WPML^{3}CP: Wasserstein Partial Multi-Label Learning with Dual Label Correlation Perspectives

Ximing Li[1,2], **Yuanchao Dai**[1,2], **Bing Wang**[1,2], **Changchun Li**[1,2,*], **Renchu Guan**[1,2], **Fangming Gu**[1,2] and **Jihong Ouyang**[1,2]
[1]College of Computer Science and Technology, Jilin University, China
[2]Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education, Jilin University, China
{liximing86, yuanchaodai, wangbing1416, changchunli93}@gmail.com, {guanrenchu, gufm, ouyj}@jlu.edu.cn

Abstract

Partial multi-label learning (PMLL) refers to a weakly-supervised classification problem, where each instance is associated with a set of candidate labels, covering its ground-truth labels but also with irrelevant ones. The current methodology of PMLL is to estimate the *ground-truth confidences* of candidate labels, *i.e.,* the likelihood of a candidate label being a ground-truth one, and induce the multi-label predictor with them, rather than the candidate labels. In this paper, we aim to estimate precise ground-truth confidences by leveraging precise label correlations, which are also required to estimate. To this end, we propose to capture label correlations from both measuring and modeling perspectives. Specifically, we measure the loss between ground-truth confidences and predictions by employing the Wasserstein distance involving label correlations; and form a label correlation-aware regularization to constrain predictive parameters. The two techniques are coupled to promote precise estimations of label correlations. Upon these ideas, we propose a novel PMLL method, namely Wasserstein Partial Multi-Label Learning with dual Label Correlation Perspectives (WPML^{3}CP). We conduct extensive experiments on several benchmark datasets. Empirical results demonstrate that WPML^{3}CP can outperform the existing PMLL baselines.

1 Introduction

Multi-label learning (MLL) [Zhang and Zhou, 2014a] refers to inducing multi-label predictors from the precisely labeled training dataset, where each instance is annotated with its ground-truth labels. However, in many real-world scenarios, *e.g.,* annotations from crowdsourcing platforms, only partially valid training dataset is available due to various reasons [Li *et al.*, 2021; Li *et al.*, 2022; Feng *et al.*, 2022; Li *et al.*, 2023; Xie *et al.*, 2023], such as the difficulty of accurate supervision collection and human effort cost reduction, *etc.* A common situation is that each instance is associated with a set of candidate labels, which cover the ground-truth labels but also with some irrelevant ones, *e.g.,* the image instance illustrated in Fig.1. Learning with such a partially valid training dataset, formally referred to as **p**artial **m**ulti-**l**abel **l**earning (**PMLL**), is naturally much more challenging than the traditional MLL since even the number of ground-truth labels is unknown.

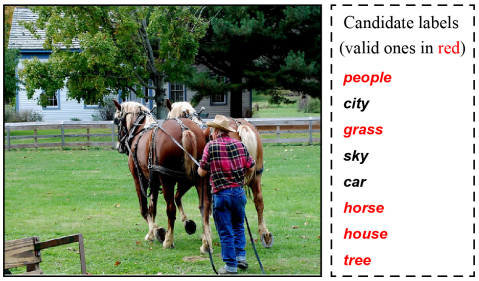

Figure 1: An example of partial multi-label learning. The example image instance is annotated with 8 candidate labels, but only 5 of them are ground-truth ones.

To handle the emerging task of PMLL, many methods have been recently developed, where the basic idea is to estimate the *ground-truth confidences* of candidate labels, *i.e.,* the likelihood of a candidate label being a ground-truth one, and induce the multi-label predictor with them rather than candidate labels [Xie and Huang, 2018a; Sun *et al.*, 2019; Li *et al.*, 2020; Li and Wang, 2020; Xie *et al.*, 2021; Sun *et al.*, 2022]. Naturally, the primary philosophy of PMLL is how to estimate precise ground-truth confidences, and to our knowledge, the prevalent spirit is to promote it by exploiting the correlations among labels. For example, the GLC method formulates the ground-truth confidences by integrating a label correlation matrix [Sun *et al.*, 2022]; the MUSER method decomposes ground-truth confidences into two parts, including a low-dimensional label matrix and a label correlation matrix [Li *et al.*, 2020]. Although this spirit is empirically effective in prior literature, it raises a new problem that label correlations are also required to estimate, and imprecise estimations may result in imprecise ground-truth confidence.

*Corresponding author.

Accordingly, in this paper, we concentrate on estimating precise label correlations to promote the precision of ground-truth confidence. To this end, we propose to capture label correlations from both measuring and modeling perspectives. Specifically, we measure the loss between ground-truth confidences and predictions by employing the Wasserstein distance involving label correlations; and form a label correlation-aware regularization to constrain predictive parameters. The two techniques are coupled to promote precise estimations of label correlations. Upon these ideas, we propose a novel PMLL method, namely **W**asserstein **P**artial **M**ulti-**L**abel **L**earning with dual **L**abel **C**orrelation **P**erspectives (**WPML^{3}CP**), which is optimized within the framework of the augmented Lagrange multiplier method. To evaluate the performance of WPML^{3}CP, we conduct extensive experiments on both real-world and synthetic datasets. Experimental results demonstrate that WPML^{3}CP can outperform the existing PMLL baselines in most cases.

To sum up, the main contributions of this paper are presented as follows:

- We propose a novel PMLL method named **WPML^{3}CP**, which promotes precise ground-truth confidences by capturing label correlations from both measuring and modeling perspectives.
- We efficiently optimize the proposed **WPML^{3}CP** method with the framework of augmented Lagrange multiplier.
- We empirically validate the effectiveness of **WPML^{3}CP** by comparing with the existing PMLL baselines on both real-world and synthetic datasets.

2 Related Works

2.1 Partial Multi-Label Learning

Recently, several PMLL methods have been proposed. Generally, they mainly focus on estimating the ground-truth confidences, so as to capture more accurate supervised signals. At present, the methods to solve PMLL are mainly divided into unified framework method and two-stage method. Specifically, the unified framework method aims to integrate label confidence learning and multi-classification model into a joint framework, then the unified framework is optimized by iterative strategy. The PMLL framework proposed by [Xie and Huang, 2018b] is built on a ranking loss objective weighted by the ground-truth confidence, which is regularized through either label correlation or feature prototype, and gives two versions, named PML-lc and PML-fp, respectively. Then both fPML [Yu *et al.*, 2018] and PML-LRS [Sun *et al.*, 2019] utilize the low-rank decomposition to capture the ground-truth label matrix from the observed candidate labels. MUSER [Li *et al.*, 2020] combines label decomposition and feature mapping to enhance the robustness of the classifier against redundant labels and noisy features. Another recent PML-NI [Xie and Huang, 2021] method jointly learns the multi-label classifier and noise label identifier which combines the label correlation exploitation and feature-induced noise model. In order to make full use of label correlation information to deal with noise labels, GLC [Sun *et al.*, 2022] uses low-rank representation and label manifold regularizer to capture the global and local label correlation, respectively. Recently, graph-based methods have been widely used to deal with PMLL problems. PARD [Hang and Zhang, 2023] first applies the probabilistic graphical model to disambiguate the candidate labels. PLAIN [Wang *et al.*, 2023] constructs the instance graph and the label graph in a data-driven way, and calculates similarity to generate pseudo-labels through label propagation.

Different from the unified framework method, the two-stage method restores the ground-truth label distribution or learns a label confidence matrix in the first stage, and then builds a multi-classification prediction model based on the recovered label distribution or label confidence matrix in the second stage. PARTICLE [Zhang and Fang, 2021] first estimates the ground-truth confidences by iterating label propagation to select credible labels. Then, it induces the prediction model with these credible labels using either VLS or MAP, respectively. In addition to PARTICLE's work, PML-LD [Xu *et al.*, 2020] uses the topological information of feature space and the correlation between labels to recover the label distribution. PML-MD [Xie *et al.*, 2021] uses confidence-weighted ranking loss to process the PMLL data in the first stage, where confidences are estimated adaptively using a meta-learning-based approach. Considering that estimating label confidences may introduce a lot of uncertainty, PAMB [Liu *et al.*, 2023] uses error-correcting output code (ECOC) technique to transform PMLL problems into multiple binary learning problems.

2.2 Learning with Wasserstein Distance

The Wasserstein distance [Villani, 2008] has drawn great attention in the machine learning community, and been widely applied to various learning tasks [Cuturi and Doucet, 2014; Rubner *et al.*, 2000; Arjovsky *et al.*, 2017; Frogner *et al.*, 2015; Kusner *et al.*, 2015; Huang *et al.*, 2016; Zhao and Zhou, 2018]. Recently, the authors of [Frogner *et al.*, 2015] propose a novel learning framework, which refers to the Wasserstein distance as the loss function. Orthogonal to WPML^{3}CP, this framework is also solved using entropic regularization approximation [Cuturi, 2013], however, it is only applicable to precisely annotated datasets. Another representative work proposed in [Zhao and Zhou, 2018] employs Wasserstein loss for the task of label distribution learning. It introduces a kernel biased regularization to simultaneously leverage the label distribution and induce label correlations. However, it is also assumed that the ground-truth supervision is known.

3 Preliminaries

In this section, we first briefly review the Wasserstein distance and then introduce its entropic regularization.

The Wasserstein distance [Bogachev and Kolesnikov, 2012] is a special case of optimal transport distance [Villani, 2008], also referred as earth mover's distance [Rubner *et al.*, 2000]. Formally, given any two probability measures $\mu(\mathbf{u})$, $\nu(\mathbf{v})$ on a space $\mathcal{K}$ (*i.e.,* $\mathbf{u}, \mathbf{v} \in \mathcal{K}$) and a cost function $c : \mathcal{K} \times \mathcal{K} \to \mathbb{R}$, the optimal transport distance measures the

cheapest way to transport the mass in μ to match that in ν:

$$W_c(\mu,\nu)=\inf_{\gamma\in\Pi(\mu,\nu)}\int_{\mathcal{K}\times\mathcal{K}} c(\mathbf{u},\mathbf{v})\,\gamma(d\mathbf{u},d\mathbf{v}), \quad (1)$$

where $\Pi(\mu,\nu)$ denotes the set of all joint probability measures on $\mathcal{K}\times\mathcal{K}$ with marginals $\mu(\mathbf{u})$ and $\nu(\mathbf{v})$. The Wasserstein distance is the case of Eq.(1), in which the cost function c is specified by $d_{\mathcal{K}}^r$, the r-th power of a metric $d_{\mathcal{K}}$ on $\mathcal{K}$ with $r\in[1,\infty)$.

Discrete Wasserstein distance. When $\mathcal{K}$ is a finite set of size $|\mathcal{K}|=K$, both $\boldsymbol{\mu}(\mathbf{u})$ and $\boldsymbol{\nu}(\mathbf{v})$ are discrete distributions (*i.e.,* histograms) in the simplex Δ^K, *i.e.,* $\{\boldsymbol{\mu},\boldsymbol{\nu}\in\mathbb{R}_+^K \mid \boldsymbol{\mu}^\top\mathbf{1}=1,\boldsymbol{\nu}^\top\mathbf{1}=1\}$, where $\mathbf{1}$ denotes the all-one vector. Given a *cost matrix* $\mathbf{M}_{\mathcal{K}}$ computed by the r-th power of pairwise distances over $\mathcal{K}$, *i.e.,* $\mathbf{M}_{\mathcal{K}}=[d_{\mathcal{K}}^r(\mathbf{u}_i,\mathbf{v}_j)]_{ij}\in\mathbb{R}^{K\times K}$, the discrete Wasserstein distance between $\boldsymbol{\mu}(\mathbf{u})$ and $\boldsymbol{\nu}(\mathbf{v})$ can be written as:

$$W_r^r(\boldsymbol{\mu},\boldsymbol{\nu};\mathbf{M}_{\mathcal{K}})=\inf_{\mathbf{T}\in U(\boldsymbol{\mu},\boldsymbol{\nu})}\langle\mathbf{T},\mathbf{M}_{\mathcal{K}}\rangle, \quad (2)$$

where $\langle\cdot,\cdot\rangle$ denotes the Frobenius dot-product, and $U(\boldsymbol{\mu},\boldsymbol{\nu})=\{\mathbf{T}\in\mathcal{R}_+^{K\times K}|\mathbf{T}\mathbf{1}=\boldsymbol{\mu},\mathbf{T}^\top\mathbf{1}=\boldsymbol{\nu}\}$ is the transportation polytope.

Regularized Wasserstein distance. The discrete Wasserstein distance of Eq.(2) is not smooth with respect to its arguments $\boldsymbol{\mu}$, $\boldsymbol{\nu}$ [Cuturi and Doucet, 2014], and its computation requires costly $O(K^3\log K)$ time [Pele and Werman, 2009]. To handle those problems, the authors of [Cuturi, 2013] incorporate an entropic regularizer into Eq.(2), leading to the regularized Wasserstein distance:

$$W_\lambda(\boldsymbol{\mu},\boldsymbol{\nu};\mathbf{M}_{\mathcal{K}})=\inf_{\mathbf{T}\in U(\boldsymbol{\mu},\boldsymbol{\nu})}\langle\mathbf{T},\mathbf{M}_{\mathcal{K}}\rangle-\frac{1}{\lambda}H(\mathbf{T}), \quad (3)$$

where $H(\mathbf{T})=-\langle\mathbf{T},\log\mathbf{T}\rangle$ is the (strictly concave) entropy function and $\lambda>0$ is the regularization parameter. Accordingly, given models with parameters of interest, *i.e.,* denoted by $\{\boldsymbol{\mu},\boldsymbol{\nu},\mathbf{M}_{\mathcal{K}}\}$, we can optimize them by the Sinkhorn's algorithm with $O(K^2)$ complexity [Cuturi, 2013; Cuturi and Doucet, 2014]. The optimization details are introduced **in the Appendix** [1] .

4 Proposed WPML^{3}CP Method

In this section, we introduce the proposed PMLL method named **WPML^{3}CP**.

Problem formulation of PMLL. Formally, given a partially valid training dataset $\mathcal{D}=\{(\mathbf{x}_i,\mathbf{y}_i)\}_{i=1}^{i=n}$ with n instances and l category labels, let $\mathbf{x}_i\in\mathbb{R}^d$ and $\mathbf{y}_i\in\{0,1\}^l$ denote the d-dimensional feature vector and the candidate label set associated with the i-th instance, respectively. For $\mathbf{y}_i$, the candidate label is signed by 1, otherwise 0. For each instance $\mathbf{x}_i$, its ground-truth label set $\mathbf{y}_i^*\in\{0,1\}^l$ is unknown, but it is covered by $\mathbf{y}_i$, *i.e.,* $\mathbf{y}_i^*\preceq\mathbf{y}_i$. The objective of PMLL is to induce a classifier over $\mathcal{D}$, which can predict the labels for unseen instances.

[1] https://github.com/daidai1118/WPML3CP/blob/main/WPML3CP_Appendix.pdf

4.1 Method Description

For each instance $\mathbf{x}_i$, its (normalized) ground-truth confidence is defined by $\mathbf{p}_i\in\mathbb{R}^l$ such that $\mathbf{0}\preceq\mathbf{p}_i\preceq\mathbf{y}_i$, $\mathbf{p}_i^\top\mathbf{1}=1$, where $\mathbf{0}$ denotes the all-zero vector. And the label correlation matrix is denoted by $\mathbf{C}\in\mathbb{R}_+^{l\times l}$ with each $\mathbf{c}_{:,j}$ being the coefficient vector for label j with respect to labels l. The basic idea of WPML^{3}CP is to estimate precise $\mathbf{p}$ by capturing precise $\mathbf{C}$ from both measuring and modeling perspectives.

Specifically, *from the measuring perspective*, we specify each $\mathbf{p}_i$ as a discrete distribution, and measure the loss between it and the prediction model by exploiting the regularized Wasserstein distance $W_\lambda(\cdot)$. Toward this goal, we employ a normalized linear model $f(\mathbf{x}_i|\mathbf{W})=\mathbf{W}\mathbf{x}_i$, $\mathbf{W}=[\mathbf{w}_1,\cdots,\mathbf{w}_l]^\top\in\mathbb{R}^{l\times d}$ such that $(\mathbf{W}\mathbf{x}_i)^\top\mathbf{1}=1$, ensuring the prediction also as a discrete distribution. *From the modeling perspective*, we form a label correlation-aware manifold regularization term to constrain the parameter $\mathbf{W}$ of the model. Upon the above ideas, we formulate the objective of WPML^{3}CP with a squared Frobenius norm of $\mathbf{W}$ as follows:

$$\begin{aligned}\min_{\mathbf{W},\mathbf{P},\mathbf{C}}\ &\sum_{i=1}^{n}W_\lambda\big(\mathbf{p}_i,\mathbf{W}\mathbf{x}_i;\mathfrak{m}(\mathbf{C})\big)\\&+\frac{\beta_1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l}\mathbf{C}_{ij}\|\mathbf{w}_i-\mathbf{w}_j\|_2^2+\frac{\beta_2}{2}\|\mathbf{W}\|_F^2\\ \mathbf{s.t.}\ &\mathbf{0}\preceq\mathbf{p}_i\preceq\mathbf{y}_i,\ \mathbf{p}_i^\top\mathbf{1}=1,\ (\mathbf{W}\mathbf{x}_i)^\top\mathbf{1}=1,\ \forall i\in[n],\end{aligned} \quad (4)$$

where $\mathfrak{m}(\cdot)=1-\text{sigmoid}(\cdot)$ is utilized to transform the label correlation matrix $\mathbf{C}$ to the cost matrix of Wasserstein distances; and $\{\beta_1,\beta_2\}$ are the regularization parameters.

Since the constraints of Eq.(4) tend to increase the optimization complexity, we leverage two surrogate heuristics to eliminate them. **First**, for the two equality constraints of $\mathbf{p}_i$ and $\mathbf{W}\mathbf{x}_i$, we exploit the softmax function $\mathfrak{s}(\cdot)$ to directly satisfy them. Specifically, for each $\mathbf{p}_i$ we introduce an auxiliary softmax parameter $\mathbf{q}_i$, *i.e.,* $\mathbf{p}_i=\mathfrak{s}(\mathbf{q}_i)$, and therefore learn $\mathbf{Q}=[\mathbf{q}_1,\cdots,\mathbf{q}_n]^\top\in\mathbb{R}^{n\times l}$, instead of $\mathbf{P}$. **Second**, for the inequality constraint of $\mathbf{p}_i$, we consider the irrelevant labels existed in the observed candidate label matrix $\mathbf{Y}=[\mathbf{y}_1,\cdots,\mathbf{y}_n]^\top$, and decompose it into the matrix $\mathbf{Q}$ and a noise matrix $\mathbf{E}$. Considering that the ground-truth labels are correlated in the multi-label learning case and the noise is always sparse in many real-world scenarios, we introduce a nuclear norm regularization for $\mathbf{Q}$ and ℓ_1-norm regularization for $\mathbf{E}$. Accordingly, we achieve the final objective of WPML^{3}CP with respect to $\{\mathbf{W},\mathbf{Q},\mathbf{C},\mathbf{E}\}$ below:

$$\begin{aligned}\min_{\mathbf{W},\mathbf{Q},\mathbf{C},\mathbf{E}}\ &\sum_{i=1}^{n}W_\lambda\big(\mathfrak{s}(\mathbf{q}_i),\mathfrak{s}(\mathbf{W}\mathbf{x}_i);\mathfrak{m}(\mathbf{C})\big)\\&+\frac{\beta_1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l}\mathbf{C}_{ij}\|\mathbf{w}_i-\mathbf{w}_j\|_2^2+\frac{\beta_2}{2}\|\mathbf{W}\|_F^2\\&+\beta_3\|\mathbf{Q}\|_*+\beta_4\|\mathbf{E}\|_1\\ \mathbf{s.t.}\quad&\mathbf{Y}=\mathbf{Q}+\mathbf{E}.\end{aligned} \quad (5)$$

Here, the equality constraint of Eq.(5) can be eliminated easily by employing the augmented Lagrange multiplier technique [Zhang *et al.*, 2017].

4.2 Optimization

The objective of WPML^{3}CP in Eq.(5) refers to four parameters of interest, *i.e.,* $\{\mathbf{W}, \mathbf{Q}, \mathbf{C}, \mathbf{E}\}$. We optimize them with the gradient decent approach under the augmented Lagrange multiplier framework. Specifically, the augmented Lagrangian of Eq.(5) is given with an auxiliary variable $\mathbf{H}$:

$$\begin{aligned}\min_{\mathbf{W},\mathbf{Q},\mathbf{C},\mathbf{E},\mathbf{H}} & \sum_{i=1}^{n} W_\lambda\big(\mathfrak{s}(\mathbf{h}_i), \mathfrak{s}(\mathbf{W}\mathbf{x}_i); \mathfrak{m}(\mathbf{C})\big) \\ & + \frac{\beta_1}{2}\mathrm{tr}(\mathbf{W}^\top\mathbf{L}\mathbf{W}) + \frac{\beta_2}{2}\|\mathbf{W}\|_F^2 + \beta_3\|\mathbf{Q}\|_* + \beta_4\|\mathbf{E}\|_1 \\ & + \langle \mathbf{Y}_1, \mathbf{Y} - \mathbf{Q} - \mathbf{E}\rangle + \frac{\mu_1}{2}\|\mathbf{Y} - \mathbf{Q} - \mathbf{E}\|_F^2 \\ & + \langle \mathbf{Y}_2, \mathbf{Q} - \mathbf{H}\rangle + \frac{\mu_2}{2}\|\mathbf{Q} - \mathbf{H}\|_F^2, \end{aligned} \tag{6}$$

where $\mathbf{L} = \mathrm{diag}(\mathbf{C1}) - \mathbf{C}$ is the Laplacian matrix of $\mathbf{C}$. According to the LADMAP method [Lin *et al.*, 2011], it can be rewritten as:

$$\begin{aligned}\min_{\mathbf{W},\mathbf{Q},\mathbf{C},\mathbf{E},\mathbf{H}} & \sum_{i=1}^{n} W_\lambda\big(\mathfrak{s}(\mathbf{h}_i), \mathfrak{s}(\mathbf{W}\mathbf{x}_i); \mathfrak{m}(\mathbf{C})\big) \\ & + \frac{\beta_1}{2}\mathrm{tr}(\mathbf{W}^\top\mathbf{L}\mathbf{W}) + \frac{\beta_2}{2}\|\mathbf{W}\|_F^2 + \beta_3\|\mathbf{Q}\|_* \\ & + \beta_4\|\mathbf{E}\|_1 + \frac{\mu_1}{2}\|\mathbf{Y} - \mathbf{Q} - \mathbf{E} + \frac{\mathbf{Y}_1}{\mu_1}\|_F^2 \\ & + \frac{\mu_2}{2}\|\mathbf{Q} - \mathbf{H} + \frac{\mathbf{Y}_2}{\mu_2}\|_F^2. \end{aligned} \tag{7}$$

Accordingly, we employ the gradient decent approach to optimize $\{\mathbf{W}, \mathbf{C}, \mathbf{H}\}$, whose gradients can be easily calculated with some simple derivations and the Sinkhorn algorithm, and update $\{\mathbf{Q}, \mathbf{E}\}$ as well as $\{\mathbf{Y}_1, \mathbf{Y}_2, \mu_1, \mu_2\}$ with the linear ADM method following [Liu *et al.*, 2010]. Due to the space limitation, the optimization details are presented **in the Appendix** [2] .

Initialization of label correlation matrix $\mathbf{C}$: We initialize the label correlation matrix $\mathbf{C}$ with the cosine similarities among label prototypes, which are calculated by averaging the training features $\mathbf{X} = [\mathbf{x}_1, \cdots, \mathbf{x}_n]^\top$ weighted with $\mathbf{Y}$. Specifically, given the label prototypes $\mathbf{O} = \mathbf{Y}^\top\mathbf{X} = [\mathbf{o}_1, \cdots, \mathbf{o}_l]^\top$, the initialization of the label correlation matrix $\mathbf{C}$ is given by:

$$\mathbf{C}_{ij} = \frac{\mathbf{o}_i^\top\mathbf{o}_j}{\|\mathbf{o}_i\|_2\|\mathbf{o}_j\|_2}, \quad \forall i, j \in [l].$$

5 Experiment

In this section, we empirically evaluate the proposed **WPML^{3}CP** method.

Datasets. To thoroughly evaluate WPML^{3}CP, we employ 5 real-world PMLL datasets,[3] including Music_emotion, Mirflickr, YeastBP, YeastCC, and YeastMF, and generate synthetic PMLL datasets by using 4 MLL datasets,[4] including Birds, Medical, Genbase, and Entertainment. For clarity, we present the detailed characteristics of all datasets in Table 1.

Dataset	$\lvert\mathcal{S}\rvert$	$dim(\mathcal{S})$	$L(\mathcal{S})$	$LCard(\mathcal{S})$	Domain
Music_emotion	6,833	98	11	2.42	Music
Mirflickr	10,433	100	7	1.77	Images
YeastBP	6,139	6,139	217	5.537	Biology
YeastCC	6,139	6,139	50	1.348	Biology
YeastMF	6,139	6,139	39	1.005	Biology
Birds	645	260	19	1.014	Audio
Medical	978	1,449	45	1.140	Text
Genbase	662	1,186	27	1.252	Biology
Entertainment	12,730	32,001	21	1.414	Text

Table 1: Detailed characteristics of datasets. $|\mathcal{S}|$: the number of instances of dataset $\mathcal{S}$. $dim(\mathcal{S})$: the number of features. $L(\mathcal{S})$: the number of category labels. $LCard(\mathcal{S})$: the average number of labels per instance.

Besides, the synthetic versions are generated by the following policies. For each instance, we randomly choose some irrelevant labels, and then mix them with the ground-truth ones, leading to the set of candidate labels. The number of irrelevant noisy labels is denoted by $m\%\|\mathbf{y}_i^*\|_0$, where the value of m is varied over $[50, 100, 150]$. Accordingly, in our experiments, we generate 12 synthetic PMLL datasets in total.

Baseline methods. We employ 7 the state-of-the-art PMLL methods for comparison. They include **PAMB** [Liu *et al.*, 2023], **GLC** [Sun *et al.*, 2022], **PMLNI** [Xie and Huang, 2021], **PML-MD** [Xie *et al.*, 2021], **PML-LRS** [Sun *et al.*, 2019], **PAR-MAP** and **PAR-VLS** [Zhang and Fang, 2021]. We set the parameters as suggested in their original papers. For our **WPML^{3}CP**, across Mirflickr and Music_emotion the parameters are set as: $\beta_1 = 1$, $\beta_2 = 10^{-3}$, $\beta_3 = 10^{-1}$, and $\beta_4 = 1$, and across YeastBP, YeastCC and YeastMF the parameters are set as: $\beta_1 = 10^{-3}$, $\beta_2 = 10^{-1}$, $\beta_3 = 10^2$, and $\beta_4 = 10^2$. Furthermore, about the parameter λ of the regularized Wasserstein distance, we set $\lambda = 1$ for Mirflickr and $\lambda = 50$ for others.

Evaluation metrics. We employ 5 widely-used MLL metrics to measure the performance, including *Average Precision* (AP), *Coverage Error* (CError), *Ranking Loss* (RLoss), *Macro Averaging F1* (Macro F1) and *Micro Averaging F1* (Micro F1). Those evaluation metrics cover both ranking- and binary-based ones and their detailed definitions can refer to the descriptions in [Zhang and Zhou, 2014b]. For CError and RLoss, the smaller value implies better performance (denoted by $\downarrow$); and for AP, Macro F1, and Micro F1, the larger value is better (denoted by $\uparrow$).

5.1 Results on Real-world Datasets

For each dataset, we conduct five-fold cross-validation and report the average results in Tables 2 and 3. We can clearly observe that WPML^{3}CP can outperform the baseline methods in most cases. Across all datasets and evaluation criteria on real-world datasets, WPML^{3}CP ranks *1st* in 80% and *2nd* in 12% cases. More detailed observations are made below.

Comparisons via the perspective of evaluation criteria: We can observe that, on the ranking based metrics (*i.e.,* AP, CError and RLoss), WPML^{3}CP ranks first in 73.3% cases. Specifically, WPML^{3}CP dominates all seven PML baselines on CError and performs similarly on AP and Rloss: It outper-

[2] https://github.com/daidai1118/WPML3CP/blob/main/WPML3CP_Appendix.pdf

[3] http://palm.seu.edu.cn/zhangml/

[4] http://mulan.sourceforge.net/datasets-mlc.html, https://github.com/fcharte/mldr.datasets

Dataset	WPML^{3}CP	PAMB	GLC	PML-NI	PML-MD	PML-LRS	PAR-MAP	PAR-VLS
				AP ↑				
YeastBP	**.552±.009**	.353±.206	.466±.016	.402±.007	.218±.013	.432±.002	.386±.203	.332±.012
YeastCC	**.836±.007**	.653±.133	.790±.018	.833±.008	.785±.009	.722±.007	.684±.121	.630±.014
YeastMF	**.779±.009**	.638±.111	.703±.008	.770±.001	.736±.016	.685±.009	.688±.108	.603±.016
Mirflickr	.858±.006	.753±.038	.765±.008	.789±.008	**.859±.007**	.793±.001	.795±.181	.673±.024
Music_emotion	.623±.004	**.642±.008**	.515±.005	.599±.007	.636±.012	.617±.008	.620±.010	.585±.008
				CError ↓				
YeastBP	**.281±.005**	.560±.189	.325±.012	.416±.011	.572±.010	.410±.008	.452±.193	.689±.012
YeastCC	**.090±.007**	.271±.119	.124±.012	.101±.003	.140±.011	.171±.004	.219±.104	.376±.015
YeastMF	**.104±.009**	.249±.096	.149±.008	.119±.010	.149±.016	.190±.005	.184±.085	.397±.017
Mirflickr	**.209±.004**	.276±.050	.289±.006	.229±.005	.220±.005	.231±.001	.242±.055	.315±.004
Music_emotion	**.405±.001**	.407±.007	.494±.004	.411±.007	.409±.007	.408±.006	.423±.006	.431±.007
				RLoss ↓				
YeastBP	**.146±.002**	.389±.119	.197±.008	.218±.005	.300±.013	.258±.005	.254±.100	.693±.012
YeastCC	**.063±.004**	.231±.103	.089±.009	.072±.004	.101±.012	.142±.002	.158±.068	.396±.015
YeastMF	**.088±.008**	.235±.083	.132±.005	.105±.007	.121±.012	.180±.006	.153±.062	.423±.017
Mirflickr	.094±.004	.147±.039	.185±.007	.123±.006	**.089±.005**	.122±.001	.134±.124	.230±.017
Music_emotion	.242±.003	**.234±.006**	.345±.006	.250±.008	.239±.011	.243±.005	.241±.005	.268±.007
				Macro F1 ↑				
YeastBP	**.339±.012**	.010±.006	.016±.001	.297±.020	.149±.009	.181±.008	.015±.007	.004±.002
YeastCC	**.536±.008**	.010±.004	.069±.004	.281±.017	.187±.025	.252±.021	.017±.006	.017±.002
YeastMF	**.404±.018**	.014±.007	.035±.006	.197±.013	.144±.036	.197±.013	.019±.007	.019±.009
Mirflickr	**.607±.016**	.567±.079	.157±.006	.564±.014	.533±.019	.571±.005	.346±.079	.480±.012
Music_emotion	**.429±.012**	.256±.008	.053±.001	.286±.013	.388±.023	.388±.010	.253±.009	.205±.006
				Micro F1 ↑				
YeastBP	**.384±.008**	.046±.026	.044±.006	.311±.018	.211±.006	.184±.007	.080±.028	.008±.004
YeastCC	**.569±.013**	.033±.007	.152±.013	.297±.018	.271±.022	.263±.021	.078±.019	.045±.006
YeastMF	**.418±.022**	.037±.014	.079±.009	.204±.014	.206±.044	.223±.014	.079±.007	.056±.008
Mirflickr	**.748±.008**	.743±.042	.553±.006	.685±.008	.748±.011	.659±.002	.531±.189	.667±.007
Music_emotion	.529±.007	.386±.005	.238±.003	.458±.012	**.538±.008**	.524±.009	.446±.009	.425±.008
Average Rank	**1.32**	5.68	5.32	3.28	3.36	3.96	5.36	7.52

Table 2: Experimental results of each comparing method (mean ± std) in terms of *AP*, *CError*, *RLoss*, *Macro F1* and *Micro F1*, where the best performance is shown in boldface.

forms most baselines on these criteria but slightly worse than PAMB in music_emotion and PML-MD in mirflickr. On binary prediction-based metrics (*i.e.,* Macro F1 and Micro F1), WPML^{3}CP ranks first in 90% cases. Specifically, WPML^{3}CP dominates all seven PML baselines under these two criteria, except in music_emotion where PML-MD outperforms WPML^{3}CP over Micro F1. This probably because that our method primarily guiding prediction model generation based on label correlation, resulting in many irrelevant labels sharing similar features being predicted as correlated. As a result, the binary prediction is perceived as more accurate but ranks slightly lower. In particular, PAMB performs very well in music_emotion on three ranking based metrics, but poorly on two binary prediction based metrics, which indicates that the binary decomposition technique still has some limitations for PMLL problems.

Comparisons via the perspective of datasets: We can observe that, on the datasets with a large number of features (i.e. YeastBP, YeastCC and YeastMF), WPML^{3}CP shows the best under all evaluation metrics. On datasets with a small number of features (i.e. Mirflickr and Music_emotion), the average rank of WPML^{3}CP ranks first, slightly better than that of PML-MD. This might be caused by the too large average number of labels per instance of Yeast compared with the number of category labels, leading to the very sparse label vector of each instance. Fortunately, our method fully considers the label correlation from both measuring and modeling.

5.2 Results on Synthetic Datasets

Across all datasets and evaluation criteria on synthetic datasets, WPML^{3}CP ranks *1st* in 76.7% and *2nd* in 20% cases. We can observe that WPML^{3}CP consistently outperforms all baselines for both medical and entertainment datasets in terms of all five evaluation metrics. For genbase, WPML^{3}CP only has a small performance gap compared with PAR-MAP on CError and RLoss metrics but still competitive. In addition, PAMB outperforms WPML^{3}CP over Macro F1 metric. One exception is in birds, where PML-NI outperforms our method over CError and RLoss evaluation metrics. This may be caused by the noise label identifier proposed in PML-NI, which leads to the improvement of performance. Besides, the performance gain of WPML^{3}CP tends to be more significant with the proportions of irrelevant labels increasing.

Additionally, we examine whether the performance gain of WPML^{3}CP is statistically significant. To answer this question, for each PMLL dataset and evaluation metric, we conduct a pairwise *t*-test over the results of the corresponding five train/test splits. The resulting win/tie/loss counts over 425 statistical tests (17 PMLL datasets × 5 evaluation metrics × 5-fold cross-validation) are reported in Table 4. In summary, out of all the tests, WPML^{3}CP significantly outperforms recent PMLL algorithms in 88.9% (PAMB), 96.5% (GLC), 77.4% (PML-NI), 88% (PML-MD), 91.8% (PML-LRS), 80.2% (PAR-MAP) and 96% (PAR-VLS) cases, respectively. This further indicates the effectiveness of WPML^{3}CP.

Dataset	AN.#Per	WPML^{3}CP	PAMB	GLC	PML-NI	PML-MD	PML-LRS	PAR-MAP	PAR-VLS
					AP ↑				
birds	50	.777±.030	.760±.030	.684±.035	**.786±.020**	.728±.035	.639±.040	.752±.024	.734±.027
	100	**.787±.024**	.743±.032	.682±.023	.781±.031	.745±.042	.671±.014	.737±.010	.745±.020
	150	**.775±.035**	.683±.024	.678±.023	.757±.036	.733±.025	.696±.021	.754±.018	.734±.029
genbase	50	**.477±.017**	.288±.044	.421±.044	.455±.020	.376±.059	.415±.034	.453±.099	.418±.107
	100	**.481±.020**	.211±.037	.420±.027	.451±.035	.381±.066	.422±.045	.458±.097	.426±.107
	150	**.466±.017**	.345±.055	.356±.050	.444±.037	.396±.058	.396±.053	.453±.097	.429±.113
medical	50	**.900±.028**	.486±.038	.657±.020	.888±.016	.820±.028	.494±.025	.810±.014	.746±.030
	100	**.894±.032**	.482±.039	.653±.028	.883±.013	.788±.042	.497±.020	.801±.025	.747±.028
	150	**.890±.031**	.482±.033	.658±.010	.863±.015	.779±.025	.492±.074	.813±.014	.711±.024
entertainment	50	**.697±.028**	–	.611±.006	.662±.009	.405±.058	.358±.003	.583±.011	.533±.006
	100	**.694±.011**	–	.610±.004	.650±.009	.445±.047	.345±.003	.570±.009	.522±.006
	150	**.691±.034**	–	.607±.005	.620±.005	.386±.053	.306±.005	.539±.004	.518±.005
					CError ↓				
birds	50	.135±.021	.147±.020	.187±.038	.136±.018	.166±.029	.289±.032	**.113±.021**	.171±.027
	100	.138±.019	.170±.027	.187±.008	.141±.020	.161±.025	.148±.021	**.108±.008**	.172±.019
	150	.145±.025	.193±.022	.194±.014	.151±.030	.163±.014	.197±.017	**.103±.014**	.179±.023
genbase	50	**.153±.020**	.322±.052	.196±.017	.176±.008	.192±.026	.247±.038	.156±.035	.222±.065
	100	**.147±.019**	.362±.064	.192±.017	.177±.024	.210±.020	.216±.053	.150±.035	.209±.054
	150	.163±.020	.319±.058	.215±.012	.196±.010	.217±.024	.273±.021	**.155±.034**	.197±.059
medical	50	**.030±.015**	.188±.015	.071±.008	.038±.002	.062±.014	.322±.028	.060±.008	.099±.011
	100	**.030±.013**	.198±.026	.077±.013	.036±.009	.067±.011	.307±.036	.074±.022	.113±.023
	150	**.034±.016**	.168±.019	.076±.018	.045±.006	.079±.010	.313±.058	.070±.013	.134±.017
entertainment	50	**.129±.012**	–	.149±.003	.172±.004	.284±.068	.400±.007	.171±.005	.206±.003
	100	**.128±.016**	–	.151±.003	.183±.001	.246±.050	.403±.004	.179±.002	.222±.002
	150	**.138±.014**	–	.153±.005	.199±.006	.261±.014	.425±.005	.199±.002	.229±.003
					RLoss ↓				
birds	50	.105±.019	.117±.009	.154±.022	.102±.015	.131±.018	.252±.028	**.090±.014**	.133±.017
	100	.105±.013	.131±.016	.156±.012	.108±.019	.127±.023	.205±.011	**.090±.004**	.132±.010
	150	.110±.016	.160±.012	.157±.012	.116±.020	.130±.010	.161±.011	**.083±.009**	.140±.013
genbase	50	**.136±.018**	.301±.047	.175±.019	.157±.009	.170±.019	.233±.040	.138±.029	.194±.053
	100	**.132±.013**	.354±.060	.171±.015	.156±.016	.184±.015	.200±.054	.133±.027	.192±.046
	150	.142±.016	.300±.052	.191±.013	.167±.010	.195±.026	.253±.016	**.135±.026**	.177±.051
medical	50	**.020±.011**	.172±.015	.058±.012	.026±.003	.048±.012	.293±.022	.045±.005	.080±.007
	100	**.020±.011**	.177±.023	.063±.010	.026±.007	.052±.011	.283±.035	.059±.010	.090±.017
	150	**.024±.012**	.151±.018	.061±.003	.033±.004	.061±.010	.033±.004	.053±.009	.105±.013
entertainment	50	**.092±.013**	–	.106±.002	.127±.003	.240±.070	.360±.007	.135±.005	.165±.002
	100	**.093±.018**	–	.106±.003	.138±.001	.204±.049	.365±.005	.141±.002	.179±.002
	150	**.100±.013**	–	.109±.002	.153±.004	.224±.014	.385±.004	.162±.002	.185±.003
					Macro F1 ↑				
birds	50	.267±.016	.173±.038	.031±.005	**.277±.041**	.211±.047	.097±.012	.189±.029	.099±.022
	100	**.305±.019**	.151±.019	.030±.005	.297±.043	.206±.034	.102±.019	.187±.028	.072±.016
	150	**.291±.046**	.105±.021	.025±.011	.213±.033	.205±.036	.179±.036	.205±.012	.104±.011
genbase	50	.272±.034	.243±.060	.209±.073	.219±.068	.253±.092	.256±.043	**.310±.067**	.287±.072
	100	.272±.033	.204±.050	.231±.059	.210±.029	.236±.045	.260±.079	**.309±.069**	.287±.072
	150	.273±.034	.226±.054	.233±.088	.258±.085	.257±.043	.244±.062	**.312±.071**	.287±.072
medical	50	**.368±.028**	.029±.007	.058±.004	.270±.020	.260±.018	.040±.003	.223±.032	.140±.016
	100	**.346±.016**	.029±.008	.053±.009	.252±.017	.215±.037	.040±.002	.207±.020	.141±.014
	150	**.341±.028**	.029±.008	.060±.007	.263±.011	.194±.017	.047±.008	.224±.017	.154±.019
entertainment	50	**.295±.021**	–	.101±.003	.236±.007	.109±.031	.160±.003	.179±.007	.152±.002
	100	**.290±.022**	–	.102±.002	.217±.004	.115±.031	.155±.000	.175±.006	.148±.006
	150	**.272±.034**	–	.101±.003	.206±.006	.105±.018	.127±.010	.184±.003	.150±.003
					Micro F1 ↑				
birds	50	**.483±.044**	.278±.028	.177±.029	.418±.034	.400±.053	.256±.035	.242±.017	.260±.030
	100	**.498±.045**	.256±.029	.174±.023	.399±.053	.428±.054	.297±.019	.249±.030	.193±.063
	150	**.480±.057**	.171±.032	.163±.063	.331±.052	.396±.063	.328±.037	.251±.025	.210±.018
genbase	50	**.264±.017**	.129±.031	.236±.042	.169±.110	.142±.086	.241±.026	.251±.116	.012±.001
	100	.245±.017	.130±.045	.236±.022	.191±.063	.176±.065	.243±.037	**.259±.111**	.012±.001
	150	**.277±.021**	.127±.021	.147±.072	.242±.039	.220±.067	.217±.062	.247±.108	.012±.001
medical	50	**.835±.036**	.291±.045	.442±.020	.758±.025	.734±.030	.390±.025	.583±.029	.433±.086
	100	**.830±.036**	.289±.042	.438±.031	.751±.025	.701±.049	.385±.016	.585±.030	.460±.051
	150	**.820±.037**	.287±.044	.444±.022	.706±.025	.682±.031	.377±.060	.571±.016	.463±.085
entertainment	50	**.563±.042**	–	.361±.009	.392±.011	.256±.042	.242±.003	.376±.008	.347±.009
	100	**.556±.015**	–	.360±.004	.361±.008	.312±.063	.233±.002	.366±.008	.332±.008
	150	**.543±.040**	–	.361±.007	.334±.003	.243±.090	.188±.009	.338±.005	.324±.005
Average Rank		**1.40**	6.87	6.11	2.71	4.24	6.36	2.82	5.44

Table 3: Results of each comparing method (mean ± std) in terms of *AP*, *CError*, *RLoss*, *Macro F1* and *Micro F1*, where the best performance is shown in boldface. The notation *"AN.#Per"* denotes the percentage of ground-truth labels added to each instance as irrelevant noisy labels.

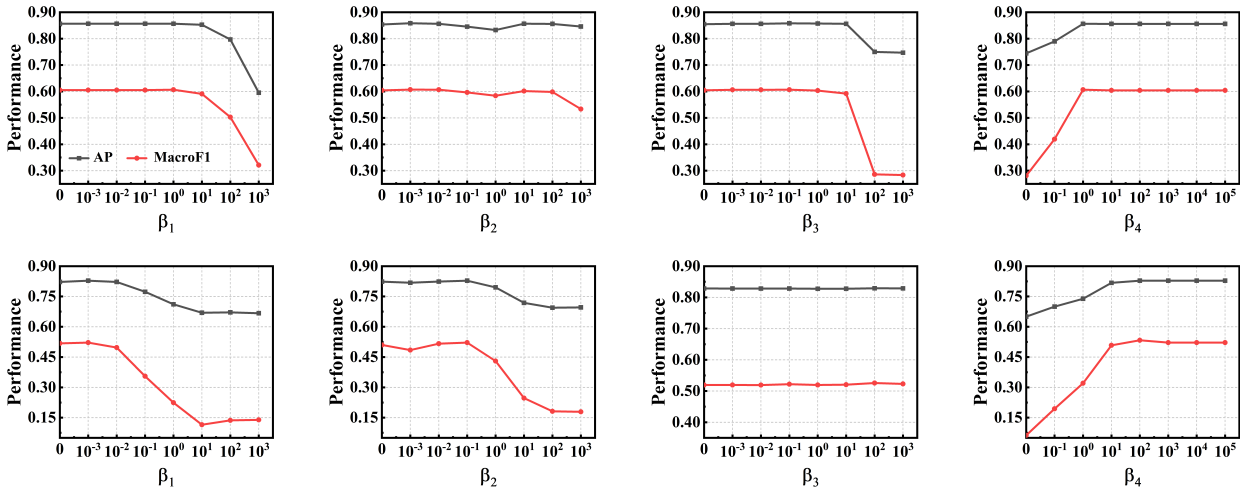

Figure 2: Sensitivity analysis of regularization parameters $\{\beta_1, \beta_2, \beta_3, \beta_4\}$ on two real PMLL datasets Mirflickr (top) and YeastCC (bottom).

Baseline Method	AP	CError	RLoss	Macro F1	Micro F1	Total
PAMB [Liu *et al.*, 2023]	**74/3/8**	**76/0/9**	**76/0/9**	**68/3/14**	**84/0/1**	**378/6/41**
GLC [Sun *et al.*, 2022]	**85/0/0**	**85/0/0**	**85/0/0**	**76/9/0**	**79/3/3**	**410/12/3**
PML-NI [Xie and Huang, 2021]	**54/7/24**	**55/5/25**	**60/3/22**	**78/6/1**	**82/0/3**	**329/21/75**
PML-MD [Xie *et al.*, 2021]	**72/6/7**	**82/3/0**	**77/1/7**	**70/15/0**	**73/8/4**	**374/33/18**
PML-LRS [Sun *et al.*, 2019]	**82/3/0**	**83/1/1**	**82/0/3**	**70/3/12**	**73/5/7**	**390/12/23**
PAR-MAP[Fang and Zhang, 2019]	**63/4/18**	**62/1/22**	**58/3/24**	**76/6/3**	**82/0/3**	**341/14/70**
PAR-VLS [Fang and Zhang, 2019]	**79/0/6**	**79/0/6**	**82/0/3**	**84/0/1**	**84/0/1**	**408/0/17**

Table 4: Win/tie/loss counts of pairwise t-test between WPML^{3}CP and each comparing method.

5.3 Parameter Evaluation

In this subsection, we evaluate the sensitivity of the regularization parameter $\{\beta_1, \beta_2, \beta_3, \beta_4\}$ of WPML^{3}CP. To this end, we examine the impact of varying $\{\beta_1, \beta_2, \beta_3\}$ values across the range $\{10^i | i = -3, \cdots, 3\}$ and varying β_4 values across the range $\{10^i | i = -1, \cdots, 5\}$ on two real PMLL datasets Mirflickr and YeastCC, measured by the ranking based metric (*i.e.,* AP) and the binary prediction based metric (*i.e.,* Macro F1). The results are presented in Figure 2. Roughly, we can observe that WPML^{3}CP tends to achieve higher scores with relatively smaller values of $\{\beta_1, \beta_2\}$ and higher value of β_4, and excepts a smaller (higher) value of β_3 on the dataset with a large (small) number of features. We empirically recommend selecting the values of $\{\beta_1, \beta_2\}$ from 10^{-3} to 10^1, β_3 inversely proportional to $dim(\mathcal{S})$ and β_4 from 10^{-1} to 10^3.

5.4 Ablation Study

In this subsection, we conduct ablative studies of different strategies (terms in the loss) by setting the corresponding regularization parameters to zero (see Figure 2). Specifically, $\{\beta_1 = 0, \beta_3 = 0, \beta_4 = 0\}$ represents the versions without the manifold regularization, $\|\mathbf{Q}\|_*$, and $\|\mathbf{E}\|_1$ terms, respectively. Furthermore, to validate the effectiveness of Wasserstein distance, we compare substituting it with Euclidean distance in Table 5, serving as an ablation experiment for the label correlation-aware Wasserstein distance. The results on all real datasets demonstrate the superiority of the Wasserstein distance over the Euclidean distance.

Dataset	Ver.	AP	CError	RLoss	MacroF1	MicroF1
YeastBP	Eucli	.370±.009	.504±.008	.286±.004	.031±.003	.114±.007
	Wass	**.552±.009**	**.281±.005**	**.146±.002**	**.339±.012**	**.384±.008**
YeastMF	Eucli	.650±.003	.215±.005	.199±.006	.024±.005	.065±.010
	Wass	**.779±.009**	**.104±.009**	**.088±.008**	**.404±.018**	**.418±.022**
YeastCC	Eucli	.661±.008	.251±.014	.195±.005	.035±.011	.091±.016
	Wass	**.836±.007**	**.090±.007**	**.063±.004**	**.536±.008**	**.569±.013**
Mirflickr	Eucli	.790±.009	.228±.005	.122±.006	.577±.010	.659±.008
	Wass	**.858±.006**	**.209±.004**	**.094±.004**	**.607±.016**	**.748±.008**
Music emotion	Eucli	.617±.004	.408±.003	.243±.003	.389±.008	.526±.004
	Wass	**.623±.004**	**.405±.001**	**.242±.003**	**.429±.012**	**.529±.007**

Table 5: Ablation study results about Wasserstein distance.

6 Conclusion

In this paper, we investigate the paradigm of PMLL with noise supervised signals. A novel PMLL method, namely WPML^{3}CP, is proposed. In WPML^{3}CP, we specify the ground-truth confidence as the normalized discrete distribution, *i.e.,* latent label distribution, to describe the probability of a candidate label being a ground-truth one. Considering it as a latent distribution variable, we jointly learn it and a normalized linear prediction model by minimizing the robust regularized Wasserstein distance. To reduce the optimization complexity, we eliminate the normalized constraints by exploiting the softmax function, leading to a surrogate objective without any constraint. The gradient descent with Adam is used to solve it. Extensive experimental results demonstrate that WPML^{3}CP outperforms the state-of-the-art baseline algorithms, and specifically it works well with high proportions of irrelevant labels.

Acknowledgements

We would like to acknowledge support for this project from the National Science and Technology Major Project (No.2021ZD0112501), the National Natural Science Foundation of China (No.62276113), and China Postdoctoral Science Foundation (No.2022M721321).

References

[Arjovsky *et al.*, 2017] Martín Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *ICML*, pages 214–223, 2017.

[Bogachev and Kolesnikov, 2012] Vladimir Igorevich Bogachev and Aleksandr Viktorovich Kolesnikov. The monge-kantorovich problem: Achievements, connections, and perspectives. *Russian Mathematical Surveys*, 67(5):785–890, 2012.

[Cuturi and Doucet, 2014] Marco Cuturi and Arnaud Doucet. Fast computation of wasserstein barycenters. In *ICML*, pages 685–693, 2014.

[Cuturi, 2013] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *NeurIPS*, pages 2292–2300, 2013.

[Fang and Zhang, 2019] Jun-Peng Fang and Min-Ling Zhang. Partial multi-label learning via credible label elicitation. In *AAAI*, pages 3518–3525, 2019.

[Feng *et al.*, 2022] Lei Feng, Jun Huang, Senlin Shu, and Bo An. Regularized matrix factorization for multilabel learning with missing labels. *IEEE TCYB*, 52(5):3710–3721, 2022.

[Frogner *et al.*, 2015] Charlie Frogner, Chiyuan Zhang, Hossein Mobahi an Mauricio Araya-Polo, and Tomaso Pogglo. Learning with a wasserstein loss. In *NeurIPS*, pages 2053–2061, 2015.

[Hang and Zhang, 2023] Jun-Yi Hang and Min-Ling Zhang. Partial multi-label learning with probabilistic graphical disambiguation. In *NeurIPS*, 2023.

[Huang *et al.*, 2016] Gao Huang, Chuan Guo, Matt J. Kusner, Yu Sun, Kilian Q. Weinberger, and Fei Sha. Supervised word mover's distance. In *NeurIPS*, pages 4862–4870, 2016.

[Kusner *et al.*, 2015] Matt J. Kusner, Yu Sun, Nicholas I. Kolkin, and Kilian Q. Weinberger. From word embeddings to document distances. In *ICML*, pages 957–966, 2015.

[Li and Wang, 2020] Ximing Li and Yang Wang. Recovering accurate labeling information from partially valid data for effective multi-label learning. In *IJCAI*, pages 1373–1380, 2020.

[Li *et al.*, 2020] Ziwei Li, Gengyu Lyu, and Songhe Feng. Partial multi-label learning via multi-subspace representation. In *IJCAI*, pages 2612–2618, 2020.

[Li *et al.*, 2021] Changchun Li, Ximing Li, and Jihong Ouyang. Semi-supervised text classification with balanced deep representation distributions. In *ACL-IJCNLP*, pages 5044–5053, 2021.

[Li *et al.*, 2022] Changchun Li, Ximing Li, Lei Feng, and Jihong Ouyang. Who is your right mixup partner in positive and unlabeled learning. In *ICLR*, 2022.

[Li *et al.*, 2023] Ximing Li, Yuanzhi Jiang, Changchun Li, Yiyuan Wang, and Jihong Ouyang. Learning with partial labels from semi-supervised perspective. In *AAAI*, pages 8666–8674, 2023.

[Lin *et al.*, 2011] Zhouchen Lin, Risheng Liu, and Zhixun Su. Linearized alternating direction method with adaptive penalty for low-rank representation. In *NeurIPS*, pages 612–620, 2011.

[Liu *et al.*, 2010] Guangcan Liu, Zhouchen Lin, and Yong Yu. Robust subspace segmentation by low-rank representation. In *ICML*, pages 663–670, 2010.

[Liu *et al.*, 2023] Bing-Qing Liu, Bin-Bin Jia, and Min-Ling Zhang. Towards enabling binary decomposition for partial multi-label learning. *IEEE TPAMI*, 2023.

[Pele and Werman, 2009] Ofir Pele and Michael Werman. Fast and robust earth mover's distances. In *ICCV*, pages 460–467, 2009.

[Rubner *et al.*, 2000] Yossi Rubner, Carlo Tomasi, and Leonidas J. Guibas. The earth mover's distance as a metric for image retrieval. *IJCV*, 40(2):99–121, 2000.

[Sun *et al.*, 2019] Lijuan Sun, Songhe Feng, Tao Wang, Congyan Lang, and Yi Jin. Partial multi-label learning with low-rank and sparse decomposition. In *AAAI*, pages 5016–5023, 2019.

[Sun *et al.*, 2022] Lijuan Sun, Songhe Feng, Jun Liu, Gengyu Lyu, and Congyan Lang. Global-local label correlation for partial multi-label learning. *IEEE Transactions on Multimedia*, 24:581–593, 2022.

[Villani, 2008] Cédric Villani. *Optimal Transport: Old and New*. Springer Berlin Heidelberg, 2008.

[Wang *et al.*, 2023] Haobo Wang, Shisong Yang, Gengyu Lyu, Weiwei Liu, Tianlei Hu, Ke Chen, Songhe Feng, and Gang Chen. Deep partial multi-label learning with graph disambiguation. *arXiv preprint arXiv:2305.05882*, 2023.

[Xie and Huang, 2018a] Ming-Kun Xie and Sheng-Jun Huang. Partial multi-label learning. In *AAAI*, pages 4302–4309, 2018.

[Xie and Huang, 2018b] Ming-Kun Xie and Sheng-Jun Huang. Partial multi-label learning. In *AAAI*, pages 4302–4309, 2018.

[Xie and Huang, 2021] Ming-Kun Xie and Sheng-Jun Huang. Partial multi-label learning with noisy label identification. *IEEE TPAMI*, 44(7):3676–3687, 2021.

[Xie *et al.*, 2021] Ming-Kun Xie, Feng Sun, and Sheng-Jun Huang. Partial multi-label learning with meta disambiguation. In *ACM SIGKDD*, pages 1904–1912, 2021.

[Xie *et al.*, 2023] Ming-Kun Xie, Jia-Hao Xiao, Hao-Zhe Liu, Gang Niu, Masashi Sugiyama, and Sheng-Jun Huang. Class-distribution-aware pseudo-labeling for semi-supervised multi-label learning. In *NeurIPS*, 2023.

[Xu *et al.*, 2020] Ning Xu, Yun-Peng Liu, and Xin Geng. Partial multi-label learning with label distribution. In *AAAI*, pages 6510–6517, 2020.

[Yu *et al.*, 2018] Guoxian Yu, Xia Chen, Carlotta Domeniconi, Jun Wang, Zhao Li, Zili Zhang, and Xindong Wu. Feature-induced partial multi-label learning. In *IEEE ICDM*, pages 1398–1403, 2018.

[Zhang and Fang, 2021] Min-Ling Zhang and Jun-Peng Fang. Partial multi-label learning via credible label elicitation. *IEEE TPAMI*, 43(10):3587–3599, 2021.

[Zhang and Zhou, 2014a] Min-Ling Zhang and Zhi-Hua Zhou. A review on multi-label learning algorithms. *IEEE TKDE*, 26(8):1819–1837, 2014.

[Zhang and Zhou, 2014b] Min-Ling Zhang and Zhi-Hua Zhou. A review on multi-label learning algorithms. *IEEE TKDE*, 26(8):1819–1837, 2014.

[Zhang *et al.*, 2017] Yongqiang Zhang, Daming Shi, Junbin Gao, and Dansong Cheng. Low-rank-sparse subspace representation for robust regression. In *IEEE CVPR*, pages 2972–2981, 2017.

[Zhao and Zhou, 2018] Peng Zhao and Zhi-Hua Zhou. Label distribution learning by optimal transport. In *AAAI*, pages 4506–4513, 2018.