IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 14 March 2023, accepted 23 April 2023, date of publication 1 May 2023, date of current version 24 May 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3271730

RESEARCH ARTICLE

A Unified Framework for Graph-Based Multi-View Partial Multi-Label Learning

JIAZHENG YUAN[1], WEI LIU[2], ZHIBIN GU[2], AND SONGHE FENG[2]
[1]College of Science and Technology, Beijing Open University, Beijing 100081, China
[2]School of Computer and Information Technology, Beijing Jiaotong University, Beijing 100044, China

Corresponding author: Jiazheng Yuan (jzyuan@139.com)

This work was supported in part by the Fundamental Research Funds for the Central Universities under Grant 2022YJS024, in part by the National Natural Science Foundation of China under Grant 61871028, in part by the Joint Key of Beijing Natural Science Foundation and Municipal Education Commission under Grant KZ201951160050, and in part by the Beijing Advanced Talents Great Wall Scholar Training Program under Grant CITTCD20190313.

ABSTRACT Multi-view partial multi-label learning (MVPML) is a fundenmental problem where each sample is linked to multiple kinds of features and candidate labels, including ground-truth and noise labels. The key problem of MVPML is how to manipulate the multiple features and recover the ground-truth labels from candidate label set. To this end, this study designs a novel **G**raph-based **M**ulti-view **P**artial **M**ulti-label model named as GMPM, which combines the multi-view information detection, valuable label selection and multi-label predictor model learning into a unified optimization model. To be specific, GMPM first exploits the consensus information across multiple views by learning the view-specific similarity graph and fuses multiple graphs into a target one. Then, we divide the observed label set into two parts: the ground-truth part and the noise part, where the latter is associated with a sparse constraint to make sure the former is clean. Furthermore, we embed the learned unified similarity graph into the process of label disambiguation to restore a more reliable ground-truth label matrix. Finally, the resulting multi-label predictive model is learned with the help of ground-truth label matrix. Extensive experiments on six common used datasets demonstrate that the proposed GMPM achieves comparable performance over the state-of-the-arts.

INDEX TERMS Multi-view learning, partial multi-label learning, graph learning, low-rank and sparse decomposition.

I. INTRODUCTION

Multi-label classification is designed to assign multiple labels to an instance, has emerged as a hot topic due to the the ubiquity of multi-view data. For instance, in image categorization, an image could contain multiple semantics objects; in movie categorization, a work can be located in both love type and funny type simultaneously. However, it is typically difficult and expensive to obtain accurate annotation in real scenes, and the label information of the given data usually contains various noises. If such ambiguous data are directly employed for model training, the learned model tends to be bias and its robustness can also not be guaranteed. To this end, *Partial Multi-label Learning* [1] (PML) attempts to learn an accurate multi-label classifier from the multi-label data with redundant labeling information, the ambiguous candidate labels are processed by assigning a confidence value to each candidate label, and then their optimization and model induction are integrated into a unified framework. Sun et al. [2] decompose the obtained labels into a ground-truth part and a noise part by utilizing the low-rank and sparse decomposition scheme-where a confidence value is utilized for each candidate label to measure how likely it is a ground-truth label of the instance.

Although the mentioned above PML models have feasible solutions, there is a common limitation that most are designed for single-view data, but are difficult to scale to multi-view scenarios. To be specific, as illustrated in Figure 1, on the one hand, an image can be characterized by diverse features, such as color, texture and shape. On the other hand, each sample is associated with an overfitting candidate labels,

The associate editor coordinating the review of this manuscript and approving it for publication was Wei Liu.

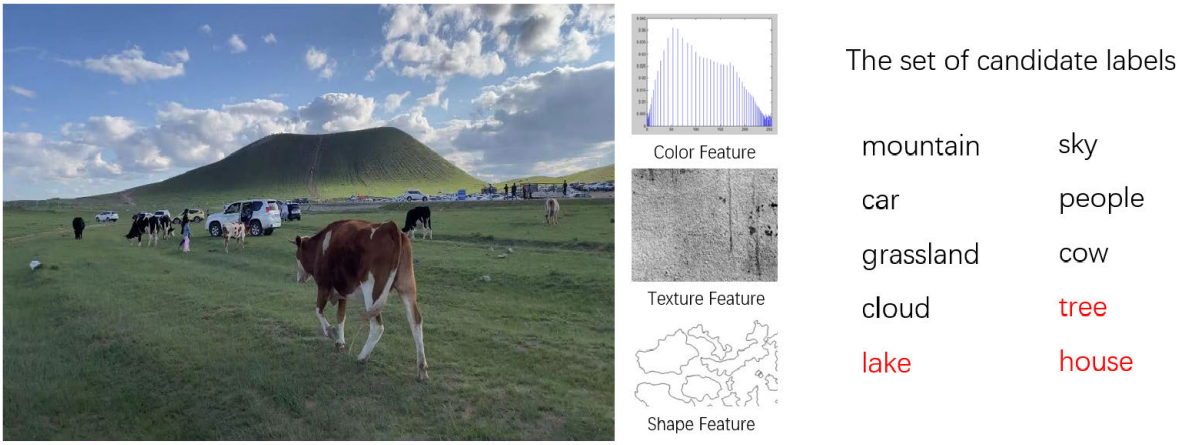


FIGURE 1. Illustration of MVPML. On the one hand, an image can be characterized by diverse features, such as color, texture and shape. On the other hand, each sample is associated with an overfitting candidate labels, including ground-truth and noise labels.

including ground-truth and noise labels. Data with the above two attributes can be called multi-view partial multi-label data, and their classification problem constitutes the problem of multi-view partial multi-label learning (MVPML). The key issues of multi-view and multi-label learning lies in the following two points: (a) How to make full use of the rich information provided by multi-view data; (b) How to effectively identify clean labels iredundant labels. For example, Chen et al. designed a novel MVPML model termed GRADIS [3], which employs the multiple graphs to filter out the ground-truth labels from the candidate label set, then adopts the obtained embedded features to learn the classification model. In additionm, FIMAN [4] fusies multiple views affinity information to arrive an aggregate structure, which is then employed to disambiguate the label space. The disambiguated labels are adopted to induce a multi-label classification model by fitting its modeling outputs. However, since both models are two-stage strategies, there is a risk of encountering local optimal solutions.

To meet the aforementioned problems, this work proposes a one-stage **G**raph-based **M**ulti-view **P**artial **M**ulti-label **L**abel (GMPM) model, which combines the abundant multiple features leveraging, noise label disambiguation and predictor model training into a unified framework, making the multiple similarity graphs learning, label disambiguation, and multi-label predictor model in a mutually reinforcing manner. To be specific, as illustrated in Figure 2, on the one hand, the proposed method first learns multiple view-specific similarity graphs to explore the similarity relationship of paired data points under different views, which can be regarded as *node similarity graph* (NSG). Then, the multiple similarity graphs are fused in to a consensus one $\mathbf{U}$ in a self-weighted way, making the consensus information of multiple features to be fully expoited. On the other hand, the given redundant label infrmation $\mathbf{Y}$ is broken down into two parts (i.e., ground-truth part $\overline{\mathbf{Y}}$ and noise part $\mathbf{E}$), where the latter is asscіated with the l_1-norm penalty term to meet the assumption that the noise information is sparse. Furthermore, the acquired global similarity graph $\mathbf{U}$ is embedded in the process of label disambiguation to ensure that the finally obtained ground-truth label matrix is reliable. Finally, we employ the mappings of heterogeneous data points and disambiguated ground-truth label to learn the multi-label prediction model. The work of this paper is an extension of the conference paper [5] that we have published before. Compared with work [5], the introduction of related work in this paper is more comprehensive, the theoretical derivation is more detailed, and the experimental results are more abundant. In summary, the main contributions of this paper are summarized as follows:

- We propose a graph-based multi-view partial multi-label learning method named GMPM, which learns multiple view-specific graphs and the fused graph jointly to exploit the consensus information of multiple features, then we decompose the candidate label set into a ground-truth part and noise part, and embed the learned unified graph into the label disambiguation to obtain a more reliable ground-truth label matrix.
- To the best of our knowledge, GMPM is the first attempt that integrate the multi-view information detection, noise label disambiguation and predictor model training as a unified optimization model, which enables multiple components to learn in a mutually reinforcing way.
- Experimental results on several data sets demonstrate that the GMPM model obtains competitive classification performance over state-of-the-arts.

The rest of the paper is structured as follows. In Section II, we briefly introduce the related work. Then we give the important notations and the proposed model GMPM in Section III, followed by the optimization method and whole process of solving GMPM in Section IV. Experimental results are reported in Section V, and Section VI concludes the paper.

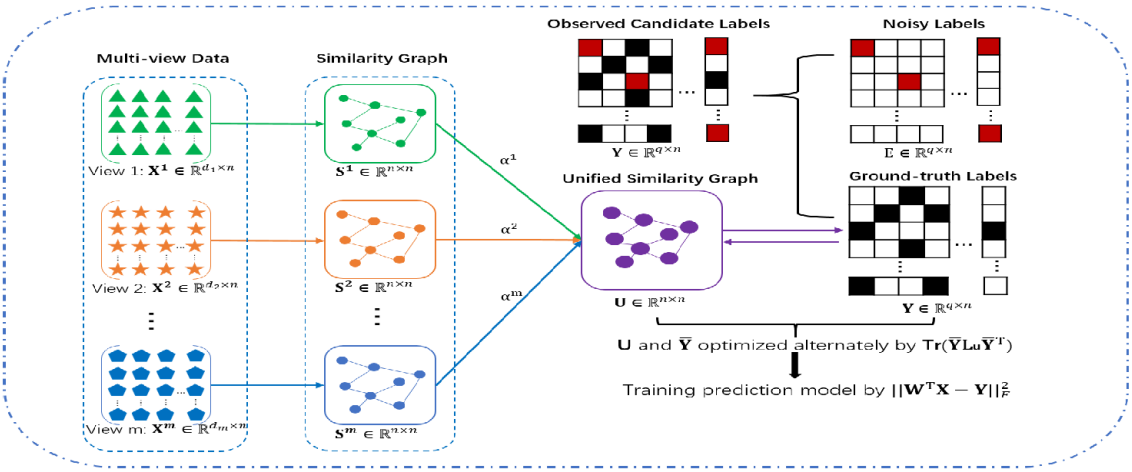


FIGURE 2. The structure diagram of the proposed GMPM. Specifically, GMPM first learns multiple view-specific similarity graphs, and fusies the multiple graphs into a consensus one **U**. On the other hand, the given label set **Y** is broken down into two parts (i.e., ground-truth part $\overline{\mathbf{Y}}$ and noise part **E**), where the latter is assciated with the l_1-norm penalty term to meet the assumption that the noise information is sparse. Finally, we employ the mappings of heterogeneous data points and disambiguated ground-truth label to learn the multi-label prediction model.

II. RELATED WORK

In this part, we briefly introduce some work related to our model, including multi-view learning and partial multi-label learning.

A. MULTI-VIEW LEARNING

Due to the fact that data can be characterized by heterogeneous features in real applications, multi-view learning has shown to be a hot topic [6]. And more recently, a lot of multi-view learning algorithms have been proposed [7], [8], [9]. From the perspective of multi-view data fusion mechanism, the existing multi-view learning methods can be divided into the following three categories: 1) co-training; 2) multiple kernel learning; and 3) graph-based multi-view learning. The first kind of multi-view learning method is the co-training based multi-view learning [10], which attempts to train the model by using alternate iteration to maximize the consistency between different views. The multiple kernel learning based methods [11] attempts to learn different kernels for different views, and then combines them with different strategies to improve learning performance. The third type of multi-view learning method is graph-based method [12], which can be further divided into two subclasses according to the construction method of the graph (i.e., subspace-based approach [13], [14] and graph-based approach [15]). Specifically, the former uses the self-representation strategy to construct the coefficient matrix to represent the similarity between data points, and the latter uses the Euclidean distance between data pairs to calculate the similarity. For example, MVSC [16] first learns the graph matrix for each view, and these graph matrices are then fused automatically to obtain the final clustering results. In order to fully discover the complementary information among different views in the fused graph, [17] introduces a diversity regularization item to explore the complementarity among different views. Liang et al. [18] model the multi-view consistency and inconsistency into the unified multi-view clustering framework to fully mine the rich information of multi-view data. Furthermore, in order to accelerate the graph constructing process, Bipartite Graph methods [16], [19], [20], [21], [22] have been proposed more recently, which can learn sparse graph by establishing the correlation between samples and the selected anchor samples. For example, Huang et al. [22] design a fast multi-view clustering algorithm, which is the first attempt to use the concept of random view groups to serves as the basic form of the flexible view-organizations.

B. PARTIAL MULTI-LABEL LEARNING

Partial multi-label learning refers to that each sample is represented by a single type of feature and associated with a redundant label set that containing ground-truth label and noise label [23], [24], [25]. According to the model construction strategy, we can divide the existing PML methods into two classes: 1) One-stage Strategy and 2) Two-Stage Strategy. One-stage approach typically learns a confidence score for each candidate label, and then distinguishes whether the corresponding label belongs to a ground-truth label or a noise label according to the confidence score. For example, Yu et al. propose fPML in [26] simultaneously factorizes the observed candidate label matrix and the feature matrix into low-rank matrices to achieve a coherent low-rank matrices and a low-rank label correlation matrix. And then the low-rank coherent matrix is utilized to estimate the label confidence. Sun et al. [2] adopt the low-rank and sparse decomposition scheme and divide the observed label set into a ground-truth label matrix and an irrelevant label matrix. The difference between two-stage strategy and one-stage strategy is that the ground-truth label screening and model training are divided into two steps. The first stage is used to filter noise and select candidate labels with high level of confidence as valuable information. The second stage is to train the model

using the valuable tags selected from the first step. The work of [27] in the first stage learn a confidence value for each candidate label by utilizing the features manifold, and then a gradient boosting model is introduced to complete the classification. Zhang et al. propose the method PARTICLE [28] which evaluate the labeling confidence of each candidate label by iterative label propagation in the first stage. And in the second stage, multi-label predictor is induced via pairwise label ranking.

III. PROPOSED METHOD

In this section, we introduce our proposed method in detail. Assume a multi-view multi-label dataset with m views and n samples $\mathbf{X} = [\mathbf{X}^{1^T}, \mathbf{X}^{2^T}, \ldots, \mathbf{X}^{m^T}]^T \in \mathbb{R}^{D\times n}$, where $\mathbf{X}^v = \left[\boldsymbol{x}_1^v, \boldsymbol{x}_2^v, \ldots, \boldsymbol{x}_n^v\right] \in \mathbb{R}^{d_v\times n}$ is v^{th} view matrix with d_v dimensions as well as n samples and $D = \sum_{v=0}^{m} d_v$. Furthermore, Denote $\mathbf{Y} \in \{0, 1\}^{q\times n}$ as the partial label matrix for all samples. Here, $\mathbf{Y}_{ji} = 1$ indicates i-th sample is annotated with j-th class, otherwise $\mathbf{Y}_{ji} = 0$.

A. NSG (NODE SIMILARITY GRAPH) MATRIX CONSTRUCTION

In our work, we first transform the data matrix of each view into a graph matrix generated from similar graph matrices. We assume that the more similar the two nodes in the view, the greater their similarity values in the NSG, otherwise the smaller they are, so we have:

$$\begin{aligned}&\min_{\{\mathbf{S}^v\}} \sum_{v=1}^{m}\sum_{i,j=1}^{n} \|\boldsymbol{x}_i^v - \boldsymbol{x}_j^v\|_2^2 s_{ij}^v + \beta \sum_{v=1}^{m}\sum_{i}^{n} \|\mathbf{s}_i^v\|_2^2 \\ &s.t.\ \forall v, s_{ii}^v = 0, s_{ij}^v \geq 0, \mathbf{1}^T\mathbf{s}_i^v = 1. \end{aligned} \tag{1}$$

where $\mathbf{S}^v$ denotes $\mathbf{S}^1, \mathbf{S}^2, \ldots, \mathbf{S}^m$ is the NSG of v^{th} view and $s_{ij} \in \mathbf{S}^v$. The first term constructs the NSG of each view. Specifically, s_{ij} represents the similarity value between each node and $\boldsymbol{x}_i$. The second term adopts ℓ_2 norm as the constraints term to avoid obtaining trivial solutions and limiting the sparsity of NSG. β is a is an implicit hyperparameter whose value depends on the number of nearest neighbors. Here, we construct each NSG matrix independently for each view, because each NSG has no relationship to the other views.

B. NSG MATRIX FUSION

Due to the multiple view-specific similarity graphs of multiple views obtained by Eq. (1) is specific to a single view, it cannot ensure that the rich information of multiple view data is fully mined. To this end, we attempt fuse the multiple graphs in to a consensus one to acquire a more comprehensive target one $\mathbf{U}$, so as to leverange the abundant information of multiple views, and its mathematical form is:

$$\begin{aligned}&\min_{\mathbf{U}} \sum_{v=1}^{m} \alpha^v \|\mathbf{U} - \mathbf{S}^v\|_F^2 \\ &s.t.\ \forall i, u_{ij} \geq 0, \mathbf{1}^T\boldsymbol{u}_i = 1. \end{aligned} \tag{2}$$

where $\boldsymbol{u}_i \in \mathbb{R}^{n\times 1}$ denotes a column vector. Considering the fact that different views may exhibit different discriminative information, α^v is introduced to automatically weight different views, where the larger α^v denotes that the v^{th} view contributes more information to the final unified graph $\mathbf{U}$.

Then, combine Eq. 1 and Eq. 2 to jointly train $\{\mathbf{S}^v\}$ and $\mathbf{U}$ so that they can help each other in a mutual reinforcement manner by the following formula:

$$\begin{aligned}&\min_{\{\mathbf{S}^v\},\mathbf{U}} \sum_{v=1}^{m}\sum_{i,j=1}^{n} \|\boldsymbol{x}_i^v - \boldsymbol{x}_j^v\|_2^2 s_{ij}^v + \sum_{v=1}^{m} \alpha^v \|\mathbf{U} - \mathbf{S}^v\|_F^2 \\ &\quad +\beta \sum_{v=1}^{m}\sum_{i}^{n} \|\mathbf{s}_i^v\|_2^2 \\ &s.t.\ \forall v, s_{ii}^v = 0, s_{ij}^v \geq 0, \mathbf{1}^T\mathbf{s}_i^v = 1, \\ &\quad \forall i, u_{ij} \geq 0, \mathbf{1}^T\boldsymbol{u}_i = 1. \end{aligned} \tag{3}$$

C. NOISE LABEL DISAMBIGUATION

Partial multi-label learning aims to solve such a situation which the observed labels contain label redundancy, i.e., annotators may roughly assign a set of candidate labels to each instance, including related labels and some unrelated labels.

The focus of this subsection is how to effectively filter out the noise label from a given set of redundant labels and restore clean labels information. To be specific, we assume that the observed candidate labels $\mathbf{Y}$ consist of the related labels $\overline{\mathbf{Y}}$ and the unrelated labels $\mathbf{E}$. The core idea of this method is to separate the $\overline{\mathbf{Y}}$ and $\mathbf{E}$, and to ensure the accuracy of separation by applying sparse constraints to the noise part. Specifically, we assume that the redundant noise label is sparse and introduce an l_1-norm regularization to eliminate the redundant label noisy matrix as $\|\mathbf{E}\|_1$.

Besides, to obtain the accurate label matrix more accurately, we introduce the unified graph to align with the accurate label matrix to remove noise labels and exploit accurate labels matrix. For multi-view partial multi-label data, the properties of the accurate label matrix need to be consistent with the properties of the unified graph. Specifically, if two samples are close in label space, they should also share similar characteristics in the unified graph:

$$\begin{aligned}&\min_{\overline{\mathbf{Y}}} \sum_{i}^{n}\sum_{j}^{n} \|\bar{\boldsymbol{y}}_i - \bar{\boldsymbol{y}}_j\|_2^2 u_{ij} \\ &\qquad = \mathbf{Tr}(\overline{\mathbf{Y}}(\mathbf{A} - \mathbf{U})\overline{\mathbf{Y}}^T) = \mathbf{Tr}(\overline{\mathbf{Y}}\mathbf{L}_u\overline{\mathbf{Y}}^T) \\ &s.t.\ \mathbf{Y} = \overline{\mathbf{Y}} + \mathbf{E}. \end{aligned} \tag{4}$$

where $\boldsymbol{y}_i$ is the i-th column in the accurate labels matrix, indicating the accurate labels of the i-th sample in $\overline{\mathbf{Y}}$. u_{ij} measures the similarity of sample i and sample j in the unified graph $\mathbf{U}$ and $\mathbf{A}$ is a diagonal matrix with $a_{ii} = \sum_{j=1}^{n} u_{ij}$ and $\mathbf{L}_u = \mathbf{A} - \mathbf{U}$ is the graph laplacian matrix.

By combining $\|\mathbf{E}\|_1$ and Eq. 4, accurate labels can be stripped from the observed candidate labels:

$$\begin{aligned}&\min_{\overline{\mathbf{Y}},\mathbf{E}} \lambda\mathbf{Tr}(\overline{\mathbf{Y}}\mathbf{L}_u\overline{\mathbf{Y}}^T) + \sigma\|\mathbf{E}\|_1 \\ &s.t.\ \mathbf{Y} = \overline{\mathbf{Y}} + \mathbf{E}. \end{aligned} \tag{5}$$

D. OBJECTIVE FUNCTION

Based on the above analysis, we can obtain an accurate label matrix through multi-view consensus information. In addition, we train the multi-label classification model by exploring multi-view data and valuable tag sets filtered from the original tag set. Therefore, we can get the improved objective function:

$$\begin{aligned}\min_{\{\mathbf{S}^v\},\mathbf{U},\mathbf{W},\overline{\mathbf{Y}},\mathbf{E},\{\alpha^v\}} & \frac{1}{2}\|\overline{\mathbf{Y}}-\mathbf{W}^T\mathbf{X}\|_F^2+\sum_{v=1}^{m}\sum_{i,j=1}^{n}\|\boldsymbol{x}_i^v-\boldsymbol{x}_j^v\|_2^2 s_{ij}^v \\ &+\sum_{v=1}^{m}\alpha^v\|\mathbf{U}-\mathbf{S}^v\|_F^2+\lambda\mathbf{Tr}(\overline{\mathbf{Y}}\mathbf{L}_u\overline{\mathbf{Y}}^T)+\sigma\|\mathbf{E}\|_1 \\ &+\gamma\|\mathbf{W}\|_F^2+\beta\sum_{v=1}^{m}\sum_{i}^{n}\|\mathbf{s}_i^v\|_2^2 \\ s.t.\ \mathbf{Y}=\overline{\mathbf{Y}}+\mathbf{E},\forall v, & s_{ii}^v=0, s_{ij}^v\geq 0, \mathbf{1}^T\mathbf{s}_i^v=1, \\ \forall i, u_{ij}\geq 0, & \mathbf{1}^T\boldsymbol{u}_i=1. \end{aligned} \tag{6}$$

where λ, σ and γ are three parameters to trade-off different regularization terms. $\mathbf{W}\in\mathbb{R}^{D\times q}$ is the prediction model to be trained, which is constrained by $\mathbf{F}$ norm to ensure that the complexity of the model is acceptable. As is shown in Eq. 6, we integrate the multi-view data into a unified graph, and adopt this graph to align the observed candidate labels to obtain accurate labels, and then complete the multi-view partial multi-label classification task.

IV. OPTIMIZATION

Due to the fact that the objective function Eq. 6 is not jointly convex, so the closed-form solution cannot be directly calculated. Thus, we use the augmented Lagrange multiplier strategy to transform the optimization of the proposed objective function into several subproblems to optimize separately.

A. UPDATE $\mathbf{S}^v$

Fixing $\mathbf{U}$, $\overline{\mathbf{Y}}$, $\mathbf{W}$, $\mathbf{E}$ and α^v, we update $\mathbf{S}^v$ by solving the following problem:

$$\begin{aligned}\min_{\{\mathbf{S}^v\}} & \sum_{v=1}^{m}\sum_{i,j=1}^{n}\|\boldsymbol{x}_i^v-\boldsymbol{x}_j^v\|_2^2 s_{ij}^v+\sum_{v=1}^{m}\alpha^v\|\mathbf{U}-\mathbf{S}^v\|_F^2 \\ &+\beta\sum_{v=1}^{m}\sum_{i}^{n}\|\mathbf{s}_i^v\|_2^2 \\ s.t.\ \forall v, & s_{ii}^v=0, s_{ij}^v\geq 0, \mathbf{1}^T\mathbf{s}_i^v=1, \\ \forall i, u_{ij}\geq 0, & \mathbf{1}^T\boldsymbol{u}_i=1. \end{aligned} \tag{7}$$

Since $\mathbf{S}^v$ of each view is independent, we simplify the above formula as follow:

$$\begin{aligned}\min_{\mathbf{S}^v} & \sum_{i,j=1}^{n}\|\boldsymbol{x}_i^v-\boldsymbol{x}_j^v\|_2^2 s_{ij}^v+\alpha^v\|\mathbf{U}-\mathbf{S}^v\|_F^2+\beta\sum_{i}^{n}\|\mathbf{s}_i^v\|_2^2 \\ s.t.\ \forall v, & s_{ii}^v=0, s_{ij}^v\geq 0, \mathbf{1}^T\mathbf{s}_i^v=1, \\ \forall i, u_{ij}\geq 0, & \mathbf{1}^T\boldsymbol{u}_i=1. \end{aligned} \tag{8}$$

We assume that each point has its neighbor nodes with similarity, so we adopt a k-nearest neighbor method to optimize and update the above formula. Specifically, we learn $\mathbf{s}_i$ in $\mathbf{S}^v$ with k nonzero values, we give the final solution below and omit the detailed steps. Denote $\boldsymbol{e}_i$ is a vector with the j-th element as $e_{ij}=\|\boldsymbol{x}_i-\boldsymbol{x}_j\|_2^2$, then the problem (8) can be simplified as as follow:

$$\begin{aligned}\min_{\boldsymbol{s}_i^v} & \frac{1}{2}\|\boldsymbol{s}_i^v+\frac{\boldsymbol{e}_i}{2\beta}\|_2^2+\frac{1}{2\beta}\alpha^v\|\boldsymbol{u}_i-\boldsymbol{s}_i^v\|_2^2 \\ s.t.\ \forall v, & s_{ii}^v=0, s_{ij}^v\geq 0, \mathbf{1}^T\mathbf{s}_i^v=1, \end{aligned} \tag{9}$$

The Lagrange multiplier method is adopted here to convert the constraint term into an augmented term in the objective function, and assume that its partial derivative with respect to the variable $\boldsymbol{s}_i^v$ is 0, and the value of β can be obtained as follows:

$$\beta=\frac{ke_{i,k+1}-\sum_{h=1}^{k}e_{ih}-2k\alpha^v u_{i,k+1}-2\alpha^v}{2} \tag{10}$$

And then, we can get the final solution for s_i^v as follows [8]:

$$s_{ij}^v=\begin{cases}\dfrac{e_{i,k+1}-e_{ij}+2w_v u_{ij}-2w_v u_{i,k+1}}{ke_{i,k+1}-\sum_{h=1}^{k}e_{ih}-2kw_v u_{i,k+1}+2\sum_{h=1}^{k}w_v u_{ih}}, \\ \qquad j\leq k \\ 0, \quad j>k\end{cases} \tag{11}$$

B. UPDATE $\mathbf{U}$

Fixing $\mathbf{S}^v$, $\overline{\mathbf{Y}}$, $\mathbf{W}$, $\mathbf{E}$ and α^v, update $\mathbf{U}$. When $\mathbf{S}^v$, $\overline{\mathbf{Y}}$, $\mathbf{W}$, $\mathbf{E}$ and α^v are fixed, update $\mathbf{U}$ is to solve the following problem:

$$\begin{aligned}\min_{\mathbf{U}} & \sum_{v=1}^{m}\alpha^v\|\mathbf{U}_{\mathbf{S}}^v\|_F^2+\lambda\mathbf{Tr}(\overline{\mathbf{Y}}\mathbf{L}_{\mathbf{u}}\overline{\mathbf{Y}}^T) \\ s.t.\ & u_{ij}\geq 0, \mathbf{1}^T\boldsymbol{u}_i=1. \end{aligned} \tag{12}$$

By taking derivative of $\mathbf{U}$ and make it to 0, we obtain

$$\mathbf{U}=\sum_{v=1}^{m}\alpha^v\mathbf{S}^v+\frac{\lambda}{2}\overline{\mathbf{Y}}^T\overline{\mathbf{Y}} \tag{13}$$

C. UPDATE $\overline{\mathbf{Y}}$

We update $\overline{\mathbf{Y}}$ with *Augmented Lagrange Multiplier(ALM)*. To optimize the objective function more conveniently, we convert Eq. 6 to the following augmented Lagrange function:

$$\begin{aligned}\min_{\{\mathbf{S}^v\},\mathbf{U},\overline{\mathbf{Y}},\mathbf{E},\alpha^v} & \frac{1}{2}\|\overline{\mathbf{Y}}-\mathbf{W}^T\mathbf{X}\|+\sum_{v=1}^{m}\sum_{i,j=1}^{n}\|\boldsymbol{x}_i^v-\boldsymbol{x}_j^v\|_2^2 s_{ij}^v \\ &+\sum_{v=1}^{m}\alpha^v\|\mathbf{U}-\mathbf{S}^v\|_F^2+\lambda\mathbf{Tr}(\overline{\mathbf{Y}}\mathbf{L}_{\mathbf{u}}\overline{\mathbf{Y}}^T) \\ &+\gamma\|\mathbf{W}\|_F^2+\sigma\|\mathbf{E}\|_1+\beta\sum_{v=1}^{m}\sum_{i}^{n}\|\mathbf{s}_i^v\|_2^2 \\ &+<\mathbf{Y}_1,\mathbf{Y}-\overline{\mathbf{Y}}-\mathbf{E}>+\frac{\mu}{2}\|\mathbf{Y}-\overline{\mathbf{Y}}-\mathbf{E}\|_F^2 \end{aligned}$$

TABLE 1. Characteristics of our employed datasets.

Datasets	Samples	Views	$D_{min-max}$	Labels
Emotions	593	2	8-64	6
Scene	2407	2	98-196	6
Yeast	2417	2	24-79	14
Corel5k	4999	4	100-4096	260
Pascal	9963	5	100-4096	20
Mirflickr	25000	5	100-4096	38

$D_{min-max}$ is the smallest-largest dimensions of features.

$$s.t.\ \forall v, s_{ii}^v = 0, s_{ij}^v \geq 0, \mathbf{1}^T \mathbf{s}_i^v = 1, \quad \forall i, u_{ij} \geq 0, \mathbf{1}^T \boldsymbol{u}_i = 1. \tag{14}$$

following equivalent problem: where $\mathbf{Y}_1 \in \mathbb{R}^{q \times n}$ is Lagrange multiplier matrix, and μ is penalty parameters.

Fixing $\mathbf{S}^v$, $\mathbf{U}$, $\mathbf{W}$, $\mathbf{E}$ and α^v, the Eq. 14 can be converted into the following equivalent problem:

$$\min_{\overline{\mathbf{Y}}} \frac{1}{2}\|\overline{\mathbf{Y}} - \mathbf{W}^T\mathbf{X}\| + \lambda \mathbf{Tr}(\overline{\mathbf{Y}}\mathbf{L_u}\overline{\mathbf{Y}}^T) + \frac{\mu}{2}\|\mathbf{Y} - \overline{\mathbf{Y}} - \mathbf{E} + \frac{\mathbf{Y}_1}{\mu}\|_F^2 \tag{15}$$

By taking the derivative with respect to $\overline{\mathbf{Y}}$ and setting it to 0, $\overline{\mathbf{Y}}$ can be updated by

$$\overline{\mathbf{Y}} = (\mu(\mathbf{Y} - \mathbf{E}) + \mathbf{Y}_1)[(1+\mu)\mathbf{I} + \lambda(\mathbf{L}_u^T + \mathbf{L}_u)]^{-1} \tag{16}$$

where $\mathbf{Y}_3 \in \mathbb{R}^{n \times n}$ is a identity matrix.

D. UPDATE E

Fixing the other variables, the subproblem of $\mathbf{E}$ can be re-written as:

$$\min_{\mathbf{E}} \frac{1}{2}\|\mathbf{E} - (\mathbf{Y} - \mathbf{W}^T\mathbf{X})\|_F^2 + \sigma\|E\|_1 \tag{17}$$

Since the first term is Frobenius norm and the second term is l1 norm, the $\mathbf{E}_{k+1}$ is given by soft-thresholding the entries of $\mathbf{G}_k = \mathbf{Y} - \mathbf{W}^T\mathbf{X}$ [29]:

$$\mathbf{E}_{k+1} = arg\min_{\mathbf{E}} \epsilon\|E\|_1 + \frac{1}{2}\|\mathbf{E} - \mathbf{G}_k\|_F^2 = \mathcal{S}_\epsilon[\mathbf{G}_k] \tag{18}$$

The details are as follows:

$$\mathbf{E}_{k+1} = \begin{cases} (\mathbf{Y} - \mathbf{W}^T\mathbf{X}) - \sigma, & (\mathbf{Y} - \mathbf{W}^T\mathbf{X}) > \sigma \\ (\mathbf{Y} - \mathbf{W}^T\mathbf{X}) + \sigma, & (\mathbf{Y} - \mathbf{W}^T\mathbf{X}) < -\sigma \\ 0, & \text{otherwise} \end{cases} \tag{19}$$

E. UPDATE W

Fixing $\mathbf{S}^v$, $\mathbf{U}$, $\overline{\mathbf{Y}}$, $\mathbf{E}$ and α^v, the optimization problem 6 is transformed into

$$\min_{\mathbf{W}} \frac{1}{2}\|\overline{\mathbf{Y}} - \mathbf{W}^T\mathbf{X}\|_F^2 + \gamma\|\mathbf{W}^T\|_F^2 \tag{20}$$

Here, $\mathbf{W}$ can be updated following $\mathbf{W}^T = \overline{\mathbf{Y}}\mathbf{X}^T(\mathbf{X}\mathbf{X}^T + 2\gamma\mathbf{I})$ and $\mathbf{I} \in \mathbb{R}^{D \times D}$ is an identity matrix.

F. UPDATE α^v

Since the α^v of each view is independent, we update each α^v separately and we adopt an adaptive method to update α^v, i.e.,

$$\alpha^v = \frac{1}{2\sqrt{\|\mathbf{U} - \mathbf{S}^v\|_F^2}} \tag{21}$$

G. UPDATE $\mathbf{Y}_1$ AND μ

We update the Lagrange multiplier matrix $\mathbf{Y}_1$ and regularization term μ by LADM:

$$\mu^{k+1} = \min(\mu_{max}, \eta\mu^k)$$
$$\mathbf{Y}_1^{k+1} = \mathbf{y}_1^k + \mu^{k+1}(\mathbf{Y} - \overline{\mathbf{Y}} - \mathbf{E}) \tag{22}$$

where η is a positive scalar. In summary, in the whole optimization process, we first initialize each variable, and then repeat the above steps until the function converges or reaches the maximum number of iterations.

V. EXPERIMENT

In this section, we conduct extensive experiments on five multi-view partial multi-label data sets to comprehensively evaluate the performance of our proposed method.

A. EXPERIMENTAL SETUP

Data sets In this paper, a multi-label classification experiment is conducted on six commonly used data sets to verify the performance of the proposed method. They are respectively: *Emotions* [30], *Yeast* [31], *Scene*, *Corel5k* [32] and *PASCAL VOC07* [33]. To be specific, the *Emotions* dataset is associated with 593 pieces of music depictedby, each of which is depicted by two kinds of features. *Yeast* is a Biological dataset which contaisd 2417 data points and the two kinds of features of each example correspond to the genetic expression and phylogenetic profile of a gene. *Scene* has 2407 images, each of which is depicted by two typies of features: the luminance and chromaticity of color. *Corel5k* and *PASCAL* are two widely used multi-view multi-label image datasets. *Corel5k* contains 4999 images, and each sample consists of 4 view features, which are GIST, HSV, HUE, DIFT. For the *PASCAL* dataset, besides the same four views daopted by *Corel5k*, the tag features are also added for each sample. Table 1 illustrates the characteristics of the above datasets, consisting of the number of samples, classes and features.

Compared methods The floowing five representative methods are adopted as baselines, and they are as follows:

- **GRADIS** [3]: An multi-view partial multi-label approach which fuses multi-view representation and disambiguating candidate label based on label propagation.
- **McWL** [34]: An multi-view multi-label approach which enforces the optimization of multi-view integration and of MC-based classification within a unified objective function.
- **ICM2L** [35]: An multi-view multi-label approach which explores the individuality and commonality

TABLE 2. Experimental results of each comparing approach in terms of *Average Precision*, where the best performance (the larger the better) on each dataset and specific value of p is shown in bold face.

Dataset	p	ours	GRADIS	ICM2L	McWI	PAR-VLS	NATAL
Emotions	0.3	**0.799±0.049**	0.736±0.041	0.680±0.033	0.577±0.015	0.779±0.032	0.772±0.031
	0.5	**0.772±0.041**	0.724±0.027	0.673±0.042	0.562±0.021	0.765±0.044	0.764±0.023
	0.7	**0.756±0.011**	0.706±0.041	0.633±0.035	0.554±0.034	0.753±0.038	0.732±0.024
Yeast	0.3	**0.731±0.009**	0.496±0.020	0.597±0.019	0.728±0.019	0.720±0.012	0.603±0.015
	0.5	**0.725±0.010**	0.485±0.015	0.590±0.016	0.723±0.020	0.712±0.011	0.598±0.012
	0.7	**0.719±0.011**	0.477±0.013	0.571±0.015	0.716±0.017	0.701±0.012	0.595±0.015
Scene	0.3	**0.803±0.011**	0.767±0.009	0.735±0.011	0.475±0.008	0.554±0.011	0.683±0.015
	0.5	**0.795±0.009**	0.755±0.008	0.718±0.010	0.461±0.010	0.548±0.013	0.675±0.011
	0.7	**0.786±0.010**	0.746±0.010	0.705±0.009	0.446±0.015	0.539±0.011	0.664±0.009
Corel5k	0.3	**0.433±0.011**	0.421±0.011	0.226±0.014	0.400±0.010	0.124±0.015	0.302±0.010
	0.5	**0.431±0.011**	0.411±0.008	0.213±0.013	0.389±0.009	0.115±0.012	0.277±0.011
	0.7	**0.426±0.012**	0.402±0.009	0.192±0.011	0.379±0.010	0.109±0.014	0.265±0.009
Pascal	0.3	**0.590±0.011**	0.554±0.017	0.506±0.014	0.529±0.056	0.536±0.010	0.358±0.011
	0.5	**0.580±0.010**	0.542±0.011	0.488±0.011	0.495±0.016	0.521±0.009	0.349±0.010
	0.7	**0.573±0.009**	0.533±0.028	0.459±0.016	0.466±0.010	0.511±0.011	0.338±0.012
Mirflickr	0.3	**0.731±0.011**	0.722±0.010	0.616±0.017	0.593±0.012	0.612±0.011	0.552±0.013
	0.5	**0.718±0.010**	0.702±0.009	0.602±0.014	0.577±0.012	0.598±0.014	0.541±0.012
	0.7	**0.701±0.014**	0.693±0.011	0.596±0.011	0.568±0.011	0.569±0.017	0.528±0.011

TABLE 3. Experimental results of each comparing approach in terms of *Hamming Loss*, where the best performance (the smaller the better) on each dataset and specific value of p is shown in bold face.

Dataset	p	ours	GRADIS	ICM2L	McWI	PAR-VLS	NATAL
Emotions	0.3	**0.204±0.011**	0.215±0.019	0.333±0.012	0.467±0.012	0.226±0.019	0.242±0.034
	0.5	**0.232±0.013**	0.237±0.016	0.341±0.015	0.553±0.019	0.235±0.021	0.255±0.016
	0.7	0.247±0.012	**0.244±0.026**	0.344±0.014	0.544±0.024	**0.244±0.017**	0.279±0.021
Yeast	0.3	**0.179±0.008**	0.193±0.010	0.285±0.011	0.271±0.010	0.285±0.014	0.290±0.009
	0.5	**0.182±0.011**	0.194±0.007	0.309±0.013	0.277±0.012	0.295±0.004	0.290±0.007
	0.7	**0.185±0.010**	0.198±0.005	0.297±0.011	0.324±0.009	0.315±0.009	0.293±0.007
Scene	0.3	**0.102±0.005**	0.108±0.006	0.122±0.004	0.221±0.008	0.155±0.012	0.195±0.005
	0.5	**0.106±0.004**	0.114±0.004	0.126±0.005	0.228±0.005	0.161±0.009	0.204±0.010
	0.7	**0.108±0.005**	0.118±0.006	0.134±0.007	0.235±0.006	0.167±0.010	0.206±0.008
Corel5k	0.3	**0.020±0.000**	0.025±0.001	0.023±0.001	0.021±0.000	0.124±0.015	0.033±0.009
	0.5	**0.020±0.000**	0.026±0.000	0.023±0.001	0.021±0.000	0.115±0.012	0.040±0.010
	0.7	**0.023±0.000**	0.028±0.001	**0.023±0.000**	0.024±0.000	0.109±0.014	0.045±0.008
Pascal	0.3	**0.136±0.002**	0.155±0.002	0.142±0.002	0.142±0.014	0.536±0.010	0.222±0.005
	0.5	**0.137±0.002**	0.160±0.003	0.144±0.001	0.148±0.002	0.521±0.009	0.235±0.006
	0.7	**0.139±0.003**	0.167±0.001	0.145±0.002	0.192±0.001	0.511±0.011	0.251±0.010
Mirflickr	0.3	**0.022±0.000**	0.027±0.000	0.023±0.000	0.112±0.004	0.085±0.008	0.062±0.005
	0.5	0.024±0.000	0.033±0.000	**0.023±0.000**	0.122±0.002	0.091±0.010	0.074±0.006
	0.7	**0.025±0.000**	0.034±0.001	0.026±0.000	0.143±0.005	0.092±0.011	0.088±0.010

TABLE 4. Experimental results of each comparing approach in terms of *One Error*, where the best performance (the smaller the better) on each dataset and specific value of p is shown in bold face.

Dataset	p	ours	GRADIS	ICM2L	McWI	PAR-VLS	NATAL
Emotions	0.3	**0.029±0.024**	0.116±0.080	0.200±0.045	0.532±0.025	0.246±0.067	0.304±0.088
	0.5	**0.039±0.020**	0.150±0.080	0.212±0.038	0.549±0.028	0.252±0.083	0.313±0.034
	0.7	**0.042±0.028**	0.166±0.136	0.224±0.046	0.579±0.057	0.263±0.077	0.387±0.040
Yeast	0.3	**0.068±0.008**	0.250±0.067	0.166±0.024	0.292±0.030	0.212±0.023	0.381±0.025
	0.5	**0.069±0.011**	0.264±0.099	0.177±0.031	0.282±0.040	0.230±0.018	0.382±0.016
	0.7	**0.047±0.016**	0.285±0.067	0.182±0.035	0.310±0.021	0.245±0.026	0.387±0.033
Scene	0.3	**0.215±0.015**	0.231±0.021	0.444±0.014	0.632±0.074	0.331±0.024	0.272±0.019
	0.5	**0.221±0.016**	0.243±0.017	0.457±0.016	0.646±0.069	0.342±0.021	0.285±0.018
	0.7	**0.229±0.014**	0.258±0.018	0.472±0.015	0.663±0.075	0.355±0.025	0.298±0.016
Corel5k	0.3	**0.174±0.013**	0.176±0.011	0.431±0.013	0.485±0.023	0.505±0.025	0.649±0.013
	0.5	**0.174±0.010**	0.177±0.013	0.448±0.015	0.489±0.019	0.515±0.019	0.653±0.011
	0.7	**0.174±0.008**	0.177±0.009	0.461±0.011	0.488±0.030	0.533±0.020	0.678±0.009
Pascal	0.3	**0.281±0.012**	0.323±0.020	0.333±0.014	0.573±0.011	0.516±0.013	0.739±0.012
	0.5	**0.288±0.015**	0.331±0.015	0.388±0.024	0.612±0.019	0.532±0.012	0.748±0.010
	0.7	**0.297±0.013**	0.344±0.018	0.444±0.021	0.640±0.013	0.544±0.015	0.755±0.008
Mirflickr	0.3	**0.498±0.015**	0.618±0.018	0.722±0.012	0.731±0.016	0.696±0.013	0.775±0.014
	0.5	**0.512±0.012**	0.622±0.016	0.744±0.016	0.750±0.019	0.712±0.015	0.783±0.016
	0.7	**0.518±0.018**	0.643±0.018	0.765±0.015	0.761±0.020	0.741±0.012	0.795±0.011

information of multi-view data in a unified subspace representation learning model.

- **PARTICLE-VLS** [36]: A partial multi-label approach which is a two-stage classification, which first

TABLE 5. Experimental results of each comparing approach in terms of *Ranking Loss*, where the best performance (the smaller the better) on each dataset and specific value of p is shown in bold face.

Dataset	p	ours	GRADIS	ICM2L	McWI	PAR-VLS	NATAL
Emotions	0.3	0.165±0.015	**0.155±0.022**	0.325±0.024	0.435±0.023	0.215±0.027	0.197±0.039
	0.5	**0.175±0.011**	0.180±0.011	0.285±0.014	0.449±0.036	0.234±0.038	0.198±0.027
	0.7	**0.184±0.014**	0.187±0.029	0.315±0.021	0.447±0.023	0.243±0.040	0.236±0.022
Yeast	0.3	**0.214±0.011**	0.232±0.024	0.367±0.011	0.389±0.014	0.223±0.008	0.346±0.022
	0.5	**0.221±0.012**	0.240±0.019	0.376±0.012	0.394±0.015	0.229±0.008	0.350±0.016
	0.7	**0.231±0.013**	0.344±0.016	0.387±0.010	0.407±0.015	0.241±0.007	0.354±0.014
Scene	0.3	**0.049±0.004**	0.061±0.005	0.107±0.006	0.203±0.044	0.133±0.011	0.165±0.010
	0.5	**0.054±0.010**	0.067±0.005	0.108±0.005	0.221±0.040	0.138±0.007	0.166±0.011
	0.7	**0.061±0.008**	0.075±0.008	0.114±0.007	0.257±0.050	0.144±0.008	0.172±0.008
Corel5k	0.3	**0.105±0.007**	0.114±0.005	0.227±0.007	0.224±0.011	0.331±0.005	0.373±0.009
	0.5	**0.111±0.007**	0.116±0.004	0.244±0.008	0.235±0.013	0.344±0.004	0.422±0.005
	0.7	**0.121±0.006**	**0.121±0.005**	0.247±0.006	0.248±0.014	0.352±0.004	0.438±0.004
Pascal	0.3	**0.178±0.008**	0.184±0.009	0.183±0.008	0.198±0.008	0.212±0.005	0.484±0.011
	0.5	**0.186±0.010**	0.190±0.005	0.199±0.010	0.220±0.010	0.231±0.004	0.498±0.013
	0.7	**0.194±0.006**	0.196±0.004	0.207±0.011	0.242±0.004	0.244±0.004	0.509±0.012
Mirflickr	0.3	**0.122±0.005**	**0.122±0.009**	0.141±0.007	0.211±0.006	0.246±0.007	0.616±0.015
	0.5	**0.129±0.006**	0.133±0.007	0.159±0.011	0.233±0.011	0.255±0.008	0.637±0.016
	0.7	**0.142±0.005**	0.147±0.006	0.166±0.009	0.256±0.010	0.271±0.005	0.655±0.016

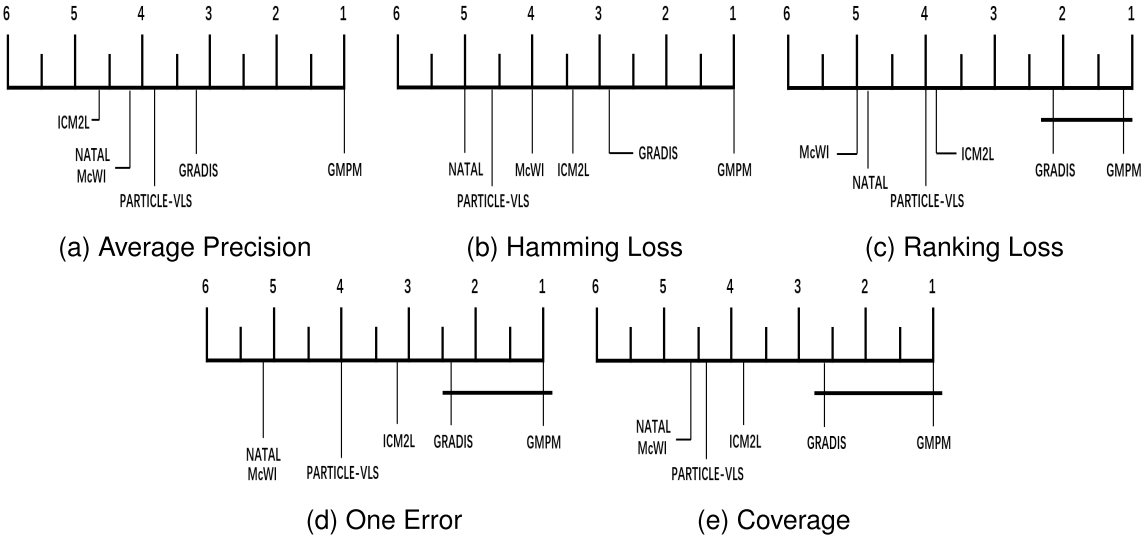


FIGURE 3. Comparison of GMPM (control approach) against other approaches with *Bonferroni-Dunn* test. Approaches not connected with PAKS are considered to have significantly distinguishable performance from GMPM (CD = 1.759 at 0.05 significance level).

transforms the partial multi-label learning into multi-label learning by a label propagation procedure and then a calibrated label ranking model is induced to the PML method **PARTICLE-VLS**.

- **NATAL** [25]: A partial multi-label approach which transfers the traditional PMP problem to a feature completion problem, and induces the multi-label classifiers by mapping the completed features to all candidate labels.

Parameter Setup For all the baseline methods, we randomly select 20% of the training data from the given multi-view data to search for the best parameters, among which the best parameters are obtained through five-fold grid search. For our proposed model, the trade-off parameters λ, γ and ϵ are chosen from $[1e-3, 1e-2, 1e-1, 1e-0, 1e-1, 1e-2]$ via five-fold cross validation, respectively.

Evaluation Metrics In order to measure the performance of multi-classification, six popular multi-label metrics are adopted here to evaluate each compared method, including Average Precision, Hamming Loss, Ranking Loss, One Error and Coverage, whose detailed definitions can be found in [37].

B. EXPERIMENTAL RESULTS

Five separate runs of five-fold cross-validation are performed on each data set, and we record their mean values and standard deviation of each evaluation measurement, the detailed comparison results of different methods on five data sets are demonstrated in Table 2 - 6. According to Table 2 - 6, we can see our proposed method achieves either superior or comparable performance against the five compared methods:

TABLE 6. Experimental results of each comparing approach in terms of *Coverage*, where the best performance (the smaller the better) on each dataset and specific value of p is shown in bold face.

Dataset	p	ours	GRADIS	ICM2L	McWI	PAR-VLS	NATAL
Emotions	0.3	**1.882±0.149**	2.157±0.047	2.400±0.112	3.221±0.144	2.020±0.227	1.965±0.184
	0.5	**1.974±0.123**	2.205±0.021	2.545±0.135	3.275±0.180	2.047±0.292	1.985±0.164
	0.7	**2.081±0.033**	2.274±0.034	2.672±0.122	3.218±0.208	2.105±0.256	2.123±0.126
Yeast	0.3	**6.614±0.337**	7.535±0.248	7.916±0.352	8.687±0.083	7.410±0.189	8.800±0.242
	0.5	**6.809±0.191**	7.821±0.320	8.124±0.295	8.942±0.334	7.661±0.124	8.864±0.265
	0.7	**7.005±0.252**	8.102±0.235	8.667±0.284	8.830±0.309	7.864±0.194	8.937±0.193
Scene	0.3	**0.308±0.052**	0.372±0.035	0.444±0.025	0.628±0.020	0.754±0.054	0.514±0.034
	0.5	**0.337±0.031**	0.395±0.046	0.504±0.035	0.675±0.031	0.781±0.064	0.535±0.028
	0.7	**0.361±0.046**	0.424±0.051	0.571±0.028	0.724±0.035	0.804±0.051	0.581±0.038
Corel5k	0.3	**68.33±3.313**	71.45±2.256	121.0±3.214	72.50±2.038	243.5±5.628	148.3±3.865
	0.5	**64.57±3.567**	73.11±3.567	134.0±3.241	74.08±1.854	255.6±6.258	158.9±2.998
	0.7	**74.26±4.035**	76.22±3.221	127.5±3.554	76.80±1.932	268.2±4.325	168.5±3.562
Pascal	0.3	**4.652±0.225**	4.895±0.221	4.944±0.321	4.998±0.184	5.168±0.167	10.546±0.214
	0.5	**4.758±0.267**	5.025±0.315	4.933±0.285	5.511±0.232	5.661±0.264	11.664±0.315
	0.7	**4.905±0.241**	5.331±0.215	5.111±0.275	5.969±0.103	6.014±0.211	13.14±0.221
Mirflickr	0.3	**12.486±0.812**	14.225±0.912	23.521±0.669	18.227±0.861	35.851±0.775	41.221±1.992
	0.5	**13.548±0.991**	16.753±0.315	29.149±0.490	21.562±0.968	36.115±1.121	47.873±2.012
	0.7	**14.224±0.768**	18.946±0.215	35.228±0.791	28.261±0.879	41.547±0.967	55.126±2.331

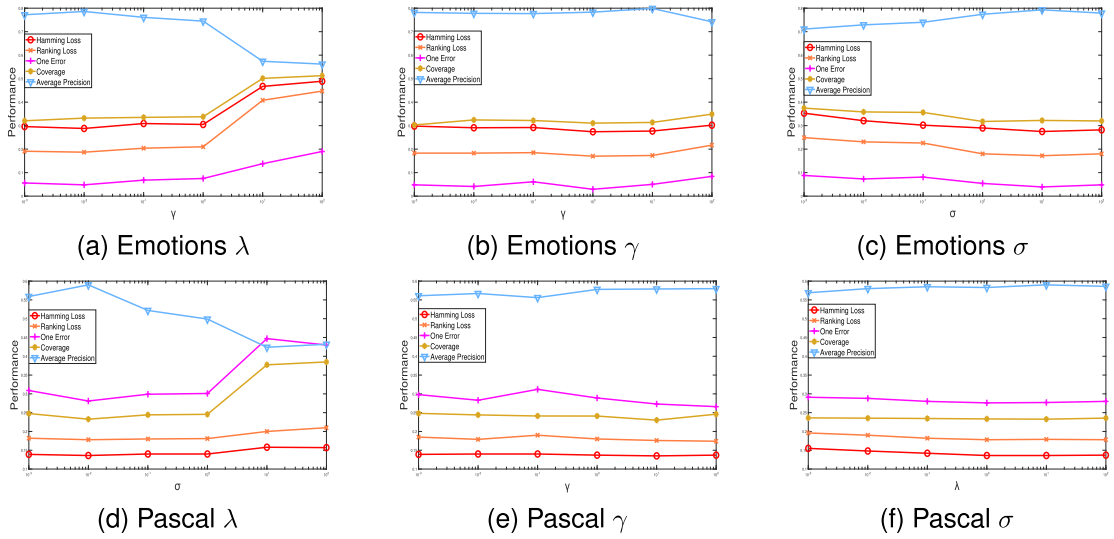


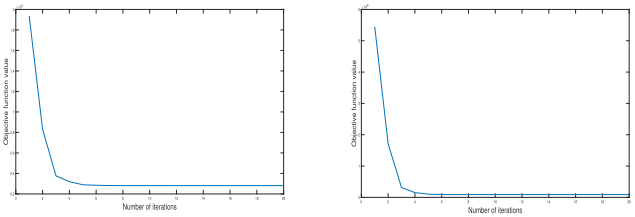
FIGURE 4. The changing trend of the performance of GMPM (*average precision, hamming loss, ranking loss, one error, coverage*) with respect to different parameters. (a) λ varies on *Emotions*; (b) γ varies on *Emotions*; (c) σ varies on *Emotions*; (d) λ varies on *Pascal*; (e) γ varies on *Pascal*; (f) σ varies on *Pascal*.

- Compared with other baseline methods, our **GMPM** model has achieved gratifying performance in many cases. For example, GMPM outperforms **GRADIS** and **PAR-VLS** in 97.3% cases and 98.6% cases, respectively.
- In terms of different evaluation indicators, our method is superior to the comparison method. For example, out proposed model is superior to all comparison methods in terms of Average Precision, One Error and Coverage metrics. In addition, **GMPM** outperforms other baselines over 96.0% on Hamming Loss metric, it is also superior or comparable to other compared methods in 98.6% cases in terms of Ranking Loss metric.
- Finally, our proposed method shows prominent advantages on all datasets. In particular, **GMPM** is superior to most baselines in nearly 80% of cases, and even achieves the best performance on *Yeast*, *Scene* and *Pascal* data sets.

To further evaluate the superiority of **GMPM** against other comparing methods, we conduct statistical analysis among all comparing methods, where *Friedman Test* [38] is employed to analyze the relative performance among all comparing approaches. Figure 3 shows the CD diagrams on five employed evaluation metrics, where the average performance rank of each comparing approach is recorded along the axis. As illustrated in Figure 3, **GMPM** always ranks 1st on all evaluation metrics, and it is comparable to **GRADIS** approach and significantly outperforms other comparing approaches.

C. FURTHER ANALYSIS

a) Complexity Analysis: In this section, we briefly analyze the computational complexity for **GMPM**. There are five main parts in our optimization procedure, i.e., $\mathbf{S}^1, \ldots, \mathbf{S}^m$, $\mathbf{U}$, $\overline{\mathbf{Y}}$, $\mathbf{E}$ and $\mathbf{W}$. For simplicity, we assume that the number of neighbour is k. Specifically,



(a) Emotions (b) Pascal

FIGURE 5. Convergence analysis.

1) the update of $\mathbf{S}^1, \ldots, \mathbf{S}^m$ has the computational complexity of $O(mnk)$.

2) the update of $\mathbf{U}$ resulting a total complexity of $O(qn)$.

3) the update of $\overline{\mathbf{Y}}$ involves matrix inversion, which leads to a computational complexity of $O(qn^2 + n^3)$.

4) the update of $\mathbf{E}$ involves SVD, which brings a computational complexity of $O(qn^2)$.

5) the update of $\mathbf{W}$ has the computational complexity of $O(qnD)$.

Consider that $n \gg D \gg q > k$, the computational complexity of the whole training process can be approximated as $O(T * qn^2)$, where T is the number of iterations.

b) Sensitivity Analysis: We analyze the sensitivity of **GMPM** in terms of its three parameters λ, γ and σ. Figure 4 shows the performance of **GMPM** on each evaluation metric under different parameter configurations on *Emotions* and *Pascal* datasets. In our experiments, we set the parameters λ, γ and σ among $[1e{-}3, 1e2]$ via cross-validation.

c) Convergence Analysis: We illustrate the convergence curves of GMPM on *Emotions* and *Pascal* datasets in Figure 5. According to Figure 5, we can observe that the value of objective function rapidly decreases after a few iterations and finally becomes relatively stable, which empirically validates the convergence of our proposed GMPM.

VI. CONCLUSION

In MVPML, each sample is described by several heterogeneous feature representations and associated with a candidate label set which are partially valid. In this paper, a one-stage method named GMPM towards MVPML is proposed. GMPM adopts the similarity graph of each view and sparse decomposition to disambiguate the candidate label set, and then the learned ground-truth labels are used for training the predictive model. To the best of our knowledge, this is the first attempt to deal with the MVPML problem by taking label disambiguation and classification as a whole. Enormous experimental results have proven that our proposed method can achieve superior performance against state-of-the-art methods. In the future, it is interesting to extend GMPM to deal with large-scale multi-view partial multi-label learning problem by using deep learning technology.

REFERENCES

[1] M. Xie and S. Huang, "Partial multi-label learning," in *Proc. AAAI Conf. Artif. Intell.*, vol. 32, no. 1, 2018, pp. 4302–4309.
[2] L. Sun, S. Feng, T. Wang, C. Lang, and Y. Jin, "Partial multi-label learning by low-rank and sparse decomposition," in *Proc. AAAI Conf. Artif. Intell.*, 2019, vol. 33, no. 1, pp. 5016–5023.
[3] Z. Chen, X. Wu, Q. Chen, Y. Hu, and M. Zhang, "Multi-view partial multi-label learning with graph-based disambiguation," in *Proc. AAAI Conf. Artif. Intell.*, 2020, vol. 34, no. 4, pp. 3553–3560.
[4] J.-H. Wu, X. Wu, Q.-G. Chen, Y. Hu, and M.-L. Zhang, "Feature-induced manifold disambiguation for multi-view partial multi-label learning," in *Proc. 26th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2020, pp. 557–565.
[5] W. Liu, S. Feng, and H. Tian, "Graph-based multi-view partial multi-label learning," in *Proc. IEEE 13th Int. Symp. Parallel Archit., Algorithms Program. (PAAP)*, Nov. 2022, pp. 1–5.
[6] J. Zhao, X. Xie, X. Xu, and S. Sun, "Multi-view learning overview: Recent progress and new challenges," *Inf. Fusion*, vol. 38, pp. 43–54, Nov. 2017.
[7] M. Chen, L. Huang, C. Wang, and D. Huang, "Multi-view clustering in latent embedding space," in *Proc. AAAI Conf. Artif. Intell.*, 2020, vol. 34, no. 4, pp. 3513–3520.
[8] H. Wang, Y. Yang, and B. Liu, "GMC: Graph-based multi-view clustering," *IEEE Trans. Knowl. Data Eng.*, vol. 32, no. 6, pp. 1116–1129, May 2019.
[9] X. Zhu, S. Zhang, W. He, R. Hu, C. Lei, and P. Zhu, "One-step multi-view spectral clustering," *IEEE Trans. Knowl. Data Eng.*, vol. 31, no. 10, pp. 2022–2034, Oct. 2019.
[10] A. Blum and T. Mitchell, "Combining labeled and unlabeled data with co-training," in *Proc. 11th Annu. Conf. Comput. Learn. theory*, Jul. 1998, pp. 92–100.
[11] A. Rakotomamonjy, F. Bach, S. Canu, and Y. Grandvalet, "SimpleMKL," *J. Mach. Learn. Res.*, vol. 9, pp. 2491–2521, Nov. 2008.
[12] H. Wang, Y. Yang, B. Liu, and H. Fujita, "A study of graph-based system for multi-view clustering," *Knowl.-Based Syst.*, vol. 163, pp. 1009–1019, Jan. 2019.
[13] C. Zhang, H. Fu, Q. Hu, X. Cao, Y. Xie, D. Tao, and D. Xu, "Generalized latent multi-view subspace clustering," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 1, pp. 86–99, Jan. 2020.
[14] C. Zhang, H. Fu, J. Wang, W. Li, X. Cao, and Q. Hu, "Tensorized multi-view subspace representation learning," *Int. J. Comput. Vis.*, vol. 128, no. 8, pp. 2344–2361, 2020.
[15] Z. Gu, S. Feng, R. Hu, and G. Lyu, "ONION: Joint unsupervised feature selection and robust subspace extraction for graph-based multi-view clustering," *ACM Trans. Knowl. Discovery From Data*, vol. 17, no. 5, pp. 1–23, Feb. 2023.
[16] H. Gao, F. Nie, X. Li, and H. Huang, "Multi-view subspace clustering," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 4238–4246.
[17] X. Cao, C. Zhang, H. Fu, S. Liu, and H. Zhang, "Diversity-induced multi-view subspace clustering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 586–594.
[18] Y. Liang, D. Huang, C.-D. Wang, and P. S. Yu, "Multi-view graph learning by joint modeling of consistency and inconsistency," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Jul. 27, 2022, doi: 10.1109/TNNLS.2022.3192445.
[19] M. Wang, W. Fu, S. Hao, D. Tao, and X. Wu, "Scalable semi-supervised learning by efficient anchor graph regularization," *IEEE Trans. Knowl. Data Eng.*, vol. 28, no. 7, pp. 1864–1877, Jul. 2016.
[20] F. Nie, W. Zhu, and X. Li, "Unsupervised large graph embedding," in *Proc. AAAI Conf. Artif. Intell.*, 2017, pp. 2422–2428.
[21] Y. Li, F. Nie, H. Huang, and J. Huang, "Large-scale multi-view spectral clustering via bipartite graph," in *Proc. AAAI Conf. Artif. Intell.*, 2015, pp. 2750–2756.
[22] D. Huang, C.-D. Wang, and J.-H. Lai, "Fast multi-view clustering via ensembles: Towards scalability, superiority, and simplicity," *IEEE Trans. Knowl. Data Eng.*, early access, Jan. 13, 2023, doi: 10.1109/TKDE.2023.3236698.
[23] Z. Li, G. Lyu, and S. Feng, "Partial multi-label learning via multi-subspace representation," in *Proc. 29th Int. Joint Conf. Artif. Intell.*, Jul. 2020, pp. 2612–2618.
[24] Y. Yan, S. Li, and L. Feng, "Partial multi-label learning with mutual teaching," *Knowl.-Based Syst.*, vol. 212, Jan. 2021, Art. no. 106624.
[25] G. Lyu, S. Feng, and Y. Li, "Noisy label tolerance: A new perspective of partial multi-label learning," *Inf. Sci.*, vol. 543, pp. 454–466, Jan. 2021.
[26] G. Yu, X. Chen, C. Domeniconi, J. Wang, Z. Li, Z. Zhang, and X. Wu, "Feature-induced partial multi-label learning," in *Proc. IEEE Int. Conf. Data Mining (ICDM)*, Nov. 2018, pp. 1398–1403.

[27] H. Wang, W. Liu, Y. Zhao, C. Zhang, T. Hu, and G. Chen, "Discriminative and correlative partial multi-label learning," in *Proc. 28th Int. Joint Conf. Artif. Intell.*, Aug. 2019, pp. 3691–3697.
[28] M.-L. Zhang and J.-P. Fang, "Partial multi-label learning via credible label elicitation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 10, pp. 3587–3599, Oct. 2021.
[29] J.-F. Cai, E. J. Candès, and Z. Shen, "A singular value thresholding algorithm for matrix completion," *SIAM J. Optim.*, vol. 20, no. 4, pp. 1956–1982, Jan. 2010.
[30] K. Trohidis, G. Tsoumakas, G. Kalliris, and I. P. Vlahavas, "Multi-label classification of music into emotions," in *Proc. ISMIR*, vol. 8, 2008, pp. 325–330.
[31] A. Elisseeff and J. Westom, "A kernel method for multi-labelled classification," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 14, 2001, pp. 681–687.
[32] P. Duygulu, K. Barnard, J. de Freitas, and D. Forsyth, "Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary," in *Proc. Eur. Conf. Comput. Vis.*, 2002, pp. 97–112.
[33] M. Everingham, S. M. A. Eslami, L. J. V. Gool, C. K. I. Williams, J. M. Winn, and A. Zisserman, "The Pascal visual object classes challenge: A retrospective," *Int. J. Comput. Vis.*, vol. 111, no. 1, pp. 98–136, 2015.
[34] Q. Tan, G. Yu, C. Domeniconi, J. Wang, and Z. Zhang, "Multi-view weak-label learning based on matrix completion," in *Proc. SIAM Int. Conf. Data Mining*, 2018, pp. 450–458.
[35] Q. Tan, G. Yu, J. Wang, C. Domeniconi, and X. Zhang, "Individuality- and commonality-based multiview multilabel learning," *IEEE Trans. Cybern.*, vol. 51, no. 3, pp. 1716–1727, Mar. 2021.
[36] J. Fang and M. Zhang, "Partial multi-label learning via credible label elicitation," in *Proc. AAAI Conf. Artif. Intell.*, 2019, vol. 33, no. 1, pp. 3518–3525.
[37] E. Gibaja and S. Ventura, "A tutorial on multi-label learning," *ACM Comput. Surv.*, vol. 47, no. 3, pp. 1–38, 2015.
[38] J. Demšar, "Statistical comparisons of classifiers over multiple data sets," *J. Mach. Learn. Res.*, vol. 7, pp. 1–30, Dec. 2006.

JIAZHENG YUAN was born in 1971. He received the Ph.D. degree from the Computer Science Department, Beijing Jiaotong University, in 2007. He is currently a Professor in software engineering with Beijing Open University, Beijing, China. His current research interests include graph and image processing, machine learning, and artificial intelligence.

WEI LIU was born in 1997. He received the B.S. degree in computer science from the Beijing University of Chemical Technology, in 2020. He is currently pursuing the master's degree with the School of Computer Science, Beijing Jiaotong University. His current research interests include machine learning and artificial intelligence.

ZHIBIN GU was born in 1992. He received the M.S. degree from the College of Information Engineering, North China University of Science and Technology, China, in 2020. He is currently pursuing the Ph.D. degree with the School of Computer and Information Technology, Beijing Jiaotong University. His research interests include computer vision and machine learning.

SONGHE FENG was born in 1981. He received the Ph.D. degree from the School of Computer and Information Technology, Beijing Jiaotong University, Beijing, China, in 2009. He has been a Visiting Scholar with the Department of Computer Science and Engineering, Michigan State University, from 2013 to 2014. In 2017, he visited the Department of Computer Science, Dresden University of Technology, Germany. He is currently a Full Professor with the School of Computer and Information Technology, Beijing Jiaotong University. His research interests include computer vision and machine learning.