New approaches for heterogeneous transfer learning

Zhou, Tianyi

2015

Zhou, T. (2015). New approaches for heterogeneous transfer learning. Doctoral thesis, Nanyang Technological University, Singapore.

<https://hdl.handle.net/10356/65532>

<https://doi.org/10.32657/10356/65532>

NEW APPROACHES FOR HETEROGENEOUS TRANSFER LEARNING

ZHOU TIANYI

School of Computer Engineering

A thesis submitted to the Nanyang Technological University
in partial fulfilment of the requirement for the degree of
Doctor of Philosophy

2015

Abstract

In many real-world problems, it is often time-consuming and expensive to collect labeled data. To alleviate this challenge, transfer learning (TL) techniques that adapt a model from a related task with ample labeled data to a task of interest with little or no additional human supervision have been proposed in recent years. Most TL methods assume that the data come from different domains having the same feature space and dimensionality. However, the assumptions may also be violated in some real world applications such as text-based image classification, cross-language document classification, and cross system recommendation. To handle situations when the assumptions do not hold, new TL approaches that utilize heterogeneous feature spaces are needed to solve the heterogeneous transfer learning (HTL) problem.

In this thesis, three novel HTL approaches are proposed to handle issues related to different settings:

- (i) **Sparse Heterogeneous Feature Representation (SHFR)** is developed to address the sparsity issue as well as the HTL multi-class classification problem. Formulating the feature transformation matrix learning problem as a compressed sensing problem, we propose a solution to learn a sparse feature transformation matrix to map each feature from one domain to another one. Moreover, a generalized HTL error bound is derived under the multi-class setting. Specifically, in order to guarantee the reconstruction performance and enhance the prediction accuracy, we can construct a sufficient number of binary classifiers based on the error correcting output correcting (ECOC) scheme. To further speed up the estimation of the transformation matrix, we also present a very efficient batch-mode algorithm to solve the corresponding non-negative sparse recovery problem.

- (ii) **Hybrid Heterogeneous Transfer Learning (HHTL)** is proposed to allow the corresponding instances across domains to be biased in either the source or target domain. In other words, HHTL works well even when the cross-domain correspondences are not identically and independently drawn from two domains. Our solution utilizes a deep learning approach to learn a robust feature mapping between cross domain heterogeneous features as well as a better feature representation for mapped data.
- (iii) **Heterogeneous Transfer Learning through Active Correspondences construction (HTLA)** is proposed to enable realistic correspondence construction and better distribution matching. HTLA is based on matrix completion with a regularization term on distribution matching. For the cross-language document classification task, the construction of correspondences between documents of different languages is very time consuming for precise translation. However, existing HTL approaches assume these correspondences are given. To reduce the budget of constructing correspondences, we extend our proposed method to actively construct correspondences between domains.

Extensive experimental results on both synthetic and real-world classification tasks are presented to verify the effectiveness of our proposed three approaches compared with the state-of-the-art baselines.

Acknowledgments

First and foremost, I would like to express my deepest gratitude to my academic supervisor, **Dr. Ho Shen-shyang** and my previous supervisor **Dr. Ivor Wai-Hung Tsang**, for patiently giving guidance and encouragement through my Ph.D research. Without their consistent support and supervision, this research would not have been possible.

This thesis would not be finished without the contribution from my consistent collaborator **Dr. Sinno Jialin Pan**. With his constructive work and motivations, this thesis becomes more comprehensive and completed. I take this opportunity to acknowledge him for helping me to make this thesis possible.

I would also like to thank all my friends and staffs in Center for Computational Intelligence Lab (C2I) for their technical supports and enthusiastic help. Many thanks are given to the members of the research group I am working in, and my roommates I am living with.

Special thanks to my parents for their constant support, understanding and love.

Contents

Abstract	i
Acknowledgments	iii
List of Figures	viii
List of Tables	1
1 Introduction	2
1.1 Weak Supervision Challenges	3
1.2 Transfer Learning	4
1.3 Heterogeneous Transfer Learning	5
1.4 Research Issues on Heterogeneous Transfer Learning	9
1.4.1 Sparsity	9
1.4.2 Multi-class Information	10
1.4.3 Bias Correspondences	11
1.4.4 Distribution Matching	14
1.4.5 Correspondences Construction	14
1.5 Contributions	16
1.6 Thesis Outline	17
2 Literature Review	20
2.1 Notations and Definitions	21
2.2 Label-based HTL	22
2.3 Cross-domain correspondences based HTL	27

3	Sparse Heterogeneous Feature Representation For Multiple Classes	33
3.1	Exploiting Sparsity and Class-Invariance in Feature Mapping	35
3.2	Sparse Heterogeneous Feature Representation	37
3.2.1	Feature Mapping Construction	37
3.2.2	Proposed Formulation	38
3.3	Generalization Error Bound for Multi-class HTL	40
3.3.1	Estimation Error of Sparse Feature Mapping	40
3.3.2	Generalization Error Bound for Binary Class HTL	42
3.3.3	Generalization Error Analysis Based on One-versus-all Scheme . .	45
3.3.4	Generalization Error Analysis Based on ECOC	46
3.3.5	Robust Transformation Learning with ECOC	52
3.4	An Efficient Batch Decoding Algorithm for SHFR	52
3.4.1	Batch-Mode GMP (BGMP)	53
3.4.2	Complexity Comparison	54
3.5	Experimental Results	55
3.5.1	Experiments on Synthetic Datasets	56
3.5.2	Experiments on Real-world Datasets	56
3.5.3	Impact of Training Sample Size of Target Domain	60
3.5.4	Comparisons on the same dimensional features	61
3.5.5	Incoherent Dictionary D construction	62
3.5.6	Error Correction Through Learning G	62
3.5.7	Impact of Dichotomizer Size on Estimation Error	63
3.5.8	Impacts of Nonnegative Constraints and ℓ_1 Norm Regularization .	64
3.5.9	Experiments on Removing Negative Alignment	66
3.5.10	Training Time Comparison	68
3.6	Summary	68
4	Hybrid Heterogeneous Transfer Learning through Deep Learning	70
4.1	Related Work	72
4.2	Problem Statement and Preliminaries	74
4.3	A Deep Learning Approach for Hybrid Heterogeneous Transfer Learning	75

4.3.1	High-Level Homogeneous Feature Learning	76
4.3.2	Heterogeneous Feature Mapping	78
4.3.3	Prediction by Stacking Layers	79
4.4	Experimental Results	79
4.4.1	Experimental Setting	79
4.4.2	Performance Comparison	82
4.4.3	Impact of Layers in the HHTL Structure	85
4.4.4	Parameter Sensitivity Study	85
4.5	Summary	86
5	Heterogeneous Transfer Learning through Active Correspondences	
	Construction	88
5.1	Related Work	90
5.2	Problem Statement and Preliminaries	91
5.2.1	Matrix Completion	92
5.2.2	Maximum Mean Discrepancy	93
5.3	Active Correspondences Construction for HTL	93
5.3.1	Distribution Matching based Matrix Completion	94
5.3.2	Active Correspondences Construction	98
5.4	Experimental Results	100
5.4.1	Experiment Setting	100
5.4.2	Performance Comparison	102
5.4.3	Effectiveness of Active Learning Strategy	102
5.4.4	Parameter Sensitivity Study	103
5.5	Summary	104
6	Conclusions and Future Work	105
6.1	Summary of Contributions	105
6.1.1	Multi-class Heterogeneous Transfer Learning	105
6.1.2	Hybrid Heterogeneous Transfer Learning	106
6.1.3	Heterogeneous Transfer Learning through Active Correspondences Construction	107

6.2	Future Work	108
6.2.1	More Discriminative Output Codes for Multi-class HTL	108
6.2.2	Unsupervised Heterogeneous Transfer Learning	109
6.2.3	Theoretical Studies on HTL in Active Learning Paradigm	110
	Publications	111
	References	112

List of Figures

1.1	Examples of Flickr Dataset.	6
1.2	Illustration of Text based Image Classification.	6
1.3	Illustration of Cross-language Document Classification.	7
1.4	Illustration of Cross-system Recommendation.	8
1.5	Illustration of a sparse feature representation matrix.	10
1.6	Examples from Office dataset	12
1.7	Hybrid Heterogeneous Transfer Learning.	13
1.8	Active Correspondences Construction for HTL	15
1.9	The organization of thesis.	18
2.1	A structure of DeviSE [40].	26
2.2	A structure of Deep Cross-Modal Transfer [87].	26
2.3	A structure of Multimodal Deep Learning [66].	30
2.4	A structure of Multimodal Log-Bilinear model [54].	32
3.1	Example of an ECOC matrix	46
3.2	Illustration of the recovered feature mappings by using different methods on the toy data. The black color represents 0 and white color represents 1.	57
3.3	Comparison of different HTL methods under varying target labeled data size and dimensions.	60
3.4	Relationships between binary base classifiers and multi-class SHFR (ECOC).	61
3.5	The performances of SHFR under (i) ECOC and (ii) ECOC with removing negative alignments.	64
3.6	Time complexity comparison	69
4.1	Deep Structures for Proposed HHTL method.	81

4.2	Parameter Analysis.	87
5.1	Distribution Matching based Matrix Completion (DMMC).	95
5.2	Heterogenous Transfer Learning through Active correspondence construction (HTLA).	99
5.3	Performance Comparison.	103
5.4	Parameter Analysis.	104
6.1	Illustration of unsupervised heterogeneous transfer learning.	110

List of Tables

2.1	Notation Table.	22
3.1	Complexity Comparison for State-of-the-art HTL baselines.	55
3.2	Multilingual Reuters Collection: comparison results in terms of classification acc (%).	59
3.3	BBC Collection: comparison results in terms of classification accuracy (%).	59
3.4	Cross-lingual Sentiment Dataset: comparison results in terms of classification accuracy(%).	60
3.5	Comparison of Correlation $\sum_{ji} \rho_{ji} $	62
3.6	Comparison results on each binary task in terms of classification accuracy (%).	62
3.7	Impact of Nonnegative Constraints and L_1 Norm Regularization: comparison among different solvers results in terms of classification accuracy (%).	65
3.8	Comparison results on each binary task after removing negative alignment in terms of classification accuracy (%).	67
4.1	Learning with unbiased target correspondence instances: comparison results in terms of testing accuracy (%).	83
4.2	Learning with biased target correspondence instances: comparison results in terms of testing accuracy (%).	84
5.1	Comparison results in classification accuracy (%).	102
6.1	An example of N -ary coding matrix M , where $N = 4$	109

Chapter 1

Introduction

For many real-world classification problems, it is often expensive or time consuming to annotate sufficient labeled training data to build machine-learning-based information systems. To overcome this challenge, many semi-supervised and active learning approaches have been proposed. However, most proposed approaches assume that training and testing samples are drawn from the same distribution. This assumption is often not realistic for many real-world problems due to changing environments. Transfer learning (TL) which aims to adapt a model from an auxiliary domain to a domain of interest with little or without additional human supervision has attracted growing attentions in recent years. In general, most TL methods assume that the data come from different domains having the same feature space and dimensionality. However, these assumptions may be violated for some applications, such as text based image classification, cross-language document classification and cross-system recommendation, where the feature spaces of auxiliary domains and domain of interest are different. To address this difficulty, one has to perform a new type of *transfer* learning called **heterogeneous transfer learning** (HTL) with totally different feature spaces.

In this thesis, we investigate research challenges in HTL, namely **sparsity, multi-class classification, bias correspondences, distribution matching and active correspondences**. Moreover, we propose three approaches to handle the five challenges.

The rest of this chapter is organized as follows. In Section 1.1, we discuss some weak supervision challenges and some existing machine learning paradigms to handle them. We present the background of transfer learning in Section 1.2. In Section 1.3, we motivate the usefulness of HTL from the perspective of real-world applications. Furthermore, in

Section 1.4, we describe the five research issues in HTL using the cross-language text classification as an example. Section 1.5 summarizes the contributions of this thesis. Section 1.6 outline the structure of the thesis.

1.1 Weak Supervision Challenges

Supervised data mining and machine learning methods are extensively studied and applied to many knowledge engineering areas. However, the success of most traditional supervised algorithms assumes that there are sufficient high quality labeled data collected to build a statistical model [17, 50]. However, labeled data are expensive to collect due to the high human labeling and annotation cost. For example, documents can be crawled from the Web, images can be obtained from surveillance cameras, and speeches can be collected from broadcast. However, the labeled data required for the prediction tasks such as face recognition [112], sentiment analysis [64], or topic categorization [59], often require expensive human annotation. The problem of weak supervision¹ is a major obstacle in the machine learning and data mining community that renders established methodology impractical in real-world applications.

To overcome this problem, semi-supervised learning techniques [23, 113, 116] (and reference therein) have been proposed where a small number of labeled data together with a large amount of unlabeled data obtained at cheap cost are used to discover the underlying data structure. These techniques have shown very promising results in enhancing generalization capability of models [79]. Nevertheless, the semi-supervised learning setting requires labeled training data and unlabeled training data are drawn from the same data distribution.

Another machine learning paradigm to address weak supervision is active learning [82]. Different from semi-supervised learning, active learning selects the most useful data to build a better predictive model to be annotated by an expert. This leads to a better performance with fewer training labeled data [61, 82]. Similar to semi-supervised learning, most active learning algorithms assume that data to be annotated are also drawn from the same distribution. Moreover, active learning setting usually assumes

¹Weak supervision refers to the problem where limited supervised information, e.g., labels, annotations, are available during the training part.

that the number of annotations (or queries) are limited (or there is a fixed budget) and determined by the funds or time available to spend on annotation. However, in some real-world scenarios, where either the queries can be biased or the budget is very limited, active learning algorithms fail to achieve satisfactory performances.

1.2 Transfer Learning

From the discussion of previous section, we know that both semi-supervised learning and active learning paradigms assumes that the training and testing data distribution are the same. However, many classification tasks in the real world violate such assumption. One example is the sentiment classification task, which one is required to classify the reviews on various products. The traditional learning algorithm cannot directly apply to cross-domain sentiment analysis since different products have different reviews. To be specific, in the book domain, we may use words like “hollow” and “moved” to express our positive feelings while “thick” and “tedious” express our negative sentiments. Different from the products in the book domain, the comments in the video game domain are more likely to include words like “hooked” and “exciting” to express users positive sentiments while words like “unrealistic” is used to indicate negative feelings. Thus, domain-specific words induce the data distribution shift across the two domains. Another example is related to RGB-D data. With the recently developed depth sensors such as Microsoft Kinect, video data with depth information (namely RGB-D) is of increasing importance to many computer vision problems, such as human activity recognition, action categorization and human-computer interaction. However, existing supervised RGB-D video data are in short supply due to expensive data collection. Thus, in this case, the models trained on these datasets are not statistically stable and cannot generalize well to the future unseen data. On the other hand, existing well established image dataset (e.g. ImageNet) or loosely labeled web videos (e.g. YouTube) are readily available. However, we cannot directly apply these well supervised datasets as training data to construct models for prediction tasks using RGB-D data since the distribution and features of the two datasets are very different.

Transfer learning is a new machine learning paradigm, which aims to transfer knowledge extracted from an auxiliary domain, i.e., a source domain, where sufficient labeled

data are available, to solve learning problems in a new domain, i.e., a target domain, with little or no additional human supervision [70]. The study of transfer learning is motivated by the fact that people can intelligently apply previously learned knowledge to solve new problems faster or with better solutions [38]. For example, the existing ability to drive a car helps a person to learn to drive a truck faster. Thus, people in machine learning community want to apply this concept to solve machine learning and data mining related tasks [72]. In this case, transfer learning algorithms correct the distribution drift and transfer knowledge from related domains to domain of interest [12, 29].

1.3 Heterogeneous Transfer Learning

Although transfer learning algorithms show promising performances in the past few years, most of them are designed for homogeneous transfer learning where the features of auxiliary domain and domain of interest are of the same dimension as well as feature spaces. However, in practice, the two domains are not necessarily the same. Recently, more researches have been focusing on transferring knowledge across heterogeneous domains where the source and target domains may have heterogeneous types of features [105]. Different from homogeneous transfer learning, which assumes that the source and target domain data are represented in the same feature space of the same dimensionality [13, 68], heterogeneous transfer learning allows the source and target domains to be from different feature spaces. There are many real-world applications where heterogeneous transfer learning is crucial. We describe three such applications below.

- (i) **Text-based image classification.** Labeled images are very expensive and are often in short supply. On the other hand, related text data are much cheaper and easier to obtain. For instance, when we search “airplane” through search engine like Google, Baidu, there are a lot of returned results, most of which are text based information. Moreover, sometimes the images are naturally associated with text tags. Figure 1.1 shows such images found on the Flickr website.

Moreover, it is well known that image classification is much harder than document classification because images are more ambiguous and feature-dependent [117]. Therefore, making use of abundant related text data becomes a promising way

Figure 1.1: Examples of Flickr Dataset.

Figure 1.2: Illustration of Text based Image Classification.

to improve the performance of image classification task. However, the knowledge transfer between text and image domains is not easy since the text data features are totally different from that of image data.

- (ii) **Cross-language document classification.** As shown in Figure 1.3, English is the most popular language in the world and there are more English documents than documents in other languages, such as German and French. In this case, labeling or annotations of the non-English documents are often obtained through expensive manual-labeling efforts. To reduce labeling cost, one attractive solution is to utilize the labeled English documents to assist the tasks on non-English documents. However, it is impossible to apply conventional transfer learning algorithms to this application due to the problem of different wordbooks (heterogeneous features) for different languages. For example, English do not share the same set of vocabularies as German if we apply the bag-of-words feature.

Figure 1.3: Illustration of Cross-language Document Classification.

- (iii) **Cross system recommendation.** Cross-system recommendation can be deemed as one special application of HTL. For instance, Netflix recommendation system mainly suggests movies and TV series, while Amazon recommends not only movies

but also other products such as books and electronics (See Figure 1.4). Note that a lot of items (or users) are exclusive to each recommendation system. Netflix only focuses on movie and TV series rating, which includes also unpopular movies and TV series. On the other hand, Amazon is very likely to exclude those unpopular items. Moreover, some items are shared among different recommendation systems, such as the popular blockbusters and movies. Similarly, the majority of users are exclusive to each recommendation system and a small subset of users can be found in both recommendation systems. The cross-system recommendation task is to transfer the knowledge from one recommendation system (source domain) to help the recommendation of another system of interest (target domain). In this case, different recommendation systems have feature spaces, which increase the difficulty of knowledge transfer.

Figure 1.4: Illustration of Cross-system Recommendation.

The above three examples demonstrate the urgent need for knowledge transfer across heterogeneous feature spaces instead of same features spaces. This machine learning task is called **Heterogeneous Transfer Learning (HTL)**. Solutions for HTL are urgently needed to overcome real world problems we described above. In this thesis, we investigate

research issues relevant to HTL task that are not considered by previous work. Moreover, we propose new methods to handle these issues.

1.4 Research Issues on Heterogeneous Transfer Learning

To achieve good performance for HTL, one needs a good mapping between two different feature spaces or a pair of mappings to map heterogeneous feature spaces to a common latent space. Based on different assumptions of most existing HLT approaches, they can be classified into two categories. In the first category, a few labeled data in the target domain are assumed to be available for training [37, 55, 96]. For example, given one topic, there are a number of documents in different languages, where different vocabularies are used. Nevertheless, these documents are highly related to each other since they describe the same topic. In the second category, a subset of correspondences of instances between domains are assumed to be known. [28, 114]. Such correspondences are available in the cross-language document classification. To be more specific, we are given English document, which can be translated into other languages either by human experts or the machines to construct cross-domain correspondences. Such “correspondences” can assist in transferring knowledge.

In this thesis, the sparsity and multi-class classification issues are investigated under the first setting. The bias correspondences, distribution matching and cross-domain correspondences issues are investigated under the second setting. These five research issues are explained and discussed.

1.4.1 Sparsity

Existing HTL methods tend to recover large number of feature mappings between domains, which is not feasible without sufficient constraints. Furthermore, these dense mappings introduce a lot of noises which affect the classification performance.

One can interpret the sparsity issue as the need to have a small number of feature mappings across domains. In other words, a feature in one domain is represented by a small subset of features in another domain. When sparsity is achieved, there will be

a significant reduction in computational cost on very high-dimensional data and better interpretability of feature relationships across domains.

Sparsity can be understood as an effective translation task. We use a two-language (e.g., English vs. Spanish) text classification task as a motivating example. Typically, the word “Sano” in Spanish has similar meaning to the words “Great”, “Good”, “Wholesome”, and “healthy”, but not all 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 only in the English domain. As illustrated in Figure 1.5, the sparse matrix $\mathbf{G}$ denotes the feature mapping from the English domain to the Spanish domain. Based on the sparse matrix $\mathbf{G}$, the word “Sano” in Spanish can be represented sparsely by only four words in English as “Sano” = $0.5 \times \text{“Great”} + 0.3 \times \text{“Good”} + 0.1 \times \text{“Wholesome”} + 0.1 \times \text{“Healthy”}$.

A solution that considers sparsity on the feature mappings will ensure better knowledge transfer between two different feature spaces.

Figure 1.5: Illustration of a sparse feature representation matrix.

1.4.2 Multi-class Information

It is observed that a lot of real-world HTL problems are not simple binary classification tasks to answer “yes” or “no”. The main challenge is to incorporate multi-class information into HTL prediction model.

The cross-language document classification task is a multi-class task. The objective of the target task is to categorize Spanish documents into six document types. Another example is the cross-domain *Office* object categorization tasks consisting of images collected from Amazon.com, Digital SLR and Webcam [55]. Different sources of images differ largely from each other on the aspects of illumination, resolution and variance, which induces heterogenous features (see Figure 1.6). In the *Office* dataset, there are 31 different classes, such as computer, lamp, calculator and etc. Most existing HTL algorithms return the multi-class predictions by simply applying the one-versus-all (OVA) scheme, which may discard underlying dependencies among multiple classes. For instance, in the cross-language document classification task, the feature mapping of the word “Sano” is invariant to different classes (See Figure 1.5). Hence, the common feature mapping invariant to classes makes sense. In this case, OVA may not always be satisfactory for multi-class classification problems [32].

1.4.3 Bias Correspondences

Apart from the label-based HTL algorithms, other HTL approaches aim to learn a feature mapping across heterogeneous feature spaces based on some cross-domain correspondences constructed either by labels [55] or *a translator* [28]. With the learned feature mapping, instances can be mapped from the target domain to the source domain or the other way round. Hence, source domain labeled data can be used to learn an accurate classifier for the target domain.

There is a common assumption for these methods: the selection of the instance-correspondences to learn the feature mapping is *unbiased*. In other words, the cross-domain corresponding instances are assumed to be representative in both source and target domains. However, in many real-world scenarios, this assumption may not hold, which means that the selected corresponding instances may be biased and able to represent neither the source nor the target domain data.

Taking cross-language document classification as a motivating example illustrated in Figure 1.7.a, the objective is to learn a text classifier on documents in one language (e.g., German) only with a set of annotated English documents. To apply HTL methods to solve this task, one can construct German-English document-correspondences by

1.6.a: Amazon: Laptop

1.6.b: Amazon: Bottle

1.6.c: Amazon: Phone

1.6.d: DSLR: Laptop

1.6.e: DSLR: Bottle

1.6.f: DSLR: Phone

1.6.g: Webcam: Laptop

1.6.h: Webcam: Bottle

1.6.i: Webcam: Phone

Figure 1.6: Examples from Office dataset

1.7.a: Case I: target corresponding instances are not biased

1.7.b: Case II: target corresponding instances are biased

Figure 1.7: Hybrid Heterogeneous Transfer Learning.

translating some German documents into English by Google translator. However, the *wordbook* of the translated English documents may be quite different from that of the original English documents. For instance, the German word “betonen” is translated into the English word “emphasize” by Google translator. However, in an original English document, its corresponding word is “highlight” or “stress”. This is referred to as the “bias” issue in word-distribution between the translated documents and the original ones in the target domain. In this case, a feature mapping learned on such correspondences may not be effective.

In the second example shown in Figure 1.7.b, we consider a multilingual sentiment

classification task, where our task is to automatically classify overall sentiment polarities of songs reviews in German given labeled books reviews in English, and some unlabeled pairs of songs reviews in German and their English translations. Though one can make use of the German-English review-correspondences to learn a feature mapping from English to German, the labeled book reviews in English after transformation may not be useful to learn an accurate sentiment classifier for songs reviews in German. The reason is that, opinion and topic words used on different types of products can be very different [62]. In the context of this example, the “bias” between the transformed source domain data and the target domain data is caused by the difference of product domains.

1.4.4 Distribution Matching

The success of transfer learning heavily relies on distribution similarity between two domains after domain adaptation. Maximum Mean Discrepancy (MMD) [48] is a very important tool to measure the distribution difference between different domains in infinite order. It has been frequently used in designing homogeneous transfer learning algorithms such as Maximum Mean Discrepancy Embedding (MMDM) [68], Transfer Component Analysis (TCA) [69], and Kernel-Mean Matching (KMM) [53]. However MMD is a distribution difference measure proposed exclusively for distributions with same feature spaces.

For the HTL problem, explicitly minimizing the distribution differences across domains is much more challenging. Although there are some existing methods such as MOMAP [49] and HeMap [84] that tried to minimize the domain difference for HTL, these methods only consider the first order information of distribution when adapting domains. An approach that explicitly minimizes the domain difference between two domains with different feature spaces in terms of MMD for HTL has not been discussed in existing literature.

1.4.5 Correspondences Construction

For the existing correspondences-based HTL methods, their common assumption is that the cross-domain correspondences are given in advance. However, in many real-world

1.8.a: OCR Error

1.8.b: Uncertainty on Finding Correspondences

Figure 1.8: Active Correspondences Construction for HTL

applications, instance-correspondences between domains are unknown. To obtain accurate correspondence construction incurs significant cost as extensive manual efforts are required. For example, for the task of topic categorization for scanned books in a minor language, e.g. German, Vietnamese and etc, there may be few annotated books as the construct of a precise hierarchy of categories is very expensive, needs to be done manually. In contrast, there are plenty of annotated books in a major language, e.g., English. Intuitively, one can apply HTL techniques to transfer annotation information from a major language to a minor language for book categorization. This usually requires a sufficient set of book correspondences between two languages as inputs.

Moreover, due to optical character recognition (OCR) error when the scanned books are converted into editable data, direct machine/human book translation fails to construct precise book correspondences between the two languages. For example, in Figure 1.8.a, the book in German “Künstliche Intelligenz” turns into unrecognized “K̈ unsti-

iche mtelligonz” by ORC readers. Such error may induce ambiguity for constructing cross-language correspondences in Figure 1.8.b. The German book named “Künstliche Intelligenz” after OCR scan and conversion can correspond to either the book “Artificial Intelligence” (“Künstliche Intelligenz” in German) or the book “Artistic Director” (“Künstlerisch Intelligenzia” in German) in English. In this case, expensive human efforts are required to construct the precise correspondences based on the content and meta information of the books. In this thesis, we propose a new approach that relaxes the assumption of existing correspondences-based HTL algorithms to enable HTL to be more feasible for real-world applications.

1.5 Contributions

In the thesis, we propose three HTL approaches to solve the above mentioned issues.

- **Sparse heterogeneous feature representation for multiple classes**

To address the first two issues, we propose a *sparse heterogeneous feature representation* (SHFR) to learn the sparse transformations for HTL, by fully exploring the common underlying structures among multiple classes of two domains. Our major result is that, according to the compressed sensing theory, a sparse feature mapping can be learned if and only if a sufficient number of classifiers are provided. Based on this observation, to accurately estimate the sparse feature mapping, we propose to generate sufficient classifiers using the *error correcting output correcting* (ECOC) scheme. To our best knowledge, this work is the first to give the generalization *error bound* for HTL in binary case. Furthermore, based on the analysis of the binary case, we give the generalization error of multi-class HTL problems based on the ECOC. Last but not least, we present an *batch-mode* pursuit method to efficiently solve the resultant non-negative sparse recovery problem. The computational cost of SHFR is independent of the data size, and scales linearly only with the target feature dimensions.

- **Hybrid heterogeneous transfer learning**

To address bias cross-domain correspondences issue, we present a new transfer learning framework called *Hybrid Heterogeneous Transfer Learning* (HHTL), which allows the corresponding instances across domains to be biased in either the source or target domain. Specifically, we propose a deep learning approach to learn a feature mapping between cross-domain heterogeneous features as well as a better feature representation for mapped data to reduce the bias issue caused by the cross-domain correspondences.

- **Heterogenous transfer learning through active correspondence construction**

To address the last two mentioned issues in HTL, we propose a novel and general HTL framework named *Heterogenous Transfer Learning through Active correspondence construction* (HTLA) to overcome the problem where the cross-domain correspondences are not given in advance. Another contribution is an efficient algorithm to solve matrix completion regularized by Maximum Mean Discrepancy (MMD) [48], which is based on stochastic sub-gradient descent (SSDG) with projection on the feasible space $\mathcal{K}$, where MMD is minimized. We solve the feasible space projection problem using a concave-convex procedure [107]. Based on this proposed method, we further propose a novel framework to actively construct correspondences between domains for effective knowledge transfer.

1.6 Thesis Outline

The organization of this thesis is shown in Figure 1.9. Below is a summary of the rest of chapters in this thesis:

Chapter 2. Literature Review.

We first introduce some related concepts and the notations of heterogeneous transfer learning (HTL). Furthermore, we review HTL approaches and categorize them into four groups. We also discuss the strengths and weaknesses of existing methods.

Chapter 3. Sparse Heterogeneous Feature Representation for Multiple Classes.

Figure 1.9: The organization of thesis.

In this chapter, we deal with multi-class classification task with sparse features mapping. We cast the multi-class HTL into a compressed sensing (CS) problem, where feature mapping from source to target domains is sparse. A generalization error bound of HTL is derived under the multi-class setting. To further speed up the estimation of the transformation matrix, we also present a very efficient batch-mode algorithm to solve the corresponding non-negative sparse recovery problem. Extensive experiments are conducted on both toy data and three real-world HTL datasets. The empirical results demonstrate the superiority of our proposed method over existing state-of-the-art HTL methods in terms of both prediction accuracy and training efficiency.

Chapter 4. Hybrid Heterogeneous Transfer Learning.

This chapter first shows some real situations where the bias issue of the correspondences in the target or (and) source domain(s) occurs. To address this issue, we propose a new transfer learning framework called Hybrid Heterogeneous Transfer

Learning (HHTL), which can tolerate the corresponding instances across domains to be biased in either the source or target domain. Specifically, we propose a deep learning approach consisting of two parts: 1) learning deep feature mapping between cross domain heterogeneous features in each layer; 2) learning better feature representation for mapped data through marginalized stacked denoised autoencoder (mSDA) [26] to reduce the distribution bias issue. Extensive experiments on several multilingual sentiment classification tasks verify the effectiveness of our proposed approach compared with some baseline methods.

Chapter 5. Heterogenous Transfer Learning via Active Correspondence Construction.

In this chapter, we aim to integrate active learning into HTL. First, we review some related active transfer learning approaches and discuss the differences between our proposed framework and their problems. Furthermore, a general framework named heterogeneous transfer learning through active correspondences construction (HTLA) is presented to integrate active learning to construct correspondences between heterogeneous domains for HTL. We first propose a new correspondence-based HTL method via matrix completion (MC), where a regularization term on distribution matching is encoded to reduce difference between heterogeneous domains. Specifically, our proposed approach can be divided into two parts: 1) stochastic sub-gradient descent for MC, and 2) projection onto feasible space. For the first part, we adopt probing matrix method in [5] to give unbiased sub-gradient descent for MC. For the second part, we adopt concave-convex procedure [107]. Based on this novel method, we further propose a strategy to actively construct correspondences between domains. We conduct extensive experiments on various multilingual sentiment classification tasks to verify the effectiveness of the proposed framework compared with some state-of-the-art baselines.

Chapter 6. Conclusion and Future Directions.

We conclude this thesis and discuss possible future work.

Chapter 2

Literature Review

Transfer learning is not the panacea for utilizing knowledge from an auxiliary domain to solve problems in any other domains. One cannot expect to transfer knowledge across irrelevant domains. In this case, brute-force transfer may be unsuccessful or even hurt the performance of learning in the target domain. Therefore, there must exist some “bridges” to link the knowledge across domains.

“Bridge” can be labeled data from a target domain with the assumption that the output space of the source domain is the same as the target domain. Hence, the data with same labels are usually related in some ways. For example, for the same object, we can use different facilities to take pictures such as digital single-lens reflex camera (DSLR) and low-pixel camera (webcam), which lead to different resolutions and qualities. When we preprocess data and extract features for different resources, it may result in different feature spaces. However, they are highly related using the same labels because their objects are same.

Cross-domain correspondences can be the second “bridge” to connect the two heterogeneous domains. For the cross-language document classification task, one translates an English document to other languages either by human experts or the machines to construct cross-domain correspondences. Either “labels” or “cross-domain correspondences” can function as “bridge” to help knowledge transfer. Their quality and quantity are vital to the performance of HTL.

In this chapter, we first describe the notations, concepts, and definitions that are used in the next few chapters. Based on these notations and definitions, we give a comprehensive review of HTL approaches for classification developed in the machine

learning and computer vision research communities. Then, we categorize the approaches into the two types of “bridges”: “labels” and “cross correspondences”.

2.1 Notations and Definitions

“Domain” can be divided into source domain $\mathcal{D}_S$ and target domain $\mathcal{D}_T$, where source domain has plenty of labeled data and target domain has few labeled or no labeled data. Let $\mathbf{X}_S = \{(\mathbf{x}_{S_i}, y_{S_i})\}_{i=1}^{n_S}$ denote a set of labeled training instances of the source domain, where $\mathbf{x}_{S_i} \in \mathcal{X}_S$ denotes the i th instance and $y_{S_i} \in \{1, 2, \dots, c\}$ denotes its 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 $n_T \ll n_S$, $\mathbf{x}_{T_i} \in \mathcal{X}_T$ and $y_{T_i} \in \{1, 2, \dots, c\}$. Except for labeled data, in some applications such as cross-language sentiment analysis, there exists a set of corresponding pairs between the source and target domains, $\{(\mathbf{x}_{S_i}^C, \mathbf{x}_{T_i}^C)\}_{i=1}^{n_C}$, where $\mathbf{x}_{S_i}^C \in R^{d_S}$ and $\mathbf{x}_{T_i}^C \in R^{d_T}$ are source- and target- domain instances, respectively. For simplicity, we denote $\mathbf{X}_{SC} \in R^{n_C \times d_S}$ and $\mathbf{X}_{TC} \in R^{n_C \times d_T}$ the matrices of corresponding instances in the source and target domains, respectively. For clarity and consistency, we summarize these notations in Table 2.1 and they are used in the rest of the thesis.

For HTL, the input spaces for source and target domain are not the same, namely $\mathcal{X}_S \neq \mathcal{X}_T$. Thus, their number of dimensions are usually not equal, namely $d_S \neq d_T$, where d_S, d_T denote the number of dimensions of source and target domain, respectively. The “task” in HTL is to give an optimal objective predictive function for target domain denoted by $f_T(\cdot) \in \mathcal{Y}_T$ by transferring the knowledge of source domain data $\mathbf{X}_S$ or predictive function $f_S(\cdot)$. The target output space $\mathcal{Y}_T$ can be continuous or categorical, which depends on the machine learning tasks. For the binary or multi-class classification task, the output space is categorical. For the regression task, the output space is continuous.

Following the definition of general transfer learning given in [70], we provide a formal definitions of HTL.

Definition 1. *Given a source domain $\mathcal{D}_S$ and learning task $\mathcal{T}_S$, a target domain $\mathcal{D}_T$ and learning task $\mathcal{T}_T$, heterogeneous transfer learning aims to help improving the learning of target predictive function $f_T(\cdot)$ in $\mathcal{D}_T$ by making use of the knowledge in $\mathcal{D}_S$, $\mathcal{D}_T$ and cross-domain knowledge, where input feature spaces must be different, namely, $\mathcal{X}_S \neq \mathcal{X}_T$.*

Table 2.1: Notation Table.

Source-, target- domain data matrix	$\mathbf{X}_S, \mathbf{X}_T$
Correspondences in source-, target- domain	$\mathbf{X}_{SC}, \mathbf{X}_{TC}$
Feature mapping for source-, target- domain	$\mathbf{P}, \mathbf{Q}$
Feature Mapping from source to target domain	$\mathbf{G}$
The i th instance in source-, target- domain	$\mathbf{x}_{S_i}, \mathbf{x}_{T_i}$
The label of i th instance in source-, target- domain	y_{S_i}, y_{T_i}

In this paper, we only consider the HTL cases which the following three assumptions hold.

- (i) There is one source domain $\mathcal{D}_S$, and one target domain $\mathcal{D}_T$.
- (ii) The same output space is shared between source domain and target domain, namely $\mathcal{Y}_S = \mathcal{Y}_T$.
- (iii) There is always a relatively large amount of labeled data (or cross-domain correspondences) in the source domain and a small amount of labeled data (or cross-domain correspondences) in the target domain.

Note that for the more general HTL, the second assumption can be relaxed to $\mathcal{Y}_S \neq \mathcal{Y}_T$. This increases the difficulties for transferring heterogeneous knowledge. In this thesis we only discuss HTL where input spaces differ while output spaces remain the same. Furthermore, all the discussion in the rest of the chapter is based on the above assumptions as these three assumptions hold for the most recent HTL works in literature.

All HTL approaches can be divided into two categories according to the third assumption. One category assumes that a small number of target domain labeled data are given. The other category assumes that some cross-domain correspondences are given in advance.

2.2 Label-based HTL

The first category of approaches is based on the labeled target domain data. In general, most labeled-based HTL approaches can be further classified into two subcategories.

The first one is to learn a pair of feature mappings to transform the source and target heterogeneous data into a common latent space. Formally, one can learn two feature mappings $\mathbf{P}$ and $\mathbf{Q}$ to transform the source domain data $\mathbf{X}_S$ and target domain data $\mathbf{X}_T$, respectively, to a new latent common feature space such that the distance between the mapped domain $\mathbf{P}\mathbf{X}_S$ and $\mathbf{Q}\mathbf{X}_T$ is minimized [37, 84, 96]. Next, we describe three HTL approaches in this subcategories.

Heterogeneous Spectral Mapping. Shi *et al.* [84] proposed a Heterogeneous Spectral Mapping (HeMap) method to learn the mappings based on the spectral embedding without using any label information. The objective is to find the projected spaces $\mathbf{B}_S$ and $\mathbf{B}_T$ for source and target domain, respectively. The source and target data are similar in distribution and at the same time, the original structures are preserved. The objective is formulated as follows:

$$\min_{\mathbf{B}_T, \mathbf{B}_S} l(\mathbf{B}_T, \mathbf{X}_T) + l(\mathbf{B}_S, \mathbf{X}_S) + \beta \text{dist}(\mathbf{B}_S, \mathbf{B}_T), \quad (2.1)$$

where the reconstruction loss for target and source domain are defined as $l(\mathbf{B}_T, \mathbf{X}_T) = \|\mathbf{B}_T \mathbf{P} - \mathbf{X}_T\|$ and $l(\mathbf{B}_S, \mathbf{X}_S) = \|\mathbf{B}_S \mathbf{Q} - \mathbf{X}_S\|$ respectively, and $\text{dist}(\mathbf{B}_S, \mathbf{B}_T) = \frac{1}{2}(l(\mathbf{B}_T, \mathbf{X}_T) + l(\mathbf{B}_S, \mathbf{X}_S))$ denotes the distance between the two domain data. HeMap discards the label information into learning feature mapping. Even though the distribution minimization is imposed on the model, the distribution difference estimation is still based on first order distribution. In the real-world applications, more sophisticated distribution differences regularization is required to capture complex higher-order information.

Manifold Alignment. Wang and Mahadevan [96] proposed a manifold alignment method called DAMA in the sequel that aligns heterogeneous features into a latent space based on the manifold regularization. In particular, DAMA aims to simultaneously match instances with same labels, to separate instances with different labels and to preserve topology of each domain. These three goals can be achieved by penalizing the following terms A , B and C respectively.

$$A = 0.5 \sum_{i=1}^{n_S} \sum_{j=1}^{n_T} \|\mathbf{P}^\top \mathbf{x}_{S_i} - \mathbf{Q}^\top \mathbf{x}_{T_j}\|^2 \mathbf{W}_s(i, j), \quad (2.2)$$

where $\mathbf{W}_s$ is a binary similarity matrix.

$$B = 0.5 \sum_{i=1}^{n_S} \sum_{j=1}^{n_T} \|\mathbf{P}^\top \mathbf{x}_{S_i} - \mathbf{Q}^\top \mathbf{x}_{T_j}\|^2 \mathbf{W}_d(i, j); \quad (2.3)$$

where $\mathbf{W}_d$ is a binary dissimilarity matrix.

$$C = 0.5\mu \sum_{k=S,T} \sum_{i=1}^{n_k} \sum_{j=1}^{n_k} \|\mathbf{P}^\top \mathbf{x}_{k_i} - \mathbf{Q}^\top \mathbf{x}_{k_j}\|^2 \mathbf{W}_k(i, j); \quad (2.4)$$

where $\mathbf{W}_k$ with $k = S, T$ is a similarity matrix derived from the manifold learning with heat kernel [7]. Thus the objective of DAMA is to minimize A , C and at same time maximize B .

$$\min_{\mathbf{P}, \mathbf{Q}} (A + C) / B. \quad (2.5)$$

The most interesting idea in DAMA is the use of labels to align manifold instead of correspondences. The above problem is easily solved by generalized eigenvalue decomposition. Unfortunately, DAMA only works on the data that have strong manifold structures. This weakness limits knowledge transferability on those data that the manifold assumption does not hold.

Heterogeneous Feature Augmentation. Duan *et al.* [37] proposed a heterogeneous feature augmentation (HFA) method to augment homogeneous common features learned by the maximum-margin approach from both source and target domains.

$$\min_{\mathbf{P}, \mathbf{Q}} \min_{\mathbf{w}, b, \xi_i^S, \xi_i^T} \frac{1}{2} \|\mathbf{w}\|^2 + C \left(\sum_{i=1}^{n_S} \xi_i^S + \sum_{i=1}^{n_T} \xi_i^T \right), \quad (2.6)$$

$$s.t. \quad y_{S_i}(\mathbf{w}^\top \varphi_S(\mathbf{x}_{S_i}) + b) \geq 1 - \xi_i^S, \xi_i^S \geq 0; \quad (2.7)$$

$$y_{T_i}(\mathbf{w}^\top \varphi_T(\mathbf{x}_{T_i}) + b) \geq 1 - \xi_i^T, \xi_i^T \geq 0; \quad (2.8)$$

$$\|\mathbf{P}\|_F^2 \leq \lambda_p, \|\mathbf{Q}\|_F^2 \leq \lambda_q. \quad (2.9)$$

However, the proposed model results in an expensive semidefinite program (SDP) problem. Moreover, it is nonconvex.

The second subcategory of HTL approaches learns a feature mapping $\mathbf{G}$ to transform heterogeneous data from one domain to another domain directly so that either the distance between $\mathbf{G}\mathbf{X}_S$ and $\mathbf{X}_T$ can be minimized or the alignment between $\mathbf{G}\mathbf{X}_S$ and $\mathbf{X}_T$ can be maximized. Next, we describe four HTL approaches in this subcategories.

Multiple Outlook Mapping (MOMAP). MOMAP [49] is proposed to learn rotation matrix $\mathbf{G}_i$ to match source data distributions to that of the target domain for each class i in an unsupervised manner which does not use the valuable training labels.

$$\{\mathbf{G}_i\} = \arg \min_{\{\mathbf{G}_i\}} \sum_{i=1}^c \|\mathbf{G}_i \mathbf{X}_S^i - \mathbf{X}_T^i\|_F^2, \quad (2.10)$$

$$s.t. \quad \mathbf{G}_i^\top \mathbf{G}_i = \mathbf{I}, \quad i = 1, 2, \dots, c, \quad (2.11)$$

where $\mathbf{X}_S^i, \mathbf{X}_T^i$ are the source and target data matrix for class i , respectively and $\mathbf{I}$ denotes the identity matrix. The model is totally unsupervised without incorporating the valuable label information into learning feature mapping. The labels are only used in training classifiers. Moreover, the model involves a series of expensive singular value decompositions.

Asymmetric Regularized Cross-domain transformation. Kulis *et al.* [55] proposed an asymmetric regularized cross-domain transformation (ARC-t) method to learn a asymmetric transformation across domains based on metric learning.

$$\min_{\{\mathbf{G}\}} \frac{1}{2} \|\mathbf{G}\|_F^2 + \lambda \sum_i c_i(\mathbf{X}_S^\top \mathbf{G} \mathbf{X}_T), \quad (2.12)$$

where c_i are the loss functions over the constraints expressed in terms of $\mathbf{X}_S^\top \mathbf{G} \mathbf{X}_T$. Specifically speaking, ARC-t aims to find the optimal solution $\mathbf{G}$ under a series of constraints that $c_i(\mathbf{X}_S^\top \mathbf{G} \mathbf{X}_T)$ is as large as possible for the same label and as small as possible for the different labels. 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.

More recently, some models based on deep learning are proposed for cross-modal classification and zero-shot learning. Both are related to the HTL problem setting.

Figure 2.1: A structure of DeviSE [40].

Figure 2.2: A structure of Deep Cross-Modal Transfer [87].

DeviSE. A deep visual-semantic embedding model (DeviSE) [40] is proposed to improve image classification with the help of transferred knowledge from unannotated text data using deep learning. Different from previous probabilistic approach [117], DeviSE tackles the visual recognition at a very large scale (all 1k ImageNet classes), especially using a fully unsupervised text model. However, the skip-gram textual model adopted in DeviSE is quite weak in the sense that no prior semantics is built in. A visual object categorization network is shown in Figure 2.1.

Deep Cross-Modal Transfer. Socher *et al.* [87] designed a novel deep learning model for zero-shot learning, where some classes are unseen in the training part. Similar to DeviSE, the key idea behind this approach is also mapping images to be close to semantic word vectors for the same classes. Different from DeviSE, a manifold is then learnt such that a non-linear transformation of the lexical feature vector and the image feature vector is projected closer to each other in this manifold. The idea is illustrated in the Figure 2.2.

2.3 Cross-domain correspondences based HTL

The second category of approaches is based on the cross-domain correspondences. Cross-domain correspondences can be divided into instances-correspondences and features-correspondences. Most existing approaches under the second category utilize instances-correspondences which are more relevant to real-world applications, such as cross-language sentiment analysis.

Kernel Canonical Component Analysis (KCCA). Vinokourov *et al.* [94] first employed KCCA on cross-language document classifications. The KCCA finds the projections in two embedding spaces so that two spaces are highly correlated. This approach can be formulated as the following optimization problem:

$$\max_{\alpha, \beta} \frac{\alpha \mathbf{K}_S \mathbf{K}_T \beta}{\|\mathbf{K}_S \alpha\| \|\mathbf{K}_T \beta\|}, \quad (2.13)$$

where α and β are the representation parameters of source and target domain data points in the reproducing kernel Hilbert space (RKHS) [27].

Translated Learning. The feature mapping is obtained by using some *translators* to construct corresponding features across domains. The new risk minimization framework is proposed based on the language model for machine translation [74]. However this model is specifically designed for text-based image classification. It is difficult to generalize to solve general HTL problem.

Manifold Alignment via Corresponding Projections. Zhai *et al.* [108] proposed a manifold alignment approaches via labeled correspondence pairs. More specifically, the objective can be formulated as follows:

$$\min_{\mathbf{Q}, \mathbf{P}} J(\mathbf{P}, \mathbf{Q}, \mathbf{X}_{SC}, \mathbf{X}_{TC}) + \alpha_S J(\mathbf{P}, \mathbf{X}_S) + \alpha_T J(\mathbf{Q}, \mathbf{X}_T), \quad (2.14)$$

where the first term $J(\mathbf{P}, \mathbf{Q}, \mathbf{X}_{SC}, \mathbf{X}_{TC}) = \sum_{(i,j) \in C} \|\mathbf{P}^\top \mathbf{x}_{S_i}^C - \mathbf{Q}^\top \mathbf{x}_{T_i}^C\|^2$ is the correspondence preserving term. The second term is the loss incurred after reconstruction from k -nearest neighbours $\mathbf{x}_{S_k}$ of the data point $\mathbf{x}_{S_i}$, where $k \in N_S(i)$, which is expressed as $J(\mathbf{P}, \mathbf{X}_S) = \sum_i^{n_S} \|\mathbf{P}^\top \mathbf{x}_{S_i} - \sum_{k \in N_S(i)} w_{i,k} \mathbf{P}^\top \mathbf{x}_{S_k}\|^2$. Similarly, the third term $J(\mathbf{Q}, \mathbf{X}_T) = \sum_i^{n_T} \|\mathbf{Q}^\top \mathbf{x}_{T_i} - \sum_{k \in N_T(i)} w_{i,k} \mathbf{Q}^\top \mathbf{x}_{T_k}\|^2$ is for the target domain data. The first term penalizes the misalignment of cross-domain manifolds, and the last two terms ensure the preservation of the intrinsic relationships and local geometric structures of data in the source and target domains, respectively.

HTL for Image Classification. Zhu *et al.* [117] proposed a framework for transferring the knowledge from text to images to assist image classification, where latent representation for auxiliary images (target domain) is learnt by using all the unlabeled auxiliary documents (source domain). The model is based on latent semantic analysis (LSA) [30] described as follows:

$$\min_{\mathbf{U}, \mathbf{V}, \mathbf{W}} \lambda \|\mathbf{C} - \mathbf{U}\mathbf{V}^\top\|_F^2 + (1 - \lambda) \|\mathbf{F} - \mathbf{W}\mathbf{V}^\top\|_F^2 + R(\mathbf{U}, \mathbf{V}, \mathbf{W}), \quad (2.15)$$

where λ is a tradeoff parameter and $\mathbf{F}$ is relation matrix between documents and tags. $\mathbf{C} = \mathbf{X}_T^\top \mathbf{X}_S$ denotes the correlation between source and target domains. $\mathbf{U}$ and $\mathbf{V}$ are obtained from the matrix factorization of G . $R(\mathbf{U}, \mathbf{V}, \mathbf{W})$ is the complexity regularization of the matrices.

Two-Step Learning. More recently, a novel two-step learning (TSL) [104] method is proposed for cross-language document classifications. The first step is to fill the missing features of two domains through matrix completion. It is then followed by latent semantic indexing (LSI) [67] to produce the low-dimensional semantic representation. Specifically, in their proposed method, some instance-correspondences between domains are assumed to be given in advance. The goal is to reconstruct corresponding “instances” for all the instances that are only observed in either the source or target domain by exploiting the sparsity and low rank of matrix by solving the following optimization problem:

$$\min_{\mathbf{X}} \frac{1}{2} \|(\mathbf{X} - \mathbf{M}) \circ \mathbf{P}\|_F^2 + \gamma \|\mathbf{X}\|_* + \mu \|\mathbf{X}\|_1 \quad (2.16)$$

where $\mathbf{M}$ is an observed matrix, $\mathbf{P}$ is an indicator matrix such that $\mathbf{P}_{ij} = 1$ if (i, j) is observed, otherwise 0, and $\circ$ is the Hadamard product. γ and μ denote tradeoff parameters for the nuclear norm $\|\cdot\|_*$ and the L_1 norm $\|\cdot\|_1$, respectively.

Coordinate System Adaptation. For cross-system recommendation applications, coordinate system adaptation (CSA) [71] is proposed to address the data sparsity in the target domain by adapting the knowledge from the heterogeneous source domain system. The first step is to apply the sparse singular value decomposition (SVD) [78] on the observed source domain data as follows:

$$\min_{\mathbf{U}_S, \mathbf{V}_S, \mathbf{B}_S} \|\mathbf{P}_S \odot (\mathbf{X}_S - \mathbf{U}_S \mathbf{B}_S \mathbf{V}_S)\|_F^2, \quad (2.17)$$

where $\mathbf{B}_S$ is diagonal matrix with descending eigenvalues, $\mathbf{U}_S, \mathbf{V}_S$ are the orthogonal matrices, $\mathbf{X}_S$ is observed source domain data matrix, and $\mathbf{P}_S$ The second step adapts the discovered basis from the source to the target domain by regularizing the target domain close to source domain basis.

$$\min_{\mathbf{U}_T, \mathbf{V}_T, \mathbf{B}_T} \|\mathbf{P}_T \odot (\mathbf{X}_T - \mathbf{U}_T \mathbf{B}_T \mathbf{V}_T)\|_F^2 + \lambda_1 \|\mathbf{U}_T - \mathbf{U}_S\|_F^2 + \lambda_2 \|\mathbf{V}_T - \mathbf{V}_S\|_F^2, \quad (2.18)$$

where $\mathbf{U}_T, \mathbf{V}_T$, and $\mathbf{B}_T$ are defined similar to the source domain. In this model, the correspondences can be either feature-based or instance-based.

Multimodal Deep Learning. In real-world applications, multiple modalities (“views”) of data, such as audio and images in single video clip, are available. Ngiam *et al.* [66] demonstrate that cross-modalities feature learning can enhance the cross heterogeneous domain classification. Their proposed bimodal deep autoencoder model extends the sparse restricted Boltzmann machines [58] by adding one additional shared representation layer to transfer features between two heterogeneous spaces shown in Figure 2.3. In this deep structure, only one shared bottleneck layer is used

Figure 2.3: A structure of Multimodal Deep Learning [66].

for knowledge transfer which may be insufficient for very complex HTL problems.

Multimodal Neural Language Models. More recently, Kiros *et al.* [54] introduce a multimodal log-bilinear model that jointly learns word representations and image features by training models together with a convolutional network. Specifically, they predict the next word representation $\hat{\mathbf{r}}$ of vocabulary word representation vectors $\mathbf{r}_{w_i}$ by adding the additive bias of image features formulated as the following model (see Figure 2.4):

$$\hat{\mathbf{r}} = \left(\sum_{i=1}^{n-1} \mathbf{C}^{(i)} r_{w_i} \right) + \mathbf{C}^{(m)} \mathbf{x}, \quad (2.19)$$

where $\mathbf{C}^{(m)}$ is a context matrix and $\mathbf{x}$ are image feature vectors. Different from existing cross-modal methods which models at the bag-of-words feature level, this

new approach can even automatically generate sentences descriptions given the images (see Figure 2.5.b) and also can retrieve images given complex sentences queries (see Figure 2.5.a). This model does not require templates, structured prediction, or syntactic trees. Different other HTL approaches that are designed for classification tasks, this type of multi-modal language model is used to 1) retrieve images given complex description queries, 2) retrieve phrase descriptions given image queries, and 3) generate text conditioned on images.

Figure 2.4: A structure of Multimodal Log-Bilinear model [54].

'nine children , some of them waving at the camera , others are a bit shy and are looking away ; one person is pointing to the camera ;

2.5.a: Sample description retrieval given images.

in this picture there is another grey pavement on the right ; three grey clouds and a blue sky in the background ; the houses and on the left before it ; a dark green , wooded slopes behind it ; grey clouds in a light blue sky in the background ; snow covered mountains

2.5.b: Image based description generation.

Chapter 3

Sparse Heterogeneous Feature Representation For Multiple Classes

As introduced in the last chapter, existing label-based HTL approaches can be classified into two categories in general. Methods in the first category learn a pair of dense feature mappings to transform the source and target heterogeneous data into a common latent space. Methods in the second category learn a dense feature mapping to transform heterogeneous data from one domain to the other domain directly.

In the first category, one can learn two dense feature mappings $\mathbf{P}$ and $\mathbf{Q}$ to transform the source domain data $\mathbf{X}_S$ and target domain data $\mathbf{X}_T$, respectively, to a new latent common feature space such that the distance between the mapped domain $\mathbf{P}\mathbf{X}_S$ and $\mathbf{Q}\mathbf{X}_T$ is minimized [37, 75, 84, 96]. For example, Shi *et al.* [84] proposed a heterogeneous spectral mapping (HeMap) method to learn the dense orthogonal mappings based on the spectral embedding without using any label information. The resultant optimization problem is a standard eigenvalue decomposition problem. However, this approach is known to have scalability issues. Wang and Mahadevan [96] proposed a manifold alignment method known as DAMA, to align heterogeneous features into a latent space based on the manifold regularization. Unfortunately, DAMA only works on the data that have strong manifold structures. This limits knowledge transferability to data that do not satisfy the manifold assumption. Furthermore, DAMA results in generalized eigenvalue decomposition of a series of dimension related matrices. More recently, Duan *et al.* [37] proposed a Heterogeneous Feature Augmentation (HFA) method to augment homogeneous common features learned using a maximum-margin approach from both source

and target domains. However, the proposed model results in an expensive semidefinite program (SDP) problem.

In the second category, HTL algorithms learn a dense feature mapping $\mathbf{G}$ to transform heterogeneous data from one domain to the other domain directly so that either the distance between $\mathbf{G}\mathbf{X}_S$ and $\mathbf{X}_T$ can be minimized or the alignment between $\mathbf{G}\mathbf{X}_S$ and $\mathbf{X}_T$ can be maximized. Specifically, in [41], a method named MOMAP was proposed to learn dense rotation matrices to match source data distributions to that of the target domain in an unsupervised manner which does not use the valuable training labels. The resultant optimization problem is solved through singular value decomposition (SVD) for each class in independent way. Besides, Kulis *et al.* [55] proposed an asymmetric regularized cross-domain transformation (ARC-t) method to learn a 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, leading to scalability issues.

Though these methods have shown promising results, they still suffer from the following three major limitations.

- (i) **High Computational Cost.** Since the size of the feature mapping $\mathbf{G}$ scales with the product of the dimensionality of source and target domains, the computational cost to estimate this mapping is extremely high, especially for high-dimensional source and target data, e.g., MOMAP, HeMap, DAMA, ARC-t. To address the computational issue, Duan *et al.* [37] and Kulis *et al.* [55] proposed kernelized versions of the feature-mapping learning methods. However, these kernelized methods still suffer from high computational cost on large-scale data.
- (ii) **Dense Feature Mappings.** Existing methods tend to recover dense feature mappings, which, however, is not feasible without sufficient constraints. As we will later show, dense feature mappings inevitably lose the interpretability and contain a lot of noises which may affect the classification performance.

- (iii) **Multi-class problem.** To address multi-class problems, most existing HTL methods [28, 37, 96] simply adopt one-vs-all strategy to learn multiple binary classifiers independently. Unfortunately, this strategy fails to fully explore the underlying structure among multiple classes. Consequently, this heuristic scheme is unable to guarantee good performance for multi-class HTL problems.

3.1 Exploiting Sparsity and Class-Invariance in Feature Mapping

We propose to address the above issues under two assumptions for HTL learning:

- (i) **Sparse feature representation:** Feature mapping $\mathbf{G}$ between two domains is row-sparse, *i.e.*, each target domain feature can be represented by a small subset of the source domain features.
- (ii) **Class-invariance transformation:** Instances belonging to all the classes adopt the same feature mapping $\mathbf{G}$.

The sparsity of feature mappings across domains means that a feature in one domain can be represented by a small subset of features in another domain. This assumption is valid for real-world multilingual text categorization applications. The sparsity characteristic, which can also facilitate a significant reduction in computational cost on very high-dimensional data, has not been explored in existing HTL methods. To encode the *class-invariance* property into the feature mapping, we propose to learn a common G underlying all the classes. This approach is similar in spirit with multi-task feature learning [3] and the construction of auxiliary tasks in the presence of domain-invariant features [47]. Specifically, we first decompose a multi-class classification problem to multiple binary classification tasks, and then jointly optimize all the classification tasks and the feature mapping G , such that G is class-invariant to all the binary classification tasks.

Based on these two assumptions, we propose a new feature mapping approach for cross-lingual multi-class text categorization. Specifically, to estimate such a feature mapping, we leverage the weight vectors of the binary classifiers learned in the source and target domains. We then formulate the problem of learning feature mapping between

domains as a compressed sensing (CS) problem, which recovers a sparse signal from a few measurements with respect to an underdetermined linear systems, by taking advantage of the signal’s sparseness or compressibility [19, 35]. In summary, the main contributions of this work are listed as follows:

- (i) We propose a **sparse heterogeneous feature representation** (SHFR) scheme to learn the sparse transformations for HTL, by exploring the common underlying structures among multiple classes of two domains.
- (ii) Our major finding is that, according to the CS theory, a sparse feature mapping can be learned if and only if a sufficient number of classifiers are provided. Based on this observation, to accurately estimate the sparse feature mapping, we propose to generate sufficient classifiers under the **error correcting output correcting** (ECOC) scheme.
- (iii) We provide the first **generalization error bound** for two-class HTL problem. Furthermore, based on the analysis of binary case, we prove that the generalization error of multi-class HTL problems is affected by the number of tasks as well as sparsity of the feature mapping matrix $\mathbf{G}$.
- (iv) We propose a **batch-mode** pursuit method to efficiently solve the resultant non-negative sparse recovery problem. The computational cost of SHFR is independent of the data size, and scales linearly with the target feature dimensions.

The remainder of the chapter is organized as follows. In Section 3.2, we cast the HTL problem as a compressed sensing problem and present details of the proposed SHFR method. Section 3.3 presents a generalization error bound of our proposed method SHFR. In Section 3.4, we propose a batch-mode efficient algorithm to solve the formulated compressed sensing problem. In Section 3.5, experimental results on toy datasets as well as three real-world datasets are presented to demonstrate the effectiveness and efficiency of SHFR. We also perform empirical studies to further verify the proposed bound. Section 3.6 summarizes this work.

3.2 Sparse Heterogeneous Feature Representation

We study the HTL problem with one source domain and one target domain under the multiple class setting.

Given a binary task $t \in \{1, 2, \dots, n_c\}$, we assume that the classifier for either source or target domain is linear, 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. Since there are sufficient labeled data in the source domain, i.e., $\{(\mathbf{x}_{S_i}, y_{S_i})\}_{i=1}^{n_S}$, we can learn a robust set of weight vectors $\{\mathbf{w}_S^t\}_{t=1}^{n_c}$ for the source classifiers on specific learning tasks decomposed from the multi-class classification problems. Similarly, we can also build the corresponding set of weight vectors $\{\mathbf{w}_T^t\}_{t=1}^{n_c}$ with limited target labeled data $\{(\mathbf{x}_{T_i}, y_{T_i})\}_{i=1}^{n_T}$.

3.2.1 Feature Mapping Construction

Recall that, for HTL 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 [55, 81], a transformation matrix $\mathbf{G} \in \mathbb{R}^{d_T \times d_S}$ is introduced to learn the similarity $\mathbf{x}_{T_i}^\top \mathbf{G} \mathbf{x}_{S_i}$ between a source instance $\mathbf{x}_{S_i} \in \mathbb{R}^{d_S}$ and a target instance $\mathbf{x}_{T_i} \in \mathbb{R}^{d_T}$. Data points can be transformed from the source feature space to the target feature space via $\mathbf{x}_{T_i} = \mathbf{G} \mathbf{x}_{S_i}$. Equivalently, we can map the target domain data into the source domain via $\mathbf{G}^\top$ matrix [81]. However, we do not have the cross-domain instances correspondences for the general HTL setting. Therefore, instead of learning the transformation using input data directly, we adapt an idea from a multi-task learning method [2]¹. We 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.

To be specific, we either can learn the transformation matrix $\mathbf{G} \in \mathbb{R}^{d_T \times d_S}$ by maximizing the dependency between the transformed weight vectors of source classifiers and weight vectors of target classifiers expressed as $\max_{\mathbf{G}} \mathbf{w}_T^{t\top} \mathbf{G} \mathbf{w}_S^t$, or by minimizing the distance between these two vectors. For the simplicity of our theoretical analysis, we adopt the second way to learn $\mathbf{G}$. Specifically, given a binary task $t \in \{1, \dots, n_c\}$ and

¹Multi-task learning aims at learning the classifiers by sharing parameters among all tasks for multiple tasks, where each task has limited labeled data. The goal of HTL is to learn classifiers only for the target domain with limited target label data by leveraging the models learned from the source domain.

a transformation $\mathbf{G} \in \mathbb{R}^{d_T \times d_S}$, the relationship between the weight vectors of the source and target classifiers is modeled as ²

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

where $\mathbf{w}_\Delta^t$ is referred to as a “delta” weight vector, and its ℓ_2 -norm $\|\mathbf{w}_\Delta^t\|_2$ is used to measure the difference between the source weight vector $\mathbf{w}_T^t$ and the transformed target weight vector $\mathbf{G}\mathbf{w}_S^t$.

3.2.2 Proposed Formulation

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 learn the transformation $\mathbf{G}$ by minimizing $\|\mathbf{w}_\Delta^t\|_2$. Moreover, as mentioned in Section 3.1, the feature mapping $\mathbf{G}$ should be class-invariant and sparse. Finally, like the multi-language text classification problem mentioned in Figure 1, in most real-world applications, the linear relationship between two domains should be non-negative, e.g., nonnegative matrix factorization (NMF) [57], multiple kernel learning (MKL) [46].

Therefore, we can jointly optimize $\mathbf{G}$ over all the tasks by imposing non-negative sparsity constraints on $\mathbf{G}$. Specifically, let $\mathbf{G} = [\mathbf{g}_1, \mathbf{g}_2, \dots, \mathbf{g}_{d_T}]^\top$, by imposing ℓ_1 -regularization on $\mathbf{g}_i$ and non-negative constraint $\mathbf{g}_i \succeq \mathbf{0}$, the learning of $\mathbf{G}$ is formulated as the following nonnegative lasso problem [85, 106]:

$$\begin{aligned} \min_{\mathbf{G}} \quad & \frac{1}{n_c} \sum_{t=1}^{n_c} \|\mathbf{w}_T^t - \mathbf{G}\mathbf{w}_S^t\|_2^2 + \sum_i^{d_T} \lambda_i \|\mathbf{g}_i\|_1, \\ \text{s.t.} \quad & \mathbf{g}_i \succeq \mathbf{0}, \end{aligned} \tag{3.1}$$

where $\lambda_i > 0$ is the regularization parameter and the non-negativity constraints is used to preserve nonnegative linear relations of two features [6, 34]. Similar to NMF, we aim to approximate “input vectors” (target features) by the nonnegative linear combination ($\mathbf{G}$) of nonnegative “basis vectors” (source features). The nonnegative linear relationship

²For the nonlinear relationship between two feature spaces can be achieved by applying explicit feature mapping $\phi(\mathbf{x})$ on the data before training the weighted vectors for each domain. Nevertheless, in this work, we focus on the cross-language text classification, where linear relationship is more suitable to high-dimension text data.

can also be found in the MKL, where the target kernel is represented by the nonnegative linear combination of base kernels. In multilingual text classification, e.g. topic categorization, nonnegative linear combination in $\mathbf{G}$ can be deemed as the global “co-occurrence” between two wordbooks, as illustrated in the Figure 1.5. On the contrary, negative correlation is indirect and document-specific.

In (3.1), the first term minimizes the difference between $\mathbf{w}_T^t$ and $\mathbf{G}\mathbf{w}_S^t$ over all the n_c tasks and the second ℓ_1 -regularizer enforces the sparsity on each row of $\mathbf{G}$, respectively. Notice that once the set of classifiers in terms of weight vectors $\{\mathbf{w}_S^t\}_{t=1}^{n_c}$ and $\{\mathbf{w}_T^t\}_{t=1}^{n_c}$ are learned offline, one can learn G without directly using the domain data, which can significantly reduce the learning complexity. Our learning scheme is typically different from most of the existing HTL methods that require source domain data to be available for learning the feature mapping across domains. More discussion on the complexity analysis is found in Section 3.4.2. Moreover, similar to multi-task feature learning where each task is equipped with very limited labeled data, the knowledge is transferred across different binary classification tasks such that a more discriminative mapping is learnt. In Section 3.3.4, we theoretically show that as long as most base tasks are not badly designed, the generalization error of the classification model is small.

Furthermore, (3.1) can be rewritten in the following equivalent form,

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

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 problem can be further reformulated as the following CS problem:

$$\min_{\mathbf{g}_i} \sum_i^{d_T} \left(\frac{1}{n_c} \|\mathbf{b}_i - \mathbf{D}\mathbf{g}_i\|_2^2 + \lambda_i \|\mathbf{g}_i\|_1 \right), \quad \text{s.t. } \mathbf{g}_i \succeq \mathbf{0}, \quad (3.2)$$

where $\mathbf{b}_i$ is the concatenated row vector containing $w_{T_i}^t$ for all the n_c tasks, and $\mathbf{D} = [\mathbf{w}_1^t \ \mathbf{w}_2^t \ \cdots \ \mathbf{w}_{n_c}^t]^\top \in \mathbb{R}^{n_c \times d_S}$.

This CS problem has been well studied and a number of algorithms have been proposed to solve it, such as non-negative least squares (NNLS) [20], accelerated proximal

gradient algorithm (APG) [92], and orthogonal matching pursuit (OMP) [110]. In this work, we propose an efficient batch-mode algorithm to solve (3.2) to be introduced in detail in Section 3.4. After the sparse transformation G is learned, for any unseen test data $\mathbf{x}_T^*$ from the target domain, we can reuse the source domain classifiers to predict its label by using $y_T^* = F(\{(G\mathbf{w}_S^t)^\top \mathbf{x}_T^*\}_{t=1}^{n_c})$, where $F(\cdot)$ is a decision function that combines the predictive results of all n_c source classifiers to make a final prediction.

3.3 Generalization Error Bound for Multi-class HTL

The risk bounds for homogeneous transfer learning have been widely studied [11, 63]. One is interesting in understanding what conditions the algorithm learned from source domain can generalize well on the target domain. However, all the analysis of these bounds are restricted under the assumption that the features from source domain and target domain have the same number of dimension. Therefore, we cannot apply these theoretical analysis directly to the HTL setting. To the best of our knowledge, there is less research giving the bound analysis for HTL problems. Theoretical analysis for the HTL is necessary for us to understand the proposed HTL algorithm and the HTL problem setting. Different from existing risk bounds for homogeneous transfer learning, we obtain the generalization error bound of our proposed SHFR method mainly from the perspectives of ECOC and compressed sensing.

3.3.1 Estimation Error of Sparse Feature Mapping

The optimization problem (3.2) is composed of d_T lasso problems. In general, we have $n_c < d_S$ for relatively high-dimensional data. Therefore, (3.2) is an underdetermined linear system [35]. However, based on the compressive sensing theory, if $\mathbf{g}_i$ is sparse, it is possible to obtain a solution with sufficient measurements and a matrix $\mathbf{D}$ which satisfies certain conditions [19, 35]. In particular, one such condition is the sparse Riesz condition or RIP condition [19, 109], which requires that any two columns of $\mathbf{D}$ should be as perfectly incoherent as possible [19, 35, 109]. Whilst the RIP conditions for random design matrices have been thoroughly investigated [19], the RIP conditions for general

deterministic design matrices are more difficult. However, sparse Riesz condition for a deterministic design matrix can be evaluated by correlations between columns of dictionary matrix according to Proposition 1.

Proposition 1. [109] Let $\rho_{ji} = \mathbf{d}_j' \mathbf{d}_i / n_c$ denote the correlation between column $\mathbf{d}_j$ and $\mathbf{d}_i$ of standardized $\mathbf{D}$. For $A \subset \{1, 2, \dots, d_S\}$. If

$$\max_{|A|=k} \inf_{\alpha \geq 1} \left\{ \sum_{j \in A} \left(\sum_{i \in A, i \neq j} |\rho_{ji}|^{\alpha/(\alpha-1)} \right)^{\alpha-1} \right\}^{1/\alpha} \leq \delta < 1,$$

then the sparse Riesz condition holds for $\mathbf{D}$ with rank k .

Remark 1. Note that ρ_{ji} can be deemed as the correlation between columns $\mathbf{d}_j$ and $\mathbf{d}_i$. Therefore, sparse Riesz condition for a deterministic design matrix can be evaluated by correlations between columns of dictionary matrix. The smaller correlation $\sum |\rho_{ji}|$ (or δ) is, the more likely the sparse Riesz condition for $\mathbf{D}$ holds.

For simplicity in presentation, let k_i denote the number of non-sparse entries of $\mathbf{g}_i$, and $\hat{\mathbf{g}}_i$ denote an estimator of $\mathbf{g}_i$. According to Theorem 3 and Remark 4 in [109], under the above condition, the estimation error $\|\mathbf{g}_i - \hat{\mathbf{g}}_i\|_2$ for each independent subproblem can be bounded by $O\left(\sqrt{\frac{k_i \log d_S}{n_c}}\right)$, where d_S denotes the dimensionality of the source domain data and n_c is the number of tasks. Hence, we obtain the following lemma.

Lemma 1. Under the RIP conditions, the estimation error $\|\Delta \mathbf{G}_{CS}\|_F = \|\mathbf{G} - \hat{\mathbf{G}}\|_F$ in (3.1) is bounded by $O\left(d_T \sqrt{\frac{k \log d_S}{n_c}}\right)$, where k , d_S , d_T , n_c , $\hat{\mathbf{G}}$ denote the largest row non-sparsity degree of $\mathbf{G}$, the dimensionality of the source and target domain data, the number of tasks, and the estimator of $\mathbf{G}$, respectively.

According to Lemma 1 and Proposition 1, to reduce the reconstruction error of the feature mapping, one needs to 1) construct as many classifiers as possible to increase the number of measurements, i.e., n_c is large enough; and 2) ensure the columns of $\mathbf{D}$ constructed from classifiers to be incoherent, i.e., $\mathbf{D}$ is incoherent. Specifically, for multi-class classification problems with labels $\{1, 2, \dots, c\}$, one can generate c binary classifiers using the one-vs-all strategy [33]. However, when c is small, this strategy is not able to generate sufficient binary classifiers. Alternatively, one may use the one-vs-one strategy

to generate $c(c-1)/2$ binary classifiers. Yet, in this way, the generated classifiers may have large redundancy, i.e., some classifiers may be highly correlated to each other. To address the above issues, we propose to use the Error Correcting Output Codes (ECOC) scheme [33] to generate sufficient binary classifiers for feature mapping estimation. With proper design of ECOC coding matrix, one can guarantee more incoherency of $\mathbf{D}$ over one-vs-all or one-vs-one. In Section 3.5.5, we show empirically that $\mathbf{D}$ constructed using ECOC has more incoherency than using one-vs-all and one-vs-one coding schemes.

3.3.2 Generalization Error Bound for Binary Class HTL

Before deriving the generalization bound for multi-class HTL problems, we first analyze the error bound of the binary case. We follow but modify some related notations and definitions in [63] to formalize the HTL for binary classification as follows:

A domain is a tuple comprising of a distribution $\mathcal{D}$ on data $\mathcal{X}$ and a hypothesis $h: \mathcal{X} \rightarrow [-1, 1]$, where there exists one weight vector $\mathbf{w}$ corresponding to the hypothesis h so that $h(\mathbf{x}) = \mathbf{w}^\top \mathbf{x}$, and $\mathbf{x}$ is a data vector drawn from the distribution $\mathcal{D}$. The distribution of target domain is denoted as $\mathcal{D}_T$.

The amount of disagreement between the target hypothesis h_1 and another hypothesis h_2 is measured by

$$\epsilon_T(h_1, h_2) = E_{\mathbf{x}_T \sim \mathcal{D}_T} [|h_1(\mathbf{x}_T) - h_2(\mathbf{x}_T)|].$$

For simplicity, we use the shorthand

$$\begin{aligned} \epsilon_T(h) &= \epsilon_T(h, \mathbf{y}_T) \\ &= E_{\mathbf{x}_T \sim \mathcal{D}_T} [|h(\mathbf{x}_T) - \mathbf{y}_T(\mathbf{x}_T)|] \\ &= E_{\mathbf{x}_T \sim \mathcal{D}_T} [|\mathbf{w}_T^\top \mathbf{x}_T - \mathbf{y}_T(\mathbf{x}_T)|] = \epsilon_T(\mathbf{w}_T) \end{aligned}$$

to define the target domain error for any target hypothesis h and its corresponding weight vector $\mathbf{w}_T$, where $\mathbf{y}_T(\mathbf{x}_T)$ denotes the ground truth label of the instance $\mathbf{x}_T$.

Before presenting the bound analysis, we state three assumptions under our setting:

- (i) There exists a global transformation $\mathbf{G}^*$ between source domain and target domain, which minimizes the difference between ideal transformed source domain weight vector $\mathbf{G}^* \widetilde{\mathbf{w}}_S$ and ideal target domain weight vector $\widetilde{\mathbf{w}}_T$, namely

$$\|\Delta \widetilde{\mathbf{w}}\|_2 = \|\widetilde{\mathbf{w}}_T - \mathbf{G}^* \widetilde{\mathbf{w}}_S\|_2 \leq \delta_1,$$

where δ_1 is a very small constant.

- (ii) The difference between $\mathbf{w}_S$ learned from the source data and the ideal source domain weight vector $\widetilde{\mathbf{w}}_S$ is very small, namely

$$\|\Delta \mathbf{w}_S\| = \|\mathbf{w}_S - \widetilde{\mathbf{w}}_S\|_2 \leq \delta_2,$$

where δ_2 is a very small constant.

- (iii) The error of ideal target domain weight vector $\epsilon_T(\widetilde{\mathbf{w}}_T)$ is small.

The first assumption is used in other proposed HTL methods [28, 49, 55]. The second assumption is based on the fact that there are sufficient labeled data in the source domain.

By multiple use of triangle inequality, we prove the following generalization bound for binary class HTL.

Theorem 1. *Let h be the target hypothesis for a transformation matrix $\mathbf{G}$ that satisfies $\mathbf{w}_T = \mathbf{G} \mathbf{w}_S$ and $\widehat{\mathbf{G}}$ be an estimator of $\mathbf{G}$. Define $\Delta \mathbf{G}_{CS} = \widehat{\mathbf{G}} - \mathbf{G}$ and $\Delta \mathbf{G}_{ST} = \mathbf{G} - \mathbf{G}^*$. The expected error of our proposed SHFR on the target domain is:*

$$\begin{aligned} \epsilon_T(h) \leq & \epsilon_T(\widetilde{\mathbf{w}}_T) + \delta_1 E_{\mathcal{D}_T}(\|\mathbf{x}_T\|_2) + \delta_2 E_{\mathcal{D}_T}(\|\mathbf{G}^{*\top} \mathbf{x}_T\|_2) \\ & + \|\Delta \mathbf{G}_{CS}\|_2 \|\mathbf{w}_S\|_2 E_{\mathcal{D}_T}(\|\mathbf{x}_T\|_2) + \|\Delta \mathbf{G}_{ST} \mathbf{w}_S\|_2 E_{\mathcal{D}_T}(\|\mathbf{x}_T\|_2) \end{aligned} \quad (3.3)$$

where δ_1 and δ_2 are very small positive constants and the above three assumptions are satisfied.

Proof: First, let $\widehat{\mathbf{G}} = \mathbf{G}^* + \Delta \mathbf{G}$. By applying triangle inequality, the expected error of the hypothesis h

$$\begin{aligned} \epsilon_T(h) &= \epsilon_T((\mathbf{G}^* + \Delta \mathbf{G}) \mathbf{w}_S) \\ &\leq \epsilon_T(\mathbf{G}^* \mathbf{w}_S) + E_{\mathcal{D}_T}(|(\Delta \mathbf{G} \mathbf{w}_S)^\top \mathbf{x}_T|). \end{aligned} \quad (3.4)$$

By applying $\mathbf{w}_S = \widetilde{\mathbf{w}}_S + \Delta \mathbf{w}_S$ to the first term in (3.4),

$$\epsilon_T(h) \leq \epsilon_T(\mathbf{G}^* \widetilde{\mathbf{w}}_S) + E_{\mathcal{D}_T}(|(\mathbf{G}^* \Delta \mathbf{w}_S)^\top \mathbf{x}_T|) + E_{\mathcal{D}_T}(|(\Delta \mathbf{G} \mathbf{w}_S)^\top \mathbf{x}_T|).$$

By substituting $\mathbf{G}^* \widetilde{\mathbf{w}}_S = \widetilde{\mathbf{w}}_T - \Delta \widetilde{\mathbf{w}}$ and again applying triangle inequality,

$$\epsilon_T(h) \leq \epsilon_T(\widetilde{\mathbf{w}}_T) + E_{\mathcal{D}_T}(|\Delta \widetilde{\mathbf{w}}^\top \mathbf{x}_T|) + E_{\mathcal{D}_T}(|(\mathbf{G}^* \Delta \mathbf{w}_S)^\top \mathbf{x}_T|) + E_{\mathcal{D}_T}(|(\Delta \mathbf{G} \mathbf{w}_S)^\top \mathbf{x}_T|). \quad (3.5)$$

By applying Cauchy-Schwarz inequality (i.e., $|\langle \mathbf{x}, \mathbf{y} \rangle| \leq \|\mathbf{x}\|_2 \|\mathbf{y}\|_2$) and using assumptions (i) and (ii) on the right hand side of (3.5),

$$\epsilon_T(h) \leq \epsilon_T(\widetilde{\mathbf{w}}_T) + \delta_1 E_{\mathcal{D}_T}(\|\mathbf{x}_T\|_2) + \delta_2 E_{\mathcal{D}_T}(\|\mathbf{G}^{*\top} \mathbf{x}_T\|_2) + \|\Delta \mathbf{G} \mathbf{w}_S\|_2 E_{\mathcal{D}_T}(\|\mathbf{x}_T\|_2). \quad (3.6)$$

Finally, substituting $\Delta \mathbf{G} = (\widehat{\mathbf{G}} - \mathbf{G}) + (\mathbf{G} - \mathbf{G}^*) = \Delta \mathbf{G}_{CS} + \Delta \mathbf{G}_{ST}$ to the last term of (3.6),

$$\begin{aligned} \epsilon_T(h) &\leq \epsilon_T(\widetilde{\mathbf{w}}_T) + \delta_1 E_{\mathcal{D}_T}(\|\mathbf{x}_T\|_2) + \delta_2 E_{\mathcal{D}_T}(\|\mathbf{G}^{*\top} \mathbf{x}_T\|_2) \\ &\quad + \|\Delta \mathbf{G}_{CS} \mathbf{w}_S + \Delta \mathbf{G}_{ST} \mathbf{w}_S\|_2 E_{\mathcal{D}_T}(\|\mathbf{x}_T\|_2), \\ &\leq \epsilon_T(\widetilde{\mathbf{w}}_T) + \delta_1 E_{\mathcal{D}_T}(\|\mathbf{x}_T\|_2) + \delta_2 E_{\mathcal{D}_T}(\|\mathbf{G}^{*\top} \mathbf{x}_T\|_2) \\ &\quad + \|\Delta \mathbf{G}_{CS}\|_F \|\mathbf{w}_S\|_2 E_{\mathcal{D}_T}(\|\mathbf{x}_T\|_2) + \|\Delta \mathbf{G}_{ST} \mathbf{w}_S\|_2 E_{\mathcal{D}_T}(\|\mathbf{x}_T\|_2). \end{aligned}$$

□

Implications: The first three terms in Theorem 1 can be very small based on our assumptions that δ_1 and δ_2 are very small constants, and $\epsilon_T(\widetilde{\mathbf{w}}_T)$ is small. Therefore, whether the bound is tight or not mainly depends on how small the last term $\|\Delta \mathbf{G}\|_F$ is. Moreover, according to Lemma 1, the estimation error $\|\Delta \mathbf{G}_{CS}\|_F = \|\widehat{\mathbf{G}} - \mathbf{G}\|_F$ can be small if we create sufficient binary classification tasks and $\mathbf{G}$ is sparse. In binary classification, one only needs to construct one classifier (i.e., $n_c = 1$). In this case, the estimation error may not be small enough and hence the generalization bound may not be tight. However, in multi-class classification with c classes, there are several ways to generate a number of binary classifiers, such as the one-vs-all strategy or the ECOC scheme. By using different schemes to decompose a multi-class classification problem

to multiple binary classification problems, the forms of the generalization error bound are slightly different. Therefore, the bound mainly relies on how small $\|\Delta \mathbf{G}_{ST} \mathbf{w}_S\|_2 = \|(\mathbf{G} - \mathbf{G}^*) \mathbf{w}_S\|_2 \simeq \|\mathbf{w}_T - \widetilde{\mathbf{w}}_T\|_2$ (the last term) is. Due to limited labeled data from the target domain, some target domain weight vectors learnt from target domain data can be biased, which leads to large $\|\Delta \mathbf{G}_{ST} \mathbf{w}_S\|_2$. However, in our latter analysis (Section 3.5.6), we show that such bias can be reduced if we jointly learn the $\mathbf{G}$. Furthermore, Theorem 2 shows that as long as the averaged bias over all the binary tasks is not large, ECOC can control the prediction error to be small in high probability due to its error correcting capability.

3.3.3 Generalization Error Analysis Based on One-versus-all Scheme

Next, we extend the theoretical analysis in Section 3.3.2 to multi-class HTL setting. We analyze multi-class HTL error on the one-versus-all classification scheme. Since the one-versus-all scheme assigns the label of an instance to the class with the largest confidence value, we use the average loss to measure the multi-class error. The expected error is formulated as follows:

$$\begin{aligned} \epsilon_T^{OVA}(h^t) &= \frac{1}{n_c} \sum_{t=1}^{n_c} \epsilon_T(h^t) \\ &\leq \frac{1}{n_c} \sum_{t=1}^{n_c} \left\{ \epsilon_T(\widetilde{\mathbf{w}}_T) + \delta_1 E_{\mathcal{D}_T}(\|\mathbf{x}_T\|_2) + \delta_2 E_{\mathcal{D}_T}(\|\mathbf{G}^{*\top} \mathbf{x}_T\|_2) \right. \\ &\quad \left. + \|\Delta \mathbf{G}_{CS}\|_2 \|\mathbf{w}_S\|_2 E_{\mathcal{D}_T}(\|\mathbf{x}_T\|_2) + \|\Delta \mathbf{G}_{ST} \mathbf{w}_S\|_2 E_{\mathcal{D}_T}(\|\mathbf{x}_T\|_2) \right\} \end{aligned} \quad (3.7)$$

where n_c is the number of binary classifiers decomposed from a multi-class classification problem with c classes. Note that based on the one-vs-all scheme, the number of constructed binary classifiers is the same as the number of classes, i.e., $n_c = c$. When c is small, n_c may not be large enough to obtain small estimation error of the cross-domain feature mapping $\mathbf{G}$ based on Lemma 1. Alternatively, one can construct much more binary classifiers for the multi-class classification problem by using other schemes, such as the ECOC scheme. In the following section, we further analyze the generalization bound of SHFR for multi-class HTL based on ECOC.

3.3.4 Generalization Error Analysis Based on ECOC

Similar to using the one-vs-all classification scheme for multi-class HTL, the multi-class hypothesis h is composed of a set of binary classifiers $\{\hat{\mathbf{G}}\mathbf{w}_S^t\}_{t=1}^{n_c}$. However, different from the one-vs-all scheme, ECOC is more general and consists of two steps: encoding, i.e., to design a coding matrix M to induce binary classifiers, and decoding, i.e., to combine results of all the binary classifiers to make predictions based on the coding matrix. In order to analyze the generalization bound of SHFR for multi-class HTL based on ECOC, we need to first analyze the principles of ECOC for multi-class classification.

As discussed in some previous works [1, 33], the generalization error of a multi-class classification model learned based on the ECOC scheme depends on the design of the coding matrix. Specifically, given a coding matrix $M = \{-1, 0, +1\}^{c \times n_c}$, each class is associated with a row of the matrix M , and each binary classifier is associated with a column of the matrix M . If an entry $M(i, j) = +1$, then it implies that in classifier f_j , class i is considered as a positive class; if $M(i, j) = -1$, then it implies that in classifier f_j , class i is considered as a negative class; otherwise, class i is not taken into account for training and testing. In this way, either the one-vs-all scheme or the one-vs-one scheme can be considered as a special case of the ECOC scheme. One example of ECOC is illustrated in the Figure 3.1.

Binary Classification Tasks								
Classes	T1	T2	T3	T4	T5	T6	T7	T8
	y1	+1	0	+1	+1	+1	+1	-1
	y2	0	+1	+1	-1	-1	0	+1
	y3	+1	-1	0	+1	+1	-1	0
	y4	-1	+1	-1	+1	-1	+1	-1

Figure 3.1: Example of an ECOC matrix

As pointed out by Dietterich *et al.* [33], a coding matrix is good if each codeword, i.e., each row, is well-separated in Hamming distance from each of the other codewords, and each bit-position function f_i , i.e., each column, is uncorrelated with the functions to be learned for the other bit positions f_j , $j \neq i$. Moreover, the power of a coding matrix

M for multi-class classification can be measured by the minimum Hamming distance between any pair of its codewords [1], denoted by $\Delta_{\min}(M)$. For instance, for the one-vs-all code, $\Delta_{\min}(M) = 2$. For the one-vs-one code, $\Delta_{\min}(M) = \left(\binom{c}{2} - 1\right)/2 + 1$. On the other hand, for a random matrix with components chosen over $\{-1, 0, 1\}$, the expected value of $\Delta_{\min}(M)$ for any distinct pair of rows is $n_c/2$.

Here, we follow the proof procedure generalization error analysis for multi-class classification described in [1] to induce a generalization error bound for multi-class HTL. Given a coding matrix $M \in \{-1, 0, +1\}^{c \times n_c}$, where each row corresponds to a class and each column corresponds to a binary task. For some $t \in \{1, 2, \dots, n_c\}$, we denote $\mathbf{w}_S^t$ and h_S^t the source domain weight vector and the hypothesis of the t -th task defined by the coding matrix M , respectively. We denote $\hat{\mathbf{G}}\mathbf{w}_S^t$ and h_T^t the estimator of the target domain weight vector and hypothesis of the t -th task, respectively³. Moreover, we define the loss function $L(z) = |1 - z|$ in terms of the margin z . For example, given a target domain instance $\mathbf{x}_{T_i}$ with label y_{T_i} , the corresponding margin $z_{T_i}^t$ for t -th task is defined as $z_{T_i}^t = M(y_{T_i}, t)h_T^t(\mathbf{x}_{T_i}) = M(y_{T_i}, t)((\hat{\mathbf{G}}\mathbf{w}_S^t)^\top \mathbf{x}_{T_i})$.

Given a sample of test data, $\{(\mathbf{x}_{T_i}, y_{T_i})\}_{i=n_T+1}^{n_{Te}}$, we define the averaged loss $\hat{\varepsilon}$ of the n_c target domain hypotheses $h_T^1, \dots, h_T^{n_c}$ with respect to the coding matrix $M \in \{-1, 0, +1\}^{c \times n_c}$ as follows:

$$\begin{aligned} \hat{\varepsilon} &= \frac{1}{n_c} \sum_{t=1}^{n_c} \hat{L}_t \\ &= \frac{1}{n_c} \sum_{t=1}^{n_c} \frac{1}{n_{Te}} \sum_{i=1}^{n_{Te}} L(M(y_{T_i}, t)h_T^t(\mathbf{x}_{T_i})) \\ &= \frac{1}{n_c} \sum_{t=1}^{n_c} \frac{1}{n_{Te}} \sum_{i=1}^{n_{Te}} L(z_{T_i}^t), \end{aligned} \tag{3.8}$$

where $\hat{L}_t$ is the empirical estimation for the expectation of the binary absolute loss L_t on the target domain with respect to task t .

Note that the expectation of the binary absolute loss L_t on the target domain is equal

³Refer to Chapter 2 Section 2.1 for notation definitions.

to the expectation of test error ϵ_T^t with respect to task t :

$$\begin{aligned}
 \epsilon_T^t &= \epsilon_T(h^t) = E_{\mathbf{x} \sim D_T} |\mathbf{y}_T(\mathbf{x}_T) - h_T^t(\mathbf{x}_T)| \\
 &= E_{\mathbf{x}_T \sim D_T} |\mathbf{y}_T(\mathbf{x}_T) \mathbf{y}_T(\mathbf{x}_T) - \mathbf{y}_T(\mathbf{x}_T) h_T^t(\mathbf{x}_T)| \\
 &= E_{\mathbf{x} \sim D_T} |1 - \mathbf{y}_T(\mathbf{x}_T) h_T^t(\mathbf{x}_T)| \\
 &= E_{\mathbf{x} \sim D_T} L(\mathbf{y}_T(\mathbf{x}_T) h_T^t(\mathbf{x}_T)) = L_t,
 \end{aligned}$$

where $t = 1, \dots, n_c$.

Therefore, we obtain the following relationship in empirical estimation:

$$\hat{\epsilon} = \frac{1}{n_c} \sum_{t=1}^{n_c} \hat{L}_t = \frac{1}{n_c} \sum_{t=1}^{n_c} \hat{\epsilon}_T^t,$$

where $\hat{\epsilon}_T^t$ denotes the averaged test error on the t^{th} binary task.

Analogous to the proof of Theorem 1 in the [1], we define the average binary loss in (3.8) over the target test data, the estimator of feature mapping $\hat{\mathbf{G}}$ and the source domain hypotheses.

To understand the error correcting ability of ECOC, we first define the distance between the codes in any distinct pair of rows, $M(r_1)$ and $M(r_2)$, in the coding matrix M as $d(M(r_1), M(r_2))$. It is the sum of the n_c distances between two entries, $M(r_1, s)$ and $M(r_2, s)$ in the same column s at two different rows, r_1 and r_2 , i.e., $d(M(r_1), M(r_2)) = \sum_{s=1}^{n_c} d(M(r_1, s), M(r_2, s))$. We further define $\rho = \min_{r_1 \neq r_2} d(M(r_1), M(r_2))$ as the minimum distance between any two rows in M .

Proposition 2. *Given a coding matrix $M \in \{-1, 0, +1\}^{c \times n_c}$, a vector of predicted labels $f(\mathbf{x}) = [f_1(\mathbf{x}), \dots, f_{n_c}(\mathbf{x})]$ by n_c base classifiers for a test instance $\mathbf{x}$ and a convex loss function $L(z)$ in terms of margin z . If $\mathbf{x}$ is misclassified by the ECOC decoding, then*

$$L(M(y)f(\mathbf{x})) \geq \rho L(0), \tag{3.9}$$

where the corresponding correct label vector is denoted as $M(y)$.

In other words, the loss for predicted label vector $f(\mathbf{x})$ is greater than $\rho L(0)$.

Proof: Let $S_\Delta = \{s : M(r, s) \neq M(y, s) \wedge M(r, s) \neq 0 \wedge M(y, s) \wedge 0\}$ be the set of columns of M in which rows r and y differ and are both zero. Let $S_0 = \{s : M(r, s) = 0 \vee M(y, s) = 0\}$ be the set of columns in which either row r or row y is zero. Suppose

that the loss-based decoding incorrectly classifies a test instance $\mathbf{x}$ with known label y . There exists a label $r \neq y$ for which

$$d(M(y), f(\mathbf{x})) \geq d(M(r), f(\mathbf{x})).$$

Hence,

$$\sum_{t=1}^{n_c} L(z'_t) \geq \sum_{s=1}^{n_c} L(z_t). \quad (3.10)$$

In other words, the loss for $z_t = M(y, t)f_t(\mathbf{x})$ should be smaller than the loss for $z'_t = M(r, t)f_t(\mathbf{x})$.

Since $z_t = z'_t$ if $t \notin S_\Delta \cup S_0$, (3.10) implies $\sum_{s \in S_\Delta \cup S_0} L(z'_t) \geq \sum_{s \in S_\Delta \cup S_0} L(z_t)$. This, in turn, implies that

$$\begin{aligned} \sum_{t=1}^{n_c} L(z_t) &\geq \sum_{t \in S_\Delta \cup S_0} L(z_t) \\ &\geq \frac{1}{2} \sum_{t \in S_\Delta \cup S_0} (L(z_t) + L(z'_t)) \\ &= \frac{1}{2} \sum_{t \in S_\Delta} (L(z_t) + L(z'_t)) + \frac{1}{2} \sum_{t \in S_0} (L(z_t) + L(z'_t)) \end{aligned} \quad (3.11)$$

If $t \in S_\Delta$, then $z'_t = -z_s$ and $(L(-z_t) + L(z_t))/2 \geq L(0)$ due to the convexity. If $t \in S_0$, then either $z_t = 0$ or $z'_t = 0$. Hence, we obtain $L(z'_t) + L(z_t) \geq L(0)$ and second term of (3.11) is at least $L(0)|S_0|/2$. Thus,

$$\begin{aligned} \sum_{t=1}^{n_c} L(z_t) &\geq \frac{1}{2} \sum_{t \in S_\Delta} (L(z_t) + L(z'_t)) + \frac{1}{2} \sum_{t \in S_0} (L(z_t) + L(z'_t)) \\ &\geq L(0) \left(|S_\Delta| + \frac{|S_0|}{2} \right) \\ &= L(0)d(M(r), M(y)) \\ &\geq \rho L(0). \end{aligned}$$

□

Implication: From Proposition 2, one notes that a mistake on a test instance $(\mathbf{x}, y)$ implies that $L(M(y)f(\mathbf{x})) \geq \rho L(0)$. In other words, the prediction codes are not required to be exactly the same as true codes for all the base classifications. As long as the loss

for predicted labels is smaller than $\rho L(0)$, ECOOC can rectify the errors committed by the base classifiers and still give an accurate prediction. This error correcting ability is very important especially when the labeled data is insufficient.

Theorem 2. *Let $\bar{B}$ be the averaged bound of binary losses of target domain hypotheses $h_T^1, \dots, h_T^{n_c}$ on the transformed target test data $\{(\mathbf{x}_{T_i}, M(y_{T_i}, t))\}_{i=n_T+1}^{n_{Te}}\}_{t=1}^{n_c}$ with respect to the coding matrix $M \in \{-1, 0, +1\}^{c \times n_c}$, where c is the cardinality of label set, and a convex loss function be $L(z) = |1 - z|$. Then the multiclass HTL generalization error ϵ_T^{ECOOC} using sparse random encoding and loss-based decoding is at most $\frac{(\bar{B}+\eta)}{\rho L(0)}$ with at least probability of $1 - \frac{\sigma^2}{n_{Te}\eta^2}$.*

Proof: According to Proposition 2, for any misclassified testing instance $\mathbf{x}_T$, its loss $L(M(\mathbf{x}_T)y_T)$ satisfy the minimal lose $\rho L(0)$, i.e.,

$$L(M(y_T)h_T^t(\mathbf{x}_T)) = \sum_{t=1}^{n_c} L(M(y_T, t)h_T^t(\mathbf{x}_T)) \geq \rho L(0).$$

Let a be the number of incorrect predictions for a set of test instances of size n_{Te} . One obtains the following inequalities:

$$\begin{aligned} \frac{n_{Te}}{n_c} \sum_{t=1}^{n_c} \hat{\epsilon}_T^t &= \frac{n_{Te}}{n_c} \sum_{i=1}^{n_{Te}} \sum_{t=1}^{n_c} L(M(y_{T_i}, t)h_T^t(\mathbf{x}_{T_i})) \\ &\geq \frac{a}{n_{Te}n_c} \sum_{i=1}^{n_{Te}} \sum_{s=1}^{n_c} L(M(y_{T_i}, t)h_T^t(\mathbf{x}_{T_i})) \\ &\geq a\rho L(0). \end{aligned}$$

From Theorem 1, we know that the expected error of our proposed SHFR on the target domain is $\epsilon_T(h) \leq B_t$, where

$$\begin{aligned} B_t &= \epsilon_T(\widetilde{\mathbf{w}}_T^t) + \delta_1 E_{\mathcal{D}_T}(\|\mathbf{x}_T\|_2) + \delta_2 E_{\mathcal{D}_T}(\|\mathbf{G}^{*\top} \mathbf{x}_T\|_2) + \|\Delta \mathbf{G}_{CS}^t\|_F \|\mathbf{w}_S^t\|_2 E_{\mathcal{D}_T}(\|\mathbf{x}_T\|_2) \\ &\quad + \|\Delta \mathbf{G}_{ST}^t \mathbf{w}_S^t\|_2 E_{\mathcal{D}_T}(\|\mathbf{x}_T\|_2) \end{aligned}$$

for $t = 1, 2, \dots, n_c$. According to the law of large numbers, we have

$$P(\hat{\epsilon}_T^t < \epsilon_T^t + \eta) \geq 1 - \frac{\sigma^2}{n_T \eta^2},$$

where $\sigma = \max_t(\sigma_t)$ such that σ^t is the standard deviation of ϵ_T^t for $t = 1, 2, \dots, n_c$. Therefore, if $\bar{B} = \frac{1}{n_c} \sum_{t=1}^{n_c} B_t$,

$$n_{Te}(\bar{B} + \eta) = \frac{n_{Te}}{n_c} \sum_{t=1}^{n_c} (B_t + \eta) \geq \frac{n_{Te}}{n_c} \sum_{t=1}^{n_c} \hat{\epsilon}_T^t \geq a\rho L(0), \quad (3.12)$$

with a probability at least $1 - \frac{\sigma^2}{n_{Te}\eta^2}$.

In other words, given any η , the number of wrongly labeled testing instances $a \leq \frac{n_{Te}(\bar{B}+\eta)}{\rho L(0)}$ with at least probability of $1 - \frac{\sigma^2}{n_{Te}\eta^2}$. Hence, the testing error rate is bounded by $\frac{(\bar{B}+\eta)}{\rho L(0)}$ with at least probability of $1 - \frac{\sigma^2}{n_{Te}\eta^2}$. $\square$

Implication: The above error bound is usually small in this setting due to the following reasons:

- The source domain hypothesis h_S^t is well estimated when there are sufficient labeled data in the source domain.
- When there are sufficient classifiers and $\mathbf{G}$ satisfies the sparsity assumption, the reconstruction error is small and bounded according to the compressive sensing theory. According to Theorem 1, we know some B_t may be large due to inaccurate estimators $\mathbf{w}_T^t$. However, according to Theorem 2, as long as the averaged term $\bar{B}$ is small, we can still predict accurately with high probability. On the other hand, if most or all the created binary classifiers are badly designed, the proposed SHFR is not expected to achieve satisfactory performances.
- When the testing instances number n_{Te} is sufficiently large, the bound holds with high probability.

In summary, we show that the multi-class HTL error ϵ_T^{ECOC} based on the ECOC scheme has a linear relationship with the average of binary classification losses when using random encoding and loss-based decoding methods. Besides theoretical analysis, Pedrajas *et al.* [42] conducted comprehensive experiments and concluded that ECOC and one-vs-one are superior to other multi-class classification schemes, and they are the best choices for either powerful learners or simpler learners. Furthermore, Montazer *et al.* [65] claimed that, among several multi-class classification schemes, sparse ECOC is

better than dense ECOC; while dense ECOC is better than one-vs-one, and one-vs-all is the worst. Moreover, Ghani *et al.* [43] proved a theoretical bound that a randomly-construct binary matrix is not well row-separated with probability at most $1/c^4$, where c is the number of classes. Therefore, in our experimental evaluations, we adopt the sparse and random ECOC scheme for multi-class HTL problems.

3.3.5 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 [33]. 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 [31], where $\lfloor \cdot \rfloor$ denotes the round-up operator. In other words, even though there are misclassifications due to incorrect base classifiers, we can still achieve good performances by using the ECOC scheme. This property is particularly important for the proposed multi-class HTL method since some of the learned binary classifiers in the target domain may not be accurate or correct due to weak 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 good learners [43]. Hence, SHFR with ECOC learns a robust transformation matrix $\mathbf{G}$ even when there are some incorrect target binary classifiers. The robustness of SHFR with ECOC will be further verified through experiments, where we demonstrate that even with inaccurate target classifiers, the class-invariant $\mathbf{G}$ based on ECOC can greatly enhance the prediction accuracy for each binary classification task.

By combining all the above theoretical analysis, we conclude that the generalization error ϵ_T^{ECOC} of our proposed method SHFR for multi-class HTL problems is small and bounded with a proper ECOC coding matrix design.

3.4 An Efficient Batch Decoding Algorithm for SHFR

Though (3.2) which contains d_T lasso problems with non-negative constraints can be solved by many existing algorithms which are either adapted from NNLS solvers [20, 85]

or lasso solvers e.g., APG [92] and OMP [110]. The complexities of most of these methods is $O(n_c d_S)$ for each lasso problem. This limitation makes the problem intractable when the source-domain dimensionality is very high. To address the computation burden, we propose an efficient batch-model algorithm for solving (3.2), which extends the general matching pursuit algorithm (GMP) proposed by Tan *et al.* [90, 91].

3.4.1 Batch-Mode GMP (BGMP)

Recall that the learning of $\mathbf{g}_i$ can be cast as the following minimization problem:

$$\begin{aligned} \min_{\mathbf{g}_i} \quad & \frac{1}{n_c} \|\mathbf{b}_i - \mathbf{D}\mathbf{g}_i\|_2^2 + \lambda_i \|\mathbf{g}_i\|_1, \\ \text{s.t.} \quad & \mathbf{g}_i \succeq \mathbf{0}, \end{aligned} \tag{3.13}$$

where $\mathbf{b}_i \in \mathbb{R}^{n_c \times 1}$ is the concatenated row vector containing $w_{T_i}^t$ for all the n_c tasks, and $\mathbf{D} = [\mathbf{w}_S^1 \ \mathbf{w}_S^2 \ \cdots \ \mathbf{w}_S^{n_c}]^\top \in \mathbb{R}^{n_c \times d_S}$. Different from the standard lasso problem, non-negative constraints $\mathbf{g}_i \succeq \mathbf{0}$ are added on $\mathbf{g}_i$'s. These constraints can be easily handled by using GMP using the proposed batch-mode GMP (BGMP) (see Algorithm 1).

Basically, GMP solves (3.13) by decomposing the standard lasso problem into several small lasso subproblems through worst-case analysis [90, 91]. For simplicity and without loss of clarity, we drop the subscript i from $\mathbf{g}_i$ and $\mathbf{b}_i$ in our discussion. Given a residual $\boldsymbol{\alpha} = \mathbf{b} - \mathbf{D}\mathbf{g}$, where $\mathbf{g}$ is a solution of the previous iteration, as shown in [90, 91], to deal with the non-negative constraints $\mathbf{g}_i \succeq \mathbf{0}$, one only needs to select B columns of $\mathbf{D}$ that maximizes $\mathbf{D}^\top \boldsymbol{\alpha}$ rather than $|\mathbf{D}^\top \boldsymbol{\alpha}|$, where $|\mathbf{D}^\top \boldsymbol{\alpha}|$ denotes the absolute value of the vector $\mathbf{D}^\top \boldsymbol{\alpha}$. Therefore, one only needs to solve the following small subproblem:

$$\begin{aligned} \min_{\mathbf{g}, \boldsymbol{\xi}} \quad & \lambda \|\mathbf{g}\|_1 + \frac{1}{2} \|\boldsymbol{\xi}\|^2, \\ \text{s.t.} \quad & \boldsymbol{\xi} = \mathbf{b} - \mathbf{D}\mathbf{g}, \\ & \mathbf{g}_{\mathcal{I}_t^c} = \mathbf{0}, \end{aligned} \tag{3.14}$$

where $\mathcal{I}_t$ denotes the index set of the selected columns in $\mathbf{D}$. This subproblem can be solved efficiently using the proximal gradient (PG) method.

Note that the decomposed lasso subproblem scales at $O(kn_c)$, where k denotes the degree of sparsity [90, 91]. In GMP, each worst-case analysis step needs to calculate

Algorithm 1 BGMP for solving the non-negative lasso problem.

- Given $\mathbf{Q} = \mathbf{D}^\top \mathbf{D}$, calculating $[\mathbf{D}^\top \mathbf{b}]$, initializing $\boldsymbol{\alpha}^0 = \mathbf{b}$, $\mathcal{I}_0 = \emptyset$, and let $t = 1$.
- 1: Perform worst-case analysis: Let $\mathbf{q} = \mathbf{D}^\top \boldsymbol{\alpha}^{t-1} = [\mathbf{D}^\top \mathbf{b}] - \mathbf{Q}_{\mathcal{I}} \mathbf{g}_{\mathcal{I}}^{t-1}$; choose the B largest q_j and record their indices by $\mathcal{J}_t$. Let $\mathcal{I}_t = \mathcal{I}_{t-1} \cup \mathcal{J}_t$.
 - 2: Solve the subproblem (3.14) and update $\mathbf{g}_{\mathcal{I}_t}^t$.
 - 3: Let $\mathbf{g}_{\mathcal{I}_t^c}^t = \mathbf{0}$ and $\boldsymbol{\alpha}^t = \boldsymbol{\xi}$.
 - 4: Terminate if the stopping condition is achieved. Otherwise, let $t = t + 1$ and go to step 1.
-

a matrix-vector product $\mathbf{D}^\top \boldsymbol{\alpha}$, which scales at $O(d_S n_c)$, and therefore dominates the complexity of the whole algorithm if n_c is very large. To reduce this computational cost, we adopt the batch-mode GMP (BGMP) algorithm instead. Specifically, BGMP pre-computes $[\mathbf{D}^\top \mathbf{b}]$ and $\mathbf{Q} = \mathbf{D}^\top \mathbf{D}$, and stores them in the memory. Let $\mathcal{I}$ be index set of nonzero components in $\mathbf{g}$ of the previous iteration, $\mathbf{D}^\top \boldsymbol{\alpha}$ can be calculated efficiently by using $\mathbf{D}^\top \boldsymbol{\alpha} = [\mathbf{D}^\top \mathbf{b}] - [\mathbf{D}^\top \mathbf{D}_{\mathcal{I}}] \mathbf{g}_{\mathcal{I}} = [\mathbf{D}^\top \mathbf{b}] - \mathbf{Q}_{\mathcal{I}} \mathbf{g}_{\mathcal{I}}$, where $\mathbf{Q}_{\mathcal{I}}$ denotes the columns of $\mathbf{Q}$ indexed by $\mathcal{I}$. Note that the $\mathbf{Q}$ matrix is shared by all the binary tasks, and only required to be calculated once. Therefore, the computational cost can be neglected for large n_c . In this way, the computational cost for $\mathbf{D}^\top \boldsymbol{\alpha}$ is reduced to $O(d_S k)$, which save a lot of computational time when $k \ll d_S$ [90, 91]. With this strategy, when d_S is large, the computational cost is significantly reduced to $O(d_T n_c k)$.

3.4.2 Complexity Comparison

Here, we use linear SVMs to build the base classifiers for the source and target domains, which is obtained using the Liblinear solver [39] and pre-trained offline. That is to say, the learning of $\mathbf{G}$ is not related to the number of training instances. According to the analysis in Section 3.4, the computational cost of SHFR is $O(d_T n_c k)$, where the sparsity degree $k \ll d_S$. Compared to the state-of-the-art HTL algorithms, our proposed method is much more efficient. For instance, the MOMAP algorithm transforms the problem into c separate singular value decomposition (SVD) problems resulting in complexity of $O(4d_S^2 d_T + 8d_S d_T^2 + 9d_T^3)$. Heterogeneous Spectral Mapping (HeMap) method by standard eigenvalue decomposition has a complexity of $O((n_S + n_T)^3)$. The ARC-t method requires to solve an optimization problem that contains $n_S n_T$ constraints by applying an alternating projection method (i.g., Bregman's algorithm [21]). The HFA method

Table 3.1: Complexity Comparison for State-of-the-art HTL baselines.

Methods	Complexity
MOMAP	$O(4d_S^2d_T + 8d_Sd_T^2 + 9d_T^3)$
HeMap	$O((n_S + n_T)^3)$
DAMA	$O((d_S + d_T)^3)$
ARC-t	$O(d_Sd_Tn_Sn_T)$
HFA	$O((n_S + n_T)^3)$
SHFR	$O(d_Sn_ck)$

adopts an alternating projection method as well to solve a semidefinite program (SDP), 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. On the other hands, DAMA first constructs a series of combinatorial Laplacian matrices in $\mathbb{R}^{(d_S+d_T) \times (d_S+d_T)}$ and then solves 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. The theoretical complexity comparison is summarized in the Table 3.1.

3.5 Experimental Results

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 HTL problems due to the small size of labeled data in the target domain, which is still an open research issue in HTL. Therefore, we tune parameters of the comparison methods on a predefined range and report their best results on the test data, respectively. We use linear SVMs as base classifiers whose the regularization parameter C is cross-validated from $\{0.01, 0.1, 1, 10, 100\}$ on the source domain for all comparison methods. For our proposed method SHFR, we empirically set the maximum number of iterations to 20, $\lambda = 0.01$ and $B = 2$ in BGMP, C for each target classifier equal to the corresponding source classifier, and generate the ECOC “code-book” matrix using sparse random matrix [33].

3.5.1 Experiments on Synthetic Datasets

We first compare the performance of different HTL methods in terms of recovering a ground-truth feature mapping $\mathbf{G}$ 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 $\mathbf{X}_S \in \mathbb{R}^{150 \times 3,000}$. After that, we construct the ground-truth sparse feature mapping $\mathbf{G} \in \mathbb{R}^{100 \times 150}$ by using the following method: for each row i , we set $\mathbf{G}_{ij} = 1/5$, where $j = i, i + 1, \dots, i + 5$, and $\mathbf{G}_{ij} = 0$ otherwise. This generation of $\mathbf{G}$ implies that each target domain feature is represented by five source domain features. The ground-truth feature mapping is displayed in Figure 3.2.a, where the dark area represents the zero entries and the bright area denote values of $\mathbf{G}$. Finally, we construct the target domain data $\mathbf{X}_T \in \mathbb{R}^{100 \times 3,000}$ by using $\mathbf{X}_T = \mathbf{G}\mathbf{X}_S$. When conducting the experiment, we randomly select five instances per class from the target domain data $\mathbf{X}_T$ as labeled training data, and apply different HTL methods on them together with all the 3,000 source domain labeled data to recover the feature mapping $\mathbf{G}$.

In this experiment, the HTL methods DAMA and ARC-t are adopted as the baselines. For ease in comparison, we present the recovered matrix $\mathbf{G}$ for the three methods in Figures 3.2.b-3.2.d, respectively. From Figure 3.2.b, we observe that DAMA fails to recover the structure of $\mathbf{G}$; while ARC-t shows better performance. However, the recovered $\mathbf{G}$'s by these two methods are not sparse. However, according to Figure 3.2.d, SHFR perfectly recovers $\mathbf{G}$ with little noise. These experimental results demonstrate that by considering the sparsity constraints, SHFR recovers the feature mapping $\mathbf{G}$ more accurately.

3.5.2 Experiments on Real-world Datasets

We further 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.

3.2.a: Ground Truth

3.2.b: DAMA

3.2.c: ARC-t

3.2.d: SHFR

Figure 3.2: Illustration of the recovered feature mappings by using different methods on the toy data. The black color represents 0 and white color represents 1.

3.5.2.1 Datasets and Experimental Setup

Multilingual Reuters Collection⁴ 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 [37], 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 has very high dimensions,

⁴<http://multilingreuters.iit.nrc.ca/ReutersMultiLingualMultiView.htm>

and the baseline methods cannot handle such high-dimensional features. To conduct the comparison, we perform principal component analysis (PCA) with 60% energy preserved on the TF-IDF features. After PCA, we obtain 1,131 for the the English documents, 1,230 features for the 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 Collection⁵ was collected for multi-view learning where each instance is represented by several views. It was constructed from a single-view BBC corpora by splitting news article into related “views” of text and consists of 3 views. 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 such that other baselines can be applied. 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 rest of the target domain instances are considered as the test data.

Cross-lingual Sentiment Dataset⁶ 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 data domain data, respectively. After PCA, the reviews have 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 rest of the 5,970 target domain instances for testing.

3.5.2.2 Overall Comparison Results

Comparison results between SHFR and other baselines on the three real-world datasets are reported in Tables 3.2-3.4 respectively. From the tables, we observe that the SVMs conducted on a small number of target domain data only, denoted by SVM-T, using

⁵<http://mlg.ucd.ie/datasets/segment.html>

⁶<http://www.uni-weimar.de/cms/medien/webis/research/corpora/corpus-webis-cls-10.html>

Table 3.2: Multilingual Reuters Collection: comparison results in terms of classification acc (%).

Source Domain	SVM-T (1vsA)	SVM-T (1vs1)	SVM-T (ECOC)	DAMA	ARC-t	HFA	SHFR (1vs1)	SHFR (ECOC)
English	65.40±4.45	65.21±7.85	68.01±2.48	63.42±2.62	66.72±2.27	68.82±1.68	69.45±1.56	72.82±1.08
French				65.12±1.28	67.67±2.05	68.14±1.71	70.77±1.57	74.01±1.25
German				66.98±2.45	68.15±1.72	68.42±1.57	71.62±1.69	74.15±1.14
Italian				67.56±2.23	67.47±2.32	69.59±2.51	70.75±1.54	73.35±1.31

Table 3.3: BBC Collection: comparison results in terms of classification accuracy (%).

Source Domain	SVM-T (1vsA)	SVM-T (1vs1)	SVM-T (ECOC)	DAMA	ARC-t	HFA	SHFR (1vs1)	SHFR (ECOC)
View 1	73.35±4.98	68.78±11.7	76.45±5.47	67.42±2.25	75.78±2.69	72.45±10.65	89.81±1.20	90.57±1.49
View 2				66.87±1.73	74.53±2.48	73.75±8.06	88.92±1.67	91.85±0.96

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 3.3, 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 3.4, 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. Compared to one-vs-all and one-vs-one, ECOC shows much more stable performances and superior results on the most domains, which results from error correcting ability of ECOC when the labeled data size is small. The performance in terms of classification accuracy of the HTL baseline methods, DAMA, ARC-t and HFA, are comparable on the three datasets except for the BCC dataset. DAMA performs much worse than the other two. This may be because the performance of DAMA is sensitive to the intrinsic manifold structure of the data. If the manifold assumption does not hold on the data, the performance of DAMA drops significantly. On the other hands, SHFR using either the one-vs-one or ECOC scheme performs best on the three datasets. Moreover, using the ECOC scheme, SHFR can further improve the performance in terms of classification accuracy. Thus superior performances are benefited from both global transformation $\mathbf{G}$ and error correcting ability of ECOC. As discussed in Section 3.3.1, with more constructed binary tasks, the recovered feature mapping $\mathbf{G}$ tends to be more accurate.

Table 3.4: Cross-lingual Sentiment Dataset: comparison results in terms of classification accuracy(%).

Target Domain	SVM-T (1vsA)	SVM-T (1vs1)	SVM-T (ECOC)	DAMA	ARC-t	HFA	SHFR (1vs1)	SHFR (ECOC)
French	48.40 $\pm$ 3.45	58.30 $\pm$ 5.01	57.45 $\pm$ 3.17	55.18 $\pm$ 3.46	53.46 $\pm$ 5.18	56.46 $\pm$ 3.42	60.80 $\pm$ 3.08	62.12$\pm$2.61
German	49.78 $\pm$ 5.12	61.62 $\pm$ 6.31	62.18 $\pm$ 4.51	56.60 $\pm$ 3.72	57.29 $\pm$ 2.17	55.14 $\pm$ 3.15	63.85 $\pm$ 3.46	65.57$\pm$2.32
Japanese	48.25 $\pm$ 6.34	56.13 $\pm$ 4.81	57.27 $\pm$ 2.61	53.56 $\pm$ 2.67	55.75 $\pm$ 3.02	55.02 $\pm$ 4.67	59.67 $\pm$ 3.75	62.58$\pm$2.86

3.3.a: Multiclass accuracy v.s. target data size.

3.3.b: Multiclass accuracy v.s. dimension.

Figure 3.3: Comparison of different HTL methods under varying target labeled data size and dimensions.

3.5.3 Impact of Training Sample Size of Target Domain

We verify the impact of the labeled training sample size of the target domain to the overall HTL 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 Reuters dataset, where we use English as the source domain and Spanish as the target domain. The experimental results are shown in Figure 3.3.a. From the figure, we 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.

Based on Theorem 2, the overall generalization error of SHFR depends on the accuracy of the binary classifiers generated based on ECOC. To empirically demonstrate this

3.4.a: Averaged binary accuracy v.s. target data size.

3.4.b: Multiclass accuracy v.s. averaged binary accuracy.

Figure 3.4: Relationships between binary base classifiers and multi-class SHFR (ECOC).

point, we firstly report the averaged accuracy of all binary classifiers based on the ECOC scheme by varying the numbers of labeled training instance in the target domain in Figure 3.4.a. Moreover, we report the overall multi-class prediction accuracy using the averaged accuracy of all binary classifiers in Figure 3.4.b. From the figures, we see that more labeled target data leads to more precise target binary classifiers, and leads to improvement in multi-class HTL classification in terms of accuracy. Particularly, we observe from Figure 3.4.b that the multi-class classification accuracy is proportional to the averaged binary task accuracy. The reason is that more precise target binary classifiers generally induce better alignment between the target and source classifiers.

3.5.4 Comparisons on the same dimensional features

For fair comparison, we choose a small subset of important features in original spaces and apply this representation to all the methods. In the experiments, we select and fix 5000 most-frequent features in English domain of Reuters as the source domain. For the target domain Spanish, we vary its dimension in the range $\{500, 1000, 2000, 3000, 5000\}$ by choosing the corresponding number of most-frequent features. The results are shown in the Figure 3.3.b. From the results, we observe that SHFR consistently outperform other baselines, especially when the dimension is low.

Table 3.5: Comparison of Correlation $\sum_{ji} |\rho_{ji}|$.

Coding Scheme	English	French	German	Italian
1 V A	0.8532 ± 0.1358	0.7557 ± 0.1289	0.2738 ± 0.0245	0.9331 ± 0.5701
1 V 1	0.3374 ± 0.0290	0.2509 ± 0.0576	0.1168 ± 0.0110	0.3322 ± 0.1679
ECOC	0.1349 ± 0.0142	0.1478 ± 0.0225	0.0732 ± 0.0145	0.1725 ± 0.1047

Table 3.6: Comparison results on each binary task in terms of classification accuracy (%).

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

3.5.5 Incoherent Dictionary $\mathbf{D}$ construction

In this experiment, we adopt the summation of absolute value of all pairs of columns in the dictionary $\mathbf{D}$ ($\sum_{ji} |\rho_{ji}|$) to evaluate its coherence degree. In other words, the smaller value $\sum_{ji} |\rho_{ji}|$ is, the more incoherent dictionary is [109]. The results of different coding schemes on the Reuters dataset are summarized in Table 3.5. From the table, we observe that ECOC significantly reduces the coherence degree to better construct the dictionary for the formulated compressed sensing problem.

3.5.6 Error Correction Through Learning $\mathbf{G}$

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 $\mathbf{G}$. To verify that SHFR can correct bias in 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 for the Reuters dataset in Table 3.6, where each column corresponds to a binary task, indexed by $k \in \{1, \dots, 15\}$. From the table, we observe that the predictions of SVM-T on some binary tasks are inaccurate due to limited labeled data. The accuracies are even below 50% (numbers in boldface). However,

by learning the transformation $\mathbf{G}$, we are able to reduce the bias of the weak binary classifiers, and increase over 30% in accuracy on average. Experiments on the other two datasets also exhibit similar results. 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 3.2-3.4 and Figure 3.3.a. ARC-t aims to align the target data with source data via a transformation $\mathbf{G}$ which is learned from some similarity and dissimilarity constraints. The constraints are constructed from plenty of source-domain labeled data and a few target-domain labeled data. Although the number of target-domain labeled data is limited, the performance of ARC-t is still acceptable. Different from ARC-t, SHFR tries to align the target classifiers with source classifiers using transformation G , which is shared by all induced binary tasks or classifiers. In other words, G is estimated through all classifiers. By borrowing the idea from multi-task feature learning which jointly optimizes all classifiers to learn a global feature transformation $\mathbf{G}$, we are still able to estimate a stable and precise G for HTL. Furthermore, as discussed in Section 3.3.5, ECOC with a well designed coding matrix has a powerful ability to correct errors as shown in Table 3.6.

3.5.7 Impact of Dichotomizer Size on Estimation Error

As proven in Section 3.3, the generalization error of SHFR for multi-class HTL depends on the estimation error of $\mathbf{G}$, which relies on two factors: the number of classifiers (measurements) and the sparsity of $\mathbf{G}$. When $\mathbf{G}$ is sparse and the constructed binary classifiers are sufficient, one can recover a precise $\mathbf{G}$ by using the dictionary constructed by $\mathbf{w}_S$.

To demonstrate how the generalization error of SHFR changes with the number of dichotomizers or classifiers, we conduct experiments on the Reuters dataset. Experimental results are showed in Figure 3.5.a. From the figure we observe that when more dichotomizers are constructed, the predictions are more accurate in the target domain. This verifies that more dichotomizers provides more information to recover the feature mapping $\mathbf{G}$. Furthermore, the standard deviation is decreasing with increasing number

3.5.a: Accuracy v.s. dichotomizers size.

3.5.b: Accuracy v.s. dichotomizers size on removing negative alignments.

Figure 3.5: The performances of SHFR under (i) ECOC and (ii) ECOC with removing negative alignments.

of dichotomizers. However, as observed from the figure, the multi-class accuracy does not increase any more when the number of classifiers reaches 31. There are two reasons for this observation: 1) the redundant in information among the constructed binary classifiers prevents improvement in estimating $\mathbf{G}$; and 2) according to Pedrajas *et al.* [42], when there are too many dichotomizers, the minimum distance in ECOC becomes small, which decreases the error correcting ability. In general, SHFR obtains better and more stable performance in terms of classification accuracy with increasing number of dichotomizers.

3.5.8 Impacts of Nonnegative Constraints and ℓ_1 Norm Regularization

Similar to the motivation of nonnegative matrix factorization (NMF), we aim to approximate “input vectors” (target features) by the non-negative linear combination ($\mathbf{G}$) of nonnegative “basis vectors” (source features). The nonnegative linear relationship can also find in the multiple kernel learning, where the target kernel is approximated by the non-negative linear combination of base kernels. In multilingual text classification, e.g topic categorization, non-negative linear combination in $\mathbf{G}$ can be deemed as

Table 3.7: Impact of Nonnegative Constraints and L_1 Norm Regularization: comparison among different solvers results in terms of classification accuracy (%).

Dataset	Domain	SHFR-LASSO	SHFR-NNLS	SHFR
Reuters	English	67.08±1.63	71.47±1.56	72.82±1.08
	French	67.17±1.26	72.08±1.19	74.01±1.25
	German	69.56±1.28	73.62±1.75	74.15±1.14
	Italian	68.19±2.51	71.98±1.78	73.35±1.31
BBC	View 1	84.72±2.18	88.66±1.29	90.57±1.49
	View 2	87.06±2.01	90.25±1.43	91.85±0.96
Sentiment	French	56.12±2.30	60.17±2.05	62.12±2.61
	German	58.83±2.24	63.47±1.49	65.57±2.32
	Japanese	56.72±1.81	61.51 ±1.76	62.58±2.86

global “co-occurrence” between two wordbooks, as illustrated in the Figure 1.5. On the contrary, negative correlation is indirect and document-specific. To verify the impact of nonnegative constraints, we report the results of SHFR without positive constraints (SHFR-LASSO) for comparison. In this case, the model reduces into typical lasso problem.

On the other hand, some studies show that the Non-Negative Least Squares (NNLS) indeed can achieve sparse recovery of non-negative signals in a noiseless setting [18, 36, 98, 99]. However, this does not hold for realistic noisy setup. As mentioned in [86], the fact that NNLS, apart from sign constraints, only consisting of a fitting term, fosters the intuition that NNLS is prone to overfitting, specifically in a high-dimensional setting. Therefore, the ℓ_1 -norm regularizer is necessary to prevent over-adaptation to noise and to enforce desired structural properties of the solution, like sparsity of the vector of coefficients [86]. In our case, the linear regression model is in high-dimension, and the constructed model contains a lot of noises due to the improper design of some binary tasks, NNLS is insufficient to achieve desirable solutions. Besides, Theorem 1 in [85] states that only the matrix satisfies the self-regularizing property, exact sparsity (i.e., ℓ_1 -norm) is not needed. Furthermore, ℓ_1 -norm is stronger regularizer compared to nonnegative constraint. Sparser solution can be achieved with ℓ_1 -norm regularization. For example, in Reuters, the averaged non-sparsity degrees for SHFR and SHFR-NNLS are 0.016 and 0.020. Furthermore, nonnegative lasso allows more flexibility in controlling the sparsity by choosing different parameter λ . To verify the necessity of ℓ_1 -norm regularizer, we

compare SHFR with one NNLS solver [20]. In practice, the nonnegative lasso problem is well studied and has many applications such as feature reduction [97] and index tracking [103]. Nevertheless, solving the nonnegative lasso problem is not the focus of our paper. For simplicity, in this paper, we address this problem by modifying the method in Tan’s work [90, 91], which has shown superior performances over many state-of-the-art solvers.

The results on three multilingual datasets are summarized in the Table 3.7. From the table we observe that SHFR-LASSO performs the worst and SHFR-NNLS can enhance its performances through the non-negative constraint. It is likely due to feature mapping $\mathbf{G}$ is global and features are naturally nonnegative for text classification. Furthermore, we notice that SHFR that explicitly imposes both the non-negativity and sparsity in learning $\mathbf{G}$ further enhances the performances by avoiding over-adaptation to noises in high-dimensional system.

3.5.9 Experiments on Removing Negative Alignment

From Table 3.6, we observe that for some constructed binary tasks, the accuracies of SHFR slightly decrease compared to those of SVM-T after applying the transformation estimator $\hat{\mathbf{G}}$, which is referred to as *negative transfer*. The negative transfer problem of SHFR is mainly caused by two aspects:

- (i) *Negative alignments* between the source and target classifiers would happen when the classification accuracies of the corresponding target classifiers are below 50%. This is due to the fact that some initial target classifiers are very unstable due to the limited target-domain labeled data. If we learn the transformation $\mathbf{G}$ by encoding the bad target classifiers to align the corresponding source classifiers, we would not be able to achieve good performance in the final multi-class classification problem.
- (ii) The transformation $\mathbf{G}$ is class-invariant. Hence, $\mathbf{G}$ is supposed to generalize well on all classifiers. $\mathbf{G}$ aims to correct the bias of the negative transfers. However, it may smooth the strength of good classifiers.

We study how the performances are affected by removing negative alignments, and only using the good target classifiers whose accuracies are above 50%, which is evaluated

Table 3.8: Comparison results on each binary task after removing negative alignment in terms of classification accuracy (%).

Binary Classifiers	1	2	3	4	5	6	7	8
SVM-T	74.42	67.98	85.67	82.53	43.32	75.95	66.30	73.92
SHFR-RN	85.39	83.26	86.45	92.43	83.79	78.19	77.45	81.28
Difference	10.97	15.28	0.87	9.90	40.47	2.24	11.15	7.36
Binary Classifiers	9	10	11	12	13	14	15	
SVM-T	42.78	76.56	78.32	65.43	72.74	46.56	49.71	
SHFR-RN	84.15	78.45	92.14	82.12	84.58	82.72	88.61	
Difference	41.37	1.89	13.82	16.69	11.84	36.16	38.90	

on test data. To be more specific, we investigate whether dropping those negative alignments can improve accuracy on all the binary tasks as well as the final multi-class HTL classification problem. Experimental results are showed in Table 3.8 and Figure 3.5.b. Compared to results in Table 3.6, from Table 3.8, we find that by removing the negative alignments we improve all the binary classification performance in terms of accuracy, and thus avoid negative transfers. Furthermore, we sort the binary tasks created from ECOC according to the accuracies evaluated on the test data and remove the bad designed tasks. From Figure 3.5.b, we observe that when the number of dichotomizers is small (i.e., less than 30), SHFR with removing negative alignment achieves better performance in terms of multi-class classification accuracy. However, we understand the difficulty of evaluating target classifiers in practice due to scarce target labeled data. Interestingly, we also find that when the number of dichotomizers is large, the classification accuracies of both SHFR with and without removing negative alignments converge to the same value. It results from the error correction ability of ECOC in multi-class HTL problem.

In summary, SHFR with Removing Negative alignments (SHFR-RN) enhances the performance in terms of classification accuracy when there are insufficient classifiers. However, the difficulty of evaluating alignments in practice makes SHFR-RN impractical due to small target labeled data size. Fortunately, the performance of SHFR remain the same as SHFR-RN when there are enough dichotomizers, which makes SHFR more practicable in the multi-class HTL problem.

3.5.10 Training Time Comparison

Finally, to verify the computational efficiency of SHFR, we conduct two experiments to compare the training time in learning the feature mapping between SHFR and other baselines under varying numbers of data dimensions and data instances respectively. These experiments are conducted on a toy dataset, where the dimensions of the source and target data are constructed to be the same. In the first experiment as shown in Figure 3.6.a, we fix the number of training data (including both source and target domain data) to 3,000 and vary the data dimensions in the range $\{100, 200, 500, 1000, 2000, 5000\}$. We can observe that overall SHFR performs faster than all the baselines. Its training time increases only linear with the number of data dimensions. Its training time increases faster with larger numbers of data dimensions. This is because that the time complexity of SHFR has a linear relationship with the number of the target-domain dimensions. Furthermore, the training time of DAMA increases dramatically when the number of data dimensions increases. In contrast, the training time of HFA and ARC-t does not increase much when then number of data dimensions increase. This is because that these two methods adopt the kernel technique such that the computational time depends on the number of data instances instead of data dimensions. In the second experiment as shown in Figure 3.6.b, we fix the data dimension to 1,000 under varying the number of training instances in the range $\{100, 500, 1000, 2000, 5000\}$. From the figure, we can observe that the training time of SHFR change slightly when the number of training instances increases. However, the training time of ARC-t and HFA increase polynomially when the number of training instances increases. These two experiments verify the superiority of SHFR in computational efficiency compared with the baseline methods.

3.6 Summary

In this chapter, we propose a **sparse heterogeneous feature representation** (SHFR) scheme to learn the sparse transformations for HTL, by exploring the common underlying structures among multiple classes of two domains for the multi-class HTL problems. In particular, SHFR exploits the sparsity and class-invariance in the feature mapping. For SHFR, the learning of feature mapping can be cast as a Compressive Sensing (CS)

3.6.a: Training time v.s. data dimensions

3.6.b: Training time v.s. data size

Figure 3.6: Time complexity comparison

problem. Based on the CS theory, we show that the number of binary learning tasks affects the multi-class HTL performance. In addition, the proposed method has superior scalability over other methods. Extensive experiments demonstrate the effectiveness, efficiency, and stability of SHFR.

Chapter 4

Hybrid Heterogeneous Transfer Learning through Deep Learning

As discussed in Chapter 2, most of the existing approaches in heterogeneous transfer learning aim to learn a feature mapping across heterogeneous feature spaces based on some cross-domain correspondences constructed by labels. With the learned feature mapping, instances are mapped from the target domain to the source domain or the other way around. In this way, source domain labelled data can be used to learn an accurate classifier for the target domain.

In this chapter, we focus on HTL setting given correspondences. A common assumption behind correspondence-based HTL methods is that the selection of instance-correspondences to learn feature mapping is *unbiased*. In other words, the corresponding instances across domains are assumed to be representative in the source and the target domains, respectively. However, in many real-world scenarios, this assumption may not hold. The selected corresponding instances may be biased and not be able to represent either the source or the target domain data.

Taking cross-language document classification as a motivating example illustrated in Figure 1.7.a, the objective is to learn a text classifier on documents in one language (e.g., German) with the help of only a set of annotated English documents. To apply HTL methods to solve this task, one can simply construct German-English document-correspondences by translating some German documents into English using Google Translate. However, the *wordbook* of the translated English documents may

be quite different from that of the original English documents. For instance, the German word “betonen” is translated into the English word “emphasize” by Google Translate. However, in an original English document, its corresponding word is “highlight” or “stress”. This is referred to as the “bias” issue in word-distribution between the translated documents and the original ones in the target domain. In this case, a feature mapping learned on such correspondences may not be effective.

As a second example (shown in Figure 1.7.b), we consider a multilingual sentiment classification task, where our task is to automatically classify overall sentiment polarities of songs reviews in German given labelled books reviews in English, and some unlabelled pairs of songs reviews in German and their English translations. Though one can make use of the German-English review-correspondences to learn a feature mapping from English to German, the labelled book reviews in English after transformation may not be useful to learn an accurate sentiment classifier for songs reviews in German. The reason is that opinion and topic words used on different types of products can be very different [62]. In the context of this example, the “bias” between the transformed source domain data and the target domain data is caused by the difference in product domains.

Motivated by the above observations, we propose a new HTL framework called “Hybrid Heterogeneous Transfer Learning” (HHTL) to ensure knowledge transfer across heterogeneous domains to be effective even though the cross-domain correspondences are biased. Specifically, the HHTL framework consists of two main components:

- (i) learn heterogeneous feature mapping between the source domain labelled data and the target unlabelled domain, and
- (ii) discover a latent representation to reduce the distribution bias between the transformed target domain unlabelled data and source domain labelled data.

Thereafter, standard classification methods can be applied on the source domain labelled data with the latent representation to build a target classifier effectively. We present a deep learning method to simultaneously learn

- a feature transformation from the target domain to the source domain, and

- two mappings from low-level features to high-level features in the source and target domains.

Note that the proposed HHTL framework is different from multi-view learning [14], where full correspondences between two views of data are required, and corresponding labels of the correspondences are assumed to be available. In HHTL, no label information of the cross-domain instance-correspondences is required. Moreover, in HHTL labeled data is only assumed to be available in the source domain, while the goal is to learn a classifier to be used in the target domain.

The remainder of the chapter is organized as follows. First of all, we briefly review some existing work related to HHTL in Section 4.1. In Section 4.3, we present our proposed deep learning approach for HHTL. In Section 4.4, we conduct extensive experiments on the cross-language sentiment dataset to demonstrate the importance of HHTL and its superiority over other state-of-the-art methods. Furthermore, parameter sensitivity analysis demonstrates that the performance of our method is insensitive to its parameters.

4.1 Related Work

“Hybrid” in HHTL refers to the combination of homogeneous transfer learning and heterogeneous transfer learning. Learning a good feature representation is crucial to both homogeneous transfer learning and heterogeneous transfer learning. With better feature representation for two domains, one learns better feature mappings across heterogeneous feature spaces. We first review some feature representation methods for homogeneous transfer learning and then introduce related heterogeneous transfer learning algorithms.

Homogeneous transfer learning aims to improve models’ generalization ability across different domains of the same feature space. Most algorithms can be categorized into two groups [70]: 1) feature-representation-transfer approach and 2) instance-transfer approach. However, feature-representation transfer approaches show more promising results than the instance-transfer approaches, especially when deep learning approaches are used [2–4, 25, 26, 44, 60, 80]. It can be attributed to the fact that all the instance-transfer

approaches assume that the given features are biased but contain no noise. If the given features are too noisy, reweighing the instances from source domain to match target domain is usually of little useful. Thus, we describe only feature-representation transfer approaches in this section. Pan *et al.* [69] proposed Transfer Component Analysis (TCA), a dimensionality reduction-based transfer learning method, to learn a low-dimensional space where the distance in distributions between a source domain and a target domain can be reduced. Raina *et al.* [77] proposed a self-taught learning framework based on sparse coding to learn high-level features for transfer learning. More recently, deep learning techniques have been proposed for transfer learning which have better performances on various transfer learning tasks. Glorot *et al.* [44] applied stack denoised autoencoder (SDA) to learn hidden feature representations for cross-domain sentiment classification. In a follow-up work, Chen *et al.* [26] proposed marginalized SDA (mSDA), a variation of SDA for transfer learning, which has been shown to be effective and efficient. Both SDA and mSDA show that deep feature representations reduce the domain differences in terms of $\mathcal{A}$ distances [11].

Heterogeneous transfer learning aims to transfer knowledge across different feature spaces. Most methods for heterogeneous transfer learning aim to learn a common feature representation based on some correspondences between domains such that both source and target domain data can be represented by homogeneous features. Specifically, one learns two different feature mappings to transform the source domain and the target domain data to a latent feature space [37, 75, 84, 96]. Alternatively, one learns an asymmetric transformation to map data from one domain to another domain [55, 115]. Moreover, there are many methods proposed to learn the mappings by incorporating instance correspondences between domains [28, 95]. Deep learning techniques have also been proposed in heterogeneous transfer learning [87], where knowledge is transferred from text to image based on a lot of text-image correspondences through deep learning. Multimodal deep learning [66] extends the sparse restricted Boltzmann machines [58] by adding an additional shared representation layer to transfer features between two heterogeneous spaces and shows that the source domain improves feature learning in target domain. Our proposed framework can be considered as a more general solution for heterogeneous transfer learning, where the bias of the correspondences between the source and target domains is considered.

4.2 Problem Statement and Preliminaries

Given a set of target domain unlabeled data $\mathbf{D}_T = \{\mathbf{x}_{T_i}\}_{i=1}^{n_1}$, a set of source domain labeled data $\mathbf{D}_S = \{(\mathbf{x}_{S_i}, y_{S_i})\}_{i=1}^{n_2}$, and an additional set of pairs of source and target domain unlabeled data, namely parallel data, $\mathbf{D}_C = \{(\mathbf{x}_{S_i}^{(c)}, \mathbf{x}_{T_i}^{(c)})\}_{i=1}^{n_c}$, where $\mathbf{x}_{S_i}$ or $\mathbf{x}_{S_i}^{(c)}$ is in $\mathbb{R}^{d_S \times 1}$, and $\mathbf{x}_{T_i}$ or $\mathbf{x}_{T_i}^{(c)}$ is in $\mathbb{R}^{d_T \times 1}$. Our objective is to learn weight matrices $\mathbf{W}_S$ and $\mathbf{W}_T$ to project the source domain labeled data and the target domain unlabeled data to hidden representations, $\mathbf{W}_S \mathbf{X}_S$ and $\mathbf{W}_T \mathbf{X}_T$, respectively, and a feature mapping $\mathbf{G}$ to map data from the target domain feature space to the source domain feature space¹, such that the difference between the original source domain data and the transferred target domain data is small. With the learned weight matrices $\mathbf{W}_S$, $\mathbf{W}_T$ and the feature mapping $\mathbf{G}$, one trains a classifier f from $\{(\mathbf{W}_S \mathbf{X}_{S_i}, y_{S_i})\}_{i=1}^{n_1}$, and predicts the labels of target domain unlabeled data $\mathbf{x}_T^*$ by applying $f(\mathbf{G}(\mathbf{W}_T \mathbf{X}_T^*))$.

In recent years, deep learning approaches has shown a great success in learning better features for transfer learning [26, 44, 58, 66, 87]. Among these approaches, autoencoder shows its superiority in many learning scenarios and application domains related to text and audio classification [51]. It takes an input $\mathbf{x} \in \mathbb{R}^d$ and first maps it with an *encoder* to a hidden representation $\mathbf{h}$ through a deterministic mapping $\mathbf{W}$, e.g.:

$$\mathbf{h} = s(\mathbf{W}\mathbf{x} + \mathbf{b}), \quad (4.1)$$

where s is a non-linearity such as the sigmoid and $\mathbf{b}$ is the bias vector. The parameters of this model, i.e., $\mathbf{W}$, $\mathbf{b}$ are optimized by minimizing the average reconstruction error as follows,

$$\min_{\mathbf{W}, \mathbf{b}} \|\mathbf{x} - \mathbf{h}\|^2. \quad (4.2)$$

However autoencoder suffers from learning hidden features over fitting to input features [9].

To avoid this problem, Denoising Autoencoders (DA) [93] was proposed to train autoencoder such that input can be reconstructed from a corrupted version of it. It can be formally expressed as follows:

$$\mathbf{h} = s(\mathbf{W}\tilde{\mathbf{x}} + \mathbf{b}), \quad (4.3)$$

¹Alternatively, one can learn a feature mapping $\mathbf{G}^\top$ to map data from the source domain to the target domain.

where $\tilde{\mathbf{x}}$ is a random corrupted version of input $\mathbf{x}$. Similarly, the parameters of DA, i.e., $\mathbf{W}, \mathbf{b}$ are optimized by minimizing the average reconstruction square error (i.e., Equation 4.2). The denoisers can be further stacked into deep learning architectures where the weights are fine-tuned with back-propagation, which is denoted as Stacked Denoising Autoencoders (SDA) [93]. In [93], the stochastic corruption process randomly sets some of the inputs to zero. Hence the SDA is trying to predict the corrupted (i.e. missing) values from the uncorrupted (i.e., non-missing) values, for randomly selected subsets of missing patterns. In order to learn robust and overcomplete features, Vincent *et al.* [93] advised to create as many corrupted versions of input as possible. On the other hand, the high computational cost will be incurred by the dimensionality of the input data and the need for computationally intensive model selection procedures to tune hyperparameters [26].

Fortunately, marginalized Stacked Denoising Autoencoder (mSDA) [26] was proposed to solve this computation challenge. The key difference between mSDA and SDA is that mSDA introduces the feature corruption probability p to simulate the random feature corruption, which results in close-form solutions for the optimal parameters. Specifically, the mSDA is formulated as follow,

$$\min_{\mathbf{W}, \mathbf{b}} = \|\mathbf{x} - s(\mathbf{W}\tilde{\mathbf{x}} + \mathbf{b})\|^2, \quad (4.4)$$

where $\tilde{\mathbf{x}} = \mathbf{x} \circ \mathbf{p}$ and $\mathbf{p} = [p, p, \dots, p]^\top \in \mathbb{R}^d$. In this way, the random feature corruption can be marginalized out over corruption probability p and close-form solutions are obtained accordingly. Thus, simplification of parameter estimation speed up significantly the learning process compared to SDA.

4.3 A Deep Learning Approach for Hybrid Heterogeneous Transfer Learning

As discussed in the two cross-language/lingual text mining examples, the selection of corresponding instances in the source domain is always biased in general. This is because the corresponding instances in the source domain are not randomly selected based on the source domain data distribution but based on the selection of the corresponding instances

in the target domain. On the other hand, the selection of the corresponding instances in the target domain can be either: 1) unbiased, e.g., in the example of cross language document classification, one can apply Google translator on a set of randomly selected documents in German to construct correspondences; or 2) bias, e.g., in the example of cross lingual sentiment classification, a set of reviews in German to be used to construct correspondences about books while our target is to learn a sentiment classifier on music reviews.

In the next few subsections, we address these two problems through the proposed hybrid heterogeneous transfer learning (HHTL) framework.

4.3.1 High-Level Homogeneous Feature Learning

Our approach is inspired by the motivations and promising results of self-taught learning and deep learning in domain adaptation [26, 44, 77]. To address the bias issue in either the source domain only or both the source and target domains caused by instance shift or feature mismatch, we propose to apply stacked denoised autoencoder (SDA) [93] on the source domain data, i.e., the source domain labeled data and the source-domain corresponding unlabeled data², and the target domain data, i.e., the target domain unlabeled data and the target-domain corresponding unlabeled data, to learn high-level representations. Specifically, SDA first randomly sets some values of source domain features to be zero, which is referred to as a “corruption” of source domain data. In total, one can obtain m different corruptions. After that SDA tries to learn high-level features by reconstructing these m corruptions. For example, German word “betonen” is translated to “emphasize” by using Google Translator. However, in human writings, one may use the words “proud” and “like” instead. SDA aims to reconstruct the machine translated word “emphasize” by using the words “proud” and “like”. Therefore, the learned high-level features have capability to reduce data or feature bias.

In particular, we adopt a recently proposed method named Marginalized Stacked Denoised Autoencoder (mSDA) [26] for high-level feature learning on homogeneous features.

²In practice, if there is an additional set of unlabeled data in the source domain, one can use it as well for high-level feature learning. However, for simplicity in description, we do not assume that additional source domain unlabeled data is available.

mSDA is an extension of SDA, which simplifies the reconstruction from two-level encoder and decoder to a single mapping. The reasons why we use mSDA are two folds:

- (i) The effectiveness of mSDA has been shown in homogeneous domain adaptation problems [26].
- (ii) Compared to the standard SDA method, mSDA has proven to be much more efficient.

For simplicity in presentation, following the notations used in [26], we absorb a constant feature into the feature vector as $\mathbf{x}_S = [\mathbf{x}_S^\top \ 1]^\top$ or $\mathbf{x}_T = [\mathbf{x}_T^\top \ 1]^\top$, and incorporate a bias term $\mathbf{b}$ within the weight matrix as $\mathbf{W} = [\mathbf{W} \ \mathbf{b}]$. We further denote $\bar{\mathbf{X}}_S = [\mathbf{X}_S^\top \ \mathbf{X}_S^c{}^\top]^\top$ the union source domain data, and $\bar{\mathbf{X}}_T = [\mathbf{X}_T^\top \ \mathbf{X}_T^c{}^\top]^\top$ the union target domain data.

First, for the source domain data, we apply mSDA on $\bar{\mathbf{X}}_S$ to learn a weight matrix $\mathbf{W}_S \in \mathbb{R}^{(d_S+1) \times (d_S+1)}$ by minimizing the squared reconstruction square loss³ as follows.

$$\sum_{i=1}^m \left\| \bar{\mathbf{X}}_S - \mathbf{W}_S \bar{\mathbf{X}}_S^{(i)} \right\|_F^2, \quad (4.5)$$

where $\bar{\mathbf{X}}_S^{(i)}$ denotes the i -th corrupted version of $\bar{\mathbf{X}}_S$. The solution to (4.5) depends on how the original features are corrupted is explicitly expressed as follows.

$$\mathbf{W}_S = \mathbf{P}\mathbf{Q}^{-1} \quad \text{with} \quad \mathbf{Q} = \tilde{\mathbf{X}}_S \tilde{\mathbf{X}}_S^\top \quad \text{and} \quad \mathbf{P} = \hat{\mathbf{X}}_S \tilde{\mathbf{X}}_S^\top, \quad (4.6)$$

where $\hat{\mathbf{X}}_S = [\bar{\mathbf{X}}_S \ \bar{\mathbf{X}}_S \ \cdots \ \bar{\mathbf{X}}_S]$ denotes the m -times repeated version of $\bar{\mathbf{X}}_S$, and $\tilde{\mathbf{X}}_S$ is the corrupted version of $\hat{\mathbf{X}}_S$. In general, to alleviate bias in estimation, a large number of m over the training data with random corruptions are required, which is computationally expensive. To address this issue, mSDA introduces a corruption probability p to model infinite corruptions, i.e., $m \rightarrow \infty$. Define a feature vector $\mathbf{q} = [1-p, \dots, 1-p, 1]^\top \in \mathbb{R}^{d_S+1}$, where $\mathbf{q}_i$ represents the probability of a feature indexed by i “surviving” after the corruption. Thus, we obtain the expectation of (4.5), and its solution is written analytically as

$$\mathbf{W}_S = \mathbb{E}[\mathbf{P}]\mathbb{E}[\mathbf{Q}]^{-1}, \quad (4.7)$$

³Alternatively, we can use other reconstruction loss, e.g., cross-entropy loss.

where $\mathbb{E}[\mathbf{P}]_{ij} = \mathbf{S}_{ij}\mathbf{q}_j$, $\mathbf{S} = \overline{\mathbf{X}}_S\overline{\mathbf{X}}_S^\top$, and

$$\mathbb{E}[\mathbf{Q}]_{ij} = \begin{cases} \mathbf{S}_{ij}\mathbf{q}_i\mathbf{q}_j, & \text{if } i \neq j, \\ \mathbf{S}_{ij}\mathbf{q}_i, & \text{otherwise.} \end{cases} \quad (4.8)$$

After $\mathbf{W}_S$ is learned, the nonlinearity of features is injected through the nonlinear encoder function $h(\cdot)$ that is learned together with the reconstruction weights $\mathbf{W}_S$, mSDA applies a nonlinear squashing-function, e.g., the hyperbolic tangent function $\tanh(\cdot)$, on the outputs of mSDA, $\overline{\mathbf{H}}_S = \tanh(\mathbf{W}_S\overline{\mathbf{X}}_S)$, to generate nonlinear features.

The above process is recursively done by replacing $\overline{\mathbf{X}}_S$ with $\overline{\mathbf{H}}_S$ to obtain a series of weight matrices $\{\mathbf{W}_S^k\}$'s for each layer of feature learning. Similarly, for the union target domain data $\overline{\mathbf{X}}_T$, we recursively apply mSDA to learn a series of reconstruction weights $\{\mathbf{W}_T^k\}$'s and generate high-level nonlinear features for each layer of feature learning, similarly. Note that when there is no data or feature bias in the target domain, one simply set $\mathbf{W}_T$ to be the identity matrix of the dimensionality $d_T + 1$, and replace $\tanh(\cdot)$ by the identical function.

4.3.2 Heterogeneous Feature Mapping

So far, in a specific layer k of feature learning, we have learned a pair of reconstruction weights $\mathbf{W}_{S,k}$ and $\mathbf{W}_{T,k}$, and higher-level feature representations $\overline{\mathbf{H}}_{S,k}$ and $\overline{\mathbf{H}}_{T,k}$ for the union source and target domain data, respectively. Let $\overline{\mathbf{H}}_{S,k}^{(c)}$ and $\overline{\mathbf{H}}_{T,k}^{(c)}$ be the higher-level feature representations of the cross-domain corresponding instances in the source and target domains, respectively. We describe how a feature mapping across heterogeneous features $\overline{\mathbf{H}}_{S,k}^{(c)}$ and $\overline{\mathbf{H}}_{T,k}^{(c)}$ is learned.

Specifically, in layer k , we aim to learn a feature transformation $\mathbf{G}_k \in \mathbb{R}^{(d_S+1) \times (d_T+1)}$, where a bias term is incorporated within the transformation, by minimizing the objective

$$\|\mathbf{H}_{S,k} - \mathbf{G}\mathbf{H}_{T,k}\|_F^2 + \lambda\|\mathbf{G}_k\|_F^2, \quad (4.9)$$

where $\lambda > 0$ is a parameter of the regularization term on $\mathbf{G}_k$, which controls the tradeoff between the alignment of heterogeneous features and the complexity of $\mathbf{G}_k$. The optimization problem (4.9) is regularized least square problem [45], which has a closed form solution

$$\mathbf{G}_k = (\mathbf{H}_{S,k}\mathbf{H}_{T,k}^\top)(\mathbf{H}_{T,k}\mathbf{H}_{T,k}^\top + \lambda\mathbf{I})^{-1}, \quad (4.10)$$

where $\mathbf{I}$ is the identity matrix of the dimensionality $d_T + 1$.

4.3.3 Prediction by Stacking Layers

Let the total number of layers be K . For each source domain labeled instance $\mathbf{x}_{S_i}$, we denote $\mathbf{h}_{S_i,k}$ to be its corresponding higher-level feature representation in the k -th layer. We also define a new feature vector $\mathbf{z}_{S_i}$ by augmenting its original features and high-level features of all layers as $\mathbf{z}_{S_i} = [\mathbf{h}_{S_i,1}^\top \cdots \mathbf{h}_{S_i,K}^\top]^\top$, where $\mathbf{h}_{S_i,1} = \mathbf{x}_{S_i}$.

One can recursively apply mSDAs and the heterogeneous feature mapping learning algorithm on the source and target domain data in each layer to generate different levels of features and feature transformations between heterogeneous features. We then apply a standard classification algorithm on $\{\mathbf{z}_{S_i}, y_{S_i}\}$'s to train a target classifier f . To perform a prediction on a target domain instance $\mathbf{x}_T^*$, we first generate its higher-level feature representations $\{\mathbf{h}_{T,k}^*\}_{k=1}^K$, where $\mathbf{h}_{T,1}^* = \mathbf{x}_T^*$, of each layer by using the weight matrices $\{\mathbf{W}_{T,k}\}_{k=1}^K$, where $\mathbf{W}_{T,1} = \mathbf{I}$, and do the feature augmentation, $\mathbf{z}_T = [(\mathbf{G}_1 \mathbf{h}_{T,1}^*)^\top \cdots (\mathbf{G}_K \mathbf{h}_{T,K}^*)^\top]^\top$. Finally, we apply the learned classifier f on $\mathbf{z}_T$ to predict $f(\mathbf{z}_T)$. The reason that we augment different layers of features for both training and testing is because we aim to incorporate additional high-level features to alleviate the bias for both domains without losing original feature information. The proposed HHTL algorithm is summarized in Algorithm 2. Figure 4.1.a and 4.1.b illustrate the learned deep structures for unbiased and biased cases discussed at the beginning of this chapter (or see Figure 1.7.a and 1.7.b).

4.4 Experimental Results

We verify the proposed framework on several cross-language classification tasks in terms of classification accuracy, impact of number of layers, and parameter sensitivity.

4.4.1 Experimental Setting

Dataset The cross-language sentiment dataset [75] comprises of Amazon product reviews of three product categories: books (B), DVDs (D) and music (M). These reviews are written in four languages: English (EN), German (GE), French (FR), and Japanese (JP). For each language, the reviews are split into a training set and a test set, including 2,000 reviews per categories. We use the English reviews in the training set as the source

Algorithm 2 Hybrid Heterogeneous Transfer Learning.

Input: target domain unlabeled data $\mathbf{D}_T = \{\mathbf{x}_{T_i}\}_{i=1}^{n_1}$, source domain labeled data $\mathbf{D}_S = \{(\mathbf{x}_{S_i}, y_{S_i})\}_{i=1}^{n_2}$, cross-domain parallel data, $\mathbf{D}_C = \{(\mathbf{x}_{S_i}^{(c)}, \mathbf{x}_{T_i}^{(c)})\}_{i=1}^{n_c}$, a feature corruption probability p in mSDA, a trade-off parameter λ , and the number of layers K .

Initializations: $\bar{\mathbf{X}}_S = [\mathbf{X}_S \ \mathbf{X}_S^{(c)}]$, $\bar{\mathbf{X}}_T = [\mathbf{X}_T \ \mathbf{X}_T^{(c)}]$, $\bar{\mathbf{H}}_{S,1} = \bar{\mathbf{X}}_S$, $\bar{\mathbf{H}}_{T,1} = \bar{\mathbf{X}}_T$, and learn $\mathbf{G}_1$ by solving

$$\min_{\mathbf{G}_1} \|\mathbf{H}_{S,1}^{(c)} - \mathbf{G}_1 \mathbf{H}_{T,1}^{(c)}\|_F^2 + \lambda \|\mathbf{G}_1\|_F^2.$$

for $k = 2, \dots, K$ **do**

1: Apply mSDA on $\bar{\mathbf{H}}_{S,k-1}$ and $\bar{\mathbf{H}}_{T,k-1}$:

$$\begin{aligned} \{\mathbf{W}_{S,k}, \bar{\mathbf{H}}_{S,k}\} &= \text{mSDA}(\bar{\mathbf{H}}_{S,k-1}, p), \\ \{\mathbf{W}_{T,k}, \bar{\mathbf{H}}_{T,k}\} &= \text{mSDA}(\bar{\mathbf{H}}_{T,k-1}, p). \end{aligned}$$

Note: if there is no data or feature bias in the target domain, simply set $\mathbf{W}_{T,k} = \mathbf{I}$, and $\bar{\mathbf{H}}_{T,k} = \bar{\mathbf{X}}_T$.

2: Learn heterogeneous feature mapping $\mathbf{G}_k$:

$$\min_{\mathbf{G}_k} \|\mathbf{H}_{S,k}^{(c)} - \mathbf{G}_k \mathbf{H}_{T,k}^{(c)}\|_F^2 + \lambda \|\mathbf{G}_k\|_F^2.$$

end for

Do feature augmentation on source domain labeled data

$$\mathbf{Z}_S = [\mathbf{H}_{S,1}^\top \ \cdots \ \mathbf{H}_{S,K}^\top]^\top,$$

and train a classifier f with $\{\mathbf{Z}_S, \mathbf{Y}_S\}$.

Output: f , $\{\mathbf{G}_k\}_{k=1}^K$, $\{\mathbf{W}_{S,k}\}_{k=2}^K$, and $\{\mathbf{W}_{T,k}\}_{k=2}^K$.

domain labeled data, non-English (each of the other three languages) reviews in a training set as target domain unlabeled data. Moreover, we apply Google translator on the non-English reviews in a test set to construct cross-domain (English v.s. non-English) unlabeled parallel data. All methods are evaluated on the target domain unlabeled data.

Baselines In our heterogeneous transfer learning setting, no labeled data is provided for the target domain. Most HTL methods which require target domain labeled data cannot be used as baselines. Therefore, we only compare the proposed HHTL framework with the following baselines:

- **SVM-SC:** We first train a classifier on the source domain labeled data and then predict on the source domain parallel data. By using the correspondence, the

4.1.a: Case I: Unbiased Instance Selection

4.1.b: Case II: Biased Instance Selection

Figure 4.1: Deep Structures for Proposed HHTL method.

predicted labels for source parallel data can be transferred into target parallel data. Next, we train a model on the target parallel data with predicted labels to make predictions on the target domain test data. We name this baseline as the SVM-Source-Correspondence transfer (SVM-SC).

- **CL-KCCA:** We apply Cross-Lingual Kernel Canonical Component Analysis (CL-

KCCA) [94] on the unlabeled parallel data to learn two projections for the source and target languages, and then train a monolingual classifier with the projected source domain labeled data.

- **HeMap:** We apply heterogeneous Spectral Mapping (HeMap) [84] to learn mappings to project two domain data onto a common feature subspace. However, HeMap does not take the instance correspondence information into consideration.
- **mSDA-CCA:** We apply Multimodal Deep Learning [66] to learn a shared feature representation for “multimodal” domains. In the experiments, for the fair comparison, instead of using restricted Boltzmann machine (RBM) [56, 66] for high-level feature learning, we adopt mSDA and perform Canonical Component Analysis (CCA) on the correspondences between domains in same layers. Different from the deep transfer architecture proposed in [66] that transfers knowledge in the one hidden layer, HHTL transfers knowledge by heterogeneous feature mapping in each layer of deep network.

For all experiments, we employ the linear support vector machine (SVM) [39] with default parameter settings. We use the cross-validation to adjust the model parameters. Specifically, we choose λ from $\{0.01, 0.1, 1, 10, 100\}$ for HHTL, choose corruption probability p from $\{0.5, 0.6, 0.7, 0.8, 0.9\}$ for mSDA, and fix the number of layers used in mSDA to be 3. We tune the parameter κ for CL-KCCA (see (5) in [94]), mSDA-CCA, and the parameter β for HeMap (See (1) in [84]) from $\{0.01, 0.1, 1, 10, 100\}$. Due to the limitation of memory and computation, we only study three layers.

4.4.2 Performance Comparison

We evaluate the performance of our proposed method under the two learning settings: 1) learning with unbiased target correspondence instances, and 2) learning with biased target correspondence instances.

4.4.2.1 Learning with unbiased target correspondence instances

In this setting, we consider bias in the source domain. During the training process, all the original English reviews consisting of three categories are used as the source

Table 4.1: Learning with unbiased target correspondence instances: comparison results in terms of testing accuracy (%).

Target	SVM-SC	CL-KCCA	HeMap	mSDA-CCA	HHTL(1)	HHTL(2)	HHTL(3)
FR	73.10	75.50	50.23	74.89	74.01	79.02	82.50
GE	73.05	75.00	49.83	76.53	73.08	78.54	82.70
JP	65.50	66.82	51.30	68.19	66.12	71.03	75.56

domain labeled data, and non-English reviews are considered as target domain data. We randomly choose 2,000 non-English reviews of all the categories, and translate them to English to form the parallel data. The remaining non-English reviews form the unlabeled target data. The average accuracy on the reviews of each non-English language over 10 repetitions are reported in Table 4.1.

From Table 4.1, we observe that the proposed HHTL method with three layers outperforms all the other baselines significantly. The reason is that HHTL benefits a lot from learning the high-level features which alleviate the data bias between the translated and original reviews in the source domain. Moreover, the performance of CL-KCCA and SVM-SC is much better than HeMap. The inferior performance of HeMap is caused by the fact that HeMap discards the valuable corresponding information in training. Moreover, mSDA-CCA performs slightly better than CL-KCCA, and much better than all the other baselines because of the powerful representations learned by the deep structure. However, it still performs worse than HHTL because the representations learned by mSDA-CCA are based on the biased correspondences. From the results, we can also see that the performance of all the methods in the Japanese domain is much worse than that in the other two language domains. It is due to the fact that Japanese is more different than English compared to German and French. French and German all come from the similar family of languages in the West. Thus, they share many similar words with English, while Japanese belongs to the family of Eastern languages. Therefore, it is more difficult to transfer knowledge from the English domain to the Japanese domain.

4.4.2.2 Learning with biased target correspondence instances

In this setting, we focus on cross-language cross-category (i.e., Books (B), DVDs (D), Music (M)) learning between English (EN) and the other 3 languages (i.e., German (GE), French (FR), Japanese(JP)). For the comprehensive comparisons, we construct

Table 4.2: Learning with biased target correspondence instances: comparison results in terms of testing accuracy (%).

TASK	SVM-SC	CL-KCCA	HeMap	mSDA-CCA	HHTL(1)	HHTL(2)	HHTL(3)
EN-B-FR-D	73.25	50.00	49.45	72.96	72.65	74.14	76.75
EN-B-FR-M	62.40	47.20	50.70	64.29	63.30	65.45	67.65
EN-B-GE-D	72.55	77.50	66.12	78.23	70.95	73.69	75.10
EN-B-GE-M	58.95	50.00	48.35	62.52	59.60	65.28	69.55
EN-B-JP-D	69.50	69.46	49.55	71.95	69.45	71.41	72.56
EN-B-JP-M	53.15	55.23	52.05	57.17	53.56	58.17	62.39
EN-D-FR-B	71.35	47.10	48.35	72.32	71.25	75.32	79.27
EN-D-FR-M	71.65	47.10	49.90	68.87	71.70	74.87	75.43
EN-D-GE-B	76.40	54.87	50.80	75.32	74.00	77.32	79.35
EN-D-GE-M	69.05	54.87	50.55	75.69	74.05	76.69	78.55
EN-D-JP-B	70.00	61.16	50.12	72.60	65.55	66.60	68.12
EN-D-JP-M	59.20	66.89	49.80	57.75	57.75	66.02	70.58
EN-M-FR-B	76.20	73.54	50.55	74.23	73.80	74.23	75.84
EN-M-FR-D	73.65	47.10	50.87	72.76	74.15	75.76	77.04
EN-M-GE-B	74.50	81.50	49.45	76.82	75.45	77.68	78.30
EN-M-GE-D	74.60	80.20	50.20	72.28	74.45	78.68	81.42
EN-B-JP-B	67.85	63.75	48.85	68.83	65.35	68.83	71.65
EN-B-JP-D	69.35	60.46	48.80	71.06	68.55	70.06	74.25

18 cross-language cross-category sentiment classification tasks as follows: EN-B-FR-D, EN-B-FR-M, EN-B-GE-D, EN-B-GE-M, EN-B-JP-D, EN-B-JP-M, EN-D-FR-B, EN-D-FR-M, EN-D-GE-B, EN-D-GE-M, EN-D-JP-B, EN-D-JP-M, EN-M-FR-B, EN-M-FR-D, EN-M-GE-B, EN-M-GE-D, EN-B-JP-B, EN-B-JP-D. For example, the task EN-B-FR-D uses all the Books reviews in French in the test file and its English translations as the parallel data, the DVD reviews in French as the target language test data, and original English Books reviews as the source domain labeled data.

The results are summarized in Table 4.2. This setting is more challenging than the previous one due to larger data bias after feature transformation. Therefore, the performance of all target languages is dropped compared to that shown in Table 4.1. Our proposed method HHTL still performs much more stable and better than the other three baselines except for a few tasks. This is because that our proposed method can largely reduce the data bias by generating the more powerful and higher-level features for both the source and target domains, which leads to better cross-language mappings

in each layer.

4.4.3 Impact of Layers in the HHTL Structure

As shown in Table 4.2, the more layers are used, the better performance is achieved. SVM-SC manifests comparable results with HHTL using one layer, where the learned high-level features are not sufficiently useful. However, with increasing number of layers, HHTL enhances the cross-language classification performance by generating more useful and higher-level features that alleviate the data bias.

4.4.4 Parameter Sensitivity Study

The proposed method has two parameters to tune: 1) the tradeoff parameter λ , and 2) the feature corruption probability p . Besides these, the number of correspondences n_c from the parallel data may also affect the performance of cross-language classification. We conduct a series of experiments to study the sensitivity issues of the parameters, λ , p and n_c .

In the first experiment, we aim to analyze how performance changes with varying values of λ in the range of $[0.001, 0.01, 1, 10, 100]$ with $p = 0.7$ and $n_c = 50$. From Figure 4.2.a, we observe that the performance is stable when λ is no more than 10. Besides, we set $\lambda = 0.01$, $n_c = 50$ and vary p in the range of $[0.5, 0.6, 0.7, 0.8, 0.9]$. The results of all the target languages are illustrated in Figure 4.2.b. We see that corruption probability cannot be either too large or too small to achieve good performance. On one hand, if we corrupt many features, a large amount of the original feature information will be discarded, resulting in failure of discovering powerful hidden features in higher layers by reconstructions. On the other hand, if we only corrupt a few features, the high-level hidden features recovered from the original features tends to be similar to the original ones. In this case, we cannot learn informative hidden features for knowledge transfer.

In the second experiment, we study the influence of the size of the unlabeled parallel data ($[50, 250, 500, 750, 1000, 1250, 1500, 1750]$) to the overall performance of our proposed HHTL method. The results for German, French and Japanese domains are reported in Figure 4.2.c, Figure 4.2.d and Figure 4.2.e, respectively. All the methods which use

the unlabeled parallel data consistently outperform HeMap, which discards the correspondence information. Nevertheless, HTLL performs the best and achieve more stable improvement with increasing size of parallel data. Different from HTLL, the accuracies of CL-KCCA and SVM-SC almost stop increasing when the size of parallel data is larger than 750. The reason is that the impact of domain distribution mismatch hinders the improvement of the overall performance even though there are more correspondences. mSDA-CCA performs even worse than SVM-SC and CL-KCCA when the number of correspondences is smaller than 750. This is because that the multimodal deep learning method requires sufficient correspondence data between two modalities to learn reliable feature representations. Though mSDA-CCA outperforms other baselines when n_c is larger than 750, it still performs much worse than HHTL. All the results demonstrate the effectiveness and robustness of HHTL for cross-language sentiment classification.

4.5 Summary

In this chapter, we describe our proposed Hybrid Heterogeneous Transfer Learning (H-HTL) framework that transfers knowledge across different feature spaces and correct the data bias on the transformed feature space. Extensive experimental results demonstrate the superiority of the proposed method on several multilingual text mining tasks. Future work includes a cross-layer heterogeneous feature mapping for multimodal classification problem to model more abstract relationships. More discussion is found in the Chapter 6. Moreover, we will extend the linear feature representation to nonlinear feature representation that can model more complicated feature relationships. Currently, high-level homogeneous feature learning and heterogeneous feature mapping are performed in a two-step paradigm. This is not an optimal choice because error in the earlier step or layer will propagate into future steps. To overcome this weakness, a unified deep learning approach for HHTL is highly desirable.

 4.2.a: Accuracy v.s. λ

 4.2.b: Accuracy v.s. p

 4.2.c: Accuracy v.s. n_c on GE

 4.2.d: Accuracy v.s. n_c on FR

 4.2.e: Accuracy v.s. n_c on JP

Figure 4.2: Parameter Analysis.

Chapter 5

Heterogeneous Transfer Learning through Active Correspondences Construction

In this chapter, we address two HTL issues. The first one is related to the explicit matching of the distribution between two heterogeneous domains. The second one is on the need to actively construct the cross-domain correspondences.

The success of transfer learning depends on the distribution similarity between two domains after domain adaptation [11]. For the homogeneous transfer learning setting, where the source and the target domain share the same feature spaces, it is natural to adapt one domain to the other in terms of existing distribution distance measures such as KL-divergence [89], Maximum Mean Discrepancy (MMD) [48], $\mathcal{H}$ -distance [8], etc. Among these distribution distance measures, MMD is a consistent and unbiased estimator, which has shown its importance in designing a number of homogeneous transfer learning algorithms, such as Maximum Mean Discrepancy Embedding (MMDM) [68], Transfer Component Analysis (TCA) [69], and Kernel-Mean Matching (KMM) [53]. In particular, MMDM and TCA are feature-based algorithms which minimize MMD between two domains by discovering the latent space. Different from feature-based methods, KMM is an instance-based method which minimizes MMD between two domains by learning density-ratio among instances of two domains.

For HTL problem, explicitly minimizing the distribution differences across domains is a much more challenging task. There are several works which try to minimize the domain difference in terms of the mean or summation, e.g. MOMAP [49], HeMap [84]. These

methods only consider the first-order moment when adapting domains. However, it is also important to consider the high-order moments. In this case, MMD is a good choice and is proven to capture up to infinite-moments. Nevertheless, we cannot directly apply MMD to HTL problems where the feature spaces of two domains are different. To address this challenge, we propose a novel algorithm based on the Matrix Completion (MC) by regularizing the MMD between two domains. This algorithm is inspired by the idea that the corresponding instances in either the source or target domain can be reconstructed by matrix completion (MC) [104]. Furthermore, we construct the “augmented” instances for two domains by concatenating the original instances and recovered corresponding instances. In this way, the new “augmented” domains are in the same feature spaces. We further regularize the solution by projecting it into the feasible space where MMD between two augmented domains is minimized. The MC and MMD regularization are performed in an iterative manner. Our proposed algorithm is called *Distribution Matching based Matrix Completion* (DMMC).

Another critical issue of correspondences-based HTL is to reduce the budget of constructing the cross-domain correspondences. In real-world applications, such correspondences are often expensive to achieve. The lack of sufficient correspondences results in several problems. One of them is distribution bias in both heterogeneous domains and is addressed using Hybrid Heterogeneous Transfer Learning (HHTL) in the last chapter. Another problem is that the given cross-domain correspondences are perhaps not most desirable for the constructed models. However, most existing HTL algorithms assume such correspondences are known in advance [28, 114]. In this chapter, we aim to relax this assumption to enable HTL to be more feasible for real-world applications. Specifically, in many real-world problems, correspondences of instances between domains are unknown, but can be identified using cost. Apart from the topic categorization for scanned books, there are also several other real-world examples. For instance, given two social media, identifying whether two user accounts from two different social media referred to the same person cannot be done automatically, and usually require intensive human effort to investigate their meta information, historical sent messages, meta information of their friends, etc. Another example, for the problem of cross-language document classification is the use of source-language labeled documents for target-language document classification, where document-level translation between languages is required to construct

document-correspondences. However, automatic translation is usually unsatisfactory in the document level, and intensive human effort is usually required to ensure precise translation.

To reduce the cost of identifying correspondences between domains for HTL, one can actively construct most “informative” and “high-quality” correspondences with limited budget for effective knowledge transfer across domains. Inspired by the active learning problem [82], we propose a novel and general HTL framework named Heterogenous Transfer Learning through Active correspondence construction (HTLA) to address the problem. In summary, the main contributions of this work are two folds:

- We propose a general correspondence-based method for HTL through matrix completion.
- We further propose a novel framework based on the proposed method to actively construct correspondences between two domains for effective knowledge transfer.

The rest of the chapter is organized as follows: We first provide formal problem statement and preliminary knowledge on matrix completion in Section 5.2. We then describe our propose Distribution Matching-based Matrix Completion (DMMC) method. For DMMC, a regularized solution is obtained by minimizing the distribution differences between two domains during the matrix completion. Section 5.3 describes the active correspondences construction strategy. This strategy is integrated into DMMC to achieve a novel HTL framework named Heterogenous Transfer Learning through Active correspondence construction (HTLA). In Section 5.4, we report extensive experiments on the cross-language sentiment dataset to demonstrate the effectiveness and superiority of DMMC algorithm as well as the proposed active correspondences construction strategy. Furthermore, parameter sensitivity analysis demonstrates that our method is insensitive to most of the parameters in the model.

5.1 Related Work

As mentioned in Chapter 2, most existing approaches to HTL are classified into two settings. One assumes few labeled data is available in the target domain [37, 55, 96, 115],

while the other setting assumes insufficient instance-correspondences between heterogeneous domains are available. For example, Dai *et al.* [28] proposed a probabilistic model to construct a “translator” to build connections between two instances from different domains, such that knowledge can be transferred across domains. Recently, Xiao *et al.* [104] applied the matrix completion technique to HTL. They assume several instance-correspondences between domains are given in advance, where the goal is to reconstruct corresponding “instances” for all the instances. However, their proposed method does not take distribution distance between domains into consideration when reconstructing the corresponding instances.

There have been research works on combining active learning and transfer learning, which focused on actively querying instances for labels in the target domain by leveraging knowledge from the source domain [24, 76, 83, 100]. A related work to actively constructing correspondences between heterogeneous domains for knowledge transfer is [111], where an active learning strategy was proposed to selectively identify user/item correspondences between different recommender systems for knowledge transfer. However, their proposed method is specific for collaborative filtering in recommender systems, and is not easy to generalize to other HTL problems.

5.2 Problem Statement and Preliminaries

Assume that we are given a set of source-domain labeled instances, $\{(\mathbf{x}_{S_i}, y_{S_i})\}_{i=1}^l$, where $\mathbf{x}_{S_i} \in \mathbb{R}^{d_S}$ denotes an instance and $y_{S_i} \in \{1, -1\}$ denotes the corresponding label, a set of source-domain unlabeled instances $\{\mathbf{x}_{S_i}\}_{i=l+1}^{n_S}$, and a set of corresponding pairs between the source and target domains, $\{(\mathbf{x}_{S_i}^C, \mathbf{x}_{T_i}^C)\}_{i=1}^{n_C}$, where $\mathbf{x}_{S_i}^C \in \mathbb{R}^{d_S}$ and $\mathbf{x}_{T_i}^C \in \mathbb{R}^{d_T}$ are source- and target-domain instances respectively. Note that n_C can be 0 at the beginning of training. The goal is to make predictions on another set of target-domain unlabeled instances $\{\mathbf{x}_{T_i}\}_{i=1}^{n_T}$. For simplicity, we denote $\mathbf{X}_{SL} \in \mathbb{R}^{l \times d_S}$, $\mathbf{X}_{SU} \in \mathbb{R}^{(n_S-l) \times d_S}$ and $\mathbf{X}_{SA} \in \mathbb{R}^{n_S \times d_S}$ the matrices of source-domain labeled, unlabeled, and all instances, respectively, where each row corresponds an instance. Similarly, we denote $\mathbf{X}_T \in \mathbb{R}^{n_T \times d_T}$ the matrices of target-domain unlabeled instances, and denote $\mathbf{X}_{SC} \in \mathbb{R}^{n_C \times d_S}$ and $\mathbf{X}_{TC} \in \mathbb{R}^{n_C \times d_T}$ the matrices of corresponding instances in the source and target domains respectively.

As our proposed method is based on matrix completion, we first construct a unified instance-feature matrix for all instances from the source and target domains as follows.

$$\mathbf{M} = \begin{bmatrix} \bar{\mathbf{X}}_S \\ \bar{\mathbf{X}}_T \end{bmatrix} = \begin{bmatrix} \bar{\mathbf{X}}_{SA} \\ \bar{\mathbf{X}}_{SC} \\ \bar{\mathbf{X}}_T \end{bmatrix} = \begin{bmatrix} \mathbf{X}_{SA} & \mathbf{0}_{n_S, d_T} \\ \mathbf{X}_{SC} & \mathbf{X}_{TC} \\ \mathbf{0}_{n_T, d_S} & \mathbf{X}_T \end{bmatrix}, \quad (5.1)$$

where $\bar{\mathbf{X}}_S = \begin{bmatrix} \bar{\mathbf{X}}_{SA}^\top & \bar{\mathbf{X}}_{SC}^\top \end{bmatrix}^\top$, and the matrices $\mathbf{0}_{n_S, d_T}$ and $\mathbf{0}_{n_T, d_S}$ denote the missing values of $\mathbf{X}_{SA}$ and $\mathbf{X}_T$ in the target and source domains respectively, which is recovered by matrix completion. We further define $n'_S = n_S + n_C$.

5.2.1 Matrix Completion

Given the partial observed cross-domain instances-feature matrix $\mathbf{M}$, we seek to estimate the missing values of the matrices $\mathbf{0}_{n_S, d_T}$ and $\mathbf{0}_{n_T, d_S}$ (i.e., to estimate the corresponding instances of $\mathbf{X}_{SA}$ and $\mathbf{X}_T$ in the target and source domains respectively) by recovering the ground-truth matrix $\mathbf{X}$ that fulfills the following properties:

- $\mathbf{X}$ is sparse. For many tasks such as cross-language text categorization, feature representation for an instance is sparse. Thus, we use the L_1 norm $\|\cdot\|_1$ to penalize its sparsity.
- $\mathbf{X}$ is row rank. For many applications, observed data of high dimensions is controlled by a few latent factors. Thus, we use the nuclear norm $\|\cdot\|_*$ to regularize its rank.

The HTL problem is then transformed into a matrix completion problem [104]:

$$\min_{\mathbf{X}} \frac{1}{2} \|(\mathbf{X} - \mathbf{M}) \circ \mathbf{P}\|_F^2 + \gamma \|\mathbf{X}\|_* + \mu \|\mathbf{X}\|_1 \quad (5.2)$$

where $\mathbf{M}$ is an observed matrix, $\mathbf{P}$ is indicator matrices such that $\mathbf{P}_{ij} = 1$ if (i, j) is observed, otherwise 0 and $\circ$ is the Hadamard product. γ and μ denote tradeoff parameters for the nuclear norm $\|\cdot\|_*$ and the L_1 norm $\|\cdot\|_1$ respectively.

5.2.2 Maximum Mean Discrepancy

Maximum Mean Discrepancy (MMD) [15] is a nonparametric measure to estimate the distance between distributions, which is expressed as follows.

$$\text{Dist}(\bar{\mathbf{X}}_S, \bar{\mathbf{X}}_T) = \left\| \frac{1}{n'_S} \sum \phi(\bar{\mathbf{x}}_{S_i}) - \frac{1}{n_T} \sum \phi(\bar{\mathbf{x}}_{T_j}) \right\|_{\mathcal{H}}^2, \quad (5.3)$$

where $\phi(\cdot)$ is a feature mapping to the Hilbert space. It can be rewritten in a more compact form:

$$\text{Dist}(\mathbf{D}_S, \mathbf{D}_T) = \text{tr}(\mathbf{K}_\mathbf{X} \mathbf{L}), \quad (5.4)$$

where $\mathbf{K}_\mathbf{X} = \begin{pmatrix} \mathbf{K}_{\bar{\mathbf{X}}_{S,S}} & \mathbf{K}_{\bar{\mathbf{X}}_{S,T}} \\ \mathbf{K}_{\bar{\mathbf{X}}_{T,S}} & \mathbf{K}_{\bar{\mathbf{X}}_{T,T}} \end{pmatrix}$, which is a kernel matrix induced by $\phi(\cdot)$, and $\mathbf{L} = \begin{pmatrix} \mathbf{L}_{S,S} & \mathbf{L}_{S,T} \\ \mathbf{L}_{T,S} & \mathbf{L}_{T,T} \end{pmatrix}$, where $\mathbf{L}_{ij} = 1/(n'_S)^2$ if $\mathbf{x}_i, \mathbf{x}_j \in \bar{\mathbf{X}}_S$, $\mathbf{L}_{ij} = 1/(n_T)^2$ if $\mathbf{x}_i, \mathbf{x}_j \in \bar{\mathbf{X}}_T$, otherwise $\mathbf{L}_{ij} = -1/(n'_S n_T)$. Here, we use gaussian kernel: $\mathbf{K}_\mathbf{X}(\bar{\mathbf{x}}_{Z_i}, \bar{\mathbf{x}}_{Z_j}) = \exp(-\|\bar{\mathbf{x}}_{Z_i} - \bar{\mathbf{x}}_{Z_j}\|^2/2\delta)$, where $Z \in \{S, T\}$.

5.3 Active Correspondences Construction for HTL

We present the overall framework of Heterogeneous Transfer Learning through Active correspondences construction (HTLA) as described in Algorithm 3. To begin with, we randomly construct several correspondences to form the matrix $\mathbf{M}$ as shown in (5.1), and apply the proposed Distribution Matching based Matrix Completion (DMMC) method to recover an initial $\mathbf{X}$. After that, we iteratively construct cross-domain correspondences based on the proposed active correspondences construction strategy, and apply DMMC to recover more precise $\mathbf{X}$ till some stopping criterion is met. After the “ground-truth” matrix $\mathbf{X}$ is recovered, we apply Singular Value Decomposition (SVD) on $\mathbf{X}$ to project both the source and target domain data to a common latent space spanned by the top r singular vectors, i.e., $\mathbf{X} = \mathbf{U}_r \mathbf{\Sigma}_r \mathbf{V}_r$, and let $\mathbf{Z} = \mathbf{X} \mathbf{V}_r$. Finally, we train a classifier on the first n_S rows of $\mathbf{Z}$, i.e., new feature representations of the source-domain labeled data, and test on the last n_T rows of $\mathbf{Z}$, i.e., new feature representations of the target-domain test data. In the next two subsections, we introduce the DMMC method and the strategy of active correspondences construction in detail.

Algorithm 3 Framework of HTLA

Initializations: Randomly construct k correspondences $\mathcal{C}^{(1)}$ to form the matrix $\mathbf{M}$, and set $\mathbf{X}^{(1)} = \mathbf{M}$.
for $t = 1, 2, \dots, T$ **do**
 1: $\mathbf{X}^{(t+1)} = \text{DMMC}(\mathbf{X}^{(t)})$.
 2: Set $\mathcal{C}^{(t+1)} = \text{ActiveLearn}(\mathbf{X}^{(t)}, \mathcal{C}^{(t)})$.
 3: Update $\mathbf{X}^{(t+1)}$ based on $\mathcal{C}^{(t+1)}$, and set $t \leftarrow t + 1$.
end for
Return: $\mathbf{X} \leftarrow \mathbf{X}^{(T+1)}$
Recover low-dimensional representations: apply a SVD on $\mathbf{X}$, $\mathbf{X} = \mathbf{U}_r \mathbf{\Sigma}_r \mathbf{V}_r$, and use the top r singular vectors to construct a projection, $\mathbf{Z} = \mathbf{XV}_r$.

5.3.1 Distribution Matching based Matrix Completion

We first assume several correspondences are given to form the matrix $\mathbf{M}$. The proposed optimization problem is expressed as follows:

$$\begin{aligned} \min_{\mathbf{X}} \quad & F(\mathbf{X}) = \gamma \|\mathbf{X}\|_* + \mu \|\mathbf{X}\|_1 + \frac{1}{2} \|(\mathbf{X} - \mathbf{M}) \circ \mathbf{P}\|_F^2 \\ \text{s.t.} \quad & \mathbf{X} \in \mathcal{K} \end{aligned} \tag{5.5}$$

where $\mathcal{K} = \{\mathbf{X} | \text{tr}(\mathbf{K}_\mathbf{X} \mathbf{L}) < k\}$, and k is a constant to constrain the distance between the source and target domain data, $\bar{\mathbf{X}}_S \bar{\mathbf{X}}_T$ measured by MMD. Note that the difference between (5.5) and (5.2) is the additional constraint based on MMD. In Section 5.4, we show empirically that the MMD constraint or regularization term can significantly improve the performance for HTL. However, this additional constraint is nonconvex, resulting in the overall optimization problem difficult to be solved by existing MC solvers.

One of our contributions is an efficient algorithm to solve (5.5) based on stochastic sub-gradient descent (SSGD) with projection on the feasible space $\mathcal{K}$. We use SSGD to solve the problem because of its ability to solve large-scale optimization problems [5, 16] and promising results on nonlinear optimization problems [52, 101]. A diagrammatic explanation of DMMC is shown in Figure 5.1. Specifically, our proposed approach is divided into three steps: 1) construction of the augmented data matrix according to Equation 5.1, 2) stochastic sub-gradient descent for MC, and 3) projection onto space $\mathcal{K}$. We iteratively perform the step 2 and 3 until the solution is convergent. The pseudo code of DMMC using SSDG is presented in Algorithm 4.

Figure 5.1: Distribution Matching based Matrix Completion (DMMC).

Algorithm 4 SSGD for DMMC

Input: Instance-feature matrix $\mathbf{M}$, trade-off parameters γ, λ, μ , maximum iteration K , upper bound of rank r .

Initializations: $t = 0$, $\mathbf{X}^{(t)}$ the initial rank- r approximation of $\mathbf{M}$, and stepsize η .

while $t \leq K$ **do**

1: Generate an $d \times p$ probing matrix $\mathbf{Y}^{(t)}$

2: Sub-gradient: $g \leftarrow \mathcal{G}(F(\mathbf{X}^{(t)}))\mathbf{Y}^{(t)}(\mathbf{Y}^{(t)})^\top$

3: Projection: $\hat{\mathbf{X}}^{(t)} \leftarrow \prod_{\mathcal{K}}(\mathbf{X}^{(t)} - \eta g)$

4: $\mathbf{X}^{(t+1)} = \hat{\mathbf{X}}^{(t)}$, and $t \leftarrow t + 1$.

end while

Output: $\mathbf{X} = \mathbf{X}^{(K+1)}$

5.3.1.1 Stochastic Sub-gradient Descent for MC

The sub-gradient of $F(\mathbf{X})$ in (5.5) for $\mathbf{X}$ is written as

$$\mathcal{G}(F(\mathbf{X})) = \gamma \frac{\partial \|\mathbf{X}\|_*}{\partial \mathbf{X}} + \mu \frac{\partial \|\mathbf{X}\|_1}{\partial \mathbf{X}} + (\mathbf{X} - \mathbf{M}) \circ \mathbf{P}. \quad (5.6)$$

Let $\mathbf{U}\Sigma\mathbf{V}$ be a compact SVD of $\mathbf{X}$ and $\frac{\partial \|\mathbf{X}\|_*}{\partial \mathbf{X}} = \mathbf{U}\mathbf{V}^\top + \mathbf{H}$. We set $\mathbf{H}$ as the zero matrix [102] and $\frac{\partial \|\mathbf{X}\|_1}{\partial \mathbf{X}_{i,j}} = \text{sign}(\mathbf{X}_{ij})$ when $\mathbf{X}_{ij} \neq 0$, otherwise $\frac{\partial \|\mathbf{X}\|_1}{\partial \mathbf{X}_{i,j}} = \theta \in [-1, 1]$.

Efficiency and Unbiased Sub-gradient Estimation: The success of SSGD relies on the unbiased estimation of sub-gradient estimation. Avron *et al.* [5] proposed to use a probing matrix to sparsify the sub-gradient estimation such that $\mathbb{E}(\tilde{g}) = g$.

Definition 2. Probing Matrix [5] : A random $d \times p$ matrix $\mathbf{Y}$ with $p < d$ is a probing matrix if $\mathbb{E}[\mathbf{Y}\mathbf{Y}^\top] = \mathbf{I}_{d \times d}$, where $\mathbf{I}_{d \times d}$ is the identity matrix, and $\mathbb{E}$ is the expectation operator.

We can randomly sample p vectors from the scaled standard basis $\{\sqrt{d}\mathbf{e}_1, \dots, \sqrt{d}\mathbf{e}_d\}$ to form the probing matrix: $\mathbf{Y} = [\sqrt{d}\mathbf{e}_1, \dots, \sqrt{d}\mathbf{e}_p]/\sqrt{p}$. Therefore, we just need to sample p column vectors from the d columns of a $n \times d$ matrix $\mathbf{X}$. Note that the time complexity for matrix completion depends on the SVD procedure on the matrix, i.e., $O(9d^3 + 8nd^2 + 4n^2d)$, where $d = d_S + d_T$ and $n = n'_S + n_T$ in our case. Compared to the method proposed by Xiao *et al.* [104], which performs a SVD on the whole matrix, our proposed method only conducts SVD on the submatrix of original matrix. Therefore, the

time complexity is reduced to $O(9p^3 + 8np^2 + 4n^2p)$. When the features are of very high dimensions, $p \ll d$, our proposed approach is much more efficient. In addition, to avoid additional parameters, we set p of the probing matrix the same as the reduced dimension r similar to [5].

5.3.1.2 Projection to $\mathcal{K}$ -space

After a step of sub-gradient descent, the recovered $\mathbf{X}$ may be out of feasible set $\mathcal{K}$. In other words, the distributions of recovered source- and target- domain data may be very different, which may make cross-domain knowledge transfer unsuccessful. Thus, the projection on the $\mathcal{K}$ -space is required. The operator of $\mathcal{K}$ -space projection for any matrix $\mathbf{X}$ is defined as

$$\prod_{\mathcal{K}}(\mathbf{X}) = \arg \min_{\hat{\mathbf{X}} \in \mathcal{K}} h(\hat{\mathbf{X}}), \quad (5.7)$$

where $h(\hat{\mathbf{X}}) = \frac{1}{2} \|\hat{\mathbf{X}} - \mathbf{X}\|_F^2$. The optimization problem (5.7) is equivalent to the following problem:

$$\min_{\hat{\mathbf{X}}} h(\hat{\mathbf{X}}) = \frac{1}{2} \|\hat{\mathbf{X}} - \mathbf{X}\|_F^2 + \lambda \text{tr}(\mathbf{K}_{\hat{\mathbf{X}}} \mathbf{L}). \quad (5.8)$$

Though (5.8) is nonconvex due to some negative terms in $\mathbf{L}$, it can be formulated as a “difference of convex” (d.c.) program, which can be solved by the concave-convex procedure (CCCP) [107] with global convergence guarantee [88] to the minimum or stationary point if the objective function is smooth. To apply CCCP to solve the problem, we further rewrite (5.8) as follows.

$$\begin{aligned} \min_{\hat{\mathbf{X}}} h(\hat{\mathbf{X}}) &= \frac{1}{2} \|\hat{\mathbf{X}} - \mathbf{X}\|_F^2 + \lambda (\text{tr}(\mathbf{K}_{\hat{\mathbf{X}}} \mathbf{L}^+) - \text{tr}(\mathbf{K}_{\hat{\mathbf{X}}} \mathbf{L}^-)), \\ &= u(\hat{\mathbf{X}}) - v(\hat{\mathbf{X}}), \end{aligned} \quad (5.9)$$

where $\mathbf{L}^+ = \begin{pmatrix} \mathbf{L}_{S,S} & \mathbf{0} \\ \mathbf{0} & \mathbf{L}_{T,T} \end{pmatrix}$, $\mathbf{L}^- = \begin{pmatrix} \mathbf{0} & -\mathbf{L}_{S,T} \\ -\mathbf{L}_{T,S} & \mathbf{0} \end{pmatrix}$, $u(\hat{\mathbf{X}}) = \frac{1}{2} \|\hat{\mathbf{X}} - \mathbf{X}\|_F^2 + \lambda \text{tr}(\mathbf{K}_{\hat{\mathbf{X}}} \mathbf{L}^+)$, and $v(\hat{\mathbf{X}}) = \lambda \text{tr}(\mathbf{K}_{\hat{\mathbf{X}}} \mathbf{L}^-)$. Note that both $u(\hat{\mathbf{X}})$ and $v(\hat{\mathbf{X}})$ are convex.

The concave-convex procedure (CCCP) aims to solve d.c. problems of the following form.

$$\min_{\mathbf{d}} u(\mathbf{d}) - v(\mathbf{d}). \quad (5.10)$$

If $v(\mathbf{d})$ is smooth, then the updates of $\mathbf{d}$ can be achieved by the majorization-minimization algorithm [88] as follows:

$$\mathbf{d}^{(l+1)} \leftarrow \arg \min_{\mathbf{d}} u(\mathbf{d}) - \mathbf{d}^\top \nabla v(\mathbf{d}^{(l)}). \quad (5.11)$$

However, all the existing CCCP solvers are based on the assumption that $\mathbf{d} \in \mathbb{R}^d$. Here, we extend CCCP in a matrix-form as follows:

$$\mathbf{X}^{(l+1)} \leftarrow \arg \min_{\mathbf{X}} u(\mathbf{X}) - \text{tr}(\nabla v(\mathbf{X}^{(l)})^\top \mathbf{X}). \quad (5.12)$$

By substituting (5.12) into (5.9), we obtain

$$\hat{\mathbf{X}}^{(l+1)} \leftarrow \arg \min_{\hat{\mathbf{X}}} \|\hat{\mathbf{X}} - \mathbf{X}\|_F^2 + \lambda \text{tr}(\mathbf{K}_{\hat{\mathbf{X}}} \mathbf{L}^+) - \lambda \text{tr}(\nabla v(\hat{\mathbf{X}}^{(l)})^\top \hat{\mathbf{X}}), \quad (5.13)$$

where the gradient of $v(\hat{\mathbf{X}})$ ($\nabla v(\hat{\mathbf{X}}^{(l)})$) with respect to the t -th row of the matrix $\hat{\mathbf{X}}$ (i.e., $\hat{\mathbf{x}}_t$) is calculated as follows:

$$\frac{\partial v(\hat{\mathbf{X}}^{(l)})}{\partial \hat{\mathbf{x}}_t} = -2 \sum_{i=1}^n \mathbf{L}_{ti}^- \exp(-\|\hat{\mathbf{x}}_t^{(l)} - \hat{\mathbf{x}}_i^{(l)}\|^2 / 2\delta) \times \frac{\hat{\mathbf{x}}_t^{(l)} - \hat{\mathbf{x}}_i^{(l)}}{\delta}.$$

Note that (5.13) is convex. Hence, the global optimal solution can be achieved by performing gradient descent approaches. For some applications with nonnegative features (i.e., bag-of-words) [104], we can simply use projected gradient descent algorithm instead to project the solution into the nonnegative space.

5.3.2 Active Correspondences Construction

In the previous subsection, we describe a new HTL method DMMC, where a set of correspondences between domains are given. Here, based on DMMC, we present a strategy to actively select a batch of the most informative instances in the source domain to query their corresponding target-domain instances to construct correspondences. The motivation is that in matrix completion, some recovered “target-domain instances” based on the corresponding source-domain instances may not be precise, resulting in large distance in distributions between domains. Therefore, these corresponding instances need to be manually constructed. The overall procedure is presented in Algorithm 5 and illustrated in Figure 5.2.

Algorithm 5 Active Correspondences Construction

Input: $\mathbf{X}^{(t)}$, a set of correspondences $\mathcal{C}^{(t)}$ after t iterations, and size k .

1: Compute α for $\bar{\mathbf{x}}_{S_i} \in \bar{\mathbf{X}}_{SU}$ based on (5.15).

2: Select $\mathcal{C}_S \subseteq \mathbf{X}_{SU}$ of size k whose corresponding $\{\bar{\mathbf{x}}_{S_i}\}$'s are of smallest $\{\alpha_i\}$'s values to query their corresponding target-domain instances to form $\mathcal{C}_k$.

3: Update $\mathcal{C}^{(t+1)} = \mathcal{C}^{(t)} \cup \mathcal{C}_k$, and $\mathbf{X}_{SU} = \mathbf{X}_{SU} \setminus \mathcal{C}_S$.

Output: $\mathcal{C}^{(t+1)}$.

Figure 5.2: Heterogenous Transfer Learning through Active correspondence construction (HTLA).

Formally, the distance in distributions between the recovered source- and target-domain data $\bar{\mathbf{X}}_S$ and $\bar{\mathbf{X}}_T$ can be formulated as follows:

$$\min_{\alpha} \left\| \frac{1}{n'_S - k} \left(\sum_{\mathbf{x}_i \in \bar{\mathbf{X}}_{SLC}} \varphi(\mathbf{x}_i) + \sum_{\mathbf{x}_i \in \bar{\mathbf{X}}_{SU}} \alpha_i \varphi(\mathbf{x}_i) \right) - \frac{1}{n_T} \sum_{\mathbf{x}_i \in \bar{\mathbf{X}}_T} \varphi(\mathbf{x}_i) \right\|_{\mathcal{H}}^2, \quad (5.14)$$

$s.t. \alpha_i \in \{0, 1\}, \alpha^\top \mathbf{1} = n_S - l - k,$

where $n'_S = n_S + n_C$, $\bar{\mathbf{X}}_{SLC} = [\bar{\mathbf{X}}_{SL}^\top \bar{\mathbf{X}}_{SC}^\top]^\top$, and α is the indicator vector for source-domain unlabeled instances, where $\alpha_i = 1$ denotes the instance is selected, which implies that the correspondingly recovered target-domain instance does not make the distance in distributions between domains large, otherwise $\alpha_i = 0$, which implies that the correspondingly recovered target-domain instance makes the distributions between domains

different.

Note that the above problem is NP-hard due to the integer constraints. Therefore, we relax the problem into the following convex problem [24, 47]:

$$\min_{\alpha \in [0, 1]} \frac{1}{2} \alpha^\top \mathbf{K}_{\bar{\mathbf{X}}_{SU}} \alpha - \frac{n'_S - k}{n_T} \mathbf{k}_{\bar{\mathbf{X}}_{SU}, \bar{\mathbf{X}}_T}^\top \alpha + \mathbf{k}_{\bar{\mathbf{X}}_{SU}, \bar{\mathbf{X}}_{SLC}}^\top \alpha, \quad (5.15)$$

where $\mathbf{K}_{\bar{\mathbf{X}}_{SU}}$ is the kernel matrix for $\bar{\mathbf{X}}_{SU}$, $\mathbf{k}_{\bar{\mathbf{X}}_{SU}, \bar{\mathbf{X}}_T}(i) = \sum_j \mathbf{K}_{\bar{\mathbf{X}}_{SU}, \bar{\mathbf{X}}_T}(i, j)$, and $\mathbf{k}_{\bar{\mathbf{X}}_{SU}, \bar{\mathbf{X}}_{SLC}}(i) = \sum_j \mathbf{K}_{\bar{\mathbf{X}}_{SU}, \bar{\mathbf{X}}_{SLC}}(i, j)$. After we solve (5.15), we sort the $\{\alpha_i\}$'s in ascending order and select the top k corresponding $\mathbf{x}_{S_i} \in \mathbf{X}_{SU}$ for querying their corresponding instances in the target domain.

5.4 Experimental Results

We aim to verify the effectiveness of the proposed framework for HTL by conducting two series of experiments. The first series of experiments compares DMMC with other HTL methods when a set of cross-domain correspondences are given. The second series of experiments compares the performance of our proposed active strategy with random selection for cross-domain correspondences construction. Furthermore, we conduct extensive experiments to study sensitivity of the parameters in our proposed framework. All the experiments are conducted on a benchmark cross-language sentiment classification dataset.

5.4.1 Experiment Setting

Dataset: The cross-language sentiment dataset [75] comprises of Amazon product reviews of three product categories: books (B), DVDs (D) and music (M). These reviews are written in four languages: English (EN), German (GE), French (FR), and Japanese (JP). For each category, reviews are split into a training set and a test set, including 2,000 reviews. Moreover, for each non-English language, there are another 2,000 unlabeled correspondences (English v.s. non-English). Each review is preprocessed into a normalized TF-IDF feature vector. Since in practice, English documents are widely accessible, we use all the English reviews in the training set as the source domain labeled data, and all non-English reviews (each of the other 3 languages) in the training set as target domain

unlabeled data. The performance of all methods are evaluated on the target-domain unlabeled data without any target-domain labeled training data.

Baselines: To verify the effectiveness of DMMC for HLT given a set of correspondences, we compare it with the following state-of-the-art baselines.

SVM-SC: We first train a classifier on the source-domain labeled data, and then make predictions on the source-domain corresponding data. In this way, the predicted labels on the source-domain corresponding data is transferred to the target-domain corresponding data. Next, we train a model on the target-domain corresponding data with the predicted labels to make predictions on the target domain unlabeled data.

OPCA: We apply Oriented Principal Component Analysis (OPCA) [73] to learn projections for the source and target domains by maximizing projected covariance of bilingual document-term matrix and at same time minimizing distances between correspondences. Classifier is trained and tested on the projected data.

LSI: We apply Latent Semantic Indexing (LSI) [67] on the matrix $\mathbf{M}$ to learn low-dimensional cross-lingual representations, and then train and test a classifier on the low-dimensional data.

KCCA: We apply Kernel Canonical Component Analysis (KCCA) [94] on the cross-domain correspondences to learn projections for bilingual data, and then train and test a classifier on the projected data.

TSL: We apply the correspondence-based HTL method proposed by Xiao *et al.* [104] to train a classifier to make predictions on the target-domain unlabeled data.

Parameter Settings: For DMMC, there are three parameters, γ , μ , λ , on the regularization terms for low-rank, sparsity, and distance matching, respectively. We vary the parameter values as follows: $\gamma \in \{10^{-2}, 10^{-1}, \dots, 10^2\}$, $\mu \in \{10^{-6}, 10^{-5}, \dots, 10^{-1}\}$, and $\lambda \in \{10^{-3}, 10^{-2}, \dots, 10^1\}$. For the parameter setting of TSL, we follow the procedure in [104]. We first set $\mu = 10^{-6}$ and $\tau = 1$, and then cross validate $\gamma \in \{10^{-3}, 10^{-2}, \dots, 10^1\}$, and $\rho \in \{10^{-6}, 10^{-5}, \dots, 10^{-1}\}$. For KCCA, we tune the parameter κ ((5) in [94]) in $\{10^{-2}, 10^{-1}, \dots, 10^2\}$. For the reduced dimensions for OPCA, LSI, KCCA, TSL and DMMC, we validate in the same range of $r \in \{20, 50, 100, 200, 500\}$, and find the best rank for all the methods. For all experiments, we employ linear support vector machines (SVMs) [22] with default parameter settings as the base classifier.

Table 5.1: Comparison results in classification accuracy (%).

TASK	SVM-SC	OPCA	KCCA	LSI	TSL	DMMC
B-FR	72.35	68.70	69.75	71.56	73.95	76.52
D-FR	74.50	65.45	72.55	71.22	74.30	76.23
M-FR	67.40	67.95	70.17	67.39	71.15	74.05
B-GE	74.12	69.00	73.69	72.43	75.98	77.47
D-GE	73.75	71.63	74.79	74.76	76.01	78.28
M-GE	73.25	65.12	69.86	72.18	74.57	76.60
B-JP	60.83	58.69	62.02	62.22	65.81	68.54
D-JP	69.36	66.60	66.17	66.47	70.72	72.12
M-JP	65.59	64.40	63.80	65.54	68.22	71.37

5.4.2 Performance Comparison

In the first series of experiments, we compare the performances in terms of classification accuracy of different methods given all the 2,000 cross-domain correspondences in the dataset. The overall results on 9 tasks, e.g., B-FR, D-GE, etc. are shown in Table 5.1, where for example B-FR denotes the task with English books reviews as source domain and French books reviews as target domain. We can see that DMMC outperforms all the baselines. SVM-SC performs slightly better than OPCA, LSI and KCCA on average due to the sufficient set of cross-domain correspondences. LSI performs poorly without filling the missing values via matrix completion. Therefore, matrix completion is necessary before dimension reduction. DMMC explicitly minimizes the distance in distributions of the recovered data between domains, which improve classification accuracy by around 3% over TSL with all the correspondences. To further analyze the effect of the MMD-based regularization term, we show how the accuracy varies along different correspondence sizes for TSL and DMMC in Figure 5.3.a. DMMC outperforms TSL significantly by around 5% especially when the correspondence size is small.

5.4.3 Effectiveness of Active Learning Strategy

In the second series of experiments, we aim to verify the effectiveness of the active learning strategy for cross-domain correspondences construction. We denote by DMMC-active and DMMC-rand DMMC with active and random correspondences construction strategies, respectively. We set $k = 100$, and show the comparison results in Figure 5.3.b,

5.3.a: TSL and DMMC.

5.3.b: Random and Active Correspondence.

Figure 5.3: Performance Comparison.

where FR, GE, and JP denote the average results of the tasks with FP, GE, and JP as the target domains, respectively. We observe that DMMC-active consistently and significantly outperforms DMMC-rand on all the tasks especially when the correspondence size is small. Moreover, the performance of DMMC-active is more stable than DMMC-rand because DMMC-active aims to choose the most informative correspondences to update the model in each iteration.

5.4.4 Parameter Sensitivity Study

For the final series of experiments, we study sensitivity of the three parameters, γ , μ , λ , in DMMC. In these experiments, we fix the correspondences size $n_c = 2,000$ and report average results of each non-English language as the target domain in Figure 5.4. In Figure 5.4.a, we show the classification accuracy under varying values of γ in the range $[10^{-2} \ 10^2]$ by fixing $\mu = 10^{-4}$ and $\lambda = 10^{-2}$. We observe that classification accuracy drops fast when γ is small (≤ 1). This implies that the low-rank regularization is important because text data can be represented by only a few latent topics. In Figure 5.4.b, we vary μ in the range $[10^{-6} \ 10^{-1}]$ by fixing $\gamma = 10$, $\lambda = 10^{-2}$. We observe that if μ is set to a large value, then the model may overfit to the observed nonzero entries, and lead to zero values for many covered entries. In Figure 5.4.c, we show classification accuracy

5.4.a: Low Rank

5.4.b: Sparsity

5.4.c: Distribution Matching

Figure 5.4: Parameter Analysis.

under different λ in the range $[10^{-3} \ 10^1]$ by fixing $\gamma = 10$, $\mu = 10^{-4}$. The figure shows that the parameter λ should be neither too large nor small to reach satisfactory performance.

5.5 Summary

Different from previous work that assumes a sufficient set of cross-domain correspondences be given in advance, we propose to actively construct cross-domain correspondences based on limited budget. Under the framework, we propose a general correspondence-based HTL method. Based on the HTL method, we propose an active learning strategy for query construction of correspondences. We conduct extensive experiments on a benchmark dataset on multi-language sentiment classification to demonstrate the superiority of the proposed method over a number of state-of-the-art baseline methods.

Chapter 6

Conclusions and Future Work

In this thesis, we identify and study five critical issues related to heterogeneous transfer learning (HTL): sparse feature mapping, multi-class information, hybrid transfer, distribution matching and active correspondences construction. We address the two issues of sparsity and multi-class information in HTL by transforming the HTL into sparse coding with the help of Error Correcting Output Codes (ECOC) to generate sufficient measurements. For hybrid transfer, we consider two cases: 1) only distribution drift in the source domain; 2) distribution drifts in both source and target domain. Then, we propose to correct distribution bias within domains and at the same time to learn the deep feature mappings through deep learning approach. For distribution matching for HTL, a novel Maximum Mean Discrepancy (MMD)-regularized matrix completion is proposed. Furthermore, we use MMD as the criterion for evaluating correspondences reconstruction quality to actively construct the cross-domain correspondences.

6.1 Summary of Contributions

6.1.1 Multi-class Heterogeneous Transfer Learning

Motivated by the language translation task where a word or phrase in one language is translated to a few words or phrases in another language, we propose to learn a sparse feature transformation matrix to map features between domains. This objective can be formulated as a compressed sensing (CS) problem, where each binary classifier constructed by decomposing a target multi-class classification problem can be considered as measurement sensors. A theoretical study shows that the estimation error of the

learned transformation matrix decreases with the increasing number of the constructed binary classifiers, and a generalization error bound of HTL under the multi-class setting can be derived. Based on this theoretical support, in order to obtain a good reconstruction of the transformation matrix between different-domain features, we need to construct a sufficient number of binary classifiers from the original multi-class classification problem.

The main research contributions are summarized as follows.

- We propose a novel **sparse heterogeneous feature representation** (SHFR) scheme to learn the sparse transformations for HTL, by fully exploring the common underlying structures among multiple classes of two domains.
- We show theoretically that a sparse feature mapping can be learned if and only if a sufficient number of classifiers are provided for the HTL task. Hence, we propose an approach that generates sufficient classifiers under Error Correcting Output Correcting **ECOC** scheme to accurately estimate the sparse feature mapping.
- We provide the first **generalization error bound** for HTL in binary case. Furthermore, based on the analysis of the binary case, we prove how the generalization error of multi-class HTL problems is affected by the number of tasks as well as the sparsity of the feature mapping matrix.
- We present a **batch-mode** pursuit method that efficiently solve the resultant non-negative sparse recovery problem used by SHFR. The computational cost of SHFR is independent of the data size, and scales linearly only with the target feature dimensions.
- Extensive experimental results on the three cross-language classification datasets demonstrate that the proposed SHFR is more effective and efficient than other state-of-the-art baselines.

6.1.2 Hybrid Heterogeneous Transfer Learning

Hybrid heterogeneous transfer learning (HHTL) is motivated by the cross-language classification task. Due to limitation in cross-language translations, the “bias” issue arises

in word-distribution between the translated documents and the original ones in the same domains. However, most previous heterogeneous transfer learning methods assume these cross-language translations are representative in the source and target domains. As a result, a classifier trained on the labeled transformed-source-domain data may not be useful for the target domain.

The main research contributions are as follows.

- We propose the new *hybrid heterogeneous transfer learning* (HHTL) framework to handle the distribution bias within the same feature spaces and transfer knowledge across heterogeneous spaces simultaneously,
- We propose an efficient deep learning approach comprising of high-level homogeneous feature learning and deep heterogeneous feature mapping. High-level homogeneous features are learned from the marginalized stacked denoised autoencoder (mSDA) to reduce the distribution bias and at the same time to generate the deep cross-domain feature mapping by quadratic regression.
- Extensive experimental results on several multilingual sentiment classification tasks verify the effectiveness of our proposed approach compared with some baseline methods. Moreover, the results of different layers show that the deep structures can help to reduce the distribution bias and to learn better cross-domain feature mapping for knowledge transfer simultaneously.

6.1.3 Heterogeneous Transfer Learning through Active Correspondences Construction

For the cross-language classification task, the cross-domain correspondences, namely document-level translation between languages, are often very expensive to obtain. To reduce the cost of the construction of such correspondences, we propose a novel and general HTL framework named Heterogeneous Transfer Learning through Active correspondence construction (HTLA).

The main research contributions are listed as follows:

- We propose a general correspondence-based method for HTL through matrix completion, where we explicitly penalize the distribution matching between two domains during the matrix completion.
- We propose a novel framework to actively construct correspondences between domains for effective knowledge transfer. Specifically, we use the Maximum Mean Discrepancy (MMD) as the criteria to actively select training instances from target domain.
- Two experimental settings are used to compare the proposed approach with various baseline methods. The first one is the traditional HTL setting with given cross-domain correspondences. The results show our proposed method outperforms other state-of-the-art methods by explicitly matching two heterogeneous domains. The second setting includes the comparison between active correspondences and randomly given correspondences. We observe from the results that actively constructing correspondences achieve a better performance than the passive paradigm.

6.2 Future Work

6.2.1 More Discriminative Output Codes for Multi-class HTL

In Chapter 3, we show that when more tasks are generated, better sparse feature heterogeneous representation mapping can be learnt. Although the investigation has shown that error correcting output codes (ECOC) scheme generates many more binary codes than one-versus-all scheme, the capability of error correction heavily relies on the minimum hamming distance between any distinct pair of rows in the coding matrix M , denoted as $\Delta_{\min}(M)$ [33]. As shown in [1], a larger $\Delta_{\min}(M)$ is more likely to rectify the errors committed by individual base classifiers. Therefore, ECOC with longer random codes lead to more robust and better predictions. However, the choice of existing codes hinders the creation of more discriminative codes.

For more flexibility in constructing longer and more discriminative codes, we are going to extend the existing binary or ternary coding scheme to N -ary coding scheme. One example of N -ary coding matrix to represent nine classes is shown in Table 6.1.

Table 6.1: An example of N -ary coding matrix M , where $N = 4$.

	M_1	M_2	M_3	M_4	M_5	M_6
C_1	1	1	2	4	1	1
C_2	2	1	1	3	2	1
C_3	3	2	1	2	3	1
C_4	4	3	1	1	4	2
C_5	4	3	2	2	4	3
C_6	4	3	3	3	3	4
C_7	3	4	4	4	2	4
C_8	2	4	3	4	2	3
C_9	3	4	2	3	3	2

Unlike traditional ECOC frameworks, a row of this coding matrix M represents the code of each class, and the code consists of N_L numbers in $\{1 \cdots N\}$, where $N > 3$; while one column M_s of M represents the N partitions of classes to be considered in the s^{th} column. The N -ary coding scheme has two major advantages over conventional ECOC: 1) It can generate more base tasks than ECOC for reducing error of estimating cross-domain feature mapping. 2) N -ary codes encode multiple classes for inducing better multi-class classifications.

6.2.2 Unsupervised Heterogeneous Transfer Learning

In this thesis, we study HTL for classification problems. As we discussed in Chapter 1, in many real-world applications the “bridge” information e.g. labels or correspondences are expensive to obtain. In our future work, we will try to further relax the “bridges” information to be cluster-dependency. This idea is taken from cross-guided clustering approach [10] which aims to enhance clustering results on target domain data with the help of source supervision. We plan to study and investigate a more challenging HTL problem called *unsupervised heterogeneous transfer learning* when there is no labeled target domain data or one-to-one cross-domain correspondences. In this case, we assume that the dependency across the clusters in the source domain remains the “same” as the target domain. This is illustrated in Figure 6.1. This assumption is reasonable in the real-world problems. Take the text-based image classification as an example, where given the unlabeled text description on the apple, pear and car, we aim to cluster all the relevant images into the corresponding three groups. Apple is a type of fruit that

resembles pear in shape and size. In contrast, the apple and pear are very different from the car in every aspect. Therefore, the apple and pear clusters should be closer to each other and farther away from the car cluster in the feature space. In other words, given the data structure in Figure 6.1.a, we aim to discover the corresponding clusters in the target domain (see Figure 6.1.b), where the cluster structure or dependency remains invariant to the rotations, expansion and contraction.

6.1.a: Dependency across the source domain clusters

6.1.b: Dependency across the target domain clusters

Figure 6.1: Illustration of unsupervised heterogeneous transfer learning.

6.2.3 Theoretical Studies on HTL in Active Learning Paradigm

In Chapter 5, we proposed a general HTLA framework to integrate active learning to construct correspondences between two heterogeneous domains for HTL. However, there are remaining problems concerning HTL under the active learning paradigm that need to be further investigated. Besides the correspondences, labels from the target domain are another valuable source of information to connect the two domains. However, the labels are usually obtained with high cost. Thus, the approach to actively select the queries from target domain becomes an important issue. Moreover, further theoretical studies on the sample complexity of HTL are required to better understand the HTL. In other words, we need to investigate the maximum number of cross-domain correspondences or the target domain labeled data that we need in order to achieve the desired performance.

Publications

- (i) **Joey Tianyi Zhou**, Ivor W. Tsang, Sinno Jialin Pan and Mingkui Tan. *Heterogeneous Domain Adaptation for Multiple Classes*. In Proceedings of the 17th International Conference on Artificial Intelligence and Statistics (AISTATS-14), Reykjavik, Iceland. Apr. 22-25, 2014.
- (ii) **Joey Tianyi Zhou**, Sinno Jialin Pan, Ivor W. Tsang, and Yan Yan. *Hybrid Heterogeneous Transfer Learning through Deep Learning*. In Proceedings of the 28th AAAI Conference on Artificial Intelligences (AAAI-14), Quebec City, Quebec, Canada. July 27-31, 2014.
- (iii) Sharath Guntuku*, **Joey Tianyi Zhou***, Sujoy Roy, Weisi Lin and Ivor W. Tsang. *Deep Representations to Model User 'Likes'*. In Asian Conference on Computer Vision (ACCV-2014), Singapore, 2014 (*denotes equal contribution).
- (iv) **Joey Tianyi Zhou**, Sinno Jialin Pan, Ivor W. Tsang and Shen-shyang Ho. *Transfer Learning for for Cross-Language Text Categorization through Active Correspondences Construction*. In Proceedings of the 30th AAAI Conference on Artificial Intelligences (AAAI-16), under submission.
- (v) **Joey Tianyi Zhou**, Ivor W. Tsang, Sinno Jialin Pan and Mingkui Tan. *Multi-class Heterogeneous Domain Adaptation*. Journal of Machine Learning Research (JMLR), in revision.

References

- [1] E. L. Allwein, R. E. Schapire, and Y. Singer. Reducing multiclass to binary: a unifying approach for margin classifiers. *J. Mach. Learn. Res.*, 1:113–141, Sept. 2001.
- [2] 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.
- [3] A. Argyriou, T. Evgeniou, and M. Pontil. Multi-task feature learning. In *NIPS*, pages 41–048. MIT Press, 2007.
- [4] A. Argyriou, M. Pontil, Y. Ying, and C. A. Micchelli. A spectral regularization framework for multi-task structure learning. In *Advances in neural information processing systems*, pages 25–32, 2007.
- [5] H. Avron, S. Kale, S. P. Kasiviswanathan, and V. Sindhvani. Efficient and practical stochastic subgradient descent for nuclear norm regularization. In *ICML*, 2012.
- [6] Y. Aytar and A. Zisserman. Tabula rasa: Model transfer for object category detection. In *ICCV*, pages 2252–2259. IEEE, 2011.
- [7] M. Belkin, P. Niyogi, and V. Sindhvani. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *J. Mach. Learn. Res.*, 7:2399–2434, Dec. 2006.
- [8] S. Ben-David, J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J. W. Vaughan. A theory of learning from different domains. *Mach. Learn.*, 79(1-2):151–175, May 2010.

- [9] Y. Bengio, P. Lamblin, D. Popovici, H. Larochelle, et al. Greedy layer-wise training of deep networks. 2007.
- [10] I. Bhattacharya, S. Godbole, S. Joshi, and A. Verma. Cross-guided clustering: Transfer of relevant supervision across domains for improved clustering. In *Data Mining, 2009. ICDM'09. Ninth IEEE International Conference on*, pages 41–50. IEEE, 2009.
- [11] J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J. Wortman. Learning bounds for domain adaptation. In *NIPS*, 2007.
- [12] J. Blitzer, M. Dredze, and F. Pereira. Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *ACL*, pages 432–439, 2007.
- [13] J. Blitzer, R. McDonald, and F. Pereira. Domain adaptation with structural correspondence learning. In *EMNLP*, pages 120–128, 2006.
- [14] A. Blum and T. M. Mitchell. Combining labeled and unlabeled data with co-training. In *COLT*, pages 92–100, 1998.
- [15] K. M. Borgwardt, A. Gretton, M. J. Rasch, H.-P. Kriegel, B. Schölkopf, and A. J. Smola. Integrating structured biological data by kernel maximum mean discrepancy. *Bioinformatics*, 22(14):e49–e57, 2006.
- [16] L. Bottou. Large-scale machine learning with stochastic gradient descent. In *COMPSTAT*, pages 177–186. 2010.
- [17] O. Bousquet, S. Boucheron, and G. Lugosi. Introduction to statistical learning theory. In *Advanced Lectures on Machine Learning*, pages 169–207. Springer, 2004.
- [18] A. M. Bruckstein, M. Elad, and M. Zibulevsky. On the uniqueness of nonnegative sparse solutions to underdetermined systems of equations. *Information Theory, IEEE Transactions on*, 54(11):4813–4820, 2008.

REFERENCES

- [19] 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, Aug. 2006.
- [20] J. Cantarella and M. Piatek. Tsnrls: A solver for large sparse least squares problems with non-negative variables. *arXiv preprint cs/0408029*, 2004.
- [21] Y. Censor and S. A. Zenios. *Parallel Optimization: Theory, Algorithms and Applications*. Oxford University Press, 1997.
- [22] C.-C. Chang and C.-J. Lin. LIBSVM: A library for support vector machines. *ACM Trans. Intell. Syst. Technol.*, 2:27:1–27:27, 2011.
- [23] O. Chapelle, B. Schölkopf, and A. Zien. *Semi-Supervised Learning*. The MIT Press, 1st edition, 2010.
- [24] R. Chattopadhyay, W. Fan, I. Davidson, S. Panchanathan, and J. Ye. Joint transfer and batch-mode active learning. In *ICML*, pages 253–261, 2013.
- [25] J. Chen, L. Tang, J. Liu, and J. Ye. A convex formulation for learning shared structures from multiple tasks. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 137–144. ACM, 2009.
- [26] M. Chen, Z. E. Xu, K. Q. Weinberger, and F. Sha. Marginalized denoising autoencoders for domain adaptation. In *ICML*, 2012.
- [27] N. Cristianini and J. Shawe-Taylor. *An introduction to support vector machines and other kernel-based learning methods*. Cambridge university press, 2000.
- [28] 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.
- [29] H. Daumé III. Frustratingly easy domain adaptation. In *ACL*, pages 256–263. ACL, June 2007.
- [30] S. Deerwester, S. Dumais, G. Furnas, T. Landauer, and R. Harshman. Indexing by latent semantic analysis. *Journal of the American society for information science*, 41(6):391–407, 1990.

- [31] T. G. Dietterich. Ensemble methods in machine learning. In *MCS*, pages 1–15. Springer-Verlag, 2000.
- [32] T. G. Dietterich and G. Bakiri. Error-correcting output codes: A general method for improving multiclass inductive learning programs. In *AAAI*, pages 572–577. AAAI Press, 1991.
- [33] T. G. Dietterich and G. Bakiri. Solving multiclass learning problems via error-correcting output codes. *J. Artif. Intell. Res.*, 2:263–286, 1995.
- [34] J. Donahue, J. Hoffman, E. Rodner, K. Saenko, and T. Darrell. Semi-supervised domain adaptation with instance constraints. In *Computer Vision and Pattern Recognition (CVPR)*, 2013.
- [35] D. L. Donoho. Compressed sensing. *IEEE Trans. Inform. Theory*, 52:1289–1306, 2006.
- [36] D. L. Donoho and J. Tanner. Counting the faces of randomly-projected hypercubes and orthants, with applications. *Discrete & computational geometry*, 43(3):522–541, 2010.
- [37] L. Duan, D. Xu, and I. W. Tsang. Learning with augmented features for heterogeneous domain adaptation. In *ICML*, 2012.
- [38] H. C. Ellis. The transfer of learning. *The Macmillan Company*, 1965.
- [39] 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.
- [40] A. Frome, G. S. Corrado, J. Shlens, S. Bengio, J. Dean, T. Mikolov, et al. Devise: A deep visual-semantic embedding model. In *Advances in Neural Information Processing Systems*, pages 2121–2129, 2013.
- [41] M. Gal-on and S. Mannor. Learning from multiple outlooks. *CoRR*, abs/1005.0027, 2010.

- [42] N. García-Pedrajas and D. Ortiz-Boyer. An empirical study of binary classifier fusion methods for multiclass classification. *Inf. Fusion*, 12(2):111–130, Apr. 2011.
- [43] R. Ghani. Using error-correcting codes for text classification. In *ICML*, pages 303–310. Morgan Kaufmann Publishers, 2000.
- [44] X. Glorot, A. Bordes, and Y. Bengio. Domain adaptation for large-scale sentiment classification: A deep learning approach. In *ICML*, pages 513–520, 2011.
- [45] G. H. Golub and C. F. Van Loan. An analysis of the total least squares problem. *SIAM Journal on Numerical Analysis*, 17(6):883–893, 1980.
- [46] M. Gönen and E. Alpaydm. Multiple kernel learning algorithms. *The Journal of Machine Learning Research*, 12:2211–2268, 2011.
- [47] B. Gong, K. Grauman, and F. Sha. Connecting the dots with landmarks: Discriminatively learning domain-invariant features for unsupervised domain adaptation. In *Proceedings of The 30th International Conference on Machine Learning*, pages 222–230, 2013.
- [48] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola. A kernel two-sample test. *The Journal of Machine Learning Research*, 13(1):723–773, 2012.
- [49] M. Harel and S. Mannor. Learning from multiple outlooks. In *ICML*, pages 401–408, 2011.
- [50] T. Hastie, R. Tibshirani, and J. Friedman. *The elements of statistical learning: data mining, inference and prediction*. Springer, 2 edition, 2008.
- [51] G. E. Hinton and R. R. Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science*, 313(5786):504–507, 2006.
- [52] C. Hu, W. Pan, and J. T. Kwok. Accelerated gradient methods for stochastic optimization and online learning. In *NIPS*, pages 781–789, 2009.
- [53] J. Huang, A. J. Smola, A. Gretton, K. M. Borgwardt, and B. Schölkopf. Correcting sample selection bias by unlabeled data. In *NIPS*, pages 601–608, 2007.

- [54] R. Kiros, R. Salakhutdinov, and R. S. Zemel. Multimodal neural language models. In *Proceedings of the 31th International Conference on Machine Learning, ICML 2014, Beijing, China, 21-26 June 2014*, pages 595–603, 2014.
- [55] 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.
- [56] H. Larochelle and Y. Bengio. Classification using discriminative restricted boltzmann machines. In *Proceedings of the 25th international conference on Machine learning*, pages 536–543. ACM, 2008.
- [57] D. D. Lee and H. S. Seung. Algorithms for non-negative matrix factorization. In *Advances in neural information processing systems*, pages 556–562, 2001.
- [58] H. Lee, C. Ekanadham, and A. Y. Ng. Sparse deep belief net model for visual area v2. In *Advances in neural information processing systems*, pages 873–880, 2008.
- [59] K. Lee, D. Palsetia, R. Narayanan, M. M. A. Patwary, A. Agrawal, and A. Choudhary. Twitter trending topic classification. In *Data Mining Workshops (ICDMW), 2011 IEEE 11th International Conference on*, pages 251–258. IEEE, 2011.
- [60] S.-I. Lee, V. Chatalbashev, D. Vickrey, and D. Koller. Learning a meta-level prior for feature relevance from multiple related tasks. In *Proceedings of the 24th international conference on Machine learning*, pages 489–496. ACM, 2007.
- [61] D. D. Lewis and W. A. Gale. A sequential algorithm for training text classifiers. In *Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 3–12. Springer-Verlag New York, Inc., 1994.
- [62] F. Li, S. J. Pan, O. Jin, Q. Yang, and X. Zhu. Cross-domain co-extraction of sentiment and topic lexicons. In *ACL*, pages 410–419, 2012.
- [63] Y. Mansour, M. Mohri, and A. Rostamizadeh. Domain adaptation: Learning bounds and algorithms. In *COLT*, 2009.

- [64] W. Medhat, A. Hassan, and H. Korashy. Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal*, 5(4):1093–1113, 2014.
- [65] G. A. Montazer, S. Escalera, et al. Error correcting output codes for multiclass classification: Application to two image vision problems. In *AISP*, pages 508–513. IEEE, 2012.
- [66] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng. Multimodal deep learning. In *ICML*, pages 689–696, 2011.
- [67] J.-Y. Nie, M. Simard, P. Isabelle, and R. Durand. Cross-language information retrieval based on parallel texts and automatic mining of parallel texts from the web. In *SIGIR*, pages 74–81, 1999.
- [68] S. J. Pan, J. T. Kwok, and Q. Yang. Transfer learning via dimensionality reduction. In *AAAI*, pages 677–682, July 2008.
- [69] 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.
- [70] S. J. Pan and Q. Yang. A survey on transfer learning. *IEEE Trans. Knowl. Data Eng.*, 22(10):1345–1359, October 2010.
- [71] W. Pan, E. W. Xiang, N. N. Liu, and Q. Yang. Transfer learning in collaborative filtering for sparsity reduction. In *Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2010, Atlanta, Georgia, USA, July 11-15, 2010*, 2010.
- [72] B. Pang, L. Lee, and S. Vaithyanathan. Thumbs up?: sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*, pages 79–86. Association for Computational Linguistics, 2002.
- [73] J. C. Platt, K. Toutanova, and W.-t. Yih. Translingual document representations from discriminative projections. In *EMNLP*, pages 251–261, 2010.

- [74] J. M. Ponte and W. B. Croft. A language modeling approach to information retrieval. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '98, pages 275–281, New York, NY, USA, 1998. ACM.
- [75] P. Prettenhofer and B. Stein. Cross-language text classification using structural correspondence learning. In *ACL*, pages 1118–1127, 2010.
- [76] P. Rai, A. Saha, H. Daumé III, and S. Venkatasubramanian. Domain adaptation meets active learning. In *NAACL HLT 2010 Workshop on Active Learning for Natural Language Processing*, pages 27–32, 2010.
- [77] R. Raina, A. Battle, H. Lee, B. Packer, and A. Y. Ng. Self-taught learning: transfer learning from unlabeled data. In *ICML*, pages 759–766, 2007.
- [78] S. RajaManickam. *Efficient algorithms for sparse singular value decomposition*. PhD thesis, University of Florida, 2009.
- [79] P. Rigollet. Generalization error bounds in semi-supervised classification under the cluster assumption. In *Journal of Machine Learning Research*, pages 1369–1392, 2007.
- [80] U. Rückert and S. Kramer. Kernel-based inductive transfer. In *Machine Learning and Knowledge Discovery in Databases*, pages 220–233. Springer, 2008.
- [81] K. Saenko, B. Kulis, M. Fritz, and T. Darrell. Adapting visual category models to new domains. In *ECCV*, pages 213–226. Springer-Verlag, 2010.
- [82] B. Settles. Active learning literature survey. *University of Wisconsin, Madison*, 52:55–66.
- [83] X. Shi, W. Fan, and J. Ren. Actively transfer domain knowledge. In *ECML/PKDD*, pages 342–357. 2008.
- [84] 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.

REFERENCES

- [85] M. Slawski and M. Hein. Sparse recovery by thresholded non-negative least squares. In *Advances in Neural Information Processing Systems*, pages 1926–1934, 2011.
- [86] M. Slawski, M. Hein, et al. Non-negative least squares for high-dimensional linear models: Consistency and sparse recovery without regularization. *Electronic Journal of Statistics*, 7:3004–3056, 2013.
- [87] R. Socher, M. Ganjoo, C. D. Manning, and A. Y. Ng. Zero-shot learning through cross-modal transfer. In *NIPS*, pages 935–943, 2013.
- [88] B. K. Sriperumbudur and G. R. G. Lanckriet. On the convergence of the concave-convex procedure. In *NIPS*, pages 1759–1767, 2009.
- [89] M. Sugiyama, S. Nakajima, H. Kashima, P. V. Buenau, and M. Kawanabe. Direct importance estimation with model selection and its application to covariate shift adaptation. In *Advances in neural information processing systems*, pages 1433–1440, 2008.
- [90] M. Tan, I. W. Tsang, and L. Wang. Matching pursuit LASSO part I: sparse recovery over big dictionary. *IEEE Transactions on Signal Processing*, 63(3):727–741, 2015.
- [91] M. Tan, I. W. Tsang, and L. Wang. Matching pursuit LASSO part II: applications and sparse recovery over batch signals. *IEEE Transactions on Signal Processing*, 63(3):742–753, 2015.
- [92] K.-C. Toh and S. Yun. An accelerated proximal gradient algorithm for nuclear norm regularized linear least squares problems. *Pacific Journal of Optimization*, 6(615-640):15, 2010.
- [93] P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol. Extracting and composing robust features with denoising autoencoders. In *ICML*, pages 1096–1103, 2008.
- [94] A. Vinokourov, J. Shawe-Taylor, and N. Cristianini. Inferring a semantic representation of text via cross-language correlation analysis. In *NIPS*, pages 1473–1480, 2002.

- [95] C. Wang and S. Mahadevan. A general framework for manifold alignment. In *AAAI Fall Symposium: Manifold Learning and Its Applications*, 2009.
- [96] C. Wang and S. Mahadevan. Heterogeneous domain adaptation using manifold alignment. In *IJCAI*, pages 1541–1546, 2011.
- [97] J. Wang and J. Ye. Two-layer feature reduction for sparse-group lasso via decomposition of convex sets. In *NIPS*, pages 2132–2140, 2014.
- [98] M. Wang and A. Tang. Conditions for a unique non-negative solution to an underdetermined system. In *Communication, Control, and Computing, 2009. Allerton 2009. 47th Annual Allerton Conference on*, pages 301–307. IEEE, 2009.
- [99] M. Wang, W. Xu, and A. Tang. A unique nonnegative solution to an underdetermined system: From vectors to matrices. *Signal Processing, IEEE Transactions on*, 59(3):1007–1016, 2011.
- [100] X. Wang, T.-K. Huang, and J. Schneider. Active transfer learning under model shift. In *ICML*, 2014.
- [101] Z. Wang, K. Crammer, and S. Vucetic. Breaking the curse of kernelization: Budgeted stochastic gradient descent for large-scale svm training. *J. Mach. Learn. Res.*, 13(1):3103–3131, 2012.
- [102] G. A. Watson. Characterization of the subdifferential of some matrix norms. *Linear Algebra and its Applications*, 170:33–45, 1992.
- [103] L. Wu, Y. Yang, and H. Liu. Nonnegative-lasso and application in index tracking. *Computational Statistics & Data Analysis*, 70:116–126, 2014.
- [104] M. Xiao and Y. Guo. A novel two-step method for cross language representation learning. In *NIPS*, pages 1259–1267, 2013.
- [105] Q. Yang, Y. Chen, G.-R. Xue, W. Dai, and Y. Yu. Heterogeneous transfer learning for image clustering via the socialweb. In *ACL/IJCNLP*, pages 1–9, 2009.

REFERENCES

- [106] 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.
- [107] A. L. Yuille and A. Rangarajan. The concave-convex procedure. *Neural Comput.*, 15(4):915–936, 2003.
- [108] D. Zhai, B. Li, H. Chang, S. Shan, X. Chen, and W. Gao. Manifold alignment via corresponding projections. In *Proceedings of the British Machine Vision Conference*, pages 3.1–3.11. BMVA Press, 2010. doi:10.5244/C.24.3.
- [109] 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.
- [110] T. Zhang. On the consistency of feature selection using greedy least squares regression. *J. Mach. Learn. Res.*, 10:555–568, 2009.
- [111] L. Zhao, S. J. Pan, E. W. Xiang, E. Zhong, Z. Lu, and Q. Yang. Active transfer learning for cross-system recommendation. In *AAAI*, 2013.
- [112] W. Zhao, R. Chellappa, P. J. Phillips, and A. Rosenfeld. Face recognition: A literature survey. *Acm Computing Surveys (CSUR)*, 35(4):399–458, 2003.
- [113] J. T. Zhou, S. J. Pan, Q. Mao, and I. W. Tsang. Multi-view positive and unlabeled learning. In *Proceedings of the 4th Asian Conference on Machine Learning, ACML 2012, Singapore, Singapore, November 4-6, 2012*, pages 555–570, 2012.
- [114] J. T. Zhou, S. J. Pan, I. W. Tsang, and Y. Yan. Hybrid heterogeneous transfer learning through deep learning. In *AAAI*, pages 2213–2220, 2014.
- [115] J. T. Zhou, I. W. Tsang, S. J. Pan, and M. Tan. Heterogeneous domain adaptation for multiple classes. In *AISTATS*, pages 1095–1103, 2014.
- [116] X. Zhu. Semi-supervised learning literature survey. *world*, 10:10, 2005.
- [117] 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.