Transfer Knowledge between Cities

Ying Wei[§], Yu Zheng[†], Qiang Yang[§]
[†]Microsoft Research, Beijing, China
[§]Hong Kong University of Science and Technology, Hong Kong
[§]{yweiad,qyang}@cse.ust.hk, [†]yuzheng@microsoft.com

ABSTRACT

The rapid urbanization has motivated extensive research on urban computing. It is critical for urban computing tasks to unlock the power of the diversity of data modalities generated by different sources in urban spaces, such as vehicles and humans. However, we are more likely to encounter the label scarcity problem and the data insufficiency problem when solving an urban computing task in a city where services and infrastructures are not ready or just built. In this paper, we propose a FLexible multimOdal tRAnsfer Learning (FLORAL) method to transfer knowledge from a city where there exist sufficient multimodal data and labels, to this kind of cities to fully alleviate the two problems. FLORAL learns semantically related dictionaries for multiple modalities from a source domain, and simultaneously transfers the dictionaries and labelled instances from the source into a target domain. We evaluate the proposed method with a case study of air quality prediction.

CCS Concepts

•**Information systems** → **Data mining;** *Geographic information systems;* •**Computing methodologies** → **Transfer learning;**

Keywords

Urban Computing; Multi-modality; Transfer Learning

1. INTRODUCTION

The rapid progress of urbanization has modernized people's lives, but also engendered many challenges in cites, such as traffic congestion and air pollution. Recently, the proliferation of big data in cities has fostered unprecedented opportunities to tackle these urban challenges by data science and computing technology, a.k.a., urban computing [33]. Given the complex setting of a city, we usually need to harness the diversity of data (i.e., multi-modality) to solve an urban computing problem. For example, to predict and tackle air pollution, we need to check air quality data from monitoring stations, pollution emission from factories and vehicles, land uses and meteorological data of different locations [36, 34]. To diagnose a city's noise situation, we need to consider human mobility, traffic conditions and layout of a neighborhood [35]. Thus, to unlock the power of knowledge from multiple disparate datasets (i.e., multi-modalities) is a key research problem in urban computing.

The problem becomes more challenging when we conduct urban computing in a "new" city where infrastructures and services are not ready or just built, thus the data required by a task are insufficient. For example, when we conduct air quality prediction in Baoding, we face the following two challenges as shown in Figure 1. 1) *The label scarcity problem*: the ground truth labels, i.e., air quality data, are very scarce because there exist only a few air quality monitoring stations in Baoding. 2) *The data insufficiency problem*: there are two types of insufficiency. One refers to *structured modality missing*. The taxi trajectory data (D4), characterizing the pollution emission from vehicles, are existing in Beijing but missing in Baoding. The other is *within-modality insufficiency*. The meteorology data (D3) in Baoding are not that sufficient as in Beijing due to limited weather stations.

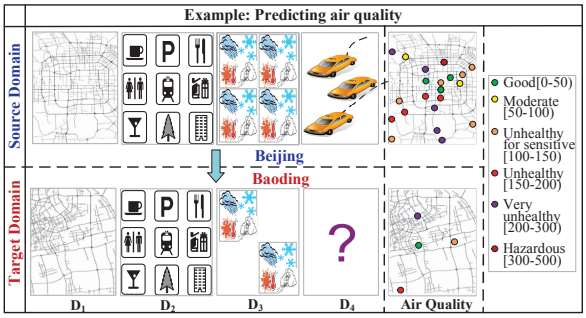


Figure 1: An example of transferring knowledge from Beijing to Baoding city.

An interesting question arises: *can we transfer knowledge from a city where data are sufficient, to a city which faces either the label scarcity or the data insufficiency problem?* As demonstrated in Figure 1, based on Beijing's data, we can learn the knowledge about underlying connections between different modalities; e.g., air pollution might be related to traffic congestion which would be caused by a dense road network structure. With such knowledge transferred from Beijing, we may be able to infer Baoding's air pollution based on road network structures even if there exists no traffic data like taxi trajectories. In this example, Beijing is a source domain where knowledge comes from, and Baoding is a target domain that we transfer knowledge to.

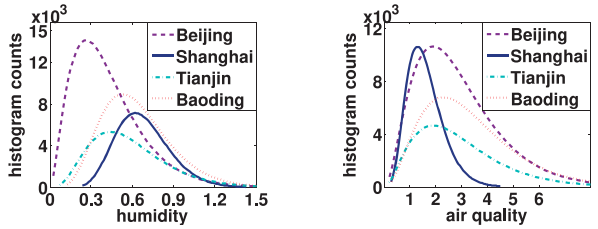
To transfer knowledge between different cities (referred to as domains in the rest of this paper) is a challenging task, as data from different cities may have different distributions in feature and label

*The paper was done when the first author was an intern in Microsoft Research under the supervision of the second author.

KDD '16, August 13-17, 2016, San Francisco, CA, USA

DOI: http://dx.doi.org/10.1145/2939672.2939830

spaces. Using the air quality inference as an example, as shown in Figure 2(a), the distributions of humidity (i.e., a kind of feature) in four cities are very different. The distributions of the four cities' air quality (i.e., labels) are also different. Though transfer learning [16] has been proposed to tackle this challenge, none of existing work can solve our problem given the following three unique challenges.



(a) Humidity distribution. (b) Air quality distribution.

Figure 2: Distribution differences across domains.

First, we transfer knowledge between source and target domains with multi-modality data rather than single-modality data. Multi-modality data have incommensurable representations. For example, the Point-Of-Interests (D2) in Figure 1 is characterized as Bool ean values indicating categories of a venue, while the meteorology (D3) is featured as real values. Simply concatenating features extracted from datasets of different modalities into a single modality compromises the performance of a transfer learning model [19, 34]. Thus, most transfer learning models [16, 28] designed for a single-modality dataset are not applicable to our problem.

Second, though a few multi-view transfer learning algorithms [5, 20, 26, 25, 30] support multi-modality data, none of them can tackle the data insufficiency problem mentioned in Figure 1. Because of within-modality insufficiency, different instances may have different modalities in a target domain. Thus, the instances cannot be treated equally. When facing the structured modality missing, we need to complement a missing modality with its knowledge representation from a source domain.

Third, data of different modalities should have different weights when transferring between different source and target domains. For example, when transferring knowledge from Beijing to Shanghai for air quality prediction, road networks may play a more important role than other modalities (like weather) as the two cities have a very similar structure of road networks (but different weather conditions). When transferring between Beijing and Tianjin (which are geographically close), however, weather conditions of the two cities are more similar than other modalities, thereby playing a more important role in the transfer. Existing transfer learning methods cannot well learn the weights for data of different modalities.

To tackle the three challenges, we propose a FLexible multi-mOdal tRAnsfer Learning (FLORAL) method with the following three contributions:

• It enforces multi-modalities to share knowledge and representation structures by learning *semantically related dictionaries* - each modality has a dictionary which consists of atoms encoding latent semantic meanings; different modalities have different dictionaries but all modalities' dictionaries share the size and latent semantic space; e.g., the third atoms of all modalities' dictionaries semantically mean "good air quality".

• It settles the data insufficiency problem, by transferring semantically related dictionaries learnt from a source to enrich feature representations of a target domain. Moreover, an algorithm called Multimodal Transfer AdaBoost (MTAB), capable of learning and differentiating different modalities' weights, is proposed to leverage labelled source instances to alleviate the label scarcity problem.

•We evaluate our method on air quality prediction in three cities, with performances outperforming six baselines.

2. RELATED WORK

In this section, we briefly review the related work in two categories: some representative research on multimodal data fusion, and state-of-the-art transfer learning methods.

2.1 Multimodal Data Fusion

There have been many attempts made towards fusing multimodal data. Some of them perform model-level fusion, i.e., generating a model for each data modality and unifying these models' outputs as the final result. Co-training [34] and multi-kernel learning [32, 36] belong to this category. The other line of research fuses different data modalities in feature level. The most naive way is to directly concatenate features from different modalities [23]. However, the performance of this method is usually inferior because it introduces overfitting and ignores non-linear interactions between modalities according to [19]. The majority of feature level fusion devote to extract a semantic latent subspace or build a translator to align different modalities. The techniques capable of aligning embrace translation [4, 21], canonical correlation analysis [6], matrix factorization [17], manifold alignment [35], coupled dictionary learning [29], and multimodal deep learning [10, 19]. Either model-level or feature-level multimodal data fusion methods require sufficient data in each modality, as well as abundant correspondence between instances across modalities. To solve urban computing tasks in a city facing the data insufficiency problem, which our work focuses on, these methods become powerless and even infeasible (imagining that a modality is missing).

2.2 Transfer Learning

Transfer learning [16] leverages knowledge from a source domain to facilitate learning in a target domain. Almost all work in this field have been motivated by the scarcity of labelled data in a target domain. Until recently, Yang et. al [28] initiated the setting called heterogeneous transfer learning which enriches the modality in a target domain with the other modality from a source by providing complementary views. This work and its follow-up [17, 22], however, can only handle the case where both source and target domains contain single modality only.

Two strands of research, i.e., multi-task multi-view learning and multi-view transfer learning, enable knowledge transfer between domains with multimodal data. Nevertheless, we first emphasize the difference between multi-task learning and transfer learning: multi-task learning assumes sufficient annotated data in each task and treats all tasks equally; while transfer learning cares only the target domain with scarce labelled data. Besides, most multi-task multi-view learning algorithms transfer model parameters, thus ignore the differences between tasks [31] or rely on enough labelled data in all tasks to learn the differences [14, 18, 24]. Though some work [7, 9, 27] transfer knowledge in feature-level, *IteM*^{2} [7] can only tackle non-negative feature values, and MAMUDA [9] and HiMLS [27] cannot fully handle the data insufficiency problem, especially the within-modality insufficiency.

To the best of our knowledge, there are only a few attempts on multi-view transfer learning. Zhang et. al [30] first proposed the MVTL-LM algorithm that transfers both model parameters and instances between domains with multi-views. The Multi-transfer [20] and DISMUTE [5] extend it to multiple source domains and multi-class classification, respectively. Blitzer et. al [2] pointed out the limitations of parameter and instance transfer in dealing with a target domain whose distribution distinctly differs from a source's.

The IMAM [26] and MDT [25] alleviate the limitations by performing feature-level knowledge transfer. Unfortunately, none of these work tackles the within-modality insufficiency, and differentiates different modalities' weights when transferring.

3. FLEXIBLE MULTIMODAL TRANSFER LEARNING

In this section, we present our method in detail. We first introduce the general framework in Figure 3, which involves two major pipelines, i.e., learning semantically related dictionaries from a source domain (represented by broken blue arrows), and transferring dictionaries and instances from a source to a target domain (shown in red solid arrows). After we introduce the notations and problem definitions, we detail how to learn semantically related dictionaries, and transfer the dictionaries and instances. The complexity analysis is given at the end of this section.

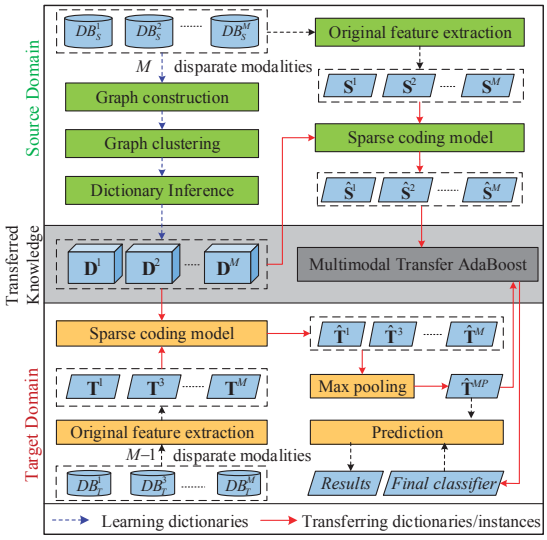


Figure 3: The framework of our proposed FLORAL method.

3.1 Overview

Learn semantically related dictionaries: To learn commensurable representations for multi-modalities, we first learn semantically related dictionaries from a source domain through a dictionary learning approach. In this approach, we build a graph that connects instances across different modalities and those in each modality. We then cluster the graph into K clusters, while ensuring that each cluster encodes a latent semantic meaning and contains instances from all modalities. Subsequently, for each modality, we build a dictionary by taking the K cluster centres of the modality as atoms. Obviously, different modalities' dictionaries have the same size K, and share the K-dimensional latent semantic space.

Transfer dictionaries and instances: To address the data insufficiency problem in a target domain, we transfer the semantically related dictionaries learnt from a source. For each modality in a target domain, we extract original features, and learn enriched representations over this modality's dictionary by sparse coding. Enriched representations make an instance more informative, thus alleviate within-modality insufficiency. As the M dictionaries may influence each other by sharing semantic meanings, the knowledge of those missing modalities (e.g., the second modality illustrated here) are preserved in the dictionaries and enriched representations of existing modalities. Therefore structured modality missing is addressed.

Transferring the dictionaries is not enough to address the label scarcity problem in a target domain. We also transfer labelled instances from a source. Before transferring, we meet the following two prerequisites: 1) learn enriched representations of labelled source instances by sparse coding, in order to make representations of source and target instances consistent; 2) perform max pooling for each target instance to aggregate enriched representations of all existing modalities, so that target instances can be treated equally regardless of within-modality insufficiency. Once these prerequisites are satisfied, we apply the Multimodal Transfer AdaBoost algorithm to transfer labelled source instances. The output of the algorithm is a classifier that can predict any target instances.

3.2 Notations and Problem Formulation

Suppose that in the target domain we are provided a very few labelled instances $\mathbf{T}_l = \{\mathbf{t}_{li}^1, \cdots, \mathbf{t}_{li}^m, \cdots, \mathbf{t}_{li}^M\}_{i=1}^{N_l^t}$ with labels $\mathbf{y} = \{y_i\}_{i=1}^{N_l^t}$ and some unlabelled instances $\mathbf{T}_u = \{\mathbf{t}_{ui}^1, \cdots, \mathbf{t}_{ui}^m, \cdots, \mathbf{t}_{ui}^M\}_{i=N_l^t+1}^{N_l^t+N_u^t}$, where $\mathbf{t}_{li}^m, \mathbf{t}_{ui}^m \in \mathbb{R}^{p^m}$ denote the feature vector of the mth modality of the ith labelled and unlabelled instance, respectively. N_l^t and N_u^t indicate the number of labelled and unlabelled instances, respectively. Meanwhile, there exists a source domain in which sufficient labelled instances $\mathbf{S}_l = \{\mathbf{s}_{lj}^1, \cdots, \mathbf{s}_{lj}^m, \cdots, \mathbf{s}_{lj}^M\}_{j=1}^{N_l^s}$ with labels $\mathbf{g} = \{g_j\}_{j=1}^{N_l^s}$ and unlabelled instances $\mathbf{S}_u = \{\mathbf{s}_{uj}^1, \cdots, \mathbf{s}_{uj}^m, \cdots, \mathbf{s}_{uj}^M\}_{j=N_l^s+1}^{N_l^s+N_u^s}$ are available. The meanings of $\mathbf{s}_{lj}^m$, $\mathbf{s}_{uj}^m$, N_l^s, N_u^s, are similar to those in the target domain. Note that M is the total number of modalities in the source domain, while in the target domain $\mathbf{t}_{li}^m$ or $\mathbf{t}_{ui}^m$ could be missing for some $1 \le m \le M$ of some $1 \le i \le N_l^t + N_u^t$ as a result of the data insufficiency. Our goal is first to learn M dictionaries $\mathbf{D}^1, \cdots, \mathbf{D}^m, \cdots, \mathbf{D}^M$ for all M modalities from $\mathbf{S}_l$ and $\mathbf{S}_u$, where $\mathbf{D}^m \in \mathbb{R}^{p^m \times K}$. Subsequently, we transfer these dictionaries to the target domain, and obtain enriched representations $\hat{\mathbf{T}}_l = \{\hat{\mathbf{t}}_{li}^1, \cdots, \hat{\mathbf{t}}_{li}^m, \cdots, \hat{\mathbf{t}}_{li}^M\}_{i=1}^{N_l^t}$ of $\mathbf{T}_l$ over the dictionaries, where $\hat{\mathbf{t}}_{li}^m \in \mathbb{R}^K$. We obtain $\hat{\mathbf{T}}_u$ and $\hat{\mathbf{S}}_l$ in the same fashion. $\hat{\mathbf{T}}_l^{MP} = \{\hat{\mathbf{t}}_{li}^{MP}\}_{i=1}^{N_l^t}$ is the max pooling result of $\hat{\mathbf{T}}_l$, by aggregating all existing modalities $\{\hat{\mathbf{t}}_{li}^m\}_{m=1}^M$ (for some $1 \le m \le M$, $\hat{\mathbf{t}}_{li}^m$ could be missing.) into $\hat{\mathbf{t}}_{li}^{MP}$ for any ith instance. The same applies to $\hat{\mathbf{T}}_u^{MP}$. Finally, we learn a classifier $h_f(\hat{\mathbf{T}}_u^{MP})$ by Multimodal Transfer AdaBoost to transfer labelled source instances, i.e., $\hat{\mathbf{S}}_l$, and adapt to target instances, i.e., $\hat{\mathbf{T}}_l$ and $\hat{\mathbf{T}}_u$. For brevity, we summarize these notations in Table 1.

Table 1: Definition of Notations

Notation	Description	No.	Set Notation
	Input		
$\mathbf{t}_{li}^m$	mth modality of ith labelled instance in the target domain	N_l^t	$\mathbf{T}_l = \{\mathbf{t}_{li}^1, \cdots, \mathbf{t}_{li}^M\}_{i=1}^{N_l^t}$
$\mathbf{t}_{ui}^m$	mth modality of ith unlabeled instance in the target domain	N_u^t	$\mathbf{T}_u = \{\mathbf{t}_{ui}^1, \cdots, \mathbf{t}_{ui}^M\}_{i=N_l^t+1}^{N_l^t+N_u^t}$
$\mathbf{s}_{lj}^m$	mth modality of jth labelled instance in the source domain	N_l^s	$\mathbf{S}_l = \{\mathbf{s}_{lj}^1, \cdots, \mathbf{s}_{lj}^M\}_{j=1}^{N_l^s}$
$\mathbf{s}_{uj}^m$	mth modality of jth unlabeled instance in the source domain	N_u^s	$\mathbf{S}_u = \{\mathbf{s}_{uj}^1, \cdots, \mathbf{s}_{uj}^M\}_{j=N_l^s+1}^{N_l^s+N_u^s}$
y_i	label of ith labelled instance in the target domain	N_l^t	$\mathbf{y} = \{y_i\}_{i=1}^{N_l^t}$
g_j	label of jth labelled instance in the source domain	N_l^s	$\mathbf{g} = \{g_j\}_{j=1}^{N_l^s}$
	Output		
$\mathbf{D}^m$	dictionary for mth modality	M	$\mathbf{D} = \{\mathbf{D}^m\}_{m=1}^M$
$\hat{\mathbf{t}}_{li}^m$	enriched representation for $\mathbf{t}_{li}^m$	N_l^t	$\hat{\mathbf{T}}_l = \{\hat{\mathbf{t}}_{li}^1, \cdots, \hat{\mathbf{t}}_{li}^M\}_{i=1}^{N_l^t}$
$\hat{\mathbf{t}}_{li}^{MP}$	max pooling of $\{\hat{\mathbf{t}}_{li}^m\}_{m=1}^M$	N_l^t	$\hat{\mathbf{T}}_l^{MP} = \{\hat{\mathbf{t}}_{li}^{MP}\}_{i=1}^{N_l^t}$
$\hat{\mathbf{t}}_{ui}^m$	enriched representation for $\mathbf{t}_{ui}^m$	N_u^t	$\hat{\mathbf{T}}_u = \{\hat{\mathbf{t}}_{ui}^1, \cdots, \hat{\mathbf{t}}_{ui}^M\}_{i=N_l^t+1}^{N_l^t+N_u^t}$
$\hat{\mathbf{t}}_{ui}^{MP}$	max pooling of $\{\hat{\mathbf{t}}_{ui}^m\}_{m=1}^M$	N_u^t	$\hat{\mathbf{T}}_u^{MP} = \{\hat{\mathbf{t}}_{ui}^{MP}\}_{i=N_l^t+1}^{N_l^t+N_u^t}$
$\hat{\mathbf{s}}_{lj}^m$	enriched representation for $\mathbf{s}_{lj}^m$	N_l^s	$\hat{\mathbf{S}}_l = \{\hat{\mathbf{s}}_{lj}^1, \cdots, \hat{\mathbf{s}}_{lj}^M\}_{j=1}^{N_l^s}$
$f(\cdot)$	classifier for $\hat{\mathbf{T}}_u^{MP}$		

3.3 Learn Semantically Related Dictionaries

Sparse coding, a technique widely used in machine learning, represents data vectors as sparse linear combinations of basis elements. The set of basis elements is called dictionary. Sparse coding provides an effective way to homogenize representation structures of multi-modalities, by enforcing all modalities' dictionaries semantically related and learning linear combination coefficients over the corresponding dictionary for each modality as new representations. There are three main categories of techniques to learn dictionaries: probabilistic learning, reconstruction error minimization, and clustering. Here we prefer clustering because of its advantage in extracting semantically related dictionaries. However, directly clustering multi-modalities in incommensurable representation structures is impossible. Instead, we propose a graph clustering algorithm as shown in Figure 4, in which we build a weighted graph to model pairwise similarities between vertices across different modalities and within each modality. Though the work [8] also learns dictionaries by graph clustering, it learns a dictionary for single modality only. Next, we will detail the graph construction on multi-modalities, the graph clustering with highly efficient submodular optimization, and the dictionaries inference.

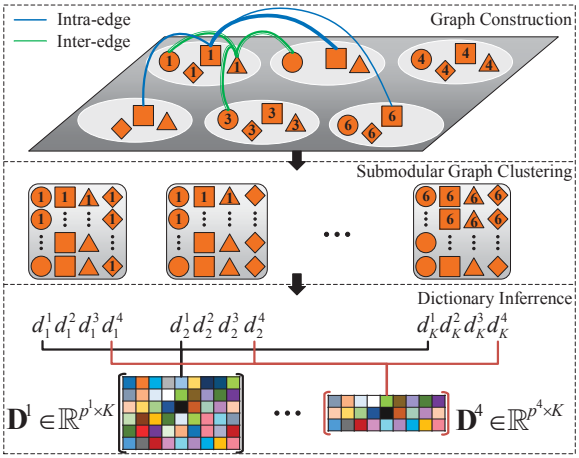


Figure 4: The procedures of dictionary learning. Different shapes represent different modalities, while the eclipses enclosing shapes denote instances. The eclipses with numbered shapes are labelled instances.

3.3.1 Graph Construction

We first build an undirected graph $G = (V, E)$. The vertex set V consists of all modalities of all instances in the source domain, i.e., $V = \mathbf{S}_l \cup \mathbf{S}_u$. We denote $|V|$, $|V^m|$, $|V_l|$ and $|V_u|$ as the number of all vertices, vertices in the mth modality, labelled vertices and unlabelled vertices, respectively. The edge set E models pairwise relations between vertices within each modality, i.e., intra-edges, and across different modalities, i.e., inter-edges.

For a pair of vertices $\mathbf{s}_i^m$ and $\mathbf{s}_j^m$ in the mth modality, we measure their similarity with the Euclidean distance between their feature vectors. The ith and jth vertices are connected with an intra-edge if each of them is among the top k similar vertices of the other vertex. This way of constructing intra-edges, i.e., mutual k-NN, has been proved to outperform traditional k-NN in semi-supervised clustering [15]. To weight each intra-edge, we apply Gaussian kernels to the similarity between two end vertices of the edge:

$$w_{ij}^m = \exp^{-\frac{\|\mathbf{s}_i^m - \mathbf{s}_j^m\|^2}{2\delta^2}}. \quad (1)$$

The more similar $\mathbf{s}_i^m$ and $\mathbf{s}_j^m$ are, the larger the weight of the intra-edge connecting them is.

As for a pair of vertices $\mathbf{s}_i^m$ and $\mathbf{s}_j^n$ in the mth and nth modality, respectively, we connect them with an inter-edge whose weight equals to 1, i.e., $w_{ij}^{m,n} = 1$, if the ith and jth instances are known to be correlated. The correlation depends on specific applications. In air quality prediction, a region (denoted by an eclipse in Figure 4) is an instance. Therefore the ith and jth instances are correlated if the two corresponding regions are geographical neighbours.

3.3.2 Submodular Graph Clustering

A natural idea of graph clustering is to partition sparsely connected dense subgraphs from each other based on the notion of intra-cluster density versus inter-cluster sparsity. Given a graph $G(V, E)$, we select $A \subseteq E$, so that the resulting graph $G(V, A)$ contains exactly K connected components. Obviously, this is a discrete optimization problem. Submodularity, oftentimes viewed as a discrete analog of convexity, is the key to effectively and efficiently solve discrete optimization problems in machine learning. Thus, we design the objective function to satisfy the "submodularity" condition. Before proceeding to the objective function, we first introduce the definitions of submodularity and monotonicity.

Definition 1. (Submodularity [12]) Let E be a finite set. A set function $F : 2^E \to \mathbb{R}$ is submodular if $F(A \cup \{a_1\}) - F(A) \geq F(A \cup \{a_1, a_2\}) - F(A \cup \{a_2\})$, for all $A \subseteq E$ and $a_1, a_2 \in E \setminus A$. This property, also named diminishing marginal gains, states that the impact of adding an element to a larger set is less.

Definition 2. (Monotonically Increasing) A set function F is monotonically increasing if $F(I_1) \leq F(I_2)$ for any $I_1 \subseteq I_2$.

In order to introduce the criteria met by our objective function, we compare a pair of graph clustering results $(C1, C2)$ for each criterion in Figure 5. $C2$ more closely complies with each criterion by enforcing $O(C1) < O(C2)$. We have determined the following four criteria. 1) The compactness guards the basic idea of graph clustering, i.e., intra-cluster density. Maximizing the objective ensures that densely rather than sparsely connected vertices constitute a cluster. 2) The homogeneity requires each cluster to be homogeneous for labelled vertices, i.e., a cluster should not mix vertices belonging to different categories. 3) The label balance states that the number of labelled vertices in each cluster stays "balanced". This constraint avoids to produce clusters without category labels, and thereby supports the homogeneity. 4) The modality diversity ensures that each cluster contains vertices from all modalities. The compactness equips the dictionaries with representation effectiveness. The homogeneity and label balance enforce each dictionary atom, i.e., each cluster center, to encode a latent semantic meaning and be discriminative. The modality diversity is crucial to couple all modalities' dictionaries to be semantically related.

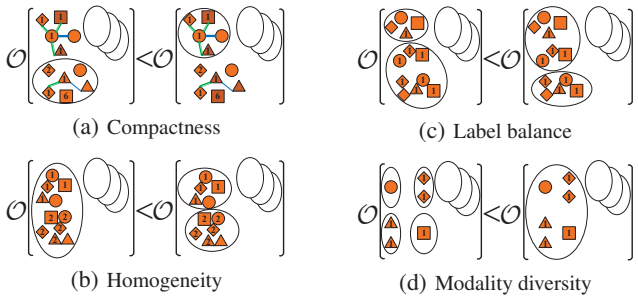


Figure 5: Illustrations of the four criteria met by our objective function. An eclipse represents a cluster.

Compactness: A random walk, starting at a vertex and then randomly travelling to a connected vertex, is more likely to stay within a cluster than travelling between. Therefore conducting random walks on the graph can discover clusters where the flow tends to gather. The transition probability from a vertex v_i to a vertex v_j is defined as a set function $P_{ij}(A) : 2^E \rightarrow R$ for the graph $G(V, A)$:

$$P_{ij}(A) = \begin{cases} 1 - \frac{\sum_{j:e_{ij}\in A} w_{ij}}{w_i} & \text{if} \quad i = j, \\ \frac{w_{ij}}{w_i} & \text{if} \quad i \neq j, e_{ij} \in A, \\ 0 & \text{if} \quad i \neq j, e_{ij} \notin A, \end{cases} \tag{2}$$

which encourages random walks within clusters ($e_{ij} \in A$) and eliminates those between clusters ($e_{ij} \notin A$). $w_i = \sum_{j:e_{ij}\in E} w_{ij}$ is the total weights incident to v_i. We add a self loop transition ($i = j$) to maintain the total transition probability out of v_i to be 1.

To satisfy compactness, we define the objective as the entropy rate of a random walk [3] to measure the uncertainty of the walk:

$$C(A) = -\sum_i \mu_i \sum_j P_{ij}(A) \log P_{ij}(A), \tag{3}$$

where μ_i is the ith element of the stationary distribution $\boldsymbol{\mu} = (\mu_1, \mu_2, \cdots, \mu_{|V|}) = (\frac{w_1}{w_{all}}, \frac{w_2}{w_{all}}, \cdots, \frac{w_{|V|}}{w_{all}})$ with $w_{all} = \sum_{i=1}^{|V|} w_i$. Intuitively, random walks on dense subgraphs are more uncertain than on sparse subgraphs. Hence maximizing the entropy rate ensures the compactness, and enforces that the edges selected into A from E can make each cluster as dense as possible. $C(A)$ has been proved to be monotonically increasing and submodular in [11].

Homogeneity: Suppose that for the graph $G(V, A)$ given by current A, we have N_A connected components, i.e., $G_1, \cdots, G_{N_A}$. In the kth connected component G_k, we denote the number of labelled vertices as $|V_{lk}|$, and the number of labelled vertices carrying the label y as $|V_{l(y)k}|$. G_k's homogeneity can be defined as $H(G_k) = \frac{1}{|V_{lk}|} \max_y |V_{l(y)k}|$, which computes the percentage of those vertices carrying the mostly assigned label in G_k. The objective function of homogeneity for the whole graph $G(V, A)$ *w.r.t.* A is straightforward by averaging over all N_A connected components:

$$\mathcal{H}(A) = \sum_{k=1}^{N_A} \frac{|V_{lk}|}{|V_l|} \times H(G_k) - N_A = \frac{1}{|V_l|} \sum_{k=1}^{N_A} \max_y |V_{l(y)k}| - N_A. \tag{4}$$

Maximizing Equation (4) encourages homogeneity (the first term), but avoids a trivial solution where each cluster contains a single vertex by restricting N_A to be as small as possible (the second term). The monotonicity and submodularity of $\mathcal{H}(A)$ are proved in [8].

Label balance: Motivated by the fact that the information entropy of a random variable achieves the maximum if this random variable is uniformly distributed, we consider the percentage of labelled vertices across clusters as a random variable and propose the objective function for label balance as:

$$\mathcal{L}(A) = -\sum_{k=1}^{N_A} L(G_k) \log L(G_k) - N_A = -\sum_{k=1}^{N_A} \frac{|V_{lk}|}{|V_l|} \log \frac{|V_{lk}|}{|V_l|} - N_A. \tag{5}$$

So that maximizing $\mathcal{L}(A)$ enforces labelled vertices to scatter uniformly across N_A clusters. In [11], the authors prove that $-\sum_k p_A(k) \log p_A(k) - N_A$ satisfies monotonicity and submodularity.

Modality diversity: We again employ the information entropy to formulate the objective for modality diversity, but consider the percentage of vertices in each modality across clusters as a random variable. By averaging over all M modalities, we obtain:

$$\begin{aligned} \mathcal{M}(A) &= \sum_{m=1}^{M} \frac{|V^m|}{|V|} \Big[-\sum_{k=1}^{N_A} M(G_k) \log M(G_k) - N_A \Big] \\ &= -\frac{1}{|V|} \sum_{m=1}^{M} \sum_{k=1}^{N_A} |V_k^m| \log \frac{|V_k^m|}{|V^m|} - N_A. \end{aligned} \tag{6}$$

Maximizing $\mathcal{M}(A)$ encourages each cluster to be diverse, i.e, including vertices from all M modalities. As mentioned above, the monotonicity and submodularity of $-\sum_k p_A(k) \log p_A(k) - N_A$ have been proved in [11]. Meanwhile, we are provided with the fact that linear combination with nonnegative coefficients preserves monotonicity and submodularity [13]. $\mathcal{M}(A)$, therefore, is also guaranteed to be monotonically increasing and submodular.

Combining the four objective functions introduced, the overall optimization problem can be written as:

$$\begin{aligned} \max_A \quad & \mathcal{O}(A) = C(A) + \lambda \mathcal{H}(A) + \gamma \mathcal{L}(A) + \mu \mathcal{M}(A) \\ s.t. \quad & A \subseteq E \quad \text{and} \quad N_A \geq K, \end{aligned} \tag{7}$$

where λ, γ, and μ are three trade-off parameters to balance the importance of the four terms. $\mathcal{O}(A)$, a linear combination of $C(A)$, $\mathcal{H}(A)$, $\mathcal{L}(A)$, and $\mathcal{M}(A)$, is monotonically increasing and submodular. Solving the optimization problem in Equation (7) is NP-hard. Fortunately, the submodularity of $\mathcal{O}(A)$ contributes a greedy approximation algorithm with effectiveness and efficiency guarantee. It initiates $A = \emptyset$ and iteratively selects the edge $e \in E \setminus A$ to maximize the marginal gain $\mathcal{O}(A \cup e) - \mathcal{O}(A)$. Fisher et al. [13] showed that the algorithm gives a 1/2-approximation bound on the optimality of the solution. Besides, the algorithm is highly efficient thanks to the diminishing marginal gains property of submodular functions. In each iteration it computes the marginal gain for only the edge who holds the second largest gain in the previous iteration, instead of all edges in the set $E \setminus A$. The implementation details and time complexity will be discussed in Section 3.5.

3.3.3 *Dictionary Inference*

In the kth cluster, we calculate the center of vertices in the mth modality as the dictionary atom d_k^m. The final dictionary of the mth modality $\mathbf{D}^m$ combines K dictionary atoms inferred from all K clusters, i.e., $\mathbf{D}^m = [d_1^m, \cdots, d_K^m]$. To wrap up, we present the pseudo code of learning semantically related dictionaries in Algorithm 1.

Algorithm 1 Learn Semantically Related Dictionaries (LSRD)

Input: $\mathbf{S}_l$, $\mathbf{S}_u$ – the labelled and unlabelled instances in the source domain; $\mathbf{g}$ – the label vector in the source domain; λ', γ', μ' – trade-off parameters for initialization; K – the dictionary size;
Output: $\mathbf{D} = \{\mathbf{D}^m\}_{m=1}^M$

```
Construct the graph G = (V, E);
Initialize A ← ∅, D^1, ··· , D^M ← ∅, λ = (max_{e∈E} C(e)−C(∅) / max_{e∈E} H(e)−H(∅)) λ', γ = (max_{e∈E} C(e)−C(∅) / max_{e∈E} L(e)−L(∅)) γ', μ = (max_{e∈E} C(e)−C(∅) / max_{e∈E} M(e)−M(∅)) μ';
while N_A > K do
    ê ← arg max_{e∈E\A} O(A ∪ e) − O(A);
    A ← A ∪ ê;
end while
for m = 1, ··· , M do
    for k = 1, ··· , K do
        D^m = D^m ∪ {(1/|V_k^m|) Σ_{j:s_j^m ∈ G_k} s_j^m};
    end for
end for
```

3.4 Transfer Dictionaries and Instances

3.4.1 *Transfer Dictionaries*

We learn M semantically related dictionaries from the source domain to unlock the power of sparse coding in homogenizing different modalities as stated in Section 3.3. More importantly, we transfer the M semantically related dictionaries to the target domain, and apply them to learn enriched representations of target instances, in order to address the data insufficiency problem. Mathematically, for the mth modality of the ith labelled instance in the target domain (if available), i.e., $\mathbf{t}_{li}^m$, we transfer the mth dictionary learnt from the source domain, i.e., $\mathbf{D}^m$, and apply sparse coding to learn the enriched representation $\hat{\mathbf{t}}_{li}^m$ by

$$\min_{\hat{\mathbf{t}}_{li}^m} \quad \|\mathbf{t}_{li}^m - \mathbf{D}^m \hat{\mathbf{t}}_{li}^m\|_F^2 + \alpha \|\hat{\mathbf{t}}_{li}^m\|_1 \qquad s.t. \quad \hat{\mathbf{t}}_{li}^m \geq 0, \tag{8}$$

where α controls the sparsity of enriched representations. We obtain the enriched representation $\hat{\mathbf{t}}_{ui}^m$ for the mth modality of the ith unlabelled instance, i.e., $\mathbf{t}_{ui}^m$, in a similar fashion.

3.4.2 Transfer Instances

After transferring the dictionaries, the label scarcity problem necessitates a much more powerful solution - transferring abundant labelled instances from the source into the target domain. To enable instance transfer, the following two prerequisites have to be met first. 1) Learn enriched representations for labelled source instances. Mathematically, for the mth modality of the jth labelled source instance, i.e., $\mathbf{s}_{lj}^m$, we learn the enriched representation $\hat{\mathbf{s}}_{lj}^m$ by performing sparse coding over the mth dictionary $\mathbf{D}^m$:

$$\min_{\hat{\mathbf{s}}_{lj}^m} \quad \|\mathbf{s}_{lj}^m - \mathbf{D}^m\hat{\mathbf{s}}_{lj}^m\|_F^2 + \alpha\|\hat{\mathbf{s}}_{lj}^m\|_1 \qquad s.t. \quad \hat{\mathbf{s}}_{lj}^m \geq 0. \tag{9}$$

Only in this way can the representation structures of labelled source instances, i.e., $\hat{\mathbf{S}}_l$, stay consistent with those of target instances, i.e., $\hat{\mathbf{T}}_l$ and $\hat{\mathbf{T}}_u$. 2) Aggregate enriched representations of all existing modalities for each target instance. We adopt max pooling, widely applied in image processing, to aggregate. For the ith labelled target instance, max pooling maximizes each feature of the enriched representation over all existing modalities, i.e.,

$$\hat{\mathbf{t}}_{li}^{MP}(k) = \max_{m=1,2,\cdots,M}\{\hat{\mathbf{t}}_{li}^m(k)\}, \quad \text{for all} \quad k = 1, 2, \cdots, K, \tag{10}$$

where $\hat{\mathbf{t}}_{li}^m$ could be missing for some $1 \leq m \leq M$. We obtain the aggregated representation for the ith unlabelled target instance, i.e., $\hat{\mathbf{t}}_{ui}^{MP}$, similarly. In this case, we obtain a uniform representation for all target instances regardless of within-modality insufficiency. Besides, the representation is robust to unreliable modalities, since max pooling chooses the most responsive modality for each feature.

Afterwards, we propose the Multimodal Transfer AdaBoost algorithm to leverage labelled source instances. The algorithm is based on TrAdaBoost [4] in terms of the basic idea, i.e., reduce the distribution differences between domains by adjusting the weights of instances for training in an adaptively boosting fashion. Specifically, the weights of mis-classified target instances increase to make sure that these instances draw enough attention to be classified right in the next iteration, while the mis-classified source instances are down weighted because they are likely the most different in distribution from target instances. However, our algorithm differs from TrAdaBoost [4] in the following two aspects: 1) for each iteration it learns M weak learners to handle M modalities, and skilfully combines M learners' results to boost the prediction accuracy; 2) more importantly, it learns and differentiates weights for different modalities besides instances. Algorithm 2 details the whole algorithm.

Algorithm 2 Multimodal Transfer AdaBoost (MTAB)

Input: $\hat{\mathbf{T}}_l^{MP}$ – enriched representations of labelled target instances; $\hat{\mathbf{T}}_u^{MP}$ – enriched representations of unlabelled target instances; $\hat{\mathbf{S}}_l$ – enriched representations of labelled source instances; $\mathbf{y}$ – the label vector in the target domain; $\mathbf{g}$ – the label vector in the source domain

Output: h_f – the final hypothesis for $\hat{\mathbf{T}}_u^{MP}$

1: Initialize the weight of the ith ($1 \leq i \leq N_l^t$) instance: $v_i(1)$ in the target domain: $v_i(1)$;
2: Initialize the weight of the mth ($1 \leq m \leq M$) modality of the jth ($1 \leq j \leq N_l^s$) instance in the source domain: $w_j^m(1)$;
3: **for** $r = 1, \cdots, R$ **do**
4: **for** $m = 1, \cdots, M$ **do**
5: Set $p_i^m(r) = \begin{cases} v_i(r)/B^m(r), & 1 \leq i \leq N_l^t, \\ w_{i-N_l^t}^m(r)/B^m(r), & N_l^t + 1 \leq i \leq N_l^t + N_l^s, \end{cases}$ where $B^m(r) = \sum_{i=1}^{N_l^t} v_i(r) + \sum_{j=1}^{N_l^s} w_j^m(r)$.
6: Train **WeakLearner** $h^m(r, \cdot)$ on $[\hat{\mathbf{T}}_l^{MP}; \hat{\mathbf{S}}_l^m]$ weighted by $\mathbf{p}^m(r) = \{p_i^m(r)\}_{i=1}^{N_l^t+N_l^s}$;
7: **end for**
8: Define the error on $\hat{\mathbf{T}}_l^{MP}$: $\varepsilon(r) = \sum_{i=1}^{N_l^t} \frac{v_i(r)\max_{m=1}^M \mathbf{I}[h^m(r,\hat{\mathbf{t}}_{li}^{MP}) \neq y_i]}{\sum_{i=1}^{N_l^t} v_i(r)}$, where $\mathbf{I}[a] = 1$ if a is true and $\mathbf{I}[a] = 0$ otherwise;
9: Define the consistency of M weak learners on $\hat{\mathbf{T}}_l^{MP}$: $consistency(r) = 1 - \frac{\sum_{m_1}^M \sum_{m_2}^M \sum_{i=1}^{N_l^t} \mathbf{I}[h^{m_1}(r,\hat{\mathbf{t}}_{li}^{MP}) \neq h^{m_2}(r,\hat{\mathbf{t}}_{li}^{MP})]}{N_l^t \times \binom{M}{2}}$;
10: Set $\epsilon(r) = \varepsilon(r) * consistency(r)$ ($\epsilon(r) < 0.5$ is compulsory);
11: Set $\beta(r) = \frac{\epsilon(r)}{1-\epsilon(r)}$ and $\beta = 1/(1 + \sqrt{2\ln N_l^s/R})$;
12: Update the weights: $v_i(r+1) = v_i(r)\beta(r)^{1-\max_{m=1}^M \mathbf{I}[h^m(r,\hat{\mathbf{t}}_{li}^{MP}) \neq y_i]}$, $1 \leq i \leq N_l^t$; $w_j^m(r+1) = w_j^m(r)\beta^{\mathbf{I}[h^m(r,\hat{\mathbf{s}}_{lj}^m) \neq g_j]}$, $1 \leq j \leq N_l^s$.
13: **end for**
14: $h_f(\hat{\mathbf{t}}_{ui}^{MP}) = \arg\min_c(\prod_{r=\lceil R/2 \rceil}^R \beta(r)^{-\max_{m=1}^M \mathbf{I}[h^m(r,\hat{\mathbf{t}}_{ui}^{MP}) \neq c]})$;

3.5 Complexity Analysis

The computational cost of the FLORAL method comprises two parts. 1) Learn semantically related dictionaries in $O(M|V^m|\log|V^m| + Mk|V^m| + c|V| + |V|\log|V|)$, where c is a constant. The first two terms together are the cost of constructing the mutual k-NN graph within each modality implemented by KD-tree [1], a space partition based approach. The third term is the cost to build inter-edges across different modalities. The last term corresponds to submodular graph clustering implemented by a max heap which stores marginal gains of all edges. Taking the full advantage of the diminishing marginal gains property, submodular clustering is highly computationally efficient by retrieving the top of the heap, re-maximizing the heap, and updating the marginal gain of the top only. 2) Transfer dictionaries and instances in $O(M(K + Z^2)(N_l^s + N_l^t + N_u^t) + RM(N_l^s + N_l^t))$, where Z is the number of non-zeros in the enriched representation. The first term is the cost to solve sparse coding in Equation (8) (9) with SPAMS[1], while Algorithm 2 runs in $O(RM(N_l^s + N_l^t))$ by training each weak learner with LIBLINEAR[2]. In conclusion, FLORAL scales linearly with the number of instances as well as the number of modalities.

4. EXPERIMENTS

In this section, we evaluate the FLORAL method with the case study of air quality prediction. In the case study, FLORAL transfers knowledge from a source city, i.e., Beijing, to improve accuracies of air quality prediction in three target cities, namely Shanghai, Tianjin and Baoding, which face either the label scarcity or the data insufficiency problem.

4.1 Datasets

We collected the following four data modalities in Beijing: 1) *road networks* from Bing Maps contain road segments each of which is described with its end points, length and level of capacity; 2) *Point-Of-Interests (POI)* from Bing Maps indicate the name, address, coordinates, category of a venue; 3) *Meteorological data* crawled from a public website every hour include weather, temperature, humidity, barometer pressure, wind strength, and etc; 4) *Taxi trajectories* generated by over 32,000 taxicabs in Beijing from February 2nd to May 26th, 2014. In the three target cities, however, only the first three modalities are available. Table 2 details the statistics of the first three modalities for all cities.

[1] http://spams-devel.gforge.inria.fr/index.html
[2] https://www.csie.ntu.edu.tw/~cjlin/liblinear/

Table 2: The statistics of three modalities for all cities.

Modalities		Cities			
		Beijing	Shanghai	Tianjin	Baoding
Road	#. Segments	249,080	313,736	97,258	69,383
	Highways	994km	2,016km	1,681km	795km
	Roads	24,643km	40,944km	18,595km	17,884km
POI	#. of POIs	379,022	433,016	152,797	88,698
Meteorology	Time span(2014)	2/1-5/31	8/1-9/10	9/10-11/30	8/1-11/30

As air quality in a city varies with time and location simultaneously, we characterize a grid region in an hour of a day as an instance by partitioning each city into grid regions in the size of 1.5km×1.5km. For each instance, we extract its features in all modalities. The feature construction for each modality follows [34] in which road network features F_r, POI features F_p, meteorological features F_m, and taxi traffic features F_t are extracted. Specifically, F_r and F_p are spatio features, and F_m and F_t are temporal features. Note that some modalities of some instances are not available, and the modality of taxi trajectories is missing for all instances of the three target cities. We label an instance with Air Quality Index (AQI) values which are collected from ground-based air quality monitor stations in the four cities every hour. The AQI values range from one to six, corresponding to six air quality states, i.e., "Good", "Moderate", "Unhealthy for sensitive groups", "Unhealthy", "Very unhealthy", and "Hazardous", respectively.

We measure the distributional difference in each modality between a source and a target domain with KL-divergence. The larger the KL-divergence is, the more different the feature distributions of a source and a target domain in a modality are. Table 3 and Figure 6 present the distributional differences between Beijing and the three target cities in the three shared modalities.

Table 3: KL-divergence in the distributions of road network and POI features.

	Target cities		
Modalities	Shanghai	Tianjin	Baoding
Road	0.541	0.7361	1.1439
POI	0.7618	0.889	1.1387

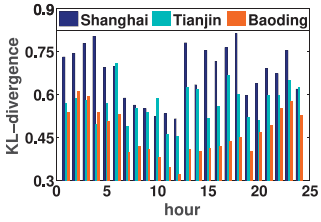


Figure 6: KL-divergence in the distributions of meteorological features, differentiated by hours.

4.2 Baselines

We compare our proposed method **FLORAL** with the following six baselines, evaluated by prediction accuracy:

Original. This method trains a classifier for each modality in a target domain. Among all classifiers, this method selects the one with the best prediction accuracy.

U-Air. This model [34] combines different modalities by co-training spatio and temporal features.

LSRD. We learn semantically related dictionaries from a source domain by applying Algorithm 1, transfer the dictionaries to enrich feature representations in a target domain according to Equation (8)(10), and train classifiers on $\hat{\mathbf{T}}_l^{MP}$.

Orig+TAB. This method performs TrAdaBoost [4], a state-of-the-art algorithm that transfers instances, on each modality with original features, and outputs the best result among all modalities.

LSRD+TAB. We perform TrAdaBoost on each modality with enriched features, and output the best result among all modalities. Original features of both source and target domains are enriched by the semantically related dictionaries according to Equation (8)(9).

MDT. Multi-view Discriminant Transfer learning (MDT) [25] transfers knowledge between domains with multiple views. We adapt MDT to solve our problem which faces the within-modality insufficiency, by discarding those instances with modalities missing.

In summary, Original and U-Air do not transfer. LSRD and Orig+TAB perform feature and instance transfer, respectively. LSRD+TAB directly combines feature and instance transfer. To make Orig+TAB and MDT applicable to our problem, we discard the modalities which are existing in a source but missing in a target domain. We use linear SVM as the base classifier. Given different feature representations for different models, the trade-off parameter C of linear SVM is set according to 10-fold cross validation.

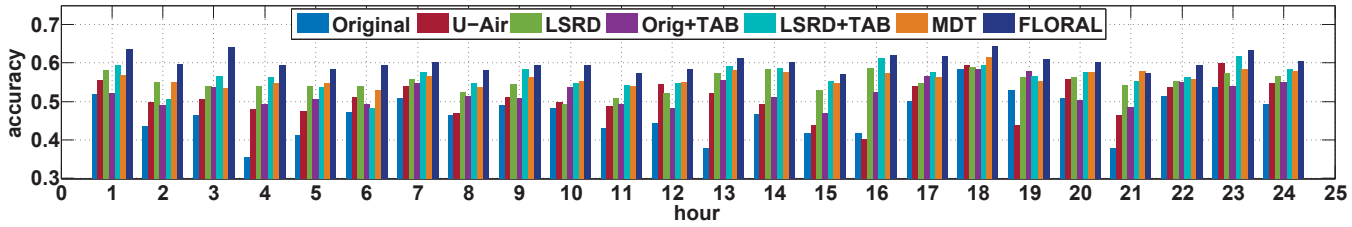
4.3 Results

Performance comparison: We differentiate the performance comparison by hours for the following two reasons: 1) distributions of temporal features for different hours, e.g., traffic features in 0am and 8am, could be distinct; 2) different numbers of instances are available in different hours. For each hour in a target domain, we first select an hour from a source so that transferring labelled instances in the hour maximizes the performance. Second, we randomly select 10% of labelled instances as training data, and the rest as test. In Figure 7, we report the average accuracy over ten such random partitions for each hour.

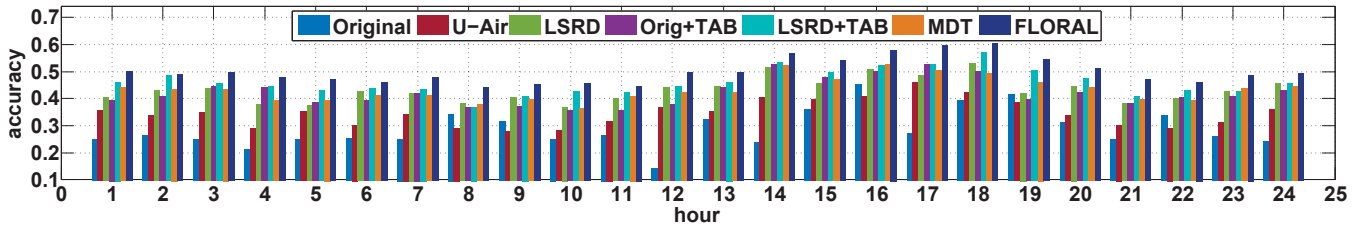
From Figure 7, we have the following observations. First, combining different modalities outperforms using single modality only. Compared to Original, U-Air unlocks the power of spatio and temporal features collectively in a co-training fashion, and thereby partially addresses the label scarcity problem. Especially, our proposed LSRD algorithm is highly effective, since it addresses the data insufficiency problem in a target domain by enriching feature representations. Second, transferring source labelled instances is also critical to improve the performance. Even though we apply TradaBoost on each modality's original features individually, i.e., Orig+TAB, we see the performance improvement. Third, performances of the multi-view transfer learning algorithm MDT are not that satisfactory, probably because it fails to tackle the structured modality missing and within-modality insufficiency. Fourth, directly combining feature and instance transfer i.e., LSRD+TAB, still falls behind our method FLORAL. LSRD+TAB cannot learn and differentiate different modalities' weights as FLORAL does. Generally speaking, FLORAL outperforms all the baselines in almost all hours of all target cities up to 50%.

The improvement of FLORAL over other baselines achieves the most significant when transferring from Beijing to Tianjin according to Figure 7(b); transferring to Baoding takes second while transferring to Shanghai ranks third. Table 3, Figure 2(b), and Figure 6 provide the explanations. The KL-divergence values between Tianjin and Beijing are averagely small for all the three modalities, i.e., road, POI, and meteorology. The distribution of labels, i.e., air quality, in Tianjin is also similar to that in Beijing. However, the feature distributions of Baoding in road and POI largely differ from those of Beijing, considering that Baoding is a small city. In this case, the meteorology which is similar for the two geographically close cities primarily accounts for the transfer. Figure 9 further confirms the fact: the smaller the KL-divergence between Baoding and Beijing in meteorology, the better FLORAL performs.

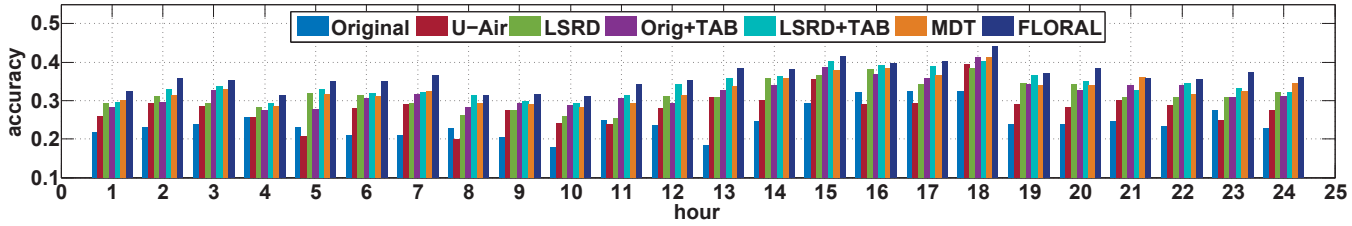
Figure 10 shows the correspondence between each hour in Baoding and the hour in Beijing selected by FLORAL. To maximize the prediction accuracy, it is expected that the hour selected from a source is the most similar to each hour in a target domain in dis-



(a) Hourly air quality prediction accuracies in Shanghai.



(b) Hourly air quality prediction accuracies in Tianjin.



(c) Hourly air quality prediction accuracies in Baoding.

Figure 7: Performance comparison of hourly air quality prediction in different target cities.

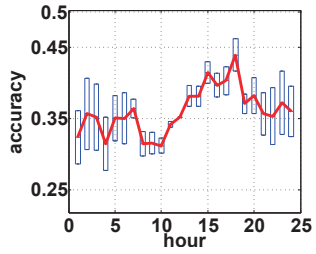


Figure 9: Hourly air quality prediction accuracies in Baoding, with the boxes denoting the scaled KL-divergence values between Baoding and Beijing in meteorology.

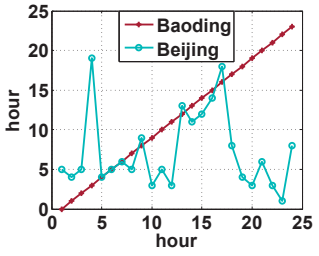


Figure 10: The correspondence between each hour in Baoding and the hour in Beijing selected by FLORAL.

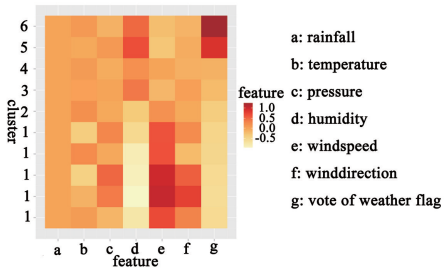
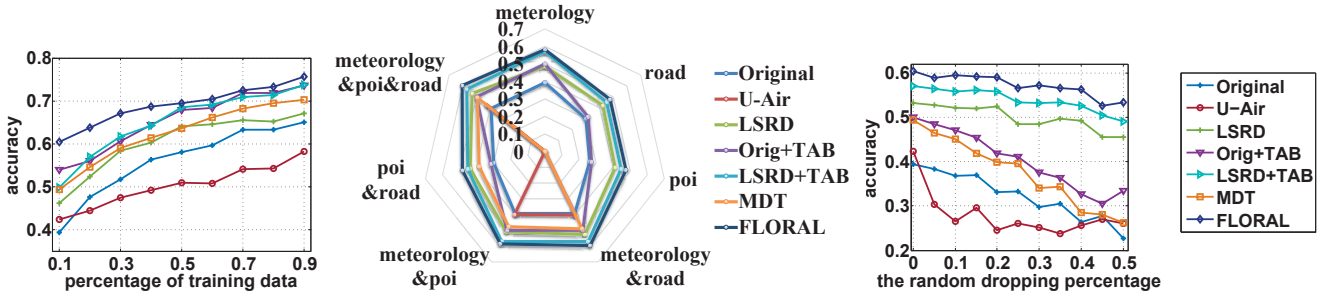


Figure 11: The meteorology dictionary learnt from Beijing during 11am-12pm. The x-axis denotes the features while the y-axis labels a dictionary atom with the mostly assigned label in the cluster.

tributions. Consequently, we conclude that during 5am-9am and 13pm-17pm Beijing is the most synchronously similar to Baoding.

Effectiveness of semantically related dictionaries: The success of FLORAL highly depends on the quality of semantically related dictionaries learnt by LSRD. In Figure 11, we examine and visualize the dictionary learnt from Beijing during 11am-12pm for the modality of meteorology with the size $K = 10$. Each dictionary atom is labelled as the mostly assigned label in the cluster which we infer the atom from. The label of an atom is regarded as the latent semantic meaning it encodes. The figure tells that the semantic meanings do make a lot of sense, and thereby the learnt dictionary is effective. For example, as the level of humidity increases and the wind speed reduces, the labels of dictionary atoms tend to increase, meaning that the air quality gets worse. It is noted that the rainfall stays unchanged across all atoms, because there is a lack of rain in Beijing and exists seldom raining days in our training data.

Dealing with the label scarcity and data insufficiency problems: In Figure 8, we verify that FLORAL is capable of dealing with the label scarcity and data insufficiency problems. We focus on the performance of air quality prediction in Tianjin during 17pm-18pm. First, we vary the percentage of labelled instances for training in the target domain, i.e., Tianjin. The smaller the percentage is, the scarcer the labelled data are. Figure 8(a) shows that when the percentage of training data increases, all algorithms perform better. Especially, when the labelled data are very scarce, say the percentage equals to 0.1, FLORAL even improves the most over the baselines. Thus we conclude that FLORAL can successfully handle the label scarcity problem, and that is why we select 10% of labelled instances as training data for performance comparison. Second, we compare different algorithms' capabilities to tackle the structured modality missing in Figure 8(b). We vary available modalities in Tianjin, ranging from single modality to three modalities together. Figure 8(b) shows that the perfor-



(a) Varying the percentage of labelled target instances.

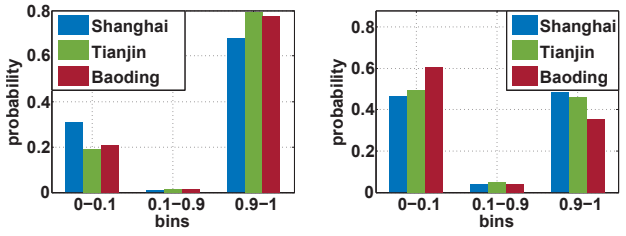
(b) Varying available modalities in the target domain.

(c) Varying the percentage of random dropping in each modality.

Figure 8: Dealing with the label scarcity and data insufficiency problems.

mance gap between FLORAL and the baselines based on LSRD, i.e., LSRD and LSRD+TAB, stays consistent, while the gap between FLORAL and the other baselines increases as more modalities are missing. Therefore we prove that learning semantically related dictionaries fully takes the advantage of the modalities which are missing in a target domain but existing in a source, and thereby effectively addresses the structured modality missing. Note that U-Air cannot handle the cases where only spatio or temporal features are available, and MDT is not applicable in the cases where only single modality is provided. Third, we investigate the capabilities of all algorithms to deal with the within-modality insufficiency in Figure 8(c), by randomly dropping a percentage of data for each modality. Reasonably, as the dropping percentage increases, the performances of all algorithms decrease. However, the performances of FLORAL and the baselines based on LSRD decrease much slower than those of the other baselines. The semantically related dictionaries complement the within-modality insufficiency by enriching feature representations.

Learning and differentiating different modalities' weights: The major reason why FLORAL wins over LSRD+TAB is that FLORAL has the ability to learn and differentiate different modalities' weights when transferring, which is further validated in Figure 12. No matter which target city FLORAL transfers to, the distribution of labelled source instances' weights in meteorology significantly differs from that in traffic. Besides, for each modality,



(a) in meteorology

(b) in traffic

Figure 12: Comparison of the distributions of labelled source instances' weights in two modalities when transferring to different target cities to predict air quality during 17pm-18pm.

the distributions of labelled source instances' weights differ for different target cities. Specifically, the modality of meteorology plays the most important role when transferring from Beijing to Tianjing because the weights are the most likely to lie in 0.9 – 1. The geographical closeness of the two cities explains this. However, the modality of traffic is weighted the highest while transferring from Beijing to Shanghai, the two of which are top two cities in China. We would also clarify why the modality of meteorology seems more important than the modality of traffic for all target cities. It is because the modality of traffic is missing in all target cities so that the meteorology is more likely to dominate.

Varying the percentage of labelled source instances: The performances of FLORAL also rely on the amount of labelled instances we transfer from a source domain. Figure 13 presents the performances of FLORAL in predicting air quality in Tianjin during 17pm-18pm, while we vary the percentage of labelled instances in the target domain, i.e., r_t, and that in the source, i.e., r_s, simultaneously. Reasonably, larger r_t and r_s lead to better performances. Besides, when $r_s = 0.6$, the performances of FLORAL start to saturate, meaning that 60% of the labelled source instances have been sufficient to improve the target domain.

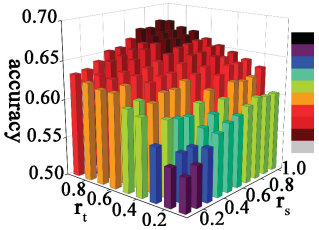
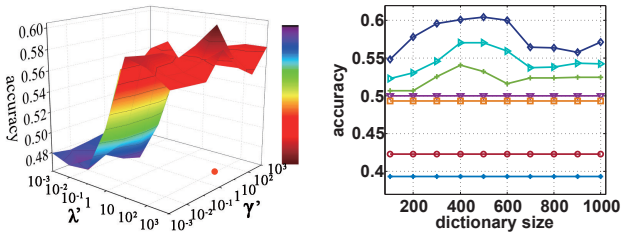


Figure 13: Varying the percentage of labelled instances in the target and source domain simultaneously.

Parameter sensitivity: We also study the influence of different parameter settings on the performances of FLORAL when transferring from Beijing to Tianjin during 17pm-18pm. We investigate three parameters: K, the size of semantically related dictionaries, λ' and γ', the trade-off parameters' initialization in Equation (7). For space limitation, we do not include the result for μ', the other trade-off parameter's initialization. We perform grid search on λ' and γ' in the range of $\{10^{-3}, 10^{-2}, 10^{-1}, 10^{0}, 10^{1}, 10^{2}, 10^{3}\}$ by fixing the dictionary size K. FLORAL gains the best accuracy at $\lambda' = 100$ and $\gamma' = 10$ as Figure 14(a) shows. In Figure 14(b), by fixing $\lambda' = 100$ and $\gamma' = 10$, we obtain the best dictionary size $K = 500$.



(a) Grid search of λ' and γ'.

(b) Varying the dictionary size.

Figure 14: Study of parameter sensitivity on air quality prediction.

Scalability: We evaluate the scalability of our LSRD algorithm, which is the major computational bottleneck of FLORAL. By using KD-tree for graph construction and submodular optimization for graph clustering, LSRD is highly efficient and capable of handling extremely large graphs involving massive vertices and hyper-edges as Figure 15 shows.

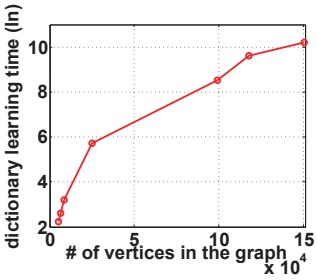


Figure 15: Scalability of FLORAL.

5. CONCLUSIONS

In this paper, we propose a novel method called FLORAL to transfer knowledge between domains with multimodal data. Particularly, FLORAL enriches feature representations in a target domain with semantically related dictionaries learnt from a source, and transfers labelled instances from the source. Extensive experimental results in the case study of air quality prediction demonstrate the superiority of FLORAL over other state-of-the-art methods. Besides air quality prediction, FLORAL could be applied whenever the target domain encounters the label scarcity and data insufficiency problems. In the future, we would like to extend FLORAL to transfer from multiple source domains. Although finding a source domain which contains all modalities in a target is not that difficult, FLORAL can be more flexible by transferring from multiple source domains.

6. ACKNOWLEDGMENTS

We thank the reviewers for their valuable comments to improve this paper. We also thank the support of China National 973 project 2014CB340304, and Hong Kong CERG projects 16211214 and 16209715.

7. REFERENCES

[1] J. L. Bentley. Multidimensional binary search trees used for associative searching. *Communications of the ACM*, 18(9):509–517, 1975.

[2] J. Blitzer, S. Kakade, and D. P. Foster. Domain adaptation with coupled subspaces. In *AISTATS*, pages 173–181, 2011.

[3] T. M. Cover and J. A. Thomas. *Elements of information theory*. 2012.

[4] W. Dai, Q. Yang, G.-R. Xue, and Y. Yu. Boosting for transfer learning. In *ICML*, pages 193–200, 2007.

[5] Z. Fang and Z. M. Zhang. Discriminative feature selection for multi-view cross-domain learning. In *CIKM*, pages 1321–1330, 2013.

[6] D. R. Hardoon, S. Szedmak, and J. Shawe-Taylor. Canonical correlation analysis: An overview with application to learning methods. *Neural computation*, 16(12):2639–2664, 2004.

[7] J. He and R. Lawrence. A graph-based framework for multi-task multi-view learning. In *ICML*, pages 25–32, 2011.

[8] Z. Jiang, G. Zhang, and L. S. Davis. Submodular dictionary learning for sparse coding. In *CVPR*, pages 3418–3425, 2012.

[9] X. Jin, F. Zhuang, H. Xiong, C. Du, P. Luo, and Q. He. Multi-task multi-view learning for heterogeneous tasks. In *CIKM*, pages 441–450, 2014.

[10] R. Kiros, R. Salakhutdinov, and R. Zemel. Multimodal neural language models. In *ICML*, pages 595–603, 2014.

[11] M.-Y. Liu, O. Tuzel, S. Ramalingam, and R. Chellappa. Entropy-rate clustering: Cluster analysis via maximizing a submodular function subject to a matroid constraint. *PAMI*, 36(1):99–112, 2014.

[12] S. T. McCormick. Submodular function minimization. *Handbooks in operations research and management science*, 12:321–391, 2005.

[13] G. L. Nemhauser, L. A. Wolsey, and M. L. Fisher. An analysis of approximations for maximizing submodular set functions. *Mathematical Programming*, 14(1):265–294, 1978.

[14] L. Nie, L. Zhang, Y. Yang, M. Wang, R. Hong, and T.-S. Chua. Beyond doctors: future health prediction from multimedia and multimodal observations. In *MM*, pages 591–600, 2015.

[15] K. Ozaki, M. Shimbo, M. Komachi, and Y. Matsumoto. Using the mutual k-nearest neighbor graphs for semi-supervised classification of natural language data. In *CoNLL*, pages 154–162, 2011.

[16] S. J. Pan and Q. Yang. A survey on transfer learning. *Knowledge and Data Engineering, IEEE Transactions on*, 22(10):1345–1359, 2010.

[17] X. Shi, Q. Liu, W. Fan, and P. S. Yu. Transfer across completely different feature spaces via spectral embedding. *Knowledge and Data Engineering, IEEE Transactions on*, 25(4):906–918, 2013.

[18] X. Song, L. Nie, L. Zhang, M. Liu, and T.-S. Chua. Interest inference via structure-constrained multi-source multi-task learning. In *IJCAI*, pages 2371–2377, 2015.

[19] N. Srivastava and R. R. Salakhutdinov. Multimodal learning with deep boltzmann machines. In *NIPS*, pages 2222–2230, 2012.

[20] B. Tan, E. Zhong, E. W. Xiang, and Q. Yang. Multi-transfer: Transfer learning with multiple views and multiple sources. In *SDM*, 2013.

[21] Y. Wei, Y. Song, Y. Zhen, B. Liu, and Q. Yang. Scalable heterogeneous translated hashing. In *SIGKDD*, pages 791–800, 2014.

[22] Y. Wei, Y. Zhu, C. W.-k. Leung, Y. Song, and Q. Yang. Instilling social to physical: Co-regularized heterogeneous transfer learning. In *AAAI*, 2016.

[23] D. Yang, D. Zhang, Z. Yu, and Z. Yu. Fine-grained preference-aware location search leveraging crowdsourced digital footprints from LBSNs. In *UbiComp*, pages 479–488, 2013.

[24] H. Yang and J. He. Learning with dual heterogeneity: a nonparametric bayes model. In *SIGKDD*, pages 582–590, 2014.

[25] P. Yang and W. Gao. Multi-view discriminant transfer learning. In *IJCAI*, pages 1848–1854, 2013.

[26] P. Yang, W. Gao, Q. Tan, and K.-F. Wong. Information-theoretic multi-view domain adaptation. In *ACL*, pages 270–274, 2012.

[27] P. Yang and J. He. Model Multiple Heterogeneity via Hierarchical Multi-Latent Space Learning. In *SIGKDD*, pages 1375–1384, 2015.

[28] Q. Yang, Y. Chen, G.-R. Xue, W. Dai, and Y. Yu. Heterogeneous transfer learning for image clustering via the social web. In *ACL*, pages 1–9, 2009.

[29] Z. Yu, F. Wu, Y. Yang, Q. Tian, J. Luo, and Y. Zhuang. Discriminative coupled dictionary hashing for fast cross-media retrieval. In *SIGIR*, pages 395–404, 2014.

[30] D. Zhang, J. He, Y. Liu, L. Si, and R. Lawrence. Multi-view transfer learning with a large margin approach. In *SIGKDD*, pages 1208–1216, 2011.

[31] J. Zhang and J. Huan. Inductive multi-task learning with multiple view data. In *SIGKDD*, pages 543–551, 2012.

[32] Z. Y. Zhao, M. Xie, and M. West. Dynamic dependence networks: Financial time series forecasting and portfolio decisions. *Applied Stochastic Models in Business and Industry*, 2016.

[33] Y. Zheng, L. Capra, O. Wolfson, and H. Yang. Urban computing: concepts, methodologies, and applications. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 5(3):38, 2014.

[34] Y. Zheng, F. Liu, and H.-P. Hsieh. U-air: When urban air quality inference meets big data. In *SIGKDD*, pages 1436–1444, 2013.

[35] Y. Zheng, T. Liu, Y. Wang, Y. Zhu, Y. Liu, and E. Chang. Diagnosing new york city's noises with ubiquitous data. In *UbiComp*, pages 715–725, 2014.

[36] Y. Zheng, X. Yi, M. Li, R. Li, Z. Shan, E. Chang, and T. Li. Forecasting Fine-Grained Air Quality Based on Big Data. In *SIGKDD*, pages 2267–2276, 2015.