label distribution learning 不仅捕获多个标签，同时发掘标签对于实例的重要程度

Progressive Enhancement of Label Distributions for Partial Multilabel Learning

Ning Xu, *Member, IEEE*, Yun-Peng Liu, Yan Zhang, and Xin Geng, *Senior Member, IEEE*

PML：标签部分已知或不完整

sharing topological structure
特征空间和标签分布空间之间存在一种共同的几何或结构关系

***Abstract*— Partial multi-label learning (PML) aims to learn a multilabel predictive model from the PML training examples, each of which is associated with a set of candidate labels where only a subset is valid. The common strategy to induce a predictive model is identifying the valid labels in each candidate label set. Nonetheless, this strategy ignores considering the essential *label distribution* corresponding to each instance as label distributions are not explicitly available in the training dataset. In this article, a novel partial multilabel learning method is proposed to recover the latent label distribution and progressively enhance it for predictive model induction. Specifically, the label distribution is recovered by considering the observation model for logical labels and the sharing topological structure from feature space to label distribution space. Besides, the latent label distribution is progressively enhanced by recovering latent labeling information and supervising predictive model training alternatively to make the label distribution appropriate for the induced predictive model. Experimental results on PML datasets clearly validate the effectiveness of the proposed method for solving partial multilabel learning problems. In addition, further experiments show the high quality of the recovered label distributions and the effectiveness of adopting label distributions for partial multilabel learning.**

存在的问题

通过交替监督，使模型逐步改进

***Index Terms*— Label distribution, label distribution learning (LDL), label enhancement (LE), partial multilabel learning.**

I. Introduction

PARTIAL multilabel learning (PML) aims to learn a multilabel predictive model from inaccurate supervised data, among which each training example is associated with a candidate label set that is partially valid. For example, in online object tagging (Fig. 1), only a subset of the candidate labels is valid due to the unreliable or irresponsible annotators. In recent years, partial multilabel learning techniques have been widely adopted in real-world applications with inaccurate supervision [1]–[6].

Manuscript received 24 May 2021; revised 29 August 2021; accepted 1 November 2021. Date of publication 16 November 2021; date of current version 4 August 2023. This work was supported in part by the National Key Research and Development Plan of China under Grant 2018AAA0100104 and Grant 2018AAA0100100, in part by the National Science Foundation of China under Grant 62125602 and Grant 62076063, in part by the Jiangsu Province Science Foundation for Youths under Grant BK20210220, and in part by the China Postdoctoral Science Foundation under Grant 2021M700023. *(Corresponding author: Xin Geng.)*

The authors are with the School of Computer Science and Engineering and the Key Laboratory of Computer Network and Information Integration (Ministry of Education), Southeast University, Nanjing 211189, China (e-mail: xning@seu.edu.cn; yunpengliu@seu.edu.cn; zhang_yan@seu.edu.cn; xgeng@seu.edu.cn).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TNNLS.2021.3125366.

Digital Object Identifier 10.1109/TNNLS.2021.3125366

Fig. 1. In object annotation, only three of the candidate labels are valid ones (in red) including *house*, *mountain*, and *tree*.

Formally, let $\mathcal{X} = \mathbb{R}^q$ denote the q-dimensional feature space and $\mathcal{Y} = \{y_1, y_2, \ldots, y_c\}$ denote the output space with c possible class labels. Given the partial multilabel learning training set $\mathcal{D} = \{(\boldsymbol{x}_i, Y_i) \mid 1 \leq i \leq n\}$, the task of partial multilabel learning is inducing a multilabel predictive model $f : \mathcal{X} \mapsto 2^{\mathcal{Y}}$ from the training set. Here, $\boldsymbol{x}_i \in \mathcal{X}$ denotes a q-dimensional feature vector and $Y_i \subseteq \mathcal{Y}$ is the candidate label set corresponding to each $\boldsymbol{x}_i$. Partial multilabel learning takes the key assumption that the ground-truth labels $\widetilde{Y}_i$ corresponding to $\boldsymbol{x}_i$ exist in its candidate label set Y_i, that is, $\widetilde{Y}_i \subseteq Y_i$, and therefore cannot be directly accessed by the common learning approach. Intuitively, the basic strategy for coping with the PML problem is disambiguation, that is, identifying the valid labels in each candidate label set. One recent attempt is utilizing the confidence degree of each candidate label to be the valid one [6]. Nonetheless, as the confidence ignores the irrelevance of the non-candidate labels, it could be error-prone especially with a label set that contains a high proportion of false-positive labels. The low-rank assumption is adopted to identify the noisy labels for disambiguation [4], [5]. For credible label elicitation techniques, the valid labels are identified from the candidate ones to make final prediction on unseen instance [1], [2]. Noisy label identification is proposed to simultaneously recover the ground-truth label and identify the noisy labels [7].

特征空间到标签空间的映射

low-rank 可以将原数据矩阵分解

In order to handle the ambiguity in partial multilabel learning, we assume that there is a real-valued description degree $d_{\boldsymbol{x}_i}^{y_j}$ associated with each label y_j. Here, the vector $\boldsymbol{d}_i = [d_{\boldsymbol{x}_i}^{y_1}, d_{\boldsymbol{x}_i}^{y_2}, \ldots, d_{\boldsymbol{x}_i}^{y_c}]^\top$ constituted by the description degrees of all the labels is called *label distribution* [8]. Note that label distribution may be the essential labeling information for PML, which describes the label ambiguity in each example more comprehensively in mainly two aspects.

假定真值向量表示标签关联程度

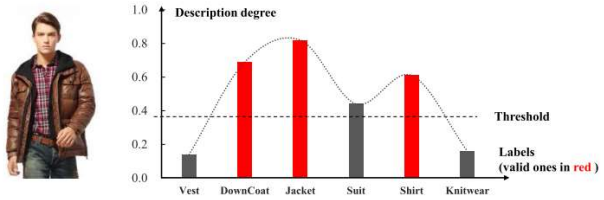

Fig. 2. Example about the differentiation between candidate labels and noncandidate labels in PML.

与实例相关但程度不同，如jacket 0.8 shirt 0.6

1) The relevance among relevant labels is diverse rather than identical. For instance, in Fig. 2, the image is annotated with the relevant label set {*Jacket*, *DownCoat*, *Shirt*}, but the relevance of each label to the image is significantly different.
2) The boundary between relevant labels and irrelevant labels is not distinct, which results in some irrelevant labels being annotated as candidate labels. For example, in Fig. 2, the threshold chosen by an unreliable annotator leads to the candidate label set {*Jacket*, *DownCoat*, *Shirt*, *Suit*} where *Suit* is not valid.

相关和不相关区分度不够

Therefore, the latent label distribution is the essential supervision in partially multilabeled data and worth to be leveraged for predictive model training. Although label distributions are not explicitly available in partial multilabel training sets, they can be somehow recovered. Accordingly, a novel partial multilabel learning algorithm named PENAD, that is, *Progressive ENhancement of lAbel Distributions for partial multilabel learning*, is proposed in this article. Specifically, the label distribution is recovered by considering the observation model for logical labels and the sharing topological structure from feature space to label distribution space. In addition, the latent label distribution is progressively enhanced by recovering latent labeling information and supervising predictive model training alternatively to make the label distribution appropriate for the induced predictive model.

Preliminary results of this article have been reported in a shorter conference version [9]. While the conference version recovers the label distributions and trains the predictive model in separated stages and only the topological information of the feature space, here we consider the observation model and the predictive model to enhance latent label distribution progressively. Moreover, we formulate the label enhancement (LE) and predictive model training into a unified framework. The recovery experiments on label distribution datasets are conducted to show that PENAD is effective to recover inherent label distributions in partial multilabel examples. Besides, the ablation experiments are conducted to validate the usefulness of recovered latent label distributions for improving the performance and the superior performance of PENAD against the conference version.

The rest of this article is organized as follows. First, some related works are briefly reviewed in Section II. Second, technical details of the proposed approach are presented in Section III. Third, the results of the extensive experiments are reported in Section IV. Finally, conclusion is shown in Section V.

II. Related Work

Two popular learning frameworks which are closely related to partial multilabel learning, namely *multilabel learning* [10]–[12] and *partial label learning* (PLL) [13], [14], are first reviewed.

In multi-label learning (MLL), each example is associated with a number of valid labels simultaneously. Multilabel learning approaches could be roughly divided into three types with the order of the correlations among labels [10] utilized for training predictive models. The simplest one is the first-order type, which disassembles the MLL problem into a number of binary classification problems [15], [16]. However, these approaches neglect the useful information of one label for another label in learning process. The second-order approaches consider the label correlations between pairs of labels [17], [18]. Nonetheless, the second-order approaches [17], [18] only consider the difference between relevant labels and irrelevant labels. The high-order approaches further focus on the label correlations among label set [19]–[22]. Both MLL and PML aim to induce a multilabel predictive model which would predict a proper set of labels for unseen instances. Compared to MLL, the task of PML is more challenging because the valid labels in candidate label sets are not directly available for learning algorithms in PML.

In PLL, a candidate label set is corresponding to each example and only one of them is valid. Therefore, the task of PLL is to learn a single-label predictive model that could predict one proper label for unseen instance. Existing PLL methods handle the PLL problem by disambiguation [13], [23]. To handle the disambiguation, progressive learning-based methods are proposed to identify the ground-truth label progressively in the learning process [24]–[26]. Both PML and PLL learn from training examples with a candidate label set that contains false-positive labels. However, PML is more difficult than PLL because PML needs to induce a multilabel predictive model rather than a single-label predictive model.

In order to handle partial multilabel learning problem, the straightforward strategy is treating all candidate labels as valid ones. Thereafter, the desired multilabel predictor can be induced by adopting any off-the-shelf MLL algorithms. Nevertheless, it is definitely seen that the performance of the strategy would be affected negatively by the candidate labels which are false-positive ones. In addition, another strategy to facilitate the PML problem is identifying the ground-truth ones in the candidate label sets. One recent attempt is utilizing the confidence degree of each candidate label to be the valid one [6]. Nonetheless, as the confidence degree ignores the irrelevance of the non-candidate labels, it could be error-prone especially with a label set that contains a high scale of false-positive labels. Low-rank assumption is adopted to identify the noisy labels for disambiguation [4], [5]. For credible label elicitation techniques, the valid labels are identified from each candidate label set to make final prediction on unseen instance [1], [2]. Noisy label identification is proposed to simultaneously recover the ground-truth label and identify the noisy labels [7].

方式1：假定所有候选标签都有效

方式2：识别有效标签

To deal with PML problem, we adopt label distribution [8] to comprehensively describe the label ambiguity in the PML

本文采用标签分布

datasets. Label distribution is a real-valued vector which can explicitly describe label ambiguity with the description degree. The learning process on the examples associated with label distributions is therefore called *label distribution learning* (LDL) [8]. According to the theoretical analysis [27], LDL is approximate to the optimal classifier via learning on the instances labeled by the ground-truth label distributions. There are several algorithms [8] designed for LDL, and the specialized algorithm is proposed by applying maximum entropy model with Kullback–Leibler divergence as loss function to learn the label distribution. LDL has been successfully utilized in many real applications, such as head pose estimation [28], facial landmark detection [29], zero-shot learning [30], age estimation [31], [32], and emotion analysis from texts [33].

However, label distributions are not explicitly available in most training sets. To overcome this situation, a process named *LE* [34] is defined for recovering label distributions from training datasets. Many LE approaches [35]–[38] are proposed to deal with MLL and LDL problems. In addition, some approaches with similar functionality for LE have been proposed [39]–[41], while there is no explicit concept of LE defined in the existing work. Most LE approaches are commonly treated as an independent stage and output label distributions for subsequent MLL or LDL straining stage, which neglects to generate a more proper label distribution for subsequent model induction. In addition, these LE approaches rely on multiple valid labels but cannot deal with the false-positive labels in PML.

In Section III, a novel partial multilabel learning approach will be introduced. The label distributions are recovered from the PML examples by partial multi-LE. In addition, partial multilabel model induction is proposed to induce a predictive model via recovered latent label distribution. Different from common LE-based strategies which treat LE as an independent stage, the PENAD approach considers LE and predictive model induction in a unified framework.

III. Proposed Approach

Formally, let $\mathcal{X} = \mathbb{R}^q$ denote the q-dimensional feature space and $\mathcal{Y} = \{y_1, y_2, \ldots, y_c\}$ denote the output space with c possible class labels. Given the partial multilabel learning training set $\mathcal{D} = \{(\boldsymbol{x}_i, Y_i) \mid 1 \le i \le n\}$, the task of partial multilabel learning is inducing a multilabel predictive model $f: \mathcal{X} \mapsto 2^{\mathcal{Y}}$ from the training set. Here, $\boldsymbol{x}_i \in \mathcal{X}$ denotes a q-dimensional feature vector and $Y_i \subseteq \mathcal{Y}$ is the candidate label set corresponding to each $\boldsymbol{x}_i$. $\boldsymbol{l}_i = [l_i^{y_1}, l_i^{y_2}, \ldots, l_i^{y_c}]^\top \in \{0, 1\}^c$ denotes the c-dimensional observed logical label vector of $\boldsymbol{x}_i$, i.e., $l_{\boldsymbol{x}_i}^{y_j} = 1$ if $y_j \in Y$, otherwise $l_{\boldsymbol{x}_i}^{y_j} = 0$. $\boldsymbol{d}_i = [d_i^{y_1}, d_i^{y_2}, \ldots, d_i^{y_c}]^\top \in [0, 1]^c$ denotes the c-dimensional label distribution. Then $\mathbf{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n]$, $\mathbf{L} = [\boldsymbol{l}_1, \boldsymbol{l}_2, \ldots, \boldsymbol{l}_n]$, and $\mathbf{D} = [\boldsymbol{d}_1, \boldsymbol{d}_2, \ldots, \boldsymbol{d}_n]$ represent the feature matrix, logical label matrix, and label distribution matrix, respectively. $\boldsymbol{\Theta}$ represents the predictive model

$$\forall_{j=1}^{c}: f(y_j \mid \boldsymbol{x}_i, \boldsymbol{\Theta}) = \hat{\boldsymbol{\theta}}_j^\top \varphi(\boldsymbol{x}_i) + b_j = \boldsymbol{\theta}_j^\top \boldsymbol{\phi}_i \tag{1}$$

where $\varphi(\boldsymbol{x})$ is a nonlinear transformation of $\boldsymbol{x}$ to a higher-dimensional feature space. For convenient describing, we let $\boldsymbol{\theta}_j = [\hat{\boldsymbol{\theta}}_j^\top, b_j]^\top$, $\boldsymbol{\phi}_i = [\varphi(\boldsymbol{x}_i)^\top, 1]^\top$, and $\Theta = [\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_c]$.

A. PENAD Framework

Note that the previous conference version [9] recovers the label distributions and trains the predictive model in separated stages and only the topological information of the feature space. Here, the PENAD approach not only focuses on the latent label distribution in PML example, but also considers the following assumptions about latent label distributions: 1) the logical label is observed from the latent label distribution; 2) the local topological structure of the label distribution space should be consistent with the feature space; and 3) the label distribution is appropriate for the predictive model induction. Thus, PENAD fuses the observation model and the predictive model to enhance label distribution progressively and formulates the LE and predictive model training into a unified framework. Then, latent label distribution $\mathbf{D}$ and the predictive model $\boldsymbol{\Theta}$ induction in PENAD are coupled by minimizing the following objective function:

$$\min_{\mathbf{D}, \boldsymbol{\Theta}} \quad \mathcal{L}(\boldsymbol{\Theta}, \mathbf{D}) + \lambda \mathcal{R}(\mathbf{D}). \tag{2}$$

PENAD not only considers the label distribution regularization, but also adjusts the label distributions to facilitate predictive model induction. On the other hand, PENAD handles the structural risk minimization with the latent label distributions.

Motivated by the assumption that the logical label vector $\boldsymbol{l}_i$ in PML data is the observed vector from the latent label distributions $\boldsymbol{d}_i$, we propose an observation factor $\boldsymbol{\delta} = [\delta_1, \delta_2, \ldots, \delta_n]^\top$ to generate the candidate labels from the latent label distribution by $\hat{\boldsymbol{l}}_i = \mathcal{B}(\delta_i \boldsymbol{d}_i)$. Here, $\mathcal{B}(\cdot)$ is a function to binarize each element in a real-value vector into $\{0, 1\}$. Therefore, the observation regularization is proposed as follows:

$$\mathcal{R}_1(\mathbf{D}, \boldsymbol{\delta}) = \sum_{i=1}^{n} \|\boldsymbol{l}_i - \mathcal{B}(\delta_i \boldsymbol{d}_i)\|^2. \tag{3}$$

To make the function differential, we employ the sigmoid function $\mathcal{S}(\cdot)$ instead of the binarization function $\mathcal{B}(\cdot)$.

We assume that $\mathbf{A} = [a_{ij}]_{n \times n}$ denotes the weight matrix in an affinity graph $G = (\mathcal{V}, \mathcal{E}, \mathbf{A})$ of features, where $\mathcal{V}$ and $\mathcal{E}$ correspond to the vertex set consisting of feature vectors and the edge set. The local topological structure of feature space could be transferred to label distribution space with smoothness assumption [42], which leads to the following error reconstruction function:

$$\mathcal{R}_2(\mathbf{D}) = \|\mathbf{D} - \mathbf{D}\mathbf{A}\|_F^2. \tag{4}$$

Here, the weight matrix $\mathbf{A} = [a_{ij}]_{n \times n}$ could be generated by modeling the relationship between k-nearest examples via feature space reconstruction

$$\begin{aligned} \min_{\mathbf{A}} \quad & \sum_{j=1}^{n} \Big|\Big|\boldsymbol{x}_j - \sum_{i \in \mathcal{N}(\boldsymbol{x}_j)} a_{ij} \cdot \boldsymbol{x}_i\Big|\Big|^2 \\ \text{s.t.} \quad & \sum_{i \in \mathcal{N}(\boldsymbol{x}_j)} a_{ij} = 1 \ (1 \le j \le n) \\ & a_{ij} \ge 0 \ \big(\forall\, i \in \mathcal{N}(\boldsymbol{x}_j)\big) \\ & a_{ij} = 0 \ \big(\forall\, i \notin \mathcal{N}(\boldsymbol{x}_j)\big) \end{aligned} \tag{5}$$

通过标签增强来恢复标签分布

LE作为一个单独的阶段，忽略了为后续模型归纳生成更合适的标签分布

LE无法解决大量假正例样本

logical label 矩阵值是通过判断标签是否属于实例获得

X 特征矩阵，L 标签矩阵，D 标签分布矩阵

θ j

假设逻辑标签值来自标签分布

标签分布在经过图结构平滑后与原始标签分布的差异

where $\mathcal{N}(\boldsymbol{x}_j)$ denotes the index set of k-nearest neighbors identified for $\boldsymbol{x}_j$ in $\mathcal{D}$. The resulting problem (5) corresponds to a quadratic programming (QP) problem.

Then, PENAD considers a least-square loss that measures how the prediction fit the latent label distribution

$$\mathcal{L}_1(\mathbf{D}, \boldsymbol{\Theta}) = \|\boldsymbol{\Theta}^\top \boldsymbol{\Phi} - \mathbf{D}\|_F^2 \tag{6}$$

where $\boldsymbol{\Phi} = [\boldsymbol{\phi}_1, \boldsymbol{\phi}_2, \ldots, \boldsymbol{\phi}_n]$. As the average value corresponding to non-candidate labels should be less than the average value corresponding to candidate labels [13], [43], an additional partial multilabel loss is designed as follows:

$$\mathcal{L}_2(\boldsymbol{\Theta}) = \mathrm{tr}\big(\mathbf{P}^\top \boldsymbol{\Theta}^\top \boldsymbol{\Phi}\big). \tag{7}$$

Here, $\mathbf{P} = [\boldsymbol{p}_1, \boldsymbol{p}_2, \ldots, \boldsymbol{p}_n]$ and $\boldsymbol{p}_i = [p_{\boldsymbol{x}_i}^{y_1}, p_{\boldsymbol{x}_i}^{y_2}, \ldots, p_{\boldsymbol{x}_i}^{y_c}]^\top$ is calculated to measure the difference between each average values

$$\forall_{j=1}^{c}: \quad p_{\boldsymbol{x}_i}^{y_j} = \begin{cases} -\dfrac{1}{|Y_i|}, & \text{if } y_j \in Y_i \\ \dfrac{1}{|\hat{Y}_i|}, & \text{if } y_j \notin Y_i \end{cases} \tag{8}$$

where Y_i denotes the candidate label set corresponding to $\boldsymbol{x}_i$ and its complementary set is denoted as $\hat{Y}_i$.

Finally, the optimization problem in (2) can be rewritten as

$$\begin{aligned} \min_{\mathbf{D},\boldsymbol{\Theta},\boldsymbol{\delta}} \quad & \beta_1 \|\boldsymbol{\Theta}^\top \boldsymbol{\Phi} - \mathbf{D}\|_F^2 + \beta_2 \mathrm{tr}\big(\mathbf{P}^\top \boldsymbol{\Theta}^\top \boldsymbol{\Phi}\big) + \|\boldsymbol{\Theta}\|_F^2 \\ & + \lambda_1 \sum_{i=1}^{n} \|\boldsymbol{l}_i - \mathcal{S}(\delta_i \boldsymbol{d}_i)\|^2 + \lambda_2 \|\mathbf{D} - \mathbf{D}\mathbf{A}\|_F^2 \\ \text{s.t.} \quad & 0 \le d_{\boldsymbol{x}_i}^{y_j} \le 1 \quad \forall 1 \le i \le n, 1 \le j \le c. \end{aligned} \tag{9}$$

A virtual label bipartition is employed to predict a proper label set for unseen instance $\boldsymbol{x}$ with the predictive model in PENAD. Specifically, an extra virtual label y_0 is added to serve as a threshold to bipartite the labels into relevant labels and irrelevant labels. Then the partial multilabel set $\mathcal{Y}$ is expanded to $\mathcal{Y}' = \mathcal{Y} \cup \{y_0\} = \{y_0, y_1, \ldots, y_c\}$. In this article, $l_{\boldsymbol{x}}^{y_0}$ is set to be 0.5. Once the latent label distributions and the predictive model have been learned on the extended the partial multilabel set, the predicted label set corresponding to the test instance $\boldsymbol{x}$ can be predicted. Let $\boldsymbol{\Theta}^* = [\boldsymbol{\theta}_0^*, \boldsymbol{\theta}_1^*, \ldots, \boldsymbol{\theta}_c^*]$ be the final predictive model after optimization, the outputs on each class y_j $(1 < j < c)$ and y_0 are

$$\forall_{u=0}^{c}: \ f(y_u|\boldsymbol{x}) = \boldsymbol{\theta}_u^{*\top} \boldsymbol{\phi}. \tag{10}$$

Then, the predicted labels corresponding to $\boldsymbol{x}$ are determined via splitting the outputs

$$\zeta(\boldsymbol{x}) = \big\{ y_j \mid f\big(y_j|\boldsymbol{x}\big) > f(y_0|\boldsymbol{x}), 1 \le j \le c \big\}. \tag{11}$$

B. Optimization

The optimization problem in (9) can be solved in an alternating way. When $\boldsymbol{\Theta}$ and $\boldsymbol{\delta}$ are fixed, (9) can be reduced as

$$\begin{aligned} \min_{\mathbf{D}} \quad & \beta_1 \|\boldsymbol{\Theta}^\top \boldsymbol{\Phi} - \mathbf{D}\|_F^2 + \lambda_1 \sum_{i=1}^{n} \|\boldsymbol{l}_i - \mathcal{S}(\delta_i \boldsymbol{d}_i)\|^2 \\ & + \lambda_2 \|\mathbf{D} - \mathbf{D}\mathbf{A}\|_F^2 \\ \text{s.t.} \quad & 0 \le d_{\boldsymbol{x}_i}^{y_j} \le 1 \quad \forall 1 \le i \le n, 1 \le j \le c. \end{aligned} \tag{12}$$

Note that (12) corresponds to a standard constrained QP process whose computational complexity would be demanding when $n \times c$ is large. Inspired by [43], we can choose to solve it with alternating optimization strategy where a series of QP subproblems, where each label distribution $\boldsymbol{d}_i$ is optimized by fixing the values of other label distributions $\boldsymbol{d}_j (j \ne i)$

$$\begin{aligned} \min_{\boldsymbol{d}_i} \quad & \beta_1 \|\boldsymbol{\Theta}^\top \boldsymbol{\phi}_i - \boldsymbol{d}_i\|^2 + \lambda_1 \|\boldsymbol{l}_i - \mathcal{S}(\delta_i \boldsymbol{d}_i)\|^2 \\ & + \lambda_2 \|\boldsymbol{d}_i - \sum_{j=1}^{n} a_{ij} \boldsymbol{d}_j\|^2 \\ \text{s.t.} \quad & 0 \le d_{\boldsymbol{x}_i}^{y_j} \le 1 \quad \forall 1 \le i \le n, 1 \le j \le c. \end{aligned} \tag{13}$$

When $\boldsymbol{\Theta}$ and $\mathbf{D}$ are fixed, (9) can be reduced as

$$\min_{\boldsymbol{\delta}} \quad \sum_{i=1}^{n} \|\boldsymbol{l}_i - \mathcal{S}(\delta_i \boldsymbol{d}_i)\|^2 \tag{14}$$

which can be solved by BFGS with the first-order gradient

$$\begin{aligned} \nabla_{\delta_i} = \sum_{j=1}^{c} \big(& -2\mathcal{S}^3\big(\delta_i d_i^{y_j}\big) + 2 d_i^{y_j} \mathcal{S}^2\big(\delta_i d_i^{y_j}\big) \\ & + 2 l_i^{y_j} d_i^{y_j} \mathcal{S}^2\big(\delta_i d_i^{y_j}\big) - 2 l_i^{y_j} d_i^{y_j} \mathcal{S}\big(\delta_i d_i^{y_j}\big) \big). \end{aligned} \tag{15}$$

When $\mathbf{D}$ and $\boldsymbol{\delta}$ are fixed, (9) can be reduced as

$$\min_{\boldsymbol{\Theta}} \quad \beta_1 \|\boldsymbol{\Theta}^\top \boldsymbol{\Phi} - \mathbf{D}\|_F^2 + \beta_2 \mathrm{tr}\big(\mathbf{P}^\top \boldsymbol{\Theta}^\top \boldsymbol{\Phi}\big) + \|\boldsymbol{\Theta}\|_F^2 \tag{16}$$

which can be solved by BFGS with the first-order gradient

$$\nabla_{\boldsymbol{\Theta}} = 2\beta_1 \boldsymbol{\Phi}\boldsymbol{\Phi}^\top \boldsymbol{\Theta} + 2\boldsymbol{\Theta} - 2\beta_1 \boldsymbol{\Phi}\mathbf{D}^\top + \beta_2 \boldsymbol{\Phi}\mathbf{P}^\top. \tag{17}$$

According to the representor's theorem [44], a learning problem can be expressed as a linear combination of the training examples in the feature space under fairly general conditions, that is, $\boldsymbol{\theta}^j = \sum_i \boldsymbol{\eta}^j \varphi(\boldsymbol{x}_i)$. This expression can be replaced into (1) and (9), which will generate the inner product $< \varphi(\boldsymbol{x}_i), \varphi(\boldsymbol{x}_j) >$ and apply the kernel trick.

To summarize, Table I gives the pseudo-code of PENAD. In the training stage, label distributions are recovered from observed logical labels and a predictive model is learned via alternating optimization (steps 1–9). In the prediction stage, predictive result on the unseen instance is made via virtual label bipartition and the resulting model (steps 10–11).

C. Complexity Analysis

In this part, we discuss the time complexity of our method, which includes three aspects, that is, the weight matrix $\mathbf{A}$ learning procedure, the initialization procedure, and the optimization procedure. First, the time cost of learning $\mathbf{A}$ is $O(q^2n^2 + T_1 n^2)$. Note that an iterative strategy is employed to solve the optimization problem in Eq. (5), and the iteration number is T_1. Second, the total time cost of initializing $\mathbf{D}^{(0)}$, $\boldsymbol{\Theta}^{(0)}$, and $\boldsymbol{\delta}^{(0)}$ is $O(cn+cm+n)$, where m is the feature dimension after the nonlinear transformation. Third, the time cost of the optimization procedure is $O(t(c^2n^2 + T_2 cn + T_3(cmn + cm^2 + m^2 n)))$, where t is the iteration number of the whole optimization procedure. Specifically, the time consumption of updating $\mathbf{D}$ in each iteration is $O(c^2n^2)$. The time consumption of updating $\boldsymbol{\delta}$ by BFGS is $O(T_2 cn)$, where T_2 is the iteration

拟牛顿法求解

解决Fig2中问题

TABLE I
PSEUDOCODE OF PENAD

Inputs:
$\mathcal{D}$: the partial multi-label training set
$\lambda_1, \lambda_2, \beta_1, \beta_2$: the hyper-parameters
$\boldsymbol{x}$: the unseen instance
Outputs:
Y: the predicted label set for $\boldsymbol{x}$
Process:
Generate the affinity graph $G = (\mathcal{V}, \mathcal{E}, \mathbf{A})$;
$\mathbf{A}$ is obtained by solving optimization problem (5) ;
Initialize $t = 0$, $\mathbf{D}^{(0)}$, $\boldsymbol{\Theta}^{(0)}$ and $\boldsymbol{\delta}^{(0)}$;
repeat
  Update $\mathbf{D}^{(t+1)}$ by adopting quadratic programming subproblems Eq. (13);
  Update $\boldsymbol{\delta}$ by BFGS with Eq. (15) ;
  Update $\boldsymbol{\Theta}^{(t+1)}$ by BFGS with Eq. (17);
  $t = t + 1$;
until convergence
The final predictive model is obtained by setting $\boldsymbol{\Theta}^* = \boldsymbol{\Theta}^{(t)}$;
Return a proper label set Y according to Eq. (11).

number of BFGS in this step. In each iteration, the time cost of updating $\boldsymbol{\Theta}$ by BFGS is $O(T_3(cmn + cm^2 + m^2n))$, where T_3 is the iteration number of BFGS. Therefore, the whole time complexity of our method is $O(q^2n^2 + T_1n^2 + cn + cm + n + t(c^2n^2 + T_2cn + T_3(cmn + cm^2 + m^2n)))$.

IV. EXPERIMENTS

A. Experimental Setup

1) Datasets: In order to thoroughly evaluate the performance of the proposed approach, a lot of synthetic and real-world partial multilabel datasets have been utilized for experimental studies. Table II shows statistics of these datasets.

Specifically, the synthetic datasets are all generated from multilabel datasets via inserting noise into labels. Some of the irrelevant labels corresponding to each example are randomly chosen to become the candidate labels along with the relevant labels. Table II shows seven benchmark multilabel datasets [10] which are utilized to generate synthetic partial multilabel datasets, including `emotions`, `image`, `scene`, `eurlex_sm`, `yeast`, `msra`, and `computer`. For each dataset, we vary the average number of candidate labels to constitute different configuration. Accordingly, 24 synthetic datasets are generated. In addition, four real-world partial multilabel datasets, that is, `music_emotion`, `yeastBP`, `music_style`, and `mirflickr` [45] are employed in the experiments. For the real-world datasets, the candidate label sets are collected from web, and valid labels are further checked by humans.

2) Methodology: PENAD is compared against six state-of-the-art partial multilabel approaches, each of which is configured with the hyperparameters suggested in respective literature except FPML since we have properly tuned these approaches and found a better configuration for FPML.

1) PML-FP [6] which optimizes labeling confidence and predictive model alternatively with feature prototypes [suggested configuration: $C_1 = 1$, C_2 with $\{1, 2, \ldots, 10\}$, C_3 with $\{1, 10, 100\}$].
2) FPML [5] which adopts noisy labels estimation to learn from partial multilabel examples via low-rank approximation [configuration: $\lambda_1 = 5$, $\lambda_2 = 1$, $\lambda_3 = 10$ (suggested configuration: $\lambda_1 = 1$, $\lambda_2 = 1$, $\lambda_3 = 10$)].
3) PARTICLE-VLS [2] which adopts credible label elicitation technique to learn from partial multilabel examples and virtual label splitting for predictive model induction [suggested configuration: $k = 10$, $\alpha = 0.95$, $thr = 0.9$].
4) PARTICLE-MAP [2] which adopts credible label elicitation technique to learn from partial multilabel examples and maximum *a posteriori* (MAP) reasoning for predictive model induction [suggested configuration: $k = 10$, $\alpha = 0.95$, $thr = 0.9$].
5) PML-LRS [4] which adopts low-rank and sparse decomposition scheme to learn from partial multilabel examples [suggested configuration: $\eta = 1$, $\gamma = 0.1$, $\beta = 1$].
6) DRAMA [46] which generates a real-valued label confidence matrix under the guidance of feature manifold and the candidate label set [suggested configuration: $\delta_1 = 0.01$, $\delta_2 = 0.5$, $k = 10$].

For PENAD, the parameter k, λ_1, λ_2, β_1, and β_2 are fix to 10, 0.01, 0.01, 1, and 10, respectively. The kernel function in PENAD is Gaussian kernel. We perform fivefold cross-validation on each dataset, where mean values and standard deviations are recorded.

3) Evaluation Metrics: Five popular multilabel metrics, that is, *Hamming loss*, *One-error*, *Coverage*, *Ranking loss*, and *Average precision* [10] are adopted for performance evaluation.

Conceptually, *Hamming loss* evaluates the fraction of the misclassified instance-label pairs, that is, a relevant label is missed or an irrelevant label is predicted, *One-error* evaluates the fraction of the instances whose top-ranked label is irrelevant, *Coverage* evaluates the average number of steps to move down the ranked label list of an instance so as to cover its relevant label set, *Ranking loss* evaluates the average fraction of the label pairs among which an irrelevant label is ranked higher than its relevant one, and *Average precision* evaluates the average fraction of relevant labels which are ranked higher than a particularly relevant one.

Note that for all metrics, the values vary between [0,1]. Furthermore, the *larger* the value of *average precision*, the better the performance. While for *Hamming loss*, *One-error*, *Coverage*, and *Ranking loss*, the *smaller* the values, the better the performance. The metrics could be adopted as well indicators for comprehensive studies, since the five metrics evaluate the performance of learned models in different aspects.

B. Experimental Results

Tables III–VII show the results of all approaches on all metrics. For brevity, the results on some synthetic configurations are given in each synthetic PML dataset, that is, avg. #CLs being 3 and 5 for `emotions`, #CLs being 3 and 5 for `scene`, 9 and 13 for `yeast`, 6 and 14 for `eurlex_sm`, 9 and 17 for `msra`, and 5 and 13 for `computer`.

Furthermore, *Friedman test* [47] is employed here to analyze the relative performance of the PML approaches.

TABLE II
STATISTICS OF THE DATASETS USED IN THE EXPERIMENTS

Dataset	#Examples	#Features	#Labels	avg. #CLs	avg. #GLs
emotions	593	72	6	3, 5	1.86
msra	1,868	898	19	9, 11, 13, 15, 17	6.31
image	2,000	294	5	2, 4	1.23
scene	2,407	294	6	3, 5	1.07
yeast	2,417	103	14	9, 11, 13	4.23
computer	11,235	880	25	5, 7, 9, 11, 13	1.51
eurlex_sm	12,679	100	15	6, 8, 10, 12, 14	1.53
yeastBP	560	5548	217	30.43	21.56
music_emotion	6,833	98	11	5.29	2.42
music_style	6,839	98	10	6.04	1.44
mirflickr	10,433	100	7	3.35	1.77

TABLE III
EXPERIMENTAL RESULTS (MEAN±STD) MEASURED BY *Ranking loss*. THE BEST PERFORMANCE (THE SMALLER THE BETTER) IS SHOWN IN BOLD FACE

Data Set	avg.#CLS	PENAD	PML-FP	PARTICLE-VLS	PARTICLE-MAP	PML-LRS	FPML	DRAMA
emotions	3	**0.163±0.016**	0.196±0.017	0.181±0.019	0.172±0.018	0.200±0.013	0.202±0.016	0.254±0.016
	5	**0.229±0.018**	0.280±0.018	0.269±0.034	0.252±0.036	0.255±0.024	0.253±0.061	0.324±0.016
image	2	**0.155±0.018**	0.192±0.017	0.193±0.019	0.172±0.018	0.186±0.016	0.180±0.017	0.263±0.020
	4	**0.196±0.008**	0.258±0.014	0.262±0.016	0.236±0.014	0.245±0.014	0.239±0.011	0.280±0.015
scene	3	**0.078±0.008**	0.151±0.011	0.119±0.006	0.104±0.010	0.125±0.012	0.100±0.008	0.205±0.014
	5	**0.119±0.014**	0.248±0.020	0.233±0.017	0.195±0.013	0.204±0.019	0.175±0.021	0.230±0.018
yeast	9	**0.181±0.007**	0.187±0.008	0.192±0.006	0.214±0.008	0.372±0.010	0.189±0.007	0.292±0.008
	13	**0.209±0.004**	0.261±0.004	0.244±0.003	0.245±0.007	0.430±0.006	0.241±0.003	0.334±0.010
eurlex_sm	6	**0.094±0.003**	0.141±0.004	0.109±0.001	0.111±0.004	0.131±0.002	0.198±0.007	0.177±0.005
	14	**0.190±0.004**	0.227±0.007	0.231±0.004	0.204±0.006	0.211±0.004	0.258±0.013	0.319±0.005
msra	9	**0.133±0.005**	0.149±0.005	0.159±0.002	0.170±0.005	0.160±0.006	0.147±0.003	0.252±0.005
	17	**0.181±0.005**	0.243±0.008	0.257±0.005	0.310±0.008	0.236±0.004	0.244±0.007	0.322±0.009
computer	5	**0.108±0.007**	0.373±0.183	0.129±0.007	0.240±0.007	0.200±0.012	0.161±0.007	0.204±0.005
	13	**0.107±0.006**	0.301±0.103	0.198±0.005	0.324±0.010	0.260±0.007	0.130±0.005	0.256±0.012
music_emotion	5.29	**0.234±0.006**	0.277±0.008	0.265±0.008	0.253±0.008	0.251±0.003	0.252±0.006	0.351±0.013
music_style	6.04	**0.134±0.004**	0.148±0.003	0.157±0.002	0.164±0.004	0.283±0.005	0.142±0.003	0.258±0.003
mirflickr	3.35	**0.094±0.030**	0.160±0.049	0.225±0.026	0.115±0.073	0.282±0.124	0.162±0.103	0.172±0.083
yeastBP	30.43	**0.256±0.028**	0.363±0.041	0.935±0.037	0.283±0.040	0.406±0.048	0.381±0.044	0.396±0.026

TABLE IV
EXPERIMENTAL RESULTS (MEAN±STD) MEASURED BY *One-error*. THE BEST PERFORMANCE (THE SMALLER THE BETTER) IS SHOWN IN BOLD FACE

Data Set	avg.#CLS	PENAD	PML-FP	PARTICLE-VLS	PARTICLE-MAP	PML-LRS	FPML	DRAMA
emotions	3	0.270±0.025	0.335±0.031	**0.224±0.040**	0.244±0.055	0.357±0.050	0.320±0.052	0.354±0.037
	5	0.331±0.042	0.426±0.040	**0.315±0.048**	0.348±0.066	0.398±0.052	0.406±0.025	0.445±0.045
image	2	0.292±0.024	0.358±0.015	**0.284±0.035**	0.317±0.041	0.345±0.023	0.351±0.029	0.415±0.013
	4	**0.358±0.023**	0.459±0.021	0.377±0.030	0.427±0.036	0.435±0.023	0.433±0.027	0.448±0.024
scene	3	**0.230±0.019**	0.382±0.021	0.240±0.021	0.286±0.019	0.320±0.020	0.277±0.022	0.389±0.027
	5	**0.330±0.032**	0.544±0.036	0.374±0.026	0.449±0.031	0.466±0.033	0.430±0.037	0.434±0.024
yeast	9	0.237±0.006	0.266±0.029	**0.229±0.006**	0.245±0.018	0.459±0.027	0.232±0.009	0.284±0.013
	13	0.254±0.011	0.355±0.021	**0.238±0.008**	0.257±0.012	0.602±0.027	0.271±0.005	0.305±0.010
eurlex_sm	6	0.264±0.006	0.361±0.005	**0.225±0.009**	0.268±0.010	0.367±0.005	0.477±0.032	0.343±0.006
	14	0.443±0.015	0.508±0.010	**0.399±0.011**	0.445±0.012	0.522±0.025	0.594±0.034	0.660±0.009
msra	9	**0.051±0.008**	0.062±0.006	0.055±0.011	0.106±0.024	0.083±0.013	0.068±0.013	0.120±0.018
	17	**0.078±0.011**	0.144±0.008	0.085±0.005	0.329±0.090	0.155±0.014	0.244±0.026	0.153±0.011
computer	5	**0.388±0.009**	0.710±0.101	0.390±0.015	0.736±0.021	0.463±0.017	0.458±0.005	0.512±0.007
	13	**0.432±0.004**	0.617±0.095	0.454±0.010	0.898±0.014	0.599±0.014	0.440±0.006	0.545±0.012
music_emotion	5.29	**0.430±0.018**	0.546±0.019	0.452±0.040	0.475±0.021	0.465±0.013	0.502±0.012	0.555±0.035
music_style	6.04	**0.334±0.007**	0.409±0.014	0.368±0.008	0.385±0.018	0.570±0.007	0.365±0.007	0.468±0.005
mirflickr	3.35	**0.107±0.046**	0.328±0.129	0.165±0.152	0.188±0.174	0.587±0.096	0.327±0.185	0.339±0.135
yeastBP	30.43	0.888±0.060	0.922±0.036	0.906±0.054	0.912±0.054	0.970±0.014	**0.884±0.043**	0.961±0.040

Table VIII shows the Friedman statistics F_F over all evaluation metrics along with the critical value at 0.05 significance level. Table VIII reports that the null hypothesis of indistinguishable performance among comparing approaches is rejected on all of the evaluation metrics across the 28 benchmark cases.

Bonferroni–Dunn test [47] is utilized as the post-hoc test to show whether PENAD have a significantly different

TABLE V

EXPERIMENTAL RESULTS (MEAN±STD) MEASURED BY *Hamming loss*. THE BEST PERFORMANCE (THE SMALLER THE BETTER) IS SHOWN IN BOLD FACE

Data Set	avg.#CLS	PENAD	PML-FP	PARTICLE-VLS	PARTICLE-MAP	PML-LRS	FPML	DRAMA
emotions	3	**0.182±0.012**	0.241±0.013	0.205±0.009	0.226±0.009	0.295±0.023	0.186±0.015	0.219±0.020
	5	**0.217±0.019**	0.271±0.013	0.344±0.034	0.264±0.020	0.355±0.013	0.221±0.019	0.236±0.013
image	2	**0.150±0.001**	0.190±0.006	0.175±0.012	0.179±0.017	0.380±0.011	0.173±0.011	0.200±0.003
	4	**0.180±0.006**	0.231±0.012	0.329±0.010	0.236±0.019	0.435±0.005	0.206±0.004	0.208±0.012
scene	3	**0.083±0.006**	0.139±0.006	0.117±0.002	0.114±0.005	0.351±0.005	0.097±0.009	0.142±0.009
	5	**0.118±0.012**	0.193±0.011	0.383±0.014	0.198±0.025	0.380±0.006	0.149±0.014	0.158±0.006
yeast	9	**0.137±0.001**	0.215±0.005	0.207±0.010	0.237±0.013	0.438±0.006	0.198±0.002	0.227±0.003
	13	**0.143±0.001**	0.268±0.004	0.697±0.004	0.232±0.006	0.465±0.006	0.192±0.001	0.269±0.005
eurlex_sm	6	0.070±0.001	0.084±0.001	**0.067±0.000**	0.078±0.002	0.085±0.001	0.140±0.001	0.082±0.001
	14	**0.091±0.002**	0.103±0.001	0.897±0.000	0.150±0.004	0.106±0.003	0.154±0.004	0.119±0.001
msra	9	**0.104±0.001**	0.116±0.002	0.117±0.002	0.145±0.001	0.194±0.003	0.188±0.002	0.143±0.003
	17	**0.104±0.001**	0.115±0.001	0.171±0.003	0.136±0.004	0.194±0.002	0.181±0.003	0.163±0.003
computer	5	0.068±0.001	0.103±0.012	**0.054±0.001**	0.124±0.003	0.072±0.003	0.099±0.001	0.074±0.001
	13	**0.060±0.001**	0.097±0.010	0.064±0.001	0.135±0.004	0.181±0.013	0.080±0.001	0.083±0.002
music_emotion	5.29	**0.123±0.002**	0.244±0.002	0.211±0.004	0.217±0.003	0.389±0.003	0.151±0.002	0.258±0.002
music_style	6.04	**0.111±0.003**	0.125±0.002	0.121±0.001	0.155±0.004	0.432±0.003	0.119±0.002	0.318±0.006
mirflickr	3.35	**0.046±0.013**	0.214±0.048	0.186±0.036	0.180±0.036	0.329±0.076	0.137±0.033	0.218±0.028
yeastBP	30.43	**0.031±0.001**	0.054±0.008	0.042±0.008	0.044±0.008	0.134±0.008	0.142±0.009	0.120±0.013

TABLE VI

EXPERIMENTAL RESULTS (MEAN±STD) MEASURED BY *Coverage*. THE BEST PERFORMANCE (THE SMALLER THE BETTER) IS SHOWN IN BOLD FACE

Data Set	avg.#CLS	PENAD	PML-FP	PARTICLE-VLS	PARTICLE-MAP	PML-LRS	FPML	DRAMA
emotions	3	**0.300±0.016**	0.323±0.022	0.307±0.019	0.318±0.020	0.320±0.015	0.339±0.020	0.383±0.019
	5	**0.364±0.015**	0.398±0.010	0.375±0.017	0.384±0.010	0.373±0.041	0.380±0.022	0.447±0.028
image	2	**0.178±0.015**	0.206±0.014	0.194±0.014	0.194±0.015	0.203±0.013	0.198±0.013	0.265±0.020
	4	**0.211±0.011**	0.261±0.017	0.242±0.018	0.242±0.017	0.250±0.016	0.244±0.012	0.278±0.015
scene	3	**0.079±0.006**	0.141±0.010	0.098±0.002	0.102±0.008	0.119±0.010	0.098±0.007	0.188±0.012
	5	**0.114±0.014**	0.223±0.016	0.179±0.010	0.178±0.011	0.187±0.018	0.160±0.019	0.208±0.015
yeast	9	**0.476±0.011**	0.489±0.014	0.477±0.007	0.549±0.016	0.636±0.014	0.491±0.012	0.627±0.011
	13	**0.515±0.008**	0.589±0.009	0.558±0.005	0.590±0.013	0.686±0.007	0.575±0.006	0.682±0.014
eurlex_sm	6	**0.148±0.003**	0.199±0.005	0.154±0.002	0.170±0.005	0.186±0.004	0.257±0.006	0.244±0.006
	14	**0.254±0.006**	0.296±0.009	0.263±0.004	0.270±0.006	0.276±0.004	0.321±0.013	0.390±0.007
msra	9	**0.546±0.011**	0.571±0.008	0.562±0.004	0.601±0.007	0.584±0.009	0.573±0.007	0.704±0.011
	17	**0.623±0.009**	0.691±0.005	0.692±0.007	0.749±0.011	0.679±0.008	0.686±0.007	0.777±0.009
computer	5	**0.160±0.010**	0.435±0.169	0.174±0.009	0.291±0.007	0.256±0.012	0.217±0.010	0.265±0.008
	13	**0.157±0.009**	0.365±0.094	0.246±0.008	0.374±0.010	0.316±0.006	0.184±0.007	0.320±0.010
music_emotion	5.29	**0.403±0.005**	0.435±0.005	0.410±0.006	0.418±0.010	0.410±0.006	0.413±0.007	0.530±0.011
music_style	6.04	**0.193±0.007**	0.204±0.006	0.199±0.006	0.224±0.006	0.338±0.004	0.199±0.006	0.209±0.001
mirflickr	3.35	0.239±0.021	0.255±0.047	0.283±0.058	**0.226±0.041**	0.346±0.142	0.264±0.042	0.276±0.055
yeastBP	30.43	0.471±0.045	0.499±0.072	0.800±0.075	**0.419±0.079**	0.604±0.038	0.623±0.041	0.710±0.050

TABLE VII

EXPERIMENTAL RESULTS (MEAN±STD) MEASURED BY *Average precision*. THE BEST PERFORMANCE (THE SMALLER THE BETTER) IS SHOWN IN BOLD FACE

Data Set	avg.#CLS	PENAD	PML-FP	PARTICLE-VLS	PARTICLE-MAP	PML-LRS	FPML	DRAMA
emotions	3	**0.804±0.021**	0.781±0.021	0.800±0.020	0.800±0.027	0.757±0.021	0.760±0.025	0.728±0.020
	5	**0.743±0.025**	0.708±0.025	0.717±0.026	0.724±0.041	0.714±0.022	0.716±0.017	0.662±0.020
image	2	**0.811±0.018**	0.769±0.013	0.790±0.024	0.789±0.024	0.776±0.016	0.777±0.019	0.717±0.014
	4	**0.768±0.013**	0.701±0.014	0.721±0.015	0.723±0.018	0.718±0.015	0.720±0.015	0.698±0.014
scene	3	**0.863±0.011**	0.762±0.015	0.830±0.009	0.826±0.013	0.801±0.015	0.832±0.013	0.736±0.017
	5	**0.780±0.020**	0.644±0.024	0.703±0.012	0.712±0.019	0.699±0.024	0.730±0.026	0.705±0.015
yeast	9	**0.747±0.006**	0.738±0.011	0.744±0.007	0.722±0.007	0.558±0.008	0.743±0.006	0.644±0.005
	13	**0.712±0.004**	0.651±0.004	0.704±0.003	0.688±0.001	0.475±0.005	0.687±0.003	0.603±0.010
eurlex_sm	6	**0.779±0.004**	0.695±0.004	0.777±0.005	0.762±0.007	0.700±0.005	0.603±0.018	0.683±0.005
	14	**0.621±0.006**	0.563±0.006	0.610±0.007	0.606±0.010	0.565±0.015	0.506±0.024	0.437±0.006
msra	9	**0.818±0.005**	0.799±0.006	0.799±0.002	0.767±0.006	0.781±0.008	0.802±0.005	0.692±0.005
	17	**0.764±0.004**	0.694±0.008	0.711±0.004	0.605±0.016	0.688±0.006	0.678±0.009	0.627±0.007
computer	5	**0.666±0.009**	0.385±0.120	0.653±0.010	0.402±0.013	0.591±0.016	0.591±0.008	0.526±0.004
	13	**0.638±0.005**	0.454±0.092	0.576±0.007	0.262±0.015	0.472±0.009	0.614±0.006	0.481±0.011
music_emotion	5.29	**0.631±0.010**	0.566±0.009	0.607±0.010	0.611±0.011	0.621±0.006	0.598±0.006	0.517±0.017
music_style	6.04	**0.746±0.004**	0.701±0.005	0.713±0.004	0.710±0.007	0.554±0.004	0.725±0.004	0.615±0.002
mirflickr	3.35	**0.867±0.034**	0.744±0.058	0.671±0.027	0.827±0.101	0.615±0.078	0.752±0.127	0.748±0.083
yeastBP	30.43	**0.177±0.045**	0.143±0.021	0.086±0.019	0.158±0.016	0.086±0.016	0.149±0.038	0.148±0.022

performance against comparing PML approaches. Here, PENAD is treated as the control approach where the difference of average rank (over all datasets) between PENAD and one comparing approach is calibrated with the *critical difference* (CD). If the average rank difference is greater than one CD (CD = 1.5231 with comparing approaches

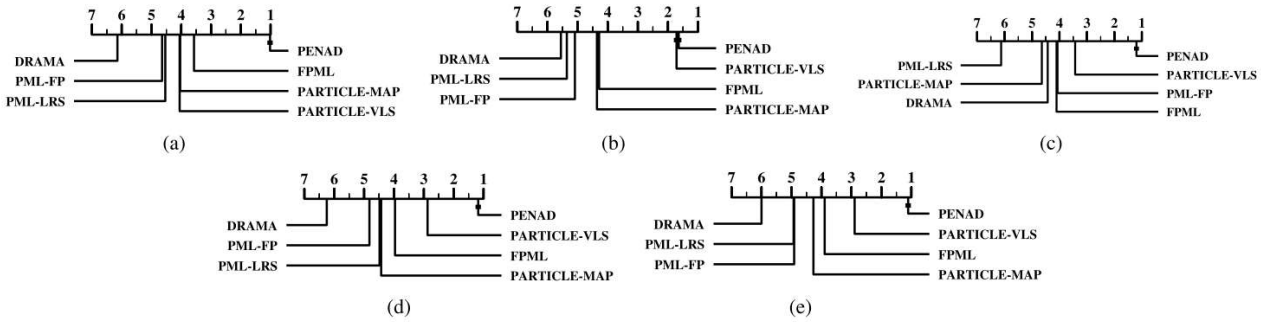

Fig. 3. Comparison of PENAD against other comparing approaches with the *Bonferroni–Dunn test*. The approaches which are not connected with PENAD in the CD diagram are considered to be significantly different from PENAD (CD = 1.5231 at 0.05 significance level). (a) Ranking loss. (b) One-error. (c) Hamming loss. (d) Coverage. (e) Average precision.

TABLE VIII
FRIEDMAN STATISTICS F_F ON ALL EVALUATION METRICS AS WELL AS THE CRITICAL VALUE (AT 0.05 SIGNIFICANCE LEVEL WITH # COMPARING APPROACHES $n = 7$ AND # BENCHMARK DATASETS $N = 28$)

Evaluation metric	F_F	critical value
Ranking loss	28.4784	
Hamming loss	24.4835	
Coverage	32.9084	2.359
One-error	38.3513	
Average precision	32.8540	

$n = 7$, and benchmark datasets $N = 28$), the performance between PENAD and one comparing approach is regarded to be different.

Fig. 3 illustrates the CD diagrams [47] on five evaluation metrics, where the average rank of each comparing PML approach is marked along the axis, where a better rank is set to the right. A thick line is used to connect the control approach and one comparing approach if their average rank difference is within CD. Otherwise, it is considered that the comparing approach has a significantly different performance against PENAD.

Based on the experimental results of comparative studies, the following observations of the comparative studies can be made.

1) Fig. 3 shows that PENAD achieves superior or at least comparable performance against all the comparing approaches on all evaluation metrics. Furthermore, PENAD achieves lowest (best) average rank on all evaluation metrics.
2) The performance of PENAD is statistically comparable to PARTICLE-VLS on *One-error*, and superior to all the comparing approaches on other metrics.
3) Tables III–VII show that the performance advantage of PENAD over comparing approaches is stable under varying the average number of candidate labels.
4) Tables III–VII show that PENAD achieves optimal performance in almost all cases (except on *Coverage* where PARTICLE-MAP outperforms PENAD on `mirflickr`, `yeastBP`, and *One-error* where FPML outperforms PENAD on `yeastBP`) on the four real-world PML datasets `yeastBP`, `music_style`, `music_emotion`, and `mirflickr`.

In summary, these experimental results clearly validate the effectiveness of PENAD for learning from partial multilabel examples.

C. Further Analysis

1) Sensitivity Analysis: In this section, performance sensitivity of the proposed PENAD approach w.r.t. its parameters λ_1, λ_2, β_1, and β_2 will be further analyzed.

Fig. 4 illustrates how PENAD performs under different parameter configurations. For clarity of illustration, three datasets `msra`, `music_style`, and `mirflickr` are chosen here for sensitivity analysis, while similar observations are also made on other datasets.

As shown in Fig. 4, it is obvious that the performance of PENAD is relatively stable across a broad range of each parameter. This property is quite desirable as PENAD could achieve robust classification performance without the need of parameter fine-tuning. Therefore, the parameter configuration for PENAD in Section IV-A naturally follows from these observations.

2) Ablation Studies: To show the helpfulness of latent label distributions to PENAD, a vanilla variant about PENAD (named as PENAD-nonLD) is adopted here which ablates the latent label distribution and follows the same procedure of PENAD with observed logical labels without considering the latent label distributions. In addition, we compare the approach of previous conference version (named as PML-LD) which only consider the topological information of the feature space for recovering latent label distribution. Following the same experimental protocol in Section IV-A3, the results of PENAD-nonLD and PML-LD are investigated.

For brevity, Table IX reports the detailed experimental results in terms of *Average precision* and *Hamming loss*, and the results on some synthetic configurations are given in each synthetic PML dataset, that is, avg. #CLs being 3 and 5 for `emotions`, #CLs being 3 and 5 for `scene`, 9 and 13 for `yeast`, 6 and 14 for `eurlex_sm`, 9 and 17 for `msra`, and 5 and 13 for `computer`. The results on other metrics are similar. In order to show whether PENAD has a significant performance than other versions, we employ

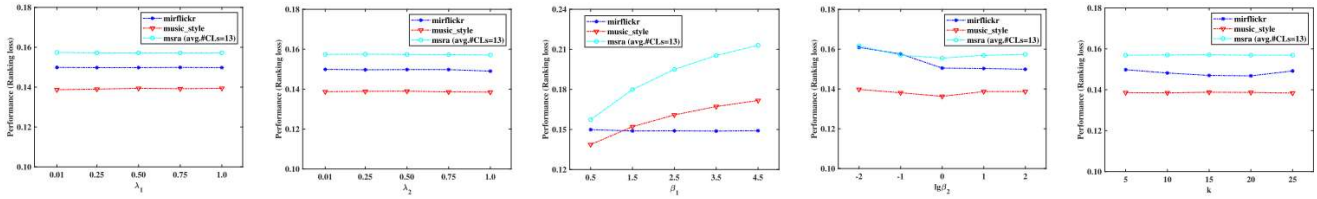

Fig. 4. Parameter sensitivity analysis for PENAD on `mirflickr`, `music_style`, and `msra`. **First**: the performance changes as λ_1 increases from 0.01 to 1.0 ($\lambda_2 = 0.01, \beta_1 = 1, \beta_2 = 10$) on *Ranking loss*. **Second**: the performance changes as λ_2 increases from 0.01 to 1.0 ($\lambda_1 = 0.01, \beta_1 = 1, \beta_2 = 10$) on *Ranking loss*. **Third**: the performance changes as β_1 increases from 0.5 to 4.5 ($\lambda_1 = 0.01, \lambda_2 = 0.01, \beta_2 = 10$) on *Ranking loss*. **Fourth**: the performance changes as β_2 increases from 0.01 to 100 ($\lambda_1 = 0.01, \lambda_2 = 0.01, \beta_1 = 1$) on *Ranking loss*. **Fifth**: the performance changes as k increases from 0.01 to 100 ($\lambda_1 = 0.01, \lambda_2 = 0.01, \beta_1 = 1, \beta_2 = 10$) on *Ranking loss*. (a) λ_1 (on *Ranking loss*). (b) λ_2 (on *Ranking loss*). (c) β_1 (on *Ranking loss*). (d) β_2 (on *Ranking loss*). (e) k (on *Ranking loss*).

TABLE IX

EXPERIMENTAL RESULTS (MEAN±STD) MEASURED BY *Average precision* AND *Hamming loss*

Dataset	avg.CLs	Average precision ↑			Hamming Loss ↓		
		PENAD	PML-LD	PENAD-nonLD	PENAD	PML-LD	PENAD-nonLD
emotions	3	**0.804±0.021**	0.804±0.021	0.783±0.010	0.182±0.012	**0.180±0.014**	0.226±0.017
	5	**0.743±0.025**	0.741±0.028	0.733±0.022	**0.217±0.019**	0.218±0.023	0.247±0.013
image	2	**0.811±0.018**	0.809±0.020	0.788±0.021	**0.150±0.001**	0.151±0.010	0.178±0.013
	4	**0.768±0.013**	0.762±0.017	0.719±0.009	**0.180±0.006**	0.186±0.009	0.228±0.005
scene	3	**0.863±0.011**	0.863±0.013	0.817±0.016	**0.083±0.006**	0.083±0.006	0.141±0.008
	5	0.780±0.020	**0.797±0.022**	0.697±0.020	**0.118±0.012**	0.119±0.012	0.182±0.008
yeast	9	**0.747±0.006**	0.746±0.007	0.738±0.006	**0.137±0.001**	0.139±0.001	0.238±0.005
	13	**0.712±0.004**	0.712±0.004	0.686±0.002	**0.143±0.001**	0.145±0.001	0.244±0.003
eulex_sm	6	**0.779±0.004**	0.778±0.005	0.771±0.005	**0.070±0.001**	0.070±0.001	0.072±0.001
	14	**0.621±0.006**	0.619±0.006	0.563±0.009	**0.091±0.002**	0.091±0.002	0.107±0.001
msra	9	**0.818±0.005**	0.812±0.005	0.798±0.003	**0.104±0.001**	0.106±0.001	0.180±0.002
	17	**0.764±0.004**	0.754±0.004	0.681±0.004	**0.104±0.001**	0.107±0.002	0.165±0.002
computer	5	**0.666±0.009**	0.647±0.010	0.636±0.010	**0.068±0.001**	0.092±0.001	0.105±0.001
	13	**0.638±0.005**	0.572±0.006	0.549±0.007	**0.060±0.001**	0.117±0.001	0.120±0.001
music_emotion	5.29	**0.631±0.010**	0.630±0.010	0.614±0.008	**0.123±0.002**	0.123±0.002	0.228±0.002
music_style	6.04	**0.746±0.004**	0.737±0.003	0.712±0.005	0.111±0.003	**0.109±0.002**	0.162±0.002
mirflickr	3.35	**0.867±0.034**	0.835±0.090	0.753±0.056	**0.046±0.013**	0.062±0.045	0.198±0.011
yeastBP	21.56	**0.177±0.045**	0.171±0.045	0.142±0.045	**0.031±0.001**	0.070±0.004	0.093±0.007

TABLE X

WILCOXON SIGNED-RANK TEST FOR PENAD AGAINST ITS VARIANT VERSION PENAD-NONLD AND CONFERENCE VERSION PML-LD ON FIVE EVALUATION METRICS (AT 0.05 SIGNIFICANCE LEVEL)

Evaluation metric	PENAD against PENAD-nonLD		PENAD against PML-LD	
	performance	p-value	performance	p-value
Ranking loss	**win**	$3.998e^{-6}$	**win**	$5.662e^{-5}$
One-error	**win**	$4.213e^{-6}$	**win**	$2.550e^{-4}$
Hamming loss	**win**	$4.000e^{-6}$	**win**	$1.790e^{-4}$
coverage	**win**	$4.228e^{-6}$	**win**	$1.182e^{-3}$
Average precision	**win**	$3.998e^{-6}$	**win**	$7.913e^{-5}$

Wilcoxon signed-rank test [47]. Wilcoxon signed-rank test is a non-parametric test, which ranks the differences in performances of two approaches for each dataset and compares the ranks for the positive and the negative differences. Table X shows the p-values for the corresponding tests and the statistical test results at 0.05 significance level.

As shown in Table X, PENAD achieves superior performance against PENAD-nonLD and PML-LD on all evaluation metrics, which clearly validates the usefulness of latent label distributions for improving performance and the substantial improvement of PENAD on its conference version.

3) Quality of Recovered Latent Label Distributions: Note that the label distributions leveraged by PENAD need to be recovered from the training data. In order to show whether PENAD can successfully exploit the label distributions, the recovering experiments are studied. Specifically, the label distributions are recovered from the partial multilabel data via PENAD and other comparing LE approaches. Then, these output label distributions are normalized by softmax normalization and compared with the ground-truth ones.

There are in total nine datasets used in the experiments. These real-world datasets[1] contain real label distributions and the corresponding logical labels [34].

1) `Yeast-alpha` to `Yeast-spo`: These datasets are collected from biological experiments on the budding yeast *Saccharomyces cerevisiae*. The result of each biological experiment is recorded by one dataset. Each dataset contains 2465 yeast genes, and each of the genes is represented by an associated phylogenetic profile vector. The labels in each dataset are corresponding to the discrete time points during one experiment. The gene

[1] http://cse.seu.edu.cn/PersonalPage/xgeng/LDL/index.htm

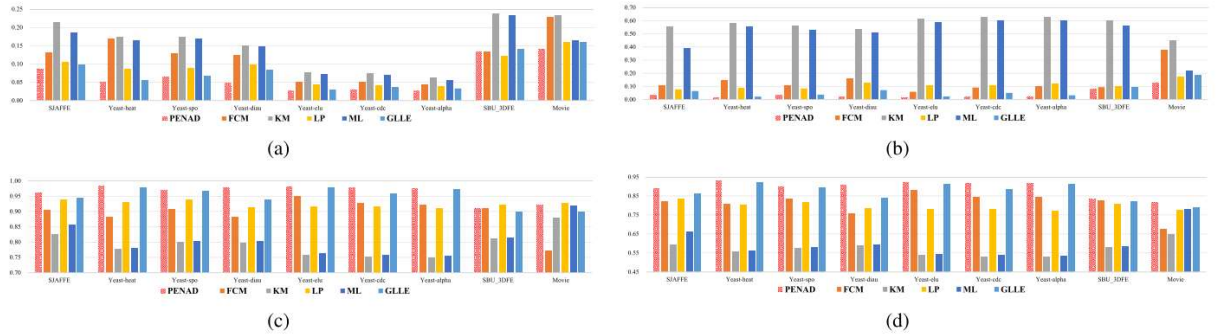

Fig. 5. Quantitative analysis on the quality of recovered latent label distributions evaluated by *Chebyshev distance*, *Kullback–Leibler divergence*, *Cosine coefficient*, and *Intersection similarity*, where the "↓" after the metrics indicates "the smaller the better" and "↑" after the metrics indicates "the larger the better." (a) *Chebyshev distance* ↓. (b) *Kullback–Leibler divergence* ↓. (c) *Cosine coefficient* ↑. (d) *Intersection similarity* ↑.

expression level at each time point exactly constitutes the label distribution of the corresponding gene. These datasets contain 2465 examples with 24 features and 6 to 18 class labels.

2) `SBU_3DFE`: This dataset is a facial expression database, where each facial expression is assigned by basic emotions such as sadness, happiness, fear, surprise, anger, and disgust. A total of 23 persons annotate the level of emotion intensity (1–5) for each facial expression, and the averaged annotation intensities are utilized to generate the ground-truth label distribution. There are in total 2500 examples with 243 features and six class labels in this dataset.
3) `SJAFFE`: This dataset is also a facial expression database, where each facial expression is assigned by basic emotions such as sadness, happiness, fear, surprise, anger, and disgust. Similarly, a total of 60 persons are asked to annotate the level of emotion intensity, and the averaged annotation intensities are utilized to generate the label distribution. This dataset contains 213 examples with 243 features and six class labels.
4) `Movie`: The dataset is a movie database, which contains 7755 movies and 54 242 292 ratings from 478 656 different users. The ratings are on a scale from 1 to 5 integral stars (five labels). The label distribution is calculated for each movie as the percentage of each rating level. This dataset contains in total 7755 examples with 1869 features and five class labels.

Specifically, the partial multilabel data is generated by adding random labeling noise. Some of the irrelevant labels corresponding to each example are randomly chosen to form the candidate label set along with the valid relevant labels.

According to Geng's suggestion [8], four evaluation metrics are selected to quantify the quality of recovered latent label distributions.

1) *Chebyshev Distance*

$$D_{\text{Cheb}} = (1/n)\sum_{i=1}^{n}\max_{j}\left|d_{\boldsymbol{x}_i}^{y_j} - \hat{d}_{\boldsymbol{x}_i}^{y_j}\right|.$$

2) *Kullback–Leibler Divergence*

$$D_{\text{KL}} = (1/n)\sum_{i=1}^{n}\sum_{j=1}^{c} d_{\boldsymbol{x}_i}^{y_j}\ln\left(d_{\boldsymbol{x}_i}^{y_j}/\hat{d}_{\boldsymbol{x}_i}^{y_j}\right).$$

3) *Cosine Coefficient*

$$S_{\text{Cos}} = (1/n)\sum_{i=1}^{n}\left(\left(\sum_{j=1}^{c} d_{\boldsymbol{x}_i}^{y_j}\hat{d}_{\boldsymbol{x}_i}^{y_j}\right)\Big/\left(\left(\sum_{j=1}^{c}\left(d_{\boldsymbol{x}_i}^{y_j}\right)^2\right)^{1/2}\left(\sum_{j=1}^{c}\left(\hat{d}_{\boldsymbol{x}_i}^{y_j}\right)^2\right)^{1/2}\right)\right)$$

4) *Intersection Similarity*

$$S_{\text{Inter}} = (1/n)\sum_{i=1}^{n}\sum_{j=1}^{c}\min\left(d_{\boldsymbol{x}_i}^{y_j}, \hat{d}_{\boldsymbol{x}_i}^{y_j}\right).$$

Here, $\boldsymbol{d}_i = [d_{\boldsymbol{x}_i}^{y_1}, d_{\boldsymbol{x}_i}^{y_2}, \ldots, d_{\boldsymbol{x}_i}^{y_c}]$ is a real label distribution, and $\hat{\boldsymbol{d}}_i = [\hat{d}_{\boldsymbol{x}_i}^{y_1}, \hat{d}_{\boldsymbol{x}_i}^{y_2}, \ldots, \hat{d}_{\boldsymbol{x}_i}^{y_c}]$ is a recovered latent label distribution. The first two are distance metrics and the last two are similarity metrics. Considering that the selected metrics all come from different families, the selected metrics are significantly different in both syntax and semantics.

Five baseline LE algorithms are employed for comparative studies.

1) FCM [48] which employs fuzzy C-means clustering technique to generate the membership degree of each instance to each cluster and adopts fuzzy composition to generate label distributions from the membership degrees [suggested configuration: $\beta = 2$].
2) KM [49] which adopts kernel function to calculate the distance between each instance and the center of each class and generate the label distributions from the distance.
3) LP [41] which employs iterative label propagation technique to generate label distributions [suggested configuration: balancing parameter $\alpha = 0.5$].
4) ML [39] which leverages feature manifold and label manifold to generate label distributions [suggested configuration: the number of neighbors $K = c + 1$].
5) GLLE [34] which recovers label distributions via leveraging the topological information of the feature space [suggested configuration: the parameter λ_1 and λ_2 are chosen among $\{10^{-2}, 10^{-1}, \ldots, 100\}$].

Fig. 5 illustrates the performance of PENAD against five baseline algorithms in terms of four evaluation metrics.

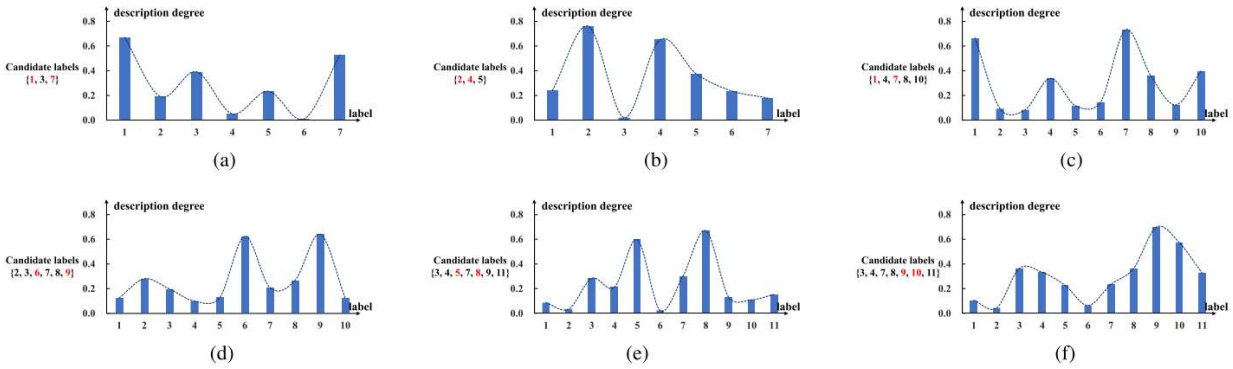

Fig. 6. Visualizations of learned label distributions, where the red label in the each candidate label set is valid. (a) `mirflickr`. (b) `mirflickr`. (c) `music_style`. (d) `music_style`. (e) `music_emotion`. (f) `music_emotion`.

For each evaluation metric, ↑ denotes the larger the better and ↓ denotes the smaller the better. PENAD ranks *1st* in 94% cases across four metrics. These results show that PENAD is effective to recover latent label distributions in partially labeled examples. In addition, the visualizations of the learned latent label distribution in real-world PML datasets are given in Fig. 6.

V. CONCLUSION

Partial multilabel learning aims to learn the multilabel predictive model from PML datasets, in which each example is associated with candidate labels but only a subset of these labels is valid. Different from existing strategies, the proposed approach PENAD considers the label distributions in the training datasets. Since the label distributions are not explicitly available in the training sets, PENAD recovers the label distributions as well as induces the predictive model simultaneously. The effectiveness of the proposed approach is validated via the PML predictive experiments. In addition, further experiments show the high quality of the recovered label distributions and the effectiveness of adopting label distributions for partial multilabel learning.

It is interesting to investigate effective ways to make full use of the label distribution in PML. Furthermore, more LE approaches need to be investigated when there are certain structures in the partial multilabel sets of PML training examples. In the future, it is also important to explore other techniques to leverage the recovered label distribution for PML.

REFERENCES

[1] M.-L. Zhang and J.-P. Fang, "Partial multi-label learning via credible label elicitation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 10, pp. 3587–3599, Oct. 2021.

[2] J.-P. Fang and M.-L. Zhang, "Partial multi-label learning via credible label elicitation," in *Proc. AAAI Conf. Artif. Intell.*, Honolulu, HI, USA, 2019, pp. 3518–3525.

[3] J. Luo and F. Orabona, "Learning from candidate labeling sets," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 23, Vancouver, QC, Canada, 2010, pp. 1504–1512.

[4] L. Sun, S. Feng, T. Wang, C. Lang, and Y. Jin, "Partial multi-label learning by low-rank and sparse decomposition," in *Proc. AAAI Conf. Artif. Intell.*, Honolulu, HI, USA, 2019, pp. 5016–5023.

[5] G. Yu *et al.*, "Feature-induced partial multi-label learning," in *Proc. IEEE Int. Conf. Data Mining (ICDM)*, Singapore, Nov. 2018, pp. 1398–1403.

[6] M.-K. Xie and S.-J. Huang, "Partial multi-label learning," in *Proc. AAAI Conf. Artif. Intell.*, New Orleans, LA, USA, 2018, pp. 4302–4309.

[7] M. K. Xie and S. J. Huang, "Partial multi-label learning with noisy label identification," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Feb. 15, 2021, doi: 10.1109/TPAMI.2021.3059290.

[8] X. Geng, "Label distribution learning," *IEEE Trans. Knowl. Data Eng.*, vol. 28, no. 7, pp. 1734–1748, Jul. 2016.

[9] N. Xu, Y.-P. Liu, and X. Geng, "Partial multi-label learning with label distribution," in *Proc. 34th AAAI Conf. Artif. Intell.*, New York, NY, USA, 2020, pp. 6510–6517.

[10] M.-L. Zhang and Z.-H. Zhou, "A review on multi-label learning algorithms," *IEEE Trans. Knowl. Data Eng.*, vol. 26, no. 8, pp. 1819–1837, Aug. 2014.

[11] E. Gibaja and S. Ventura, "A tutorial on multilabel learning," *ACM Comput. Surv.*, vol. 47, no. 3, pp. 1–38, Apr. 2015.

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

[13] T. Cour, B. Sapp, and B. Taskar, "Learning from partial labels," *J. Mach. Learn. Res.*, vol. 12, pp. 1501–1536, Apr. 2011.

[14] L. Liu and T. Dietterich, "A conditional multinomial mixture model for superset label learning," in *Proc. Adv. Neural Inf. Process. Syst.*, Cambridge, MA, USA, 2012, pp. 557–565.

[15] M. R. Boutell, J. Luo, X. Shen, and C. M. Brown, "Learning multi-label scene classification," *Pattern Recognit.*, vol. 37, no. 9, pp. 1757–1771, Sep. 2004.

[16] M.-L. Zhang and Z.-H. Zhou, "ML-KNN: A lazy learning approach to multi-label learning," *Pattern Recognit.*, vol. 40, no. 7, pp. 2038–2048, Jul. 2007.

[17] A. Elisseeff and J. Weston, "A kernel method for multi-labelled classification," in *Proc. Adv. Neural Inf. Process. Syst.*, Vancouver, BC, Canada, 2002, pp. 681–687.

[18] J. Fürnkranz, E. Hüllermeier, E. L. Mencía, and K. Brinker, "Multilabel classification via calibrated label ranking," *J. Mach. Learn.*, vol. 73, no. 2, pp. 133–153, Nov. 2008.

[19] J. Read, B. Pfahringer, G. Holmes, and E. Frank, "Classifier chains for multi-label classification," *Mach. Learn.*, vol. 85, no. 3, p. 333, Dec. 2011.

[20] G. Tsoumakas, I. Katakis, and L. Vlahavas, "Random K-labelsets for multilabel classification," *IEEE Trans. Knowl. Data Eng.*, vol. 23, no. 7, pp. 1079–1089, Jul. 2011.

[21] C. Gong, D. Tao, J. Yang, and W. Liu, "Teaching-to-learn and learning-to-teach for multi-label propagation," in *Proc. AAAI Conf. Artif. Intell.*, vol. 30, no. 1, 2016, pp. 1610–1616.

[22] C. Gong, D. Tao, W. Liu, L. Liu, and J. Yang, "Label propagation via teaching-to-learn and learning-to-teach," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 28, no. 6, pp. 1452–1465, Jun. 2017.

[23] F. Yu and M.-L. Zhang, "Maximum margin partial label learning," *Mach. Learn.*, vol. 106, no. 4, pp. 573–593, Apr. 2017.

[24] L. Feng *et al.*, "Provably consistent partial-label learning," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 33, Vienna, Austria, 2020.

[25] J. Lv, M. Xu, L. Feng, G. Niu, X. Geng, and M. Sugiyama, "Progressive identification of true labels for partial-label learning," in *Proc. Int. Conf. Mach. Learn.*, Vienna, Austria, 2020, pp. 6500–6510.

[26] Y. Yao, C. Gong, J. Deng, and J. Yang, "Network cooperation with progressive disambiguation for partial label learning," in *Proc. Mach. Learn. Knowl. Discovery Databases-Eur. Conf. (PKDD)*, Ghent, Belgium, Sep. 2020, pp. 471–488.
[27] J. Wang and X. Geng, "Theoretical analysis of label distribution learning," in *Proc. 33rd AAAI Conf. Artif. Intell.*, Honolulu, HI, USA, 2019, pp. 5256–5263.
[28] X. Geng and Y. Xia, "Head pose estimation based on multivariate label distribution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Columbus, OH, USA, Jun. 2014, pp. 1837–1842.
[29] K. Su and X. Geng, "Soft facial landmark detection by label distribution learning," in *Proc. 33rd AAAI Conf. Artif. Intell.*, Honolulu, HI, 2019, pp. 5008–5015.
[30] Z. Huo and X. Geng, "Ordinal zero-shot learning," in *Proc. 26th Int. Joint Conf. Artif. Intell.*, Melbourne, VIC, Australia, Aug. 2017, pp. 1331–1337.
[31] B.-B. Gao, H.-Y. Zhou, J. Wu, and X. Geng, "Age estimation using expectation of label distribution learning," in *Proc. 27th Int. Joint Conf. Artif. Intell.*, Stockholm, Sweden, Jul. 2018, pp. 712–718.
[32] X. Geng, C. Yin, and Z.-H. Zhou, "Facial age estimation by learning from label distributions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 10, pp. 2401–2412, Oct. 2013.
[33] D. Zhou, X. Zhang, Y. Zhou, Q. Zhao, and X. Geng, "Emotion distribution learning from texts," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, Austin, TX, USA, 2016, pp. 638–647.
[34] N. Xu, Y.-P. Liu, and X. Geng, "Label enhancement for label distribution learning," *IEEE Trans. Knowl. Data Eng.*, vol. 33, no. 4, pp. 1632–1643, Apr. 2021.
[35] N. Xu, J. Shu, Y.-P. Liu, and X. Geng, "Variational label enhancement," in *Proc. Int. Conf. Mach. Learn.*, Vienna, Austria, 2020, pp. 10597–10606.
[36] H. Tang, J. Zhu, Q. Zheng, J. Wang, S. Pang, and Z. Li, "Label enhancement with sample correlations via low-rank representation," in *Proc. AAAI Conf. Artif. Intell.*, 2020, vol. 34, no. 4, New York, NY, USA, pp. 5932–5939.
[37] S. Chen, J. Wang, Y. Chen, Z. Shi, X. Geng, and Y. Rui, "Label distribution learning on auxiliary label space graphs for facial expression recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2020, pp. 13984–13993.
[38] F. Zhang, X. Jia, and W. Li, "Tensor based multi-view label enhancement for multi-label learning," in *Proc. Int. Joint Conf. Artif. Intell.*, Yokohama, Japan, 2020, pp. 2369–2375.
[39] P. Hou, X. Geng, and M.-L. Zhang, "Multi-label manifold learning," in *Proc. 30th AAAI Conf. Artif. Intell.*, Phoenix, AZ, USA, 2016, pp. 1680–1686.
[40] M.-L. Zhang, Q.-W. Zhang, J.-P. Fang, Y.-K. Li, and X. Geng, "Leveraging implicit relative labeling-importance information for effective multi-label learning," *IEEE Trans. Knowl. Data Eng.*, vol. 35, no. 5, pp. 2057–2070, May 2021.
[41] Y.-K. Li, M.-L. Zhang, and X. Geng, "Leveraging implicit relative labeling-importance information for effective multi-label learning," in *Proc. IEEE Int. Conf. Data Mining*, Atlantic City, NJ, USA, Nov. 2015, pp. 251–260.
[42] X. Zhu, J. Lafferty, and R. Rosenfeld, *Semi-Supervised Learning With Graphs*. Pittsburgh, PA, USA: Carnegie Mellon Univ., 2005.
[43] M.-L. Zhang, B.-B. Zhou, and X.-Y. Liu, "Partial label learning via feature-aware disambiguation," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2016, pp. 1335–1344.
[44] A. J. Smola, "Learning with kernels," Ph.D. dissertation, GMD, Birlinghoven, German, 1999.
[45] M. J. Huiskes and M. S. Lew, "The MIR Flickr retrieval evaluation," in *Proc. 1st ACM Int. Conf. Multimedia Inf. Retr. (MIR)*, Vancouver, BC, Canada, 2008, pp. 39–43.
[46] H. Wang, W. Liu, Y. Zhao, C. Zhang, T. Hu, and G. Chen, "Discriminative and correlative partial multi-label learning," in *Proc. 28th Int. Joint Conf. Artif. Intell.*, Macao, China, Aug. 2019, pp. 3691–3697.
[47] J. Demšar, "Statistical comparisons of classifiers over multiple data sets," *J. Mach. Learn. Res.*, vol. 7, pp. 1–30, Jan. 2006.
[48] N. E. Gayar, F. Schwenker, and G. Palm, "A study of the robustness of KNN classifiers trained using soft labels," in *Proc. 2nd Int. Conf. Artif. Neural Netw. Pattern Recognit.*, Ulm, Germany, 2006, pp. 67–80.
[49] X. Jiang, Z. Yi, and J. C. Lv, "Fuzzy SVM with a new fuzzy membership function," *Neural Comput. Appl.*, vol. 15, nos. 3–4, pp. 268–276, Jun. 2006.

Ning Xu (Member, IEEE) received the B.Sc. degree from the University of Science and Technology of China, Hefei, China, in 2010, the M.Sc. degree from the Chinese Academy of Sciences China, Beijing, China, in 2013, and the Ph.D. degree from Southeast University, Nanjing, China, in 2020.

He is currently an Assistant Professor with the School of Computer Science and Engineering, Southeast University. His research interests mainly include machine learning and pattern recognition. His research results have been published at prestigious journals and leading conferences such as IEEE TKDE, NeurIPS, ICML, AAAI, IJCAI, and so on.

Yun-Peng Liu was born in 1994. He is currently pursuing the master's degree with Southeast University, Nanjing, China.

His main research interests include machine learning and computer vision.

Yan Zhang was born in 1997. She is currently pursuing the master's degree with Southeast University, Nanjing, China.

Her main research interests include machine learning and computer vision.

Xin Geng (Senior Member, IEEE) received the B.Sc. and M.Sc. degrees in computer science from Nanjing University, Nanjing, China, in 2001 and 2004, respectively, and the Ph.D. degree in computer science from Deakin University in 2008.

He is currently a Professor and the Dean of the School of Computer Science and Engineering, Southeast University, Nanjing. He has authored or coauthored over 70 refereed articles in these areas, including those published in prestigious journals and top international conferences. His research interests include machine learning, pattern recognition, and computer vision.

Dr. Geng is a Distinguished Fellow of IETI. He has been an Associate Editor of IEEE T-MM, FCS, and MFC, a Steering Committee Member of PRICAI, a Program Committee Chair for conferences such as PRICAI'18, VALSE'13, and so on, an Area Chair for conferences such as CVPR, ACMMM, PRCV, CCPR, and a Senior Program Committee Member for conferences such as IJCAI, AAAI, ECAI, and so on.