Pattern Recognition 131 (2022) 108839

Contents lists available at ScienceDirect

Pattern Recognition

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

Semi-supervised partial multi-label classification via consistency learning

Anhui Tan[a,b], Jiye Liang[a,*], Wei-Zhi Wu[b,c], Jia Zhang[d]

[a] Key Laboratory of Computational Intelligence and Chinese Information Processing of Ministry of Education, School of Computer and Information Technology, Shanxi University, Taiyuan 030006, China
[b] School of Information Engineering, Zhejiang Ocean University, Zhoushan, Zhejiang 316022, PR China
[c] Key Laboratory of Oceanographic Big Data Mining and Application of Zhejiang Province, Zhoushan, Zhejiang 316022, PR China
[d] College of Information Science and Technology, Jinan University, Guangzhou, Guangdong 510632, PR China

ARTICLE INFO

Article history:
Received 23 October 2021
Revised 10 May 2022
Accepted 7 June 2022
Available online 9 June 2022

Keywords:
Semi-supervised partial multi-label learning
Label correlation
HSIC

ABSTRACT

Partial multi-label learning refers to the problem that each instance is associated with a candidate label set involving both relevant and noisy labels. Existing solutions mainly focus on label disambiguation, while ignoring the negative effect of the inconsistency between feature information and label information. Specifically, the existence of completely unlabeled instances makes the estimation of label co-occurrence difficult. To tackle these problems, we propose a novel framework for partial multi-label learning in semi-supervised scenarios by solving the inconsistency between features and labels. In the first stage, the label-level correlation matrix on both labeled and unlabeled instances is derived via Hilbert-Schmidt Independence Criterion (HSIC). The correlation matrix can characterize the label correlation of labeled instances and can propagate the label correlation of unlabeled instances. In the second stage, the proposed framework achieves the training of feature mapping, the recovery of ground-truth labels, and the alleviation of noisy labels in a mutually beneficial manner, and develops an alternative optimization procedure to optimize them. In addition, a nonlinear version is extended by using kernel trick. Experimental studies demonstrate that the proposed methods can achieve competitive superiority against existing well-established methods.

1. Introduction

In recent years, multi-label learning has received increasing research and application in various real-word domains such as image annotation, text categorization and medical diagnosis [1,2]. In multi-label learning, each instance is associated with a set of labels simultaneously, e.g., a landscape may be crowdly-tagged by annotators with *street*, *sea*, *building*, *cloud*, and *tree*. Conventional multi-label learning assumes that the labeling information of each instance is complete, i.e., each instance has been precisely annotated with all relevant labels. However, in many real-world scenarios, this assumption may be infeasible and relevant labels of instances are usually partially correct and noise-corrupted. In order to combat with those problems, a weakly-supervised learning framework called partial multi-label learning [3] has been formalized whose purpose is to train a classification model from partially labeled multi-label data that is able to predict relevant labels for unseen instances as accurately as possible.

In reality, it is typical that the notations of training examples may be incomplete because the acquisition of labels suffers from various difficulties such as the costs of times and labors. Such scenario is realized as a weak multi-label learning problem with missing labels [4,5]. Especially, annotators may give relevant labels for a few training examples. In this case, labeled instances are partially available and many training instances are completely unlabeled, which is formalized as semi-supervised multi-label learning [6,7]. In recent years, partial multi-label learning and semi-supervised multi-label learning are combined to generate a new generalized framework, i.e., semi-supervised partial multi-label learning [8], where a part of training instances are associated with candidate noise-corrupted label sets and the rest are completely unlabeled. In general, the disambiguation of labels for partially labeled instances and the prediction of unlabeled instances are crucial to the success of semi-supervised partial multi-label learning tasks.

The motivations of this work are mainly established on the following considerations. As labels are usually co-related by semantic meanings [5], the strategies considering the manifold structures of instances and the label correlation can promote the label prediction accuracy. Most related studies assume the smoothness of la-

[*] Corresponding author.
E-mail addresses: tananhui86@163.com (A. Tan), ljy@sxu.edu.cn (J. Liang), wuwz@zjou.edu.cn (W.-Z. Wu), jiazhang@jnu.edu.cn (J. Zhang).

https://doi.org/10.1016/j.patcog.2022.108839
0031-3203/© 2022 Elsevier Ltd. All rights reserved.

bel co-occurrence of neighbor instances and neighbor categories. It should be noticed that those studies usually leverage the instance distribution in feature-level rather than in label-level for the inference of latent label co-occurrence. Nevertheless, there is always strong inconsistency and bias of the instance correlations between the feature-level and the label-level. The smoothness assumption of feature-level label correlation may give rise to structural bias and hinder the multi-label learning performance. To this end, a class of pertinent strategy is the dimensionality reduction that aims to extract the label-specific feature information [9–11]. However, many dimensionality reduction approaches suffer from insufficient and inaccurate labels, and still have limitations to obtain the label-level instance correlation information. Another is the embedding strategy that pursues the latent shared representation of features and labels so as to capture the consistent structural information of data. For example, Lv et al. [12] designed a compact embedding approach to combine the feature space and the label space as mutual guidance. Sun et al. [13] proposed a feature subspace representation and label disambiguation approach for partial multi-label learning. Skrlj et al. [14] explored the manifold-based embedding of the feature and the label spaces using the Relief algorithms where a given embedding dimensionality is intrinsic to the dimensionality of the dataset. Liu et al. [15] proposed solving multi-label classification through multi-output residual embedding that can learn a more low-rank representation by analyzing the residual structures of the feature and label spaces. However, it is hard to find an appropriate shared embedding space by the influence of the structural differences of features and labels. Moreover, the existence of missing labels makes the disclosure of the latent structure and the concealed label-level instance correlation more difficult, especially in the semi-supervised setting where the entire labeling information of many instances could be missing. To cope with this challenge, this paper leverages the HSIC [16] to yield the overall instance correlation on both labeled and unlabeled instances. The basic idea is to achieve the consistency between the feature information and the label information so as to realize the estimation of label-level instance correlation instead of feature-level instance correlation. Moreover, a semi-supervised partial multi-label learning method is proposed, which is expected to be more superior than many supervised methods only modeling labeled data because the structures of more data can be employed.

For a comprehensive considerations, the main contributions of this work include:

- The HSIC is utilized to disambiguate the inconsistency of instance correlations from feature-level and label-level so that the label-level instance correlation can be derived on labelled and unlabelled instances.
- A semi-supervised learning framework for partial multi-label learning is constructed in which the training of classifier, the recovery of ground-truth labels and the alleviation of noisy labels are implemented simultaneously in a mutually beneficial manner.
- We conduct extensive experiments on 16 real-world datasets to demonstrate the superiority of the proposed methods over the state-of-the-art algorithms.

The rest of this paper is organized as follows. Section 2 reviews related works on partial multi-label classification. Sections 3, 4 and 5 present details of the proposed approaches. Experimental results and analyses are reported in Section 6. Finally, the paper is concluded in Section 7.

2. Related work

Semi-supervised partial multi-label learning is a new emerging learning paradigm evolving from partial multi-label learning and semi-supervised multi-label learning. In this section, we overview the existing methods of semi-supervised multi-label learning and partial multi-label learning. In addition, we illustrate the features of the proposed work.

In previous studies of semi-supervised multi-label learning, the exploration of label co-occurrence is found to be effective for improving the tagging performance. Most related methods usually employ the intrinsic manifold structures by encoding the similarities of labeled and unlabeled data points. Existing semi-supervised learning methods mainly include generative approaches, co-training approaches, transductive margin machines, and etc [17]. For instance, Liu et al. [18] regarded the confidence score matrix of assigned class labels as a weight factor for spanning the instance correlation matrix. Chen et al. [6] constructed a transductive graph-based semi-supervised multi-label learning method that incorporated the regularization term of label consistency. Kong et al. [19] formulated a transductive multi-label learning algorithm to assign label sets to unlabeled instances via label set propagation. In many real world applications, the requirement that all labeled instances are available during training could not be satisfied. To adapt to this situation, Wu and Zhang [20] formulated an inductive semi-supervised multi-label learning via maximum margin assumption. To effectively utilize the information from both labeled and unlabeled data, Mikalsen et al. [9] introduced the HSIC measurement for semi-supervised classification which is applied well to multi-label dimensionality reduction. It should be pointed out that, in the work of [9], the label prediction is accomplished via label propagation based on the feature-level similarity of instances. Hence, the negative effect of the inconsistency between features and labels is ignored, which could bring prediction bias of label co-occurrence.

Partial multi-label learning is a new learning framework of inaccurate supervision in the presence that each instance is annotated with a noise-corrupted candidate label set [3]. A variety of partial multi-label learning methods have been proposed, which mainly focus on the recovery of ground-truth labeling information and the alleviation of noisy labeling information. A type of popular learning strategy is based on the disambiguation of labels that treats the ground-truth label matrix as a latent variable and refines the variable iteratively according to special criteria. For instance, Xie et al. [3] exploited the structural information from either the feature space or the label space and introduced two confidence label refinement algorithms. Sun et al. [21,22] utilized the property of low-rank and sparsity to decompose the ground-truth label matrix and the irrelevant label matrix. Note that two instances have a low feature similarity but with a high semantic similarity, their ground-truth labels can be overlapped to some extent. Yu et al. [23] reconstructed the label confidence matrix by collaboratively preserving both feature-level and label-level instance similarities simultaneously. According to the smoothness assumption that the feature and label spaces are prone to share the same topological structure, Wang et al. [24] presented a discriminative and correlative partial multi-label learning algorithm. To avoid misguiding by false positive labels concealed in the candidate label set, Zhang et al. [25] developed two credible label elicitation methods with virtual label splitting or maximum posteriori reasoning. More recently, Lyu [26] utilized the self-representation of instances and the prior knowledge of labels to learn the subspace representations of distinct instances and the high-order correlation of instances. Liu et al. [27] proposed a shared subspace learning framework to address the noisy labels and the missing views of multiview multi-label data. Without considering the structure of data, Xie et al. [28] tried to disambiguate the ground-truth and noisy labels in a meta-learning fashion. Xie et al. [8] formalized the semi-supervised partial multi-label learning problem by introducing the low-dimensional embedding from the feature space to the label space. The main idea of the work is to recover the label matrix by implementing the instance reconstruction via feature similarity.

However, the instance reconstruction is vulnerable with the change of the restored label matrix in the iterations and the negative effect of the inconsistency between features and labels is not addressed.

The aforementioned partial multi-label solutions assume that the instances having similar features will have similar labels, while ignoring the structural inconsistency between the feature space and the label space. Moreover, it is difficult to acquire the label-level correlation of instances in semi-supervised scenarios where many instances are absolutely unlabeled. The label-level correlation of labeled instances and the linkage between the feature and label information should be well accounted for effective learning the label-level similarities of unlabeled instances. Note that the HSIC is a non-parametric index that can well measure the independence of two data distributions and can learn a latent embedding representation of data by maximizing the independence. We employ HSIC to maximize the dependency between the feature and label information and to propagate the label-level correlation of labeled and unlabeled instances. Furthermore, we introduce a novel framework to implement the interaction of the label and feature information and to disambiguate the label matrix based on the propagation of the label-level instance correlation and the sparse and low-rank decomposition of the label matrix, which will be detailed in the following contents.

3. Problem statement and notations

In semi-supervised partial multi-label learning, denote that $\mathcal{D}_p = \{\mathbf{x}_i, \mathbf{y}_i\}_{i=1}^{p}$ is a set of p partially labeled instances, $\mathcal{D}_u = \{\mathbf{x}_i, \mathbf{y}_i\}_{i=p+1}^{p+u}$ is a set of u unlabeled instances, and $\mathcal{L} = \{\mathbf{l}_1, \cdots, \mathbf{l}_q\}$ is a set of labels. In the setting, $\mathbf{x}_i = (\mathbf{x}_{i1}, \cdots, \mathbf{x}_{id})^T$ represents the d-dimensional feature vector of the ith instance and $\mathbf{y}_i = (\mathbf{y}_{i1}, \cdots, \mathbf{y}_{iq})^T$ represents the q-dimensional binary label vector of the ith instance (T is the transpose operator).

Formally, denote $\mathbf{X}_p = [\mathbf{x}_1, \cdots, \mathbf{x}_p]^T \in \mathbb{R}^{p\times d}$ by the feature matrix of partially labeled instances and $\mathbf{X}_u = [\mathbf{x}_{p+1}, \cdots, \mathbf{x}_{p+u}]^T \in \mathbb{R}^{u\times d}$ by the feature matrix of unlabeled labeled instances. Correspondingly, denote $\mathbf{Y}_p = [\mathbf{y}_1, \cdots, \mathbf{y}_p]^T \in \{-1, 1\}^{p\times q}$ by the label matrix of partially labeled instances and $\mathbf{Y}_u = [\mathbf{y}_{p+1}, \cdots, \mathbf{y}_{p+u}]^T \in \{0\}^{u\times q}$ by the label matrix of unlabeled instances. In the label matrix, if the jth label $\mathbf{l}_j$ is a candidate label of the ith instance, then $\mathbf{Y}_{ij} = 1$, and $\mathbf{Y}_{ij} = -1$ otherwise. Especially, each entry of $\mathbf{Y}_u$ is 0 indicating that the labels are absolutely unknown for unlabeled instances. Moreover, let $n = p + u$ be the total number of instances, $\mathbf{X} = [\mathbf{x}_1, \cdots, \mathbf{x}_n]^T \in \mathbb{R}^{n\times d}$ and $\mathbf{Y} = [\mathbf{y}_1, \cdots, \mathbf{y}_n]^T \in \{-1, 0, 1\}^{n\times q}$ be the entire feature matrix and label matrix of instances, respectively.

3.1. Label-level instance correlation estimation

Existing methods typically assume that similar instances are more likely to share the same labels. However, there always exists inconsistency of instance distributions between the feature space and the label space. To this end, we acquire the label-level correlation of instances in semi-supervised scenarios via HSIC.

According to graph-based learning, a weight graph is appropriate to be constructed to measure the correlation among instances in the label space. Let $G = (V, E)$ be a connected graph corresponding to the n instances, where $V = \{1, \cdots, p, p+1 \cdots, n\}$ is the set of nodes corresponding to the p partially labeled instances and u unlabeled instances. Define an $n \times n$ symmetric instance weight matrix $\mathbf{S} \in \mathbb{R}^{n\times n}$ on the edges of the graph:

$$\mathbf{S} = \begin{bmatrix} \mathbf{P} & \mathbf{B} \\ \mathbf{B}^T & \mathbf{U} \end{bmatrix} \tag{1}$$

where $\mathbf{P} \in \mathbb{R}^{p\times p}$, $\mathbf{B} \in \mathbb{R}^{p\times u}$ and $\mathbf{U} \in \mathbb{R}^{u\times u}$ are the sub-matrices corresponding to the weights between pairwise labeled instances, the weights between labeled and unlabeled instances, and the weights between pairwise unlabeled instances, respectively.

We first define the cosine similarity matrix $\widetilde{\mathbf{P}}$ to measure the label-level similarity between labeled instances $\mathbf{x}_i, \mathbf{x}_j \in \mathcal{D}_p$ as:

$$\begin{aligned} \widetilde{\mathbf{P}}_{ij} &= \max\{0, cos(\mathbf{y}_i, \mathbf{y}_j)\} \\ &= \max\{0, \frac{<\mathbf{y}_i, \mathbf{y}_j>}{||\mathbf{y}_i||||\mathbf{y}_j||}\} \end{aligned} \tag{2}$$

We further re-scaled $\widetilde{\mathbf{P}}$ by normalizing it into $\mathbf{P} = \mathbf{D}^{-\frac{1}{2}}\widetilde{\mathbf{P}}\mathbf{D}^{-\frac{1}{2}}$. Here, $\mathbf{D} = \mathbf{diag}(d_{11}, \cdots, d_{nn})$ is a diagonal matrix satisfying $d_{ii} = \sum_{j=1}^{p} \widetilde{\mathbf{P}}_{ij}$.

Due to the existence of unlabeled instances, $\mathbf{B}$ and $\mathbf{U}$ are both unknown that need to learn via valid instance information. To this end, the HSIC is utilized to characterize the dependence between the instance correlations generated by the embedding space and by the partially known label space, which is defined as:

$$(n-1)^{-2} tr(\mathbf{HEHS}) \tag{3}$$

where $tr(\cdot)$ is the trace of a matrix, $\mathbf{E} = (\mathbf{XM})(\mathbf{XM})^T$, $\mathbf{M} \in \mathbb{R}^{d\times q}$ is the embedding matrix from the feature matrix to the label matrix, $\mathbf{H} = \mathbf{I}^{n\times n} - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^T$, $\mathbf{I}^{n\times n}$ is the $n \times n$ identity matrix, and $\mathbf{1}_n$ is the $n-$dimensional all-one vector. Here, $\mathbf{H}$ is used to align the variables $\mathbf{E}$ and $\mathbf{S}$ to both have zero means, and Eq. (3) aims to achieve a consistency of instance correlations from both feature-level and label-level. To extract more compact feature information related to label-level instance correlation, the basis of the embedding matrix is constrained to be orthonormal. The following optimization problem is induced:

$$\begin{aligned} \max_{\mathbf{M},\mathbf{B},\mathbf{U}} \quad & tr(\mathbf{HXMM}^T\mathbf{X}^T\mathbf{HS}) \\ s.t. \quad & \mathbf{M}^T\mathbf{M} = \mathbf{I}^{q\times q} \\ & \mathbf{S} = \begin{bmatrix} \mathbf{P} & \mathbf{B} \\ \mathbf{B}^T & \mathbf{U} \end{bmatrix} \end{aligned} \tag{4}$$

We can see from Eq. (5) that matrix $\mathbf{S}$ consists of four sub-matrices, within which $\mathbf{P}$ provides prior information for determining the unknown $\mathbf{B}$ and $\mathbf{U}$. To achieve the same scale, we assume that $\frac{1}{up}||\mathbf{B}||_F^2 \leq \frac{1}{p^2}||\mathbf{P}||_F^2$ and $\frac{1}{u^2}||\mathbf{U}||_F^2 \leq \frac{1}{p^2}||\mathbf{P}||_F^2$. Moreover, we always assume the nonnegativity of $\mathbf{S}$. Then, the following optimization problem is formulated:

$$\begin{aligned} \max_{\mathbf{M},\mathbf{B},\mathbf{U}} \quad & tr(\mathbf{HXMM}^T\mathbf{X}^T\mathbf{HS}) \\ s.t. \quad & \mathbf{M}^T\mathbf{M} = \mathbf{I}^{q\times q} \\ & \mathbf{S} = \begin{bmatrix} \mathbf{P} & \mathbf{B} \\ \mathbf{B}^T & \mathbf{U} \end{bmatrix} \\ & \tfrac{1}{up}||\mathbf{B}||_F^2 \leq \tfrac{1}{p^2}||\mathbf{P}||_F^2 \\ & \tfrac{1}{u^2}||\mathbf{U}||_F^2 \leq \tfrac{1}{p^2}||\mathbf{P}||_F^2 \\ & \mathbf{B} \geq 0 \\ & \mathbf{U} \geq 0 \end{aligned} \tag{5}$$

3.2. Solution of **M** *with other variables fixed*

When keeping the variables $\mathbf{B}$ and $\mathbf{U}$ fixed in Eq. (5), the optimization problem w.r.t. $\mathbf{M}$ is simplified as follows:

$$\begin{aligned} \max_{\mathbf{M}} \quad & tr(\mathbf{HXMM}^T\mathbf{X}^T\mathbf{HS}) \\ s.t. \quad & \mathbf{M}^T\mathbf{M} = \mathbf{I}^{q\times q} \end{aligned} \tag{6}$$

The optimal solution of $\mathbf{M}$ consists of q normalized eigenvectors corresponding to the top q largest eigenvalues of $\mathbf{X}^T\mathbf{HSHX}$.

3.3. Solution of **B** and **U** with other variables fixed

We first simplify Eq. (5) by separating the variables **B** and **U** from it.

Proposition 1. *The following equation holds.*

$$\begin{aligned} & tr(\mathbf{HXMM}^T\mathbf{X}^T\mathbf{HS}) \\ = & tr((\mathbf{X}_p\mathbf{M} - \frac{1}{n}\mathbf{1}_p\mathbf{1}_n^T\mathbf{XM})(\mathbf{X}_p\mathbf{M} - \frac{1}{n}\mathbf{1}_p\mathbf{1}_n^T\mathbf{XM})^T\mathbf{P}) \\ & +2tr((\mathbf{X}_u\mathbf{M} - \frac{1}{n}\mathbf{1}_u\mathbf{1}_n^T\mathbf{XM})(\mathbf{X}_p\mathbf{M} - \frac{1}{n}\mathbf{1}_p\mathbf{1}_n^T\mathbf{XM})^T\mathbf{B}) \\ & +tr((\mathbf{X}_u\mathbf{M} - \frac{1}{n}\mathbf{1}_u\mathbf{1}_n^T\mathbf{XM})(\mathbf{X}_u\mathbf{M} - \frac{1}{n}\mathbf{1}_u\mathbf{1}_n^T\mathbf{XM})^T\mathbf{U}) \end{aligned} \tag{7}$$

Proof. We have that $tr(\mathbf{HXMM}^T\mathbf{X}^T\mathbf{HS}) = \sum_{i,j=1}^{n}(\mathbf{X}_{i\cdot}\mathbf{M} - \frac{1}{n}\mathbf{1}_n^T\mathbf{XM})(\mathbf{X}_{j\cdot}\mathbf{M} - \frac{1}{n}\mathbf{1}_n^T\mathbf{XM})^T\mathbf{S}_{ij}$, where $\mathbf{X}_{i\cdot}$ is the ith row of **X**. Then the equation is not hard to prove. □

According to Eq. (7), the optimization subproblem w.r.t. **B** is simplified as follows:

$$\begin{aligned} \max_{\mathbf{B}} \; & tr((\mathbf{X}_u\mathbf{M} - \tfrac{1}{n}\mathbf{1}_u\mathbf{1}_n^T\mathbf{XM})(\mathbf{X}_p\mathbf{M} - \tfrac{1}{n}\mathbf{1}_p\mathbf{1}_n^T\mathbf{XM})^T\mathbf{B}) \\ s.t. \; & \tfrac{1}{up}||\mathbf{B}||_F^2 \le \tfrac{1}{p^2}||\mathbf{P}||_F^2 \\ & \mathbf{B} \ge 0 \end{aligned} \tag{8}$$

By solving Eq. (8), we obtain that $\mathbf{B} = \max\{\sqrt{\frac{u}{p}}\frac{||\mathbf{P}||_F}{||\mathbf{C}||_F}\mathbf{C}, \mathbf{0}\}$, where $\mathbf{C} = (\mathbf{X}_p\mathbf{M} - \frac{1}{n}\mathbf{1}_p\mathbf{1}_n^T\mathbf{XM})((\mathbf{X}_u\mathbf{M} - \frac{1}{n}\mathbf{1}_u\mathbf{1}_n^T\mathbf{XM}))^T$.

According to Eq. (7), the optimization subproblem imposed on **U** is considered by:

$$\begin{aligned} \max_{\mathbf{U}} \; & tr((\mathbf{X}_u\mathbf{M} - \tfrac{1}{n}\mathbf{1}_u\mathbf{1}_n^T\mathbf{XM})(\mathbf{X}_u\mathbf{M} - \tfrac{1}{n}\mathbf{1}_u\mathbf{1}_n^T\mathbf{XM})^T\mathbf{U}) \\ s.t. \; & \tfrac{1}{u^2}||\mathbf{U}||_F^2 \le \tfrac{1}{p^2}||\mathbf{P}||_F^2 \\ & \mathbf{U} \ge 0 \end{aligned} \tag{9}$$

By solving Eq. (9), we obtain that $\mathbf{U} = \max\{\frac{u}{p}\frac{||\mathbf{P}||_F}{||\mathbf{R}||_F}\mathbf{R}, \mathbf{0}\}$, where $\mathbf{R} = (\mathbf{X}_u\mathbf{M} - \frac{1}{n}\mathbf{1}_u\mathbf{1}_n^T\mathbf{XM})((\mathbf{X}_u\mathbf{M} - \frac{1}{n}\mathbf{1}_u\mathbf{1}_n^T\mathbf{XM}))^T$.

By alternatively solving Eq. (5), we can estimate the label-level instance correlation concealed in the semi-supervised multi-label data, while overcoming the inconsistency of the instance information between the feature space and the label space.

3.4. Instance correlation-based Laplacian manifold

To capture the label confidence of partially labeled instances and disclose the labels of unlabeled instances, we introduce a label vector $\mathbf{z}_i$ for measuring how likely each label is a ground-truth label of the instance $\mathbf{x}_i$. Suppose $\mathbf{Z} = [\mathbf{z}_1, \cdots, \mathbf{z}_n]^T = \begin{pmatrix}\mathbf{Z}_p \\ \mathbf{Z}_u\end{pmatrix} \in \{-1, 1\}^{n\times q}$ is the ground-truth label matrix, where $\mathbf{Z}_p \in \{-1, 1\}^{p\times q}$ and $\mathbf{Z}_u \in \{-1, 1\}^{u\times q}$ represent the label sub-matrices corresponding to the partially labeled instances and unlabeled instances, respectively.

The instance correlations on partially labeled and unlabeled instances may not be synchronous in scale to each other due to independence. Hence, the following regularization terms are proposed by separately considering the label-level instance correlations on partially labeled and unlabeled instances:

$$\begin{aligned} \frac{1}{2}\sum_{i,j=1}^{p}\mathbf{P}_{ij}||\mathbf{Z}_{pi\cdot} - \mathbf{Z}_{pj\cdot}||^2 &= tr(\mathbf{Z}_p^T\mathbf{L}_p\mathbf{Z}_p) \\ &= tr(\begin{pmatrix}\mathbf{Z}_p \\ \mathbf{Z}_u\end{pmatrix}^T\begin{pmatrix}\mathbf{L}_p & \mathbf{0} \\ \mathbf{0} & \mathbf{0}\end{pmatrix}\begin{pmatrix}\mathbf{Z}_p \\ \mathbf{Z}_u\end{pmatrix}) \\ &= tr(\mathbf{Z}^T\begin{pmatrix}\mathbf{L}_p & \mathbf{0} \\ \mathbf{0} & \mathbf{0}\end{pmatrix}\mathbf{Z}) \end{aligned} \tag{10}$$

where $\mathbf{Z}_{pi\cdot}$ is the ith row vector of $\mathbf{Z}_p$, $\mathbf{L}_p = \widehat{\mathbf{P}} - \mathbf{P}$ is the graph Laplacian matrix of **P**, and $\widehat{\mathbf{P}} \in \mathbb{R}^{p\times p}$ is a diagonal matrix such that $\widehat{\mathbf{P}}_{ii} = \sum_{j=1}^{p}\mathbf{P}_{ij}$.

The regularization term for label propagation from partially labeled to unlabeled instances is formulated as:

$$\begin{aligned} \sum_{\substack{j=u \\ i=p}}^{i,j=1}\mathbf{B}_{ij}\;||\mathbf{Z}_{pi\cdot} - \mathbf{Z}_{uj\cdot}||^2 &= \sum_{i=1}^{p}\sum_{j=1}^{u}\mathbf{B}_{ij}\mathbf{Z}_{pi\cdot}\mathbf{Z}_{pi\cdot}^T \\ & + \sum_{j=1}^{u}\sum_{i=1}^{p}\mathbf{B}_{ij}\mathbf{Z}_{uj\cdot}\mathbf{Z}_{uj\cdot}^T - 2\sum_{i=1}^{p}\sum_{j=1}^{u}\mathbf{B}_{ij}\mathbf{Z}_{pi\cdot}\mathbf{Z}_{uj\cdot}^T \\ &= tr(\mathbf{Z}_p^T\widehat{\mathbf{B}}_1\mathbf{Z}_p) + tr(\mathbf{Z}_u^T\widehat{\mathbf{B}}_2\mathbf{Z}_u) - 2tr(\mathbf{Z}_p^T\mathbf{BZ}_u) \end{aligned} \tag{11}$$

where $\widehat{\mathbf{B}}_1 \in \mathbb{R}^{p\times p}$ and $\widehat{\mathbf{B}}_2 \in \mathbb{R}^{u\times u}$ are diagonal matrices satisfying $\widehat{\mathbf{B}}_{1ii} = \sum_{j=1}^{u}\mathbf{B}_{ij}$ and $\widehat{\mathbf{B}}_{2jj} = \sum_{i=1}^{p}\mathbf{B}_{ij}$.

Eq. (11) can be rewritten as:

$$\begin{aligned} \sum_{\substack{j=u \\ i=p}}^{i,j=1}\mathbf{B}_{ij}||\mathbf{Z}_{pi\cdot} - \mathbf{Z}_{uj\cdot}||^2 &= tr(\begin{pmatrix}\mathbf{Z}_p \\ \mathbf{Z}_u\end{pmatrix}^T\begin{pmatrix}\widehat{\mathbf{B}}_1 & -\mathbf{B} \\ -\mathbf{B}^T & \widehat{\mathbf{B}}_2\end{pmatrix}\begin{pmatrix}\mathbf{Z}_p \\ \mathbf{Z}_u\end{pmatrix}) \\ &= tr(\mathbf{Z}^T\begin{pmatrix}\widehat{\mathbf{B}}_1 & -\mathbf{B} \\ -\mathbf{B}^T & \widehat{\mathbf{B}}_2\end{pmatrix}\mathbf{Z}) \end{aligned} \tag{12}$$

The regularization term on unlabeled instances is formulated as:

$$\begin{aligned} \frac{1}{2}\sum_{i,j=1}^{u}\mathbf{U}_{ij}||\mathbf{Z}_{ui\cdot} - \mathbf{Z}_{uj\cdot}||^2 &= tr(\mathbf{Z}_u^T\mathbf{L}_u\mathbf{Z}_u) \\ &= tr(\begin{pmatrix}\mathbf{Z}_p \\ \mathbf{Z}_u\end{pmatrix}^T\begin{pmatrix}\mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{L}_u\end{pmatrix}\begin{pmatrix}\mathbf{Z}_p \\ \mathbf{Z}_u\end{pmatrix}) \\ &= tr(\mathbf{Z}^T\begin{pmatrix}\mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{L}_u\end{pmatrix}\mathbf{Z}) \end{aligned} \tag{13}$$

where $\mathbf{L}_p = \widehat{\mathbf{P}} - \mathbf{P}$ is the graph Laplacian matrix of **P** and $\widehat{\mathbf{P}}$ is a diagonal matrix such that $\widehat{\mathbf{P}}_{ii} = \sum_{j=1}^{p}\mathbf{P}_{ij}$.

If we incorporate the regularization terms in Eqs. (10), (12) and (13), the label-level instance correlations encourage us to generate the following graph-based semi-supervised optimization term:

$$\min_{\mathbf{Z}} tr(\mathbf{Z}^T\mathbf{AZ}) \tag{14}$$

where

$$\mathbf{A} = \begin{pmatrix}\mathbf{L}_p & \mathbf{0} \\ \mathbf{0} & \mathbf{0}\end{pmatrix} + \gamma\begin{pmatrix}\mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{L}_u\end{pmatrix} + \gamma\begin{pmatrix}\widehat{\mathbf{B}}_1 & -\mathbf{B} \\ -\mathbf{B}^T & \widehat{\mathbf{B}}_2\end{pmatrix} \tag{15}$$

and $\gamma = \frac{p}{u}$ is a parameter to make a balance between different scales.

Proposition 2. *Matrix* **A** *in Eq. (15) is positive semidefinite for any* $\gamma \ge 0$.

Proof. This can be concluded from Eqs. (10), (12), and (13). □

3.5. Noise degradation and classifier training

To leverage the feature information for label prediction, a projection function $\mathbf{W}$ is introduced to map the feature matrix to the latent ground-truth label matrix $\mathbf{Z}$. On the other hand, due to the existence of noise, the original label matrix $\mathbf{Y}$ is decomposed as the sum of the ground-truth label matrix $\mathbf{Z}$ and the noisy label matrix $\mathbf{N}$ with $\mathbf{Z}$ constrained to be low-rank and $\mathbf{N}$ to be sparse. By incorporating these considerations, the following optimization problem is obtained:

$$\begin{aligned} \min_{\mathbf{W},\mathbf{b},\mathbf{Z},\mathbf{N}} \quad & ||\mathbf{X}\mathbf{W} + \mathbf{1}_n\mathbf{b}^T - \mathbf{Z}||_F^2 + \delta||\mathbf{W}||_F^2 + \eta||\mathbf{Z}||_* + \theta||\mathbf{N}||_1 \\ s.t. \quad & \mathbf{Z}_p = \mathbf{Y}_p - \mathbf{N}_p \end{aligned} \tag{16}$$

where $\mathbf{Z} = \begin{pmatrix} \mathbf{Z}_p \\ \mathbf{Z}_u \end{pmatrix}$, $\mathbf{N} = \begin{pmatrix} \mathbf{N}_p \\ \mathbf{0}_{u\times q} \end{pmatrix}$ is the noisy label matrix concealed in the partially labeled matrix, and δ, η and θ are trade-off parameters.

The constraint condition $\mathbf{Z}_p = \mathbf{Y}_p - \mathbf{N}_p$ can be approximately substituted by the minimization of the square loss $||\mathbf{Z}_p - \mathbf{Y}_p + \mathbf{N}_p||_F^2$. Moreover, if we incorporate the Laplacian regularization term Eq. (14) into Eq. (16), this encourages us to formulate the following optimization framework:

$$\begin{aligned} \min_{\mathbf{W},\mathbf{b},\mathbf{Z},\mathbf{N}} \quad & tr((\mathbf{Z} - \mathbf{Y} + \mathbf{N})^T\mathbf{K}(\mathbf{Z} - \mathbf{Y} + \mathbf{N})) + \alpha tr(\mathbf{Z}^T\mathbf{A}\mathbf{Z}) \\ & + \delta||\mathbf{W}||_F^2 + \eta||\mathbf{Z}||_* + \theta||\mathbf{N}||_1 + \lambda||\mathbf{X}\mathbf{W} + \mathbf{1}_n\mathbf{b}^T - \mathbf{Z}||_F^2 \end{aligned} \tag{17}$$

where $\mathbf{K} = \begin{pmatrix} \mathbf{I}_{p\times p} & \mathbf{0}_{p\times u} \\ \mathbf{0}_{u\times p} & \mathbf{0}_{u\times u} \end{pmatrix}$.

4. Solving the optimization problem

For solving Eq. (17), we first equivalently convert it into the following form:

$$\begin{aligned} \min_{\mathbf{W},\mathbf{b},\mathbf{Z},\mathbf{F},\mathbf{N}} \quad & tr((\mathbf{F} - \mathbf{Y} + \mathbf{N})^T\mathbf{K}(\mathbf{F} - \mathbf{Y} + \mathbf{N})) + \alpha tr(\mathbf{F}^T\mathbf{A}\mathbf{F}) \\ & + \delta||\mathbf{W}||_F^2 + \eta||\mathbf{Z}||_* + \theta||\mathbf{N}||_1 + \lambda||\mathbf{X}\mathbf{W} + \mathbf{1}_n\mathbf{b}^T - \mathbf{F}||_F^2 \\ s.t. \quad & \mathbf{Z} = \mathbf{F} \end{aligned} \tag{18}$$

According to the LADMAP methods [29], Eq. (18) can be rewritten as:

$$\begin{aligned} \min_{\mathbf{W},\mathbf{b},\mathbf{Z},\mathbf{F},\mathbf{N}} \quad & tr((\mathbf{F} - \mathbf{Y} + \mathbf{N})^T\mathbf{K}(\mathbf{F} - \mathbf{Y} + \mathbf{N})) + \alpha tr(\mathbf{F}^T\mathbf{A}\mathbf{F}) \\ & + \delta||\mathbf{W}||_F^2 + \eta||\mathbf{Z}||_* + \theta||\mathbf{N}||_1 \\ & + \lambda||\mathbf{X}\mathbf{W} + \mathbf{1}_n\mathbf{b}^T - \mathbf{F}||_F^2 + \tfrac{\mu}{2}||\mathbf{Z} - \mathbf{F} + \tfrac{\Phi}{\mu}||_F^2 \end{aligned} \tag{19}$$

where $\Phi \in \mathbb{R}^{n\times q}$ is the Lagrange multiplier matrix and μ is the Lagrange parameter.

The proposed objective function of Eq. (19) is convex w.r.t. each of the involved variables. Here, we introduce an alternating iterative algorithm to obtain the optimal solution, and present the detailed solving procedure in the next content.

4.1. Solving **W** and **b** with other variables fixed

By setting the derivative of Eq. (19) w.r.t. $\mathbf{b}$ to zero, the solution of $\mathbf{b}$ can be obtained as:

$$\mathbf{b} = \frac{1}{n}(\mathbf{F} - \mathbf{X}\mathbf{W})^T\mathbf{1}_n \tag{20}$$

Plugging Eq. (20) into Eq. (19), we can rewrite Eq. (19) as:

$$\begin{aligned} \min_{\mathbf{W},\mathbf{Z},\mathbf{F},\mathbf{N}} \quad & tr((\mathbf{F} - \mathbf{Y} + \mathbf{N})^T\mathbf{K}(\mathbf{F} - \mathbf{Y} + \mathbf{N})) + \alpha tr(\mathbf{F}^T\mathbf{A}\mathbf{F}) \\ & + \delta||\mathbf{W}||_F^2 + \eta||\mathbf{Z}||_* + \theta||\mathbf{N}||_1 \\ & + \lambda||\mathbf{H}\mathbf{X}\mathbf{W} - \mathbf{H}\mathbf{F}||_F^2 + \tfrac{\mu}{2}||\mathbf{Z} - \mathbf{F} + \tfrac{\Phi}{\mu}||_F^2 \end{aligned} \tag{21}$$

where $\mathbf{H} = \mathbf{I}^{n\times n} - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^T$.

We can examine that $\mathbf{H}^2 = \mathbf{H}$ and $\mathbf{H}^T = \mathbf{H}$. By setting the derivative of the objective function in Eq. (21) w.r.t. $\mathbf{W}$ to zero, we have:

$$\mathbf{W} = \Gamma\mathbf{F} \tag{22}$$

where $\Gamma = (\mathbf{X}^T\mathbf{H}\mathbf{X} + \frac{\delta}{\lambda}\mathbf{I})^{-1}\mathbf{X}^T\mathbf{H}$.

Plugging Eq. (22) into Eq. (21), we can rewrite Eq. (21) as:

$$\begin{aligned} \min_{\mathbf{Z},\mathbf{F},\mathbf{N}} \quad & tr((\mathbf{F} - \mathbf{Y} + \mathbf{N})^T\mathbf{K}(\mathbf{F} - \mathbf{Y} + \mathbf{N})) + \alpha tr(\mathbf{F}^T\mathbf{A}\mathbf{F}) \\ & + \delta||\Gamma\mathbf{F}||_F^2 + \eta||\mathbf{Z}||_* + \theta||\mathbf{N}||_1 \\ & + \lambda||\mathbf{H}(\mathbf{X}\Gamma - \mathbf{I})\mathbf{F}||_F^2 + \tfrac{\mu}{2}||\mathbf{Z} - \mathbf{F} + \tfrac{\Phi}{\mu}||_F^2 \end{aligned} \tag{23}$$

Ultimately, Eq. (23) contains three variables $\mathbf{N}$, $\mathbf{F}$ and $\mathbf{Z}$ that need to solve, which is computationally less expensive than the original version in Eq. (19).

4.2. Solving **N** with other variables fixed

When fixing other variables, the optimization of Eq. (23) over $\mathbf{N}$ is written as:

$$\min_{\mathbf{N}} tr((\mathbf{F} - \mathbf{Y} + \mathbf{N})^T\mathbf{K}(\mathbf{F} - \mathbf{Y} + \mathbf{N})) + \theta||\mathbf{N}||_1 \tag{24}$$

Since the noisy labeling information only conceals in the partially labeled matrix, it is enough to solve the sub-matrix $\mathbf{N}_p$, which is simplified as:

$$\min_{\mathbf{N}_p} ||\mathbf{F}_p - \mathbf{Y}_p + \mathbf{N}_p||_F^2 + \theta||\mathbf{N}_p||_1 \tag{25}$$

where $\mathbf{F}_p$ is the sub-matrix consisting of the first p rows of $\mathbf{F}$.

The optimizing rule for $\mathbf{N}_p$ can be calculated as:

$$\mathbf{N}_p = \mathcal{S}_{\frac{\theta}{2}}(\mathbf{Y}_p - \mathbf{F}_p) \tag{26}$$

where $\mathcal{S}$ is the element-wise shrinkage operator, which is defined as $\mathcal{S}_w(a) = (a - w)_+ - (-a - w)_+$ [30].

4.3. Solving **F** with other variables fixed

When fixing other variables, the optimization of Eq. (23) over $\mathbf{F}$ is written as:

$$\begin{aligned} \min_{\mathbf{F}} \quad & tr((\mathbf{F} - \mathbf{Y} + \mathbf{N})^T\mathbf{K}(\mathbf{F} - \mathbf{Y} + \mathbf{N})) + \alpha tr(\mathbf{F}^T\mathbf{A}\mathbf{F}) \\ & + \lambda||\mathbf{H}(\mathbf{X}\Gamma - \mathbf{I})\mathbf{F}||_F^2 + \tfrac{\mu}{2}||\mathbf{Z} - \mathbf{F} + \tfrac{\Phi}{\mu}||_F^2 \end{aligned} \tag{27}$$

By setting the derivative of the objective function in Eq. (27) w.r.t. $\mathbf{F}$ to zero, we have:

$$\mathbf{O}_1\mathbf{F} = \mathbf{O}_2 \tag{28}$$

where

$$\begin{aligned} \mathbf{O}_1 &= \mathbf{K} + \lambda(\mathbf{X}\Gamma - \mathbf{I})^T\mathbf{H}(\mathbf{X}\Gamma - \mathbf{I}) + \alpha\mathbf{A} + \delta\Gamma^T\Gamma + \tfrac{\mu}{2}\mathbf{I} \\ \mathbf{O}_2 &= \mathbf{K}(\mathbf{Y} - \mathbf{N}) + \tfrac{\mu}{2}\mathbf{Z} + \tfrac{1}{2}\Phi \end{aligned} \tag{29}$$

It is obvious that $\mathbf{O}_1$ is positive definite, whose inverse matrix is unique. Hence, Eq. (24) can be efficiently solved by current software libraries. Although the scale of $\mathbf{O}_1$ is $n \times n$, its inverse matrix is constant in the iteration and can be calculated only once before the iteration starts.

4.4. Solving **Z** with other variables fixed

Fixing other variables, the sub-problem in Eq. (23) over variable $\mathbf{Z}$ is reformulated as:

$$\min_{\mathbf{Z}} \frac{\mu}{2}||\mathbf{Z} - \mathbf{F} + \frac{\Phi}{\mu}||_F^2 + \eta||\mathbf{Z}||_* \tag{30}$$

The optimizing rules is given as:

$$\mathbf{Z} = \mathcal{T}_{\frac{\eta}{\mu}}(\mathbf{F} - \frac{\Phi}{\mu}) \tag{31}$$

Here, $\mathcal{T}$ is the single value thresholding operator [31], which first implements the singular value decomposition (SVD) of $\mathbf{F} - \frac{\Phi}{\mu}$ and then applies the soft thresholding on the singular values.

Finally, the Lagrange multiplier matrix Φ and the penalty parameter μ at the tth iteration are updated by the following rules:

$$\begin{aligned}\Phi^{t+1} &= \Phi^{t} + \mu^{t+1}(\mathbf{Z} - \mathbf{F}) \\ \mu^{t+1} &= \min(\mu_{\max}, \rho\mu^{t})\end{aligned} \tag{32}$$

where ρ is a positive scalar.

In brief, the solution is designed by executing the iterative process for unknown variables, and the pseudo code of the proposed semi-supervised partial multi-label learning algorithm for achieving consistency (CS2PML) is shown in Algorithm 1.

Algorithm 1 CS2PML Algorithm.

```
Require: ~~Feature matrix of partially labeled instances $\mathbf{X}_p \in \mathbb{R}^{p\times d}$,
    feature matrix of unlabeled labeled instances $\mathbf{X}_u \in \mathbb{R}^{u\times d}$, observed label
    matrix $\mathbf{Y}_p \in \mathbb{R}^{p\times q}$ of partially labeled instances, and parameters $\lambda, \alpha, \eta, \delta, \theta$;
Ensure: Learned label confidence matrix $\mathbf{Z} \in \mathbb{R}^{n\times q}$.
Set $\mathbf{X} = \begin{pmatrix}\mathbf{X}_p \\ \mathbf{X}_u\end{pmatrix}$ and $\mathbf{K} = \begin{pmatrix}\mathbf{I}_{p\times p} & \mathbf{0}_{p\times u} \\ \mathbf{0}_{u\times p} & \mathbf{0}_{u\times u}\end{pmatrix}$;
Initialize $\mathbf{F} = \mathbf{Z} = \begin{pmatrix}\mathbf{Y}_p \\ \mathbf{0}_{u\times q}\end{pmatrix}$, $\mathbf{N}_p = \mathbf{0}$ and $\mathbf{N} = \begin{pmatrix}\mathbf{N}_p \\ \mathbf{0}_{u\times q}\end{pmatrix}$;
Compute $\mathbf{P}$ by Eq. (4) and randomly initialize $\mathbf{B}$ and $\mathbf{U}$;
While not converged do
    ~~~~Compute $\mathbf{M}$ by solving Eq. (6);
    ~~~~Compute $\mathbf{B}$ and $\mathbf{U}$ by solving Eq.(8) and Eq. (8);
End While
Obtain the instance weight matrix $\mathbf{S} = \begin{pmatrix}\mathbf{P} & \mathbf{B} \\ \mathbf{B}^T & \mathbf{U}\end{pmatrix}$;
Compute $\mathbf{A}$ by Eq. (15);
Compute $\mathbf{H} = \mathbf{I}^{n\times n} - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^T$ and $\Gamma = (\mathbf{X}^T\mathbf{H}\mathbf{X} + \frac{\delta}{\lambda}\mathbf{I})^{-1}\mathbf{X}^T\mathbf{H}$;
While not converged do
    ~~~~ Update $\mathbf{N}_p$ by solving Eq. (26);
    ~~~~ Update $\mathbf{F}$ and $\mathbf{Z}$ based on Eqs. (28) and (31) by using the
        Lagrange multiplier method;
End While
Return the results.
```

4.5. Time complexity

The computations of $\mathbf{M}$, $\mathbf{B}$, and $\mathbf{U}$ can be respectively done within $\mathcal{O}(n^2d + d^3)$, $\mathcal{O}(n^2q)$, and $\mathcal{O}(n^2q)$, respectively. In the iteration of the proposed CS2PML, the main time-consuming steps include the matrix inverse operation when solving $\mathbf{F}$ and the SVD for solving $\mathbf{Z}$. In Eq. (24), the inverse matrix of $\mathbf{O}_1$ is constant with iteration and only needs to be calculated for one time. Hence, the computational complexity for solving $\mathbf{F}$ relies in the matrix multiplication of $\mathbf{O}_1^{-1}\mathbf{O}_2$, which can be done within $\mathcal{O}(n^2q)$, and the SVD in Eq. (31) that needs about $\mathcal{O}(nq^2)$. In summarize, the total complexity of CS2PML is no more than $\mathcal{O}((n^2d + d^3 + n^2q + nq^2)t)$, in which t is the iteration number.

5. Kernelization for CS2PML

We derive a nonlinear version of the proposed method by utilizing kernel tricks [32], denoted by CS2PML-n. Assume the projection matrix $\mathbf{W}$ can be spanned by kernel feature vectors, i.e. $\mathbf{W} = \Psi(\mathbf{X})\mathbf{T}$. In the setting, $\Psi(\mathbf{X}) = [\varphi(\mathbf{x}_1)\cdots\varphi(\mathbf{x}_n)]$, $\varphi : \mathbb{R}^d \to \mathbb{R}^N$ is the nonlinear mapping from the original feature space to the N-dimensional Reproducing Kernel Hilbert Space, and $\mathbf{T} \in \mathbb{R}^{N\times q}$ is the matrix of the corresponding linear combination coefficients.

Table 1
Description of multi-label data sets.

Data sets	Instances	Features	Labels	Card.	Dens.	Domain
Birds	645	260	20	1.470	0.074	audio
Medical	978	1449	45	1.245	0.028	text
Emotions	1702	1001	53	3.378	0.064	music
Langlog	1460	1004	75	1.180	0.016	text
Image	2000	294	5	1.236	0.247	image
Scene	2407	294	6	1.074	0.179	image
Yeast	2417	103	14	4.238	0.303	biology
Slashdot	3782	1079	22	1.181	0.054	text
Arts	5000	462	26	1.636	0.063	text
Computers	5000	681	33	1.509	0.046	text
Corel5k	5000	499	374	3.522	0.009	image
Enron	5000	550	33	1.461	0.044	text
Health	5000	612	32	1.662	0.052	text
Science	5000	743	40	1.4506	0.0363	text
Society	5000	636	27	1.692	0.063	text
Bibtex	7395	1836	159	2.402	0.015	text

Let $\Sigma = \Psi(\mathbf{X})^T\Psi(\mathbf{X})$ be the kernel matrix. Then, in Eq. (21), we can calculate that:

$$\begin{aligned}\Psi(\mathbf{X})^T\mathbf{W} &= \Psi(\mathbf{X})^T\Psi(\mathbf{X})\mathbf{T} \\ &= \Sigma\mathbf{T} \\ ||\mathbf{W}||_F^2 &= tr(\mathbf{W}^T\mathbf{W}) \\ &= tr(\mathbf{T}^T\Psi(\mathbf{X})^T\Psi(\mathbf{X})\mathbf{T}) \\ &= tr(\mathbf{T}^T\Sigma\mathbf{T})\end{aligned} \tag{33}$$

By replacing $\mathbf{XW}$ with $\Psi(\mathbf{X})^T\mathbf{W}$ in Eq. (21) and setting the derivative of the objective function w.r.t. $\mathbf{T}$ to zero, we have

$$\mathbf{T} = \widetilde{\Gamma}\mathbf{F} \tag{34}$$

where $\widetilde{\Gamma} = (\Sigma\mathbf{H}\Sigma + \frac{\delta}{\lambda}\Sigma)^{-1}\Sigma\mathbf{H} = (\mathbf{H}\Sigma + \frac{\delta}{\lambda}\mathbf{I})^{-1}\mathbf{H}$.

The updating rule of $\mathbf{F}$ is reformulated as:

$$\widetilde{\mathbf{O}_1}\mathbf{F} = \widetilde{\mathbf{O}_2} \tag{35}$$

where

$$\begin{aligned}\widetilde{\mathbf{O}_1} &= \mathbf{K} + \lambda(\Sigma\widetilde{\Gamma} - \mathbf{I})^T\mathbf{H}(\Sigma\widetilde{\Gamma} - \mathbf{I}) + \alpha\mathbf{A} + \delta\widetilde{\Gamma}^T\widetilde{\Gamma} + \frac{\mu}{2}\mathbf{I} \\ \widetilde{\mathbf{O}_2} &= \mathbf{K}(\mathbf{Y} - \mathbf{N}) + \frac{\mu}{2}\mathbf{Z} + \frac{1}{2}\Phi\end{aligned} \tag{36}$$

Similar to Eq. (24), the inverse matrix of $\widetilde{\mathbf{O}_1}$ only needs to be calculated for one time and the solution of Eq. (35) is definite. The detailed solving process of CS2PML-n is similar to that of the linear case, which is omitted due to limited space.

6. Experiments

6.1. Experiment preparation

In this section, we perform a sequence of experiments on totally 16 benchmark multi-label data sets downloaded from the websites of Mulan[1] and Uco[2]. The data sets cover five diverse domains: audio, music, image, biology and text. All features are normalized into the interval [0,1]. Table 1 presents a brief introduction of each data in detail, where “Instances”, “Features” and “Labels” represent the number of instances, the number of features, and the number of labels, respectively, “Card.” means the label cardinality, i.e., the number of labels distributed evenly to all instances, and “Dens.” denotes the normalization of label cardinality by the number of labels. To conduct semi-supervised partial multi-label learning, we randomly select the percentage of 40% instances from

[1] Data: http://mulan.sourceforge.net/datasets.html

[2] Data: http://www.uco.es/kdis/mllresources/

each data set as training data and the others as unlabeled data. In each training step, we randomly add noisy labels on each instance with the percentage of $\{50\%, 100\%, 150\%\}$ of ground-truth labels. All methods repeat five times in each case. As the validation is iterated, five pieces of experimental results are obtained. The mean metric value and the standard deviation of the results on each data are recorded.

The proposed methods are compared with some current state-of-art multi-label learning algorithms. The configuration parameters of all algorithms are set as suggested in the literature. The detailed information of the algorithms is described as follows:

The proposed methods: Two substantial algorithms are constructed, including the linear version CS2PML and the nonlinear version CS2PML-n using RBF kernel. Parameters α, θ, λ and δ are searched in $\{10^{-2}, 10^{-1}, \cdots, 10^{2}\}$. Parameter η is searched in $\{10^{-3}, 10^{-2}, \cdots, 10^{1}\}$. The influences when different parameters vary are also examined in the following contents.

SSPML [3]: SSPML is a semi-supervised partial multi-label learning method, where a latent label variable is introduced as a low-dimensional embedding of the feature space. Meanwhile, the multi-label classifier is jointly trained under the supervision of label variables. As suggestion, balancing parameters are set as $\lambda = 1$, $\beta = 1$, $\gamma = 1$, $\mu = 1$ and $\alpha = 0.1$.

PML-NI[3] [8]: PML-NI is a partial multi-label learning method that enables the identification of noisy labels caused by the ambiguous contents of instances. The parameters are set as $\alpha = 0.5$ and $\gamma = 0.5$. Moreover, parameter λ is selected from $\{1, 10, 100\}$.

MSWL[4] [33]: It generates a predicted label matrix for labeled and unlabeled data via semi-supervised multi-label learning. The algorithm first fills missing labels by global label correlation with a one-to-all style, and then uses feature manifold to build the regularizer. The parameters α and β are tuned in $\{10^{-3}, 10^{-2}, \cdots, 10^{3}\}$, and γ is tuned in $\{10^{-5}, 10^{-4}, \cdots, 10^{5}\}$.

TRAM[5] [19]: It formulates the transductive multi-label learning as an optimization problem of estimating label concept compositions and estimates the label sets of unlabeled instances by utilizing the information from both labeled and unlabeled data. The parameter $k = 10$ as the default setting.

PML-LRS [21]: PML-LRS is a partial multi-label learning approach by using low-rank and sparse decomposition. The observed label set is reformulated into a ground-truth label matrix and an irrelevant label matrix. As suggestion, the parameters are set as $\gamma = 0.01$, $\beta = 0.1$, and $\eta = 1$.

PARTICLA[6] [25]: PARTICLA is a two-stage partial multi-label learning model within which credible labels are first elicited from the candidate label set via label propagation, and then the label sets of unlabeled instances are produced via pairwise label ranking with virtual label splitting or maximum a posteriori reasoning. The model contains two substantial algorithms, i.e., PML-VLS and PML-MAP. The credible label elicitation threshold *thr* is set to 0.9, and the number of neighboring instances *k* is set to 10.

Five commonly used rank-based metrics, including *Macro average AUC* (*AUC*), *Coverage*, *Ranking Loss*, *Average Precision*, and *Hamming Loss* are employed to examine the performance of different multi-label learning algorithms. These metrics measure the predictive accuracy from multiple aspects, and their detailed definitions can be found in [1]. For the *AUC* and *Average Precision*, the greater the values, the better the performances; whereas for the *Hamming Loss*, *Coverage*, and *Ranking Loss*, the smaller the values, the better the performances.

6.2. Performance analysis

The experimental results under the case when the percentage of labeled data is 40% in listed in Tables 2,3,4,5,6, where the best performance on each data set is highlighted in bold. After a thorough and careful observation, we arrive at a couple of facts: All the comparing algorithms can achieve relatively remarkable performances in terms of the five metrics. In terms of *AUC*, CS2PML frequently outperforms other methods and slightly loses to PML-NI, TRAM and PML-LRS on Birds, Emotions, Image, Corel5k, Eron, and Bibtex datasets. In terms of *Coverage*, CS2PML-n slightly outperforms CS2PML while they both fail 2 or 3 cases when compared with PML-NI, MSWL, TRAM, PML-LRS, PML-VLS, and PML-MAP. PML-LRS and MSWL both employ the low-rank feature of the label matrix via label space representation. As a result, they sometimes obtain similar results. In terms of *Average Precision*, PML-VLS and PML-MAP outperform CS2PML and CS2PML-n on Yeast, Slashdot, Corel5k, Eron, Health, and Sciences datasets. The difference between PML-VLS and PML-MAP is not significant. The reason is mainly that PML-VLS and PML-MAP utilize the neighbor label distribution of a label for the credible label elicitation of the label, which can better explore the hidden labels when the rate of noisy labels is not too high. On the other hand, their performance is lack of stability in some cases because they only use the label information of nearest neighbors while ignoring the feature relevance. Although SSPML and PML-NI can not get the best results in many cases, their rankings remain high out of all the statistical tests. It is because the two algorithms jointly utilize the label information and the feature information and are effective when most noise has some special characteristics rather than is generated randomly. On the other hand, the overall performance of CS2PML and CS2PML-n are superior or highly competitive to the other comparing methods. Specifically, for the linear version, CS2PML achieves the best accuracy for 17 times in average over the 80 comparison tests; whereas for the nonlinear version, CS2PML-n achieves the best accuracy for 16 times in average over the 80 comparison tests.

The success of the proposed algorithms mainly relies in the following considerations: In the proposed semi-supervised learning methods, the overall label-level instance correlation is estimated through the evolution of the local instance correlation, which is more precise than using the feature-level instance correlation. The feature structures of labeled and unlabeled instances and the label structures can be jointly utilized to explore the label co-occurance of instances. Hence, the proposed methods can well account the inconsistence information between features and labels and can utilize more unsupervised information for effective partial multi-label learning.

To reveal the statistical significance, the Friedman test [34] and Bonferroni-Dunn test [35] are employed over all performances of the algorithms and the results are reported in Table 7 and Fig. 1 respectively. Denote r_j by the average rank of algorithm j on all data sets, N by the number of multi-label data sets, and K by the number of all algorithms. Under the null hypothesis, the Friedman statistic F_F follows a Fisher distribution with $(K-1)(N-1)$ degrees of freedom, which is defined as:

$$F_F = \frac{(N-1)\chi_F^2}{N(K-1)-\chi_F^2}, \text{ where}$$

$$\chi_F^2 = \frac{12N}{K(K+1)}\left(\sum_{j=1}^{K} r_j^2 - \frac{K(K+1)^2}{4}\right) \tag{37}$$

As seen in Table 7 that the null hypothesis that the performance of all methods is equivalent is rejected on each evaluation metric at a significance level of $\alpha = 0.05$. Moreover, in Fig. 1, Crit-

[3] Code: http://milkxie.github.io/code/MIPMLNIcode.zip

[4] Code: https://jiazhang-ml.pub/MSWL-master.zip

[5] Code: http://www.lamda.nju.edu.cn/Data.ashx

[6] Code: http://palm.seu.edu.cn/zhangml/files/PARTICLE.rar

Table 2
Comparative analysis of prediction performance of different algorithms in terms of *AUC*, where the best results (the larger the better) are shown in bold.

Methods	*AUC* ↑								
	CS2PML	CS2PML-n	SSPML	PML-NI	MSWL	TRAM	PML-LRS	PML-VLS	PML-MAP
Birds	0.647±.029	0.628±.025	0.595±.027	**.658±.020**	0.626±.023	0.623±.034	0.541±.028	0.592±.019	0.563±.035
Medical	**.702±.051**	0.682±.064	0.601±.034	0.697±.030	0.610±.041	0.697±.031	0.582±.067	0.612±.016	0.676±.034
Emotions	0.617±.031	0.609±.042	0.608±.016	**.637±.026**	0.610±.028	0.496±.045	0.572±.037	0.500±.091	0.500±.028
Langlog	**.593±.021**	0.592±.028	0.511±.035	0.531±.030	0.522±.043	0.567±.043	0.517±.019	0.519±.027	0.534±.040
Image	0.593±.040	0.640±.044	0.552±.018	**.598±.035**	0.580±.027	0.434±.032	0.433±.024	0.468±.028	0.591±.032
Scene	**.695±.017**	0.661±.012	0.598±.033	0.659±.034	0.603±.027	0.465±.016	0.465±.023	0.499±.021	0.499±.028
Yeast	**.718±.017**	0.700±.023	0.617±.020	0.678±.021	0.665±.025	0.637±.021	0.707±.017	0.561±.027	0.568±.025
Slashdot	0.681±.020	**.698±.019**	0.517±.018	0.539±.015	0.623±.017	0.624±.016	0.638±.015	0.523±.021	0.502±.018
Arts	0.573±.033	**.662±.037**	0.580±.027	0.638±.032	0.590±.033	0.603±.027	0.660±.028	0.543±.021	0.552±.026
Computers	0.730±.023	**.752±.021**	0.621±.023	0.727±.022	0.638±.032	0.654±.018	0.703±.028	0.537±.023	0.549±.025
Corel5k	0.572±.020	0.570±.018	0.533±.025	0.541±.019	0.639±.024	**.654±.012**	0.560±.020	0.499±.010	0.502±.020
Enron	0.600±.035	0.620±.027	0.551±.021	0.593±.017	0.518±.029	0.691±.027	**.695±.025**	0.541±.028	0.582±.013
Health	**.631±.020**	0.617±.021	0.568±.029	0.623±.012	0.565±.027	0.572±.016	0.562±.019	0.567±.008	0.587±.027
Science	**.643±.016**	0.625±.015	0.572±.019	0.637±.017	0.568±.024	0.567±.025	0.617±.020	0.540±.021	0.611±.021
Society	0.675±.024	**.707±.023**	0.603±.030	0.651±.024	0.503±.025	0.617±.021	0.663±.019	0.558±.028	0.549±.020
Bibtex	0.557±.025	0.552±.030	0.548±.039	0.558±.033	0.532±.042	**.604±.025**	0.536±.021	0.525±.037	0.550±.022

Table 3
Comparative analysis of prediction performance of different algorithms in terms of *Ranking Loss*, where the best results (the smaller the better) are shown in bold.

Methods	*Ranking Loss* ↓								
	CS2PML	CS2PML-n	SSPML	PML-NI	MSWL	TRAM	PML-LRS	PML-VLS	PML-MAP
Birds	0.313±.011	0.340±.017	0.378±.028	0.307±.011	0.329±.020	0.360±.014	0.355±.016	0.268±.004	**.264±.024**
Medical	0.291±.043	0.311±.042	0.392±.019	0.294±.018	0.383±.032	0.297±.020	0.421±.028	**.289±.011**	0.343±.013
Emotions	0.359±.036	0.369±.030	0.374±.009	0.340±.027	0.369±.023	**.241±.031**	0.406±.033	0.464±.182	0.415±.013
Langlog	0.404±.018	0.403±.016	0.490±.031	0.423±.023	0.385±.018	**.224±.026**	0.406±.015	0.477±.021	0.385±.030
Image	0.351±.027	0.281±.030	0.390±.007	0.416±.019	0.385±.012	**.260±.025**	0.289±.030	0.455±.001	0.321±.022
Scene	0.277±.006	0.313±.009	0.380±.023	0.375±.030	0.375±.021	**.233±.017**	0.305±.004	0.497±.016	0.491±.025
Yeast	0.278±.008	0.297±.005	0.376±.004	0.314±.009	0.328±.010	0.357±.015	0.295±.011	**.241±.017**	0.285±.012
Slashdot	0.282±.015	0.274±.012	0.429±.004	0.438±.008	0.366±.015	0.365±.015	**.255±.012**	0.565±.034	0.563±.005
Arts	0.575±.035	**.313±.029**	0.407±.015	0.337±.027	0.398±.043	0.376±.083	0.435±.034	0.412±.030	0.323±.017
Computers	0.418±.022	**.216±.019**	0.363±.017	0.243±.008	0.353±.013	0.320±.104	0.416±.019	0.286±.006	0.339±.008
Corel5k	0.451±.012	**.430±.008**	0.468±.002	0.460±.001	0.467±.008	0.461±.017	0.472±.011	0.507±.035	0.729±.006
Enron	0.395±.018	0.358±.019	0.439±.012	0.391±.016	0.480±.028	**.294±.040**	0.362±.018	0.389±.023	0.322±.004
Health	0.354±.006	0.370±.015	0.424±.014	0.362±.014	0.427±.014	0.419±.018	0.352±.014	0.321±.001	**.304±.014**
Science	**.334±.007**	0.355±.005	0.416±.012	0.340±.013	0.421±.014	0.423±.027	0.347±.007	0.394±.008	0.397±.009
Society	0.220±.022	0.220±.013	0.345±.018	0.288±.010	0.496±.029	0.356±.039	**.218±.015**	0.310±.015	0.373±.012
Bibtex	0.434±.030	0.439±.014	0.450±.026	0.437±.013	0.463±.028	**.382±.024**	0.443±.016	0.437±.022	0.485±.013

Table 4
Comparative analysis of prediction performance of different algorithms in terms of *Coverage*, where the best results (the smaller the better) are shown in bold.

Methods	*Coverage* ↓								
	CS2PML	CS2PML-n	SSPML	PML-NI	MSWL	TRAM	PML-LRS	PML-VLS	PML-MAP
Birds	7.41±.26	7.87±.25	8.58±.34	7.27±.07	7.69±.31	8.26±.16	7.88±.26	6.56±.03	**5.89±.46**
Medical	14.73±1.84	15.54±2.07	19.20±.59	14.83±.53	18.70±1.49	14.72±.60	**14.44±2.26**	14.47±.11	16.16±.39
Emotions	2.81±.19	2.85±.16	2.85±.03	**2.68±.07**	2.83±.14	3.29±.11	2.99±.16	3.21±.32	3.29±.08
Langlog	29.12±.72	29.13±.74	34.27±1.69	32.86±.88	13.84±.74	**13.64±1.20**	32.94±.80	33.43±1.36	29.09±1.36
Image	1.51±.11	**1.38±.13**	1.80±.03	1.79±.07	2.06±.04	2.06±.18	1.42±.13	1.90±.01	1.67±.06
Scene	**1.48±.01**	1.66±.01	1.99±.10	1.66±.12	1.97±.09	2.51±.07	1.71±.01	2.56±.01	2.51±.09
Yeast	8.46±.20	8.71±.18	9.52±.15	8.82±.10	8.88±.25	9.41±.10	9.60±.19	**7.86±.15**	8.17±.12
Slashdot	214.10±3.29	213.57±2.92	272.52±.84	251.65±.36	243.18±3.61	**206.82±2.28**	211.30±3.02	306.92±11.50	287.40±2.44
Arts	12.21±.74	10.25±.68	12.62±.28	10.78±.46	12.36±1.06	12.83±2.06	13.28±.65	12.58±.54	**9.77±.31**
Computers	10.89±.59	**8.93±.55**	13.84±.49	9.91±.16	13.46±.45	12.43±3.32	15.49±.57	11.05±.16	12.54±.10
Corel5k	268.66±3.15	**265.33±3.07**	279.74±.75	277.27±.37	278.81±3.70	280.56±2.34	270.81±3.05	305.80±1.55	341.13±2.74
Enron	33.60±.93	32.17±.96	36.29±.24	33.91±.76	36.37±2.12	**28.12±2.38**	33.68±.88	37.14±.99	32.99±.23
Health	16.49±.23	17.10±.26	19.23±.39	16.82±.16	19.33±.59	19.03±.01	16.33±.26	15.10±.01	**14.31±.27**
Science	15.63±.10	16.51±.01	18.90±.38	15.93±.20	18.99±.45	19.20±.98	**12.03±.02**	17.56±.36	13.62±.32
Society	8.18±.45	**8.16±.46**	10.60±.22	10.26±.14	14.80±.62	11.86±.79	8.98±.47	10.20±.15	10.87±.26
Bibtex	92.54±.45	92.48±.46	94.96±.22	93.05±.14	97.82±.62	**86.28±.80**	96.92±.47	93.70±.15	101.51±.27

ical Distance (CD) diagrams of the Bonferroni-Dunn test are depicted to show the overall performance of the proposed methods. At the 0.05 significance level, we can compute:

$$CD_\alpha = q_\alpha \sqrt{\frac{K(K+1)}{6N}} = 2.6375 \tag{38}$$

where $q_\alpha = 2.724$, $K = 9, N = 16$. If the average ranks of two methods differ by one CD, the two methods are regarded as different. As indicated in Fig. 1, CS2PML and CS2PML-n significantly outperform the comparing methods more than one time on some of the evaluation metrics.

Table 5
Comparative analysis of prediction performance of different algorithms in terms of *Average Precision*, where the best results (the larger the better) are shown in bold.

Methods	*Average Precision* ↑								
	CS2PML	CS2PML-n	SSPML	PML-NI	MSWL	TRAM	PML-LRS	PML-VLS	PML-MAP
Birds	0.390±.010	0.355±.015	0.309±.044	**.399**±**.035**	0.397±.012	0.338±.011	0.236±.015	0.376±.037	0.373±.045
Medical	**.251**±**.043**	0.229±.056	0.171±.014	0.241±.013	0.167±.028	0.242±.030	0.240±.051	0.218±.007	0.212±.021
Emotions	0.631±.025	0.632±.030	0.606±.011	**.638**±**.017**	0.618±.027	0.530±.053	0.594±.032	0.569±.049	0.530±.008
Langlog	0.123±.038	0.124±.039	0.077±.004	0.083±.011	0.229±.014	**.230**±**.015**	0.113±.024	0.083±.018	0.122±.022
Image	0.560±.027	**.642**±**.028**	0.561±.018	0.606±.026	0.595±.011	0.526±.037	0.601±.033	0.515±.002	0.594±.025
Scene	**.626**±**.005**	0.587±.008	0.516±.026	0.576±.026	0.521±.028	0.430±.008	0.591±.009	0.418±.003	0.430±.023
Yeast	0.653±.006	0.630±.012	0.533±.008	0.596±.008	0.578±.016	0.548±.019	0.521±.009	**.679**±**.012**	0.657±.015
Slashdot	**.151**±**.014**	0.145±.009	0.030±.001	0.031±.002	0.055±.014	0.032±.006	0.137±.009	0.028±.007	0.028±.006
Arts	0.257±.035	**.348**±**.033**	0.242±.016	0.314±.026	0.240±.027	0.265±.052	0.214±.035	0.197±.021	0.232±.020
Computers	0.420±.019	**.474**±**.011**	0.250±.028	0.418±.023	0.245±.017	0.282±.059	0.204±.014	0.220±.027	0.148±.012
Corel5k	0.048±.013	0.052±.015	0.029±.006	0.031±.010	0.029±.015	0.041±.015	0.049±.011	**.068**±**.009**	0.021±.007
Enron	0.209±.023	0.230±.024	0.165±.005	0.202±.012	0.108±.033	0.316±.021	0.244±.023	0.254±.018	**.326**±**.008**
Health	0.247±.009	0.222±.011	0.173±.024	0.229±.019	0.160±.031	0.175±.004	0.148±.008	0.211±.008	**.267**±**.010**
Science	0.239±.006	0.215±.010	0.165±.014	0.225±.014	0.152±.026	0.158±.037	0.139±.001	0.236±.008	**.254**±**.012**
Society	**.494**±**.008**	0.476±.008	0.316±.021	0.375±.018	0.148±.019	0.288±.029	0.459±.006	0.353±.017	0.262±.012
Bibtex	0.063±.014	0.062±.013	0.057±.022	0.061±.024	0.057±.024	**.135**±**.035**	0.052±.009	0.076±.019	0.056±.017

Table 6
Comparative analysis of prediction performance of different algorithms in terms of *Hamming Loss*, where the best results (the smaller the better) are shown in bold.

Methods	*Hamming Loss* ↓								
	CS2PML	CS2PML-n	SSPML	PML-NI	MSWL	TRAM	PML-LRS	PML-VLS	PML-MAP
Birds	**.074** ±**.009**	0.087±.008	0.132±.004	0.187±.005	0.099±.006	0.400±.007	0.113±.009	0.120±.024	0.077±.013
Medical	**.028**±.010	0.062±.009	0.171±.013	0.277±.009	0.162±.009	0.399±.007	0.157±.005	0.075±.006	0.032±.010
Emotions	0.311±.032	0.307±.028	0.311±.015	0.336±.006	0.310±.009	0.450±.027	**.128**±**.028**	0.324±.030	0.311±.011
Langlog	0.017±.001	**.016**±**.005**	0.905±.006	0.355±.006	0.225±.008	0.405±.014	0.062±.008	0.278±.007	0.017±.011
Image	0.234±.001	**.230**±**.004**	0.733±.008	0.284±.008	0.750±.005	0.498±.026	0.244±.001	0.235±.001	0.299±.019
Scene	**.180**±**.000**	0.189±.006	0.226±.005	0.305±.006	0.195±.008	0.480±.110	0.412±.003	0.489±.005	0.191±.008
Yeast	0.301±.013	0.279±.011	0.303±.009	0.319±.015	0.301±.000	0.337±.020	0.396±.012	**.257**±**.015**	0.295±.017
Slashdot	**.009**±**.004**	**.009**±**.003**	0.027±.002	0.271±.005	0.058±.007	0.404±.004	0.019±.010	0.016±.010	0.019±.004
Arts	0.068±.004	**.066**±**.008**	0.068±.006	0.132±.005	0.107±.001	0.405±.011	0.372±.002	0.156±.007	0.097±.009
Computers	0.049±.006	**.047**±**.001**	0.054±.008	0.093±.009	0.099±.000	0.308±.016	0.384±.003	0.125±.013	0.072±.005
Corel5k	**.010**±**.002**	0.048±.005	0.039±.009	0.291±.005	0.035±.002	0.041±.008	0.368±.008	0.015±.010	0.019±.006
Enron	0.103±.001	0.161±.005	0.352±.008	0.368±.007	0.302±.005	0.391±.008	0.466±.009	0.098±.007	**.066**±**.004**
Health	**.036**±**.008**	0.037±.010	0.046±.010	0.091±.009	0.057±.000	0.040±.006	0.040±.005	0.082±.004	0.037±.004
Science	0.036±.005	0.035±.004	0.046±.010	0.091±.001	0.064±.009	**.031**±**.005**	0.040±.004	0.086±.001	0.038±.014
Society	0.060±.002	0.057±.010	0.058±.007	0.112±.006	0.422±.005	**.040**±**.006**	0.055±.003	0.178±.012	0.179±.013
Bibtex	0.018±.004	**.015**±**.009**	0.230±.014	0.371±.010	0.239±.009	0.400±.010	0.433±.006	0.036±.010	**.015**±**.008**

Table 7
Summary of the Friedman statistics F_F (# comparing algorithms $K = 9$, # data sets $N = 16$) and the critical value at 0.05 with different evaluation metrics.

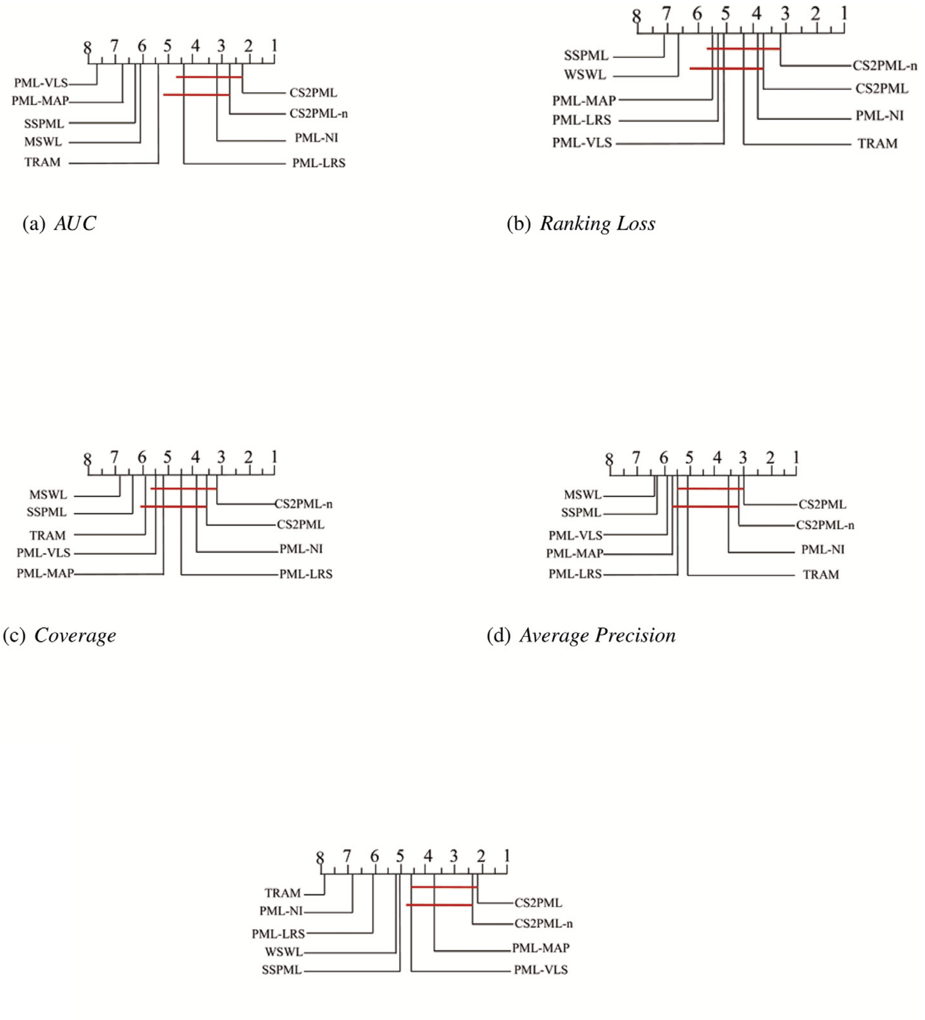
Metrics	F_F	Critical value
AUC	18.5862	2.016
Ranking loss	7.9524	
Coverage	93.6871	
Averageprecision	10.5236	
Hamming loss	26.2573	

6.3. Runtime analysis

It is also important to compare the running times of these algorithms, particularly for the datasets with large feature dimensions. From Table 1, we can see that the feature dimensions of the datasets are all larger than their label dimensions. Table 8 reports the runtimes and the rankings of the algorithms. TRAM and PML-NI are fast because of their simple operability. PML-VLS and PML-MAP are relatively slow on the datasets of high dimensionality because they need to traverse through all the instances and to search the neighbors of each instance for label confidence reasoning. CS2PML is also desirable which only loses to PML-NI and TRAM in terms of the average runtime. Since CS2PML-n utilizes a nonlinear mapping from the original feature space to the high dimensional space, the amount of calculations is increased. Its average runtime and average rank are respectively 5.44 and 75.46, which is less efficient than TRAM, PML-NI, CS2PML, and PML-LRS. However, the nonlinear version is still a desirable option for linearly inseparable datasets and can achieve good performance in terms of the evaluation metrics.

6.4. Sensitivity analysis

We test the influence of involved parameters by varying one while keeping others fixed and illustrate the variations of performance in Fig. 2. Parameter α is used to reflect the influence of the label-level instance correlation. It can be seen that the performance of CS2PML achieves the best in some intermediate regions and gradually deteriorates toward the outer regions. This is because that the label-level instance correlation can be fully utilized when α falls into a fitting area; while when α gets too large, the effect of label correlation would overwhelm other principal terms. The curve changes in a similar way while λ and δ vary. Parameters λ and δ are used to realize the utilization of feature information for label prediction. When they are too small or too large, the influence of feature information can not be reasonably utilized for label inference. This illustrates that the selection of suitable bounds for λ and δ can effect the accuracy when utilizing feature information for labeling matrix prediction. Parameters η and θ are used to



(a) *AUC*

(b) *Ranking Loss*

(c) *Coverage*

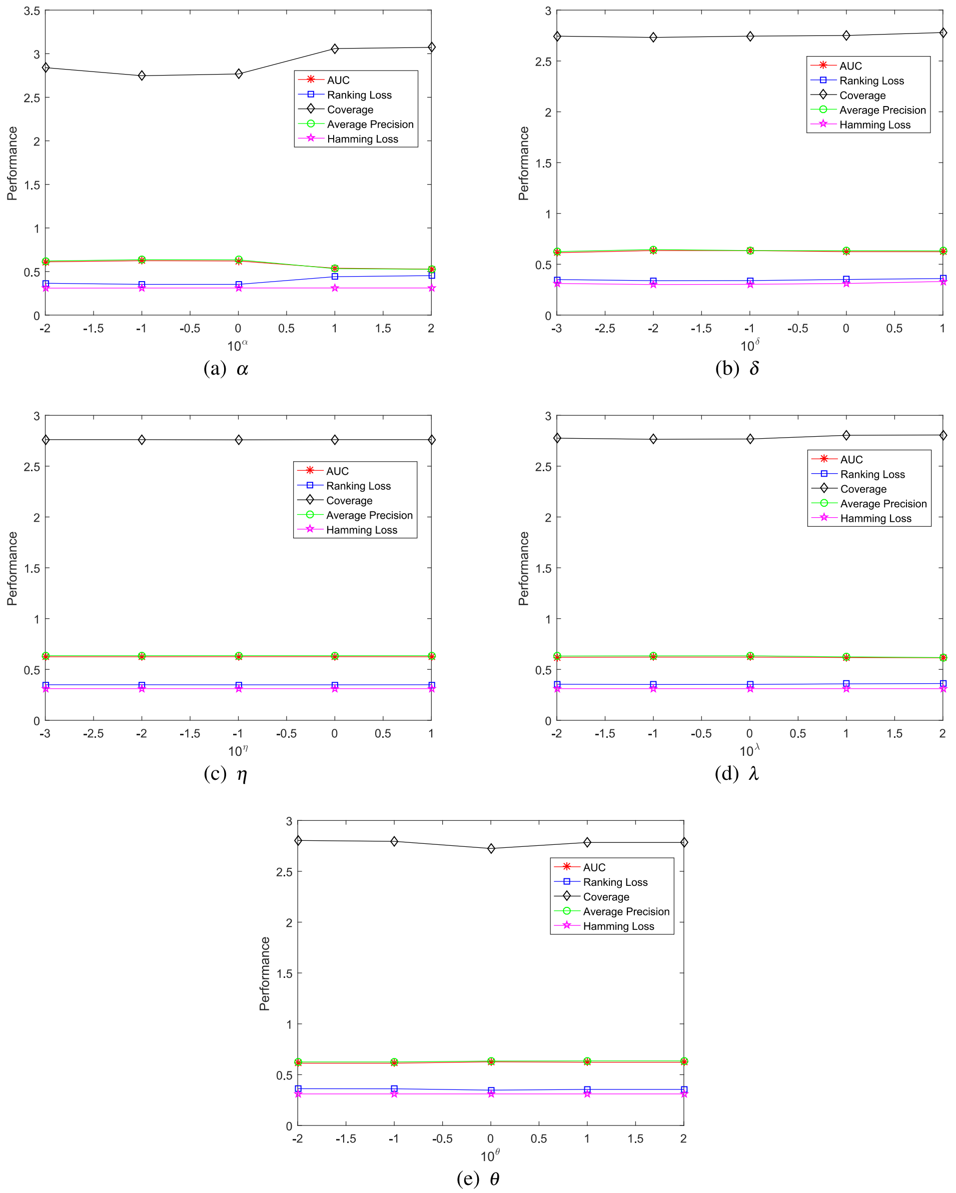
(d) *Average Precision*

(e) *Hamming Loss*

Fig. 1. Comparison of CS2PML and CS2PML-n against algorithms under comparison with the Bonferroni-Dunn test ($CD = 2.6375$ at 0.05 significance level). Algorithms not connected within one CD diagram are considered to have a significantly different performance from the control approach.

Table 8
Runtime of different algorithms, where the best results are shown in bold.

Methods	*Runtime* (*Ranking*)								
	CS2PML	CS2PML-n	SSPML	PML-NI	MSWL	TRAM	PML-LRS	PML-VLS	PML-MAP
Birds	0.84 (2)	1.16 (6)	15.95 (8)	0.81 (1)	27.05 (9)	1.59 (7)	0.95 (3)	0.95 (4)	1.11 (5)
Medical	5.67 (2)	6.56 (4)	61.30 (8)	16.53 (6)	65.61 (9)	2.34 (1)	6.11 (3)	16.14 (5)	19.97 (7)
Emotions	1.06 (6)	0.80 (4)	12.16 (9)	1.25 (7)	5.81 (8)	0.53 (3)	0.91 (5)	0.38 (2)	0.28 (1)
Langlog	1.13 (5)	0.89 (3)	12.18 (8)	1.32 (6)	83.08 (9)	2.16 (7)	1.00 (4)	0.42 (2)	0.33 (1)
Image	1.09 (6)	0.81 (3)	12.21 (9)	1.32 (7)	8.91 (8)	0.88 (4)	0.90 (5)	0.44 (2)	0.35 (1)
Scene	8.23 (4)	17.48 (8)	35.62 (9)	8.67 (5)	9.38 (6)	1.27 (1)	9.73 (7)	4.22 (3)	3.27 (2)
Yeast	6.08 (4)	13.55 (7)	35.62 (9)	7.58 (5)	15.66 (8)	1.64 (1)	9.84 (6)	4.56 (2)	5.55 (3)
Slashdot	7.01 (4)	14.35 (7)	35.72 (8)	10.40 (5)	76.27 (9)	4.72 (1)	10.70 (6)	5.39 (2)	6.46 (3)
Arts	45.02 (2)	116.88 (7)	122.02 (8)	52.36 (4)	130.70 (9)	6.34 (1)	50.89 (3)	69.78 (5)	85.83 (6)
Computers	52.70 (2)	121.94 (7)	2199.59 (9)	54.84 (3)	95.52 (5)	6.94 (1)	67.33 (4)	109.17 (6)	127.47 (8)
Corel5k	79.69 (4)	142.41 (5)	242.70 (6)	42.44 (1)	948.98 (9)	54.95 (2)	61.05 (3)	396.98 (7)	433.33 (8)
Enron	6.94 (2)	12.64 (5)	44.55 (6)	8.45 (3)	82.48 (9)	2.09 (1)	9.90 (4)	46.83 (7)	58.24 (8)
Health	53.03 (3)	129.75 (5)	157.11 (6)	12.22 (2)	58.66 (4)	6.97 (1)	691.31 (9)	238.50 (7)	299.19 (8)
Science	56.36 (3)	130.33 (6)	158.23 (7)	15.17 (2)	62.92 (4)	7.30 (1)	63.36 (5)	259.98 (8)	321.61 (9)
Society	57.03 (3)	131.02 (5)	218.85 (7)	15.99 (2)	163.30 (6)	4.23 (1)	64.25 (4)	260.81 (8)	322.38 (9)
Bibtex	243.73 (4)	366.78 (5)	389.43 (6)	42.05 (1)	948.14 (7)	48.20 (2)	205.29 (3)	4875.06 (8)	5596.83 (9)
Average	39.10 (3.50)	75.46 (5.44)	234.58 (7.69)	18.21 (3.75)	173.90 (7.44)	9.51 (2.19)	78.34 (4.63)	393.10 (4.88)	455.14 (5.50)



(a) α

(b) δ

(c) η

(d) λ

(e) θ

Fig. 2. Parameter sensitivity analysis of CS2PML on Emotions dataset with varying the trade-off parameters α, δ, η, λ, and θ.

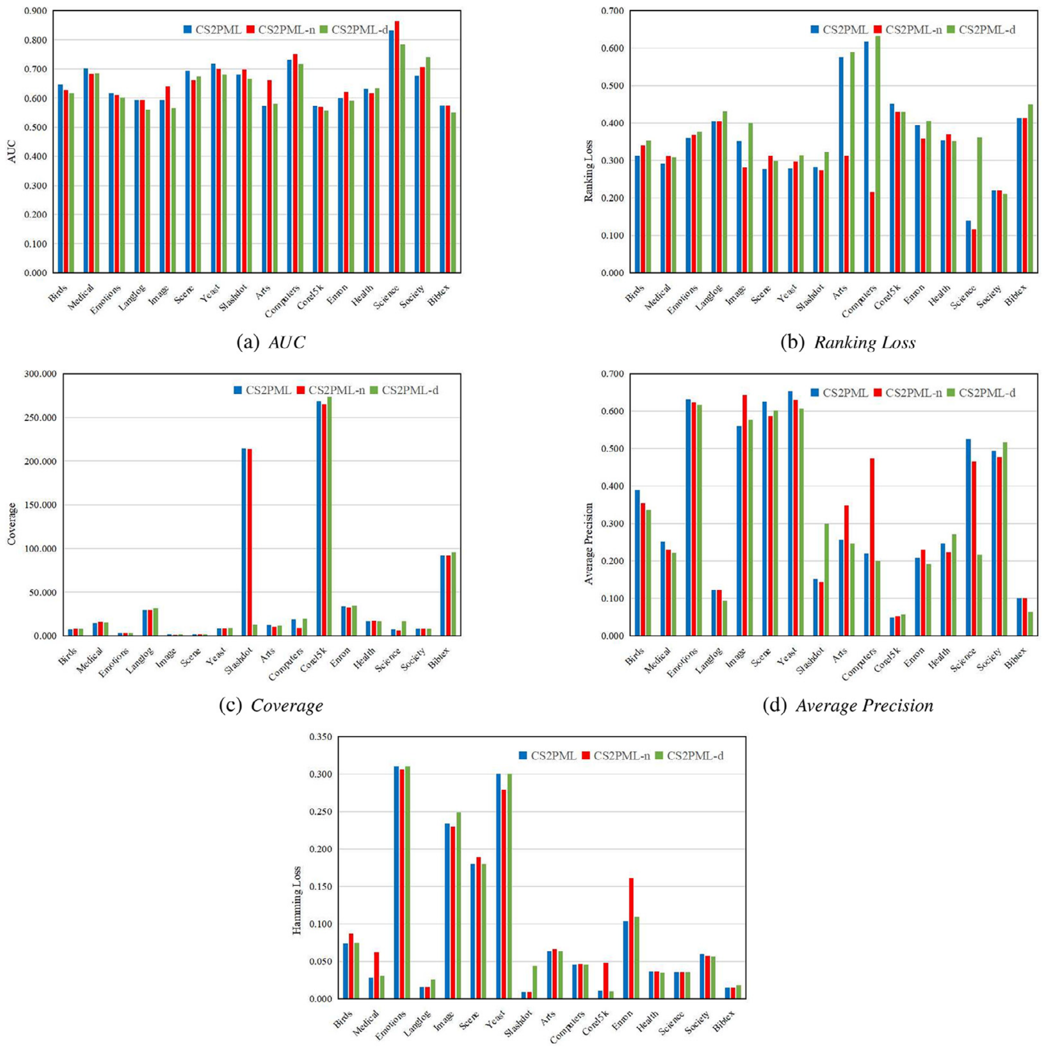
realize the low-rank of the latent label matrix and the sparsity of the noisy label matrix. The performance of each valuation metric will be stable when η and θ get too large. It is because that the noisy label matrix and the rank of the latent label matrix tend to be invariable as η and θ vary. It is suggested that there are suitable bounds for the parameters, which can enforce the proposed methods to be stable.

6.5. Ablation study

We further investigate the importance of the HISC technique for label-level instance correlation estimation on unseen instances. A degenerated version of the proposed approach named CS2PML-d without using the HSIC-based instance correlation matrix is generated. Fig. 3 outlines the comparison results of CS2PML, CS2PML-n and WPML-d in terms of each metric. As seen in the figure, CS2PML and CS2PML-n can achieve superior or at least comparable performance to the degenerated version in most cases. This strongly verifies the benefit of incorporating the HSIC-based instance correlation matrix for improving the generalization performance of the proposed models. Furthermore, it can be seen that the degenerated approach can obtain better performance in several cases. This may be caused by the fact that the existence of massive tail labels makes the estimation of label-level correlation more difficult and the HSIC-based strategy fails to consider this influence. In such cases, it would be more practical to implement multiple cross training so as to disambiguate the inconsistency of distributions between the unlabeled and labeled data.

6.6. Convergence analysis

Fig. 4 outlines the change in the objective function value w.r.t. each iteration on the Medical and Yeast datasets. From the figure, we can observe that the loss curve falls fast within the number of iterations and then tends towards stability. Hence, the results empirically verify the convergence of the proposed algorithm in practice.



(a) *AUC*

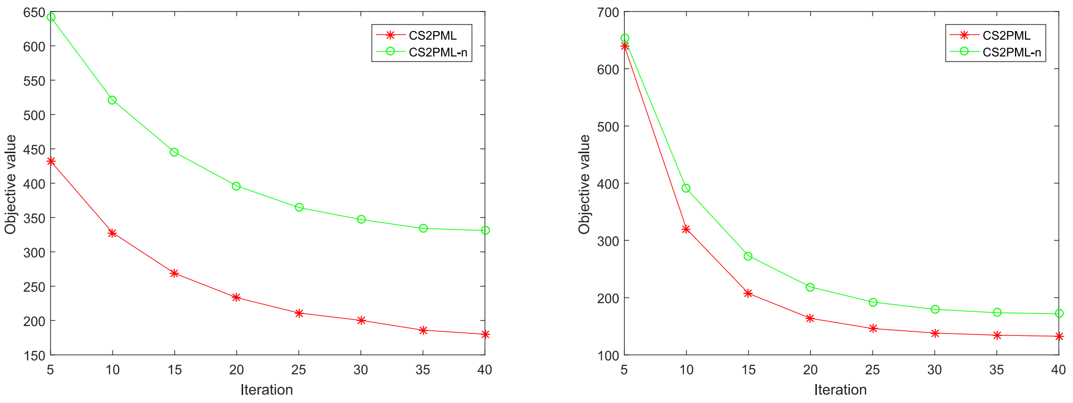
(b) *Ranking Loss*

(c) *Coverage*

(d) *Average Precision*

(e) *Hamming Loss*

Fig. 3. Comparisons of CS2PML, CS2PML-n, and the degenerated version CS2PML-d in terms of each evaluation metric.



(a) Medical

(b) Yeast

Fig. 4. Convergence analysis of CS2PML and CS2PML-n on Medical and Yeast data sets.

7. Conclusion

In this paper, we address the problem of partial multi-label learning in semi-supervised setting, where the training instances are either associated with full candidate labels or are without supervised labels. The relation between the feature and the label spaces is established via HSIC, based on which the label-level instance correlation can be yielded on both labeled and unlabeled instances. Then, a new method and its kernel version are proposed that employ three components including feature mapping, label-level correlation maintenance, and low-rank and sparse schemes. The proposed methods are enable to resolve the inconsistency of the feature and the label structures and have generalizations. The experiments demonstrate that the proposed methods can achieve competitive superiority against the state-of-the-art methods. Furthermore, ablation study further analyzes their effectiveness.

In the future, we will explore the following problems:

(1) How to automatically achieve the consistency between labels and features within the training of the desired predictors for partial multi-label classification.

(2) Introduce more powerful and efficient learning models by considering tail labels and various types of noisy labels in reality.

(3) Leverage the consistency idea to multi-view multi-label learning and hierarchical classification.

Declaration of Competing Interest

No conflict of interest.

Acknowledgements

This work was supported by National Key Research and Development Program of China (No. 2020AAA0106100) and National Natural Science Foundation of China (Nos. 62076221 and 61976194).

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.patcog.2022.108839.

References

[1] M.-L. Zhang, Z.H. Zhou, A review on multilabel learning algorithms, IEEE Trans. Knowl. Data Eng. 26 (2014) 1819–1837.

[2] Z.-H. Zhou, M.-L. Zhang, C. Sammut, G.I. Webb, Multi-label learning, Encyclopedia of Machine Learning and Data Mining, second ed., Springer, Eds. Berlin:, 2017.

[3] M. Xie, S.J. Huang, Partial multi-label learning with noisy label identification, IEEE Trans. Pattern Anal. Mach. Intell. (2021), doi:10.1109/TPAMI.2021.3059290.

[4] S. Bucak, R. Jin, A. Jain, Multilabel learning with incomplete class assignments, in: 2011 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2011, Providence, RI, USA, pp. 2801–2808, 2011.

[5] B. Wu, S. Lyu, B. Hu, Q. Ji, Multi-label learning with missing labels for image annotation and facial action unit recognition, Pattern Recognit. 48 (2015) 2279–2289.

[6] G. Chen, Y. Song, F. Wang, C. Zhang, Semi-supervised multilabel learning by solving a sylvester equation, in: Proceedings of the 2008 SIAM International Conference on Data Mining, 2008, pp. 410–419.

[7] Y. Guo, D. Schuurmans, Semi-supervised multi-label classification, ECML/PKDD (2012) 355–370.

[8] M. Xie, S.J. Huang, Semi-supervised partial multi-label learning, in: Proceedings of the 20th IEEE International Conference on Data Mining, p. in press, 2020. 10.1109/ICDM50108.2020.00078

[9] K. Mikalsena, C. Soguero-Ruiz, F. Bianchi, R. Jenssen, Noisy multi-label semi–supervised dimensionality reduction, Pattern Recognit. 90 (2019) 257–270.

[10] P. Zhu, Q. Xu, Q. Hu, C. Zhang, H. Zhao, Multi-label feature selection with missing labels,multi-label feature selection with missing labels, Pattern Recognit. 74 (2018) 488–502.

[11] P. Zhang, G. Liu, W. Gao, J. Song, Multi-label feature selection considering label supplementation, Pattern Recognit. 120 (2021) 108137.

[12] J. Lv, T. Wu, C. Peng, Y. Liu, N. Xu, X. Geng, Compact learning for multi-label classification, Pattern Recognit. 113 (2021) 107833.

[13] Y. Sun, G. Lyu, S. Feng, Partial label learning via subspace representation and global disambiguation, ECML/PKDD 2 (2020) 439–454.

[14] B. Skrlj, S. Dzeroski, N. Lavrac, M. Petkovic, Reliefe: feature ranking in high-dimensional spaces via manifold embeddings, Mach. Learn. 111 (2022) 273–317.

[15] S. Liu, X. Song, Z. Ma, E. Ganaa, X. Shen, MoRE: multi-output residual embedding for multi-label classification, Pattern Recognit. 126 (2022) 108584.

[16] A. Gretton, O. Bousquet, A. Smola, B. Schölkopf, Measuring statistical dependence with hilbert-schmidt norms, in: Proceedings of the 16th International Conference on Algorithmic Learning Theory,Singapore, 2005, pp. 63–77.

[17] Y. Li, D. Liang, Safe semi-supervised learning: a brief introduction, Fronter Comput. Sci. 13 (2019) 669–676.

[18] Y. Liu, R. Jin, L. Yang, Semi-supervised multi-label learning by constrained non-negative matrix factorization, in: Proceedings of the National Conference on Artificial Intelligence, 1, 2006, pp. 421–426.

[19] X. Kong, M.K. Ng, Z.H. Zhou, Transductive multilabel learning via label set propagation, IEEE Trans. Knowl. Data Eng. 25 (2013) 704–719.

[20] L. Wu, M.L. Zhang, Multi-label classification with unlabeled data: an inductive approach, J. Mach. Learn. Res. 29 (2013) 197–212.

[21] L. Sun, S. Feng, T. Wang, C. Lang, Y. Jin, Partial multi-label learning by low-rank and sparse decomposition, in: Proceedings of the 33rd AAAI Conference on Artificial Intelligence, Honolulu, HI, 2019, pp. 5016–5023.

[22] L. Sun, S. Feng, Partial multi-label learning with noisy side information, Knowl. Inf. Syst. 63 (2021) 541–564.

[23] T. Yu, G. Yu, J. Wang, M. Guo, Partial multi-label learning with label and feature collaboration, in: International Conference on Database Systems for Advanced Applications, 12112, 2020, pp. 621–637.

[24] H. Wang, W. Liu, Y. Zhao, C. Zhang, T. Hu, G. Chen, Discriminative and correlative partial multi-label learning, in: Proceedings of the 28th International Joint Conference on Artificial Intelligence Macau, China, 2019, pp. 3691–3697.

[25] M.-L. Zhang, J.P. Fang, Partial multi-label learning via credible label elicitation, IEEE Trans. Pattern Anal. Mach. Intell. 43 (2021) 3587–3599.

[26] G. Lyu, S. Feng, Y. Jin, T. Wang, C. Lang, Y. Li, Prior knowledge regularized self-representation model for partial multilabel learning, IEEE Trans. Cybern. (2021). P. in press

[27] X. Liu, L. Sun, S. Feng, Incomplete multi-view partial multi-label learning, Appl. Intell. 52 (2022) 3289–3302.

[28] M. Xie, F. Sun, S.J. Huang, Partial multi-label learning with meta disambiguation, in: Proceedings of the 21st ACM SIGKDD Conference on Knowledge Discovery and Data Mining, p. in press, 2021.

[29] Z. Lin, R. Liu, Z. Su, Linearized alternating direction method with adaptive penalty for low-rank representation, Advances in Neural Information Processing Systems, pp. 612–620, 2011.

[30] X. Zhang, Y. Ma, Z. Lin, H. Gao, L. Zhuang, N. Yu, Non-negative low rank and sparse graph for semi-supervised learning, in: IEEE Conference on Computer Vision and Pattern Recognition, 2012, pp. 2328–2335.

[31] S. Ji, J. Ye, An accelerated gradient method for trace norm minimization, in: Proceedings of the 26th International Conference on Machine Learning, 2009, pp. 457–464.

[32] B. Schölkopf, A. Smola, Learning with Sernels: Support Vector Machines, Regularization, Optimization and Beyond, Camridge, MA: MIT Press, 2001.

[33] J. Zhang, S. Li, M. Jiang, K. Tan, Learning from weakly labeled data based on manifold regularized sparse model, IEEE Transactions on Cybernetics, 2021. in press

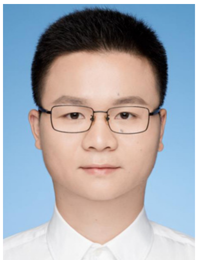
[34] M. Friedman, A comparison of alternative tests of significance for the problem of m rankings, Ann. Math. Stat. 11 (1940) 86–92.

[35] J. Demsar, Statistical comparisons of classifiers over multiple data sets, J. Mach. Learn. Res. 7 (2006) 1–30.

Anhui Tan received his M.S. degree from the School of Mathematics at Shandong University, Jinan, China, in 2012, and the Ph.D. degree from the School of Mathematics at Xiamen University, Xiamen, China, in 2015. He is currently an associate professor with the School of Information Engineering, Zhejiang Ocean University, China. He is actively pursuing research in machine learning and granular computing.

Jiye Liang received the PhD degree from Xi'an Jiaotong University, Xi'an, China. He is currently a professor in Key Laboratory of Computational Intelligence and Chinese Information Processing of Ministry of Education, the School of Computer and Information Technology, Shanxi University, Taiyuan, China. His research interests include artificial intelligence, granular computing, and machine learning. He has published more than 160 papers in his research fields, including TPAMI, TKDE, KDD, and Artificial Intelligence.

Wei-Zhi Wu received the M.Sc. degree in Mathematics from East China Normal University, Shanghai, China, in 1992, and the Ph.D. degree in Applied Mathematics from Xi'an Jiaotong University, Xi'an, China, in 2002. He is currently a professor with the School of Information Engineering, Zhejiang Ocean University. He has published 3 monographs and more than 150 articles in international journals and book chapters. His current research interests include granular computing, approximate reasoning, and data mining. Dr. Wu also serves in the editorial boards of several international journals.



Jia Zhang received the M.S. degree from the School of Computer Science, Minnan Normal University, Zhangzhou, China, in 2016, and the Ph.D. degree in artificial intelligence from Xiamen University, Xiamen, China, in 2020. He is currently a lecturer with the College of Information Science and Technology, Jinan University. He is broadly interested in machine learning, data mining, and artificial intelligence. He is currently working on multi-label learning, data fusion, feature selection, and weakly supervised learning.