# Wrapped Partial Label Dimensionality Reduction via Dependence Maximization

**Xiang-Ru Yu**[1,3] , **Deng-Bao Wang**[2,3] , **Min-Ling Zhang**[2,3*]
[1]School of Cyber Science and Engineering, Southeast University, China
[2]School of Computer Science and Engineering, Southeast University, China
[3]Key Laboratory of Computer Network and Information Integration (Southeast University), Ministry of Eduction, China
{yuxr, wangdb, zhangml}@seu.edu.cn

## Abstract

Partial label learning induces classifier from data with ambiguous supervision, where each instance is associated with a set of candidate labels but only one of which is valid. As a classic data preprocessing strategy, dimensionality reduction contributes to enhance the generalization capabilities of learning algorithms. Due to the ambiguity of supervision, existing works on partial label dimensionality reduction are confined to two separate stages: dimensionality reduction and partial label disambiguation. However, the decoupling of dimensionality reduction from partial label disambiguation can lead to severe performance degradation. In this paper, we present a novel approach called *Wrapped Partial Label Dimensionality Reduction* (WPLDR) to address this challenge. Specifically, WPLDR integrates the dimensionality reduction and partial label disambiguation within a unified framework, employing alternating optimization to concurrently perform dimensionality reduction and partial label disambiguation. WPLDR maximizes the interdependence between features in the embedded space and confidence-based label information, while simultaneously ensuring the manifold consistency between the embedded feature space and label space. Extensive experiments over a broad range of synthetic and real-world partial label data sets validate that the performance of well-established partial label learning algorithms can be significantly improved by the proposed WPLDR.

## 1 Introduction

Weakly supervised learning learns from data with limited supervision, where the annotations are usually incomplete (only a subset of data is labeled), inexact (ambiguous labels exist in data) or inaccurate (instances may be mislabeled) [Zhou *et al.*, 2018]. Partial label learning is a typical weakly-supervised paradigm, where each instance is associated with a candidate label set, among which only one is true [Tian *et al.*, 2023]. Compared with multi-class learning, partial label learning is more realistic and challenging [Gong *et al.*, 2021b]. Accordingly, the need to learn from partial labeled data naturally arises in many real-world applications such as crowdsoucing tagging [Ren *et al.*, 2024], part-of-speech tagging [Chen *et al.*, 2017] and face naming [Xu *et al.*, 2021], etc.

One intuitive approach towards partial label learning is label disambiguation, which aims to identify the only ground-truth label within candidate label set [He *et al.*, 2022; Xu *et al.*, 2023]. Generally, label disambiguation is pursued through two distinct strategies: averaging-based strategy [Ni *et al.*, 2021] and identification-based strategy [Lyu *et al.*, 2020a]. The former treats all potential positive labels in candidate label set equivalently, which distinguishes the averaged output on candidate labels labels from the outputs on non-candidate labels. On the other hand, identification-based disambiguation endeavors to recover the ground-truth label from candidate label set. This strategy treats the ground-truth label as latent variable and performs label disambiguation by optimizing the corresponding objective functions [Yu and Zhang, 2016], which are defined based on the maximum likelihood criterion $log(\sum_{y\in S_i} p(y|\boldsymbol{x_i},\boldsymbol{\theta}))$ or maximum margin criterion $max_{y_j\in S_i} f(\boldsymbol{x}_i,y_j) - max_{y_k\notin S_i} f(\boldsymbol{x}_i,y_k)$.

Dimensionality reduction is an effective technique for enhancing the generalization capability across various learning systems through alleviating the issue of curse of dimensionality [Ximendes *et al.*, 2022; Zhao *et al.*, 2023]. Existing works [Huang *et al.*, 2019] are mainly classified into two categories: unsupervised and supervised dimensionality reduction. Unsupervised dimensionality reduction aims to preserve the underlying structure and patterns in data without the help of supervision information [Niu *et al.*, 2023; Yao *et al.*, 2023]. As a representative algorithm, Principal component analysis (PCA) induces projection matrix by maximizing the variance of projected data [Hasan and Abdulazeez, 2021]. Supervised dimensionality reduction focuses on preserving class discriminative information, which utilizes the supervision information to guide the process of dimensionality reduction [Vogelstein *et al.*, 2021]. For example, Linear Discriminant Analysis (LDA) achieves this by maximizing the intra-class scatter and minimizing the inter-class scatter simultaneously [Sachin and others, 2015].

Due to the requirement for explicit class membership in defining objective functions, prior works on supervised dimensionality reduction heavily rely on ground-truth labels to induce projection matrices. Consequently, the intrinsic challenge of ambiguous supervision within partially labeled data hinders the application of supervised dimensionality reduction methods. Remarkably, the adaptation of dimensionality reduction techniques to address partial label learning remains a relatively unexplored problem. To the best of our knowledge, DELIN [Zhang *et al.*, 2022], CENDA [Bao *et al.*, 2021] and PLDA [Yu *et al.*, 2024] are the only existing dimensionality reduction designed for partial label learning, which induce projection matrix by adapting supervised dimensionality reduction methods into partial label learning. DELIN achieves dimensionality reduction through an alternating procedure that optimizes the LDA projection matrix based on disambiguation-guided labeling confidences. CENDA achieves dimensionality reduction by maximizing the dependence between projected features and confidence vectors of candidate labels, where the dependence is quantified by the Hilbert-Schmidt Independence Criterion (HSIC). PLDA further utilizes the weakly-supervised characteristics of partially labeled data.

Given the inherent ambiguity of partially labeled data, DELIN, CENDA and PLDA substitute the ground-truth labels with label confidence. Additionally, these label confidences are dynamically updated based on weighted voting from neighboring instances in projected feature space. Consequently, these methods are confined to operate within a two-stage framework encompassing dimensionality reduction and partial label disambiguation. In this process, the dimensionality reduction phase operates independently from the partial label disambiguation, which may lead to the result of dimensionality reduction being less satisfactory. This paper introduces an wrapped framework that unifies dimensionality reduction and partial label disambiguation, which enables their simultaneous execution in a cohesive manner within one stage. To attain this goal, we present a novel approach termed WPLDR, signifying Wrapped Partial label Dimensionality Reduction through dependence maximization. Specifically, by incorporating the manifold consistency in both the embedded feature space and label space, we propose a unified objective function to conduct confidence-based dimensionality reduction and similarity-based label disambiguation simultaneously. WPLDR maximizes the interdependence between the embedded features and the label confidences, while ensures the manifold consistency between projected instances and the corresponding label confidence vectors. Comprehensive experiments over a broad range of synthetic and real-world partial label data sets validate the effectiveness of proposed WPLDR.

## 2 Related Work

Partial label learning aims to learn from instances with ambiguous supervision, where the ground-truth labels are concealed in candidate label set [Gong *et al.*, 2022; Jia *et al.*, 2024; Wang and Zhang, 2022]. Accordingly, one intuitive way is label disambiguation [Xie *et al.*, 2021], which is usually achieved by two different strategies: averaging-based disambiguation [Cour *et al.*, 2011] and identification-based disambiguation [Jin and Ghahramani, 2002]. Averaging-based disambiguation equally treats each label in candidate label set and attempts to differentiate between the candidate and non-candidate labels [Tang and Zhang, 2017] [Cour *et al.*, 2011], and the prediction is made by aggregating the voting among the candidate labels of its neighboring examples [Xu *et al.*, 2019; Zhang *et al.*, 2016]. Although averaging-based disambiguation is intuitive and easy to implement, the output of ground-truth label is prone to be overwhelmed by the false positive labels in candidate label set, thereby leading to a degradation in the final predictive performance [Gong *et al.*, 2021a].

Identification-based disambiguation aims to recover the ground-truth label, which is treated as a latent variable, achieving disambiguation through iterative optimization of the objective function that includes these latent variables. For example, maximum likelihood methods identify the model parameter by solving $\boldsymbol{\theta}^* = \arg\max_{\boldsymbol{\theta}} \sum_{i=1}^{m} log(\sum_{y\in S_i} p(y|\boldsymbol{x}_i, \boldsymbol{\theta}))$, where the posterior probability is characterized by probabilistic graphical model [Liu and Dietterich, 2012; Dempster *et al.*, 1977]. Maximum margin methods define the objective function by maximizing the margin between the ground-truth label and other labels, i.e., $\big(\boldsymbol{w}_{y_i}^T \cdot \boldsymbol{x}_i + b_{y_i}\big) - \max_{\widetilde{y}_i \neq y_i}(\boldsymbol{w}_{\widetilde{y}_i}^T \cdot \boldsymbol{x}_i + b_{\widetilde{y}_i})$ [Yu and Zhang, 2016; Lyu *et al.*, 2020b]. Despite identification-based disambiguation attempts to recover the ground-truth label, the recovery process can be affected by false positive labels, leading to the error accumulation. In addition, contrary to the above approaches, disambiguation-free approaches induce classification model from partial labeled instances via problem transformation. Specifically, this kind of approaches transform the partial label learning problem into a series of binary classification problems by exploring the opposite relationship between candidate label set $S_i$ and non-candidate labels $\mathcal{Y} \setminus S_i$ [Zhang *et al.*, 2017; Lin *et al.*, 2022; Wu and Zhang, 2018].

The existing partial label learning literature mainly emphasizes the manipulation of label space. As a classic data preprocessing technique, dimensionality reduction exploits the manipulation in feature space, which is usually helpful to improve the generalization ability of learning algorithms. Depending on whether the label information is used, dimensionlity reduction can be classified into two categories, namely, unsupervised and supervised. Generally, unsupervised dimensionality reduction algorithms directly identify the projection matrix by preserving the underlying data structure. This kind of methods usually utilize manifold learning to achieve dimensionality reduction, including isometric mapping (ISOMAP) [Tenenbaum *et al.*, 2000], locally linear embedding (LLE) [Roweis and Saul, 2000], laplacian eigenmaps (LE) [Belkin and Niyogi, 2006] and locality preserving projection (LPP) [Jia *et al.*, 2023]. Supervised dimensionality reduction depends on ground-truth labels to determine within-class or between-class relationship and define objective function [Jia *et al.*, 2022]. LDA is a representative algorithm, which induces the projection matrix by maximizing the intra-class similarity and minimizing the inter-class similarity simultaneously. In the past few decades, some advances in supervised dimensionality reduction have been studied, such as canonical correlation analysis (CCA), partial least square and latent semantic indexing [Wang *et al.*, 2023]. However, due to the constraints of ambiguous supervision, existing supervised dimensionality reduction approaches are rarely used in partial label learning problem. To the best of our knowledge, DELIN, CENDA and PLDA are the only existing supervised dimensionality reduction approaches towards solving partial label learning. However, constrained by ambiguity supervision, these approaches are forced to utilize a two-stage learning strategy, which means that the dimensionality reduction process and partial label disambiguation process are independent from each other. Therefore, the inconsistency between these two processes may degrade the final performance.

## 3 The Proposed Approach

In this section, we first present our WPLDR framework, which performs dimensionality reduction and partial label disambiguation simultaneously. Then, an alternating optimization algorithm is introduced to solve the optimization problem.

### 3.1 Wrapped Partial Label Dimensionality Reduction

Let $\mathcal{X} = \mathbb{R}^d$ be the $d$-dimensional instance space and $\mathcal{Y} = \{y_1, y_2, \ldots, y_q\}$ denote the label space with $q$ labels. A partially labeled training set is denoted as $\mathcal{D} = \{(\boldsymbol{x}_i, S_i) \mid 1 \le i \le m\}$, where $\boldsymbol{x}_i \in \mathcal{X}$ is a $d$-dimensional feature vector $(x_{i1}, x_{i2}, \ldots, x_{id})^\top$ and $S_i \subseteq \mathcal{Y}$ is the corresponding candidate label set among in which the ground-truth label $y_i$ is concealed. The task of partial label learning is to induce a *multi-class* classifier $f : \mathcal{X} \mapsto \mathcal{Y}$ from training set $\mathcal{D}$.

Denote $\mathbf{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_m] \in \mathbb{R}^{d\times m}$ as the instance matrix and $\mathbf{Y} = [\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_m] \in \mathbb{R}^{m\times q}$ as the partial label matrix, where $y_{ij} = 1$ indicates that the $j$-th label belongs to the candidate label set of $\boldsymbol{x}_i$. Dimensionality reduction aims to seek a projection matrix $\mathbf{P} = [\boldsymbol{p}_1, \boldsymbol{p}_2, \ldots, \boldsymbol{p}_{d'}] \in \mathbb{R}^{d\times d'}$ $(d' \ll d)$ to map the instance matrix $\mathbf{X}$ into an embedded feature space characterized by $d'$-dimensional features, calculated as $\mathbf{X}' = \mathbf{P}^\top\mathbf{X}$. Constrained by the ambiguous supervision, previous approaches work in two-stage manner by firstly optimizing the projection matrix, and then leveraging the projected data to conduct candidate label disambiguation. Instead of two-stage dimensionality reduction framework, we attempt to perform dimensionality reduction and partial label disambiguation simultaneously in a unified framework.

To this end, we present a wrapped framework to jointly optimize label confidence, projection matrix and similarity weights to enhance the generalization performance. In this paper, we embrace the confidence-based HSIC as the dimensionality reduction term to induce projection matrix, and use the similarity matrix $\mathbf{S} \in \mathbb{R}^{m\times m}$ to characterize the manifold consistency between embedded feature space and label space. Then the objective function of proposed approach WPLDR is shown as follows:

$$\begin{aligned}\max_{\mathbf{P},\mathbf{S},\mathbf{F}} \ & \frac{1}{2}tr(\mathbf{H}\mathbf{X}^\top\mathbf{P}\mathbf{P}^\top\mathbf{X}\mathbf{H}\mathbf{F}\mathbf{F}^\top) \\ & - \frac{\alpha}{2}tr(\mathbf{P}^\top\mathbf{X}(\mathbf{I}_{m\times m}-\mathbf{S})(\mathbf{I}_{m\times m}-\mathbf{S})^\top\mathbf{X}^\top\mathbf{P}) \\ & - \frac{\beta}{2}tr(\mathbf{F}^\top(\mathbf{I}_{m\times m}-\mathbf{S})(\mathbf{I}_{m\times m}-\mathbf{S})^\top\mathbf{F}) \qquad (1) \\ s.t. \ \ & \mathbf{S}^\top\mathbf{1_m} = \mathbf{1_m}, \mathbf{0}_{m\times m} \le \mathbf{S} \le \mathbf{W}, \\ & \mathbf{F}\mathbf{1_q} = \mathbf{1_m}, \mathbf{0}_{m\times q} \le \mathbf{F} \le \mathbf{Y}, \\ & \boldsymbol{p}_i{}^\top(\mu\mathbf{X}\mathbf{X}^\top + (1-\mu)\mathbf{I})\boldsymbol{p}_j = \delta_{ij},\end{aligned}$$

where $\mathbf{F}$ denotes the label confidence matrix. $\mathbf{W} \in \{0,1\}^{m\times m}$ is adjacency matrix, in which $w_{ij} = 1$ if there exists an edge iff $\boldsymbol{x}'_i$ is among the $k$ nearest neighbors of $\boldsymbol{x}'_j$, otherwise, $w_{ij} = 0$. $\mathbf{S}$ is a non-negative similarity matrix. In addition, $\alpha$ and $\beta$ are the trade-off parameters to balance the dimensionality reduction and manifold information in label and projected feature space.

In the initial stage, the label confidence matrix $\mathbf{F}$, owing to the deficiency of discriminative ground-truth label, is initialized as $\mathbf{F}_0$ according to candidate label set as follows:

$$\forall\, 1 \le i \le m,\ 1 \le j \le q: \quad f_{ij} = \begin{cases} \frac{1}{|S_i|}, & \text{if } l_j \in S_i \\ 0, & \text{otherwise} \end{cases} \qquad (2)$$

The processed data set with lower-dimensional features is denoted as $\mathcal{D}' = \{(\boldsymbol{x}'_i, \boldsymbol{f}_i)|1 \le i \le m\}$, where $\boldsymbol{x}'_i = \mathbf{P}^\top\boldsymbol{x}_i$. A weighted graph $\mathcal{G} = \{\mathbf{V}, \boldsymbol{\mathcal{E}}, \mathbf{S}\}$ is built on the low dimensional feature space. Here, $\mathbf{V} = \{\boldsymbol{x}'_i|0 \le i \le m\}$ denotes the set of vertices, and $\boldsymbol{\mathcal{E}} = \{(\boldsymbol{x}'_i, \boldsymbol{x}'_j)|\boldsymbol{x}'_i \in k\mathrm{NN}(\boldsymbol{x}'_j)\}$ represents the set of edges between $\boldsymbol{x}'_i$ and $\boldsymbol{x}'_j$. For the construction of $\boldsymbol{\mathcal{E}}$, there exists an edge iff $\boldsymbol{x}'_i$ is among the $k$ nearest neighbors of $\boldsymbol{x}'_j$. $\mathbf{S}$ characterizes the similarity weight between $(\boldsymbol{x}'_i, \boldsymbol{x}'_j)$, where $s_{ij} > 0$ if $(\boldsymbol{x}'_i, \boldsymbol{x}'_j) \in \boldsymbol{\mathcal{E}}$, $s_{ij} = 0$ when $(\boldsymbol{x}'_i, \boldsymbol{x}'_j) \notin \boldsymbol{\mathcal{E}}$. Then, based on the graph structure and low-dimensional features, the similarity graph weight matrix $\mathbf{S}$ can be calculated by solving the following linear least square problem, which is defined as the reconstruction loss in embedded feature space:

$$\begin{aligned}\min_{\mathbf{S}} \ & tr(\mathbf{P}^\top\mathbf{X}(\mathbf{I}_{m\times m}-\mathbf{S})(\mathbf{I}_{m\times m}-\mathbf{S})^\top\mathbf{X}^\top\mathbf{P}) \\ s.t. \ & \mathbf{S}^\top\mathbf{1}_m = \mathbf{1}_m, \\ & s_{ij} \ge 0, (\boldsymbol{x}'_i, \boldsymbol{x}'_j) \in \boldsymbol{\mathcal{E}}, s_{ij} = 0, (\boldsymbol{x}'_i, \boldsymbol{x}'_j) \notin \boldsymbol{\mathcal{E}},\end{aligned} \qquad (3)$$

where $\mathbf{1}_m$ is a $m$-dimensional all-ones column vector, the value of $s_{ij}$ implies the relative contribution of instance $\boldsymbol{x}'_i$ on $\boldsymbol{x}'_j$.

Following the smoothness assumption, which states that the manifold structure in projected feature space should be preserved in label space [Wang *et al.*, 2021; Song *et al.*, 2022], label confidence matrix $\mathbf{F} = [\boldsymbol{f}_1, \boldsymbol{f}_2, \ldots, \boldsymbol{f}_m]^\top$ can be optimized by solving the following problem:

$$\begin{aligned}\min_{\mathbf{F}} \ & tr(\mathbf{F}^\top(\mathbf{I}_{m\times m}-\mathbf{S})(\mathbf{I}_{m\times m}-\mathbf{S})^\top\mathbf{F}) \\ s.t. \ & \mathbf{S}^\top\mathbf{1}_m = \mathbf{1}_m, \mathbf{F}\mathbf{1_q} = \mathbf{1_m}, \\ & s_{ij} \ge 0, (\boldsymbol{x}'_i, \boldsymbol{x}'_j) \in \boldsymbol{\mathcal{E}}, s_{ij} = 0, (\boldsymbol{x}'_i, \boldsymbol{x}'_j) \notin \boldsymbol{\mathcal{E}} \\ & f_{il} \ge 0, (0 \le l \le q), f_{il} = 0, (\forall y_{il} = 0).\end{aligned} \qquad (4)$$

In partial label learning, feature vector and label confidence vector elucidates each instance from two perspectives. In our framework, we achieve partial label dimensionality reduction by maximizing the dependence between projected feature and label information. WPLDR employs the HSIC to measure the dependence between them, and the corresponding empirical estimate of HSIC is denoted as:

$$\mathrm{HSIC}(\mathcal{F}, \mathcal{Q}) = (m-1)^{-2}tr(\mathbf{HKHL}) \qquad (5)$$

where $tr$ is the trace operator of matrix. $\mathbf{H} = \mathbf{I} - \frac{1}{m}\boldsymbol{e}\boldsymbol{e}^\top$, and $\boldsymbol{e}$ is a column vector with the same value 1. $\mathcal{F}$ and $\mathcal{Q}$ denote

**Inputs:**

$\mathcal{D}$: partial label training data set $\{(\boldsymbol{x}_i, S_i) \mid 1 \leq i \leq m\}$ ($\mathcal{X} \in \mathbb{R}^d, \mathcal{Y} = \{l_1, l_2, \ldots, l_q\}, \boldsymbol{x}_i \in \mathcal{X}, S_i \subseteq \mathcal{Y}$)
$d'$: the number of retained dimension after dimensionality reduction
$k$: the number of nearest neighbors used to update the label confidence matrix
$\alpha$: the feature space trade-off parameter
$\beta$: the label space trade-off parameter
$\mu$: the constraints trade-off parameter

**Outputs:**

$\mathbf{P}$: the $d \times d'$ projection matrix via the WPLDR
$\mathcal{D}'$: the transformed lower-dimensional partial label training set $\{(\boldsymbol{x}'_i, S_i) \mid 1 \leq i \leq m\}$

**Process:**

Initialize the $m \times q$ label confidence matrix $\mathbf{F}_0$ as shown in Eq. (2);
Cascade the training data into the instance matrix $\mathbf{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_m]$;
Initialize the $d \times d'$ projection matrix $\mathbf{P}_0$ via dependence maximization between embedded feature and label information as shown in Eq.(7);
**repeat**
  Calculate the similarity matrix $\mathbf{S}$ according to the embedded feature vectors and label confidence matrix via Eq.(8);
  Calculate $\mathbf{H} = \mathbf{I} - \frac{1}{m}\boldsymbol{e}\boldsymbol{e}^T$;
  Update the label confidence matrix $\mathbf{F}$ as shown in Eq. (12), which is a transformed problem of WPLDR in Eq. (1);
  Update projection matrix $\mathbf{P}$, and solve the transformed problem in Eq.(13). Given the generalized eigenvalue problem in Eq.(16), then the projection matrix is obtained by concatenating the $d'$ eigenvectors w.r.t. the top $d'$ eigenvalues;
**until** convergence
Derive the lower-dimensional partial label training data sets $\mathcal{D}'$ with $d'$ features via the projection matrix $\mathbf{P}$, $\mathbf{X}' = \mathbf{P}^\top\mathbf{X}$;

Table 1: The pseudo-code of WPLDR.

the reproducing kernel Hilbert space mapped from $\mathcal{X}$ and $\mathcal{Y}$ respectively. Substituting $\mathbf{K} = \mathbf{X}^\top\mathbf{P}\mathbf{P}^\top\mathbf{X}$ and $\mathbf{L} = \mathbf{F}\mathbf{F}^\top$ into Eq.(5) and dropping the normalization term, the objective function is rewritten as follows:

$$\boldsymbol{p}^* = \arg\max_{\mathbf{P}} tr(\mathbf{H}\mathbf{X}^\top\boldsymbol{p}\boldsymbol{p}^\top\mathbf{X}\mathbf{H}\mathbf{F}\mathbf{F}^\top) \tag{6}$$

Then, to avoid the impact of scaling, constraint on $l_2$ norm is imposed on $\boldsymbol{p}$, i.e. $\boldsymbol{p}^\top\boldsymbol{p} = 1$. Furthermore, considering that the features in projected feature space still remains some redundant information, we introduce a constraint to ensure projected features maintain uncorrelated as ${\boldsymbol{p}_i}^\top\mathbf{X}\mathbf{X}^\top\boldsymbol{p}_j = \delta_{ij}$, where $\delta_{ij}$ is Kronecker delta function. By combining the above two constraints, we can rewrite the optimization problem in Eq.(5) as:

$$\begin{aligned}&\max_{\mathbf{P}}\ tr(\mathbf{H}\mathbf{X}^\top\mathbf{P}\mathbf{P}^\top\mathbf{X}\mathbf{H}\mathbf{F}\mathbf{F}^\top)\\ &s.t.\ {\boldsymbol{p}_i}^\top(\mu\mathbf{X}\mathbf{X}^\top + (1-\mu)\mathbf{I})\boldsymbol{p}_j = \delta_{ij},\end{aligned} \tag{7}$$

where $\mu \in (0, 1)$ is the trade-off parameter to balance the weight of two constraints in inducing projection matrix.

### 3.2 Alternative Optimization

As shown in Eq.(1), the objective function of WPLDR contains three sets of variables with different constraints, which is hard to be solved directly. Thus, in this paper, we leverage alternative optimization to tackle this problem. Specifically, each set of variables will be iteratively optimized by fixing other sets of variables until convergence or the maximum number of iterations reaches.

**Update** $\mathbf{S}$ with fixed $\mathbf{F}$ and $\mathbf{P}$, the objective function of WPLDR is rewritten as:

$$\begin{aligned}\min_{\mathbf{S}}\ &\frac{\alpha}{2}tr(\mathbf{P}^\top\mathbf{X}(\mathbf{I}_{m\times m} - \mathbf{S})(\mathbf{I}_{m\times m} - \mathbf{S})^\top\mathbf{X}^\top\mathbf{P})\\ &+ \frac{\beta}{2}tr(\mathbf{F}^\top(\mathbf{I}_{m\times m} - \mathbf{S})(\mathbf{I}_{m\times m} - \mathbf{S})^\top\mathbf{F})\\ &s.t.\ \ \mathbf{S}^\top\mathbf{1_m} = \mathbf{1_m}, \mathbf{0}_{m\times m} \leq \mathbf{S} \leq \mathbf{W}.\end{aligned} \tag{8}$$

For each instance, the similarity vector is independent, which indicates that each column in $\mathbf{S}$ is independent of other columns. Thus, we optimize the similarity vector one by one. Accordingly, for the $j$-th instance $\boldsymbol{x}'_j$, the optimization problem of $j$-th column vector in $\mathbf{S}$ is defined as follows:

$$\begin{aligned}&\min_{\mathbf{S_{.j}}} \alpha \mid\mid \boldsymbol{x}'_j - \sum_{(\boldsymbol{x}'_i,\boldsymbol{x}'_j)\in\mathcal{E}} s_{ij}\boldsymbol{x}'_i \mid\mid_2^2 + \beta \mid\mid \boldsymbol{f}_j - \sum_{(\boldsymbol{x}'_i,\boldsymbol{x}'_j)\in\mathcal{E}} s_{ij}\boldsymbol{f}_i \mid\mid_2^2\\ &s.t.\ \ \mathbf{S}^\top\mathbf{1}_m = \mathbf{1}_m, \mathbf{0}_{m\times m} \leq \mathbf{S} \leq \mathbf{W},\end{aligned} \tag{9}$$

where $\boldsymbol{x}'_i \in \mathbb{R}^{d'}$ denotes the feature vector from embedded feature space. The value of $s_{ij}$ in $\mathbf{S}$ characterizes the relative importance of neighboring instance $\boldsymbol{x}'_i$ on $\boldsymbol{x}'_j$. Furthermore, denote matrix $\mathbf{E}^{f_j} = [\boldsymbol{f}_j - \boldsymbol{f}_{\mathcal{N}_{j(1)}}, \boldsymbol{f}_j - \boldsymbol{f}_{\mathcal{N}_{j(2)}}, \ldots, \boldsymbol{f}_j - \boldsymbol{f}_{\mathcal{N}_{j(k)}}]^\top \in \mathbb{R}^{k\times q}$ and $\mathbf{E}^{x'_j} = [\boldsymbol{x}'_j - \boldsymbol{x}'_{\mathcal{N}_{j(1)}}, \boldsymbol{x}'_j - \boldsymbol{x}'_{\mathcal{N}_{j(2)}}, \ldots, \boldsymbol{x}'_j - \boldsymbol{x}'_{\mathcal{N}_{j(k)}}]^\top \in \mathbb{R}^{k\times d'}$, then the optimization problem in Eq. (9) can be rewritten as follows:

$$\begin{aligned}&\min_{\hat{\boldsymbol{s}}_j} \hat{\boldsymbol{s}}_{\boldsymbol{j}}^\top(\alpha\mathbf{G}^{f_j} + \beta\mathbf{G}^{x_j})\hat{s}_j\\ &s.t.\ \ \hat{\boldsymbol{s}}_{\boldsymbol{j}}^\top\mathbf{1}_k = 1, \mathbf{0}_k \leq \hat{\boldsymbol{s}}_{\boldsymbol{j}} \leq \mathbf{1}_k,\end{aligned} \tag{10}$$

where $\mathbf{G}^{\boldsymbol{f}_j}$ and $\mathbf{G}^{\boldsymbol{x}'_j}$ are Gram matrices on label space and projected feature space, i.e. $\mathbf{G}^{\boldsymbol{f}_j} = \mathbf{E}^{\boldsymbol{f}_j}(\mathbf{E}^{\boldsymbol{f}_j})^T$ and $\mathbf{G}^{\boldsymbol{x}'_j} = \mathbf{E}^{\boldsymbol{x}'_j}(\mathbf{E}^{\boldsymbol{x}'_j})^T$. The optimization problem in Eq. (10) is a standard Quadratic Programming (QP) problem, which can be efficiently solved by off-the-shelf QP tools.

**Update F** with fixed **P** and **S**, the objective problem in Eq.(1) can be stated as follows:

$$\begin{aligned} \max_{\mathbf{F}} \ & \frac{1}{2} tr(\mathbf{F}^\top \mathbf{H}\mathbf{X}^\top \mathbf{P}\mathbf{P}^\top \mathbf{X}\mathbf{H}\mathbf{F}) \\ & - \frac{\beta}{2} tr(\mathbf{F}^\top (\mathbf{I}_{m\times m} - \mathbf{S})(\mathbf{I}_{m\times m} - \mathbf{S})^\top \mathbf{F}), \\ & s.t. \ \ \mathbf{F}\mathbf{1_q} = \mathbf{1_m}, \mathbf{0}_{m\times q} \leq \mathbf{F} \leq \mathbf{Y}. \end{aligned} \tag{11}$$

Here, we define a square matrix $\mathbf{T} = \frac{\beta}{2}(\mathbf{I}_{m\times m} - \mathbf{S})(\mathbf{I}_{m\times m} - \mathbf{S})^\top - \frac{1}{2}\mathbf{H}\mathbf{X}^\top\mathbf{P}\mathbf{P}^\top\mathbf{X}\mathbf{H}$. Although $\mathbf{T}$ is symmetric, it may be a indefinite matrix for some datasets. In fact, some optimization tools attempt to solve the indefinte QP problem like Gurobi, but the efficiency is less satisfactory. Fortunately, under close scrutiny, the first term in Eq.(12) is convex, and the last term is concave, thus, it is a constrained convex-concave problem. Accordingly, we can utilize the Convex-Concave Procedure (CCCP) to solve the problem, which solves the original nonconvex problem by optimizing a sequence of convex problems. Specifically, in each iteration, the second term in Eq. (12) is replaced by its first order Taylor approximation, which can be rewritten as the following form:

$$\begin{aligned} \mathbf{F}^{i+1} = & \arg\min_{\mathbf{F}} \frac{\beta}{2} tr(\mathbf{F}^\top (\mathbf{I}_{m\times m} - \mathbf{S})(\mathbf{I}_{m\times m} - \mathbf{S})^\top \mathbf{F}) \\ & - tr(\mathbf{F}^\top \mathbf{H}\mathbf{X}^\top \mathbf{P}\mathbf{P}^\top \mathbf{X}\mathbf{H}\mathbf{F}^i) \\ & s.t. \ \ \mathbf{F}\mathbf{1_q} = \mathbf{1_m}, \mathbf{0}_{m\times q} \leq \mathbf{F} \leq \mathbf{Y}. \end{aligned} \tag{12}$$

**Update P** with fixed **S** and **F**, the objective function can be stated as follows:

$$\begin{aligned} \max_{\mathbf{P}} \ & \frac{1}{2} tr(\mathbf{P}^\top \mathbf{X}\mathbf{H}\mathbf{F}\mathbf{F}^\top \mathbf{H}\mathbf{X}^\top \mathbf{P}) \\ & - \frac{\alpha}{2} tr(\mathbf{P}^\top \mathbf{X}(\mathbf{I}_{m\times m} - \mathbf{S})(\mathbf{I}_{m\times m} - \mathbf{S})^\top \mathbf{X}^\top \mathbf{P}) \\ & s.t. \ \ \boldsymbol{p}_i{}^\top (\mu \mathbf{X}\mathbf{X}^\top + (1-\mu)\mathbf{I})\boldsymbol{p}_j = \delta_{ij}, \end{aligned} \tag{13}$$

where $\mu \in (0,1)$ is a trade-off parameter which balances the importance of the above two constraints.

By Lagrange method, the Lagrange function is induced as:

$$\begin{aligned} \mathcal{L}(\mathbf{P}) = & \frac{1}{2} tr(\mathbf{P}^\top \mathbf{X}\mathbf{H}\mathbf{F}\mathbf{F}^\top \mathbf{H}\mathbf{X}^\top \mathbf{P}) \\ & - \frac{\alpha}{2} tr(\mathbf{P}^\top \mathbf{X}(\mathbf{I}_{m\times m} - \mathbf{S})(\mathbf{I}_{m\times m} - \mathbf{S})^\top \mathbf{X}^\top \mathbf{P}) \\ & + tr(\Lambda(\mathbf{I} - \mathbf{P}^\top(\mu \mathbf{X}\mathbf{X}^\top + (1-\mu)\mathbf{I})\mathbf{P})), \end{aligned} \tag{14}$$

where $\Lambda$ is a diagonal matrix whose entries are Lagrange multipliers. By setting the derivative of Eq.(14) as 0, we can obtain:

$$\begin{aligned} & \frac{\partial \mathcal{L}}{\partial \mathbf{P}} = -2(\mu \mathbf{X}\mathbf{X}^\top + (1-\mu)\mathbf{I})\mathbf{P}\Lambda \\ & +2(\mathbf{X}\mathbf{H}\mathbf{F}\mathbf{F}^\top \mathbf{H}\mathbf{X}^\top - \alpha \mathbf{X}(\mathbf{I}_{m\times m} - \mathbf{S})(\mathbf{I}_{m\times m} - \mathbf{S})^\top \mathbf{X}^\top). \end{aligned} \tag{15}$$

Finally, we induce the projection matrix by solving the following generalized eigenvalue problem:

$$\begin{aligned} & (\mathbf{X}\mathbf{H}\mathbf{F}\mathbf{F}^\top \mathbf{H}\mathbf{X}^\top - \alpha \mathbf{X}(\mathbf{I}_{m\times m} - \mathbf{S})(\mathbf{I}_{m\times m} - \mathbf{S})^\top \mathbf{X}^\top)\boldsymbol{p} \\ & = \lambda(\mu \mathbf{X}\mathbf{X}^\top + (1-\mu)\mathbf{I})\boldsymbol{p}, \end{aligned} \tag{16}$$

where $\mathbf{P}$ is obtained by selecting the $d'$ eigen vectors of the top $d'$ eigenvalues.

# 4 Experiment

## 4.1 Experimental Setup

To evaluate the effectiveness of proposed WPLDR, we couple five state-of-the-art partial label learning algorithms with four partial dimensionality reduction approaches, DELIN, CENDA, PLDA and the proposed WPLDR. For each partial label learning method $\mathcal{L}$, the coupled version is denoted as $\mathcal{L}$-DELIN, $\mathcal{L}$-CENDA and $\mathcal{L}$-WPLDR respectively.

In this paper, we instantiate $\mathcal{L}$ with five well-established partial label learning algorithms, and their parameter configurations are set based on the recommendations provided in corresponding literatures.

- PL-KNN [Hüllermeier and Beringer, 2006]: an averaging-based partial label learning algorithm, which makes prediction by weighted voting on candidate labels from $k$NN instances [suggested configuration: $k$=10].
- PL-SVM [Nguyen and Caruana, 2008]: an identification-based partial label learning approach which induces classification model by adapting maximum margin [suggested configuration: regularization parameter pool with $\{10^{-3}, \dots, 10^3\}$].
- IPAL [Zhang and Yu, 2015]: a disambiguation-based partial label learning method, which determines the valid label via label propagation on weighted graph [suggested configuration: $k$=10, balancing parameter $\alpha = 0.95$].
- SURE [Feng and An, 2019]: a self-training partial label learning algorithm, under proper constraints, which unifies the model training and identification of pseudo label into one formulation [suggested configuration: regularization parameters $\lambda$ = 0.3, $\beta$ = 0.05].
- PL-AGGD [Wang *et al.*, 2021]: an adaptive graph guided disambiguation algorithm, which jointly performs graph construction, model training and partial label disambiguation in a framework. [suggested configuration: $k = 10$, $\mu = 1$ and $\gamma = 0.05$ ].

In the following subsections, for each dataset, we perform ten-fold cross-validation, while the mean and standard deviation of classification results are reported.

## 4.2 Results on Real-world Data Sets

Seven real-world partial label data sets have been collected from different tasks and domains. Due to page limit, Table 3 reports the experimental results on real-world as well as synthetic partial label data sets with different configurations. As shown in Table 3, the experimental results illustrate the classification accuracy of partial label algorithms before and after employing three dimensionality reduction approaches DELIN,

| Comparing Algorithm | Data Set | | | | | | |
|---|---|---|---|---|---|---|---|
| | Amazon | Enron | Dermatology | Winerate | Zoo | Segm-2500 | Segm-3000 |
| | $r = 2$ (two false positive label) | | | | | | |
| PL-KNN | 0.024±0.010 | 0.543±0.045 | 0.896±0.038 | 0.853±0.093 | 0.491±0.039 | 0.179±0.018 | 0.218±0.032 |
| PL-KNN-DELIN | 0.483±0.049 | 0.503±0.039 | 0.911±0.042 | 0.872±0.115 | 0.513±0.056 | 0.329±0.021 | 0.331±0.033 |
| PL-KNN-CENDA | 0.493±0.045 | 0.529±0.060 | 0.908±0.051 | 0.931±0.048 | 0.490±0.033 | 0.344±0.025 | 0.332±0.021 |
| PL-KNN-PLDA | 0.504±0.027 | 0.482±0.050 | 0.923±0.045 | 0.912±0.053 | 0.492±0.047 | 0.348±0.022 | 0.359±0.021 |
| PL-KNN-WPLDR | **0.520±0.028** | **0.607±0.039** | **0.944±0.047** | **0.931±0.067** | **0.517±0.059** | **0.348±0.030** | **0.376±0.037** |
| PL-SVM | 0.067±0.019 | 0.594±0.090 | 0.873±0.032 | 0.812±0.129 | 0.468±0.048 | 0.190±0.028 | 0.223±0.035 |
| PL-SVM-DELIN | 0.481±0.052 | 0.378±0.188 | 0.831±0.053 | 0.863±0.114 | 0.490±0.049 | 0.312±0.023 | 0.231±0.026 |
| PL-SVM-CENDA | 0.491±0.045 | 0.216±0.035 | 0.879±0.070 | 0.931±0.048 | 0.525±0.069 | **0.371±0.025** | 0.353±0.020 |
| PL-SVM-PLDA | 0.492±0.028 | 0.154±0.116 | 0.829±0.067 | 0.861±0.097 | 0.497±0.050 | 0.332±0.021 | 0.226±0.012 |
| PL-SVM-WPLDR | **0.517±0.031** | **0.604±0.041** | **0.899±0.053** | **0.931±0.067** | **0.731±0.250** | 0.348±0.029 | **0.366±0.042** |
| IPAL | 0.099±0.019 | 0.553±0.036 | 0.905±0.039 | 0.752±0.136 | 0.430±0.037 | 0.327±0.013 | 0.311±0.037 |
| IPAL-DELIN | 0.484±0.050 | 0.486±0.057 | 0.905±0.046 | 0.715±0.160 | 0.444±0.051 | 0.331±0.021 | 0.271±0.024 |
| IPAL-CENDA | 0.491±0.047 | 0.258±0.048 | 0.897±0.040 | 0.734±0.161 | 0.409±0.037 | 0.321±0.020 | 0.332±0.021 |
| IPAL-PLDA | 0.503±0.028 | 0.276±0.041 | 0.911±0.046 | 0.716±0.200 | 0.446±0.055 | 0.339±0.037 | 0.281±0.036 |
| IPAL-WPLDR | **0.519±0.031** | **0.581±0.039** | **0.929±0.049** | **0.753±0.164** | **0.457±0.047** | **0.346±0.026** | **0.366±0.042** |
| SURE | 0.109±0.023 | **0.674±0.042** | 0.929±0.028 | 0.921±0.079 | 0.599±0.051 | 0.208±0.023 | 0.289±0.033 |
| SURE-DELIN | 0.486±0.049 | 0.509±0.053 | 0.929±0.045 | 0.931±0.082 | 0.595±0.045 | 0.330±0.020 | 0.327±0.029 |
| SURE-CENDA | 0.491±0.045 | 0.256±0.039 | 0.932±0.046 | 0.941±0.051 | **0.603±0.046** | **0.366±0.024** | 0.339±0.018 |
| SURE-PLDA | 0.503±0.028 | 0.352±0.035 | 0.932±0.040 | 0.941±0.051 | 0.350±0.020 | 0.351±0.030 | 0.352±0.028 |
| SURE-WPLDR | **0.521±0.030** | 0.603±0.041 | **0.944±0.045** | **0.951±0.052** | 0.600±0.034 | 0.348±0.028 | **0.366±0.042** |
| PL-AGGD | 0.131±0.024 | **0.651±0.047** | 0.938±0.029 | 0.931±0.067 | 0.592±0.047 | 0.220±0.026 | 0.286±0.032 |
| PL-AGGD-DELIN | 0.485±0.051 | 0.510±0.054 | 0.923±0.045 | 0.931±0.067 | 0.597±0.049 | 0.330±0.020 | 0.327±0.029 |
| PL-AGGD-CENDA | 0.492±0.046 | 0.343±0.049 | 0.923±0.049 | 0.951±0.052 | 0.599±0.043 | 0.345±0.020 | 0.396±0.026 |
| PL-AGGD-PLDA | 0.503±0.027 | 0.380±0.046 | 0.932±0.050 | 0.941±0.051 | 0.584±0.051 | **0.358±0.023** | 0.352±0.028 |
| PL-AGGD-WPLDR | **0.521±0.030** | 0.605±0.040 | **0.944±0.043** | **0.951±0.052** | **0.600±0.040** | 0.350±0.028 | **0.366±0.042** |

Table 2: Classification accuracy (mean ± std) of comparing algorithms on controlled synthetic data sets with two false positive labels($r = 2$).

CENDA, PLDA and WPLDR. According to the reported results on these real-world data sets, the following observations can be concluded:

- Compared with partial label learning algorithms $\mathcal{L}$, across 35 statistical comparisons (7 data sets × 5 algorithms), the prediction accuracy has been significantly improved by employing WPLDR in 26 cases in pairwise $t$-test at 0.05 significance level. Furthermore, FG-NET is a challenging data set since it holds least number of examples but the second largest average number in candidate labels. The classification improvement by WPLDR indicates that WPLDR can significantly improve the performance in difficult circumstance with insufficient examples and high rate of false positive labels.
- Compared with DELIN, $\mathcal{L}$-WPLDR achieves superior or at least statistically comparable performance against $\mathcal{L}$-DELIN across 31 cases out of 35 cases.
- Compared with CENDA, $\mathcal{L}$-WPLDR achieves comparable or better performance in all cases, furthermore, among 35 cases, $\mathcal{L}$-WPLDR achieves significant performance improvement in 22 cases in pairwise $t$-test at 0.05 significance level.

### 4.3 Synthetic Data Sets

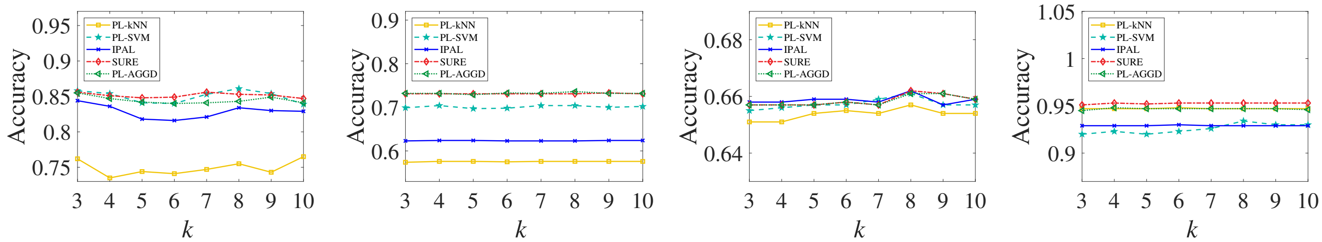
Following the widely used controlling protocal in partial label learning, synthetic partial label data sets are generated from UCI multi-class data sets with controlling parameter $r$, which indicates the number of false positive labels added in candidate label set.

For each synthetic data set, we set $r$ as $\{1, 2, 3\}$ to evaluate the performance under different ambiguity levels.The detailed experimental results with $r = 2$ are reported in Table 2. In addition, the pairwise $t$-test at 0.05 significance level is conducted to show whether the performance difference between two comparison methods is significant in statistics, and the results of win/tie/loss counts with $r = 1/2/3$ are reported in Table 3. Based on these comparative results, the following observations can be concluded:

- Compared with partial label learning algorithms $\mathcal{L}$, across the 105 statistical comparison cases (7 synthetic data sets × 3 configurations × 5 algorithms), the proposed WPLDR achieves superior or comparable classification performance in 98 cases.
- Compared with the existing partial label dimensionality reduction method DELIN, WPLDR achieves comparable or better performance in all cases, while it achieves significant performance improvement in 97 cases in pairwise $t$-test at 0.05 significance level.
- Compared with the existing PLDA, WPLDR achieves comparable or better performance in 95 cases, while the improvement is more impressive in most cases.
- Compared with the existing CENDA, among all the 105 cases, $\mathcal{L}$-WPLDR achieves comparable or better classification performance in 95 cases.
- For high dimensional dataset amazon, where the dimen-

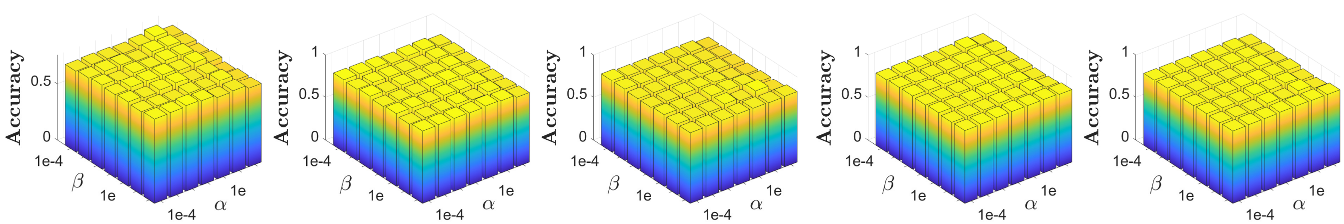
| **Data Set** | $\mathcal{L}$-WPLDR against $\mathcal{L}$-DELIN | | | | | $\mathcal{L}$-WPLDR against $\mathcal{L}$-CENDA | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | $\mathcal{L}$=PL-KNN | $\mathcal{L}$= PL-SVM | $\mathcal{L}$=IPAL | $\mathcal{L}$=SURE | $\mathcal{L}$=PL-AGGD | $\mathcal{L}$= PL-KNN | $\mathcal{L}$= PL-SVM | $\mathcal{L}$=IPAL | $\mathcal{L}$=SURE | $\mathcal{L}$=PL-AGGD |
| $r = 1$ | 6/1/0 | 6/1/0 | 7/0/0 | 6/1/0 | 7/0/0 | 6/0/1 | 2/3/2 | 4/3/0 | 3/3/1 | 3/3/1 |
| $r = 2$ | 6/1/0 | 6/1/0 | 7/0/0 | 6/1/0 | 6/1/0 | 5/2/0 | 4/1/2 | 7/0/0 | 5/1/1 | 3/3/1 |
| $r = 3$ | 7/0/0 | 7/0/0 | 7/0/0 | 6/1/0 | 7/0/0 | 7/0/0 | 5/2/0 | 6/1/0 | 6/1/0 | 5/1/1 |
| **Real-world** | 3/2/2 | 7/0/0 | 7/0/0 | 5/1/1 | 4/2/1 | 5/2/0 | 5/2/0 | 4/3/0 | 4/3/0 | 4/2/1 |
| In Total | **22/4/2** | **26/2/0** | **28/0/0** | **23/4/1** | **24/3/1** | **23/4/1** | **16/8/4** | **21/7/0** | **18/8/2** | **15/9/4** |
| **Data Set** | $\mathcal{L}$-WPLDR against $\mathcal{L}$ | | | | | $\mathcal{L}$-WPLDR against $\mathcal{L}$-PLDA | | | | |
| | $\mathcal{L}$=PL-KNN | $\mathcal{L}$= PL-SVM | $\mathcal{L}$=IPAL | $\mathcal{L}$=SURE | $\mathcal{L}$=PL-AGGD | $\mathcal{L}$= PL-KNN | $\mathcal{L}$= PL-SVM | $\mathcal{L}$=IPAL | $\mathcal{L}$=SURE | $\mathcal{L}$=PL-AGGD |
| $r = 1$ | 7/0/0 | 6/1/0 | 5/2/0 | 5/2/0 | 5/2/0 | 5/1/1 | 5/1/1 | 5/1/1 | 3/3/1 | 2/4/1 |
| $r = 2$ | 7/0/0 | 7/0/0 | 6/1/0 | 5/1/1 | 4/2/1 | 6/1/0 | 6/0/1 | 6/1/0 | 6/1/0 | 6/1/0 |
| $r = 3$ | 7/0/0 | 6/0/1 | 5/1/1 | 5/1/1 | 5/0/2 | 5/1/1 | 7/0/0 | 6/1/0 | 5/1/1 | 4/1/2 |
| **Real-world** | 5/2/0 | 6/1/0 | 5/1/1 | 5/2/0 | 4/2/1 | 2/4/2 | 6/0/1 | 6/0/1 | 6/0/1 | 3/2/2 |
| In Total | **26/2/0** | **25/2/1** | **21/5/2** | **20/6/2** | **18/6/4** | **18/7/4** | **24/1/3** | **23/3/2** | **20/5/3** | **15/8/5** |

Table 3: Win/tie/loss counts (pairwise $t$-test at 0.05 significance level) of $\mathcal{L}$-WPLDR against $\mathcal{L}$-DELIN, $\mathcal{L}$-CENDA and and $\mathcal{L}$-PLDA under different configurations on the number of false positive labels ($r = 1, 2, 3$).




(a) Lost (b) Italian (c) Amazon ($r = 1$) (d) Dermatology($r = 1$)

Figure 1: Parameter sensitivity analysis for $\mathcal{L}$-WPLDR, classification accuracy changes as $k$ on real-world and synthetic partial label datasets.




(a) PL-KNN (b) PL-SVM (c) IPAL (d) SURE (e) PL-AGGD

Figure 2: Parameter sensitivity analysis of varying $\alpha$ and $\beta$ for $\mathcal{L}$-WPLDR on `Lost`.

sion of feature vector exceeds 1,300, compared with $\mathcal{L}$, the classification performance has been improved with WPLDR by more than **0.3** in 14 cases among 15 cases (3 configurations × 5 algorithms). These results indicate the superior performance of WPLDR in difficult settings.

### 4.4 Sensitivity Analysis

For WPLDR, $k$ (the number of nearest neighbors) is an important paramter. Fig. 1 illustrates how the classification accuracy of each partial label learning algorithm changes as $k$ increases from 3 to 10 with interval 1. As is shown, on these four datasets, the classification accuracy of all partial label learning algorithms coupled with WPLDR is very stable cross different settings of $k$. Furthermore, the trade-off factors $\alpha$ and $\beta$ serve as important parameters. In Fig. 2, the values of $\alpha$ and $\beta$ increase from 0.0001 to 100. As is shown, when coupling with WPLDR, classification accuracy of each partial label learning algorithm is relatively stable across different values of $\alpha$ and $\beta$. According to the empirical studies, we suggest the value of $\alpha$ and $\beta$ can be simply set as 0.01 and 0.01 in practice.

## 5 Conclusion

In this paper, we propose a wrapped partial label dimensionality reduction approach, which is the fist attempt towards integrating dimensionality reduction and partial label disambiguation in one stage. To achieve this, WPLDR maximizes the interdependence between the embedded feature space and confidence-based label information, while ensures the manifold consistency between the embedded feature space and label space. Extensive experiments over a broad range of synthetic and real-world partial label data sets validate that WPLDR can significantly enhance the generalization performance of well-established partial label learning algorithms. In future work, we will further investigate how to extend WPLDR to other weakly-supervised learning frameworks.

## Acknowledgements

The authors wish to thank the anonymous reviewers for their helpful comments and suggestions. This work was supported by the National Science Foundation of China (62225602) and the Big Data Computing Center of Southeast University.

## References

[Bao *et al.*, 2021] Wei-Xuan Bao, Jun-Yi Hang, and Min-Ling Zhang. Partial label dimensionality reduction via confidence-based dependence maximization. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pages 46–54, 2021.

[Belkin and Niyogi, 2006] Mikhail Belkin and Partha Niyogi. Convergence of laplacian eigenmaps. *Advances in Neural Information Processing Systems*, 19, 2006.

[Chen *et al.*, 2017] Ching-Hui Chen, Vishal M Patel, and Rama Chellappa. Learning from ambiguously labeled face images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(7):1653–1667, 2017.

[Cour *et al.*, 2011] Timothee Cour, Ben Sapp, and Ben Taskar. Learning from partial labels. *The Journal of Machine Learning Research*, 12:1501–1536, 2011.

[Dempster *et al.*, 1977] Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: series B (methodological)*, 39(1):1–22, 1977.

[Feng and An, 2019] Lei Feng and Bo An. Partial label learning with self-guided retraining. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 3542–3549, 2019.

[Gong *et al.*, 2021a] Xiuwen Gong, Jiahui Yang, Dong Yuan, and Wei Bao. Generalized large margin $k$ nn for partial label learning. *IEEE Transactions on Multimedia*, 24:1055–1066, 2021.

[Gong *et al.*, 2021b] Xiuwen Gong, Dong Yuan, and Wei Bao. Discriminative metric learning for partial label learning. *IEEE Transactions on Neural Networks and Learning Systems*, 2021.

[Gong *et al.*, 2022] Xiuwen Gong, Dong Yuan, and Wei Bao. Partial label learning via label influence function. In *Proceedings of the 39th International Conference on Machine Learning*, pages 7665–7678, 2022.

[Hasan and Abdulazeez, 2021] Basna Mohammed Salih Hasan and Adnan Mohsin Abdulazeez. A review of principal component analysis algorithm for dimensionality reduction. *Journal of Soft Computing and Data Mining*, 2(1):20–30, 2021.

[He *et al.*, 2022] Shuo He, Lei Feng, Fengmao Lv, Wen Li, and Guowu Yang. Partial label learning with semantic label representations. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 545–553, 2022.

[Huang *et al.*, 2019] Xuan Huang, Lei Wu, and Yinsong Ye. A review on dimensionality reduction techniques. *International Journal of Pattern Recognition and Artificial Intelligence*, 33(10):1950017, 2019.

[Hüllermeier and Beringer, 2006] Eyke Hüllermeier and Jürgen Beringer. Learning from ambiguously labeled examples. *Intelligent Data Analysis*, 10(5):419–439, 2006.

[Jia *et al.*, 2022] Weikuan Jia, Meili Sun, Jian Lian, and Sujuan Hou. Feature dimensionality reduction: a review. *Complex & Intelligent Systems*, 8(3):2663–2693, 2022.

[Jia *et al.*, 2023] Yuheng Jia, Jiahao Jiang, and Yongheng Wang. Semantic dissimilarity guided locality preserving projections for partial label dimensionality reduction. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 964–973, 2023.

[Jia *et al.*, 2024] Yuheng Jia, Fuchao Yang, and Yongqiang Dong. Partial label learning with dissimilarity propagation guided candidate label shrinkage. *Advances in Neural Information Processing Systems*, 36, 2024.

[Jin and Ghahramani, 2002] Rong Jin and Zoubin Ghahramani. Learning with multiple labels. *Advances in Neural Information Processing Systems*, 15, 2002.

[Lin *et al.*, 2022] Guang-Yi Lin, Zi-Yang Xiao, Jia-Tong Liu, Bei-Zhan Wang, Kun-Hong Liu, and Qing-Qiang Wu. Feature space and label space selection based on error-correcting output codes for partial label learning. *Information Sciences*, 589:341–359, 2022.

[Liu and Dietterich, 2012] Liping Liu and Thomas Dietterich. A conditional multinomial mixture model for superset label learning. *Advances in Neural Information Processing Systems*, 25, 2012.

[Lyu *et al.*, 2020a] Gengyu Lyu, Songhe Feng, Wenying Huang, Guojun Dai, Hua Zhang, and Baifan Chen. Partial label learning via low-rank representation and label propagation. *Soft Computing*, 24:5165–5176, 2020.

[Lyu *et al.*, 2020b] Gengyu Lyu, Songhe Feng, Tao Wang, and Congyan Lang. A self-paced regularization framework for partial-label learning. *IEEE Transactions on Cybernetics*, 52(2):899–911, 2020.

[Nguyen and Caruana, 2008] Nam Nguyen and Rich Caruana. Classification with partial labels. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 551–559, 2008.

[Ni *et al.*, 2021] Peng Ni, Su-Yun Zhao, Zhi-Gang Dai, Hong Chen, and Cui-Ping Li. Partial label learning via conditional-label-aware disambiguation. *Journal of Computer Science and Technology*, 36(3):590–605, 2021.

[Niu *et al.*, 2023] Fan Niu, Xiaowei Zhao, Jun Guo, Mei Shi, Xiaoxia Liu, and Baoying Liu. Fast and robust unsupervised dimensionality reduction with adaptive bipartite graphs. *Knowledge-Based Systems*, 276:110680, 2023.

[Ren *et al.*, 2024] Lijuan Ren, Liangxiao Jiang, Wenjun Zhang, and Chaoqun Li. Label distribution similarity-based noise correction for crowdsourcing. *Frontiers of Computer Science*, 18(5):185323, 2024.

[Roweis and Saul, 2000] Sam T Roweis and Lawrence K Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500):2323–2326, 2000.

[Sachin and others, 2015] Deshmukh Sachin et al. Dimensionality reduction and classification through pca and lda. *International Journal of Computer Applications*, 122(17), 2015.

[Song *et al.*, 2022] Zixing Song, Xiangli Yang, Zenglin Xu, and Irwin King. Graph-based semi-supervised learning: A comprehensive review. *IEEE Transactions on Neural Networks and Learning Systems*, 2022.

[Tang and Zhang, 2017] Cai-Zhi Tang and Min-Ling Zhang. Confidence-rated discriminative partial label learning. In *Proceedings of the 31st AAAI Conference on Artificial Intelligence*, volume 31, 2017.

[Tenenbaum *et al.*, 2000] Joshua B Tenenbaum, Vin de Silva, and John C Langford. A global geometric framework for nonlinear dimensionality reduction. *science*, 290(5500):2319–2323, 2000.

[Tian *et al.*, 2023] Yingjie Tian, Xiaotong Yu, and Saiji Fu. Partial label learning: Taxonomy, analysis and outlook. *Neural Networks*, 2023.

[Vogelstein *et al.*, 2021] Joshua T Vogelstein, Eric W Bridgeford, Minh Tang, Da Zheng, Christopher Douville, Randal Burns, and Mauro Maggioni. Supervised dimensionality reduction for big data. *Nature communications*, 12(1):2872, 2021.

[Wang and Zhang, 2022] Wei Wang and Min-Ling Zhang. Partial label learning with discrimination augmentation. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 1920–1928, 2022.

[Wang *et al.*, 2021] Deng-Bao Wang, Min-Ling Zhang, and Li Li. Adaptive graph guided disambiguation for partial label learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(12):8796–8811, 2021.

[Wang *et al.*, 2023] Kui-Kui Wang, Gong-Ping Yang, Lu Yang, Yu-Wen Huang, and Yi-Long Yin. Ecg biometrics via enhanced correlation and semantic-rich embedding. *Machine Intelligence Research*, 20(5):697–706, 2023.

[Wu and Zhang, 2018] Xuan Wu and Min-Ling Zhang. Towards enabling binary decomposition for partial label learning. In *IJCAI*, pages 2868–2874, 2018.

[Xie *et al.*, 2021] Ming-Kun Xie, Feng Sun, and Sheng-Jun Huang. Partial multi-label learning with meta disambiguation. In *Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining*, pages 1904–1912, 2021.

[Ximendes *et al.*, 2022] Erving Ximendes, Riccardo Marin, Luis Dias Carlos, and Daniel Jaque. Less is more: dimensionality reduction as a general strategy for more precise luminescence thermometry. *Light: Science & Applications*, 11(1):237, 2022.

[Xu *et al.*, 2019] Ning Xu, Jiaqi Lv, and Xin Geng. Partial label learning via label enhancement. In *Proceedings of the 33rd AAAI Conference on Artificial Intelligence*, volume 33, pages 5557–5564, 2019.

[Xu *et al.*, 2021] Ning Xu, Congyu Qiao, Xin Geng, and Min-Ling Zhang. Instance-dependent partial label learning. *Advances in Neural Information Processing Systems*, 34:27119–27130, 2021.

[Xu *et al.*, 2023] Ning Xu, Biao Liu, Jiaqi Lv, Congyu Qiao, and Xin Geng. Progressive purification for instance-dependent partial label learning. In *Proceedings of the 40th International Conference on Machine Learning*, pages 38551–38565, 2023.

[Yao *et al.*, 2023] Yuqin Yao, Hua Meng, Yang Gao, Zhiguo Long, and Tianrui Li. Linear dimensionality reduction method based on topological properties. *Information Sciences*, 624:493–511, 2023.

[Yu and Zhang, 2016] Fei Yu and Min-Ling Zhang. Maximum margin partial label learning. In *Proceedings of the 7th Asian Conference on Machine Learning*, pages 96–111. PMLR, 2016.

[Yu *et al.*, 2024] Xiang-Ru Yu, Deng-Bao Wang, and Min-Ling Zhang. Dimensionality reduction for partial label learning: A unified and adaptive approach. *IEEE Transactions on Knowledge and Data Engineering*, 2024.

[Zhang and Yu, 2015] Min-Ling Zhang and Fei Yu. Solving the partial label learning problem: An instance-based approach. In *Proceedings of the 24th International Joint Conference on Artificial Intelligence*, pages 4048–4054, 2015.

[Zhang *et al.*, 2016] Min-Ling Zhang, Bin-Bin Zhou, and Xu-Ying Liu. Partial label learning via feature-aware disambiguation. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1335–1344, 2016.

[Zhang *et al.*, 2017] Min-Ling Zhang, Fei Yu, and Cai-Zhi Tang. Disambiguation-free partial label learning. *IEEE Transactions on Knowledge and Data Engineering*, 29(10):2155–2167, 2017.

[Zhang *et al.*, 2022] Min-Ling Zhang, Jing-Han Wu, and Wei-Xuan Bao. Disambiguation enabled linear discriminant analysis for partial label dimensionality reduction. *ACM Transactions on Knowledge Discovery from Data*, 16(4):1–18, 2022.

[Zhao *et al.*, 2023] Yongwei Zhao, Zidong Du, Qi Guo, Zhiwei Xu, and Yunji Chen. Rescue to the curse of universality. *Science China Information Sciences*, 66(9):192102, 2023.

[Zhou *et al.*, 2018] De-Yu Zhou, Zhi-Kai Zhang, Min-Ling Zhang, and Yu-Lan He. Weakly supervised pos tagging without disambiguation. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 17(4):1–19, 2018.