are the number of patterns in the i-th source and target domains, $\{\boldsymbol{x}_p^{S_i}\}_{p=1}^{N_i}, \{\boldsymbol{x}_q^{\mathrm{T}}\}_{q=1}^{N_T} \in \mathbb{R}^k$ are the patterns in the i-th source and target domains, k is the feature space dimension of the source and target domains, and $\{y_p^{S_i}\}_{p=1}^{N_i} \in \{\omega_1, \ldots, \omega_c\}$ are real labels. The patterns in the source and target domains are in the same feature space but drawn from different distributions as $\mathcal{X}_{S_1} = \mathcal{X}_{S_2} = \cdots = \mathcal{X}_{S_n} = \mathcal{X}_T$ and $P(X_{S_1}) \neq P(X_{S_2}) \neq \cdots \neq P(X_{S_n}) \neq P(X_T)$. The difference among the source and target domains in such case is shown in Figure 1.

One can see that the patterns in the source and target domains are in the same feature space but do not satisfy the independent and identically distributed (i.i.d.) assumption. The classification models learnt in different source domains cannot be directly used to classify objects in the target domain. Domain adaptation techniques should be employed to reduce the distribution discrepancy between the source and target domains. The work presented in [5, 29] aimed to match/align distributions by learning a new feature representation for patterns in the source and target domains, i.e., by mapping patterns into one new feature space. Traditional machine learning approaches can work in this new feature space because the distributions become close. The patterns in different source domains provide diverse complementary information for classifying unseen objects in the target domain, so the combination of classification results provided by multiple source domains should improve the accuracy of the classification result.

In practice, more information sources can generate more classification results and should lead to good combination results. If we can obtain more information sources, more complementary knowledge could be integrated in the fusion process for improving the classification performance. The distribution of the union of several source domain data sets is different from that of singleton source domain data set. Thus, the discernment information in the union of several source domain data sets after matching distribution will be different from that in the singleton source domain data set. We simply merge the data sets in different source domains as new information sources to provide extra complementary knowledge (i.e., information). More classification results can be obtained by the help of the union of several source domain data sets for combination. However, the classification results produced by the low-quality information sources often have a bad influence on the fusion result. Therefore, we select some high-quality information sources for exploiting only reliable information. The combination of the classification results yielded by

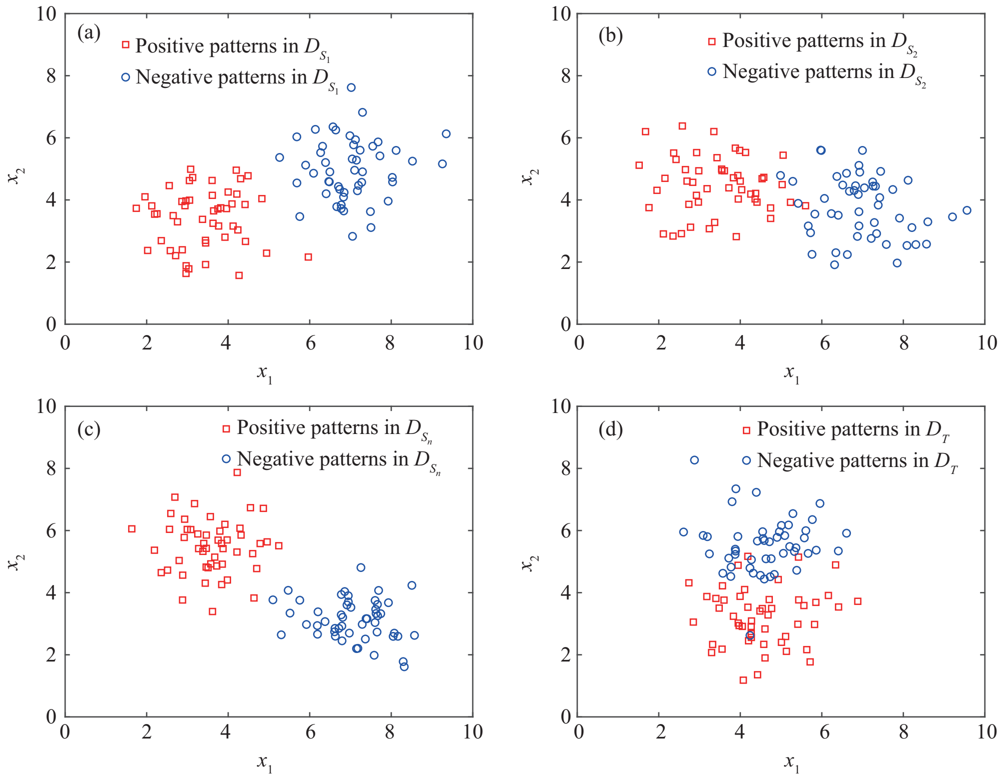

Figure 1 (Color online) Patterns in different domains. (a) Patterns in $\mathcal{D}_{S_1}$; (b) patterns in $\mathcal{D}_{S_2}$; (c) patterns in $\mathcal{D}_{S_n}$; (d) patterns in $\mathcal{D}_T$.

the auxiliary of high-quality information sources is expected to give higher accuracy than classification only based on a singleton source domain data set (information source). Nevertheless, the weights of classification results are usually different because the distribution discrepancy among domains is diverse. Moreover, the information sources may not be distinct, so we also use the discounting process to reduce the negative influence of reliability and correlation before combining multiple classification results. The weighting factors are obtained by an optimization procedure by the labeled data in the source domain described in Subsection 3.2.

3.1 Selection of effective information sources

In applications, the combination results could be improved if more information sources are available to produce classification results. However, the new source domain data sets usually are difficult to obtain for multi-source domain adaptation issue. We regard the union of several existing source domain data sets, e.g., $D_{S_1 \cup S_2} = D_{S_1} \cup D_{S_2}$, as new information sources to produce extra classification results. The interest and justification of this operation to augment information sources are introduced in the sequel.

As example, we consider the TCA [3] which is a classical shallow domain adaptation method. TCA learns a new feature representation across domains by minimizing the reconstruction error of source domain data and target domain data. Let $\boldsymbol{X}_S = [\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_{n_s}] \in \mathbb{R}^{k \times n_s}$, $\boldsymbol{X}_T = [\boldsymbol{x}_{n_s+1}, \boldsymbol{x}_{n_s+2}, \ldots, \boldsymbol{x}_{n_s+n_t}] \in \mathbb{R}^{k \times n_t}$ and $\boldsymbol{X} = [\boldsymbol{X}_S, \boldsymbol{X}_T] = [\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_{(n_s+n_t)}] \in \mathbb{R}^{k \times (n_s+n_t)}$ be the set of patterns in the source domain, the target domain and the two domains respectively, where n_s and n_t are the number of patterns in the source and target domains and k is the feature dimension. One transformation matrix $\boldsymbol{A} \in \mathbb{R}^{k \times \tilde{k}}$ ($\tilde{k} \ll k$) which maps data into a ($\tilde{k}$-dimension) common feature space can be obtained by minimizing

$$\hat{\boldsymbol{A}} = \arg\min_{\boldsymbol{A}} \left\| \frac{1}{n_s} \sum_{p=1}^{n_s} \boldsymbol{A}^{\mathrm{T}} \boldsymbol{x}_p - \frac{1}{n_t} \sum_{q=n_s+1}^{n_s+n_t} \boldsymbol{A}^{\mathrm{T}} \boldsymbol{x}_q \right\|^2 = \arg\min_{\boldsymbol{A}} \operatorname{tr}(\boldsymbol{A}^{\mathrm{T}} \boldsymbol{X} \boldsymbol{M} \boldsymbol{X}^{\mathrm{T}} \boldsymbol{A}), \tag{1}$$

where $\mathrm{tr}(\cdot)$ denotes the trace of a matrix, $\boldsymbol{M} \in \mathbb{R}^{(n_s+n_t)\times(n_s+n_t)}$ is the maximum mean discrepancy (MMD) matrix, and the element $(M)_{pq}$ of this matrix is computed by

$$(M)_{pq} = \begin{cases} \frac{1}{n_s n_s}, & \boldsymbol{x}_p, \boldsymbol{x}_q \in \mathcal{D}_S, \\ \frac{1}{n_t n_t}, & \boldsymbol{x}_p, \boldsymbol{x}_q \in \mathcal{D}_T, \\ -\frac{1}{n_s n_t}, & \text{otherwise}. \end{cases} \tag{2}$$

The distributions between patterns $\boldsymbol{X}_S$ in the source domain and objects $\boldsymbol{X}_T$ in the target domain under new feature representation $\boldsymbol{Z}_S = \hat{\boldsymbol{A}}^{\mathrm{T}}\boldsymbol{X}_S = [\boldsymbol{z}_1, \boldsymbol{z}_2, \ldots, \boldsymbol{z}_{n_s}] \in \mathbb{R}^{\tilde{k}\times n_s}$ and $\boldsymbol{Z}_T = \hat{\boldsymbol{A}}^{\mathrm{T}}\boldsymbol{X}_T = [\boldsymbol{z}_{n_s+1}, \boldsymbol{z}_{n_s+2}, \ldots, \boldsymbol{z}_{n_s+n_t}] \in \mathbb{R}^{\tilde{k}\times n_t}$ are drawn close. One classifier learnt by patterns in the source domain under new feature representation can be used to classify objects in the target domain mapped into this new space. It is worth noting that the transformation matrix $\hat{\boldsymbol{A}}$ acquired by (1) varies with the input $\boldsymbol{X}$. Thus, the new feature representation of patterns will be different when matching distributions of the target domain data and different source domain data, and the discernment information contained in these data under new representation are diverse.

If there are two source domain data sets $D_{S_i}, D_{S_j}, i \neq j$, and one target domain data set D_T, the transformation matrices $\hat{\boldsymbol{A}}_i, \hat{\boldsymbol{A}}_j \in \mathbb{R}^{k\times\tilde{k}}$ can be obtained by matching distributions with two inputs $[\boldsymbol{x}_1^{S_i}, \boldsymbol{x}_2^{S_i}, \ldots, \boldsymbol{x}_{N_i}^{S_i}, \boldsymbol{x}_1^{\mathrm{T}}, \boldsymbol{x}_2^{\mathrm{T}}, \ldots, \boldsymbol{x}_{N_T}^{\mathrm{T}}] \in \mathbb{R}^{k\times(N_i+N_T)}$ and $[\boldsymbol{x}_1^{S_j}, \boldsymbol{x}_2^{S_j}, \ldots, \boldsymbol{x}_{N_j}^{S_j}, \boldsymbol{x}_1^{\mathrm{T}}, \boldsymbol{x}_2^{\mathrm{T}}, \ldots, \boldsymbol{x}_{N_T}^{\mathrm{T}}] \in \mathbb{R}^{k\times(N_j+N_T)}$. The new feature representation of patterns in the i-th, j-th source and target domains is given by

$$\begin{cases} \hat{\boldsymbol{x}}_p^{S_i} = \hat{\boldsymbol{A}}_i^{\mathrm{T}}\boldsymbol{x}_p^{S_i}, \ p = 1, \ldots, N_i, \\ \hat{\boldsymbol{x}}_p^{S_j} = \hat{\boldsymbol{A}}_j^{\mathrm{T}}\boldsymbol{x}_p^{S_j}, \ p = 1, \ldots, N_j, \end{cases} \tag{3}$$

and

$$\begin{cases} \hat{\boldsymbol{x}}_q^{T_i} = \hat{\boldsymbol{A}}_i^{\mathrm{T}}\boldsymbol{x}_q^{\mathrm{T}}, \ q = 1, \ldots, N_T, \\ \hat{\boldsymbol{x}}_q^{T_j} = \hat{\boldsymbol{A}}_j^{\mathrm{T}}\boldsymbol{x}_q^{\mathrm{T}}, \ q = 1, \ldots, N_T. \end{cases} \tag{4}$$

Similarly, let the union of the i-th and j-th source domain data sets $D_{S_i\cup S_j} = D_{S_i} \cup D_{S_j}$ and the target domain data set D_T be input data. The transformation matrix $\hat{\boldsymbol{A}}_{i,j} \in \mathbb{R}^{k\times\tilde{k}}$ can be acquired using (1) with the input $[\boldsymbol{x}_1^{S_i}, \boldsymbol{x}_2^{S_i}, \ldots, \boldsymbol{x}_{N_i}^{S_i}, \boldsymbol{x}_1^{S_j}, \boldsymbol{x}_2^{S_j}, \ldots, \boldsymbol{x}_{N_j}^{S_j}, \boldsymbol{x}_1^{\mathrm{T}}, \boldsymbol{x}_2^{\mathrm{T}}, \ldots, \boldsymbol{x}_{N_T}^{\mathrm{T}}] \in \mathbb{R}^{k\times(N_i+N_j+N_T)}$. One gets the new feature representation by

$$\begin{cases} \tilde{\boldsymbol{x}}_p^{S_i} = \hat{\boldsymbol{A}}_{i,j}^{\mathrm{T}}\boldsymbol{x}_p^{S_i}, \ p = 1, \ldots, N_i, \\ \tilde{\boldsymbol{x}}_p^{S_j} = \hat{\boldsymbol{A}}_{i,j}^{\mathrm{T}}\boldsymbol{x}_p^{S_j}, \ p = 1, \ldots, N_j, \\ \tilde{\boldsymbol{x}}_q^{T_{i,j}} = \hat{\boldsymbol{A}}_{i,j}^{\mathrm{T}}\boldsymbol{x}_q^{\mathrm{T}}, \ q = 1, \ldots, N_T. \end{cases} \tag{5}$$

It is obvious that the new feature representations of one pattern $\boldsymbol{x}_p^{S_i}$ using $\hat{\boldsymbol{A}}_i$ and $\hat{\boldsymbol{A}}_{i,j}$ are different, i.e., $\{\hat{\boldsymbol{x}}_p^{S_i} \neq \tilde{\boldsymbol{x}}_p^{S_i}\}_{p=1}^{N_i}$. In other words, the important properties (geometric properties, statistical properties, or side information) obtained by matching the singleton source data set and the target domain data set will be different from those obtained by adapting the union of several source domain data sets and the target domain data set. Therefore, the discernment information involved in the union of several source domain data sets under new feature representation is different from that in the singleton source domain data sets. The classification results obtained by the auxiliary of the union of source domain data sets will provide some extra complementary knowledge for combination. It is interesting and judicious to use this operation for augmenting information sources.

If there are n original source domain data sets, one can obtain $2^n - n - 1$ new source domain data sets by simple merging operation; e.g., three singleton source domain data sets as $D_{S_1}, D_{S_2}, D_{S_3}$ can yield extra four new source domain data sets as $D_{S_1\cup S_2}, D_{S_1\cup S_3}, D_{S_2\cup S_3}, D_{S_1\cup S_2\cup S_3}$. Thus, there will be $2^n - 1$ information sources (i.e., n singleton source domain data sets and $2^n - n - 1$ unions of several source domain data sets) that can help to classify objects in the target domain. In applications, the classification results with low reliabilities usually lead to bad combination result using Dempster's rule [20]. The combination result is expected to be very close to the ground truth when the (high-quality) information sources provide reliable classification results. Therefore, it is desirable to select some high-quality information sources to produce better classification results for combination.

For singleton source domain data set selection, there will be little useful information in some individual source domains with low domain-consistency (it can be measured by MMD [30] or $\mathcal{A}$-distance [31], and the bigger the MMD value or $\mathcal{A}$-distance value, the smaller the domain-consistency) for object classification. The accuracy in the target domain is very low by the auxiliary of singleton source domain data set with very big distribution discrepancy to the target domain data set; i.e., the classification results obtained by low-quality singleton source domain data sets have a bad influence on the combination result. Thus, we select the most consistent one to get reliable classification results for reducing the negative influence as much as possible.

For unions of several source domain data sets selection, the augmented new information sources involve all knowledge in multiple singleton source domains. More information will be preserved after matching distributions compared with the singleton source domain data sets. The accuracy can be improved to some degree using the new information sources, so we select some (more than one) unions of several source domain data sets to produce extra classification results for combination. In practice, the new information sources which involve the most consistent source domain data set usually have much useful discernment information for classifying objects. Therefore, the unions of several source domain data sets insisting of the singleton source domain data set with the highest domain-consistency are simply selected.

We select the most consistent singleton source domain data set and some unions of several (i.e., $2, 3, \ldots, n$) source domain data sets as high-quality information sources. When n source domain data sets are available, the number of selected information sources will be $1+C_{n-1}^{1}+C_{n-1}^{2}+\cdots+C_{n-1}^{n-1}=2^{n-1}$. We aim to effectively combine the multiple classification results yielded by the selected 2^{n-1} high-quality information sources. We use the belief functions framework here to combine multiple classification results because it can well model uncertainty and it has been already successfully used in real applications [32, 33], e.g., data classification [34], data clustering [35], decision making [32], fault prediction [36, 37] and information fusion [21].

3.2 Weighted combination of transfer classification

We consider a frame of discernment (FoD) $\Omega=\{\omega_1,\omega_2,\ldots,\omega_c\}$ consisting of c exclusive and exhaustive elements. In c-class classification problem, the element ω_i represents the i-th class, and the FoD Ω is the class space. The power set denoted by 2^{Ω} is the set of all subsets of Ω. It contains $2^{|\Omega|}$ elements, where $|\Omega|$ is the cardinality of Ω. For example, if the FoD is $\Omega=\{\omega_1,\omega_2,\omega_3\}$, the power set of Ω can be denoted by $2^{\Omega}=\{\emptyset,\{\omega_1\},\{\omega_2\},\{\omega_3\},\{\omega_1,\omega_2\},\{\omega_1,\omega_3\},\{\omega_2,\omega_3\},\Omega\}$.

The basic belief assignment (BBA) also called mass function is defined as a mapping from 2^{Ω} to $[0,1]$, and satisfies $\sum_{A\in 2^{\Omega}} m(A)=1$ and $m(\emptyset)=0$, where $m(A)$ represents the belief that one is willing to commit exactly A and not to any of its subsets. The element A is called a focal element of the BBA when $m(A)>0$. If all the focal elements are singletons, the BBA $m(\cdot)$ is said to be Bayesian BBA. In c-class classification problem, $m(A)$ represents the support degree of one object assigned to the singleton class (e.g., $A=\omega_i, i=1,\ldots,c$) or the union of several classes (e.g., $A=\{\omega_i,\omega_j\}, i\neq j$) which characterizes the partial ambiguity. The upper and lower bounds of probability called belief function $\mathrm{Bel}(\cdot)$ and the plausibility function $\mathrm{Pl}(\cdot)$ are respectively defined by $\mathrm{Bel}(A)=\sum_{B\subseteq A} m(B)$ and $\mathrm{Pl}(A)=\sum_{B\cap A\neq\emptyset} m(B)$.

In this work, the classification results (i.e., sources of evidence) characterized by simple Bayesian BBA under the class space Ω are combined by Dempster's rule (a.k.a., DS rule) [20] because of its associative and commutative properties. Let us consider two distinct pieces of evidence $\boldsymbol{m}_1$ and $\boldsymbol{m}_2$ ($\boldsymbol{m}_1=[m_1(\{\omega_1\}), m_1(\{\omega_2\}),\ldots,m_1(\Omega)], \boldsymbol{m}_2=[m_2(\{\omega_1\}), m_2(\{\omega_2\}),\ldots,m_2(\Omega)]$), the combination results of DS rule denoted by $\boldsymbol{m}=\boldsymbol{m}_1\oplus\boldsymbol{m}_2$ is mathematically defined by

$$\begin{cases} m(A)=m_1\oplus m_2(A)=\frac{\sum_{B,C\in 2^{\Omega}|B\cap C=A} m_1(B)m_2(C)}{1-K}, \\ m(\emptyset)=0, \end{cases} \tag{6}$$

where $K = \sum_{B\cap C=\emptyset|B,C\in 2^\Omega} m_1(B)m_2(C)$, and it measures the degree of conflict between two BBAs. The symbol $\oplus$ denotes the DS combination operator.

Let us assume that $D_{S_i} = \{(\boldsymbol{x}_p^{S_i}, y_p^{S_i})\}_{p=1}^{N_i}$ is the most consistent source domain data set and the selected singleton source domain data set and the unions of several source domain data sets can be denoted by $\tilde{D}_{S_i}^r, r = 1,\ldots,2^{n-1}$. Let $\{\tilde{\boldsymbol{m}}_{r,p}\}_{p=1}^{N_i}, r = 1,\ldots,2^{n-1}$ and $\{\boldsymbol{m}_{r,q}\}_{q=1}^{N_T}$ be the classification results characterized by Bayesian BBAs of source domain data $\{(\boldsymbol{x}_p^{S_i}, y_p^{S_i})\}_{p=1}^{N_i}$ and the unseen object $\boldsymbol{x}_q^{\rm T}$ in the target domain. The 2^{n-1} classification results $\{\boldsymbol{m}_{r,q}\}_{q=1}^{N_T}$ can be combined by DS rule. It is worth noting that the classification results obtained by the selected (high-quality) information sources should have diverse weights because the domain-consistency of different information sources is diverse and they may correlate to each other to some extent. Shafer [20] proposed a discounting operation before combining multiple classification results with weight $\beta \in [0,1]$. It is employed here before combining multi-source information to make the combination results more representative of what is expected (i.e., ground truth). The discounting method is computed by

$$\begin{cases} {}^{\beta}m(A) = \beta \cdot m(A),\ A \subset \Omega, A \neq \Omega, \\ {}^{\beta}m(\Omega) = 1 - \beta + \beta \cdot m(\Omega). \end{cases} \tag{7}$$

If the BBA is completely reliable, one takes $\beta = 1$ and gets ${}^{\beta}m(A) = m(A), A \subseteq \Omega$. If $\beta = 0$, it indicates that the source of evidence is not reliable at all, and the mass value of all the focal elements will be discounted to the total ignorance. One can obtain ${}^{\beta}m(A) = 0, {}^{\beta}m(\Omega) = 1, A \subset \Omega, A \neq \Omega$, and it plays a neural role on the combination.

We want to learn the weighting factors to discount the multiple sources of evidence for reducing the bad influence of reliability and correlation on the combination. The weights will be learnt by an optimization procedure with DS rule based on the knowledge (labeled patterns) in the source domain, i.e., the optimal weights should make the combination results as close as possible to the ground truth for patterns in the most consistent source domain. Let $\{\boldsymbol{t}_p = [t_{p,1}, t_{p,2}, \ldots, t_{p,c}, t_{p,c+1}]\}_{p=1}^{N_i} \in \mathbb{R}^{c+1}$ be the ground truth[1)] of source domain data $\{(\boldsymbol{x}_p^{S_i}, y_p^{S_i})\}_{p=1}^{N_i}$, the optimal weights $\boldsymbol{\beta} = [\beta_1, \ldots, \beta_{2^{n-1}}]$ ($\beta_r \in [0,1], r = 1,\ldots,2^{n-1}$) will be obtained by minimizing

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \sum_{p=1}^{N_i} \left\| \bigoplus_{r=1}^{2^{n-1}} {}^{\beta_r}\tilde{\boldsymbol{m}}_{r,p} - \boldsymbol{t}_p \right\|^2. \tag{8}$$

The discounting operation aims to reduce the bad influence on the fusion by tuning the ignorance degree of each BBA, and it can take full advantage of complementary knowledge in the augmentation information sources. The weighting factors are optimized by the patterns in the most consistent source domain which are all involved in the selected information sources. It is a classical constrained nonlinear least squares problem that can be solved by active-set algorithm [38] or heuristic algorithm. The fmincon function in MATLAB is used to compute the weighting factors. The time cost of obtaining $\hat{\boldsymbol{\beta}}$ is related to the size of data sets and the number of the augmented information sources. The larger the size of data sets or the bigger the number of the information sources, the higher the time cost. In applications, the weighting factors should be learnt from scratch for different tasks with different source combination. The discounted results $\{{}^{\hat{\beta}_r}\boldsymbol{m}_{r,q}\}_{q=1}^{N_T}$ using (7) with weighting factors $\hat{\beta}_r$ can be combined by

$$\boldsymbol{m}_q = \bigoplus_{r=1}^{2^{n-1}} {}^{\hat{\beta}_r}\boldsymbol{m}_{r,q}. \tag{9}$$

In this work we use the argument of the max of plausibility for making the classification decision; i.e., the unseen object $\boldsymbol{x}_q^{\rm T}$ will be committed to the class with the biggest plausibility value by $\omega = \arg\min_l \mathrm{Pl}(\omega_l)$.

The proposed ECAMI method also can jointly work with deep domain adaptation techniques. In applications, the multiple classification results for combination are yielded by different augmented information

1) It is characterized by one binary vector, and all elements will be zero but t_{pl} =1 if the real class is ω_l. In this work, the combination result contains an ignorance class Ω, so one extra component $t_{p,c+1} = 0$ is included to compute the mean square error.

sources after aligning the distribution between the source and target domains by shallow or deep domain adaptation techniques. The pseudo code of ECAMI is shown in Algorithm 1, and one toy example to clearly illustrate the proposed method is given as follows.

Algorithm 1 ECAMI

Input: The n source domain data sets: $D_{S_i} = \{(\boldsymbol{x}_p^{S_i}, y_p^{S_i})\}_{p=1}^{N_i}, i = 1, \ldots, n$ and one target domain data set: $D_T = \{\boldsymbol{x}_q^{\mathrm{T}}\}_{q=1}^{N_T}$.
1: Generate more information sources by simply merging source domain data sets;
2: Select some high-quality information sources;
3: Obtain the weighting factors by (8);
4: **for** $q = 1$ to N_T **do**
5: Obtain multiple classification results by the auxiliary of selected high-quality information sources;
6: Discount these sources of evidence using (7);
7: Combine discounted results by DS rule with (6);
8: Make class decision using plausibility function value $\mathrm{Pl}(\cdot)$;
9: **end for**
Output: Class decisions.

Example 1. Let us consider that there are three source domain data sets D_{S_1}, D_{S_2}, D_{S_3} and one target domain data set D_T. The class space is $\Omega = \{\omega_1, \omega_2, \omega_3\}$, and the most consistent source domain data set is D_{S_1}. The augmentation operation yields four new information sources $D_{S_1 \cup S_2}$, $D_{S_1 \cup S_3}$, $D_{S_2 \cup S_3}$ and $D_{S_1 \cup S_2 \cup S_3}$. The selected high-quality information sources are D_{S_1}, $D_{S_1 \cup S_2}$, $D_{S_1 \cup S_3}$ and $D_{S_1 \cup S_2 \cup S_3}$. For one unseen object $\boldsymbol{x}$, the classification results obtained by these high-quality information sources based on domain adaptation are given by

$$m_1(\omega_1) = 0.20,\ m_1(\omega_2) = 0.50,\ m_1(\omega_3) = 0.30;\ m_2(\omega_1) = 0.40,\ m_2(\omega_2) = 0.30,\ m_2(\omega_3) = 0.30;$$

$$m_3(\omega_1) = 0.10,\ m_3(\omega_2) = 0.50,\ m_3(\omega_3) = 0.40;\ m_4(\omega_1) = 0.50,\ m_4(\omega_2) = 0.20,\ m_4(\omega_3) = 0.30.$$

It is assumed that the weighting factors learnt by the optimization operation are $\hat{\boldsymbol{\beta}} = [0.7\ 0.8\ 0.9\ 0.7]$. The discounted results using (7) are given by

$$^{\hat{\beta}_1}m_1(\omega_1) = 0.14,\ ^{\hat{\beta}_1}m_1(\omega_2) = 0.35,\ ^{\hat{\beta}_1}m_1(\omega_3) = 0.21, ^{\hat{\beta}_1}m(\Omega) = 0.30;$$

$$^{\hat{\beta}_2}m_2(\omega_1) = 0.32,\ ^{\hat{\beta}_2}m_2(\omega_2) = 0.24,\ ^{\hat{\beta}_2}m_2(\omega_3) = 0.24, ^{\hat{\beta}_2}m(\Omega) = 0.20;$$

$$^{\hat{\beta}_3}m_3(\omega_1) = 0.09,\ ^{\hat{\beta}_3}m_3(\omega_2) = 0.45,\ ^{\hat{\beta}_3}m_3(\omega_3) = 0.36, ^{\hat{\beta}_3}m(\Omega) = 0.10;$$

$$^{\hat{\beta}_4}m_4(\omega_1) = 0.35,\ ^{\hat{\beta}_4}m_4(\omega_2) = 0.14,\ ^{\hat{\beta}_4}m_4(\omega_3) = 0.21, ^{\hat{\beta}_4}m(\Omega) = 0.30.$$

The combination of the four pieces of discounted classification results using (6) are computed by $\boldsymbol{m} = \oplus_{r=1}^{4} {}^{\hat{\beta}_r}\boldsymbol{m}_r = [0.08\ 0.58\ 0.33\ 0.01]$. One can find that the plausibility value $\mathrm{Pl}(\omega_2) = 0.58 + 0.01 = 0.59$ is the biggest, so the unseen object $\boldsymbol{x}$ is committed to class ω_2.

4 Experiment

4.1 Benchmark datasets

The widely used datasets Office+Caltech-10, VLSC and Office-31 were regarded as the benchmarks to evaluate the effectiveness of ECAMI. Four domains are involved in Office+Caltech-10: Amazon (A), Caltech (C), DSLR (D), Webcam (W). VLSC has four different domains, i.e., VOC2007 (V), LabelMe (L), SUN09 (S), Caltech101 (C). Office-31 is a database for object recognition, and consists of three domains, Amazon (A), DSLR (D), Webcam (W). The basic information about them is shown in Table 1.

Table 1 Basic information of the selected benchmark datasets

Dataset	Domain	Feature	Sample	Class
Office+Caltech-10	Amazon (A)	800	958	10
	Caltech (C)	800	1123	10
	DSLR (D)	800	157	10
	Webcam (W)	800	295	10
VLSC	VOC2007 (V)	4096	3376	5
	LabelMe (L)	4096	2656	5
	SUN09 (S)	4096	3282	5
	Caltech101 (C)	4096	1415	5
Office-31	Amazon (A)	800	2715	31
	DSLR (D)	800	482	31
	Webcam (W)	800	776	31

4.2 Domain adaptation techniques and fusion methods

Several state-of-the-art domain adaptation techniques were used for reducing distribution difference between the source and target domains. They are briefly listed and summarized as below.

- Geodesic flow kernel (GFK) [6]: GFK models the domain shift by integrating an infinite number of subspaces that characterize changes in geometric and statistical properties to learn new representations.
- Correlation alignment (CORAL) [39]: It minimizes the domain shift by aligning the second-order statistics of domain distributions.
- Transfer component analysis (TCA) [3]: This method tries to discover a new feature representation across domains by mining the marginal distribution difference.
- Joint distribution adaptation (JDA) [4]: The marginal and conditional distributions are both adopted to acquire a robust new feature representation.
- Transfer joint matching (TJM) [40]: It jointly matches the distributions and re-weights patterns to learn the feature representation which is invariant to both distribution difference and irrelevant patterns.
- Balanced distribution adaptation (BDA) [7]: This method adaptively uses the importance of the marginal and conditional distribution difference at iteration when learning the new representation.
- Deep adaptation network (DAN): The DAN enhances substantially by mean-embedding matching of the multi-layer representations across domains. The deep features are made more transferable, while the domain discrepancy is further reduced via the use of multiple kernel learning.
- Adversarial discriminative domain adaptation (ADDA): It combines the discriminative modeling, untied weight sharing, and a GAN loss to match the distributions.

We tested the performance of ECAMI method using Dempster's rule with learnt weights (i.e., ECAMI-WDS) with respect to related fusion approaches, i.e., MV method, weighted majority vote (WMV) method, average fusion (AF) method, weighted average fusion (WAF) method and DS rule without considering weights (e.g., ECAMI-DS).

4.3 Experiment setting and implementation details

The experiments to verify the effectiveness of ECAMI method are briefly introduced in the sequel.

- K-nearest neighbor (KNN)/support vector machine (SVM)[2]: The standard machine learning methods directly use labeled patterns in the source domain to classify objects in the target domain without any pre-processing.
- GFK/CORAL/TCA/JDA/TJM/BDA: In these experiments, a new feature representation is obtained by adapting distributions, and the KNN classification model is used to classify unseen objects.

2) The standard machine learning techniques are used in the domain adaptation methods, and the selection of base classification model is out of scope of this paper. In previous studies [3, 4, 7, 40], the KNN is often regarded as the base classifier in the domain adaptation task, so we also use it to classify objects in our proposed ECAMI-WDS method. The classification performance of related combination methods by SVM are also tested in the experiments.

• Selected best source domain (SBSD): In this experiment, the source domain which is most consistent to the target domain is used to provide knowledge for building classification model, and the unseen objects in the target domain can be classified by this model.

• TCA/JDA/TJM/BDA+MV/WMV/AF/WAF/DS: In these experiments, the domain adaptation method (i.e., TCA, JDA, TJM, BDA) is first applied to match distributions, and the fusion operation (i.e., MV, WMV, AF, WAF, DS) is employed to combine multiple classification results yielded by different individual source domains. They do not augment the information sources.

• TCA/JDA/TJM/BDA/DAN/ADDA+ECAMI-MV/WMV/AF/WAF/DS: These experiments considering information sources augmentation are used to combine classification results yielded by high-quality information sources using different fusion methods.

We randomly selected one domain as the target domain, and the remainder as the source domains. There will be $4 \times (2^3 - 1) = 28$, $4 \times (2^3 - 1) = 28$ and $3 \times (2^2 - 1) = 9$ cross-domain classification tasks for Office+Caltech-10, VLSC and Office-31 respectively. The tasks are shown in the first column of Tables A1–A7 in Appendix A. The item in the right side of the arrow is the data set in the target domain, and the left represents the singleton source domain data set (e.g., A, V) or the union of several source domain data sets (e.g., AC, VL). The weights of WMV and WAF were estimated by the accuracy in the source domains η_i as $w_i = \frac{\eta_i}{\sum_i \eta_i}$ for testing the excellent classification performance of ECAMI-WDS. The classification performances of the benchmarks based on different domain adaptation techniques are shown in Tables A1–A7. The maximum classification accuracy are marked in bold, and the reserved information sources are characterized by gray background for convenience.

4.4 Experiment results and analyses

From Tables A1–A11, one can see that the classification performance is poor when directly using KNN or SVM because standard machine learning methods cannot work well when the distributions between domains are quite different, and the shallow or deep domain adaptation techniques improve the accuracy with different degree by adopting distributions. The classification accuracy in the target domain via TCA/JDA/TJM/BDA+MV/WMV/AF/WAF/DS lies between the maximum and minimum accuracies based on (existing) multiple individual source domains. The accuracy of the union of several source domain data sets (e.g., CW $\rightarrow$ A) is higher than only using the corresponding singleton source domain data set (e.g., C$\rightarrow$A and W$\rightarrow$A) in many cases. Our experiment results show that the simple merging operation improves the accuracy more or less because all information in multiple individual source domains is included in the union of several source domain data sets. More useful information is obtained after adopting distributions, so the union of several source domain data sets can provide some extra knowledge for classifying objects in the target domain. We also find the accuracy based on ECAMI-MV/WMV/AF/WAF/DS is higher than that of MV/WMV/AF/WAF/DS in general because more useful information is obtained from the augmented information sources. These experiment results validate the interest of information sources augmentation for improving the classification performance. When high-quality information sources are used to yield classification results for combination, the ECAMI-WDS method improves the accuracy a lot compared with other fusion methods. In applications, the classification results usually have different qualities and may be correlated to each other in some degree. The discounting operation can successfully reduce the bad influence of quality and correlation among information sources. In some extreme cases, ECAMI-WDS produces lower accuracy than other fusion methods because the negative influence may not be completely eliminated by only using the classical Shafer's discounting operation. Overall, the ECAMI-WDS method produces higher accuracy than other fusion methods in general. We also tested the classification performance of ECAMI-WDS and related combination methods on data sets Office+Caltech-10 and Office-31 when regarding SVM as the base classifier, and the experiment results are reported in Tables A10 and A11 in Appendix A. One can see that the ECAMI-WDS usually can achieve the highest classification accuracy. The classification performance of ECAMI-WDS and related combination methods varies with the selection of base classifier, but the improvement trend of accuracy does not change. The analysis of the results of Tables A1–A7 shows

Table 2 Time cost (s) of proposed and related combination methods with TCA on Office+Caltech-31

Task	ECAMI-MV	ECAMI-WMV	ECAMI-AF	ECAMI-WAF	ECAMI-DS	ECAMI-WDS
{C, D, W}→A	0.06	0.06	0.05	0.06	0.36	5.09
{A, D, W}→C	0.05	0.05	0.05	0.07	0.22	6.06
{A, C, W}→D	0.07	0.05	0.05	0.05	0.30	1.88
{A, C, D}→W	0.05	0.07	0.05	0.06	0.19	1.01

Table 3 Time cost (s) of proposed and related combination methods with TCA on VLSC

Task	ECAMI-MV	ECAMI-WMV	ECAMI-AF	ECAMI-WAF	ECAMI-DS	ECAMI-WDS
{L, S, C}→V	0.06	0.05	0.05	0.05	0.29	2.09
{V, S, C}→L	0.05	0.06	0.04	0.05	0.20	2.42
{V, L, C}→S	0.06	0.05	0.05	0.05	0.30	2.53
{V, L, S}→C	0.05	0.05	0.04	0.05	0.21	1.82

Table 4 Time cost (s) of proposed and related combination methods with TCA on Office-31

Task	ECAMI-MV	ECAMI-WMV	ECAMI-AF	ECAMI-WAF	ECAMI-DS	ECAMI-WDS
{D, W}→A	0.05	0.07	0.06	0.06	0.34	0.83
{A, W}→D	0.05	0.08	0.08	0.05	0.15	0.60
{A, D}→W	0.05	0.05	0.06	0.05	0.26	0.44

that the proposed method improves the average classification accuracy at least 2.91%, 4.85%, 4.97%, 2.29%, 1.99%, 3.48% for Office+Caltech-10, 4.75% for VLSC and 7.67%, 4.81% for Office-31 compared with different state-of-the-art domain adaptation techniques. In Tables A8 and A9 in Appendix A, one can see that the accuracy is improved a lot when using deep domain adaptation techniques because the CNN can extract more transferable feature across domains. The classification results yielded by DAN or ADDA are more reliable for combination, so the classification accuracy of ECAMI-WDS with deep domain adaptation methods is higher than that with shallow domain adaptation methods. The experiment results demonstrate the ECAMI-WDS method jointly working with shallow or deep domain adaptation methods can effectively improve the classification performance.

4.5 Time cost experiment

In this work, we focus on the combination of classification results yielded by augmented information sources with different shallow or deep domain adaptation methods. The time of distribution alignment operation with different shallow or deep domain adaptation is all included in the proposed and comparative methods, so we only tested the time cost of different combination methods for comparison. We tested the time cost on a laptop with Intel (R) Core (TM) i7-7700HQ CPU@2.80 GHz, 8 G RAM. The experiment results of different data sets with TCA are reported in Tables 2–4. One can see that the time cost of ECAMI-WDS is higher than other related methods. The combination methods (e.g., ECAMI-MV, ECAMI-WMV, ECAMI-AF, ECAMI-WAF) are linear, and the complexity is low. The time cost is related to the number of information sources and the size of data sets. The proposed ECAMI-WDS needs to learn the weighting factors and this operation results in some extra time cost. Thus, the time cost of ECAMI-WDS is higher than the other methods.

5 Conclusion

A new method called ECAMI for solving multi-source domain adaptation issue has been proposed in this paper. In fusion systems, more information sources could generate good combination results, and we regard the union of several source domain data sets as new information sources to provide extra classification results for combination. More extra information can be acquired from the union of source domain data sets to get higher classification accuracy. However, the low-quality information sources will yield unreliable classification results, and lead to bad influence on the fusion result. In practice, high-quality information sources can produce quite reliable classification results, so we select some high-quality information sources for improving accuracy as much as possible. The singleton source domain data set which has the highest domain-consistency to the target domain data set is chosen, and the unions of several source domain data sets involving the most consistent source domain data set are also selected. The reliabilities/weights of the classification results obtained by the auxiliary of selected information sources are usually different because the domain-consistency between domains is diverse, and the information sources may be correlated to each other in some degree. Thus, the classification results must be discounted before fusing by appropriate weighting factors to reduce the influence of domain-consistency and correlation. The optimal weights are computed by an optimization operation based on the patterns in the source domain. The effectiveness of ECAMI method has been clearly demonstrated by comparing its classification performance with respect to several state-of-the-art domain adaptation approaches and some related fusion methods on three benchmark datasets. The experiment results show that ECAMI based on belief functions and information sources augmentation provides higher accuracy than other tested methods. The ECAMI method will be further evaluated with other combination methods, e.g., PCR rules [41]. One interesting and challenging research direction is to consider a totally unseen target domain (without both data and labels), and we will focus on this topic with multiple source domains in our future work.

Acknowledgements This work was partially supported by National Natural Science Foundation of China (Grant Nos. 61672431, 61790552, 61790554, 61701409), Shaanxi Science Fund for Distinguished Young Scholars (Grant No. 2018JC-006), and Fundamental Research Funds for the Central Universities.

Supporting information Tables A1–A11. The supporting information is available online at info.scichina.com and link.springer.com. The supporting materials are published as submitted, without typesetting or editing. The responsibility for scientific accuracy and content remains entirely with the authors.

References

1 Pan S J, Yang Q. A survey on transfer learning. IEEE Trans Knowl Data Eng, 2010, 22: 1345–1359
2 Dai W, Yang Q, Xue G-R, et al. Boosting for transfer learning. In: Proceedings of the International Conference on Machine Learning, 2007. 193–200
3 Pan S J, Tsang I W, Kwok J T, et al. Domain adaptation via transfer component analysis. IEEE Trans Neural Netw, 2011, 22: 199–210
4 Long M, Wang J, Ding G, et al. Transfer feature learning with joint distribution adaptation. In: Proceedings of the IEEE International Conference on Computer Vision, 2013. 2200–2207
5 Fernando B, Habrard A, Sebban M, et al. Unsupervised visual domain adaptation using subspace alignment. In: Proceedings of the IEEE International Conference on Computer Vision, 2014. 2960–2967
6 Gong B, Shi Y, Sha F, et al. Geodesic flow kernel for unsupervised domain adaptation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2012. 2066–2073
7 Wang J, Chen Y, Hao S, et al. Balanced distribution adaptation for transfer learning. In: Proceedings of the IEEE International Conference on Data Mining, 2017. 1129–1134
8 Deng W Y, Lendasse A, Ong Y S, et al. Domain adaption via feature selection on explicit feature map. IEEE Trans Neural Netw Learn Syst, 2019, 30: 1180–1190
9 Wen L, Gao L, Li X. A new deep transfer learning based on sparse auto-encoder for fault diagnosis. IEEE Trans Syst Man Cybern Syst, 2019, 49: 136–144
10 Long M, Cao Y, Cao Z, et al. Transferable representation learning with deep adaptation networks. IEEE Trans Pattern Anal Mach Intell, 2019, 41: 3071–3085
11 Tzeng E, Hoffman J, Saenko K, et al. Adversarial discriminative domain adaptation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017. 7167–7176
12 Sun S, Shi H, Wu Y. A survey of multi-source domain adaptation. Inf Fusion, 2015, 24: 84–92
13 Liu H, Shao M, Fu Y. Structure-preserved multi-source domain adaptation. In: Proceedings of the IEEE International Conference on Data Mining, 2016. 1059–1064
14 Duan L X, Xu D, Tsang I W. Domain adaptation from multiple sources: a domain-dependent regularization approach. IEEE Trans Neural Netw Learn Syst, 2012, 23: 504–518
15 Ding Z, Shao M, Fu Y. Incomplete multisource transfer learning. IEEE Trans Neural Netw Learn Syst, 2018, 29: 310–323
16 Zhu Y, Zhuang F, Zhuang D. Aligning domain-specific distribution and classifier for cross-domain classification from multiple sources. In: Proceedings of the AAAI Conference on Artificial Intelligence, 2019. 5989–5996

17 Xu R, Chen Z, Zuo W, et al. Deep cocktail network: multi-source unsupervised domain adaptation with category shift. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018. 3964–3973

18 Zhao H, Zhang S, Wu G, et al. Multiple source domain adaptation with adversarial learning. In: Proceedings of the International Conference on Learning Representations, 2018. 8559–8570

19 Pizzi N J, Pedrycz W. Aggregating multiple classification results using fuzzy integration and stochastic feature selection. Int J Approximate Reasoning, 2010, 51: 883–894

20 Shafer G. A Mathematical Theory of Evidence. Princeton: Princeton University Press, 1976. 42

21 Liu Z G, Pan Q, Dezert J, et al. Combination of classifiers with optimal weight based on evidential reasoning. IEEE Trans Fuzzy Syst, 2018, 26: 1217–1230

22 Liu Z G, Huang L Q, Zhou K, et al. Combination of transferable classification with multisource domain adaptation based on evidential reasoning. IEEE Trans Neural Netw Learn Syst, 2020. doi: 10.1109/TNNLS.2020.2995862

23 Kobetski M, Sullivan J. Apprenticeship learning: transfer of knowledge via dataset augmentation. In: Proceedings of the Scandinavian Conference on Image Analysis, 2013. 432–443

24 Aftab U, Siddiqui G. Big data augmentation with data warehouse: a survey. In: Proceedings of the IEEE International Conference on Big Data, 2018. 2775–2784

25 Shao L, Zhu F, Li X L. Transfer learning for visual categorization: a survey. IEEE Trans Neural Netw Learn Syst, 2015, 26: 1019–1034

26 Chen Y, Song S, Li S, et al. Domain space transfer extreme learning machine for domain adaptation. IEEE Trans Cybern, 2019, 49: 1909–1922

27 Raina R, Battle A, Lee H, et al. Self-taught learning: transfer learning from unlabeled data. In: Proceedings of the International Conference on Machine learning, 2007. 759–766

28 Xu Y, Pan S J, Xiong H, et al. A unified framework for metric transfer learning. IEEE Trans Knowl Data Eng, 2017, 29: 1158–1171

29 Li S, Song S, Huang G, et al. Cross-domain extreme learning machines for domain adaptation. IEEE Trans Syst Man Cybern Syst, 2019, 49: 1194–1207

30 Borgwardt K M, Gretton A, Rasch M J, et al. Integrating structured biological data by Kernel Maximum Mean Discrepancy. Bioinformatics, 2006, 22: e49–e57

31 Ben-David S, Blitzer J, Crammer K, et al. Analysis of representations for domain adaptation. In: Proceedings of the International Conference on Neural Information Processing Systems, 2007. 137–144

32 Denœux T. Decision-making with belief functions: a review. Int J Approximate Reasoning, 2019, 109: 87–110

33 Denœux T. Logistic regression, neural networks and Dempster-Shafer theory: a new perspective. Knowledge-Based Syst, 2019, 176: 54–67

34 Denœux T. A k-nearest neighbor classification rule based on Dempster-Shafer theory. IEEE Trans Syst Man Cybern, 1995, 25: 804–813

35 Liu Z, Pan Q, Dezert J, et al. Credal c-means clustering method based on belief functions. Knowledge-Based Syst, 2015, 74: 119–132

36 Zhou Z J, Liu T Y, Hu G Y, et al. Fault-alarm-threshold optimization method based on interval evidence reasoning. Sci China Inf Sci, 2019, 62: 089202

37 Si X S, Zhou C H, Hu Z J. Fault prediction model based on evidential reasoning approach. Sci China Inf Sci, 2010, 53: 2032–2046

38 Han S P. A globally convergent method for nonlinear programming. J Opt Theor Appl, 1977, 22: 297–309

39 Sun B, Feng J, Saenko K. Return of frustratingly easy domain adaptation. In: Proceedings of the AAAI Conference on Artificial Intelligence, 2016. 2058–2065

40 Long M, Wang J, Ding G, et al. Transfer joint matching for unsupervised domain adaptation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014. 1410–1417

41 Dezert J, Tchamova A. On the validity of Dempster's fusion rule and its interpretation as a generalization of Bayesian fusion rule. Int J Intell Syst, 2014, 29: 223–252