

Article

Not peer-reviewed version

Partial Multi-Label Learning with Missing Labels via Feature-Label Disentanglement

Yuzhi Tao and [Anhui Tan](#) *

Posted Date: 23 March 2026

doi: 10.20944/preprints202603.1692.v1

Keywords: multi-label learning; missing labels; noisy labels; ;abel correlation; label-specific features



Preprints.org is a free multidisciplinary platform providing preprint service that is dedicated to making early versions of research outputs permanently available and citable. Preprints posted at Preprints.org appear in Web of Science, Crossref, Google Scholar, Scilit, Europe PMC.

Copyright: This open access article is published under a [Creative Commons CC BY 4.0 license](#), which permit the free download, distribution, and reuse, provided that the author and preprint are cited in any reuse.

Disclaimer/Publisher's Note: The statements, opinions, and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions, or products referred to in the content.

Article

Partial Multi-label Learning with Missing labels via Feature-Label Disentanglement

Yuzhi Tao ¹ and Anhui Tan ^{2,*}

¹ College of Business Administration, Huaqiao University, Quanzhou 362021 China

² School of Mathematical Sciences, Huaqiao University, Quanzhou 362021 China

* Correspondence: 20100@hqu.edu.cn; Tel.: +86-17805808846

Abstract

Partial multi-label learning addresses scenarios where each instance is associated with a set of candidate labels that include both relevant and irrelevant ones. In practical scenarios, such label sets are often simultaneously incomplete and noisy, which severely hampers the ability of models to extract compact and discriminative features. To address these issues, we propose an integrated learning paradigm that simultaneously enhances feature compactness and improves robustness against label noise. Our method constructs a similarity graph through the fuzzy C-means algorithm to capture the intrinsic relationships among instances. The resulting graph enables reliable label propagation, which effectively rectifies incorrect annotations and infers missing labels. In addition, we introduce a feature disentanglement mechanism that isolates reliable label-related feature representations from spurious ones introduced by noisy supervision. By integrating feature learning and label refinement into a joint optimization process, the proposed approach achieves a synergistic improvement in both representation quality and label reliability. Extensive theoretical analysis and empirical studies on multiple benchmark datasets demonstrate that our framework consistently outperforms state-of-the-art methods in terms of accuracy, stability and robustness to annotation noise.

Keywords: multi-label learning; missing labels; noisy labels; label correlation; label-specific features

1. Introduction

Multi-label learning (MLL) [1] aims to assign multiple semantic labels to each instance, enabling models to capture the complex correlations among categories. Traditional MLL algorithms usually assume that all training samples are annotated with complete and noise-free label sets. However, this assumption rarely holds in realistic scenarios, where label annotations are often noisy or incomplete due to ambiguous visual content, low-quality data, or inconsistent human judgments. The presence of such imperfect annotations can significantly degrade the performance of conventional MLL models, as they tend to overfit to unreliable supervision. To mitigate this problem, partial multi-label learning (PMLL) [2] has emerged as a practical paradigm for weakly supervised learning. In the PMLL setting, each instance is accompanied by a set of candidate labels that includes both relevant and irrelevant ones. The learning task is to identify the true labels from these candidates while simultaneously learning a robust predictive model. Although this framework alleviates the dependency on exhaustive manual labeling, it introduces new challenges—specifically, how to recover missing labels and suppress the adverse effects of false-positive annotations.

A straightforward strategy for PMLL treats all candidate labels as correct and directly applies standard MLL techniques. Such an approach may yield acceptable results when noisy labels are scarce, but it quickly deteriorates when noise becomes dominant. To address this limitation, recent studies have reformulated PMLL as a latent label inference problem, iteratively refining the candidate label set through probabilistic or graph-based modeling. These methods have demonstrated notable improvements by leveraging structural dependencies among instances to propagate label information and reduce uncertainty.

In real-world learning scenarios, obtaining a completely labeled dataset is often infeasible because manual annotation demands extensive human effort and domain expertise. Annotators may inadvertently neglect certain relevant categories or misassign irrelevant ones, resulting in label sets that are both incomplete and corrupted by noise. Such imperfect supervision has become increasingly common across diverse domains, including medical diagnosis, multimedia retrieval, and social media content tagging. In medical image analysis, clearly discernible pathological patterns allow clinicians to make confident diagnoses. Yet, when the manifestations are subtle or ambiguous, establishing a clear diagnosis becomes difficult, leading to uncertain or missing annotations that may require further expert validation. Similar issues occur in other application areas, where annotation reliability is easily affected by semantic ambiguity or poor data quality.

These practical constraints have motivated growing interest in learning paradigms that integrate the advantages of PMLL and MLL with incomplete supervision, aiming to enhance both robustness and generalization. The goal of such approaches is to construct models that are both robust to noisy labels and capable of generalizing effectively to unseen data. A central challenge is that label noise is seldom purely random. Instead, errors are often correlated with specific regions of the feature space, arising from ambiguous patterns, overlapping semantic categories, or low-quality input data. However, most existing approaches concentrate primarily on correcting or propagating candidate labels, while paying limited attention to learning compact, discriminative representations capable of disentangling reliable label-specific information from misleading features introduced by noisy annotations. Specifically, Zhong et al. [3] proposed the exploration of misleading features from noisy annotations and the disambiguation driven by negative label and noise information. However, the method relies heavily on manually constructed negative label information as the main guidance, which only focuses on disambiguating label-specific local features, whereas fails to integrate global cross-label features and local label-specific features. Moreover, existing approaches often neglect the simultaneous presence of missing and incorrect annotations, as well as the critical task of extracting reliable and informative features. Chen et al. [4] proposed a universal approach for unifying the handling of label noise and incompleteness in multi-label scenarios, where the compatibility with incomplete labels is achieved through the generalization of noise-handling mechanisms. However, the model adopts a deterministic feature extraction module that lacks the ability to quantify the uncertainty of extracted features, making it difficult to distinguish between reliable informative features and noisy irrelevant ones. To address these challenges, in this paper, we propose a novel integrated learning paradigm that jointly enhances feature compactness and robustness against label noise. Specifically, our framework constructs a similarity graph among instances using the fuzzy C-means algorithm, enabling effective label propagation and correction. In addition, we introduce a feature disentanglement mechanism to separate reliable label-specific representations from misleading information caused by noisy annotations. By integrating feature learning and label refinement within a unified optimization process, the proposed approach improves both representation quality and label reliability. Extensive experiments on multiple benchmark datasets demonstrate that our method consistently outperforms state-of-the-art approaches in terms of accuracy, stability, and robustness to annotation noise. The main contributions of this work are summarized as follows:

- We propose a unified learning paradigm that simultaneously promotes compact, discriminative feature representations and mitigates the impact of label noise, addressing the dual challenges of missing and incorrect annotations in PMLL.
- By constructing a similarity graph among instances using the fuzzy C-means algorithm, our framework effectively propagates label information, enabling the correction of inaccurate annotations and the inference of missing labels.
- We design a strategy to separate reliable label-specific features from misleading or noisy signals, allowing the model to focus on informative representations and reduce the adverse effects of corrupted supervision.

- The proposed approach integrates feature learning and label refinement into a single optimization process, ensuring synergistic improvements in both representation quality and label reliability.

The remainder of this paper is organized as follows. Section 2 reviews related work. Sections 3-4 present the proposed framework and detail the solution methodology. Section 5 reports experimental results and performance analysis. Finally, Section 6 concludes the paper and outlines future research directions.

2. Related Work

This section reviews prior research relevant to our work, focusing on PMLL under incomplete and noisy annotations, and MLL with missing labels.

2.1. Partial Multi-Label Learning (PMLL)

Partial multi-label learning (PMLL) studies scenarios where each instance is associated with a set of candidate labels containing both correct and incorrect entries [5]. It can be viewed as a combination of traditional multi-label learning (MLL) [6] and partial label learning. The key challenge in PMLL is to identify the true labels while learning robust models under weak supervision.

A key line of research in PMLL focuses on disambiguating candidate labels through iterative refinement strategies. For example, PARTICLE [7] estimates label probabilities by leveraging neighborhood information, while Sun et al. [8] introduced a fuzzy similarity-based label enhancement approach to improve label reliability. PML-MT [9] employs a mutual teaching framework with dual self-ensembling networks for collaborative label refinement, and [10] extends PMLL to semi-supervised settings, enabling effective learning even when a portion of the training data is unlabeled. More recently, Li et al. [11] proposed calibrated label disambiguation, which integrates probabilistic estimation with iterative refinement to accurately separate true labels from spurious ones. Collectively, these approaches aim to enhance the quality of candidate labels prior to model training. Another significant category of methods focuses on joint modeling of latent feature-label subspaces to capture high-order semantic relationships and structural dependencies. Wang et al. [12] proposed a framework that jointly models feature and label latent spaces, improving robustness under partial supervision. Zhong et al. [3] introduced a noise-aware mechanism that leverages both reliable and unreliable label signals to guide disambiguation more effectively. Complementary studies [13] exploit semantic embeddings and label co-occurrence patterns to strengthen feature-label interactions, further enhancing representation learning. By enhancing label reliability through disambiguation and refinement, and by modeling structured feature-label relationships via latent subspaces, these strategies contribute to more robust and accurate multi-label prediction under weak supervision.

Recent research has focused on leveraging label confidence and structured feature refinement to improve model robustness under weak supervision. Han et al. [14] utilized label confidence to guide feature selection, emphasizing highly reliable features during training. Hang and Zhang [15] proposed label-specific feature correction, disentangling trustworthy label-related representations from misleading signals caused by noisy annotations. Jalali and Kasneci [17] proposed an expert selection multi-label classification method by adapting to instance-specific characteristics.

In parallel, methods addressing noisy and incomplete annotations have been developed. Yang et al. [18] introduced a strategy for noisy label removal, filtering unreliable annotations prior to model training, Qian et al. [19] designed a noise-tolerant broad learning system that combines label enhancement with dimensionality reduction to improve generalization, while Li et al. [20] proposed to generate pseudo-labels for the instances with missing or uncertain labels while preserving local geometric consistency. Collectively, these studies highlight the emerging trends of exploiting label confidence for feature selection and correction and developing noise-tolerant frameworks for more robust and generalizable learning under weak supervision.

2.2. Multi-Label Learning with Missing Labels

A seminal line of work focuses on matrix completion and low-rank modeling, which assume that the label matrix exhibits a low-rank structure due to inter-label correlations. Early studies, such as Bucak et al. [21] and Yu et al. [22], formulated MLL as a matrix recovery problem, estimating unobserved labels through joint feature–label reconstruction. Chen et al. [23] extended this idea by solving a Sylvester equation to propagate label confidences in semi-supervised settings, while Jain et al. [24] proposed scalable generative models capable of handling large-scale incomplete annotations. Kong et al. [25] further improved scalability through efficient optimization for incomplete label assignments. More recent research has explored nonlinear and hierarchical representations. Wang et al. [26] proposed a two-level nonlinear mapping fusion model, and Jiang et al. [27] performs missing label recovery and feature selection via label compression and local feature correlations. These approaches collectively advance matrix-based paradigms by enriching representational flexibility and incorporating uncertainty modeling.

Another major direction seeks to model semantic and structural dependencies among labels. Yang et al. [28] leveraged structured semantic correlations to propagate information between related labels, while Wu et al. [29] introduced mixed dependency graphs to jointly capture feature and label relations. Braytee et al. [30] further addressed class imbalance and incomplete label spaces via correlated multi-label classification. These methods emphasize the importance of exploiting label dependencies and relational structures for more accurate label recovery. In parallel, feature selection and subspace modeling have been widely explored to enhance discriminative representation learning under missing supervision. Ma and Chow [31] proposed label-specific feature selection combined with two-level label recovery, enabling the model to focus on label-relevant attributes. Yin et al. [32] designed a multi-scale fuzzy uncertainty measure for robust feature selection, and Dai et al. [33] introduced a weak-label fusion approach using fuzzy discernibility pairs to guide feature relevance estimation. Similarly, Sun et al. [34] employed multilabel fuzzy neighborhood rough sets with maximum relevance–minimum redundancy to ensure feature compactness and label consistency. These strategies demonstrate how integrating fuzzy reasoning and feature relevance modeling can improve MLL-ML performance under uncertain or incomplete supervision.

2.3. Multi-Label Learning with Noisy and Incomplete Supervision

Most prior studies in multi-label learning address either noisy labels or missing labels independently, and relatively few explicitly consider their coexistence. Recent work in weakly supervised MLL has highlighted the challenges posed by incomplete, noisy, or partial labels. Sun et al. [35] proposed one of the earliest unified frameworks for weakly supervised multi-label learning, jointly modeling label incompleteness and corruption through reconstruction-based consistency regularization. Ding et al. [36] decomposed noisy features and leveraged low-rank label structures to mitigate the interference of missing and corrupted annotations. Wei et al. [37] developed a safe prediction mechanism to reduce the influence of unreliable labels under weak supervision. Moreover, Fang et al. [38] proposed an online multi-label active learning algorithm that integrates uncertainty and diversity querying and exploits label correlations via a co-occurrence guided classifier chain.

Although significant progress has been made in multi-label learning with missing or noisy labels, several challenges remain. Few approaches explicitly integrate feature representation learning and label refinement to handle both missing and noisy labels simultaneously, especially in scenarios with weak supervision or partial annotations. To address these gaps, in this paper, we propose a unified learning framework that jointly enhances feature compactness and label reliability. Our approach constructs a similarity graph among instances using fuzzy C-means clustering to propagate and correct labels while employing a feature disentanglement mechanism to separate reliable label-specific representations from misleading or noisy signals. By integrating feature learning and label refinement in a single optimization process, the proposed method improves both representation quality and prediction robustness under weak supervision.

Table 1. Notations frequently used.

Notation	Meaning
n	Instance number
d	Feature size
q	Label number
m	Embedded feature size
$\mathbf{x} \in \mathbb{R}^d$	Feature vector
$\mathbf{X} \in \mathbb{R}^{n \times d}$	Feature matrix
$\mathbf{y} \in \mathbb{R}^q$	Label vector
$\mathbf{Y} \in \{-1, 0, 1\}^{n \times q}$	Observed label matrix, where '0' means unknown annotation
$\mathbf{S} \in \mathbb{R}^{n \times n}$	Fuzzy similarity matrix
$\mathbf{S}_i \in \mathbb{R}^{n \times 1}$	i th row vector of \mathbf{S}
$\mathbf{S}_j \in \mathbb{R}^{1 \times n}$	j th column vector of \mathbf{S}
S_{ij}	ij th entry of \mathbf{S}
$\mathcal{N}_k(\mathbf{x})$	Top k -nearest neighbors of \mathbf{x}
$\ \cdot\ _1$	ℓ_1 -norm
$\ \cdot\ _F$	Frobenius norm
$\ \cdot\ _2$	L2-norm
\circ	Hadamard product (entry-wise product)
\oslash	Hadamard division (entry-wise division)
$tr(\cdot)$	Trace norm
T	Transposition
$\langle \cdot \rangle$	Hilbert-Schmidt inner product
\mathbf{D}_v	Diagonal matrix of vector \mathbf{v}
$\mathbf{F} \in \mathbb{R}^{n \times q}$	Predicted label matrix
$\mathbf{J} \in \{0, 1\}^{n \times q}$	Label occurrence indicator matrix
$\mathbf{Y}_0 \in \mathbb{R}^{n \times q}$	Noisy label matrix
$\mathbf{T} \in \mathbb{R}^{d \times m}$	Embedding projection matrix
$\mathbf{P} \in \mathbb{R}^{m \times q}$	Latent mapping matrix
$\mathbf{W} \in \mathbb{R}^{d \times q}$	Noisy mapping matrix

3. The Proposed Methodology

3.1. Problem Statement and Notations

Consider a multi-label dataset $\mathcal{D} = (\mathbf{X}, \mathbf{Y})$ with a label set $L = \{l_1, l_2, \dots, l_q\}$, where $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n]^T \in \mathbb{R}^{n \times d}$ denotes the feature matrix, and $\mathbf{Y} \in \{-1, 0, 1\}^{n \times q}$ is the observed label matrix. Here, \top denotes the matrix transpose, and $\mathbf{x}_i \in \mathbb{R}^d$ represents the d -dimensional feature vector of the i -th instance. For the entry \mathbf{Y}_{ij} corresponding to the i -th instance and j -th label, we define:

$$\mathbf{Y}_{ij} = \begin{cases} 1, & \text{if } l_j \text{ is a candidate label for } \mathbf{x}_i, \\ -1, & \text{if } l_j \text{ is explicitly excluded for } \mathbf{x}_i, \\ 0, & \text{if the label is unobserved or missing (true label unknown).} \end{cases}$$

In this work, we aim to learn a mapping from \mathbf{X} to \mathbf{Y} that can accurately predict the true labels for each instance, despite the presence of missing (0) and potentially noisy (incorrect 1 or -1) annotations. Table 1 summarizes the key notations used throughout the paper.

3.2. Instance-level Graph Adjacency Matrix

To capture the pairwise similarity among instances, we define a graph adjacency matrix $\mathbf{S} \in \mathbb{R}^{n \times n}$. The similarity between two instances \mathbf{x}_i and \mathbf{x}_j is assumed to decrease as their squared Euclidean distance $\|\mathbf{x}_i - \mathbf{x}_j\|^2$ increases, meaning that closer points are more similar. To better reflect local consistency and account for uncertainty in the feature space, we employ fuzzy C-means clustering [39, 40], which assigns each instance a set of membership degrees across clusters rather than a single

hard label. Using these soft memberships, we construct a similarity graph that captures nuanced relationships between instances. Formally, this leads to the following optimization problem:

$$\begin{aligned} \min_{\mathbf{S}} \quad & \sum_{\mathbf{x}_j \in \mathcal{N}_k(\mathbf{x}_i)} \|\mathbf{x}_i - \mathbf{x}_j\|^2 \mathbf{S}_{ij}^r \\ \text{s.t.} \quad & \sum_{\mathbf{x}_j \in \mathcal{N}_k(\mathbf{x}_i)} \mathbf{S}_{ij} = 1, \forall i = 1, \dots, n \\ & \mathbf{S}_{ii} = 1, \forall i = 1, \dots, n \\ & \mathbf{S}_{ij} \geq 0, \forall i, j = 1, \dots, n \end{aligned} \quad (1)$$

where $\|\cdot\|$ denotes the vector norm, $\mathcal{N}_k(\mathbf{x}_i)$ denotes the set of k nearest neighbors of \mathbf{x}_i , and r is the fuzzifier controlling the degree of membership weighting. Let $\mathbf{N} \in \mathbb{R}^{n \times n}$ denote the instance distance matrix, which satisfies the following:

$$\mathbf{N}_{ij} = \begin{cases} \|\mathbf{x}_i - \mathbf{x}_j\|^2, & \mathbf{x}_j \in \mathcal{N}_k(\mathbf{x}_i) \\ \infty, & \text{else} \end{cases} \quad (2)$$

Equation (1) can equivalently be expressed in the following form:

$$\begin{aligned} \min_{\mathbf{S}} \quad & \sum_{\mathbf{x}_j \in \mathcal{N}_k(\mathbf{x}_i)} \mathbf{N}_{ij} \mathbf{S}_{ij}^r \\ \text{s.t.} \quad & \sum_{\mathbf{x}_j \in \mathcal{N}_k(\mathbf{x}_i)} \mathbf{S}_{ij} = 1, \forall i = 1, \dots, n \\ & \mathbf{S}_{ii} = 1, \forall i = 1, \dots, n \\ & \mathbf{S}_{ij} \geq 0, \forall i, j = 1, \dots, n \end{aligned} \quad (3)$$

We can express the Lagrangian function associated with Equation (3) as:

$$\min_{\mathbf{S}} \left\{ \sum_{\mathbf{x}_j \in \mathcal{N}_k(\mathbf{x}_i)} \mathbf{N}_{ij} \mathbf{S}_{ij}^r - \theta \left(\sum_{\mathbf{x}_j \in \mathcal{N}_k(\mathbf{x}_i)} \mathbf{S}_{ij} - 1 \right) - \sum_{\mathbf{x}_j \in \mathcal{N}_k(\mathbf{x}_i)} \mathbf{b}_j \mathbf{S}_{ij} \right\} \quad (4)$$

where θ and \mathbf{b} are the Lagrangian multipliers.

According to the Karush-Kuhn-Tucker (KKT) conditions, the optimal solution is given by:

$$\mathbf{S}_{ij} = \begin{cases} \frac{(\frac{1}{\mathbf{N}_{ij}})^{\frac{1}{r-1}}}{\sum_{\mathbf{x}_j \in \mathcal{N}_k(\mathbf{x}_i)} (\frac{1}{\mathbf{N}_{ij}})^{\frac{1}{r-1}}}, & \mathbf{x}_j \in \mathcal{N}_k(\mathbf{x}_i) \\ 0, & \text{otherwise} \end{cases} \quad (5)$$

Denote $\hat{\mathbf{S}} = \frac{1}{2}(\mathbf{S} + \mathbf{S}^\top)$ as the symmetric adjacency matrix representing pairwise instance relationships. This matrix serves as a guidance for subsequent structural modeling in the label space. Unlike conventional manifold learning approaches [41] that rely on predefined graphs for label propagation, we directly infer $\hat{\mathbf{S}}$ from the data via fuzzy clustering, allowing the model to adaptively capture the intrinsic and potentially scattered distribution of the instances.

3.3. Compact Feature-Label Collaboration

High-dimensional input data often include redundant or irrelevant features, which can degrade model performance. Therefore, it is essential to extract a compact set of features that capture the most informative and label-relevant patterns. Compact features are typically decorrelated, encoding distinct information that improves both interpretability and generalization. In addition, semantically abstract features, represented as embedding vectors, can jointly capture feature and label semantics.

To facilitate the learning of label-specific representations and promote effective collaboration between feature extraction and classifier training, we define the following optimization objective:

$$\begin{aligned} \min_{\mathbf{P}, \mathbf{T}, \mathbf{F}} & \frac{1}{2} \|\mathbf{XTP} - \mathbf{F}\|_F^2 + \frac{\alpha}{2} \|\mathbf{T}\|_F^2 \\ \text{s.t.} & \quad \mathbf{PP}^\top = \mathbf{I} \end{aligned} \quad (6)$$

where $\mathbf{F} \in \mathbb{R}^{n \times q}$ is the label confidence matrix, with each entry measuring the degree to which an instance possesses a given label. The parameter α is a balancing factor.

The objective in Equation (6) jointly optimizes three components, feature projection, label mapping, and semantic representation, under an orthogonality constraint. Specifically, the term $\|\mathbf{XTP} - \mathbf{F}\|_F^2$ enforces consistency between the transformed feature representations and the label confidence matrix, thereby encouraging \mathbf{T} and \mathbf{P} to collaboratively capture discriminative and label-relevant information. The regularization term $\frac{\alpha}{2} \|\mathbf{T}\|_F^2$ prevents overfitting by controlling the complexity of the transformation from the original high-dimensional space to the latent space. The constraint encourages \mathbf{P} to form a set of orthogonal semantic bases that disentangle label-related factors, thereby strengthening the interaction between features and labels and effectively suppressing redundant correlations. Consequently, the model achieves a principled balance between compactness and discriminability.

Through this formulation, a compact feature-label collaboration mechanism naturally emerges: the latent representation \mathbf{T} learns to capture intrinsic structural regularities of the input, while \mathbf{P} aligns these representations with the label semantics in a mutually reinforcing manner. The synergy between them promotes more robust generalization and allows the learned embeddings \mathbf{F} to manifest both compactness in the feature space and specificity in the label space.

3.4. Label Recovery via Graph Propagation

To mitigate the adverse effects of unknown label noise, we decompose the observed label matrix \mathbf{Y} into two complementary components: a reliable label confidence matrix $\mathbf{F} \in \mathbb{R}^{n \times q}$ and a sparse noise matrix $\mathbf{Y}_0 \in \mathbb{R}^{n \times q}$. Formally, the decomposition is expressed as

$$\mathbf{Y} = \mathbf{F} \circ \mathbf{J} + \mathbf{Y}_0,$$

where \circ denotes the Hadamard (element-wise) product, and $\mathbf{J} \in \{0, 1\}^{n \times q}$ is the binary indicator matrix of observed label occurrences, such that $\mathbf{J}_{ij} = 0$ if $\mathbf{Y}_{ij} = 0$, and $\mathbf{J}_{ij} = 1$ otherwise. Based on this formulation, the optimization problem can be defined as:

$$\begin{aligned} \min_{\mathbf{F}, \mathbf{Y}_0} & \quad \|\mathbf{Y}_0\|_1 \\ \text{s.t.} & \quad \mathbf{Y} = \mathbf{F} \circ \mathbf{J} + \mathbf{Y}_0, \\ & \quad \mathbf{Y}_0 = \mathbf{Y}_0 \circ \mathbf{J}, \\ & \quad \mathbf{Y}_0 \geq \mathbf{0}, \end{aligned} \quad (7)$$

where $\|\mathbf{Y}_0\|_1$ denotes the ℓ_1 -norm of \mathbf{Y}_0 , enforcing sparsity on the noisy label matrix to suppress spurious annotations. The constraint $\mathbf{Y}_0 = \mathbf{Y}_0 \circ \mathbf{J}$ restricts the noise to positions corresponding to observed labels, while $\mathbf{Y}_0 \geq \mathbf{0}$ guarantees that corrections are non-negative and remain within the feasible annotation range.

This formulation enables the model to isolate and suppress unreliable labels while preserving the informative confidence structure within \mathbf{F} . Once the denoised label confidences are obtained, graph propagation is employed to diffuse the recovered confidence values across semantically similar instances, thereby leveraging the manifold structure of the data. In this way, label recovery becomes a smooth, graph-regularized refinement process that restores latent label consistency and enhances overall annotation reliability. To align the instance-level fuzzy similarity with the label-level semantic

similarity, we introduce a graph regularization term based on the Laplacian of the similarity graph. Specifically, the following loss function is employed:

$$\text{tr}(\mathbf{F}^\top \mathbf{C}_S \mathbf{F}) = \frac{1}{2} \sum_{i,j} \widehat{\mathbf{S}}_{ij} \|\mathbf{F}_i - \mathbf{F}_j\|_2^2, \quad (8)$$

where $\mathbf{C}_S = \mathbf{D}_S - \widehat{\mathbf{S}}$ is the graph Laplacian constructed from the normalized similarity matrix $\widehat{\mathbf{S}}$, and \mathbf{D}_S is a diagonal degree matrix with entries $[\mathbf{D}_S]_{ii} = \sum_{j=1}^n \widehat{\mathbf{S}}_{ij}$.

Intuitively, Equation (8) enforces local smoothness in the label confidence space: if two instances \mathbf{x}_i and \mathbf{x}_j exhibit high fuzzy similarity in the feature space (i.e., large $\widehat{\mathbf{S}}_{ij}$), their corresponding label confidence vectors \mathbf{F}_i and \mathbf{F}_j are encouraged to be close to each other. This mechanism propagates similarity information from the instance manifold to the label domain, ensuring that semantically related samples share consistent label representations. Consequently, the graph Laplacian regularizer effectively bridges the feature and label spaces, leading to smoother and more reliable label recovery across the data manifold.

3.5. Ambiguous Feature Identification

Prior studies [42] have highlighted that noisy labels in real-world datasets are often not generated randomly. Rather, they are frequently induced by ambiguous or confounding content concealed in the feature space. To uncover this relationship, we model the dependence between features and sparse noisy labels by optimizing the following objective:

$$\min_{\mathbf{W}, \mathbf{Y}_0} \frac{1}{2} \|\mathbf{X}\mathbf{W} - \mathbf{Y}_0\|_F^2 + \beta \|\mathbf{W}\|_{2,1}, \quad (9)$$

where the reconstruction term $\|\mathbf{X}\mathbf{W} - \mathbf{Y}_0\|_F^2$ isolates ambiguous features that may cause labeling errors, and the $\ell_{2,1}$ -norm regularization $\|\mathbf{W}\|_{2,1}$ promotes row-wise sparsity, effectively selecting the most informative feature dimensions. The hyperparameter $\beta > 0$ balances these two objectives.

By integrating the previous formulations in Eqs. (6), (7), and (8) with Equation (9), we obtain a unified framework that simultaneously performs feature mapping, label consistency refinement, instance manifold learning, and ambiguous feature identification. The resulting joint optimization problem is expressed as:

$$\begin{aligned} \min_{\mathbf{T}, \mathbf{P}, \mathbf{W}, \mathbf{F}, \mathbf{Y}_0} & \frac{1}{2} \|\mathbf{X}\mathbf{T}\mathbf{P} - \mathbf{F}\|_F^2 + \frac{1}{2} \|\mathbf{X}\mathbf{W} - \mathbf{Y}_0\|_F^2 + \frac{\lambda}{2} \text{tr}(\mathbf{F}^\top \mathbf{C}_S \mathbf{F}) \\ & + \frac{\alpha}{2} \|\mathbf{T}\|_F^2 + \beta \|\mathbf{W}\|_{2,1} + \gamma \|\mathbf{Y}_0\|_1 \\ \text{s.t.} & \mathbf{Y} = \mathbf{F} \circ \mathbf{J} + \mathbf{Y}_0, \\ & \mathbf{Y}_0 = \mathbf{Y}_0 \circ \mathbf{J}, \\ & \mathbf{P}\mathbf{P}^\top = \mathbf{I}, \\ & \mathbf{Y}_0 \geq \mathbf{0}. \end{aligned} \quad (10)$$

By adopting a block coordinate descent strategy, each variable is iteratively updated while holding the others fixed, ensuring convergence to a stable solution. The detailed algorithmic procedure is provided in the subsequent section.

4. Solutions to the Optimization Problem

To solve the joint optimization problem in Equation (10), we first introduce the augmented Lagrangian function:

$$\begin{aligned} \mathcal{L}(\mathbf{T}, \mathbf{P}, \mathbf{W}, \mathbf{F}, \mathbf{Y}_0, \mathbf{A}) &= \frac{1}{2} \|\mathbf{XTP} - \mathbf{F}\|_F^2 + \frac{1}{2} \|\mathbf{XW} - \mathbf{Y}_0\|_F^2 + \frac{\lambda}{2} \text{tr}(\mathbf{F}^\top \mathbf{C}_5 \mathbf{F}) \\ &\quad + \frac{\alpha}{2} \|\mathbf{T}\|_F^2 + \beta \|\mathbf{W}\|_{2,1} + \gamma \|\mathbf{Y}_0\|_1 \\ &\quad + \langle \mathbf{A}, \mathbf{Y} - \mathbf{F} \circ \mathbf{J} - \mathbf{Y}_0 \rangle + \frac{\mu}{2} \|\mathbf{Y} - \mathbf{F} \circ \mathbf{J} - \mathbf{Y}_0\|_F^2 \end{aligned} \quad (11)$$

s.t. $\mathbf{Y}_0 = \mathbf{Y}_0 \circ \mathbf{J}$,
 $\mathbf{PP}^\top = \mathbf{I}$,
 $\mathbf{Y}_0 \geq \mathbf{0}$,

where $\mathbf{A} \in \mathbb{R}^{n \times q}$ is the Lagrange multiplier matrix and $\mu > 0$ is a trade-off parameter controlling the penalty for constraint violation.

Following the Linearized Alternating Direction Method with Adaptive Penalty (LADMAP) framework, Equation (11) can be equivalently reformulated as:

$$\begin{aligned} \mathcal{L}(\mathbf{T}, \mathbf{P}, \mathbf{W}, \mathbf{F}, \mathbf{Y}_0, \mathbf{A}) &= \frac{1}{2} \|\mathbf{XTP} - \mathbf{F}\|_F^2 + \frac{1}{2} \|\mathbf{XW} - \mathbf{Y}_0\|_F^2 + \frac{\lambda}{2} \text{tr}(\mathbf{F}^\top \mathbf{C}_5 \mathbf{F}) \\ &\quad + \frac{\alpha}{2} \|\mathbf{T}\|_F^2 + \beta \|\mathbf{W}\|_{2,1} + \gamma \|\mathbf{Y}_0\|_1 \\ &\quad + \frac{\mu}{2} \left\| \mathbf{Y} - \mathbf{F} \circ \mathbf{J} - \mathbf{Y}_0 + \frac{\mathbf{A}}{\mu} \right\|_F^2 \end{aligned} \quad (12)$$

s.t. $\mathbf{Y}_0 = \mathbf{Y}_0 \circ \mathbf{J}$,
 $\mathbf{PP}^\top = \mathbf{I}$,
 $\mathbf{Y}_0 \geq \mathbf{0}$.

This reformulation linearizes the augmented term, allowing each variable to be updated efficiently in a block coordinate descent manner while adaptively adjusting the penalty parameter μ . The LADMAP approach ensures convergence to a stable solution under standard convexity conditions, while respecting the orthogonality and sparsity constraints inherent in the problem.

4.1. Updating \mathbf{P} by Fixing Other Variables

The subproblem for \mathbf{P} is formulated as:

$$\begin{aligned} \min_{\mathbf{P}} \mathcal{L} &= \frac{1}{2} \|\mathbf{XTP} - \mathbf{F}\|_F^2 \\ \text{s.t. } \mathbf{PP}^\top &= \mathbf{I} \end{aligned} \quad (13)$$

The closed-form solution is obtained via singular value decomposition (SVD) of $\mathbf{F}^\top \mathbf{X} \mathbf{T} = \mathbf{B} \mathbf{\Sigma} \mathbf{R}^\top$, yielding $\mathbf{P} = \mathbf{R} \mathbf{B}^\top$.

4.2. Updating \mathbf{T} by Fixing Other Variables

Optimizing \mathbf{T} corresponds to solving:

$$\min_{\mathbf{T}} \mathcal{L} = \frac{1}{2} \|\mathbf{XTP} - \mathbf{F}\|_F^2 + \frac{\alpha}{2} \|\mathbf{T}\|_F^2. \quad (14)$$

The closed-form solution can be written as:

$$\mathbf{T} = \mathbf{U} \left[(\mathbf{U}^\top \mathbf{X}^\top \mathbf{F} \mathbf{P}^\top \mathbf{V}) \oslash (\Lambda_1 \mathbf{1}_d \mathbf{1}_m^\top \Lambda_2 + \alpha \mathbf{1}_d \mathbf{1}_m^\top) \right] \mathbf{V}^\top, \quad (15)$$

where \mathbf{U} and \mathbf{V} are the eigenvector matrices of $\mathbf{X}^\top \mathbf{X}$ and $\mathbf{P}\mathbf{P}^\top$, respectively, and Λ_1, Λ_2 are the corresponding diagonal eigenvalue matrices.

4.3. Updating \mathbf{W} by Fixing Other Variables

With $\mathbf{T}, \mathbf{P}, \mathbf{F}, \mathbf{Y}_0$ fixed, the subproblem for \mathbf{W} is:

$$\min_{\mathbf{W}} \mathcal{L} = \frac{1}{2} \|\mathbf{X}\mathbf{W} - \mathbf{Y}_0\|_{\mathbb{F}}^2 + \beta \|\mathbf{W}\|_{2,1}. \quad (16)$$

Its gradient is

$$\nabla_{\mathbf{W}} \mathcal{L} = \mathbf{X}^\top (\mathbf{X}\mathbf{W} - \mathbf{Y}_0) + \beta \Omega \mathbf{W}, \quad (17)$$

where $\Omega \in \mathbb{R}^{d \times d}$ is diagonal:

$$\Omega = \begin{bmatrix} \frac{1}{\|\mathbf{W}_1\|_2} & & \\ & \ddots & \\ & & \frac{1}{\|\mathbf{W}_d\|_2} \end{bmatrix} \quad (18)$$

Treating Ω as fixed at iteration t (denoted $\Omega^{(t)}$), the closed-form update for \mathbf{W} is:

$$\mathbf{W}^{(t+1)} = (\mathbf{X}^\top \mathbf{X} + \beta \Omega^{(t)})^{-1} \mathbf{X}^\top \mathbf{Y}_0. \quad (19)$$

4.4. Updating \mathbf{F} by Fixing Other Variables

The subproblem for \mathbf{F} is:

$$\min_{\mathbf{F}} \mathcal{L} = \frac{1}{2} \|\mathbf{X}\mathbf{T}\mathbf{P} - \mathbf{F}\|_{\mathbb{F}}^2 + \frac{\lambda}{2} \text{tr}(\mathbf{F}^\top \mathbf{C}_S \mathbf{F}) + \frac{\mu}{2} \|\mathbf{Y} - \mathbf{F} \circ \mathbf{J} - \mathbf{Y}_0 + \frac{\mathbf{A}}{\mu}\|_{\mathbb{F}}^2. \quad (20)$$

Setting the gradient to zero gives a linear system for each column $\mathbf{F}_{:,j}$:

$$(\mathbf{I} + \lambda \mathbf{C}_S + \mathbf{D}_{J,j}) \mathbf{F}_{:,j} = \mathbf{Q}_{:,j}, \quad 1 \leq j \leq q, \quad (21)$$

where $\mathbf{D}_{J,j} = \text{diag}(\mathbf{J}_{:,j})$ and $\mathbf{Q} = \mathbf{X}\mathbf{T}\mathbf{P} + \mathbf{Y} - \mu \mathbf{Y}_0 + \mathbf{A}$. Since the coefficient matrix is invariant across iterations, its inverse can be precomputed for efficiency.

4.5. Updating \mathbf{Y}_0 by Fixing Other Variables

The subproblem over \mathbf{Y}_0 is:

$$\begin{aligned} \min_{\mathbf{Y}_0} \mathcal{L} &= \frac{1}{2} \|\mathbf{X}\mathbf{W} - \mathbf{Y}_0\|_{\mathbb{F}}^2 + \gamma \|\mathbf{Y}_0\|_1 + \frac{\mu}{2} \|\mathbf{Y} - \mathbf{F} \circ \mathbf{J} - \mathbf{Y}_0 + \frac{\mathbf{A}}{\mu}\|_{\mathbb{F}}^2 \\ \text{s.t. } &\mathbf{Y}_0 = \mathbf{Y}_0 \circ \mathbf{J} \\ &\mathbf{Y}_0 \geq \mathbf{0} \end{aligned} \quad (22)$$

This admits the closed-form solution:

$$\mathbf{Y}_0 = \mathcal{S}_{\frac{\gamma}{1+\mu}}^+ \left[\frac{\mathbf{X}\mathbf{W} + \mu \mathbf{Y} - \mu \mathbf{F} \circ \mathbf{J} + \mathbf{A}}{1 + \mu} \right] \circ \mathbf{J}, \quad (23)$$

where $\mathcal{S}_w^+(a) = \max(a - w, 0)$ is the element-wise positive soft-thresholding operator.

Finally, the Lagrange multiplier \mathbf{A} and penalty parameter μ are updated as:

$$\begin{aligned} \mathbf{A}^{(t+1)} &= \mathbf{A}^{(t)} + \mu^{(t+1)} (\mathbf{Y} - \mathbf{F} \circ \mathbf{J} + \mathbf{Y}_0) \\ \mu^{(t+1)} &= \min(\mu_{\max}, \rho \mu^{(t)}) \end{aligned} \quad (24)$$

where $\rho > 0$ is a scalar controlling the adaptive increment of μ .

4.6. The Solution of Equation (14)

Following [43], Equation (14) can be solved in closed form for latent feature embedding. By setting the derivative of Equation (14) with respect to \mathbf{T} to zero, we obtain:

$$\mathbf{X}^\top \mathbf{X} \mathbf{T} \mathbf{P} \mathbf{P}^\top - \mathbf{X}^\top \mathbf{F} \mathbf{P}^\top + \alpha \mathbf{T} = \mathbf{0}. \quad (25)$$

Let the eigenvalue decompositions of the symmetric matrices $\mathbf{X}^\top \mathbf{X}$ and $\mathbf{P} \mathbf{P}^\top$ be

$$\mathbf{X}^\top \mathbf{X} = \mathbf{U} \Lambda_1 \mathbf{U}^\top, \quad \mathbf{P} \mathbf{P}^\top = \mathbf{V} \Lambda_2 \mathbf{V}^\top,$$

where $\mathbf{U} \in \mathbb{R}^{d \times d}$ and $\mathbf{V} \in \mathbb{R}^{m \times m}$ are orthonormal matrices of eigenvectors, and $\Lambda_1 \in \mathbb{R}^{d \times d}$, $\Lambda_2 \in \mathbb{R}^{m \times m}$ are diagonal matrices of corresponding eigenvalues.

Substituting these decompositions into Equation (25) yields

$$\mathbf{U} \Lambda_1 \mathbf{U}^\top \mathbf{T} \mathbf{V} \Lambda_2 \mathbf{V}^\top - \mathbf{X}^\top \mathbf{F} \mathbf{P}^\top + \alpha \mathbf{T} = \mathbf{0}. \quad (26)$$

Multiplying both sides by \mathbf{U}^\top on the left and \mathbf{V} on the right gives

$$\Lambda_1 (\mathbf{U}^\top \mathbf{T} \mathbf{V}) \Lambda_2 - \mathbf{U}^\top \mathbf{X}^\top \mathbf{F} \mathbf{P}^\top \mathbf{V} + \alpha (\mathbf{U}^\top \mathbf{T} \mathbf{V}) = \mathbf{0}. \quad (27)$$

Considering each i, j -th element, we have

$$(\Lambda_{1,ii} \Lambda_{2,jj} + \alpha) [\mathbf{U}^\top \mathbf{T} \mathbf{V}]_{ij} = [\mathbf{U}^\top \mathbf{X}^\top \mathbf{F} \mathbf{P}^\top \mathbf{V}]_{ij}. \quad (28)$$

Equivalently, we have in matrix form that:

$$\mathbf{U}^\top \mathbf{T} \mathbf{V} = \frac{\mathbf{U}^\top \mathbf{X}^\top \mathbf{F} \mathbf{P}^\top \mathbf{V}}{\Lambda_1 \mathbf{1}_d \mathbf{1}_m^\top \Lambda_2 + \alpha \mathbf{1}_d \mathbf{1}_m^\top}. \quad (29)$$

Finally, the closed-form solution for \mathbf{T} is

$$\mathbf{T} = \mathbf{U} \left(\frac{\mathbf{U}^\top \mathbf{X}^\top \mathbf{F} \mathbf{P}^\top \mathbf{V}}{\Lambda_1 \mathbf{1}_d \mathbf{1}_m^\top \Lambda_2 + \alpha \mathbf{1}_d \mathbf{1}_m^\top} \right) \mathbf{V}^\top. \quad (30)$$

This completes the proof. \square

4.7. Convergence Analysis

To theoretically verify the convergence of the proposed optimization framework, we first introduce a useful lemma.

Lemma 1. [40] Given any two real numbers $x, y \geq 0$, the following inequality holds:

$$x - \frac{x^2}{2y} \leq y - \frac{y^2}{2y}.$$

Theorem 1. Algorithm 1 monotonically decreases the objective function with respect to \mathbf{W} in each iteration.

Proof. Define the surrogate function for \mathbf{W} at iteration t as

$$\mathcal{F}(\mathbf{W}) = \frac{1}{2} \|\mathbf{X} \mathbf{W} - \mathbf{Y}_0\|_F^2 + \beta \text{tr}(\mathbf{W}^\top \Omega^{(t)} \mathbf{W}). \quad (31)$$

The gradient of $\mathcal{F}(\mathbf{W})$ with respect to \mathbf{W} is

$$\nabla_{\mathbf{W}} \mathcal{F}(\mathbf{W}) = \mathbf{X}^\top (\mathbf{X} \mathbf{W} - \mathbf{Y}_0) + 2\beta \Omega^{(t)} \mathbf{W}. \quad (32)$$

Let $\mathbf{W}^{(t+1)} = \arg \min_{\mathbf{W}} \mathcal{F}(\mathbf{W})$. Then, by definition,

$$\mathcal{F}(\mathbf{W}^{(t+1)}) \leq \mathcal{F}(\mathbf{W}^{(t)}). \quad (33)$$

By Lemma 1, for each row $\mathbf{W}_{i \cdot}$, we have

$$\|\mathbf{W}_{i \cdot}^{(t+1)}\|_2 - \frac{\|\mathbf{W}_{i \cdot}^{(t+1)}\|_2^2}{2\|\mathbf{W}_{i \cdot}^{(t)}\|_2} \leq \|\mathbf{W}_{i \cdot}^{(t)}\|_2 - \frac{\|\mathbf{W}_{i \cdot}^{(t)}\|_2^2}{2\|\mathbf{W}_{i \cdot}^{(t)}\|_2}. \quad (34)$$

Summing over all rows yields

$$\|\mathbf{W}^{(t+1)}\|_{2,1} - \frac{1}{2} \text{tr}(\mathbf{W}^{(t+1)\top} \Omega^{(t)} \mathbf{W}^{(t+1)}) \leq \|\mathbf{W}^{(t)}\|_{2,1} - \frac{1}{2} \text{tr}(\mathbf{W}^{(t)\top} \Omega^{(t)} \mathbf{W}^{(t)}). \quad (35)$$

Combining Eqs. (33) and (35), we obtain

$$\mathcal{F}(\mathbf{W}^{(t+1)}) + \beta \|\mathbf{W}^{(t+1)}\|_{2,1} \leq \mathcal{F}(\mathbf{W}^{(t)}) + \beta \|\mathbf{W}^{(t)}\|_{2,1}. \quad (36)$$

Equation (36) demonstrates that the objective function $\mathcal{L}(\mathbf{W})$ decreases monotonically at each iteration, completing the proof. \square

Algorithm 1 The Proposed WPML Algorithm

Require: Feature matrix \mathbf{X} , observed label matrix \mathbf{Y} , parameters $\lambda, \alpha, \beta, \gamma$, and the size of nearest neighbors k ;

Ensure: Learned model coefficient \mathbf{W} and label confidence matrix \mathbf{F} .

- 1: Compute \mathbf{S} according to Equation (5);
 - 2: $t \leftarrow 0$;
 - 3: Initialize $\mathbf{F}^{(0)} = \mathbf{Y}$, $\mathbf{T}^{(0)} = \mathbf{I}$, $\Omega^{(0)} = \mathbf{I}$, $\mathbf{Y}_0^{(0)} = \mathbf{0}$;
 - 4: Compute the indicator matrices \mathbf{J} and \mathbf{N} ;
 - 5: **Repeat**
 - 6: Update $\mathbf{P}^{(t+1)}$ according to Equation (13);
 - 7: Update $\mathbf{T}^{(t+1)}$ according to Equation (15);
 - 8: Update $\mathbf{W}^{(t+1)}$ according to Equation (19);
 - 9: Compute the diagonal matrix $\Omega^{(t+1)}$ according to Equation (18);
 - 10: Update $\mathbf{F}^{(t+1)}$ according to Equation (21);
 - 11: Update $\mathbf{Y}_0^{(t+1)}$ according to Equation (23);
 - 12: Update $\mathbf{A}^{(t+1)}$ and $\mu^{(t+1)}$ according to Equation (24);
 - 13: $t \leftarrow t + 1$;
 - 14: **Until** convergence;
 - 15: **Return** the results.
-

4.8. Complexity Analysis

We analyze the computational complexity of each major step in the proposed WPML algorithm. Specifically, solving \mathbf{P} involves SVD, resulting in to a complexity of $O(nqd + ndm + qm^2)$. The update of \mathbf{T} requires an eigendecomposition operation, with a cost of $O(d^3 + m^2)$. Updating \mathbf{W} incurs a complexity of $O(d^3 + ndq)$, while the computation of \mathbf{F} incurs a higher cost of $O(n^2q)$. Additionally, updating \mathbf{Y}_0 requires $O(ndq)$. In summary, the overall time complexity per iteration is $O(nqd + ndm + qm^2 + d^3 + m^2 + n^2q)$. It is worth noting that matrix inversion for solving \mathbf{W} and \mathbf{F} only needs to be performed once, which can be precomputed and reused, helping reduce the actual computational overhead in practice. Moreover, the dominant term $O(n^2q)$ arises from label propagation, suggesting that for large-scale datasets, approximations such as sparse neighbors or low-rank propagation could further accelerate computation without sacrificing performance.

Table 2. The information of datasets used in the experiment.

Data set	n	d	q	ℓ_c	ℓ_d
Birds	645	260	19	1.014	.053
Business	5000	438	30	1.588	.053
Bibtex	7395	1836	159	2.402	.015
CAL500	502	68	174	26.044	.150
Computers	5000	681	33	1.509	.048
Education	5000	550	33	1.461	.044
Emotions	593	72	6	1.868	.311
Entertainment	5000	640	21	1.414	.067
Eron	1702	1001	53	3.378	.064
Genbase	662	1186	27	1.252	.046
Health	5000	612	32	1.662	.052
Image	2000	294	5	1.236	.247
Medical	978	1449	45	1.245	.028
Reference	5000	793	33	1.169	.035
Scene	2407	294	6	1.074	.179
Yeast	2417	103	14	4.237	.303

5. Experiments

5.1. Experimental Setup

We evaluate the proposed WPML method on a benchmark collection of 16 publicly available multi-label datasets, sourced from the Mulan¹ and Uco² repositories. All feature values are normalized to the range $[0, 1]$ to ensure comparability across datasets. Table 2 summarizes the key characteristics of each dataset. In particular, n , d , and q denote the number of instances, features, and labels, respectively. We also report ℓ_c , the average number of labels per instance, and ℓ_d , the normalized label cardinality computed by dividing ℓ_c by the total number of labels q .

5.2. Compared Methods and Evaluation Metrics

We benchmark the proposed WPML algorithm against several state-of-the-art multi-label learning (MLL) methods that are capable of handling noisy or missing labels. All comparative algorithms are obtained from publicly available repositories, with hyperparameters configured according to the original papers. A brief description of each method and its hyperparameter setting is provided below:

- **WPML (Algorithm 1):** The hyperparameters λ and α are selected from $\{10^{-2}, 10^{-1}, \dots, 10^1\}$, β is tuned within $\{10^{-3}, 10^{-2}, \dots, 10^1\}$, and γ controlling noisy label sparsity is chosen from $\{0.1, 0.2, \dots, 1\}$. The neighbor size k is fixed at 10.
- **PML-NI³ [42]:** This partial multi-label learning method identifies noisy labels caused by ambiguous features. Parameters α and γ are set to 0.5, and λ is selected from $\{1, 10, 100\}$.
- **PML-VLS and PML-MAP⁴ [7]:** Both algorithms belong to the PARTICLA framework, using label propagation for initial label refinement, followed by virtual label splitting or MAP reasoning for final prediction. The credible label threshold is set to 0.9, with neighbor size $k = 10$.
- **DM2L-l and DM2L-nl⁵ [31]:** Designed to handle missing labels, these methods exploit local low-rank (DM2L-l) and global high-rank (DM2L-nl) structures using linear and Gaussian kernels, respectively. Regularization parameter λ_d is searched over 10^{-5} to 10^5 , with $\delta = 0.005$.
- **Glocal⁶ [45]:** Leverages global and local label correlations for mapping high-dimensional features to a low-rank label space. Parameters $\lambda = 1$, $\lambda_2 = 10^{-3}$, while λ_3 and λ_4 are tuned over $\{10^{-4}, \dots, 1\}$. Neighbor size k and group size g are selected from $\{5, 10, 15, 20\}$.

¹ <http://mulan.sourceforge.net/datasets.html>

² <http://www.uco.es/kdis/mlresources/>

³ <http://milkxie.github.io/code/PML-NIcode.zip>

⁴ <http://palm.seu.edu.cn/zhangml/files/PARTICLE.rar>

⁵ <https://github.com/John986/Multi-label-Learning-with-Missing-Labels>

⁶ <http://www.lamda.nju.edu.cn/Data.ashx>

- **ESMC**⁷ [46]: Employs non-linear embedding for tail label prediction. Hyperparameters λ , ρ , and σ_z are grid-searched over five orders of magnitude (10^1 to 10^5).
- **MSWL**⁸ [47]: Handles limited labeled data via a label manifold and semi-supervised learning. Hyperparameters α and β are evaluated over $\{10^{-3}, \dots, 10^3\}$, and γ is grid-searched in $\{10^{-4}, \dots, 10^4\}$.

To quantitatively assess performance, we adopt five widely used ranking-based metrics: *Macro Average AUC (AUC)*, *Coverage*, *Ranking Loss*, *Average Precision*, and *Hamming Loss*. Higher values of *AUC* and *Average Precision* indicate better performance, while lower values are preferable for *Hamming Loss*, *Coverage*, and *Ranking Loss*.

To simulate MLL scenarios with noisy or incomplete labels, we designed a controlled label perturbation strategy, which was applied uniformly to all datasets in our experiments. Specifically, for missing labels, each label in the training set was independently removed with a probability of 20%, 40% or 60%; removed labels were marked with '0' to indicate their missing status. For noisy labels, we corrupted the remaining non-missing labels at the same three rates 20%, 40% or 60% by randomly flipping the label value from -1 to 1. Due to space limitations, we only report the detailed results under the configuration with 60% missing and noisy labels, as the results under other proportion configurations are similar. A ten-fold cross-validation is conducted on each dataset. The mean and standard deviation of all evaluation metrics are reported in Tables 3–5 for comprehensive analysis.

Table 3. A comparative analysis of prediction performance among different algorithms in terms of *AUC*, with the best results (the larger, the better) highlighted in bold.

Methods	WPML	PML- NI	PML- VLS	PML- MAP	D2ML	Glocal	ESMC	MSWL
Birds	.846	.840	.692	.729	.765	.806	.728	.813
Business	.920	.766	.540	.538	.897	.775	.737	.663
Bibtex	.751	.728	.496	.496	.680	.660	.721	.560
CAL500	.746	.716	.514	.514	.709	.703	.597	.714
Computer	.867	.809	.559	.551	.823	.807	.777	.788
Education	.890	.822	.676	.661	.843	.801	.778	.826
Emotions	.718	.708	.502	.502	.686	.716	.594	.766
Entertainment	.858	.853	.591	.574	.817	.778	.807	.802
Eron	.877	.792	.501	.501	.793	.797	.846	.536
Genbase	.986	.985	.502	.502	.972	.980	.958	.984
Health	.926	.912	.636	.604	.888	.851	.882	.874
Image	.793	.781	.790	.786	.743	.797	.600	.762
Medical	.958	.918	.896	.896	.914	.919	.874	.650
Reference	.851	.874	.626	.620	.868	.857	.834	.857
Scene	.545	.357	.504	.504	.387	.559	.453	.767
Yeast	.653	.562	.505	.505	.618	.601	.564	.509

⁷ https://github.com/Akbarnejad/ESMC_Implementation

⁸ <https://jiazhang-ml.pub/MSWL-master.zip>

Table 4. A comparative analysis of prediction performance among different algorithms in terms of *Ranking Loss*, with the best results (the smaller, the better) highlighted in bold.

Methods	WPML	PML- NI	PML- VLS	PML- MAP	D2ML- nl	Glocal	ESMC	MSWL
Birds	.121	.126	.156	.137	.192	.157	.235	.152
Business	.051	.244	.115	.046	.077	.232	.270	.337
Bibtex	.234	.257	.304	.344	.325	.337	.269	.434
CAL500	.257	.284	.614	.194	.291	.297	.400	.286
Computer	.101	.164	.160	.111	.140	.180	.193	.206
Education	.104	.165	.093	.108	.150	.180	.209	.157
Emotions	.259	.272	.480	.461	.292	.266	.384	.209
Entertainment	.110	.114	.160	.162	.158	.190	.159	.168
Eron	.105	.187	.192	.127	.183	.186	.137	.464
Genbase	.003	.004	.009	.140	.012	.007	.022	.004
Health	.051	.063	.153	.108	.091	.119	.088	.196
Image	.165	.176	.198	.186	.214	.186	.369	.350
Medical	.037	.070	.140	.141	.074	.070	.116	.178
Reference	.124	.105	.159	.101	.117	.125	.143	.125
Scene	.426	.629	.427	.498	.593	.417	.531	.209
Yeast	.346	.432	.265	.272	.373	.391	.431	.499

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

Methods	WPML	PML- NI	PML- VLS	PML- MAP	D2ML- nl	Glocal	ESMC	MSWL
Birds	.723	.707	.695	.688	.569	.653	.563	.651
Business	.860	.573	.854	.866	.679	.536	.570	.468
Bibtex	.256	.207	.243	.223	.224	.239	.232	.116
CAL500	.443	.418	.382	.468	.418	.398	.289	.383
Computer	.654	.593	.513	.586	.544	.599	.573	.494
Education	.565	.541	.583	.543	.427	.543	.500	.556
Emotions	.726	.722	.537	.566	.702	.711	.634	.759
Entertainment	.664	.660	.515	.556	.465	.591	.594	.605
Eron	.636	.531	.480	.490	.504	.554	.630	.477
Genbase	.994	.994	.936	.945	.931	.990	.893	.994
Health	.757	.751	.660	.612	.611	.685	.719	.699
Image	.795	.779	.777	.782	.735	.554	.595	.763
Medical	.864	.776	.714	.797	.759	.794	.653	.606
Reference	.666	.658	.555	.595	.558	.660	.617	.662
Scene	.512	.348	.426	.418	.405	.507	.393	.759
Yeast	.619	.572	.523	.627	.584	.593	.564	.575

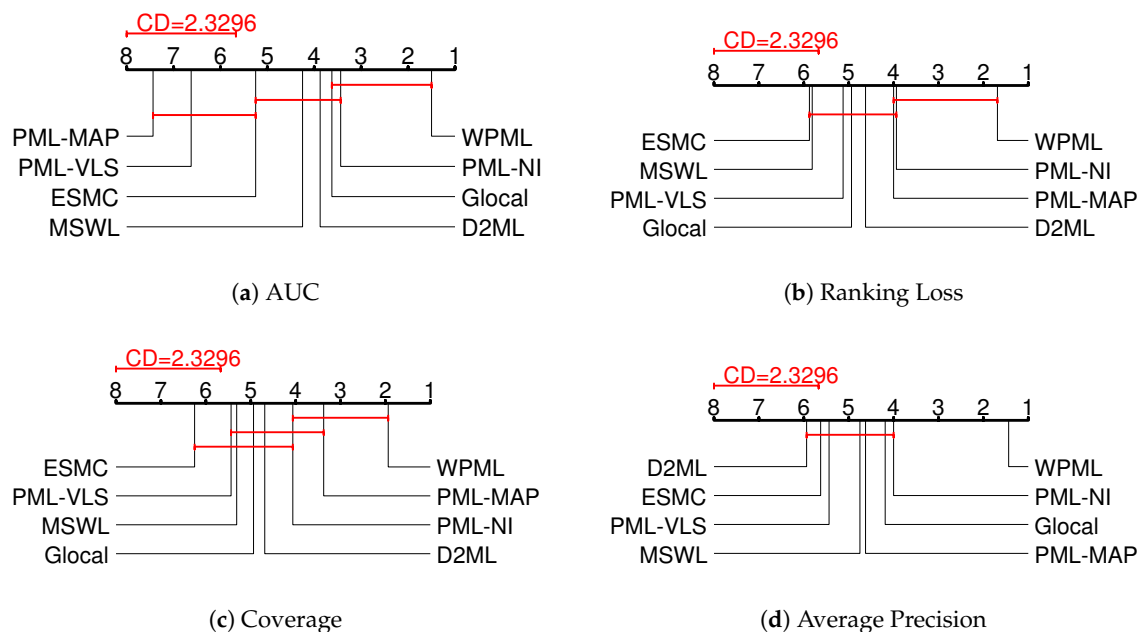


Figure 1. Comparison of WPML against algorithms under comparison with the Bonferroni-Dunn test ($CD = 2.3296$ at 0.05 significance level). Algorithms not connected with WPML in the CD diagram are considered to have a significantly different performance from the control approach.

5.3. Performance Analysis

Tables 3–5 provide a detailed comparison of the evaluated algorithms across five widely used metrics in multi-label learning: *AUC*, *Ranking Loss*, *Coverage* and *Average Precision*. The best-performing algorithm for each dataset is highlighted in bold. From these results, several notable insights can be drawn:

(1) All algorithms demonstrate competitive performance on at least a subset of datasets, highlighting their effectiveness in addressing multi-label learning challenges. Algorithms such as PML-NI, PML-VLS, and MSWL exhibit strong performance on datasets with moderate label sparsity, while ESMC and Glocal achieve improvements on datasets with long-tailed label distributions. However, the performance varies considerably across datasets, reflecting the diversity in label cardinality, feature dimensionality, and noise levels.

(2) The proposed WPML consistently achieves near-optimal performance across the majority of datasets. In terms of *AUC*, WPML obtains the highest score on 75% of datasets, demonstrating its superior discriminative capability. For *Ranking Loss*, WPML ranks best on 56.25% of datasets, indicating that it more accurately preserves label orderings. Similarly, WPML excels in *Average Precision* (62.5%), suggesting that the model maintains strong predictive accuracy even under label sparsity and noise. While WPML shows a slightly smaller advantage in *Coverage* (50%), its performance remains highly competitive and often close to the best-performing algorithms.

(3) On text datasets such as Bibtex and Business, WPML demonstrates substantial improvements in *AUC* and *Ranking Loss* compared to baseline methods, likely due to its robust handling of noisy label assignments common in user-annotated tags. On music and audio datasets such as CAL500 and Emotions, WPML's *Average Precision* gains suggest that the model effectively captures correlations among semantically related labels.

These results provide strong evidence that WPML exhibits exceptional robustness in handling incomplete and noisy label assignments. This superior performance is largely due to WPML's ability to model intricate interactions between the feature and label spaces. By jointly analyzing both spaces from multiple perspectives, WPML captures higher-order dependencies that are often overlooked by other methods, yielding more compact and informative representations that enhance generalization to unseen data. Furthermore, WPML's design leverages a synergistic combination of feature space

regularization, label space exploration, and sparse label correction. This integrated approach not only suppresses noise and recovers missing information but also reinforces meaningful label patterns, enabling the model to deliver robust predictions and effective label disambiguation in a unified framework.

5.4. Statistical Analysis of Algorithm Performance

To quantitatively assess the effectiveness of the proposed WPML algorithm, we conducted rigorous non-parametric tests, namely the Friedman test [48] and the Bonferroni-Dunn test [49]. These tests allow us to determine whether the performance differences between WPML and existing multi-label learning methods are statistically significant.

Let r_i denote the average rank of the i -th algorithm across all datasets, N the total number of datasets, and K the number of algorithms compared. Under the null hypothesis that assuming all algorithms perform equivalently, the Friedman statistic F_F follows an F -distribution with $(K - 1)(N - 1)$ degrees of freedom. The Friedman statistic is formally defined as:

$$F_F = \frac{(N - 1)\chi_F^2}{N(K - 1) - \chi_F^2}, \quad \text{where} \quad (37)$$

$$\chi_F^2 = \frac{12N}{K(K + 1)} \left(\sum_{i=1}^K r_i^2 - \frac{K(K + 1)^2}{4} \right).$$

Table 6. Friedman test results and critical value ($\alpha = 0.05$, $K = 8$, $N = 16$) for various performance metrics.

Metric	F_F	Critical Value
AUC	49.3722	3.031
Ranking Loss	9.5255	3.031
Coverage	9.7286	3.031
Average Precision	11.2720	3.031
Hamming Loss	8.1667	3.031

Table 6 presents the Friedman statistics for each evaluation metric alongside the corresponding critical value at a significance level of $\alpha = 0.05$ ($K = 8$, $N = 16$). All computed Friedman statistics substantially exceed the critical value, indicating that the null hypothesis of equal performance can be confidently rejected. This confirms that statistically significant differences exist among the algorithms across all evaluation metrics. To further examine pairwise differences between algorithms, we employed the Bonferroni-Dunn test. The critical difference (CD) for distinguishing performance is computed as:

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

where $q_\alpha = 2.098$ for $\alpha = 0.05$, yielding $CD_\alpha = 2.3296$ with $K = 8$ and $N = 16$. Figure 1 illustrates the CD diagrams for all metrics, where algorithms not linked by a line are considered statistically different.

From the CD diagrams, it is evident that WPML consistently ranks higher than competing methods across all metrics. Specifically, WPML demonstrates statistically significant superiority in *Ranking Loss* and *Average Precision*, highlighting its robustness and effectiveness in handling noisy or incomplete label scenarios. These results provide strong quantitative evidence that the proposed WPML algorithm not only achieves high predictive performance but also maintains stability and reliability across diverse multi-label datasets.

5.5. Sensitivity Analysis

To evaluate the robustness of WPML with respect to its hyperparameters, we conducted a systematic sensitivity analysis. Each parameter's effect was examined by varying one or two parameters at a time while keeping the others fixed, thereby isolating their individual contributions to model

performance. Figures 2 and 3 illustrate the impact of the hyperparameters λ , α , β , and γ on the Emotion dataset.

The results indicate that WPML achieves peak performance within intermediate ranges of the parameter space, while extreme values lead to performance deterioration. Specifically, a very small λ fails to adequately preserve local instance structures, diminishing the model's ability to capture critical relationships in the data. Conversely, an excessively large λ disproportionately emphasizes local structure, suppressing the effect of other regularization terms and thereby degrading overall performance. Similar patterns are observed for α , β , and γ , where both overly small and large values result in suboptimal outcomes. These observations underscore the model's sensitivity to hyperparameter selection and highlight the necessity of choosing appropriate ranges to ensure stable and balanced performance. While WPML exhibits robustness within well-chosen parameter intervals, improper tuning can significantly impair its effectiveness. Consequently, efficient hyperparameter optimization strategies are recommended for practical deployment to maintain optimal predictive performance.

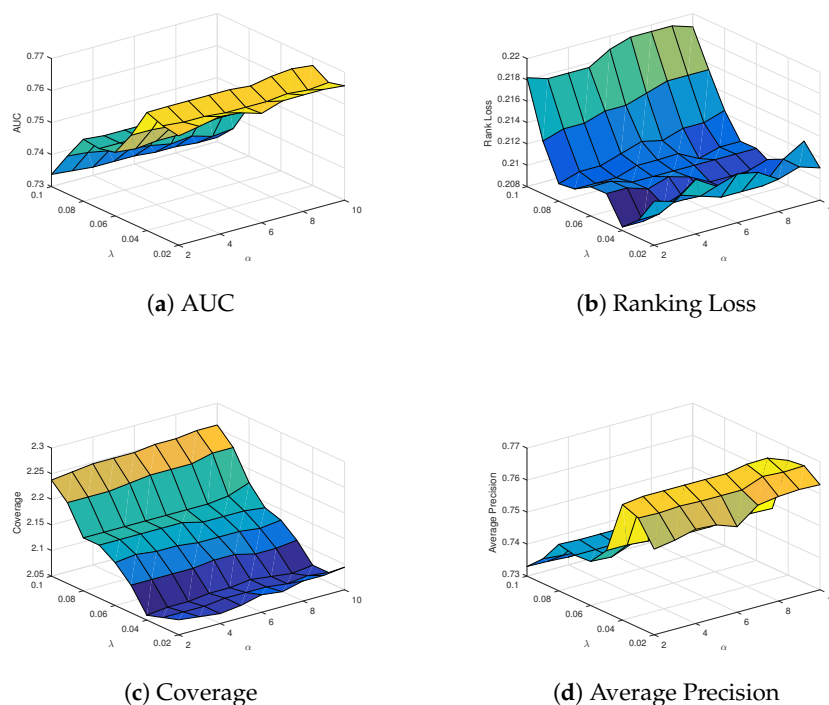


Figure 2. Sensitivity analysis of WPML on Emotion dataset by considering the variations of parameters λ and α .

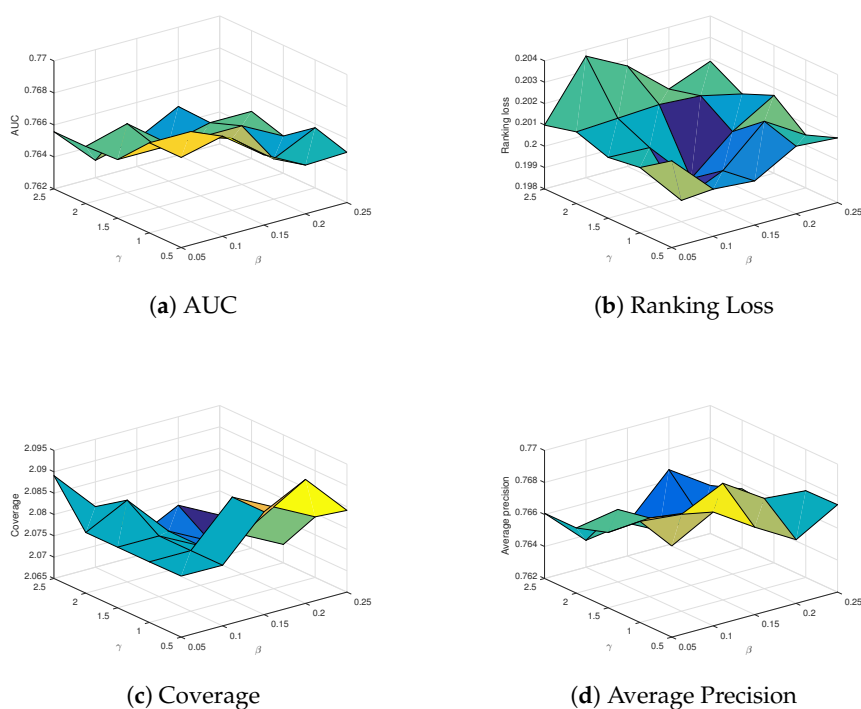


Figure 3. Sensitivity analysis of WPML on Emotion dataset by considering the variations of parameters β and γ .

5.6. Ablation Study

To evaluate the contribution of local instance structure within the feature space, we introduce a simplified variant of WPML, denoted as WPML-d, which omits local structure information. Figure 4 compares the performance of WPML and WPML-d across multiple evaluation metrics. The results demonstrate that WPML generally outperforms WPML-d, confirming that integrating local instance relationships enhances the model's ability to preserve critical patterns and improves overall predictive accuracy. Interestingly, WPML-d occasionally surpasses the full WPML model in specific cases, particularly when the local feature structure conflicts with the label space. This suggests that mismatches between feature and label structures can introduce noise, and in such scenarios, ignoring local structure enables the model to rely on more informative instance-level cues. Consequently, while incorporating local structure typically strengthens generalization and robustness, careful consideration of feature-label alignment is necessary, and adaptive strategies may further mitigate inconsistencies in heterogeneous datasets.

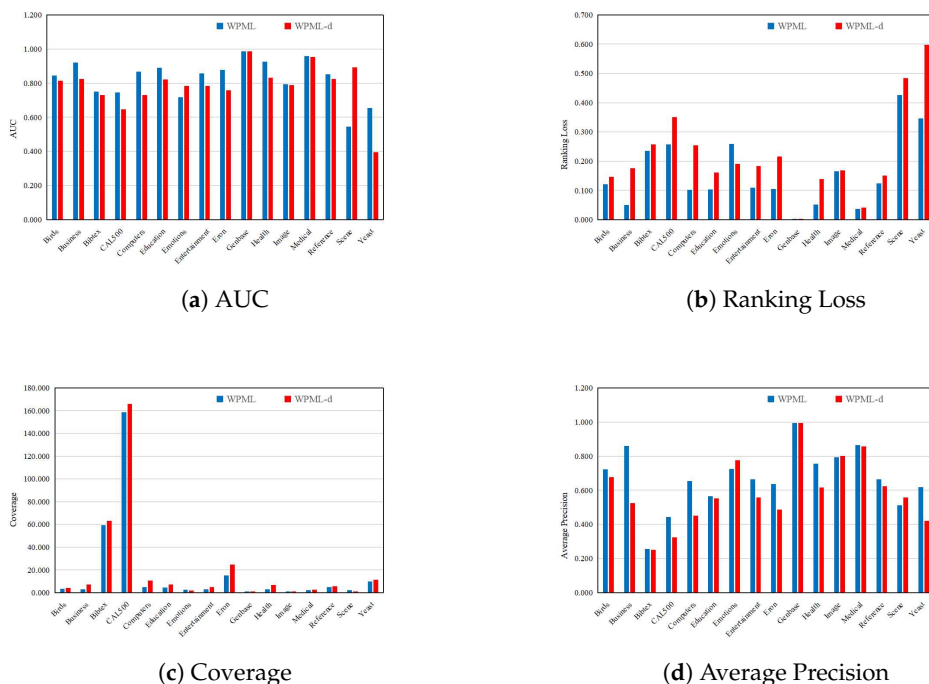


Figure 4. Comparisons of WPML and the degenerated version WPML-d in terms of each evaluation metric.

5.7. Convergence Analysis

Figure 5 presents the convergence trends of the objective function on the Image and Yeast datasets. The algorithm exhibits a rapid decrease in the objective value during the initial iterations, indicating its ability to quickly capture the dominant structures in the data. As training proceeds, the rate of decrease slows, and the objective function gradually stabilizes, suggesting that the algorithm is approaching a near-optimal solution.

This convergence pattern reflects a two-stage optimization process: an initial phase of fast descent, followed by a refinement phase with slower, more precise adjustments. Such behavior demonstrates WPML's capacity to balance convergence speed with solution accuracy, efficiently navigating the trade-off between rapid optimization and fine-grained precision. The empirical evidence indicates that WPML reliably attains stable and optimal solutions across diverse datasets, highlighting its robustness and practical applicability. The combination of swift early adaptation and gradual fine-tuning ensures both efficiency and reliability, making WPML particularly suitable for complex multi-label learning tasks where both speed and stability are critical.

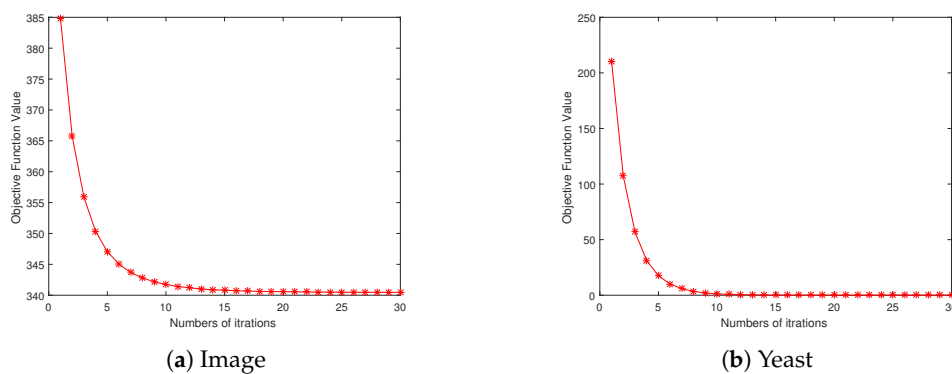


Figure 5. Convergence analysis of WPML on Image and Yeast datasets.

6. Conclusions

In this work, we introduced a novel framework for partial multi-label learning that simultaneously addresses missing and noisy labels. By combining graph-based similarity propagation with collaborative compact feature learning, our approach improves label prediction accuracy while mitigating the adverse effects of weak annotations. Specifically, a graph similarity matrix constructed via fuzzy C-means clustering enables effective label consistency propagation, facilitating the detection of noisy labels and recovery of missing ground-truth labels. Additionally, the model separates refined, label-specific features from ambiguous, noise-related ones, ensuring that only informative signals guide classifier training. Through joint optimization of feature representations and label predictions, the proposed method exhibits enhanced robustness and predictive performance. Both theoretical insights and extensive empirical evaluations demonstrate its clear advantage over existing approaches in weakly annotated multi-label learning scenarios.

Future work may focus on several directions: (1) exploring alternative structural clustering techniques to potentially improve similarity-based label propagation; (2) enhancing the scalability of the framework for large-scale, high-dimensional datasets through computational optimizations; and (3) extending the approach to real-world applications such as healthcare, where data is often noisy and incomplete, for tasks including medical diagnosis, patient monitoring, and predictive modeling of clinical outcomes.

Author Contributions: Conceptualization, Y.T. and A.T.; methodology, Y.T. and A.T.; software, Y.T. and A.T.; validation, Y.T. and A.T.; formal analysis, Y.T. and A.T.; investigation, Y.T. and A.T.; resources, Y.T. and A.T.; data curation, Y.T. and A.T.; writing—original draft preparation, Y.T. and A.T.; writing—review and editing, Y.T. and A.T.; visualization, Y.T. and A.T.; supervision, Y.T. and A.T.; project administration, Y.T. and A.T.; funding acquisition, A.T. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Research initiation project of Huaqiao University (No. 24BS114), and Open Project Foundation of Key Laboratory of Computation Intelligence and Chinese Information Processing of Ministry of Education and Key Laboratory of Data Intelligence and Cognitive Computing of Shanxi Province (No. CICIP2024006).

Institutional Review Board Statement: Not applicable.

Data Availability Statement: All multi-label datasets used in the experiments are from a multi-label classification dataset repository that is available at <http://mulan.sourceforge.net/datasets.html> and <http://www.uco.es/kdis/mlresources/>.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Liu, W.; Wang, H.; Shen, X.; Tsang, I.W. The emerging trends of multi-label learning. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *44*(11), 7955–7974.
2. Xie, M.; Huang, S. Partial multi-label learning. In Proc. AAAI Conf. Artif. Intell., 2018; pp. 4302–4309.
3. Zhong, J.; Shang, R.; Zhao, F.; Zhang, W.; Xu, S. Negative label and noise information guided disambiguation for partial multi-label learning. *IEEE Trans. Multim.* **2024**, *26*, 9920–9935.
4. Chen, J.-Y.; Li, S.-Y.; Huang, S.-J.; Chen, S.; Wang, L.; Xie, M.-K. UNM: A universal approach for noisy multi-label learning. *IEEE Trans. Knowl. Data Eng.* **2024**, *36*(9), 4968–4980.
5. Zhang, M.; Yu, F.; Tang, C. Disambiguation-free partial label learning. *IEEE Trans. Knowl. Data Eng.* **2017**, *29*(10), 2155–2167.
6. Gibaja, E.; Ventura, S. A tutorial on multilabel learning. *ACM Comput. Surv.* **2015**, *47*(3), 1–38.
7. Zhang, M.; Fang, J. Partial multi-label learning via credible label elicitation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *43*, 3587–3599.
8. Sun, L.; Du, W.; Ding, W.; Long, Q.; Xu, J. Granular ball-based fuzzy multineighborhood rough set for feature selection via label enhancement. *Eng. Appl. Artif. Intell.* **2025**, *145*, 110191.
9. Yan, Y.; Li, S.; Feng, L. Partial multi-label learning with mutual teaching. *Knowl.-Based Syst.* **2021**, *212*, 106624.

10. Xie, M.; Huang, S. Semi-supervised partial multi-label learning. In IEEE Int. Conf. Data Mining (ICDM), 2020; pp. 691–700.
11. Li, Z.; Jia, Y.; Yu, M.; Miao, Z. Calibrated disambiguation for partial multi-label learning. In Proc. AAAI Conf. Artif. Intell., vol. 39, 2025; pp. 18620–18628.
12. Wang, C.; Zhang, Y.; An, S.; Deng, T. Adaptive feature selection based on fuzzy rough set fusion model with class variance. *Pattern Recognit.* 2026, 170, 112014.
13. Wang, C.; Wang, Y.; Deng, T.; Huang, Y. A nonlinear multi-label learning model based on tanh mapping. *Eng. Appl. Artif. Intell.* 2023, 126, 106837.
14. Han, Q.; Hu, L.; Gao, W. Integrating label confidence-based feature selection for partial multi-label learning. *Pattern Recognit.* 2025, 161, 111281.
15. Hang, J.-Y.; Zhang, M.-L. Partial multi-label learning via label-specific feature corrections. *Sci. China Inf. Sci.* 2025, 68(3), 132104.
16. Sun, Z.; Chen, Z.; Liu, J.; Chen, Y.; Yu, Y. Partial multi-label feature selection via low-rank and sparse factorization with manifold learning. *Knowl.-Based Syst.* 2024, 296, 111899.
17. Jalali, H.; Kasneci, G. Multilabel Classification for Entry-Dependent Expert Selection in Distributed Gaussian Processes. *Entropy* 2025, 27, 307.
18. Yang, F.; Jia, Y.; Liu, H.; Dong, Y.; Hou, J. Noisy label removal for partial multi-label learning. In Proc. 30th ACM SIGKDD Conf. Knowl. Discov. Data Min., 2024; pp. 3724–3735.
19. Qian, W.; Tu, Y.; Huang, J.; Shu, W.; Cheung, Y.-M. Partial multilabel learning using noise-tolerant broad learning system with label enhancement and dimensionality reduction. *IEEE Trans. Neural Netw. Learn. Syst.* 2025, 36(2), 3758–3772.
20. Li, R. J.; Ma, Y. C.; Chen, H.; Yang, X. F.; Xing, Z. W. Coordinate descent for top-k multi-label feature selection with pseudo-label learning and manifold learning. *Neurocomput.* 2025, 658, 131640.
21. Bucak, S.S.; Jin, R.; Jain, A.K. Multi-label learning with incomplete class assignments. In Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2011; pp. 2801–2808.
22. Yu, H.-F.; Jain, P.; Kar, P.; Dhillon, I. Large-scale multi-label learning with missing labels. In Int. Conf. Mach. Learn., PMLR, 2014; pp. 593–601.
23. Chen, G.; Song, Y.; Wang, F.; Zhang, C. Semi-supervised multi-label learning by solving a sylvester equation. In Proc. 2008 SIAM Int. Conf. Data Mining, 2008; pp. 410–419.
24. Jain, V.; Modhe, N.; Rai, P. Scalable generative models for multi-label learning with missing labels. In Int. Conf. Mach. Learn., PMLR, 2017; pp. 1636–1644.
25. Kong, X.; Wu, Z.; Li, L.-J.; Zhang, R.; Yu, P.S.; Wu, H.; Fan, W. Large-scale multi-label learning with incomplete label assignments. In Proc. 2014 SIAM Int. Conf. Data Mining, 2014; pp. 920–928.
26. Wang, C.; Wang, Y.; Deng, T.; Ding, W. Missing multi-label learning based on the fusion of two-level nonlinear mappings. *Inf. Fusion* 2024, 103, 102105.
27. Jiang, L.; Yu, G. X.; Guo, M. Z.; Wang, J. Feature selection with missing labels based on label compression and local feature correlation. *Neurocomput.* 2020, 395, 95–106.
28. Yang, H.; Zhou, J.T.; Cai, J. Improving multi-label learning with missing labels by structured semantic correlations. In *Comput. Vis.–ECCV 2016: 14th Eur. Conf.*, Springer, 2016; pp. 835–851.
29. Wu, B.; Jia, F.; Liu, W.; Ghanem, B.; Lyu, S. Multi-label learning with missing labels using mixed dependency graphs. *Int. J. Comput. Vis.* 2018, 126(8), 875–896.
30. Braytee, A.; Liu, W.; Anaissi, A.; Kennedy, P.J. Correlated multi-label classification with incomplete label space and class imbalance. *ACM Trans. Intell. Syst. Technol.* 2019, 10(5), 1–26.
31. Ma, J.; Chow, T.W. Label-specific feature selection and two-level label recovery for multi-label classification with missing labels. *Neural Netw.* 2019, 118, 110–126.
32. Yin, T.; Chen, H.; Wang, Z.; Liu, K.; Yuan, Z.; Horng, S.-J.; Li, T. Feature selection for multilabel classification with missing labels via multi-scale fusion fuzzy uncertainty measures. *Pattern Recognit.* 2024, 110580.
33. Dai, J.; Li, M.; Zhang, C. Multi-label feature selection with missing labels by weak-label fusion fuzzy discernibility pair. *Inf. Fusion* 2025, 117, 102921.
34. Sun, L.; Yin, T.; Ding, W.; Qian, Y.; Xu, J. Feature selection with missing labels using multilabel fuzzy neighborhood rough sets and maximum relevance minimum redundancy. *IEEE Trans. Fuzzy Syst.* 2021, 30(5), 1197–1211.
35. Sun, L.; Lyu, G.; Feng, S.; Huang, X. Beyond missing: Weakly-supervised multi-label learning with incomplete and noisy labels. *Appl. Intell.* 2021, 51(3), 1552–1564.

36. Ding, J.; Zhang, Y.; Jia, L.; Fu, X.; Jiang, Y. Noisy feature decomposition-based multi-label learning with missing labels. *Inf. Sci.* **2024**, *662*, 120228.
37. Wei, T.; Guo, L.; Li, Y.; Gao, W. Learning safe multi-label prediction for weakly labeled data. *Mach. Learn.* **2018**, *107*, 703–725.
38. Fang, Q.; Xiang, C.; Duan, J.; Soufiyan, B.; Shao, C.; Yang, X.; Xu, S.; Yu, H. OMAL: A Multi-Label Active Learning Approach from Data Streams. *Entropy* **2025**, *27*, 363–363.
39. Bezdek, J. C. Pattern recognition with fuzzy objective function algorithms; Springer Sci. & Bus. Media, 2013.
40. Nie, F.; Xu, D.; Tsang, I.; Zhang, C. Flexible manifold embedding: A framework for semi-supervised and unsupervised dimension reduction. *IEEE Trans. Image Process.* **2010**, *19*(7), 1921–1932.
41. Liu, G.; Li, Q.; Yang, X.; Xing, Z.; Ma, Y. Partial multi-label feature selection based on label matrix decomposition. *Neural Comput. Appl.* **2025**, *37*(6), 4207–4227.
42. Xie, M.; Huang, S. Partial multi-label learning with noisy label identification. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *44*, 3676–3687.
43. Yu, Z.; Zhang, M. Multi-label classification with label-specific feature generation: A wrapped approach. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *44*, 5199–5210.
44. Ma, Z.; Chen, C. Expand globally, shrink locally: Discriminant multilabel learning with missing labels. *Pattern Recognit.* **2021**, *111*, 107675.
45. Zhu, Y.; Kwok, J.; Zhou, Z. Multilabel learning with global and local label correlation. *IEEE Trans. Knowl. Data Eng.* **2018**, *30*(6), 1081–1094.
46. Akbarnejad, A.; Baghshah, M. An efficient semi-supervised multilabel classifier capable of handling missing labels. *IEEE Trans. Knowl. Data Eng.* **2019**, *31*(2), 229–242.
47. Zhang, J.; Li, S.; Jiang, M.; Tan, K. Learning from weakly labeled data based on manifold regularized sparse model. *IEEE Trans. Cybern.* **2022**, *52*, 3841–3854.
48. Friedman, M. A comparison of alternative tests of significance for the problem of m rankings. *Ann. Math. Stat.* **1940**, *11*(1), 86–92.
49. Demsar, J. Statistical comparisons of classifiers over multiple data sets. *J. Mach. Learn. Res.* **2006**, *7*, 1–30.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.