Heterogeneous Domain Adaptation for Multiple Classes

Joey Tianyi Zhou[†] **Ivor W. Tsang**[‡] **Sinno Jialin Pan**[§] **Mingkui Tan**[†]
[†]Center for Computational Intelligence, Nanyang Technological University, Singapore
[‡]Centre for Quantum Computation & Intelligent Systems, University of Technology, Sydney
[§]Institute for Infocomm Research, Singapore
tzhou1@ntu.edu.sg, ivor.tsang@gmail.com, jspan@i2r.a-star.edu.sg, tanmingkui@gmail.com

Abstract

In this paper, we present an efficient multi-class heterogeneous domain adaptation method, where data from source and target domains are represented by heterogeneous features of different dimensions. Specifically, we propose to reconstruct a sparse feature transformation matrix to map the weight vector of classifiers learned from the source domain to the target domain. We cast this learning task as a compressed sensing problem, where each binary classifier induced from multiple classes can be deemed as a measurement sensor. Based on the compressive sensing theory, the estimation error of the transformation matrix decreases with the increasing number of classifiers. Therefore, to guarantee reconstruction performance, we construct sufficiently many binary classifiers based on the error correcting output coding. Extensive experiments are conducted on both a toy dataset and three real-world datasets to verify the superiority of our proposed method over existing state-of-the-art HDA methods in terms of prediction accuracy.

1 Introduction

In many real-world problems, it is often expensive to collect labeled data for training predictive models. To address this issue, transfer learning or domain adaptation (DA) [25], which aims to adapt a model from an auxiliary domain (i.e., a source domain) to a domain of interest (i.e., a target domain) with little or without additional human supervision, has attracted growing attention in recent years. Towards this goal, a lot of DA methods have been successfully applied to various real-world applications, such as WiFi-based localization [23, 24], image classification [35], video concept detection [14], sentiment analysis [5, 28], coreference resolution [32], vehicle routing [18], game playing [4], etc.

In general, most of the existing domain adaptation methods assume that data of different domains are of the same dimensionality or represented by the same feature space [25, 35]. However, this assumption may not hold for many applications. Taking cross-language document classification as an example, documents in English do not share the same feature representation with those in German due to different vocabularies. Another example comes from image classification, where two images of the same object with different illuminations and resolutions may be of different dimensions of features.

Recently, more and more attention has been shifted to domain adaptation across heterogeneous feature spaces, which is referred to as heterogeneous domain adaptation (HDA) [26, 35]. To address this problem, most existing HDA methods aim to learn a common feature representation such that both source and target domain data can be represented by homogeneous features. Formally, one can learn two feature mappings P and Q to transform the source domain data X_S and target domain data X_T to a new latent feature space such that the difference between the mapped domain data PX_S and QX_T is reduced [29, 26, 31, 15]. Alternatively, one can also learn an asymmetric transformation G to map data from one domain to another so that either the difference between GX_S and X_T can be minimized or the alignment between GX_S and X_T can be maximized [8, 22]. Though these methods have shown promising results, they still suffer from the following three major limitations.

Firstly, since the size of the feature mapping G scales with the product of the dimensions of source and target domains, the computational cost to estimate G is extremely high, especially for high-dimensional source and target domain data. To address the computational issue, Duan *et al.* [15] and Kulis *et al.* [22] proposed kernelized versions of the

Appearing in Proceedings of the 17^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2014, Reykjavik, Iceland. JMLR: W&CP volume 33.

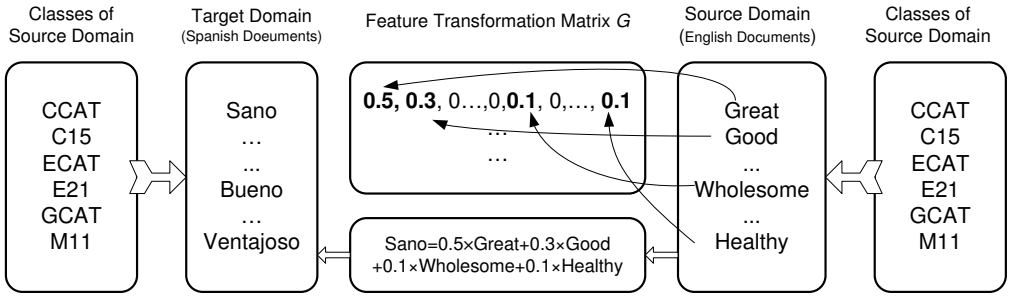


Figure 1: Illustration of a sparse feature representation matrix.

feature-mapping learning methods, respectively. However, these kernelized methods still suffer from high computational cost on large-scale data in terms of data volume.

Secondly, most existing methods tend to recover a dense feature mapping, which, however, is not feasible without enough constraints or information. Besides, in many real-world scenarios, a dense feature mapping is not necessary.

Thirdly, most existing HDA methods simply adopt one-vs-rest strategy to learn multiple binary classifiers independently to address the multi-class issue [8, 31, 15]. In this way, the underlying structure among multiple classes fails to be fully explored. Consequently, the one-vs-rest scheme may limit the ability of knowledge transfer in a multi-class classification manner.

1.1 Encoding Sparsity and Class-Invariance in Learning Feature Mapping

In this paper, we propose to overcome the above limitations for HDA under two assumptions:

1. Sparse feature representation: the feature mapping G between the two domains is highly sparse. In other words, each source domain feature can be represented by only a small subset of the target domain features.

2. Class-invariant transformation: all the classes share the same feature mapping G. Different from [15], we do not aim to learn class-specific feature mappings.

It can be shown that the above two assumptions are satisfied in many real-world HDA problems. Firstly, the sparsity of a feature mapping across domains means that a feature in one domain can be represented by only several features in another domain, which is also known as the feature selection problem. To demonstrate this fact, we use multi-language (e.g., English v.s. Spanish) text classification as a motivating example. Typically, the word "Sano" in Spanish has a similar meaning to the words "Great", "Good", "Wholesome", and "healthy", but not all of the words in English. Therefore, by assuming that the feature mapping across domains is linear, a feature or word in the Spanish domain can be represented by a linear combination of several features or words in the English domain only. As illustrated in Figure 1, the sparse matrix G denotes the feature mapping from the English domain to the Spanish domain. Based on the sparse matrix G, the word "Sano" in Spanish can be represented sparsely by only four words in Spanish as "Sano" $= 0.5 \times$ "Great" $+ 0.3 \times$ "Good" $+ 0.1 \times$ "Wholesome" $+ 0.1 \times$ "Healthy". Such sparsity, which can also facilitate a significant reduction in computational cost on very high-dimensional data, has not been explored in previous HDA methods. Moreover, the feature mapping of the word "Sano" is invariant to different classes. To preserve such *class-invariance*, the feature mapping across domains is learned underlying all the classes, which shares a similar spirit of multi-task feature learning [2].

1.2 Our Contributions

Based on these two assumptions, we propose to learn a sparse and class-invariant feature mapping for multi-class HDA. Specifically, to estimate such a feature mapping, we leverage the weight vectors of the binary classifiers learned in the source and target domains. As will be shown later, this learning task can be cast as a compressed sensing (CS) problem [13, 6]. In summary, the main contributions of this paper are two-folds:

1. We propose a sparse heterogeneous feature representation (SHFR) algorithm to learn a sparse feature transformation for HDA by fully exploring the shared underlying structures among multiple classes between domains.

2. Based on the CS theory, it can be shown that the sparse feature mapping can be learned precisely if and only if a sufficient number of classifiers are provided. Therefore, we further propose to use the ECOC scheme to generate a sufficient number of binary classifiers from a set of classes.

The remainder of this paper is organized as follows. In Section 2, we briefly review some related work. In Section 3,

we firstly cast the proposed learning problem of feature mapping as a compressed sensing problem, and present the details of the proposed SHFR. In Section 4, we conduct a series of experiments on both a toy dataset and three real-world datasets to demonstrate the effectiveness of SHFR. Finally, we conclude this paper and point out some future directions in Section 5.

2 Related Work

In general, approaches to HDA can be classified into two categories. The first group is to learn a pair of feature mappings to transform source and target domain heterogeneous data to a common latent space respectively. For example, Shi *et al.* [29] proposed a Heterogeneous Spectral Mapping (HeMap) method to learn the mappings based on spectral embedding without using any label information. Wang and Mahadevan [31] proposed a manifold alignment method, which is denoted by DAMA in the sequel, to align heterogeneous features in a latent space based on manifold regularization. However, DAMA only works on the data that have strong manifold structures, which limits its transferability on those data where the manifold assumption does not hold. More recently, Duan *et al.* [15] proposed a Heterogeneous Feature Augmentation (HFA) method to augment homogeneous common features that learned by a maximum-margin approach from both the source and target domains. However, the proposed model requires to solve an expensive semidefinite program (SDP) problem.

Another group of HDA algorithms is to learn a feature mapping to transform heterogeneous data from one domain to another domain directly. Specifically, in [21], a method was proposed to learn rotation matrices to match source data distributions to that of the target domain in an unsurprised manner. Dai *et al.* [8] proposed to learn a feature mapping by construct some feature correspondences between domains based on *translators*. However, in general, such translators for feature correspondences are not available or difficult to be constructed in real-world applications. Kulis *et al.* [22] proposed an Asymmetric Regularized Cross-domain transformation (ARC-t) method to learn an asymmetric transformation across domains based on metric learning. Similar to DAMA, ARC-t also utilizes the label information to construct the similarity and dissimilarity constraints between instances from the source and target domains respectively. However, the computational complexities of ARC-t and its kernelized version depend quadratically on the feature dimensions and the data size respectively, which are difficult to be scaled up.

3 Multi-class HDA via Sparse Mapping

In this paper, we study a heterogeneous domain adaptation problem with one source domain and one target domain in a multi-class setting. Specifically, let $\{(\mathbf{x}_{S_i}, y_{S_i})\}_{i=1}^{n_S}$ denote a set of labeled training instances of the source domain, where $\mathbf{x}_{S_i} \in \mathbb{R}^{d_S}$ denotes the i-th instance and $y_{S_i} \in \{1, 2, \cdots, c\}$ denotes the corresponding label. Similarly, let $\{(\mathbf{x}_{T_i}, y_{T_i})\}_{i=1}^{n_T}$ be a set of labeled training instances of the target domain, where $n_T \ll n_S$, $\mathbf{x}_{T_i} \in \mathbb{R}^{d_T}$ and $y_{T_i} \in \{1, 2, \cdots, c\}$. Since there are sufficient labeled data in the source domain, one can build a set of robust predictors $\{\mathbf{w}_S^t\}_{t=1}^{n_c}$ regarding specific binary learning tasks $\{t\}$'s decomposed from the multi-class classification problem. Similarly, one can also build the corresponding predictors $\{\mathbf{w}_T^t\}_{t=1}^{n_c}$ with limited target labeled data. Given a binary task $t \in \{1, 2, \cdots, n_c\}$, we assume that the predictive classifier for either the source or target domain is linear, which can be written as $f^t(\mathbf{x}) = \mathbf{w}^{t\top}\mathbf{x}$, where $\mathbf{w}^t$ is the weight vector of the t-th classifier.

3.1 Problem Formulation

Recall that, in HDA problems, the feature dimensions of the source and target domains are not equal, i.e., $d_S \neq d_T$. To make the learning of heterogeneous domains possible, in ARC-t [22, 27], a transformation matrix $G \in \mathbb{R}^{d_T \times d_S}$ is introduced to learn the similarity $\mathbf{x}_{T_i}^\top G \mathbf{x}_{S_i}$ between a source domain instance $\mathbf{x}_{S_i} \in \mathbb{R}^{d_S}$ and a target domain instance $\mathbf{x}_{T_i} \in \mathbb{R}^{d_T}$. Essentially, data points can be transformed from the source feature space to the target one via $\mathbf{x}_{T_i} = G\mathbf{x}_{S_i}$. Equivalently, we can map the target domain data into the source domain via $G^\top$ [27].

Instead of learning the transformation using metric learning, we borrow an idea from a multi-task learning method [1], and propose to learn the feature mapping across heterogeneous features based on the source and target predictive structures, i.e., $\{\mathbf{w}_S^t\}$'s and $\{\mathbf{w}_T^t\}$'s. Therefore, we can either learn the transformation $G \in \mathbb{R}^{d_T \times d_S}$ by maximizing the dependency between the transformed weight vectors of source classifiers and the weight vectors of target classifiers as follows,

$$\max_G \mathbf{w}_T^{t\top} G \mathbf{w}_S^t,$$

or alternatively by minimizing the distance between the two weight vectors as

$$\min_G \|\mathbf{w}_T^t - G\mathbf{w}_S^t\|.$$

For simplicity in theoretical analysis, we adopt the latter approach to learn G. Specifically, given a binary task $t \in \{1, \cdots, n_c\}$ and a transformation $G \in \mathbb{R}^{d_T \times d_S}$, the relationship between the weight vectors of the source and target classifiers can be modeled as

$$\mathbf{w}_T^t - G\mathbf{w}_S^t = \mathbf{w}_\Delta^t, \tag{1}$$

where $\mathbf{w}_\Delta^t$ is referred to as a "delta" weight vector, and its ℓ_2-norm $\|\mathbf{w}_\Delta^t\|_2$ can be used to measure the difference

between the source weight vector $\mathbf{w}_T^t$ and the transformed target weight vector $G\mathbf{w}_S^t$. Our motivation is that in order to use the robust weight vector $\mathbf{w}_S^t$ to make predictions on the target domain data, one should minimize the difference between domains after transformation. In this sense, we propose to learn the transformation G by minimizing $\|\mathbf{w}_\Delta^t\|_2$. Moreover, as mentioned in Section 1, the feature mapping G should be sparse and class-invariant. Finally, like the multi-language text classification problem mentioned in Figure 1, in most real-world applications, the transformation between two domains should be nonnegative. Therefore, we propose to jointly optimize G over all the binary tasks by imposing non-negative sparsity constraints on G. Specifically, let $G = [\mathbf{g}_1, \mathbf{g}_2, \cdots, \mathbf{g}_{d_T}]^\top$, by imposing ℓ_1-regularization on $\mathbf{g}_i$ and non-negative constraints $\mathbf{g}_i \succeq \mathbf{0}$, the problem of learning G can be formulated as the following nonnegative LASSO problem [33, 30]:

$$\begin{aligned} \min_{G} \quad & \frac{1}{n_c}\sum_{t=1}^{n_c}\|\mathbf{w}_T^t - G\mathbf{w}_S^t\|_2^2 + \sum_{i}^{d_T}\lambda_i\|\mathbf{g}_i\|_1, \quad (2) \\ s.t. \quad & \mathbf{g}_i \succeq \mathbf{0}, \end{aligned}$$

where $\lambda_i > 0$ is the regularization parameter, and the nonnegativity constraints are to preserve nonnegative correlation of the source and target weight vectors [3, 12]. The first term of the objective is to minimize the difference between $\mathbf{w}_T^t$ and $G\mathbf{w}_S^t$ over all the n_c tasks, and the second term is a ℓ_1-regularization term on $\{\mathbf{g}_i\}$'s to enforce sparsity on each row of G respectively. Notice that once the set of source classifiers in terms of the weight vectors $\{\mathbf{w}_S^t\}_{t=1}^{n_c}$ are learned offline, one can learn G directly without reusing the source domain data, which can significantly reduce the learning complexity. Our learning strategy is typically different from most of the existing methods which require the source domain data to be available for learning the feature mapping across domains.

Finally, after learning the sparse transformation G, for any unseen test data $\mathbf{x}_T^*$ from the target domain, we can reuse the source domain classifiers to predict its label by

$$y_T^* = F\left(\{(G\mathbf{w}_S^t)^\top \mathbf{x}_T^*\}_{k=1}^{n_c}\right),$$

where $F(\cdot)$ is a decision function that combines the predictive results of all the n_c source classifiers to make a final prediction.

3.2 Error Bound of Reconstruction

Before presenting a solution to solve the proposed optimization (2), we first analyze the error bound of the reconstruction of G in (2), which will be used to guide our algorithm design. It can be shown that the objective of (2) can be rewritten as the following equivalent form,

$$\min_{\mathbf{g}_i \succeq \mathbf{0}} \frac{1}{n_c}\sum_{t=1}^{n_c}\sum_{i=1}^{d_T}(w_{T_i}^t - \mathbf{w}_S^{t\top}\mathbf{g}_i)^2 + \sum_{i}^{d_T}\lambda_i\|\mathbf{g}_i\|_1,$$

where $w_{T_i}^t$ is the i-th element of the vector $\mathbf{w}_T^t$. If we exchange the summation sequences of the first term, the above formulation can be further rewritten as follows,

$$\min_{\mathbf{g}_i \succeq \mathbf{0}} \sum_{i}^{d_T}\left(\frac{1}{n_c}\|\mathbf{b}_i - \mathbf{D}\mathbf{g}_i\|_2^2 + \lambda_i\|\mathbf{g}_i\|_1\right),$$

where $\mathbf{b}_i$ is the concatenated row vector containing $w_{T_i}^t$ for all the n_c tasks, and $\mathbf{D} = [\mathbf{w}_1^t\ \mathbf{w}_2^t\ \cdots\ \mathbf{w}_{n_c}^t]^\top \in \mathbb{R}^{n_c \times d_S}$. Note that (3) contains d_T nonnegative LASSO problems. In general, we have $n_c < d_S$ for relatively high-dimensional problems. Therefore, it is an underdetermined linear system [13]. However, if $\mathbf{g}_i$ is sparse, as suggested by the compressive sensing theory, it can be possibly recovered if the measurements are sufficient and the matrix $\mathbf{D}$ satisfies some restricted conditions [13, 6, 34].

For convenience in presentation, let k_i denote the sparsity of $\mathbf{g}_i$ and $\hat{\mathbf{g}}_i$ be the estimator of $\mathbf{g}_i$. According to Theorem 4 in [34], under some restricted conditions, the estimation error $\|\mathbf{g}_i - \hat{\mathbf{g}}_i\|_2$ for each independent subproblem can be bounded by $O(\sqrt{\frac{k_i \log d_S}{n_c}})$, where d_S denotes the dimensions of the source domain, and n_c is the number of tasks. Therefore, the estimation error of G in (2) is bounded by $O(d_T\sqrt{\frac{k_i \log d_S}{n_c}})$. According to this bound, we can observe that if more binary classification tasks are generated (i.e., n_c is large), and the sparsity of each row vector $\mathbf{g}_i$ is high (i.e., k_i is small), the estimation error bound will become relatively small. Note that, the above bound holds under some restricted conditions, such as the sparse Riesz condition or RIP condition [34, 6]. More specifically, the sparse Riesz condition requires that any two columns of $\mathbf{D}$ should be as perfectly incoherent as possible [13, 6].[1] In the following sections, we will concentrate on building an incoherent $\mathbf{D}$ for the transformation learning.

3.3 Many Binary Classification Tasks Construction

As discussed in Section 3.2, to reduce the reconstruction error, we need to 1) build as more classifiers as possible to increase the number of measurements, and 2) construct incoherent classifiers to generate an incoherent $\mathbf{D}$. For multi-class classification problems with a set of classes $\{1, 2, \cdots, c\}$, one can generate c binary classifiers using the one-vs-all strategy [11]. However, this strategy fails to generate sufficient number of classifiers if the number of classes is relatively small. Alternatively, one can also use the one-vs-one strategy to generate $c(c-1)/2$ binary classification tasks. However, the generated classifiers may have large redundancy, e.g. some classifiers are highly correlated with others.

To address the above issues, we propose to use the Error Correcting Output Codes (ECOC) scheme to generate

[1] More details of the exact recovery conditions can be found in the cited reference papers.

sufficient binary classifiers [11]. Basically, ECOC aims to construct correspondences between classes and codewords by designing a "code book" generate a number of binary classifiers for multi-class classification problems. Generally speaking, ECOC consist of two steps: encoding and decoding. In the encoding step, one can construct a both row-well-separated and column-well-separated "code book" [10, 16]. In other words, the classifiers generated under the ECOC scheme are incoherent. After the "code book" is constructed, one can use some decoding techniques to assign labels to the corresponding codes for predictions. In this paper, we use the sparse random design coding and loss-based decoding techniques [11].

3.4 Robust Transformation Learning with ECOC

The robustness of ECOC is another important motivation to use ECOC in our method. Recall that, the error-correcting codes can be viewed as a compact form of voting, and a certain number of incorrect votes can be corrected through the corrected votes [11]. Specifically, given a total of T classifiers, the voting-based methods guarantee to make a correct decision as long as there is $\lfloor \frac{T}{2} + 1 \rfloor$ correct classifiers [9], where $\lfloor\ \rfloor$ denotes the round-up operator. In other words, even though there are part of misclassified tasks due to the incorrect base classifiers, we can still achieve good performances by using the ECOC scheme. This property is particularly important for the proposed multi-class HDA method since some of the learned binary classifiers in the target domain may not be accurate or correct, which may be due to the poor learning, limited training data or bad feature representations. Fortunately, it has been shown that though the bit errors are unavoidable in real applications, using the error-correcting codes can still make correct decisions with enough correct learners [20]. Accordingly, with the help of ECOC, SHFR can still learn a robust transformation matrix G even with some incorrect target binary classifiers. The robustness of SHFR with ECOC will be further verified in experiments, where we demonstrate that even with inaccurate target classifiers, the obtained class-invariant G based on ECOC can greatly enhance the prediction accuracy for each binary classification task.

3.5 Complexity Comparison

In this paper, we use linear SVMs [17] to build the base classifiers for the source and target domains, which can be pre-trained offline. In other words, the learning of G is independent to the number of source domain training instances. It can be shown that the computational cost of our method SHFR is $O(d_T n_c d_S)$, which is cost by solving the nonnegative LASSO problem 2. Compared to state-of-the-art HDA algorithms, our proposed method is much more efficient. The ARC-t method requires to solve an optimization problem that contains $n_S n_T$ constraints by applying an alternating projection method (i.e., Bregman's algorithm [7]). The HFA method adopts an alternating projection method as well to solve a SDP problem, where the transformation matrix to be learned is in $\mathbb{R}^{(n_S+n_T)\times(n_S+n_T)}$, resulting in time complexity bounded by $O(n_S + n_T)^3$. Therefore, ARC-t and HFA perform inefficiently when the data size is large. Differently, DAMA first constructs a series of combinatorial Laplacian matrices in $\mathbb{R}^{(d_S+d_T)\times(d_S+d_T)}$ and then solve a generalized eigenvalue decomposition problem of time complexity bounded by $O(d_S + d_T)^3$. Therefore, DAMA is very computational expensive when the data dimensionality is high.

4 Experiments

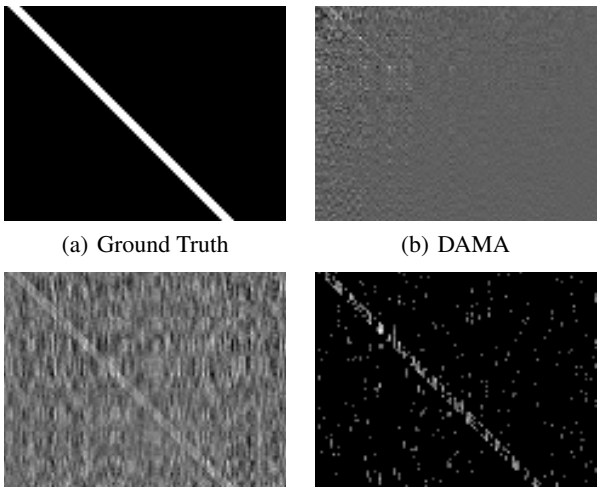
In experiments, we use linear SVMs as base classifiers whose regularization parameter C is set to 1 for all comparison methods. Furthermore, for ARC-t and HFA, which require the use of kernel functions to measure data similarities, we use the RBF kernel for learning the transformation. For parameter tuning, cross-validation is not applicable in HDA problems due to the small size of labeled data in the target domain, which is still an open research issue in HDA. Therefore, we tune parameters of the comparison methods on a predefined range and report their best results, respectively. For SHFR, we generate the ECOC "code-book" matrix using sparse random matrix [11].

4.1 Experiments on Toy Dataset

In this section, we first compare the performance of different HDA methods in terms of recovering a ground-truth feature mapping G on a 20-class toy dataset. To generate the toy dataset, we first randomly generate 150 instances of 150 features for each class from different Gaussian distributions to form a source domain $X_S \in \mathbb{R}^{150\times 3{,}000}$. After that, we construct the ground-truth sparse feature mapping $G \in \mathbb{R}^{100\times 150}$ by using the following method: for each row i, we set $G_{ij} = 1/5$, where $j = i, i+1, \cdots, i+5$, and $G_{ij} = 0$ otherwise. This generation of G implies that each target domain feature is represented by 5 source domain features. The ground-truth feature mapping is displayed in Figure 2(a), where the dark area represents the zero entries and the bright area denote nonzero values of G. Finally, we construct the target domain data $X_T \in \mathbb{R}^{100\times 3{,}000}$ by using $X_T = GX_S$. When conducting the experiment, we randomly select 5 instances per class from the target domain data X_T as labeled training data, and apply different HDA methods on them together with all the 3,000 source domain labeled data to recover the feature mapping G.

In this experiment, the HDA methods DAMA and ARC-t are adopted as the baselines. For ease in comparison, we present the recovered matrix G for the three methods in Figures 2(b)-2(d), respectively. From Figure 2(b), we can observe that DAMA fails to recover the structure of G; while ARC-t shows relatively better performance. Howev-

er, the recovered G's by these two methods are not sparse. On the contrary, according to Figure 2(d), the proposed method SHFR can perfectly recover G with little noise. These experimental results demonstrate that by considering the sparsity constraints, the proposed method SHFR can recover the feature mapping G more precisely.



(a) Ground Truth (b) DAMA

(c) ARC-t (d) SHFR

Figure 2: Illustrations of recovered feature mappings by different methods on the toy dataset. Pixels in black color represent 0, and those in white color represent 1.

4.2 Experiments on Real-world Datasets

In this section, we conduct experiments on three real-world datasets: Multilingual Reuters Collection, BBC Collection, and Cross-lingual Sentiment Dataset, to verify the effectiveness and efficiency of SHFR. The reported results are averaged over 10 independent data-split procedures.

4.2.1 Datasets and Experimental Settings

Multilingual Reuters Collection[2] is a text dataset with over 11,000 news articles from 6 categories in 5 languages (i.e., English, French, German, Italian and Spanish), which are represented by a bag-of-words weighted by TF-IDF. Following the setting in [15], we use Spanish as the target domain and the other four languages as source domains. For each class, we randomly select 100 instances from the source domain and 10 instances from the target domain for training. Furthermore, we randomly select 10,000 instances from the target domain as the test data. Note that the original data is in very high dimensions, and the baseline methods cannot handle such high-dimensional features. To conduct the comparison, we perform PCA with 60% energy preserved on the TF-IDF features. After PCA, we obtain 1,131 for English documents, 1,230 features for French documents, 1,417 features for German documents, 1,041 features for Italian documents, and 807 features for Spanish documents. In contrast to the baseline methods, we use original features for our proposed method SHFR since it can efficiently handle high-dimensional data.

BBC Collection[3] was collected for multi-view learning where each instance is represented by three views. These views were constructed from a single-view BBC corpora by splitting news article into related "views" of text. We consider **View 3** as the target domain, and **View 1** and **View 2** as source domains respectively. Similar to the pre-processing on the Reuters dataset, we perform PCA on the original data to reduce dimensions for other baselines. Consequently, the reduced dimensions for **View 1**, **View 2** and **View 3** are 203, 205 and 418, respectively. We randomly select 70% source domain instances, and 10 target domain instances for each class for training. The remaining target domain instances are used for testing.

Cross-lingual Sentiment Dataset[4] consists of Amazon product reviews of three product categories: books, DVDs and music. These reviews are written in four languages: English, German, French, and Japanese. We treat English reviews as the source domain data and the other language reviews as the target domain data respectively. After PCA, the reviews are of 715, 929, 964, 874 features for English, German, French and Japanese, respectively. We randomly select 1,500 source domain instances and 10 target domain instances per class for training, and use the remaining 5,970 target domain instances for testing.

4.2.2 Overall Performance Comparison

Comparison results between SHFR and other baselines on the three real-world datasets are reported in Tables 1-3 respectively. From the tables, we can observe that the SVMs conducted on a small number of target domain data only, denoted by SVM-T, using either one-vs-one or one-vs-all strategy performs the worst on average. Moreover, the results of SVM-T using one-vs-one and one-vs-all, respectively, are not consistent. For instance, on the BCC dataset in Table 2, SVM-T using the one-vs-all strategy performs much better than that using the one-vs-one strategy, while on the sentiment dataset in Table 3, SVM-T using the one-vs-one strategy performs much better than that using one-vs-all strategy. The reason is that the size of labeled training data is too limited to train a precise and stable classifier in the target domain. The testing accuracy of the HDA baseline methods, DAMA, ARC-t and HFA, are comparable on the three datasets except for the BCC dataset. On the BCC dataset, DAMA performs much worse than the other two. This may be because the performance of DAMA

[2] http://multilingreuters.iit.nrc.ca/ReutersMultiLingualMultiView.htm

[3] http://mlg.ucd.ie/datasets/segment.html

[4] http://www.uni-weimar.de/cms/medien/webis/research/corpora/corpus-webis-cls-10.html

Table 1: Multilingual Reuters Collection: comparison results in terms of classification accuracy (%).

Source Domain	SVM-T(1vsR)	SVM-T(1vs1)	DAMA	ARC-t	HFA	SHFR(1vs1)	SHFR(ECOC)
English	64.40±4.45	64.94±8.10	63.42±2.62	65.56±2.48	66.78±1.56	69.45±1.56	**72.79±1.10**
French			64.32±1.86	65.30±1.83	67.09±1.63	70.51±1.32	**73.82±1.12**
German			66.56±2.34	67.45±1.65	68.42±1.74	71.23±1.76	**74.15±1.14**
Italian			67.48±2.17	66.41±2.18	68.19±2.51	70.75±1.54	**73.35±1.31**

Table 2: BBC Collection: comparison results in terms of classification accuracy (%).

Source Domain	SVM-T(1vsR)	SVM-T(1vs1)	DAMA	ARC-t	HFA	SHFR(1vs1)	SHFR(ECOC)
View 1	73.35±4.98	66.69±12.61	67.42±2.25	75.93±2.54	71.72±11.11	89.81±1.20	**90.45±1.00**
View 2			66.35±1.76	74.23±2.12	72.24±8.14	88.56±1.02	**91.82±0.84**

Table 3: Cross-lingual Sentiment Dataset: comparison results in terms of classification acc. (%).

Target Domain	SVM-T(1vsR)	SVM-T(1vs1)	DAMA	ARC-t	HFA	SHFR(1vs1)	SHFR(ECOC)
French	47.23±3.89	58.14±4.44	52.12±3.67	50.01±5.2	55.16±3.71	60.12±3.56	**62.09±2.15**
German	48.54±4.95	60.06±5.28	54.51±3.64	55.30±2.83	54.84±3.63	63.20±2.74	**65.22±2.03**
Japanese	47.10±6.21	54.87±4.81	52.12±2.45	54.45±3.65	53.42±4.74	59.40±3.71	**62.05±3.16**

is sensitive to the intrinsic manifold structure of the data. If the manifold assumption does not hold on the data, the testing accuracy of DAMA drops a lot. On the contrary, our proposed method SHFR using either the one-vs-one or ECOC scheme performs the best on these three datasets. Moreover, using the ECOC scheme, SHFR can further improve the performance in terms of classification accuracy. As discussed in Section 3.2, this is because that with more constructed binary tasks, the recovered feature mapping G tends to be more accurate.

4.3 Impact of the Number of Binary Classifiers on Estimation Error

As discussed in Section 3.2, estimation error of G depends on two factors: the number of classifiers (measurements) and the sparsity of G. When G is sparse and the constructed binary classifiers are sufficient, one can possibly recover a precise G by using the dictionary constructed by $\mathbf{w}_S$.

To analyze the error estimation bound w.r.t. the number of binary classifiers of our proposed SHFR, we conduct an experiment on the Reuters dataset. Experimental results are showed in Figure 3(a). From the figure, we can observe that the more binary classifiers are constructed, the higher accuracy of the predictions can be achieved in the target domain. This verifies that more classifiers can provide more discriminative information to recover the feature mapping G. Furthermore, the standard deviation of testing accuracy is also decreasing with the increasing number of classifiers.

In general, SHFR can obtain better and more stable performance in terms of classification accuracy with increasing number of binary classifiers. However, as observed from the figure, the multi-class accuracy does not further increase any more when the number of classifiers reaches 31. This may be caused by two reasons: 1) The redundancy among the increasingly constructed binary classifiers may hinder the estimation of G from being more accurate. 2) According to [19], when there are too many binary classifiers, the minimum distances in ECOC becomes small, which may decrease the ability of correcting errors.

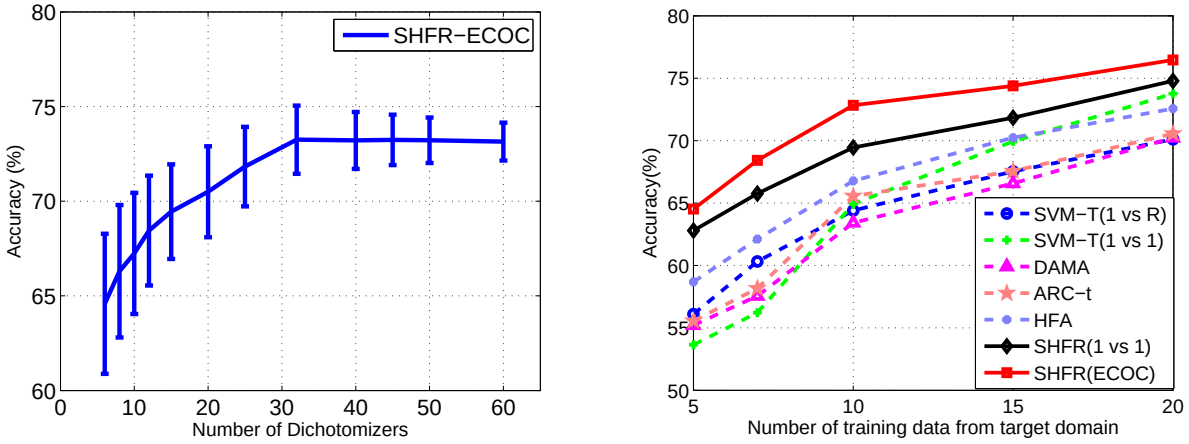
4.4 Impact of Target Domain Training Size

In this experiment, we verify the impact of the labeled training sample size of the target domain to the overall H-DA performance in terms of classification accuracy. We vary the number of target domain training instances from 5 to 20. Here, we only report the results of the Reuter dataset, where we use English as the source domain and Spanish as the target domain. The experimental results are reported in Figure 3(b). From the figure, we can observe that SHFR consistently outperforms the baseline methods under different numbers of labeled training instance in the target domain. Particularly, when the size of the target domain labeled data is smaller than 10, SHFR shows significantly better performance than the baseline methods.

4.5 Error Corrections Through Learning G

The weight vectors of the binary classifiers constructed in the target domain (i.e., $\mathbf{w}_T$'s) may be very unreliable due to the lack of target labeled data, which may affect the estimation of G. To verify how SHFR can correct bias committed by some binary classifiers, we conduct experiments to show the comparison results between SVM-T and SHFR in term of classification accuracy on each binary task on the Reuters dataset in Table 4, where each column corresponds to a binary task, indexed by $k \in \{1, \cdots, 15\}$.

From the table, we can observe that the predictions of SVM-T on some binary tasks are inaccurate due to limited labeled data, whose accuracies are even below 50% (num-



(a) Accuracy v.s. #binary classifiers.

(b) Accuracy v.s. target training sample size.

Figure 3: Comparison results of SHFR in different settings on the Reuters dataset.

Table 4: Comparison results of binary Classifiers in terms of classification accuracy (%).

Binary Classifiers	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
SVM-T	74.42	67.98	85.67	82.53	**43.32**	75.95	66.30	73.92	**42.78**	76.56	78.32	65.43	72.74	**46.56**	**49.71**
SHFR	65.39	84.81	82.45	94.28	80.76	76.75	79.74	84.63	84.45	70.07	90.35	81.35	83.97	80.75	86.43
Difference	-9.03	16.82	-3.22	11.75	37.44	0.80	13.44	10.70	41.67	-6.49	12.03	15.92	11.23	34.20	36.72

bers in boldface). However, through learning the transformation G, we are able to reduce the bias of the weak binary classifiers, and increase over 30% in accuracy on average. Other experiments on the other two datasets also exhibit similar results, e.g., for each binary task, the performance of classifiers obtained through G is increased by 26.53% on the BBC dataset and 3.01% on the Sentiment dataset on average, respectively, compared to those based on SVM-T.

This experiment also verifies why SHFR outperforms ARC-t in the experiments shown in Tables 1-3 and Figure 3(b). ARC-t aims to align the target data with source data via a transformation G which is learned from some similarity and dissimilarity constraints. These constraints are constructed from plenty of source-domain labeled data and only a few target-domain labeled data. In other words, if those limited target-domain labeled data are noisy, the estimation in ARC-t may fail. Different from ARC-t, SHFR tries to align the target classifiers with source classifiers through a transformation G, which is shared by all induced binary tasks or classifiers. In other words, G is estimated through all classifiers. Inspired by multi-task feature learning which jointly optimizes all classifiers to learn a robust global feature transformation G, SHFR can still estimate a stable and precise G for HDA. Furthermore, as discussed in Section 3.4, ECOC with a proper design of a coding matrix has the ability to rectify errors committed by individual binary classifiers as shown in Table 4. In summary, SHFR takes an advantages of multi-class information and reduces the bias caused by the limited training instances in the target domain, resulting in robust and better prediction performance in multi-class HDA.

5 Conclusion and Future Work

In this paper, by exploiting the sparsity and class-invariance in learning a feature mapping, we propose a novel method, namely SHFR, for multi-class HDA problems. In the proposed method, the learning of feature mapping can be cast as a compressive sensing (CS) problem. Based on the CS theory, we show that how the number of constructed binary learning tasks can affect the multi-class HDA performance. In addition, by exploring the sparsity, the proposed method has superior scalability over other methods. Extensive experiments demonstrate the effectiveness, efficiency and stability of SHFR in multi-class HDA. In future work, we plan to study more theoretical analysis of SHFR.

Acknowledgment

This research is partially supported by Multi-plAtform Game Innovation Centre (MAGIC), funded by the Singapore National Research Foundation under its IDM Futures Funding Initiative and administered by the Interactive & Digital Media Programme Office, Media Development Authority, and DSO grant DSOCL10021.

References

[1] R. K. Ando and T. Zhang. A framework for learning predictive structures from multiple tasks and unlabeled data. *J. Mach. Learn. Res.*, 6:1817–1853, 2005.

[2] A. Argyriou, T. Evgeniou, and M. Pontil. Multi-task feature learning. In *NIPS*, pages 41–048. 2007.

[3] Y. Aytar and A. Zisserman. Tabula rasa: Model transfer for object category detection. In *ICCV*, pages 2252–2259, 2011.

[4] B. Banerjee and P. Stone. General game learning using knowledge transfer. In *IJCAI*, 2007.

[5] J. Blitzer, R. McDonald, and F. Pereira. Domain adaptation with structural correspondence learning. In *EMNLP*, pages 120–128, 2006.

[6] E. J. Candès, J. K. Romberg, and T. Tao. Stable signal recovery from incomplete and inaccurate measurements. *Comm. Pure Appl. Math.*, 59(8):1207–1223, August 2006.

[7] Y. Censor and S. A. Zenios. *Parallel Optimization: Theory, Algorithms and Applications*. Oxford University Press, 1997.

[8] W. Dai, Y. Chen, G.-R. Xue, Q. Yang, and Y. Yu. Translated learning: Transfer learning across different feature spaces. In *NIPS*, pages 353–360, 2008.

[9] T. G. Dietterich. Ensemble methods in machine learning. In *MCS*, pages 1–15. Springer-Verlag, 2000.

[10] T. G. Dietterich and G. Bakiri. Error-correcting output codes: A general method for improving multiclass inductive learning programs. In *AAAI*, pages 572–577, 1991.

[11] T. G. Dietterich and G. Bakiri. Solving multiclass learning problems via error-correcting output codes. *J. Artif. Intell. Res.*, 2:263–286, 1995.

[12] J. Donahue, J. Hoffman, E. Rodner, K. Saenko, and T. Darrell. Semi-supervised domain adaptation with instance constraints. In *CVPR*, pages 668–675, 2013.

[13] D. L. Donoho. Compressed sensing. *IEEE Trans. Inform. Theory*, 52:1289–1306, 2006.

[14] L. Duan, I. W. Tsang, D. Xu, and T.-S. Chua. Domain adaptation from multiple sources via auxiliary classifiers. In *ICML*, 2009.

[15] L. Duan, D. Xu, and I. W. Tsang. Learning with augmented features for heterogeneous domain adaptation. In *ICML*, 2012.

[16] S. Escalera, P. Radeva, and O. Pujol. Forest extension of error correcting output codes and boosted landmarks. In *ICPR*, pages 104–107, 2006.

[17] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin. LIBLINEAR: A library for large linear classification. *J. Mach. Learn. Res.*, 9:1871–1874, 2008.

[18] L. Feng, Y.-S. Ong, I. W. Tsang, and A.-H. Tan. An evolutionary search paradigm that learns with past experiences. In *WCCI*, 2012.

[19] N. García-Pedrajas and D. Ortiz-Boyer. An empirical study of binary classifier fusion methods for multiclass classification. *Inf. Fusion*, 12(2):111–130, April 2011.

[20] R. Ghani. Using error-correcting codes for text classification. In *ICML*, pages 303–310. Morgan Kaufmann Publishers, 2000.

[21] M. Harel and S. Mannor. Learning from multiple outlooks. In *ICML*, pages 401–408, 2011.

[22] B. Kulis, K. Saenko, and T. Darrell. What you saw is not what you get: Domain adaptation using asymmetric kernel transforms. In *CVPR*, pages 1785–1792, 2011.

[23] S. J. Pan, D. Shen, Q. Yang, and J. T. Kwok. Transferring localization models across space. In *AAAI*, pages 1383–1388, 2008.

[24] S. J. Pan, I. W. Tsang, J. T. Kwok, and Q. Yang. Domain adaptation via transfer component analysis. *IEEE Trans. Neural Netw.*, 22(2):199–210, 2011.

[25] S. J. Pan and Q. Yang. A survey on transfer learning. *IEEE Trans. Knowl. Data Eng.*, 22(10):1345–1359, October 2010.

[26] P. Prettenhofer and B. Stein. Cross-language text classification using structural correspondence learning. In *ACL*, pages 1118–1127, 2010.

[27] K. Saenko, B. Kulis, M. Fritz, and T. Darrell. Adapting visual category models to new domains. In *ECCV*, pages 213–226. Springer-Verlag, 2010.

[28] C.-W. Seah, I. W. Tsang, and Y.-S. Ong. Transfer ordinal label learning. *IEEE Trans. Neural Netw. Learn. Syst.*, 24(11):1863–1876, Nov 2013.

[29] X. Shi, Q. Liu, W. Fan, P. S. Yu, and R. Zhu. Transfer learning on heterogenous feature spaces via spectral transformation. In *ICDM*, pages 1049–1054, 2010.

[30] M. Slawski and M. Hein. Sparse recovery by thresholded non-negative least squares. In *NIPS*, pages 1926–1934, 2011.

[31] C. Wang and S. Mahadevan. Heterogeneous domain adaptation using manifold alignment. In *IJCAI*, pages 1541–1546, 2011.

[32] J.-B. Yang, Q.-L. Xiang Q. Mao, K. M. A. Chai I. W. Tsang, and H. L. Chieu. Domain adaptation for coreference resolution: An adaptive ensemble approach. In *EMNLP-CoNLL*, 2012.

[33] M. Yuan and Y. Lin. On the non-negative garrotte estimator. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(2):143–161, 2007.

[34] C.-H. Zhang and J. Huang. The sparsity and bias of the LASSO selection in high-dimensional linear regression. *Ann. Statist.*, 36(4):1567–1594, 2008.

[35] Y. Zhu, Y. Chen, Z. Lu, S. J. Pan, G.-R. Xue, Y. Yu, and Q. Yang. Heterogeneous transfer learning for image classification. In *AAAI*, 2011.