Pattern Recognition 145 (2024) 109919

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

Mutual Domain Adaptation

Sunghong Park [a], Myung Jun Kim [b], Kanghee Park [c], Hyunjung Shin [d,e,*]

[a] Department of Psychiatry, Ajou University School of Medicine, Suwon 16499, South Korea
[b] SODA Team, Inria Saclay, Palaiseau 91120, France
[c] Technology Intelligence Research Team, Korea Institute of Science and Technology Information, Seoul 02456, South Korea
[d] Department of Industrial Engineering, Ajou University, Suwon 16499, South Korea
[e] Department of Artificial Intelligence, Ajou University, Suwon 16499, South Korea

ARTICLE INFO

Keywords:
Domain adaptation
Semi-supervised learning
Label propagation
Pseudo-labeling

ABSTRACT

To solve the label sparsity problem, domain adaptation has been well-established, suggesting various methods such as finding a common feature space of different domains using projection matrices or neural networks. Despite recent advances, domain adaptation is still limited and is not yet practical. The most pronouncing problem is that the existing approaches assume source-target relationship between domains, which implies one domain supplies label information to another domain. However, the amount of label is only marginal in real-world domains, so it is unrealistic to find source domains having sufficient labels. Motivated by this, we propose a method that allows domains to mutually share label information. The proposed method finds a projection matrix that matches the respective distributions of different domains, preserves their respective geometries, and aligns their respective class boundaries. The experiments on benchmark datasets show that the proposed method outperforms relevant baselines. In particular, the results on varying proportions of labels present that the fewer labels the better improvement.

1. Introduction

Domain adaptation is a representative strategy solving the label sparsity problem [1–4]. For the target domain sparsely labeled, a related domain containing many labels is used as the source. This allows the target domain to improve the performance of classification or regression from the information provided in the label [5–7]. In a word, domain adaptation is the process by which less-labeled domains resemble more labeled domains in order to borrow label information.

However, the two domains cannot transmit information directly because their data properties are different. A transformation is required to view the two domains as one [8,9]. The objective of domain adaptation is generally set to search a common feature space by a neural network or projection matrix. There have been interesting studies for respective approaches. In [10,11], the adversarial learning with neural networks is employed to perform domain adaptation. The method consists of the feature extractor and the classifier: the former trains domain labels (source or target), and the latter trains class labels. The feature extractor includes reverse gradient layers which make different domain samples indistinguishable [12]. The methods yield good performance. In contrast, the extracted feature space hardly represents the original properties of each domain such as distributional or geometrical traits. As a result, it creates a feature space in which different domain samples are mixed and arranged only for classification performance, but intrinsic domain specific properties are lost. This is due to the nonlinear embedding of neural networks and the resulting inexplicability.

On the other hand, approaches using projection matrices preserve the properties of original domain. For this, the objective function pursues to keep distributional or geometrical traits. In transfer component analysis (TCA) [13]—the well-known method, the objective function seeks for a set of projectors to minimize the maximum mean discrepancy (MMD) [14–16] between the source and target sample distributions. On the other hand, semi-supervised domain adaptation (SSDA) [17] performs manifold regularization based on TCA by integrating unlabeled data with similarity constraints into the objective function. In [18], manifold embedded distribution alignment (MEDA) was suggested which learns a domain-invariant classifier by minimizing the structural risk in manifold and performing the dynamic distribution alignment simultaneously. As such, when domain adaptation is performed with the projection matrix, it is possible to derive common representation in a feature space while preserving distributional and geometrical properties.

* Corresponding author at: Department of Industrial Engineering, Ajou University, Worldcup-ro 206, Yeongtong-gu, Suwon 16499, South Korea.
E-mail address: shin@ajou.ac.kr (H. Shin).

https://doi.org/10.1016/j.patcog.2023.109919
Received 3 February 2022; Received in revised form 7 June 2023; Accepted 23 August 2023
Available online 27 August 2023

In recent, the studies on domain adaptation with the projection matrix have proposed various methods to match not only feature properties of data but also class information of samples. Li et al. proposed domain invariant and class discriminative feature learning for visual domain adaptation (DICD), which learns a latent feature space while reducing the domain difference by jointly matching the marginal and class-conditional distributions [19]. Similarly, Zhang et al. proposed joint geometrical and statistical alignment for visual domain adaptation (JGSA) in which the marginal and conditional distribution divergences between domains and the projections for each domain are constrained to reduce the domain shift statistically and geometrically, respectively [20]. Likewise, Zhou et al. proposed label-guided heterogeneous domain adaptation (LHDA) that matches the marginal and conditional distributions of different data with the adaptation of the combination of labeled data to the unlabeled data [21]. These methods have in common to iteratively perform pseudo-labeling of samples in the target domain using the labels in the source domain to obtain converged prediction results. By generating artificial labels, those methods obtain more sophisticated results.

Despite these successes, domain adaptation in real-world scenarios is still difficult for a couple of reasons. *First*, labels are lacking in most domains. This is more evident in recent data, where both quantity and size are rapidly increasing, but label annotation is time consuming and expensive. In this situation, it is unrealistic to assume the existence of a source domain full of labeled data as in previous studies. It would be desirable for the adaptation method not to assume a source-target relationship between domains. *Second*, the amount of label is only marginal in both domains. Fewer labels make sharing information less efficient across domains. It would be nice to supplement additional labels with labels predicted with high confidence, i.e., pseudo-labels, by a classifier.

Motivated by the limitations, we propose a domain adaptation method called mutual domain adaptation (MDA) based on using a projection matrix. As the name suggests, this method does not assume a source-target relationship between domains. The projection matrix is regularized to preserve the distributional and geometric properties of the original domain. Also, the proposed method includes a method for pseudo-labeling to compensate for deficiency in the label. Labeled and pseudo-labeled samples in one domain are only paired with those labeled (and pseudo-labeled) in the other domain. This strategy provides the projection matrix to align the class boundaries of the two domains, preventing them from intersecting or awkwardly located. In summary, MDA allows two different domains to share label information with each other by matching respective their distributions, preserving their respective geometries, and aligning their respective class boundaries.

In the next section, we sketch the main ideas of the proposed method, MDA: *distribution matching, manifold preserving,* and *consistency mapping*. And accordingly, each property is mathematically implemented as a term of a single objective function. The experimental section shows the performance of MDA on benchmark datasets with varying proportions of labeled data. Experiments on the ablation of three properties are also presented. Section 4 concludes the paper with remarks on contributions and limitations of MDA.

2. Proposed method

2.1. Synopsis

To share label information between domains, it is a priority to find a common feature space because domains have different dimensions as well as feature properties. Let $\mathscr{Z}$ be the common feature space of domain $\mathscr{A}$ and domain $\mathscr{B}$. Here are some desirable premises while projecting two domains to $\mathscr{Z}$. *First*, the projection adjusts the underlying distribution in domain $\mathscr{A}$ so that it maps analogously to the distribution in domain $\mathscr{B}$, and vice versa. *Second*, the projection should be carefully conducted not to disrupt the inherent manifold structure of each domain. That is, the neighbors in the original domain must be neighbors in $\mathscr{Z}$, maintaining topological order. However, the premises above may distort the shape of distribution or change the orientation of class boundary. One of the worst-case scenarios is that the class boundaries of two domains are crossed or awkwardly deployed in $\mathscr{Z}$. To alleviate this, *third*, we restrict the labeled data to be paired only with data labeled in the opposite domain. Pairing labeled data by class across domains can have the effect of aligning class boundaries in projected space. However, having only a few labels blurs the class boundary. To make clearer class boundary, we can use pseudo-labeling prior to pairing, that is, the labels predicted by a particular classifier with high confidence. Then, the information between domains is then passed to other domains through class-by-class pairing of labeled samples with labeled samples. The above premise is named in the order of *distribution matching, manifold preservation*, and *consistency mapping*.

Fig. 1 depicts the process how the MDA implements the premises. For convenience of manifold representation and label propagation between domains, domain data sets are represented as graphs. Fig. 1(a) and (b) show the shape, distribution, and manifold of domain $\mathscr{A}$ and domain $\mathscr{B}$, respectively. Outlined circles or triangles indicate labeled samples and blue or red indicates classes. If no premise is made, data from both domains will be messed up in the common space $\mathscr{Z}$. Fig. 1(c) shows this case which depicts an ill-posed feature space. However, if the first premise, *distribution matching,* is applied, the difference between the two domains is reduced by matching the centers of distributions. The effect is shown in Fig. 1(d). This approach is known as MMD, which minimizes gaps between means of different distributions [15]. However, since it does not consider the topology between samples, there is no guarantee that the relationships between samples in the original domain $\mathscr{A}$ or $\mathscr{B}$ will still holds in $\mathscr{Z}$. To complement the limitation, the MDA employs the graph Laplacian and implements the premise of *manifold preservation.* And Fig. 1(e) depicts that each manifold in domain $\mathscr{A}$ and $\mathscr{B}$ still holds even in the reshaped distribution in $\mathscr{Z}$. The grey solid edges indicate the relationship between samples. Fig. 1(f) shows *consistency mapping* where the red and blue thick solid lines represent pairs between domains. Consistency mapping is additionally described in Fig. 2. The figure illustrates how class-wise pairing between labeled samples reasonably align class boundaries.

2.2. Formulation and optimization

Here we provide the mathematical formulation to find the common feature space $\mathscr{Z}$ satisfying the three premises, by finding projection matrix $\boldsymbol{P}$ from domain $\mathscr{A}$ and domain $\mathscr{B}$.

Distribution matching: To minimizing the gap between two domain distributions in the projected feature space, the MMD is adopted [9]. It explicitly reduces their marginal distribution difference by the term below:

$$\| \frac{1}{n_{\mathscr{A}}} \sum_{i=1}^{n_{\mathscr{A}}} \boldsymbol{P}_{\mathscr{A}}^{\mathrm{T}} \boldsymbol{X}_{\mathscr{A}_{(i)}} - \frac{1}{n_{\mathscr{B}}} \sum_{j=1}^{n_{\mathscr{B}}} \boldsymbol{P}_{\mathscr{B}}^{\mathrm{T}} \boldsymbol{X}_{\mathscr{B}_{(j)}} \|_2^2 \tag{1}$$

where $n_{\mathscr{A}}$, $d_{\mathscr{A}}$, $\boldsymbol{X}_{\mathscr{A}} \in \mathbb{R}^{d_{\mathscr{A}} \times n_{\mathscr{A}}}$, and $\boldsymbol{P}_{\mathscr{A}} : \mathbb{R}^{d_{\mathscr{A}}} \to \mathbb{R}^k$ denote the number of data, dimension, data matrix, and the projection matrix of domain $\mathscr{A}$. They are similarly denoted in domain $\mathscr{B}$. Instead of finding each projection, we can simplify the representation by concatenating data matrices as $\boldsymbol{X} = \begin{bmatrix} \boldsymbol{X}_{\mathscr{A}} & 0 \\ 0 & \boldsymbol{X}_{\mathscr{B}} \end{bmatrix} \in \mathbb{R}^{d \times n}$ where $d = d_{\mathscr{A}} + d_{\mathscr{B}}$ and $n = n_{\mathscr{A}} + n_{\mathscr{B}}$, and also the projection matrices as $\boldsymbol{P} = \begin{bmatrix} \boldsymbol{P}_{\mathscr{A}} & 0 \\ 0 & \boldsymbol{P}_{\mathscr{B}} \end{bmatrix} \in \mathbb{R}^{d \times k}$ where k is the dimension of feature space $\mathscr{Z}$. (1) can be rephrased as a form of trace

$$\mathrm{tr}\left(\boldsymbol{P}^{\mathrm{T}} \boldsymbol{X} \mathscr{D} \boldsymbol{X}^{\mathrm{T}} \boldsymbol{P}\right) \tag{2}$$

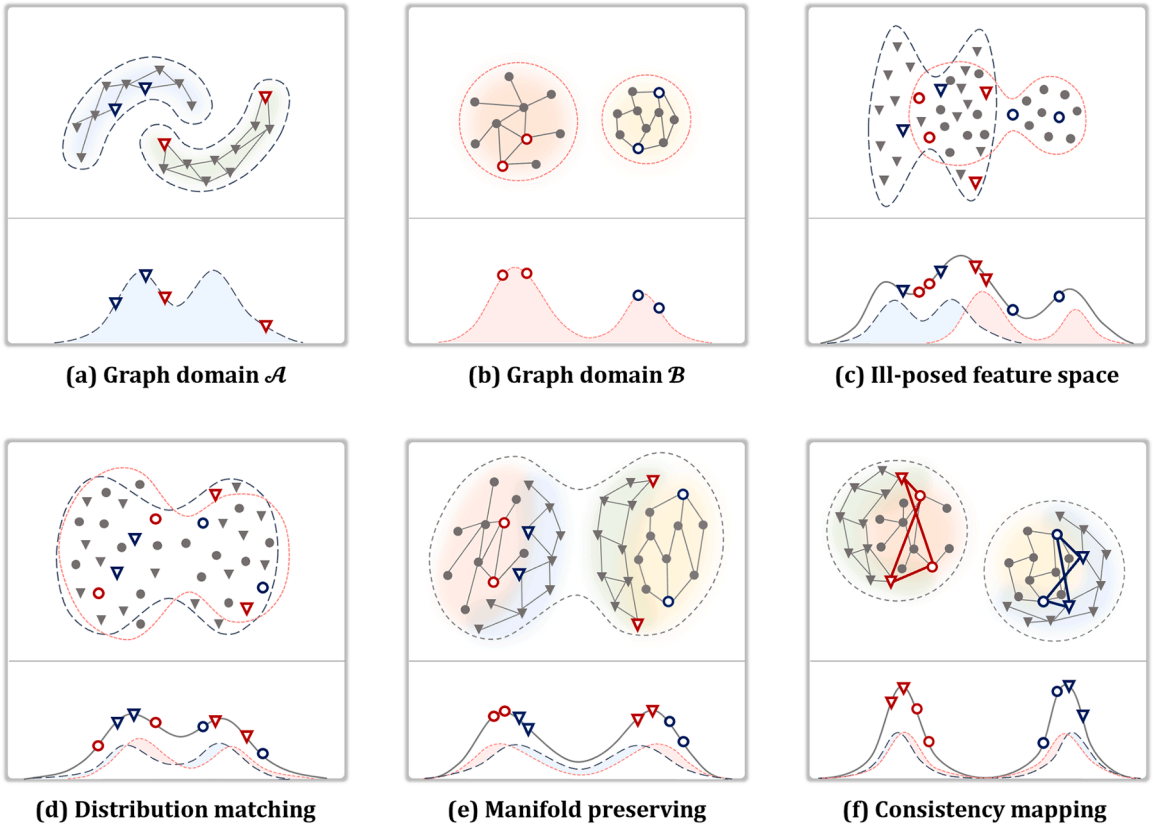

Fig. 1. Mutual domain adaptation. MDA allows two different domains to share label information with each other by matching respective their distributions (distribution matching), preserving their respective geometries (manifold preserving), and aligning their respective class boundaries (consistency mapping).

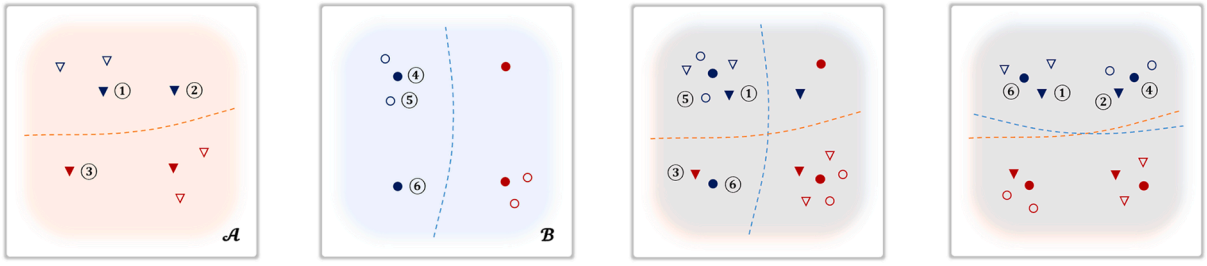

Fig. 2. The effect of consistency mapping. (a) and (b) represent the distributions in the feature space mapped from domains $\mathscr{A}$ and $\mathscr{B}$, respectively. The dotted lines indicate the class boundaries, the filled circles or triangles stand for labeled samples, otherwise unlabeled samples. (c) A common feature space that overlaps the two domains as they are, shows that two distributions are well matched and that each manifold is also well preserved. However, the class boundaries are misaligned, which can lead to confusion in determining class boundaries after domain adaptation. Meanwhile, (d) shows the mapping result according to 'class consistency': labeled samples are paired by class, for instance, ①-⑥ and ②-④. This leads to better alignment of class boundaries of domains $\mathscr{A}$ and $\mathscr{B}$, and thus better domain adaptation can be expected.

by defining $\mathscr{D} = \begin{bmatrix} \frac{\mathbf{1}_{n_{\mathscr{A}} \times n_{\mathscr{A}}}}{{n_{\mathscr{A}}}^2} & -\frac{\mathbf{1}_{n_{\mathscr{A}} \times n_{\mathscr{B}}}}{n_{\mathscr{A}} n_{\mathscr{B}}} \\ -\frac{\mathbf{1}_{n_{\mathscr{B}} \times n_{\mathscr{A}}}}{n_{\mathscr{B}} n_{\mathscr{A}}} & \frac{\mathbf{1}_{n_{\mathscr{B}} \times n_{\mathscr{B}}}}{{n_{\mathscr{B}}}^2} \end{bmatrix} \in \mathbb{R}^{n \times n}$.

Manifold preserving: To preserve manifolds, the smoothness of each domain is regularized and summed as

$$\sum_{i,j} \boldsymbol{W}_{\mathscr{A}_{(i,j)}} \left(\boldsymbol{P}_{\mathscr{A}}^{\mathrm{T}} \boldsymbol{X}_{\mathscr{A}_{(i)}} - \boldsymbol{P}_{\mathscr{A}}^{\mathrm{T}} \boldsymbol{X}_{\mathscr{A}_{(j)}} \right)^2 + \sum_{i,j} \boldsymbol{W}_{\mathscr{B}_{(i,j)}} \left(\boldsymbol{P}_{\mathscr{B}}^{\mathrm{T}} \boldsymbol{X}_{\mathscr{B}_{(i)}} - \boldsymbol{P}_{\mathscr{B}}^{\mathrm{T}} \boldsymbol{X}_{\mathscr{B}_{(j)}} \right)^2 \tag{3}$$

where $\boldsymbol{W}_{\mathscr{A}}$ is the similarity matrix of domain $\mathscr{A}$, and $\boldsymbol{W}_{\mathscr{A}_{(i,j)}}$ is the similarity between $\mathscr{A}_{(i)}$ and $\mathscr{A}_{(j)}$ calculated by the Gaussian function [22, 23], $\exp(-|\boldsymbol{X}_{\mathscr{A}_{(i)}} - \boldsymbol{X}_{\mathscr{A}_{(j)}}|^2 / \sigma^2)$. That of domain $\mathscr{B}$ is similarly defined.

(3) can also be simplified as

$$\mathrm{tr}\left(\boldsymbol{P}^{\mathrm{T}} \boldsymbol{X} \mathscr{M} \boldsymbol{X}^{\mathrm{T}} \boldsymbol{P}\right) \tag{4}$$

by introducing the graph Laplacian [22], $\mathscr{M} = \begin{bmatrix} \mathscr{M}_{\mathscr{A}} & 0 \\ 0 & \mathscr{M}_{\mathscr{B}} \end{bmatrix}$ where $\mathscr{M}_{\mathscr{A}} = \mathrm{diag}(\boldsymbol{W}_{\mathscr{A}}) - \boldsymbol{W}_{\mathscr{A}}$ and $\mathscr{M}_{\mathscr{B}} = \mathrm{diag}(\boldsymbol{W}_{\mathscr{B}}) - \boldsymbol{W}_{\mathscr{B}}$.

Consistency mapping: To be consistent of label classes, samples with labels are paired by each class. The class-wise pairing between labeled samples is performed across both domains and consists of all combinations of samples belonging to the same class. This pairing can be done by defining the indicator matrix $\boldsymbol{E}$ where $\boldsymbol{E}_{(i,j)}$ is 1 if the labeled samples in the different domains $\mathscr{A}_{(i)}$ and $\mathscr{B}_{(j)}$ belong to the same class, 0 otherwise.

$$\sum_{i,j} \boldsymbol{E}_{(i,j)} \left(\boldsymbol{P}_{\mathscr{A}}^{\mathrm{T}} \boldsymbol{X}_{\mathscr{A}_{(i)}} - \boldsymbol{P}_{\mathscr{B}}^{\mathrm{T}} \boldsymbol{X}_{\mathscr{B}_{(j)}} \right)^2 \tag{5}$$

Also, (5) can be rephrased as

$$\mathrm{tr}\left(\boldsymbol{P}^{\mathrm{T}} \boldsymbol{X} \mathscr{C} \boldsymbol{X}^{\mathrm{T}} \boldsymbol{P} \right) \tag{6}$$

where $\mathscr{C} = \mathrm{diag}(\boldsymbol{E}) - \boldsymbol{E}$. (6) stands for the L_2-norm, which means the distance in the feature space between class-wisely paired labeled samples between domains. In the proposed method, to enhance the effectiveness of consistency mapping, pseudo-labeling is applied prior to projection from the original domain into feature space. The detail of pseudo-labeling is described in Appendix S1.

Optimization: The objective function is derived by linearly combining above equations,

$$\underset{\boldsymbol{P}}{\mathrm{argmin.}}\ \gamma_{\mathscr{D}} \mathrm{tr}\left(\boldsymbol{P}^{\mathrm{T}} \boldsymbol{X} \mathscr{D} \boldsymbol{X}^{\mathrm{T}} \boldsymbol{P} \right) + \gamma_{\mathscr{M}} \mathrm{tr}\left(\boldsymbol{P}^{\mathrm{T}} \boldsymbol{X} \mathscr{M} \boldsymbol{X}^{\mathrm{T}} \boldsymbol{P} \right) + \gamma_{\mathscr{C}} \mathrm{tr}\left(\boldsymbol{P}^{\mathrm{T}} \boldsymbol{X} \mathscr{C} \boldsymbol{X}^{\mathrm{T}} \boldsymbol{P} \right) + \mathrm{tr}\left(\boldsymbol{P}^{\mathrm{T}} \boldsymbol{P} \right) \tag{7}$$

$$\text{s.t. } \boldsymbol{P}^{\mathrm{T}} \boldsymbol{X} \boldsymbol{H} \boldsymbol{X}^{\mathrm{T}} \boldsymbol{P} = \boldsymbol{I}$$

The last term, $\mathrm{tr}(\boldsymbol{P}^{\mathrm{T}}\boldsymbol{P})$ regularizes the complexity, and the constraints plays role of centering data in the feature space where $\boldsymbol{H} = \boldsymbol{I} - \frac{1}{n}\mathbb{1}$. The $\gamma_{\mathscr{D}}$, $\gamma_{\mathscr{M}}$, and $\gamma_{\mathscr{C}}$ are combining coefficients ($\gamma_* \geq 0$). Note that (7) is quadratic, and thus the problem is convex [24]. By denoting $\boldsymbol{\Theta}$ be the Lagrange multiplier and defining $\boldsymbol{\Psi} = \gamma_{\mathscr{D}} \mathscr{D} + \gamma_{\mathscr{M}} \mathscr{M} + \gamma_{\mathscr{C}} \mathscr{C}$, the Lagrangian $\mathbb{L}$ of (7) is derived as follows:

$$\mathbb{L} = \mathrm{tr}\left(\boldsymbol{P}^{\mathrm{T}} \left(\boldsymbol{X} \boldsymbol{\Psi} \boldsymbol{X}^{\mathrm{T}} + \boldsymbol{I} \right) \boldsymbol{P} \right) + \mathrm{tr}\left(\left(\boldsymbol{I} - \boldsymbol{P}^{\mathrm{T}} \boldsymbol{X} \boldsymbol{H} \boldsymbol{X}^{\mathrm{T}} \boldsymbol{P} \right) \boldsymbol{\Theta} \right). \tag{8}$$

The solution can be obtained by $\partial \mathbb{L} / \partial \boldsymbol{P} = 0$,

$$\left(\boldsymbol{X} \boldsymbol{\Psi} \boldsymbol{X}^{\mathrm{T}} + \boldsymbol{I} \right) \boldsymbol{P} = \boldsymbol{X} \boldsymbol{H} \boldsymbol{X}^{\mathrm{T}} \boldsymbol{P} \boldsymbol{\Theta}. \tag{9}$$

By defining $\boldsymbol{S} = (\boldsymbol{X}\boldsymbol{H}\boldsymbol{X}^{\mathrm{T}})^{-1}(\boldsymbol{X}\boldsymbol{\Psi}\boldsymbol{X}^{\mathrm{T}} + \boldsymbol{I})$, it boils down to

$$\boldsymbol{S}\boldsymbol{P} = \boldsymbol{P}\boldsymbol{\Theta} \tag{10}$$

which is the eigen-decomposition problem for $\boldsymbol{S}$ where $\boldsymbol{\Theta} = \mathrm{diag}(\theta_1, \cdots, \theta_k)$ contains the k largest eigenvalues and $\boldsymbol{P} = [\boldsymbol{p}_1, \cdots, \boldsymbol{p}_k]$ consists of the corresponding eigenvectors [25,26]. By solving (10), the optimal solution $\boldsymbol{P}$ of the can be simply obtained.

2.3. Mutual domain adaptation

Mutual domain adaptation is performed by bridging two domains in feature space. By means of (7) and (10), the samples of domain $\mathscr{A}$ and domain $\mathscr{B}$ are projected to $\mathscr{Z}$ via $\boldsymbol{P}$. Now, the two domains share a feature space, and are also ready to share label information. Graph-based SSL is employed for this. In $\mathscr{Z}$, a giant graph $G_{\mathscr{Z}} = (V_{\mathscr{Z}}, W_{\mathscr{Z}})$ is constructed, where $V_{\mathscr{Z}}$ is the node set of $n(= n_{\mathscr{A}} + n_{\mathscr{B}})$ cardinality and $W_{\mathscr{Z}} = \{w_{ij}\}$ is the similarity matrix of n samples calculated in the common feature space $\mathscr{Z}$ of dimension k. The Gaussian function [22,23] is used to calculate the w_{ij} with the transformed feature set $\boldsymbol{Z} = \boldsymbol{P}^{\mathrm{T}}\boldsymbol{X}$.

$$w_{ij} = \begin{cases} \exp\left(\dfrac{-\left| \mathbf{Z}_i - \mathbf{Z}_j \right|^2}{\sigma^2} \right) & \text{if } i \sim j \ (i \text{ and } j \text{ are } k-\text{nearest neighbors}) \\ 0 & \text{otherwise} \end{cases}$$

Note that there are no more domain boundaries. That is, in $\mathscr{Z}$, a sample from domain $\mathscr{A}$ can be linked with a sample not only from its own domain but also from domain $\mathscr{B}$, as long as they are similar.

By denoting the label set as $\boldsymbol{y} = (y_1, y_2, \cdots, y_l, y_{l+1} = 0, \cdots, y_n = 0)^{\mathrm{T}}$ where $y_i \in \{-1, +1\}$ for $i = 1, \cdots, l$ and $y_i = 0$ for the rest, the predicted value set defined as $\boldsymbol{f} = (f_1, f_2, \cdots, f_l, f_{l+1}, \cdots, f_n)^{\mathrm{T}}$ where $0 \leq f_i \leq 1$ can be obtained by solving the quadratic objective function

$$\underset{\boldsymbol{f}}{min}\ (\boldsymbol{f} - \boldsymbol{y})^{\mathrm{T}} (\boldsymbol{f} - \boldsymbol{y}) + \mu \boldsymbol{f}^{\mathrm{T}} \boldsymbol{L}_{\mathscr{Z}} \boldsymbol{f}$$

where $\boldsymbol{L}_{\mathscr{Z}}$ is the graph Laplacian, defined as $\boldsymbol{L}_{\mathscr{Z}} = \mathrm{diag}(W_{\mathscr{Z}}) - W_{\mathscr{Z}}$, and μ is a user-specified parameter which trades off the loss(the first term) and the smoothness(the second term). The solution is obtained as a closed form as below. The graph-based SSL has been well-established, and so further details can be found in [27,28].

$$\boldsymbol{f} = (\boldsymbol{I} + \mu \boldsymbol{L}_{\mathscr{Z}})^{-1} \boldsymbol{y}$$

As a result, all nodes in domain $\mathscr{A}$ and $\mathscr{B}$ make label predictions by sharing information through the gigantic graph $G_{\mathscr{Z}}$, which is mutually beneficial to both domains. The overall procedure for MDA is summarized in Algorithm 1.

3. Experimental results

The experimental results on the proposed method are described in this section. We present various empirical results for feature space alignment (Section 3.2), mutual domain adaptation (Section 3.3) and the ablation study (Section 3.4) on benchmark datasets.

3.1. Datasets

The proposed method was applied to various benchmark datasets. There are a total of five datasets, three text datasets (Amazon, Email, News) and two image datasets (Digit, Object). The summary and details for each are described in the Table 1 and below.

Amazon dataset is a collection of product reviews from Amazon.com [29]. There are four domains namely *books, DVDs, electronics,* and *kitchen appliances,* and each domain contains 2000 labeled reviews encoded in 400-dimensional feature vectors of unigrams and bigrams. One half of reviews are ranked 4 or 5 stars (labeled as +1) and the other half of them are ranked up to 3 stars (labeled as -1).

News dataset is a collection of news documents from 20Newsgroup [30]. There are 18,774 documents encoded in 26,214-dimensional bag-of-words vectors with 6 top categories and 20 subcategories in a hierarchical structure [31,32]. In our experiments, we recategorized newsgroups of 3 main categories and 12 subcategories. The task was to classify main categories (*comp, rec,* and *sci*). The details of News data are shown in Table 2 below.

Email dataset is a collection of personal emails in 2006 ECML-PKDD discovery challenge [33]. There are three inboxes (each has 2,500 emails) by different three users. Emails are encoded in 10,588-dimensional bag-of-words vectors, and one-half of them are *non-spam* (labeled as +1) and the other half of them are *spam* (labeled as -1) [34].

Algorithm 1
Mutual Domain Adaptation.

Input: Datasets for each domain $\{X_*, y_*\}$ ($*$: $\mathscr{A}$, $\mathscr{B}$)
Output: Label prediction f on the common feature space
(1) Pseudo-labeling for each domain
Construct graph $G_* = (V_*, W_*)$
Predict labels $f_* = (I + \mu L_*)^{-1} y_*$
Select samples as pseudo-labels with $|2f_* - 1| \geq \delta$
(2) Finding an optimal projection matrix
Concatenate data matrices as $X = \begin{bmatrix} X_{\mathscr{A}} & 0 \\ 0 & X_{\mathscr{B}} \end{bmatrix}$
Derive the objective function by (7)
Solve the Lagrangian $\mathbb{L} = \mathrm{tr}(P^{\mathrm{T}}(X\Psi X^{\mathrm{T}} + I)P) + \mathrm{tr}((I - P^{\mathrm{T}}XHX^{\mathrm{T}}P)\Theta)$
Eigen-decompose the matrix $(XHX^{\mathrm{T}})^{-1}(X\Psi X^{\mathrm{T}} + I)$ by (10)
Select k leading eigenvectors and construct the projection matrix P
(3) Mutual label propagation
Transform the original data to $Z = XP$
Construct graph $G_{\mathscr{Z}} = (V_{\mathscr{Z}}, W_{\mathscr{Z}})$ for mutual domains
Predict labels $f = (I + \mu L_{\mathscr{Z}})^{-1} y$
return Label prediction f on the common feature space

Table 1
Summary of benchmark datasets.

Dataset	Domain	Type	# of data	# of class
Amazon	Books, DVDs, Electronics, and Kitchens.	Text	8,000	2
News	Newsgroup A, B, C, and D	Text	12,000	3
Email	Inbox A, B, and C	Text	6,000	2
Digit	MNIST, USPS, and SynNumber	Image	15,000	10
Object	CIFAR10 and STL10	Image	8,100	9

Table 2
Categories of newsgroups of News dataset.

Newsgroup	Categories		
A	*comp.graphics*	*rec.autos*	*sci.crypt*
B	*comp.sys.ibm.pc.hardware*	*rec.motorcycles*	*sci.electronics*
C	*comp.sys.mac.hardware*	*rec.sport.baseball*	*sci.med*
D	*comp.windows.x*	*rec.sport.hockey*	*sci.space*

In our experiments, each inbox was set as a domain with 2,000 emails.

Digit dataset is a collection of digit images, and it consists of MNIST [35], USPS [36], and synthetic numbers (SynNumber) [37]. MNIST is a handwritten digit data containing 70,000 of 28 × 28 pixels grayscale images. USPS is automatically scanned digit data containing 9298 of 16 × 16 pixels grayscale images. SynNumber data is synthetically generated digits of English digits containing 12,000 of 32 × 32 pixels images. In our experiments, SynNumber data was transformed into grayscale images, and every data was set as a domain with 5000 images.

Object dataset is a collection of various object image data including CIFAR10 [38] and STL10 [39]. CIFAR10 data is 60,000 images of 32 × 32 pixels with 10 classes. STL10 data is 113,000 images of 96 × 96 pixels with 10 classes. In our experiments, the task was to classify the 9 common classes: *airplane, bird, automobile* (or *car*), *cat, deer, dog, horse, ship,* and *truck,* from chosen 5400 and 2700 images (600 and 300 per class) of CIFAR10 and STL10, respectively.

3.2. Results for feature space alignment

In this subsection, we describe results for domain adaptation depending on alignment of feature space. The effect of adaptation on features was validated on the Amazon data, and Fig. 3 depicts feature spaces between the books and DVDs. There two cases with t-SNE [40,41] visualizations: the left case when domain adaptation was not performed and the right case when our mutual domain adaptation was performed. First, mutual domain adaptation seems to make two distributions much closer. Also, it could be seen that manifolds of two domains are noticeable with the adaptation. Moreover, labeled samples gather and lie on the shared manifold by same class.

Also, in other domains in the Amazon dataset, features got much closer through mutual domain adaptation. The closeness between two domains was calculated by the Proxy $\mathscr{A}$-distance (PAD) [42]. The smaller the value of PAD, the closer the distance between domains, indicating that domain adaptation was successfully performed. Fig. 4 shows the PAD results according to various dimensions of the common feature space. In the books panel, the values on the bars represent the average PAD for all possible domain pairs with the books. The proposed method reduced PAD between domains in all dimensional settings and showed that the domain adaptation performed better with lower dimensions. In the proposed method, even if the feature dimensions of two domains are different, they are projected into the common feature space of the same dimension. In the k-dimensional space where the two domains are projected, the PAD becomes small due to the role of the MMD in making the feature distributions of the two domains similar with each other. The MMD forces the distance between two distributions to decrease even if the value of k is set to the same size as the original feature dimension: for example, the PAD decreases for an Amazon dataset with 400 dimensions, when k is 400. However, since k is a user-specified parameter, further research on how to derive an optimized value is needed. Therefore, in the experiments of this paper, the results according to various k values are presented in Fig. 4. As the value of k gradually decreases from 400 to 100 and 10, the feature dimension decreases, so the PAD also decreases. Furthermore, the feature projection into the low-dimensional space indicates that only similar information remains between the two domains. Although results could be good in terms of reducing the difference between domain, but it means loss of information, which decrease the discriminative power. The detailed experimental results for feature space alignment are reported in Table S1 of Appendix.

3.3. Results for mutual domain adaptation

In this subsection, we compare the performance of MDA based SSL (MDA-SSL), with that of SSL on a single domain (Single SSL). Every node in graphs was connected to five nearest neighbors. Edges between nodes were calculated by Gaussian function. For Single SSL, the input data

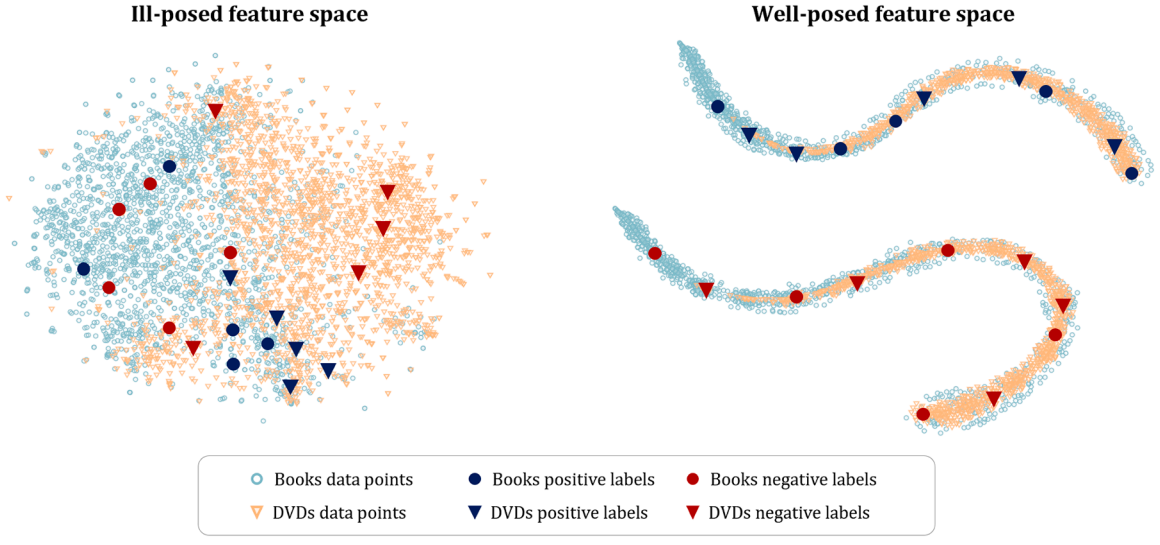

Fig. 3. Results for domain adaptation depending on alignment of feature space.

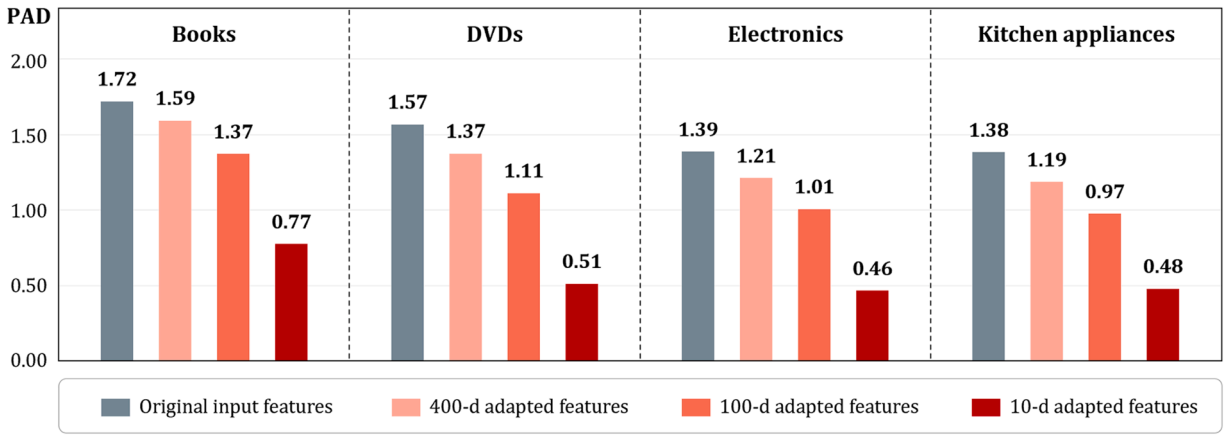

Fig. 4. Results for Proxy $\mathscr{A}$-distances.

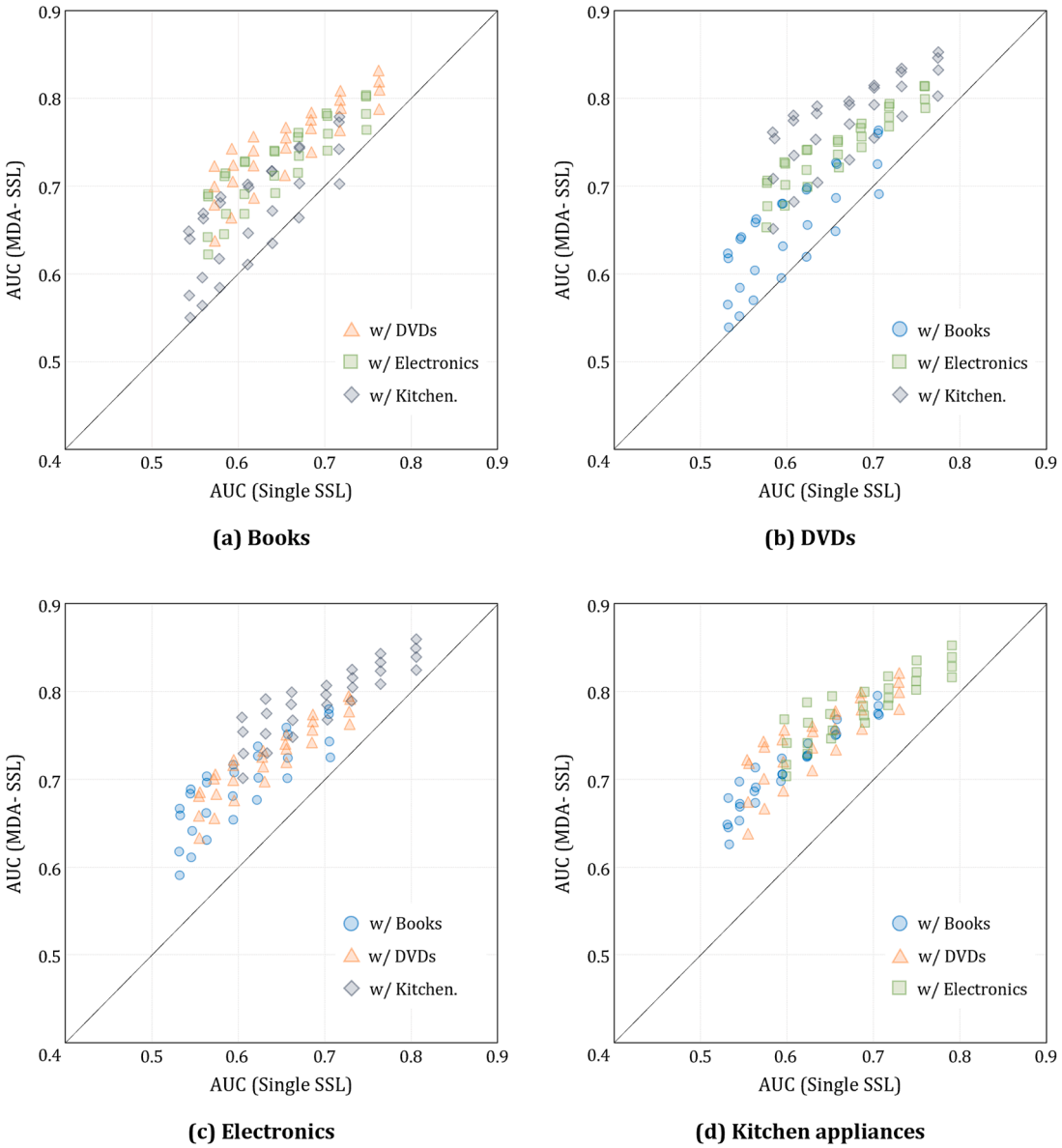

Fig. 5. Individual AUC comparison for Single SSL and MDA-SSL.

were original features. For MDA-SSL, the input data were the projected features where the number of dimensions varied over {10, 20, 50, 100}. Also, our method was performed on all combination of domains in each dataset $\{_4C_2, _4C_2, _3C_2, _3C_2, _2C_2\}$. Results were obtained with various proportions of labels {0.5%, 1%, 2%, 5%, 10%, 20%, 50%} on each graph. The parameter μ in SSL, which controls the degree of label smoothing, was determined to be the value that yielded the best results of each method in the range of {0.01, 0.1, 1, 10, 100} after the validation. The performance was measured by the area under receiving operating characteristic curve (AUC), and the entire experiment was repeated 100 times for each setting.

Fig. 5 shows the individual AUC comparison results for each domain in Amazon dataset. One point on the scatterplot represents the average result of 100 iterations for the experimental setup, and each plot has 84 points. The diagonal line is a criterion for performance comparison, and when a point is located above the line, the vertical axis performs better. In all plots in Fig. 5, since most points are located above the diagonal line, it can be seen that the four domains complementarily utilized the label information for prediction by mutual domain adaptation. In addition, in Fig. 5(c) and (d), the electronics and kitchen appliances showed the better classification results and stable performance compared to other domains. In particular, through mutual domain adaptation, those two domains make a significant contribution to improving each other's performance. It can be inferred that this result is due to the similarity between domains. Since the electronics and the kitchen appliances have in common that they are electronic products in a wide range, consumers' evaluation standards and expression methods for products have a similar context. As a result, it may be easier to perform domain adaptation on review data for the electronics and the kitchen appliances. With these results, mutual domain adaptation shows the performance improvement in all domains related to each other. It can also be inferred that the more similar the domains, the better the proposed method works.

Fig. 6 presents the details of performance improvement of Amazon dataset according to the proportions of labels. The thick dotted and solid lines in the plot represent the AUCs of Single SSL and MDA-SSL, respectively, and the bars represent the difference in AUC between two methods. Experimental results showed that MDA-SSL outperforms Single SSL in all proportions of labels with the average improvement rate of 16.2%. In particular, the improvement of 0.5% labels was the highest with 23.1% in that 43 and 182% higher than the improvement of 5% labels (16.1%) and that of 50% labels (8.2%) respectively. This result indicates that MDA-SSL improves better with fewer labels. Therefore, sparsely labeled domains may be more accurate when they adapt their label information mutually.

In addition, the SSL classification performance of MDA was compared with other methods: DICD [19], JGSA [20], and LHDA [21]. The overall comparison results are shown in Table 3, and the detailed results by each dataset are reported in Table S2 to S6 of Appendix. The results indicate that the proposed method generally performs better than other methods. In particular, it can be seen that the smaller the proportion of labels, the more pronounced the improvement of the proposed method. Sometimes the comparison methods perform better when there are many labels. It is considered that the pseudo-labeling included in the proposed method acted as noisy information for prediction. In other words, MDA utilizes pseudo-labeling when there are few labels to amplify the effect of consistency mapping to make prediction results more accurate, whereas when labels are sufficient, class-consistency between labeled nodes occurs more than necessary due to pseudo-labeling and is rather will have the opposite effect. Consequently, through the proposed method, the classification performance can be improved by mutually propagating labels between different domains, and the fewer the labels, the more robust the performance improvement.

Here are some remarks on the similarities and differences between the proposed method, MDA, and the comparison methods, DICD, JGSA,

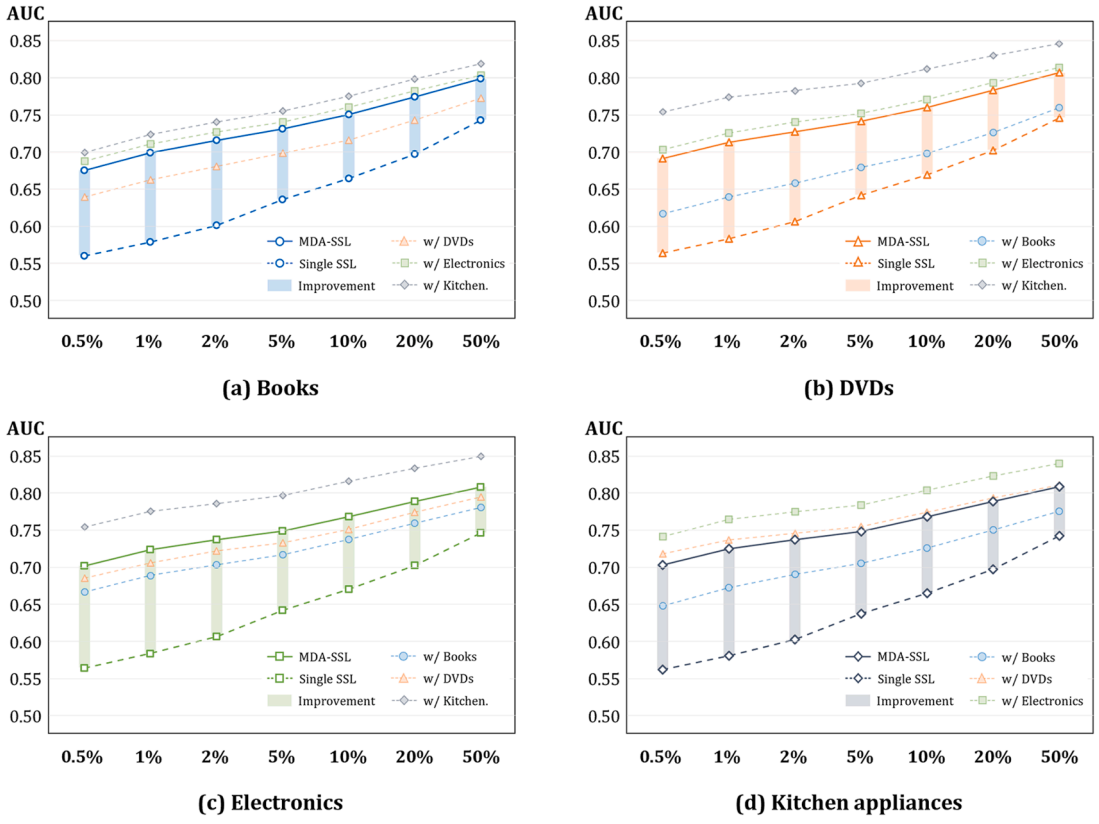

Fig. 6. Performance improvement by proportions of labels.

Table 3
Performance comparison for SSL classification.

Dataset	Method	Proportions of labels						
		0.5 %	1 %	2 %	5 %	10 %	20 %	50 %
Amazon	Single SSL	.563	.582	.604	.639	.668	.700	.744
	DICD	.639	.671	.711	.740	.764	.793	.837
	JGSA	.593	.665	.696	.721	.760	.791	.863
	LHDA	.596	.638	.663	.691	.727	.749	.761
	MDA-SSL	.693	.715	.729	.742	.762	.784	.805
News	Single SSL	.747	.792	.837	.891	.926	.953	.974
	DICD	.759	.801	.836	.890	.911	.928	.943
	JGSA	.751	.809	.812	.828	.870	.907	.926
	LHDA	.753	.766	.786	.812	.832	.857	.883
	MDA-SSL	.781	.816	.847	.883	.908	.927	.945
Email	Single SSL	.852	.887	.913	.939	.956	.968	.977
	DICD	.839	.853	.880	.915	.936	.956	.973
	JGSA	.715	.765	.805	.889	.938	.978	.987
	LHDA	.727	.776	.790	.863	.887	.919	.931
	MDA-SSL	.927	.937	.946	.955	.961	.964	.968
Digit	Single SSL	.887	.919	.943	.963	.975	.984	.991
	DICD	.883	.895	.922	.952	.967	.982	.991
	JGSA	.787	.807	.860	.910	.946	.973	.992
	LHDA	.780	.792	.807	.833	.869	.903	.984
	MDA-SSL	.948	.959	.969	.979	.984	.988	.992
Object	Single SSL	.564	.584	.605	.634	.659	.691	.733
	DICD	.596	.619	.644	.673	.698	.721	.744
	JGSA	.602	.622	.645	.675	.702	.728	.754
	LHDA	.579	.594	.612	.644	.672	.709	.735
	MDA-SSL	.608	.629	.653	.682	.708	.732	.756

and LHDA. The four methods are similar in that they project different domains onto a common feature space and perform similar labeling in common to clarify class regions. On the other hand, the main difference with MDA is that they distinguish between source and target domains and assume that the source domain is fully labeled. However, MDA does not require premise. That is, there is no distinction between source and target domains. In addition, MDA works when only a few labels are given, regardless of whether the labels belong to the source domain or the target domain. MDA thus provides more flexibility in designating source and target domains and is not limited by the amount of labeled data.

3.4. Ablation study

In this subsection, we describe results of the ablation study designed to figure out the effect of three objectives for mutual domain adaptation: *distribution matching, manifold preserving,* and *consistency mapping*. Each objective is formulated as a term in (7), and for this ablation study, we varied the objective function by combinations of those terms. Thereafter, seven projection matrices were derived by optimization, and same classification tasks were performed. Fig. 7 depicts the results of ablation study on Amazon dataset.

At first, the overall trend indicates that performances get better when the more terms are added on the objective function. For instance, by denoting the objective function as f, the AUCs are $\max(f_{\mathscr{D}}, f_{\mathscr{M}}) < f_{\mathscr{D}+\mathscr{M}}$, $\max(f_{\mathscr{M}}, f_{\mathscr{C}}) < f_{\mathscr{M}+\mathscr{C}}$, and $\max(f_{\mathscr{D}+\mathscr{M}}, f_{\mathscr{D}+\mathscr{C}}, f_{\mathscr{M}+\mathscr{C}}) < f_{\mathscr{D}+\mathscr{M}+\mathscr{C}}$. Therefrom, the distribution matching, manifold preserving, and consistency mapping are independent and helpful for adapting domains mutually. Next, the contribution of consistency mapping $\mathscr{C}$ is remarkable. In the case of 0.5% labels in book domain as shown in Fig. 7(a), the performance is above 0.6 even when the label information is adapted, which is about 15% higher than in $f_{\mathscr{D}}$ or $f_{\mathscr{M}}$. The remainder domains also show similar patterns. Additionally, with the more sparsely labeled data, $\mathscr{C}$ has the more significant role. Let us compare the case of 0.5% labels (the dotted line at the bottom) and the case of 50% labels. The increments to the worst performance are (a) 0.15 and 0.08, (b) 0.18 and 0.08, (c) 0.24 and 0.07, and (d) 0.30 and 0.08 where the averages of these increments are 0.22 with 0.5% labels and 0.08 with 50% labels. As a result, for domain adaptation tasks, the mutual connection of label information can have a greater discriminative power of features than only considering the properties of the original input data. Moreover, it could be more powerful when there are more sparsely labeled data.

Next, the effect of pseudo-labeling was empirically analyzed. We compared the SSL classification performance with and without pseudo-labeling for $f_{\mathscr{D}+\mathscr{M}+\mathscr{C}}$, and the results were derived from various proportions of labels as in the previous experiments, as shown in Table 4. The comparison results in Table 4 represent that the pseudo-labeling helps to improve performance when there are few labels, but it is not necessary when sufficient labels are given. It can be seen that the pseudo-labeling contributes to domain adaptation by generating labels when the data are labeled sparsely. The detailed experimental results for ablation study are reported in Table S7 of Appendix.

Furthermore, we compared results with three methods: TCA [9], SSDA [17], and MEDA [18]. TCA minimize the MMD between the source and target domains, SSDA regularizes manifold with similarity constraints integrating unlabeled data, and MEDA performs the dynamic distribution alignment and minimizes the structural risk in manifold together. The results of comparison are shown in Table 5, and they indicate that the proposed method outperforms three methods. In particular, by comparing results with the proportion of labels, our method seems to make features more discriminative in the sparsely labeled data. For the 0.5% labeled data, the average AUC of competitive methods was about 0.57. On the other hand, we improved 19.2% to 0.68. Domain adaptation aims to make predictions by adding other domain information when there is no label information. Considering this, it can be seen that the proposed method meets the purpose of domain adaptation.

4. Conclusion

Domain adaptation is one of strategies to solve the label sparsity problem. The label-sufficient source domain can transfer its information to the label-deficient target domain through a common feature space where two domains are represented as like one. Although existing methods are well-established, in real-world scenarios, most domains prefer to be the target because recent data becomes more sparsely labeled. In that there are very few sources domain, domain adaptation needs to be improved to be realistic.

In this paper, we presented the realistic domain adaptation method, which we refer to as mutual domain adaptation, MDA. The purpose of MDA is to search a common feature space of different domains where label information can be shared. The projection matrix transforms original feature sets and represents on the space. The projected features are derived by three objectives: *distribution matching, manifold preserving,* and *consistency mapping*. The distribution matching minimizes the distributional gap between two domains. By applying MMD, two feature sets are induced to resemble so that they are seen as to share one distribution. The manifold preserving minimizes the loss of original topologies. With the graph Laplacian, samples are smoothed along the manifold. The consistency mapping seeks the common feature space being consistent between samples by each class. Labeled samples are class-wise paired, and the projection matrix aligns the class boundaries. To enhance the effectiveness of consistency mapping, pseudo-labeling is applied prior to projection from the original domain into feature space. By applying semi-supervised learning, only prediction values with high confidence are selected as pseudo labels. At last, the three objectives for MDA are mathematically defined as an optimization problem, and the projection matrix for a common feature space is derived. We validated MDA on benchmark datasets with varying the proportion of labels. The experimental results show that the proposed method outperforms the relevant baselines, indicating the better for the sparser labeled data.

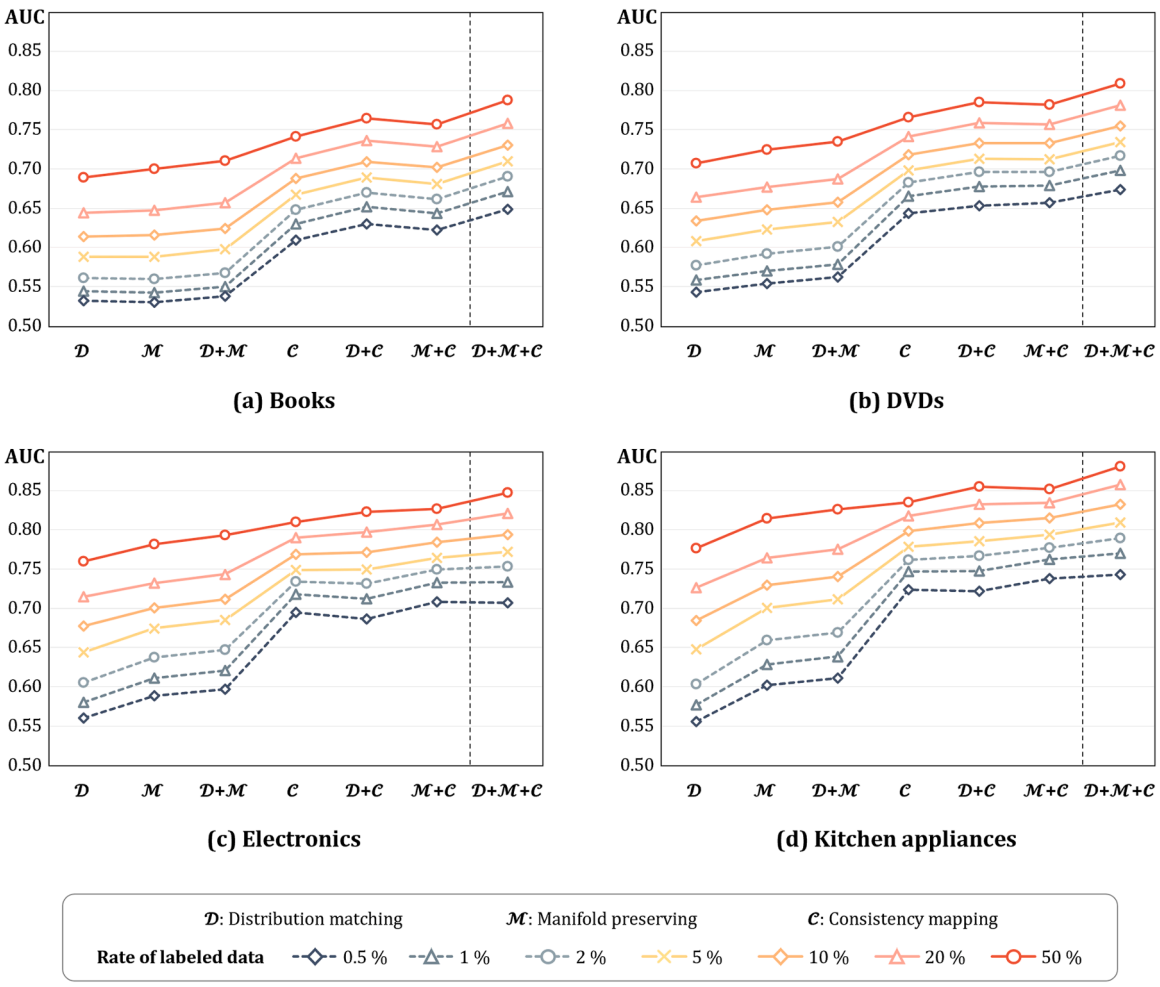

Fig. 7. Results for the ablation study.

Table 4
Results for the empirical analysis on the effect of pseudo-labeling.

Objective	Method	Proportions of labels						
		0.5 %	1 %	2 %	5 %	10 %	20 %	50 %
$\mathscr{D}+\mathscr{M}+\mathscr{C}$	w/o pseudo-labeling	.687	.703	.723	.757	.779	.807	.855
	w/ pseudo-labeling	.694	.719	.738	.757	.779	.805	.832

Further analysis of each objective in MDA showed that the more terms added in the objective function, the better the results.

In sum, our main contributions are summarized as follows. (a) We tackle a realistic problem setting of domain adaptation, where most domains are label-deficient and need to be helped and recent data become more sparsely labeled which makes the learning even more difficult. (b) To tackle this problem, we propose a mutual domain adaptation, which transfer label information both-way, to search a common feature space that matches different data distributions, preserves original manifolds, and maximize consistency between labeled samples with pseudo-labeling via semi-supervised learning. (c) We validate MDA on benchmark datasets for domain adaptation with varying the rate of labels, on which it outperforms relevant baselines and is especially better for the sparsely labeled data so as to be suitable for real-world scenarios.

However, a number of limitations, including future works, need to be addressed. First, the objective function includes several hyper-parameters to be optimized. Accordingly, considerable analyses and heuristics on the selection of its value should be further conducted. Second, it is required to further exploit non-linear projection. For the more discriminative and informative feature space, the non-linear projection would be better than the linear. Third, the integration of multiple domains and their optimal combination is reserved for our future study.

Table 5
Comparison results with other methods.

Competitive method	0.5% labeled data				5% labeled data				50% labeled data			
	B	D	E	K	B	D	E	K	B	D	E	K
TCA	.533	.544	.561	.556	.589	.609	.645	.648	.690	.708	.760	.777
SSDA	.531	.555	.589	.603	.589	.624	.675	.701	.701	.725	.782	.815
MEDA	.539	.563	.598	.612	.598	.633	.686	.712	.711	.736	.794	.827
Ours	.649	.675	.708	.744	.710	.735	.773	.810	.789	.810	.848	.881

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The data that has been used is confidential.

Acknowledgments

This study was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (MOE) (NRF-2022R1A6A3A01086784), the BK21 FOUR program of the NRF funded by the MOE (NRF5199991014091), the NRF grant funded by the MOE (2021R1A2C2003474), and the Ajou University research fund. This research was also supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by Ministry of Science & ICT (MSIT) (No. 2022-0-00653, Voice Phishing Information Collection and Processing and Development of a Big Data Based Investigation Support System), the NRF grant funded by the MSIT (NRF-2019R1A5A2026045), and the grant funded by the MSIT (KISTI Project No. K-23-L03-C02 and J-23-RD-CR02-S01).

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.patcog.2023.109919.

References

[1] S. Ben-David, J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, J.W. Vaughan, A theory of learning from different domains, Mach. Learn. 79 (2010) 151–175.
[2] J. Huang, A. Gretton, K. Borgwardt, B. Schölkopf, A. Smola, Correcting sample selection bias by unlabeled data, Adv. Neural Inf. Process. Syst. 19 (2006) 601–608.
[3] J. Wang, X.L. Zhang, Improving pseudo labels with intra-class similarity for unsupervised domain adaptation, Pattern Recognit. 138 (2023), 109379.
[4] W. Wang, H. Wang, Z. Zhang, C. Zhang, Y. Gao, Semi-supervised domain adaptation via Fredholm integral based kernel methods, Pattern Recognit. 85 (2019) 185–197.
[5] S.J. Pan, Q. Yang, A survey on transfer learning, IEEE Trans. Knowl. Data Eng. 22 (2009) 1345–1359.
[6] M. Chen, K.Q. Weinberger, J. Blitzer, Co-training for domain adaptation, nips, Citeseer 2011, pp. 2456–2464.
[7] G.R. Xue, W. Dai, Q. Yang, Y. Yu, Topic-bridged PLSA for cross-domain text classification, in: Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 2008, pp. 627–634.
[8] L. Bruzzone, M. Marconcini, Domain adaptation problems: A DASVM classification technique and a circular validation strategy, IEEE Trans. Pattern Anal. Mach. Intell. 32 (2009) 770–787.
[9] S.J. Pan, J.T. Kwok, Q. Yang, Transfer Learning via Dimensionality Reduction, AAAI, 2008, pp. 677–682.
[10] M.Y. Liu, O. Tuzel, Coupled generative adversarial networks, Adv. Neural Inf. Process. Syst. 29 (2016) 469–477.
[11] E. Tzeng, J. Hoffman, K. Saenko, T. Darrell, Adversarial discriminative domain adaptation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 7167–7176.
[12] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, V. Lempitsky, Domain-adversarial training of neural networks, J. Mach. Learn. Res. 17 (2016), 2096-2030.
[13] S.J. Pan, I.W. Tsang, J.T. Kwok, Q. Yang, Domain adaptation via transfer component analysis, IEEE Trans. Neural Netw. 22 (2010) 199–210.
[14] A. Gretton, K.M. Borgwardt, M.J. Rasch, B. Schölkopf, A. Smola, A kernel two-sample test, J. Mach. Learn. Res. 13 (2012) 723–773.
[15] A. Smola, A. Gretton, L. Song, B. Schölkopf, A Hilbert space embedding for distributions, in: Proceedings of the International Conference on Algorithmic Learning Theory, Springer, 2007, pp. 13–31.
[16] P. Ge, C.X. Ren, X.L. Xu, H. Yan, Unsupervised domain adaptation via deep conditional adaptation network, Pattern Recognit. 134 (2023), 109088.
[17] C. Wang, S. Mahadevan, Heterogeneous domain adaptation using manifold alignment, in: Proceedings of the 22nd International Joint Conference on Artificial Intelligence, 2011.
[18] J. Wang, W. Feng, Y. Chen, H. Yu, M. Huang, P.S. Yu, Visual domain adaptation with manifold embedded distribution alignment, in: Proceedings of the 26th ACM International Conference on Multimedia, 2018, pp. 402–410.
[19] S. Li, S. Song, G. Huang, Z. Ding, C. Wu, Domain invariant and class discriminative feature learning for visual domain adaptation, IEEE Trans. Image Process. 27 (2018) 4260–4273.
[20] J. Zhang, W. Li, P. Ogunbona, Joint geometrical and statistical alignment for visual domain adaptation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 1859–1867.
[21] Z. Zhou, Y. Wang, C. Niu, J. Shang, Label-guided heterogeneous domain adaptation, Multimedia Tools Appl. (2022) 1–22.
[22] X. Zhu, Z. Ghahramani, J.D. Lafferty, Semi-supervised learning using gaussian fields and harmonic functions, in: Proceedings of the 20th International Conference on Machine Learning (ICML-03), 2003, pp. 912–919.
[23] D. Zhou, O. Bousquet, T. Lal, J. Weston, B. Schölkopf, Learning with local and global consistency, Adv. Neural Inf. Process. Syst. 16 (2003).
[24] S. Boyd, S.P. Boyd, L. Vandenberghe, Convex Optimization, Cambridge University Press, 2004.
[25] L.N. Trefethen, D. Bau, Numerical Linear Algebra, Siam, 1997.
[26] K.R. Muller, S. Mika, G. Ratsch, K. Tsuda, B. Scholkopf, An introduction to kernel-based learning algorithms, IEEE Trans. Neural Netw. 12 (2001) 181–201.
[27] A. Subramanya, P.P. Talukdar, Graph-based semi-supervised learning, Synthesis Lectures on Artificial Intelligence and Machine Learning, 8 (2014) 1–125.
[28] Y. Chong, Y. Ding, Q. Yan, S. Pan, Graph-based semi-supervised learning: a review, Neurocomputing 408 (2020) 216–230.
[29] J. Blitzer, M. Dredze, F. Pereira, Biographies, bollywood, boom-boxes and blenders: domain adaptation for sentiment classification, in: Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, 2007, pp. 440–447.
[30] K. Lang, Newsweeder: learning to filter netnews. Machine Learning Proceedings 1995, Elsevier, 1995, pp. 331–339.
[31] T. Mitchell, Machine Learning (1997).
[32] T. Joachims, A probabilistic analysis of the rocchio algorithm with TFIDF for text categorization, carnegie-mellon univ pittsburgh pa dept of computer science 1996.
[33] S. Bickel, ECML-PKDD discovery challenge 2006 overview, in: Proceedings of the ECML-PKDD Discovery Challenge Workshop, 2006, pp. 1–9.
[34] B. Pfahringer, A Semi-Supervised Spam Mail Detector, Discovery Challenge Workshop, 2006.
[35] Y. LeCun, L. Bottou, Y. Bengio, P. Haffner, Gradient-based learning applied to document recognition, Proc. IEEE 86 (1998) 2278–2324.
[36] J.J. Hull, A database for handwritten text recognition research, IEEE Trans. Pattern Anal. Mach. Intell. 16 (1994) 550–554.
[37] Y. Ganin, V. Lempitsky, Unsupervised domain adaptation by backpropagation, in: Proceedings of the International Conference on Machine Learning, PMLR2015, 2023, pp. 1180–1189.
[38] A. Krizhevsky, G. Hinton, Learning multiple layers of features from tiny images (2009).
[39] A. Coates, A. Ng, H. Lee, An analysis of single-layer networks in unsupervised feature learning, in: Proceedings of the 14th International Conference on Artificial Intelligence and Statistics, JMLR Workshop and Conference Proceedings, 2011, pp. 215–223.
[40] L. Van der Maaten, G. Hinton, Visualizing data using t-SNE, J. Mach. Learn Res. 9 (2008).
[41] L. Van Der Maaten, Accelerating t-SNE using tree-based algorithms, J. Mach. Learn. Res. 15 (2014) 3221–3245.
[42] S. Ben-David, J. Blitzer, K. Crammer, F. Pereira, Analysis of representations for domain adaptation, Adv. Neural Inf. Process. Syst. 19 (2007) 137.

Sunghong Park received Ph.D. degree in Artificial Intelligence from Ajou University in 2022. His current research interest is on domain adaptation and biomedical informatics using various techniques of machine learning algorithms.

Myung Jun Kim received Ph.D. degree in Artificial Intelligence from Ajou University in 2021. His research interest is on machine learning, especially semi-supervised learning, algorithms and applications for networks with multi-layered structure.

Kanghee Park received Ph.D. degree in Industrial Engineering from Ajou University in 2014. His research interest is on graph-based semi-supervised learning, especially time-series prediction, and its application for stock price.

Hyunjung Shin received the Ph.D. degree in Data Mining from Seoul National University, and further majored in Machine Learning during her Post-Doc at Max Planck Institute (MPI) Tübingen in Germany. Since 2006, she joined Ajou University as a faculty member of the Department of Industrial Engineering. Currently, she provides academic services as a member of board of directors in Business Intelligence & Data Mining Society, Korean Institute of Information Scientists and Engineers (KIISE), Artificial Intelligence Society at KIISE, Korean Institute of Industrial Engineers, and Korean Society for Bioinformatics and Systems Biology. Also, she has been the vice chairman for Billing Software Inspection and Review Committee of Health Insurance Review and Assessment Service. Theory interest of her is focused on Machine Learning algorithms, particularly in Kernel and Semi-Supervised Learning methods. Her research activities range across diverse areas including network analytics, biomedical informatics, hospital fraud detection, etc.