Knowledge-Based Systems 245 (2022) 108601

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

Partial multi-label learning based on sparse asymmetric label correlations

Peng Zhao *, Shiyi Zhao, Xuyang Zhao, Huiting Liu, Xia Ji

Key Laboratory of Intelligent Computing and Signal Processing of Ministry of Education, Anhui University, Hefei 230601, China
School of Computer Science and Technology, Anhui University, Hefei 230601, China

A R T I C L E I N F O

Article history:
Received 11 August 2021
Received in revised form 12 March 2022
Accepted 14 March 2022
Available online 28 March 2022

Keywords:
Partial multi-label learning
Label confidence
Sparse asymmetric label correlations

A B S T R A C T

In many real-world applications, an instance from the training dataset of multi-label learning (MLL) often has some irrelevant labels. Traditional MLL and partial label learning (PLL) cannot deal with this problem very well. This has given rise to partial multi-label learning (PML). In this setting, it is very challenging to distinguish between the ground-truth labels and noisy labels. Most of the existing PML methods focus on identifying the ground-truth labels using label correlations, while they ignore the fact that the real label correlations often have been corrupted due to the noisy labels. Moreover, the existing PML methods usually consider the label correlations to be symmetric. However, in real-world applications, the label correlations are asymmetric. To address the above problems, we present partial multi-label learning based on sparse asymmetric label correlations (PML-SALC). PML-SALC integrates asymmetric label correlation learning and multi-label classifier learning into a unified framework. It utilizes the sparse asymmetric label correlation matrix to alleviate the negative influence of noisy labels to obtain label confidence. Moreover, PML-SALC models the relationship between the feature and label confidence, which makes the model smoother and more robust. The extensive experimental results show that the PML-SALC achieves state-of-the-art performance, which validates the effectiveness of the proposed method.

1. Introduction

The traditional classification task usually assumes that each instance has only one label. However, in real-world applications, the scenarios are often more complex and each instance often will have more than one label. For example, a text may belong to multiple fields, and an image may contain multiple objects. Therefore, multi-label learning (MLL) has emerged as a research topic in recent years [1]. MLL has been widely applied to various fields, such as image annotation [2] and information retrieval [3]. In traditional MLL, it is assumed that each training instance is accurately annotated with the relevant labels. However, in reality, the assumption often does not hold. Due to the low quality of the instance or an unreliable annotator, the instance often includes some irrelevant labels. Fig. 1 shows an example in which only parts of the annotated labels are relevant to the image. Learning in this scenario is called partial multi-label learning (PML) [4]. In contrast to partial label learning (PLL) which learns a model from partially labeled data for single label tasks, PML learns a model from partially labeled data for multi-label tasks. In this sense, PML can be regarded as a combination of MLL and PLL.

In the MLL task, each training instance has been accurately annotated with a set of ground-truth labels [1,5,6]. MLL methods can be generally divided into the following three categories. (1) MLL based on a first-order strategy: This kind of MLL usually transfers the MLL problem to q independent binary classification problems directly, where q is the number of labels. MLL based on a first-order strategy does not consider the correlations among labels [7,8]. (2) MLL based on a second-order strategy: This kind of MLL considers pairwise label correlations. Rank-SVM [9] constructs a set of pairwise labels with a relevant label and an irrelevant label for each instance and utilizes a support vector machine (SVM)-like model to solve the MLL problem. CRPC (Calibrated Ranking by Pairwise Comparison) [10] distinguishes the relevant labels from the irrelevant labels by introducing a manual calibration label for each instance in pairwise comparison. (3) MLL based on a high-order strategy: This kind of MLL considers the high-order correlations among labels. RAkEL (RAndom k-labELsets) [11] randomly divides the original label set into several small-sized subsets and trains a classifier for each label subset with the Label Powerset (LP) method. CC (Classifier Chains) [12] considers the high-order correlations among labels by taking the 0/1 label of previous classifiers as the extension of the attribute space to be input into the subsequent classifier. In addition, [13] proposes LEVI (Label Enhancement via Variational

* Corresponding author at: School of Computer Science and Technology, Anhui University, Hefei 230601, China.
E-mail address: zhaopeng_ad@163.com (P. Zhao).

https://doi.org/10.1016/j.knosys.2022.108601
0950-7051/

Fig. 1. An example of instance from the training dataset of PML. The candidate label set of the instance includes ground-truth labels (marked with black color) and noisy labels (marked with red color).

Inference) to address the MLL problem, which infers the label distributions from the logical labels. GLLE (Graph Laplacian Label Enhancement) [14] utilizes the topological information of the feature space and the correlation among the labels to address the MLL problem.

In the PLL task, each training instance has been annotated with a set of candidate labels, where there is only one ground-truth label [15–18]. PLL mainly focuses on disambiguation to discover the ground-truth label. In [19], the EM+Prior Model is proposed to distinguish the ground-truth label from the candidate labels. In [20], a dictionary-based learning method learns intermediate dictionaries by updating the confidences and dictionaries iteratively, which are then used to distinguish the ground-truth label. M3PL (MaxiMum Margin Partial Label learning) [21] distinguishes the ground-truth label and maximizes the margin iteratively with an alternating optimization process. PL-LEAF (Partial Label LEArning via Feature-aware disambiguation) [22] utilizes the feature manifold structure of instances to learn the label confidences. VALEN (VAriational Label Enhancement for instance-dependent partial label learning) [23] iteratively recovers the latent label distribution for each instance and trains the predictive model by leveraging the recovered label distribution to address the PLL problem. However, these PLL methods based on disambiguation are seriously affected by noisy labels [24]. The larger the candidate label set is, the more serious the negative effect of noisy labels.

In the PML task, each training instance is annotated with a set of candidate labels, where there are some ground-truth labels. Because the number of ground-truth labels is unknown, PML is more challenging than MLL and PLL. A simple and lazy strategy of PML is to take all the candidate labels as ground-truth labels directly and utilizes conventional MLL algorithms such as RankSVM [9], ML-KNN [25], and LIFT [26] to train models. However, such a strategy does not consider the negative effect of noisy labels from the set of candidate labels, which leads to performance degeneration. To address this problem, some algorithms have been designed specifically for the PML task. The label confidence is widely used in PML methods to assess for each candidate label, the probability of it being a ground-truth label. PARTICLE (PARTIal multi-label learning via Credible Label Elicitation) [27] identifies credible labels with high confidence by adapting label propagation based on a weighted graph. PML-*lc* (Partial Multi-Label Learning with Label Correlations) [4] learns the confidence matrix with label correlations. PML-*fp* (Partial Multi-Label Learning with Feature Prototypes) [4] learns the confidence matrix with a feature prototype. Recently, low-rank and sparse decomposition have been applied to PML. PML-LRS (Partial Multi-label Learning by Low-Rank and Sparse decomposition) [28] utilizes l_1-norm regularization to constrain the irrelevant labels to be sparse and trace norm regularization to constrain the ground-truth labels to be low-rank. PML-NI (Partial Multi-label Learning with Noisy label Identification) [29] defines two mapping matrices as the multi-label classifier and the noisy label identifier. The multi-label classifier is constrained to be low-rank, and the noisy label identifier is constrained to be sparse.

Label correlations are beneficial to MLL. However, in PML, due to noisy labels, real label correlations are corrupted. Therefore, most of the existing PML methods ignore the label correlations and mainly focus on label disambiguation to learn a ground-truth label matrix directly. PML-*lc* [4] employs label correlations, but it directly employs the damaged co-occurrence rate of two labels as their correlation, which may degrade the performance of PML. In addition, most of the multi-label learning methods consider the label correlations to be symmetric. However, this is not true in most cases. For example, the label "bee" often occurs with the label "flower", while the label "flower" does not often occur with the label "bee". Learning the real label correlations from the candidate labels with noisy labels is very challenging. Based on the above considerations, we make full use of the structural information of the label space and the feature space and present a Partial Multi-Label Learning based on Sparse Asymmetric Label Correlation (PML-SALC). To the best of our knowledge, we are the first to integrate sparse asymmetric label correlation learning and multi-label classifier learning into a unified framework in PML. PML-SALC learns the asymmetric label correlations from the concurrence corrupted by noisy labels with sparse, global and local consistency constraints and obtains label confidence with the learned label correlations to realize label disambiguation. Different from previous PML methods, label correlation learning and label disambiguation are performed simultaneously in PML-SALC. The main contributions of our work are as follows:

- PML-SALC integrates sparse asymmetric label correlation learning and multi-label classifier learning into a unified framework, which makes full use of the advantage of each learning submodule and benefits to each other with the joint optimization.
- PML-SALC combines global asymmetric label structure information and local feature structure information to capture the real label correlations. At the same time, the label correlations are constrained to be sparse, which further reduces the negative effect of the noisy labels.
- PML-SALC models the relationship between the feature and the real-value label confidence instead of the relationship between the feature and the 0/1 label. This transfers the classification problem to a regression problem and makes the model smoother and more robust.
- Extensive experiments on various datasets validate the effectiveness of the proposed PML-SALC.

The remainder of this paper is organized as follows. In Section 2, we review the related work. In Section 3, the proposed PML-SALC is described in detail. In Section 4, the experimental results are analyzed and discussed. In Section 5, the conclusion is drawn.

2. Related work

How to utilize weakly supervised information to explore instance-label correlations is a very important problem in PML. Correlation mining is widely applied in many cross-modal tasks, such as social image understanding [30–32], social image retrieval [33], facial expression recognition [34] and multi-view multi-label learning [35–38]. [32] incorporates the weakly supervised information, semantic structure and visual structure from heterogeneous data resources in a deep collaborative embedding framework (DCE) to learn a common latent space as a bridge between instance and label spaces. WDMF (weakly supervised deep matrix factorization) [31] collaboratively explores weakly supervised labels and visual and semantic structures to uncover the latent visual and label representations embedded in the latent subspace. RSSL (robust structure subspace learning) [30] exploits geometric structure and the local and global structure consistencies to learn a latent discriminative representation of images. SGH (Semantic Guided Hashing) [33] jointly exploits hash code learning, semantic information mining and data structure discovery for social image retrieval. EDL-LRL (Emotion Distribution Learning method by exploiting Low-Rank label correlations Locally) [34] utilizes label distribution learning for facial expression recognition and considers label correlations by determining the Pearson's correlation coefficients. CDMM (consistency and diversity neural network multi-view multi-label learning) [35] considers the diversity and consistency of views and integrates the correlation of labels and different contributing factors of views into the classification and prediction models. GLMVML-IVL (global and local multi-view multi-label learning with incomplete views and labels) [36] combines global and local label correlation and complementary information of different views and learns a pseudoclass label matrix to address the problem of incomplete labels. MVRL (multi-view representation learning) [37] simultaneously learns a set of view weights to identify the quality of each view via a matrix factorization model. [38] proposes a neuron-wise correlation-maximizing regularizer for deep multi-view representation learning.

In recent years, PML has attracted increasing attention and some advances have been achieved. Based on the methods of learning, PML methods can be roughly divided into two categories [39]: two-stage learning-based PML and end-to-end learning-based PML.

Two-stage leaning-based PML, as the name suggests, consists of two stages of learning. In the first stage, the label confidences are learned. The larger the label confidence is, the more likely the label is a ground-truth label. In the second stage, the label confidences are utilized to learn a classifier using conventional MLL. For instance, PARTICLE [27] contains credible label elicitation and predictive model induction. In the stage of credible label elicitation, PARTICLE obtains the label confidence by adapting the label propagation based on a weighted graph and takes the labels with high confidence as the credible labels. In the stage of predictive model induction, the predictive model is learned with credible labels. DRAMA (DiscRiminative and correlAtive partial Multi label leArning) [40] consists of confidence learning and predictor learning. In the first stage, DRAMA employs a feature manifold to learn a label confidence matrix. In the second stage, it makes use of the learned confidence matrix to train a predictor using a gradient boosting algorithm. In each boosting round, the initial feature space is augmented with the previously learned labels for extracting the label correlations. The performance of two-stage learning-based PML largely depends on the confidence learning stage. If the label confidences cannot be learned correctly, the model performance will significantly degenerate.

On the other hand, end-to-end learning-based PML integrates label disambiguation and model induction into a single stage. For instance, PML-*lc* [4] and PML-*fp* [4] combine label ranking and confidence matrix to learn multi-label classification models. In the optimization step, the label confidence matrix and classifier are optimized alternately. PML-LRS [28] decomposes the label matrix into a ground-truth label matrix and a noisy label matrix. It then incorporates these two matrices into the objective function to learn a multi-label classifier with low-rank and sparse constraints. PML-NI [29] defines two mapping matrices as the multi-label classifier and the noisy label identifier. The multi-label classifier and the noisy label identifier are learned simultaneously with alternating optimization steps. The previous sparse methods ignore the correlation between labels in the multi-label learning task and this will affect the classification performance. The sparse method used in this paper is different from the previous methods. From the perspective of label correlation, we consider the correlation between labels to be sparse, which obtains the confidence of labels and improves the performance of multi-label classification. MUSER (partial multi-label learning via MUlti-SubspacE Representation) [41] assumes that noisy labels are partly caused by noisy features. Accordingly, it introduces a latent label subspace and feature subspace to reduce the negative impact of noisy information. fPML (feature-induced partial multi-label learning) [42] decomposes the label matrix and the feature matrix into low-rank matrices. It simultaneously distinguishes noisy labels and learns multi-label classifier. NATAL (Noisy lAbel Tolerated pArtial multi-label Learning) [43] treats the noisy labels as redundant labels and assumes that the redundant labels originated from the absence of feature information. Therefore, the PML is converted to a feature completion problem and the multi-label classification model with the completed features and the candidate labels is directly trained. PML-MT (Partial multi-label Learning with Mutual Teaching) [44] refines the label confidence matrix iteratively with a couple of self-ensemble teacher works and trains two prediction networks simultaneously. End-to-end learning-based PML methods fuse label disambiguation and model induction with iterative optimization, which is simple and direct. However, during label disambiguation, most of the existing end-to-end learning-based PML methods ignore the facts that the real label correlations have been corrupted by the noisy labels and the label correlations are usually asymmetric. This compromises computation of the real label correlations and the performance of multi-label classifier.

3. Our approach

To better learn the real label correlations from the candidate labels with noisy labels and improve the performance of partial multi-label learning, we propose partial multi-label learning based on sparse asymmetric label correlations (PML-SALC). PML-SALC is an end-to-end learning-based PML method. Different from the existing PML methods, PML-SALC takes advantage of the global asymmetric label structure information and local feature structure information to capture the real label correlations. Due to the fact that each label is usually associated with a few other labels, we constrain the label correlation matrix to be sparse, which can also alleviate the negative effect of noisy labels. Moreover, PML-SALC maps a feature to a real-value label confidence instead of assigning a 0/1 label. This transfers the classification problem to a regression problem and makes the model smoother and more robust. In this section, we will present the details of PML-SALC.

3.1. Notation

First, we give the notations used in this paper. Let $\mathcal{X} \in \mathbb{R}^d$ denote the d-dimensional feature space, and $\mathcal{Y} \in \{0, 1\}^q$ denote the q-dimensional label space. $\mathbf{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_i, \ldots, \boldsymbol{x}_n] \in \mathbb{R}^{d\times n}$ is the feature matrix of the training instances, where n is the number of instances and $\boldsymbol{x}_i \in \mathcal{X}$ is the feature vector of the ith instance. $\mathbf{Y} = [\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_i, \ldots, \boldsymbol{y}_n] \in \{0, 1\}^{q\times n}$ is the corresponding candidate label matrix of the training instances, where $\boldsymbol{y}_i = \left(y_{1i}, y_{2i}, \ldots, y_{ji}, \ldots, y_{qi}\right)^T \in \mathcal{Y}$ is a q-dimensional candidate label vector of the ith instance, q is the number of labels, and y_{ji}=1 means that the jth label is in the candidate label set of the ith instance; otherwise, $y_{ji} = 0$.

3.2. Formulation

The goal of PML is to learn a classifier or predictor from the training instances annotated with the candidate labels. Label disambiguation is very challenging. Instead of taking the co-occurrence of two labels as their correlations directly, we learn real asymmetric label correlations with global consistency, local consistency and sparse constraints to achieve label disambiguation. In this paper, we define $\tilde{\mathbf{Y}} = [\tilde{\boldsymbol{y}}_1, \tilde{\boldsymbol{y}}_2, \ldots, \tilde{\boldsymbol{y}}_i, \ldots, \tilde{\boldsymbol{y}}_n] \in \mathbb{R}^{q\times n}$ as the unknown label confidence matrix, where $\boldsymbol{y}_i = \left(\tilde{y}_{1i}, \tilde{y}_{2i}, \ldots, \tilde{y}_{ji}, \ldots, \tilde{y}_{qi}\right)^T$ is a q-dimensional label confidence vector of the ith instance. $\tilde{y}_{ji}$ is the real-value confidence of the jth label of the ith instance. The larger $\tilde{y}_{ji}$ is, the more likely the jth label is the ground-truth label of the ith instance. It is well known that there are correlations among labels. The more relevant the two labels are, the more likely they occur simultaneously. The capture of the correlations among labels will benefit partial multi-label learning. Therefore, we introduce the unknown label correlation matrix $\mathbf{Q} \in \mathbb{R}^{q\times q}$, where $\mathbf{Q}_{ij}$ represents the correlation between the ith label and the jth label. Then, we can obtain Eq. (1).

$$\tilde{\mathbf{Y}} = \mathbf{QY} \tag{1}$$

The ith row of $\mathbf{Q}$ represents the correlation between the ith label and other labels. The jth column of $\mathbf{Y}$ represents all candidate labels of the jth instance. From Eq. (1), it can be seen that $\tilde{\mathbf{Y}}_{ij}$ is the dot product of the ith row of $\mathbf{Q}$ and the jth column of $\mathbf{Y}$. The greater the sum of correlations between the ith label and all candidate labels of the jth instance, the greater$\tilde{\mathbf{Y}}_{ij}$, which means that the ith label has greater confidence to be the ground-truth label of the jth instance. Since the noisy label is usually less correlated with the other labels of the instance, the confidence of the noisy label of the instance computed using Eq. (1) will be low, which will reduce the negative impact of the noisy label and to some extent achieve label disambiguation.

Because some noisy labels are concealed in the candidate label matrix of the training instances, we introduce a linear mapping matrix $\mathbf{W} \in \mathbb{R}^{q\times d}$ to learn a classifier, which maps the feature to the label confidence instead of the label. Thus, the objective function is as follows.

$$\min_{\mathbf{W},\mathbf{Q}} \; L(\mathbf{X}, \mathbf{Y}; \mathbf{W}, \mathbf{Q}) + \alpha R(\mathbf{W}) \tag{2}$$

where the first term is the loss function of the classifier, $R(\mathbf{W})$ is the regularization term to control the complexity of the model, and α is a balance parameter. The introduction of the label confidence matrix can reduce the negative impact of noisy labels concealed in the candidate label matrix. In addition, it transfers a classification problem to a regression problem by mapping the instance feature to a real-value label confidence instead of a 0/1-value label, which can make the model smoother. For simplicity, we choose the quadratic loss function and utilize the Frobenius norm to control the model complexity. Then, Eq. (2) can be rewritten as follows.

$$\min_{\mathbf{W},\mathbf{Q}} \frac{1}{2}\|\mathbf{WX} - \mathbf{QY}\|_F^2 + \frac{\alpha}{2}\|\mathbf{W}\|_F^2 \tag{3}$$

where $\|\cdot\|_F$ represents the Frobenius norm.

Moreover, because of the correlation among labels [29], we constrain the mapping matrix $\mathbf{W}$ to be low-rank. Thus, Eq. (3) can be rewritten as follows.

$$\min_{\mathbf{W},\mathbf{Q}} \frac{1}{2}\|\mathbf{WX} - \mathbf{QY}\|_F^2 + \varphi rank(\mathbf{W}) + \frac{\alpha}{2}\|\mathbf{W}\|_F^2 \tag{4}$$

where $rank(\cdot)$ represents the rank of the matrix and φ is a trade-off parameter.

Due to the noisy labels concealed in the candidate labels, learning the real label correlation matrix $\mathbf{Q}$ in PML is very challenging. Most of the existing methods assume that the greater the co-occurrence of labels, the more relevant the labels are. Therefore, the label correlations are extracted using the label co-occurrence rate, and the obtained label correlation matrix is symmetric. However, in reality, label correlations are usually asymmetric; that is, the relation of label p to label q is not equal to the relation of label q to label p. In this paper, we introduce a conditional probability matrix $\mathbf{P} \in \mathbb{R}^{q\times q}$ and move $\mathbf{Q}$ close to $\mathbf{P}$ to capture the global asymmetric structure information of the label space. Thus, Eq. (4) can be rewritten as follows.

$$\min_{\mathbf{W},\mathbf{Q}} \frac{1}{2}\|\mathbf{WX} - \mathbf{QY}\|_F^2 + \frac{\beta}{2}\|\mathbf{Q} - \mathbf{P}\|_F^2 + \varphi rank(\mathbf{W}) + \frac{\alpha}{2}\|\mathbf{W}\|_F^2 \tag{5}$$

where the second term is the global consistency constraint, β is a trade-off parameter, and the element p_{ij} in $\mathbf{P}$ is the conditional probability of the ith label occurrence given that the jth label occurs, which can be computed as follows.

$$p_{ij} = \frac{\frac{N(l_j,l_i)}{n}}{\frac{N(l_j)}{n}} = \frac{N\left(l_j, l_i\right)}{N\left(l_j\right)} \tag{6}$$

where $N(l_j, l_i)$ represents the concurrence number of the ith label l_i and the jth label l_j. $N(l_j)$ represents the occurrence number of the jth label l_j. It can be seen from Eq. (6) that the conditional probability matrix is asymmetric. Due to the randomness of the concurrence of noisy labels and other ground-truth labels, the corresponding conditional probabilities of noisy labels are very low. Constraining $\mathbf{Q}$ to be close to $\mathbf{P}$ will make the correlation between noisy labels and other ground-truth labels in the learned $\mathbf{Q}$ as small as possible, which will mitigate the negative impact of the noisy label.

Furthermore, it is easily observed that similar instances often have similar labels [45]. Therefore, we combine the local manifold structure information of the feature space with the global asymmetric structure information of the label space to learn the correlation matrix. In addition, we constrain $\mathbf{Q}$ to be sparse with l_0 norm, which further alleviates the negative impact of the noisy labels on label correlation extraction. Since the proportion of label pairs with correlation in all label pairs is relatively small, the sparseness constraint will make the original low correlation between noisy labels and ground-truth labels tend to 0. In this paper, we propose a sparse approach based on label correlation, which utilizes the correlation between labels to guide classification more effectively in the partial multi-label learning task. Thus, the objective function can be rewritten as follows.

$$\begin{aligned}\min_{\mathbf{W},\mathbf{Q}} & \frac{1}{2}\|\mathbf{WX} - \mathbf{QY}\|_F^2 + \frac{\beta}{2}\|\mathbf{Q} - \mathbf{P}\|_F^2 + \\ & \frac{\sigma}{2n}\sum_{i=1}^{n}\left\|\mathbf{Q}\boldsymbol{y}_i - \frac{1}{|nei(\boldsymbol{x}_i)|}\sum_{\boldsymbol{y}_j \in nei(\boldsymbol{x}_i)}\mathbf{Q}\boldsymbol{y}_j\right\|_2^2 + \\ & \gamma\|\mathbf{Q}\|_0 + \varphi rank(\mathbf{W}) + \frac{\alpha}{2}\|\mathbf{W}\|_F^2\end{aligned} \tag{7}$$

where the third term is the local consistency constraint, the fourth term is the sparse constraint, $\|\cdot\|_0$ represents the l_0 norm, σ, γ are trade-off parameters, and $nei(x_i)$ is the label set of the k-nearest neighbors of x_i.

Since solving Eq. (7) is an NP hard problem, we approximate the l_0 norm by the l_1 norm and approximate the solution of the low-rank problem using the trace norm. We then obtain the final objective function as follows.

$$\min_{\mathbf{W},\mathbf{Q}} \frac{1}{2}\|\mathbf{WX}-\mathbf{QY}\|_F^2+\frac{\beta}{2}\|\mathbf{Q}-\mathbf{P}\|_F^2+ \frac{\sigma}{2n}\sum_{i=1}^{n}\left\|\mathbf{Q}\boldsymbol{y}_i-\frac{1}{|nei(\boldsymbol{x}_i)|}\sum_{\boldsymbol{y}_j\in nei(\boldsymbol{x}_i)}\mathbf{Q}\boldsymbol{y}_j\right\|_2^2+ \gamma\|\mathbf{Q}\|_1+\varphi\|\mathbf{W}\|_{tr}+\frac{\alpha}{2}\|\mathbf{W}\|_F^2 \tag{8}$$

where $\|\cdot\|_1$ represents the l_1 norm and $\|\cdot\|_{tr}$ represents the trace norm.

3.3. Optimization

We adopt alternating optimization to obtain the optimal solution. The values of **W**, **Q** are updated individually while keeping the other variable fixed. There are two steps:

Step1: Fix **Q**, and update **W**.

Eq. (8) can be transformed into the following optimization problem:

$$\min_{\mathbf{W}} \frac{1}{2}\|\mathbf{WX}-\mathbf{QY}\|_F^2+\varphi\|\mathbf{W}\|_{tr}+\frac{\alpha}{2}\|\mathbf{W}\|_F^2 \tag{9}$$

We utilize the accelerated proximal gradient descend [46] to minimize the trace norm. The main steps are as follows:

1. Let $f(\mathbf{W})=\frac{1}{2}\|\mathbf{WX}-\mathbf{QY}\|_F^2+\frac{\alpha}{2}\|\mathbf{W}\|_F^2$, $\nabla f(\mathbf{Z})=(\mathbf{ZX}-\mathbf{QY})\mathbf{X}^T+\alpha\mathbf{Z}$, $g(\mathbf{W},\mathbf{Z})=f(\mathbf{Z})+\langle\nabla f(\mathbf{Z}),\mathbf{W}-\mathbf{Z}\rangle+\varphi\|\mathbf{W}\|_{tr}$, $\theta_0=\theta_{-1}\in(0,1]$, $L>1$, $\mathbf{W}_0=\mathbf{W}_{-1}$, $\rho>1$.

2. Do the following iteration until convergence.

In the tth iteration, there are the following three steps.

(1) Let $\mathbf{Z}_t=\mathbf{W}_t+\theta_t\left(\theta_{t-1}^{-1}-1\right)(\mathbf{W}_t-\mathbf{W}_{t-1})$, we can obtain the $\mathbf{W}_{t+1}$ by using Singular Value Thresholding (SVT) [47] to optimize the following Eq. (10):

$$\mathbf{W}_{t+1}=\min_{\mathbf{W}}\{g(\mathbf{W},\mathbf{Z}_t)+\frac{L}{2}\|\mathbf{W}-\mathbf{Z}_t\|_F^2\} \tag{10}$$

(2) When $f\left(\mathbf{W}_{t+1}+\varphi\|\mathbf{W}_{t+1}\|_{tr}\right)$ is larger than $g(\mathbf{W}_{t+1},\ \mathbf{Z}_t)+\frac{L}{2}\|\mathbf{W}_{t+1}-\mathbf{Z}_k\|_F^2$, the following iteration is performed.

① Update $L=\rho L$

② Update the $\mathbf{W}_{t+1}$ by using Singular Value Thresholding (SVT) [47] to optimize Eq. (10)

(3) Let $\theta_{t+1}=\left(\sqrt{\theta_t^4+4\theta_t^2}-\theta_t^2\right)/2$, updating $t=t+1$

Step2: Fix **W**, and update **Q**.

Eq. (8) can be transformed into the following optimization problem:

$$\min_{\mathbf{Q}} \frac{1}{2}\|\mathbf{WX}-\mathbf{QY}\|_F^2+\frac{\beta}{2}\|\mathbf{Q}-\mathbf{P}\|_F^2+ \frac{\sigma}{2n}\sum_{i=1}^{n}\left\|\mathbf{Q}\boldsymbol{y}_i-\frac{1}{|nei(\boldsymbol{x}_i)|}\sum_{\boldsymbol{y}_j\in nei(\boldsymbol{x}_i)}\mathbf{Q}\boldsymbol{y}_j\right\|_2^2+\gamma\|\mathbf{Q}\|_1 \tag{11}$$

We utilize the shrinkage operator [48] to solve the above optimization problem, and the main process is as follows.

First, let

$$J(\mathbf{Q})=\frac{1}{2}\|\mathbf{WX}-\mathbf{QY}\|_F^2+\frac{\beta}{2}\|\mathbf{Q}-\mathbf{P}\|_F^2+ \frac{\sigma}{2n}\sum_{i=1}^{n}\left\|\mathbf{Q}\boldsymbol{y}_i-\frac{1}{|nei(\boldsymbol{x}_i)|}\sum_{\boldsymbol{y}_j\in nei(\boldsymbol{x}_i)}\mathbf{Q}\boldsymbol{y}_j\right\|_2^2$$

$$\mathbf{H}=\mathbf{Q}-\frac{1}{L_p}\nabla J(\mathbf{Q}) \tag{12}$$

where L_p is the Lipschitz constant and can be obtained by Eq. (13).

$$L_p=\left\|\mathbf{Y}\mathbf{Y}^T+\beta\mathbf{E}+\frac{\sigma}{n}\sum_{i=1}^{n}\left(\left(\boldsymbol{y}_i-\frac{1}{|nei(\boldsymbol{x}_i)|}\sum_{\boldsymbol{y}_j\in nei(\boldsymbol{x}_i)}\boldsymbol{y}_j\right)\left(\boldsymbol{y}_i-\frac{1}{|nei(\boldsymbol{x}_i)|}\sum_{\boldsymbol{y}_j\in nei(\boldsymbol{x}_i)}\boldsymbol{y}_j\right)^T\right)\right\|_F \tag{13}$$

where **E** is an identity matrix.

Then, we obtain the analytical solution of $\mathbf{Q}^*$ with Eq. (14) [49].

$$Q_{ki}^*=\begin{cases} H_{ki}-\frac{\gamma}{L_p} & if\ H_{ki}>\frac{\gamma}{L_p} \\ 0 & if\ |H_{ki}|\le\frac{\gamma}{L_p} \\ H_{ki}+\frac{\gamma}{L_p} & if\ H_{ki}<-\frac{\gamma}{L_p}\end{cases}\quad(1\le k,i\le q) \tag{14}$$

3.4. Algorithm

The complete algorithm is summarized in Algorithm 1.

Algorithm 1 PML-SALC

Input: Feature matrix **X**, labels matrix **Y**, trade-off parameters $\alpha,\beta,\gamma,\sigma,\varphi$
Output: **W**, **Q**

```
construct the k-nn graph and initialize.
repeat
    for t=0 to T do do
        Update W according to Eq. (10) by using SVT algorithm;
        while f(W_{t+1} + φ||W_{t+1}||_tr) > g(W_{t+1},
            Z_t) + L/2 ||W_{t+1} − Z_k||_F^2 do
            L = ρL;
            Update W according to Eq. (10) by using SVT algorithm;
        end while
        θ_{t+1} = (sqrt(θ_t^4 + 4θ) − θ_t^2)/2;
        t = t + 1
    end for
    Compute L_p according to Eq. (13);
    Compute H according to Eq. (12);
    Update Q according to Eq. (14);
until convergence
return W, Q
```

3.5. Multi-label classification

After the model has been trained using Algorithm 1 described in Section 3.4, we obtain the outputs of Algorithm 1, namely, **W** and **Q**. Now, we can realize multi-label classification. First, the new instance $\hat{\boldsymbol{x}}$ is mapped to the label confidence space using **W**. Thus, we obtain the label confidence vector $\mathbf{W}\hat{\boldsymbol{x}}=(lc_1,lc_2,\ldots,lc_q)^T$, where lc_i is the label confidence of the ith label. Then, if lc_i is larger than the threshold t, $\hat{\boldsymbol{x}}$ is annotated with the ith label. In this paper, we choose the threshold t according to Ref. [9]. The specific process is as follows. First, for each instance

Table 1
The details of six datasets used in the experiments.

Datasets	#Instance	Dim	#Class labels	Cardinality	Field
Genbase	662	1186	27	1.252	Biology
CAL500	502	68	174	26.044	Music
Bibtex	7395	1836	159	2.402	Text
Medical	978	1449	45	1.245	Text
Birds	645	260	19	1.014	Audio
Emotions	593	72	6	1.869	Music

$\boldsymbol{x}_i$ in the training dataset, a threshold t_i is chosen by the following equation.

$$t_i = \arg\min_t |\{j \in L \,\text{s.t.}\, w_j \boldsymbol{x}_i \le t\}| + |\{j \in \bar{L} \,\text{s.t.}\, w_j \boldsymbol{x}_i \ge t\}|$$

where L is the candidate label set of $\boldsymbol{x}_i$ and $\bar{L}$ is the complementary set of L. If the minimum is a numerical interval, the middle of this interval is taken as the threshold t_i. Thus, we obtain a dataset $\Omega = \{(\boldsymbol{x}_i, t_i)\}_{i=1}^n$.

Second, we learn a mapping function $\Gamma : \mathbf{X} \to \mathbb{R}$ from the dataset Ω.

Finally, for the new instance $\hat{\boldsymbol{x}}$, we choose $\Gamma(\hat{\boldsymbol{x}})$ as the threshold.

3.6. Computational complexity analysis

In this section, we discuss the computational cost of the proposed algorithm and utilize big O notation to represent the complexity of the algorithm. As depicted in Algorithm 1, the k-nn graph is first constructed, and the corresponding time complexity is $O(dn^2)$. Then, the proposed optimization problem is solved iteratively. For each iteration, the complexity of SVT is $(T_0(2d^2q + d(2nq+q^2+2q)+n(q^2+q)+q^3))$, where T_0 is the iteration number of SVT. The complexity of computing L_p is $O(2n(q^2+q)+3q^2)$. The complexity of computing $\mathbf{H}$ is $O(dnq+n(q^3+3q^2+3q)+3q^2)$. The complexity of updating $\mathbf{Q}$ is $O(q^2)$. Therefore, the total computational complexity is $O(T_1((dnq+n(q^3+6q^2+6q)+q^3+7q^2)+T_0(2d^2q+d(2nq+q^2+2q)+n(q^2+q)+q^3))+dn^2)$, where T_1 is the iteration number of our algorithm.

4. Experiments

4.1. Datasets and settings

We perform extensive experiments to evaluate the proposed PML-SALC. We synthesize the PML datasets from six widely-adopted MLL datasets[1]: Genbase [50], CAL500 [51], Bibtex [52], Medical [53], Birds [54] and Emotions [55]. These datasets cover different fields: biology, music, text and audio. Table 1 shows the details of the six datasets, where #Instance is the number of instances in the dataset, Dim is the dimension of the instance feature, #Class Label is the number of labels in the label space, Cardinality is the average number of labels to each instance and Field refers the field of the dataset. We convert these MLL datasets into PML datasets by adding r random noisy labels to each training instance, where r is the number of noisy labels. For the Genbase, CAL500, Bibtex, Medical, and Birds datasets, we set $r \in \{1, 2, 3\}$. Many instances in the Emotions dataset have 3 labels, while the label set of the Emotions dataset only has 6 labels. If we add 3 noisy labels to each training instance in the Emotions dataset, many instances will include all the labels of the label set, which is not reasonable for PML. Therefore, we set $r \in \{1, 2\}$ for the Emotions dataset.

As in previous works, we use five evaluation metrics to evaluate the performance of the proposed PML-SALC and all baseline methods, which are Ranking Loss, Hamming Loss, One Error, Coverage, and Average Precision [1]. For the average precision metric, the higher the value is, the better the performance. For the other four metrics, the lower the value is, the better the performance. The detailed definitions of the five evaluation metrics can be found in [1].

4.2. Comparison with the state-of-the-art methods

To validate the effectiveness of the proposed PML-SALC, we adopt the following seven state-of-the-art methods as baselines to demonstrate the performance of PML-SALC: ML-KNN [25], LIFT [26], IPAL [56], PML-*fp* [4], fPML [42], PAR-VLS [27], and PAR-MAP [27]. The parameter settings of these baseline methods follow the recommended settings in their original paper. For the proposed PML-SALC, we use the following parameter values: $\alpha = 1$, $\beta = 1000$, $\gamma = 400$, $\sigma = 1000$ and $\varphi = 1$.

For fair evaluation, we transform each dataset into PML datasets five times, and each time we take five-fold cross-validation. The comparison experimental results are displayed in Tables 2, 3, and 4, where each experimental result includes the mean and standard deviation. For each dataset and each evaluation metric, the best result among the proposed PML-SALC and baselines is displayed in boldface.

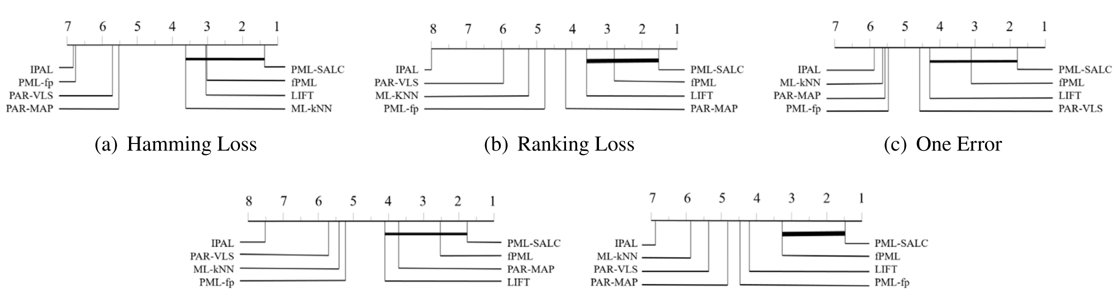
ML-KNN and LIFT are two MLL methods, IPAL is a PLL method, and the other baselines are PML methods. ML-KNN first finds the k-nearest neighbors for each unseen instance. Then, according to the labels of neighbor instances, the label set of the unseen instance is determined using the MAP criterion. LIFT first constructs label-specific features using clustering techniques and then trains the classification models with the constructed label-specific features. However, ML-KNN and LIFT do not consider the correlations between labels and the negative effect of noisy labels. IPAL identifies the ground-truth label and classifies the unseen instance based on minimum error reconstruction from its nearest neighbors. However, it also does not consider the correlations among labels. PML-*fp* combines label ranking and label confidence to learn a multi-label classification model and optimizes the label confidence using a feature prototype. fPML utilizes correlations between labels and features to distinguish the noisy labels and incorporates the noisy label identification into the multi-label classifier training. PAR-VLS extracts credible labels and employs pairwise label ranking to induce multi-label classification model with the help of virtual label splitting. PAR-MAP extracts credible labels and employs pairwise label ranking to induce multi-label classification model using MAP reasoning. However, the above PML methods ignore the fact that the real label correlations have been corrupted by the noisy labels hidden in the candidate label set. The proposed PML-SALC utilizes the global asymmetric label structure information and local feature structure information to capture the real label correlations and constrains the label correlation matrix to be sparse to alleviate the negative effect of noisy labels. From Tables 2–4, out of 85 (5 datasets × 3 configurations + 1 dataset × 2 configurations in 5 evaluation metrics) comparison cases, the following can be observed: (1) The proposed PML-SALC achieves the best performance in 70.6% of the cases. (2) The proposed PML-SALC ranks second in 12.9% of the cases. Compared with other methods, our method achieves the best performance in most cases.

Furthermore, we employ the Friedman test to analyze the relative performance among the proposed PML-SALC and baseline methods. The Friedman test [57] is an effective statistical test for comparing more than two methods on multiple datasets. Let

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

Table 2
Experimental results (mean ± standard deviation) of each comparison method on six synthesized PML datasets, where 1 noisy label is added to each instance.

Datasets	ML-KNN	LIFT	IPAL	PML-*fp*	fPML	PAR-VLS	PAR-MAP	PML-SALC
Ranking Loss (the smaller, the better)								
Genbase	0.008 ± 0.006	0.009 ± 0.005	0.031 ± 0.013	0.017 ± 0.006	0.010 ± 0.005	0.025 ± 0.010	0.014 ± 0.006	**0.005 ± 0.003**
CAL500	0.185 ± 0.006	0.184 ± 0.006	0.831 ± 0.059	0.182 ± 0.009	0.181 ± 0.005	0.538 ± 0.01	**0.176 ± 0.005**	0.178 ± 0.008
Bibtex	0.225 ± 0.005	0.110 ± 0.005	0.716 ± 0.053	0.108 ± 0.007	**0.105 ± 0.009**	0.325 ± 0.010	0.320 ± 0.007	0.107 ± 0.008
Medical	0.061 ± 0.012	0.036 ± 0.008	0.247 ± 0.080	0.098 ± 0.019	0.050 ± 0.012	0.106 ± 0.018	0.080 ± 0.013	**0.025 ± 0.006**
Birds	0.336 ± 0.025	0.263 ± 0.029	0.634 ± 0.087	0.426 ± 0.040	0.226 ± 0.023	0.319 ± 0.019	0.316 ± 0.022	**0.171 ± 0.029**
Emotions	0.312 ± 0.026	0.263 ± 0.027	0.417 ± 0.116	0.386 ± 0.021	0.206 ± 0.014	0.242 ± 0.016	0.230 ± 0.051	**0.193 ± 0.020**
Hamming Loss (the smaller, the better)								
Genbase	0.005 ± 0.003	0.005 ± 0.004	0.013 ± 0.013	0.032 ± 0.000	0.004 ± 0.004	0.012 ± 0.006	0.006 ± 0.003	**0.003 ± 0.003**
CAL500	0.139 ± 0.006	0.139 ± 0.007	0.163 ± 0.005	0.162 ± 0.005	0.140 ± 0.008	0.150 ± 0.006	0.145 ± 0.006	**0.138 ± 0.007**
Bibtex	0.015 ± 0.000	**0.012 ± 0.000**	0.028 ± 0.000	0.019 ± 0.000	0.013 ± 0.000	0.016 ± 0.000	0.022 ± 0.000	0.013 ± 0.000
Medical	0.016 ± 0.002	0.012 ± 0.002	0.049 ± 0.005	0.035 ± 0.003	0.050 ± 0.011	0.021 ± 0.001	0.029 ± 0.003	**0.011 ± 0.003**
Birds	0.054 ± 0.006	0.058 ± 0.010	0.081 ± 0.006	0.125 ± 0.021	0.054 ± 0.006	0.144 ± 0.018	0.120 ± 0.006	**0.051 ± 0.008**
Emotions	0.335 ± 0.024	0.310 ± 0.024	0.309 ± 0.028	0.376 ± 0.023	0.224 ± 0.011	0.228 ± 0.018	0.268 ± 0.021	**0.223 ± 0.015**
One Error (the smaller, the better)								
Genbase	0.011 ± 0.011	**0.002 ± 0.003**	0.013 ± 0.011	0.005 ± 0.004	**0.002 ± 0.003**	0.005 ± 0.005	0.011 ± 0.010	**0.002 ± 0.003**
CAL500	0.116 ± 0.019	0.124 ± 0.024	0.303 ± 0.139	0.130 ± 0.008	0.116 ± 0.019	0.324 ± 0.051	**0.116 ± 0.018**	0.119 ± 0.022
Bibtex	0.624 ± 0.012	0.401 ± 0.017	0.625 ± 0.014	0.377 ± 0.006	0.453 ± 0.017	0.588 ± 0.012	0.743 ± 0.017	**0.359 ± 0.012**
Medical	0.276 ± 0.035	0.174 ± 0.039	0.301 ± 0.049	0.345 ± 0.043	0.194 ± 0.038	0.252 ± 0.051	0.441 ± 0.073	**0.164 ± 0.041**
Birds	0.738 ± 0.036	0.777 ± 0.081	0.473 ± 0.085	0.859 ± 0.058	0.669 ± 0.033	0.618 ± 0.047	0.689 ± 0.036	**0.401 ± 0.046**
Emotions	0.419 ± 0.047	0.386 ± 0.048	0.336 ± 0.057	0.469 ± 0.042	0.349 ± 0.040	**0.288 ± 0.048**	0.355 ± 0.064	0.326 ± 0.052
Coverage (the smaller, the better)								
Genbase	0.024 ± 0.014	0.027 ± 0.013	0.043 ± 0.021	0.043 ± 0.010	0.030 ± 0.014	0.050 ± 0.024	0.028 ± 0.015	**0.019 ± 0.009**
CAL500	0.751 ± 0.016	0.758 ± 0.015	0.978 ± 0.004	0.754 ± 0.015	0.752 ± 0.014	0.966 ± 0.006	**0.701 ± 0.016**	0.742 ± 0.012
Bibtex	0.366 ± 0.007	0.189 ± 0.004	0.521 ± 0.015	0.201 ± 0.013	**0.173 ± 0.005**	0.277 ± 0.005	0.204 ± 0.017	0.196 ± 0.011
Medical	0.083 ± 0.019	0.055 ± 0.011	0.154 ± 0.052	0.122 ± 0.023	0.070 ± 0.017	0.114 ± 0.021	0.102 ± 0.017	**0.040 ± 0.010**
Birds	0.217 ± 0.027	0.213 ± 0.024	0.238 ± 0.037	0.268 ± 0.042	0.149 ± 0.021	0.198 ± 0.027	0.206 ± 0.022	**0.121 ± 0.023**
Emotions	0.434 ± 0.028	0.398 ± 0.037	0.410 ± 0.041	0.486 ± 0.028	0.339 ± 0.024	0.349 ± 0.034	0.357 ± 0.026	**0.325 ± 0.023**
Average Precision (the larger, the better)								
Genbase	0.983 ± 0.010	0.982 ± 0.011	0.895 ± 0.011	0.985 ± 0.008	0.981 ± 0.010	0.959 ± 0.022	0.980 ± 0.011	**0.991 ± 0.004**
CAL500	0.490 ± 0.010	0.497 ± 0.013	0.203 ± 0.024	0.504 ± 0.015	0.506 ± 0.011	0.380 ± 0.019	0.511 ± 0.011	**0.513 ± 0.014**
Bibtex	0.321 ± 0.005	0.528 ± 0.014	0.251 ± 0.027	0.546 ± 0.008	0.487 ± 0.015	0.482 ± 0.007	0.477 ± 0.008	**0.572 ± 0.003**
Medical	0.772 ± 0.026	0.855 ± 0.029	0.649 ± 0.093	0.694 ± 0.041	0.833 ± 0.029	0.733 ± 0.034	0.651 ± 0.051	**0.871 ± 0.033**
Birds	0.371 ± 0.020	0.360 ± 0.044	0.434 ± 0.042	0.273 ± 0.027	0.433 ± 0.029	0.408 ± 0.025	0.402 ± 0.025	**0.629 ± 0.041**
Emotions	0.672 ± 0.022	0.703 ± 0.025	0.696 ± 0.041	0.626 ± 0.020	0.752 ± 0.019	0.758 ± 0.017	0.736 ± 0.030	**0.769 ± 0.029**



(a) Hamming Loss (b) Ranking Loss (c) One Error

(d) Coverage (e) Average Precision

Fig. 2. Comparison of PML-SALC (control method) and baseline methods with the Nemenyi test. Methods not connected with PML-SALC in the CD diagram are considered to have a significantly different performance from PML-SALC. (CD = 2.5465 at 0.05 significance level).

N denote the total number of datasets, k denote the number of algorithms, and r_i denote the rank of the ith algorithm. The Friedman statistic is computed as follows.

$$F_F = \frac{(N-1)\chi_F^2}{N(k-1)-\chi_F^2} \tag{15}$$

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

The null hypothesis for the Friedman test is that there are no distinguishable differences among the performances of different algorithms. The calculated average ranks of each algorithm for the five evaluation metrics are shown in Table 5. The Friedman statistics F_F of the five evaluation metrics and the corresponding critical value are listed in Table 6. For each evaluation metric, the null hypothesis is rejected at the 0.05 significance level. Afterwards, we proceed with a post-hoc Nemenyi test and Bonferroni–Dunn test [57] to illustrate the relative performance among the proposed PML-SALC and baseline methods. If the average ranks of the two methods differ by at least the critical difference (CD), the performance of the two methods is considered to be significantly different. Figs. 2 and 3 show the CD diagrams on each evaluation metric for two post-hoc tests. The average ranks of the PML-SALC and baseline methods are marked along the axis of the CD diagram. The methods not connected with the proposed PML-SALC in the CD diagram are considered to have significantly different performance from PML-SALC. As shown in Figs. 2 and 3, in most cases, our proposed PML-SALC is significantly different from other baseline methods.

Table 3
Experimental results (mean ± standard deviation) of each comparison method on five synthesized PML datasets, where 2 noisy labels are added to each instance.

Datasets	ML-KNN	LIFT	IPAL	PML-*fp*	fPML	PAR-VLS	PAR-MAP	PML-SALC
Ranking Loss (the smaller, the better)								
Genbase	0.013 ± 0.009	0.010 ± 0.005	0.051 ± 0.016	0.008 ± 0.003	0.010 ± 0.005	0.024 ± 0.007	0.018 ± 0.005	**0.005 ± 0.002**
CAL500	0.186 ± 0.005	0.184 ± 0.006	0.836 ± 0.057	0.184 ± 0.008	0.181 ± 0.005	0.553 ± 0.022	**0.170 ± 0.005**	0.178 ± 0.014
Bibtex	0.230 ± 0.004	0.120 ± 0.008	0.723 ± 0.039	0.108 ± 0.004	**0.106 ± 0.012**	0.320 ± 0.008	0.326 ± 0.007	0.116 ± 0.006
Medical	0.074 ± 0.011	0.043 ± 0.010	0.263 ± 0.106	0.052 ± 0.019	0.050 ± 0.013	0.103 ± 0.016	0.085 ± 0.014	**0.027 ± 0.008**
Birds	0.352 ± 0.024	0.329 ± 0.045	0.670 ± 0.275	0.427 ± 0.043	0.277 ± 0.024	0.320 ± 0.018	0.312 ± 0.022	**0.184 ± 0.035**
Emotions	0.358 ± 0.021	0.345 ± 0.037	0.461 ± 0.115	0.443 ± 0.022	0.234 ± 0.020	0.263 ± 0.019	0.262 ± 0.057	**0.223 ± 0.015**
Hamming Loss (the smaller, the better)								
Genbase	0.006 ± 0.003	0.004 ± 0.004	0.052 ± 0.013	0.037 ± 0.000	0.004 ± 0.004	0.019 ± 0.006	0.011 ± 0.004	**0.003 ± 0.002**
CAL500	0.139 ± 0.006	0.139 ± 0.006	0.164 ± 0.005	0.162 ± 0.004	0.141 ± 0.009	0.150 ± 0.006	0.145 ± 0.006	**0.138 ± 0.007**
Bibtex	0.015 ± 0.000	**0.013 ± 0.000**	0.033 ± 0.008	0.021 ± 0.000	**0.013 ± 0.000**	0.016 ± 0.000	0.022 ± 0.000	0.014 ± 0.000
Medical	0.017 ± 0.000	0.013 ± 0.002	0.054 ± 0.004	0.029 ± 0.003	0.012 ± 0.003	0.022 ± 0.001	0.030 ± 0.005	**0.011 ± 0.002**
Birds	0.056 ± 0.004	0.058 ± 0.010	0.083 ± 0.005	0.122 ± 0.021	0.056 ± 0.005	0.155 ± 0.029	0.131 ± 0.007	**0.055 ± 0.009**
Emotions	0.638 ± 0.022	0.642 ± 0.044	0.347 ± 0.025	0.420 ± 0.019	0.248 ± 0.016	0.249 ± 0.018	0.296 ± 0.026	**0.246 ± 0.015**
One Error (the smaller, the better)								
Genbase	0.015 ± 0.014	0.003 ± 0.004	0.011 ± 0.009	0.004 ± 0.005	**0.002 ± 0.003**	0.007 ± 0.011	0.013 ± 0.009	**0.002 ± 0.003**
CAL500	0.117 ± 0.018	0.128 ± 0.033	0.305 ± 0.141	0.126 ± 0.012	0.116 ± 0.019	0.282 ± 0.050	**0.116 ± 0.018**	0.118 ± 0.020
Bibtex	0.634 ± 0.005	0.418 ± 0.013	0.638 ± 0.016	0.389 ± 0.009	0.458 ± 0.017	0.588 ± 0.002	0.749 ± 0.023	**0.373 ± 0.011**
Medical	0.283 ± 0.030	0.191 ± 0.050	0.326 ± 0.053	0.241 ± 0.054	0.196 ± 0.040	0.234 ± 0.032	0.445 ± 0.065	**0.164 ± 0.043**
Birds	0.745 ± 0.053	0.771 ± 0.078	0.483 ± 0.115	0.806 ± 0.055	0.648 ± 0.041	0.619 ± 0.043	0.691 ± 0.050	**0.445 ± 0.064**
Emotions	0.480 ± 0.038	0.479 ± 0.062	0.334 ± 0.054	0.553 ± 0.051	0.391 ± 0.038	**0.316 ± 0.040**	0.418 ± 0.059	0.355 ± 0.052
Coverage (the smaller, the better)								
Genbase	0.031 ± 0.017	0.029 ± 0.013	0.054 ± 0.019	0.027 ± 0.013	0.030 ± 0.011	0.045 ± 0.018	0.042 ± 0.014	**0.021 ± 0.007**
CAL500	0.760 ± 0.015	0.760 ± 0.017	0.979 ± 0.004	0.764 ± 0.011	0.753 ± 0.017	0.947 ± 0.006	**0.701 ± 0.016**	0.754 ± 0.015
Bibtex	0.369 ± 0.011	0.216 ± 0.007	0.536 ± 0.013	0.206 ± 0.019	**0.178 ± 0.008**	0.281 ± 0.003	0.205 ± 0.011	0.211 ± 0.006
Medical	0.098 ± 0.014	0.062 ± 0.013	0.167 ± 0.054	0.077 ± 0.024	0.070 ± 0.017	0.115 ± 0.018	0.109 ± 0.017	**0.042 ± 0.011**
Birds	0.227 ± 0.026	0.213 ± 0.024	0.249 ± 0.040	0.265 ± 0.041	0.184 ± 0.025	0.198 ± 0.024	0.206 ± 0.021	**0.131 ± 0.031**
Emotions	0.473 ± 0.023	0.464 ± 0.052	0.436 ± 0.050	0.518 ± 0.026	0.366 ± 0.039	0.364 ± 0.037	0.384 ± 0.039	**0.352 ± 0.028**
Average Precision (the larger, the better)								
Genbase	0.978 ± 0.013	0.980 ± 0.013	0.661 ± 0.067	0.988 ± 0.005	0.980 ± 0.012	0.958 ± 0.026	0.966 ± 0.082	**0.991 ± 0.004**
CAL500	0.489 ± 0.010	0.496 ± 0.012	0.200 ± 0.024	0.503 ± 0.016	0.500 ± 0.031	0.393 ± 0.016	**0.511 ± 0.011**	0.511 ± 0.013
Bibtex	0.318 ± 0.006	0.488 ± 0.013	0.226 ± 0.024	0.543 ± 0.013	0.486 ± 0.011	0.480 ± 0.005	0.474 ± 0.015	**0.553 ± 0.007**
Medical	0.756 ± 0.024	0.839 ± 0.037	0.606 ± 0.096	0.800 ± 0.048	0.833 ± 0.036	0.744 ± 0.026	0.646 ± 0.047	**0.869 ± 0.032**
Birds	0.331 ± 0.026	0.360 ± 0.043	0.410 ± 0.039	0.315 ± 0.040	0.441 ± 0.026	0.416 ± 0.019	0.403 ± 0.034	**0.559 ± 0.048**
Emotions	0.631 ± 0.018	0.636 ± 0.031	0.669 ± 0.040	0.579 ± 0.021	0.726 ± 0.015	0.737 ± 0.014	0.702 ± 0.022	**0.742 ± 0.022**

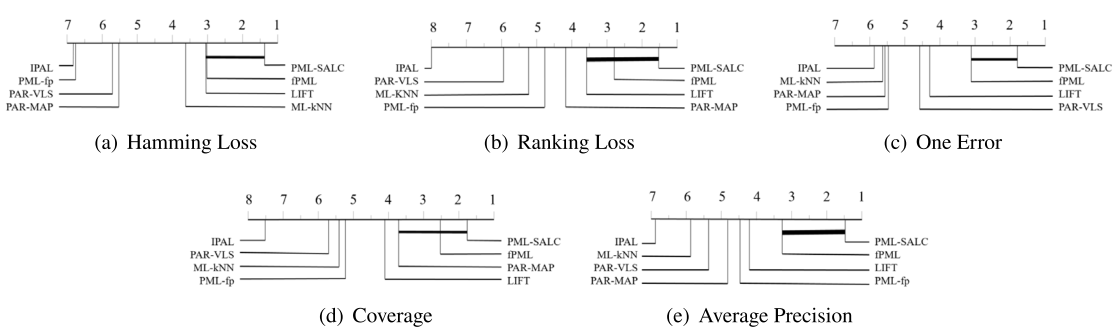


Fig. 3. Comparison of PML-SALC (control method) and baseline methods with the Bonferroni–Dunn test. Methods not connected with PML-SALC in the CD diagram are considered to have a significantly different performance from PML-SALC. (CD = 2.2601 at 0.05 significance level).

4.3. Ablation study

To verify the effectiveness of each constraint term in the objective function, we perform an ablation study. Specifically, we compare PML-SALC with its degenerated algorithms, which are sparse-free, correlation-free and neighbor-free. For the sparse-free case, we remove the term that controls the sparse label correlations from PML-SALC. For the correlation-free case, we remove the term that controls the global label correlations from PML-SALC. For the neighbor-free case, we remove the term that controls similar instances with similar labels from PML-SALC. We convert each dataset into PML datasets with r=1 and perform five-fold cross-validation. We also set $\gamma = 0$, $\beta = 0$ and $\sigma = 0$. The experimental results are shown in Tables 7–9. We observe that PML-SALC is superior to each term-free algorithm in all cases, which demonstrates the effectiveness of each term on PML.

4.4. Parameter analysis

There are five hyper-parameters used in the optimization of PML-SALC, which are α, β, γ, σ and φ. To test the effect of the parameters and demonstrate how to select the proper value, we fix four of the five parameters and run PML-SALC with varying values of the remaining parameter. Fig. 4 gives the experimental results on the Birds dataset.

First, to explore how the parameter impacts the performance of partial multi-label learning, we set $\beta = 1000$, $\gamma = 400$, $\sigma = 1000$, $\varphi = 1$, and search $\alpha \in \{0.1, 1, 10, 100\}$. Theoretically, α

Table 4
Experimental results (mean ± standard deviation) of each comparison method on five synthesized PML datasets, where 3 noisy labels are added to each instance.

Datasets	ML-KNN	LIFT	IPAL	PML-*fp*	fPML	PAR-VLS	PAR-MAP	PML-SALC
Ranking Loss (the smaller, the better)								
Genbase	0.013 ± 0.010	0.011 ± 0.009	0.065 ± 0.013	0.009 ± 0.003	0.009 ± 0.005	0.023 ± 0.006	0.017 ± 0.013	**0.004** ± **0.003**
CAL500	0.188 ± 0.008	0.185 ± 0.006	0.884 ± 0.027	0.183 ± 0.009	0.182 ± 0.005	0.543 ± 0.018	**0.176** ± **0.005**	0.179 ± 0.015
Bibtex	0.237 ± 0.003	0.124 ± 0.005	0.728 ± 0.059	0.112 ± 0.006	**0.108** ± **0.004**	0.320 ± 0.007	0.335 ± 0.006	0.129 ± 0.006
Medical	0.087 ± 0.015	0.043 ± 0.011	0.261 ± 0.102	0.053 ± 0.013	0.049 ± 0.013	0.109 ± 0.018	0.089 ± 0.013	**0.030** ± **0.008**
Birds	0.379 ± 0.022	0.358 ± 0.050	0.758 ± 0.055	0.442 ± 0.051	0.391 ± 0.026	0.354 ± 0.025	0.337 ± 0.035	**0.217** ± **0.028**
Hamming Loss (the smaller, the better)								
Genbase	0.007 ± 0.004	0.004 ± 0.004	0.062 ± 0.010	0.041 ± 0.001	0.005 ± 0.006	0.027 ± 0.005	0.014 ± 0.004	**0.004** ± **0.003**
CAL500	0.140 ± 0.006	**0.139** ± **0.007**	0.156 ± 0.009	0.162 ± 0.005	0.140 ± 0.008	0.150 ± 0.006	0.145 ± 0.006	0.139 ± 0.009
Bibtex	0.015 ± 0.000	**0.013** ± **0.000**	0.033 ± 0.000	0.021 ± 0.000	**0.013** ± **0.000**	0.017 ± 0.000	0.022 ± 0.000	0.014 ± 0.000
Medical	0.018 ± 0.001	0.013 ± 0.002	0.084 ± 0.009	0.030 ± 0.002	0.013 ± 0.002	0.024 ± 0.001	0.084 ± 0.003	**0.012** ± **0.003**
Birds	0.059 ± 0.007	0.061 ± 0.013	0.088 ± 0.014	0.122 ± 0.022	0.109 ± 0.013	0.227 ± 0.029	0.148 ± 0.008	**0.053** ± **0.007**
One Error (the smaller, the better)								
Genbase	0.016 ± 0.014	0.005 ± 0.006	0.022 ± 0.007	0.019 ± 0.006	**0.003** ± **0.003**	0.008 ± 0.009	0.018 ± 0.013	0.004 ± 0.008
CAL500	0.117 ± 0.017	0.122 ± 0.018	0.371 ± 0.140	0.127 ± 0.012	0.117 ± 0.022	0.291 ± 0.047	**0.116** ± **0.018**	0.117 ± 0.020
Bibtex	0.640 ± 0.009	0.434 ± 0.011	0.642 ± 0.016	0.396 ± 0.016	0.460 ± 0.007	0.588 ± 0.007	0.751 ± 0.018	**0.386** ± **0.015**
Medical	0.312 ± 0.037	0.193 ± 0.042	0.341 ± 0.052	0.255 ± 0.050	0.197 ± 0.042	0.280 ± 0.057	0.453 ± 0.074	**0.164** ± **0.037**
Birds	0.778 ± 0.049	0.814 ± 0.065	0.492 ± 0.122	0.817 ± 0.045	0.599 ± 0.044	0.617 ± 0.059	0.730 ± 0.055	**0.468** ± **0.053**
Coverage (the smaller, the better)								
Genbase	0.032 ± 0.019	0.032 ± 0.048	0.065 ± 0.017	0.029 ± 0.012	0.029 ± 0.015	0.046 ± 0.022	0.043 ± 0.012	**0.018** ± **0.009**
CAL500	0.766 ± 0.017	0.766 ± 0.013	0.980 ± 0.002	0.960 ± 0.019	0.757 ± 0.018	0.967 ± 0.004	**0.703** ± **0.016**	0.760 ± 0.017
Bibtex	0.399 ± 0.004	0.227 ± 0.006	0.542 ± 0.019	0.206 ± 0.008	**0.184** ± **0.007**	0.283 ± 0.004	0.205 ± 0.010	0.221 ± 0.010
Medical	0.116 ± 0.017	0.062 ± 0.015	0.173 ± 0.050	0.073 ± 0.016	0.070 ± 0.031	0.124 ± 0.019	0.114 ± 0.017	**0.045** ± **0.011**
Birds	0.240 ± 0.031	0.232 ± 0.034	0.322 ± 0.037	0.263 ± 0.038	0.148 ± 0.022	0.251 ± 0.025	0.216 ± 0.020	**0.151** ± **0.025**
Average Precision (the larger, the better)								
Genbase	0.974 ± 0.017	0.981 ± 0.013	0.572 ± 0.045	0.982 ± 0.005	0.981 ± 0.013	0.957 ± 0.015	0.960 ± 0.011	**0.989** ± **0.006**
CAL500	0.481 ± 0.009	0.497 ± 0.013	0.183 ± 0.012	0.503 ± 0.017	0.501 ± 0.011	0.397 ± 0.018	**0.512** ± **0.011**	0.510 ± 0.017
Bibtex	0.312 ± 0.003	0.486 ± 0.009	0.208 ± 0.029	0.530 ± 0.010	0.486 ± 0.008	0.479 ± 0.004	0.473 ± 0.006	**0.533** ± **0.009**
Medical	0.728 ± 0.029	0.836 ± 0.032	0.562 ± 0.092	0.789 ± 0.035	0.832 ± 0.033	0.708 ± 0.042	0.641 ± 0.053	**0.865** ± **0.028**
Birds	0.338 ± 0.024	0.342 ± 0.042	0.350 ± 0.029	0.289 ± 0.036	0.390 ± 0.029	0.337 ± 0.025	0.369 ± 0.039	**0.561** ± **0.034**

Table 5
The ranks of each comparison method for five metrics.

	MLKNN	LIFT	IPAL	PML-*fp*	fPML	PAR-VLS	PAR-MAP	PML-SALC (Ours)
Average Precision	5.8824	4.2059	6.8824	4.4706	3.2647	5.3529	4.8235	1.4706
One Error	5.5588	4.2941	5.6471	5.4706	3.1176	4.5882	5.5882	1.7941
Ranking Loss	5.2353	3.5588	8.0000	4.7647	2.7941	5.9412	4.1765	1.5294
Coverage	5.3529	4.1176	7.5294	5.2353	5.5294	5.7059	3.7059	1.7647
Hamming Loss	3.6491	3.0294	6.8235	6.7647	3.0588	5.5294	5.7059	1.3824

Table 6
Friedman statistics F_F for each evaluation metric at 0.05 significance level (8 algorithms and 17 datasets)).

Evaluation metric	F_F	Critical value
Ranking Loss	43.1650	
Hamming Loss	40.6803	
One Error	9.7694	2.0924
Coverage	26.1143	
Average Precision	23.0639	

controls the relative importance of model complexity. When α is too large, the importance of other terms is almost ignored, which makes it difficult to train an effective model. When α is too small, the importance of complexity control is almost ignored, which easily causes model to be overfitting. Therefore, PML-SALC cannot train the model very well in both extreme cases. From Fig. 4(a), the following can be observed: (1) The average precision is generally stable and increases slightly when α increases. However, after α reaches the value 1, the average precision decreases. (2) The other four metrics are generally stable and decrease slightly when α increases. However, after α reaches the value 1, they increase. From the above observations, we can see that when α is set to 1, PML-SALC achieves the best performance with respect to the five evaluation metrics.

Second, to explore how the parameter β impacts the performance of partial multi-label learning, we set $\alpha = 1$, $\gamma = 400$, $\sigma = 1000$, $\varphi = 1$, and search $\beta \in \{500, 1000, 1500, 2000\}$. Theoretically, β controls the proximity of label correlation matrix $\mathbf{Q}$ to the conditional probability matrix $\mathbf{P}$. When β is too large, the importance of other terms is almost ignored, which makes it difficult to train an effective model. When β is too small, the importance of the proximity of $\mathbf{Q}$ to $\mathbf{P}$ is almost ignored, which cannot make full use of the global asymmetric label structure information to capture the real label correlation. From Fig. 4(b), we can observe the same performance trends with respect to the five evaluation metrics noted below: The performance of PML-SALC increases when β increases. However, after β reaches a certain value, the performance decreases. When β is set to 1000, PML-SALC achieves the best performance.

Third, to explore how the parameter γ impacts the performance of partial multi-label learning, we set $\alpha = 1$, $\beta = 1000$, $\sigma = 1000$, $\varphi = 1$, and search $\gamma \in \{200, 400, 600, 800\}$. Theoretically, γ controls the sparse of label correlation matrix $\mathbf{Q}$. When γ is too large, the importance of other terms is almost ignored, which makes it difficult to train an effective model. When γ is too small, the importance of the sparse of $\mathbf{Q}$ is almost ignored, which cannot further reduce the negative impact of noisy label in label correlation learning. From Fig. 4(c), we can see when γ is set to 600, the performance is best with respect to the evaluation

Table 7
The experimental comparison between the proposed PML-SALC and sparse-free algorithm on PML datasets, where 1 noisy label is added to each instance.

Dataset	Method	Ranking Loss	One error	Average Precision	Coverage	Hamming Loss
Birds	PML-SALC	**0.177 ± 0.025**	**0.408 ± 0.041**	**0.623 ± 0.032**	**0.124 ± 0.021**	**0.052 ± 0.008**
	sparse-free	0.189 ± 0.020	0.529 ± 0.075	0.556 ± 0.051	0.130 ± 0.017	0.058 ± 0.008
Emotions	PML-SALC	**0.179 ± 0.008**	**0.315 ± 0.044**	**0.778 ± 0.014**	**0.311 ± 0.023**	**0.221 ± 0.015**
	sparse-free	0.233 ± 0.019	0.452 ± 0.028	0.703 ± 0.021	0.356 ± 0.028	0.257 ± 0.016
CAL500	PML-SALC	**0.177 ± 0.006**	**0.116 ± 0.017**	**0.513 ± 0.014**	**0.747 ± 0.016**	**0.138 ± 0.007**
	sparse-free	0.177 ± 0.007	0.118 ± 0.020	0.511 ± 0.014	0.749 ± 0.014	0.140 ± 0.008
Genbase	PML-SALC	**0.004 ± 0.002**	**0.002 ± 0.003**	**0.990 ± 0.006**	**0.019 ± 0.010**	**0.003 ± 0.003**
	sparse-free	0.005 ± 0.003	0.003 ± 0.003	0.885 ± 0.004	0.021 ± 0.012	0.004 ± 0.003
Medical	PML-SALC	**0.022 ± 0.005**	**0.150 ± 0.042**	**0.886 ± 0.028**	**0.036 ± 0.010**	**0.011 ± 0.003**
	sparse-free	0.028 ± 0.006	0.194 ± 0.068	0.850 ± 0.046	0.043 ± 0.012	0.013 ± 0.003
bibtex	PML-SALC	**0.106 ± 0.006**	**0.367 ± 0.013**	**0.570 ± 0.008**	**0.200 ± 0.010**	**0.014 ± 0.000**
	sparse-free	0.138 ± 0.005	0.436 ± 0.010	0.520 ± 0.006	0.225 ± 0.009	0.016 ± 0.000

Table 8
The experimental comparison between the proposed PML-SALC and correlation-free algorithm on PML datasets, where 1 noisy label is added to each instance.

Dataset	Method	Ranking Loss	One error	Average Precision	Coverage	Hamming Loss
Birds	PML-SALC	**0.177 ± 0.025**	**0.408 ± 0.041**	**0.623 ± 0.032**	**0.124 ± 0.021**	**0.052 ± 0.008**
	correlation-free	0.483 ± 0.030	0.907 ± 0.027	0.219 ± 0.022	0.293 ± 0.036	0.076 ± 0.010
Emotions	PML-SALC	**0.179 ± 0.008**	**0.315 ± 0.044**	**0.778 ± 0.014**	**0.311 ± 0.023**	**0.221 ± 0.015**
	correlation-free	0.524 ± 0.073	0.713 ± 0.117	0.487 ± 0.059	0.605 ± 0.063	0.387 ± 0.057
CAL500	PML-SALC	**0.177 ± 0.006**	**0.116 ± 0.017**	**0.513 ± 0.014**	**0.747 ± 0.016**	**0.138 ± 0.007**
	correlation-free	0.486 ± 0.030	0.787 ± 0.018	0.183 ± 0.015	0.958 ± 0.011	0.154 ± 0.005
Genbase	PML-SALC	**0.316 ± 0.090**	**0.00728 ± 0.113**	**0.385 ± 0.009**	**0.340 ± 0.079**	**0.044 ± 0.007**
	correlation-free	0.005 ± 0.003	0.003 ± 0.003	0.885 ± 0.004	0.021 ± 0.012	0.004 ± 0.003
Medical	PML-SALC	**0.022 ± 0.005**	**0.150 ± 0.042**	**0.886 ± 0.028**	**0.036 ± 0.010**	**0.011 ± 0.003**
	correlation-free	0.497 ± 0.067	0.971 ± 0.017	0.101 ± 0.020	0.527 ± 0.059	0.037 ± 0.003
bibtex	PML-SALC	**0.106 ± 0.006**	**0.367 ± 0.013**	**0.570 ± 0.008**	**0.200 ± 0.010**	**0.014 ± 0.000**
	sparse-free	0.413 ± 0.008	0.752 ± 0.021	0.154 ± 0.010	0.548 ± 0.012	0.103 ± 0.056

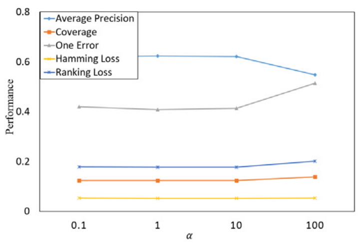
Table 9
The experimental comparison between the proposed PML-SALC and neighbor-free algorithm on PML datasets, where 1 noisy label is added to each instance.

Dataset	Method	Ranking Loss	One error	Average Precision	Coverage	Hamming Loss
Birds	PML-SALC	**0.177 ± 0.025**	**0.408 ± 0.041**	**0.623 ± 0.032**	**0.124 ± 0.021**	**0.052 ± 0.008**
	neighbor-free	0.180 ± 0.024	0.507 ± 0.066	0.572 ± 0.045	0.125 ± 0.019	0.058 ± 0.008
Emotions	PML-SALC	**0.179 ± 0.008**	**0.315 ± 0.044**	**0.778 ± 0.014**	**0.311 ± 0.023**	**0.221 ± 0.015**
	neighbor-free	0.282 ± 0.024	0.440 ± 0.148	0.678 ± 0.061	0.401 ± 0.025	0.277 ± 0.035
CAL500	PML-SALC	**0.177 ± 0.006**	**0.116 ± 0.017**	**0.513 ± 0.014**	**0.747 ± 0.016**	**0.138 ± 0.007**
	neighbor-free	0.178 ± 0.008	0.120 ± 0.023	0.512 ± 0.014	0.749 ± 0.013	0.138 ± 0.008
Genbase	PML-SALC	**0.004 ± 0.002**	**0.002 ± 0.003**	**0.990 ± 0.006**	**0.019 ± 0.010**	**0.003 ± 0.003**
	neighbor-free	0.004 ± 0.003	0.002 ± 0.004	**0.990 ± 0.006**	**0.019 ± 0.010**	**0.003 ± 0.003**
Medical	PML-SALC	**0.022 ± 0.005**	**0.150 ± 0.042**	**0.886 ± 0.028**	**0.036 ± 0.010**	**0.011 ± 0.003**
	neighbor-free	0.023 ± 0.006	0.154 ± 0.049	0.879 ± 0.034	0.037 ± 0.011	0.012 ± 0.003
bibtex	PML-SALC	**0.106 ± 0.006**	**0.367 ± 0.013**	**0.570 ± 0.008**	**0.200 ± 0.010**	**0.014 ± 0.000**
	sparse-free	0.110 ± 0.005	0.396 ± 0.012	0.545 ± 0.007	0.226 ± 0.009	0.015 ± 0.000

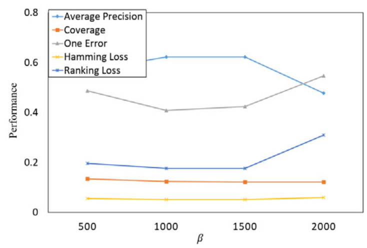
metric of Hamming Loss. While γ is set to 400, the performances are best with respect to the other four evaluation metrics and the performance is also very closed to the best performance with respect to the evaluation metric of Hamming Loss.

Fourth, to explore how the parameter σ impacts the performance of partial multi-label learning, we set $\alpha = 1$, $\beta = 1000$, $\gamma = 400$, $\varphi = 1$, and search $\sigma \in \{1, 100, 1000, 10000\}$. Theoretically, σ controls the proximity of each instance to its neighbors with respect to the label confidence. When σ is too large, the importance of other terms is almost ignored, which makes it difficult to train an effective model. When σ is too small, the importance of the proximity of each instance to its neighbors with respect to the label confidence is almost ignored, which cannot make full use of the local feature structure information to capture the real label correlation. From Fig. 4(d), we can observe the same performance trends with respect to the five evaluation metrics noted below: The performance of PML-SALC increases when σ increases. However, after σ reaches a certain value, the performance decreases. When σ is set to 1000, PML-SALC achieves the best performance.

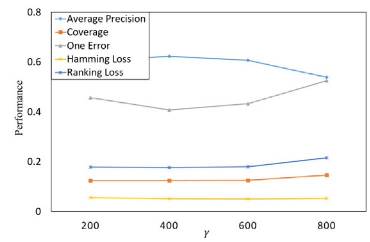
Last, to explore how the parameter φ impacts the performance of partial multi-label learning, we set $\alpha = 1$, $\beta = 1000$, $\gamma = 400$, $\sigma = 1000$, and search $\varphi \in \{0.1, 1, 10, 100\}$. Theoretically, φ controls the low rank of the mapping matrix $\mathbf{W}$. When φ is too large, the importance of other terms is almost ignored, which makes it difficult to train an effective model. When φ is too small, the importance of the low rank of the mapping matrix $\mathbf{W}$ is almost ignored, which cannot make full use of label correlation



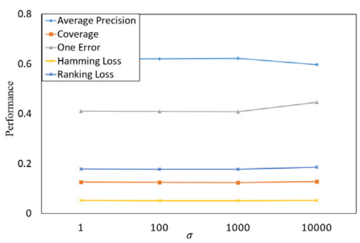
(a) Performance curve with α changes



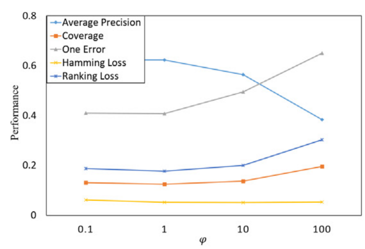
(b) Performance curve with β changes



(c) Performance curve with γ changes

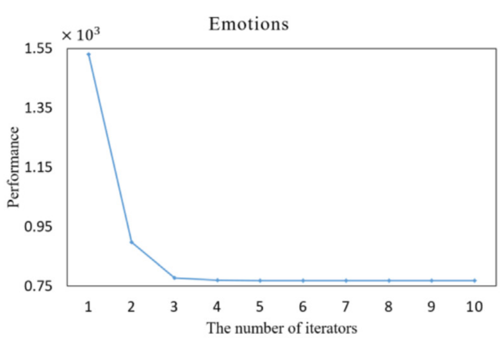


(d) Performance curve with σ changes

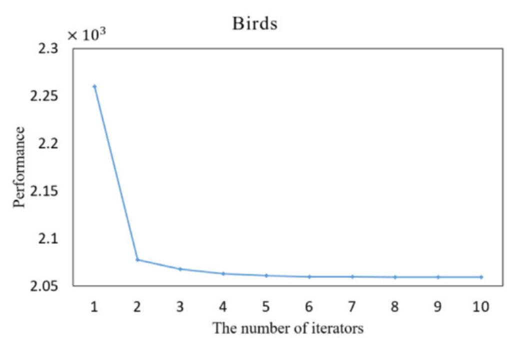


(e) Performance curve with φ changes

Fig. 4. Results of PML-SALC with varying values of trade-off parameters on Birds..



(a)



(b)

Fig. 5. The convergence curves of PML-SALC on Emotions and Birds.

to learn the classifier. From Fig. 4(e), we can observe that when φ is set to 1, PML-SALC achieves the best performance.

4.5. Convergence analysis

Eq. (8) is not convex for $\mathbf{W}$ and $\mathbf{Q}$ simultaneously, and its optimal solution cannot be obtained directly. However, if the variables $\mathbf{W}$ and $\mathbf{Q}$ are considered separately, Eq. (8) can be converted to a convex programming problem. Therefore, we employ alternating optimization to solve it. We conduct the convergence analysis of PML-SALC on Emotions and Birds. The experimental results are shown in Fig. 5. From each subfigure of Fig. 5, it can be easily observed that the loss decreases to a certain value as the number of iterations increases, which demonstrates the convergence of PML-SALC.

5. Conclusion

Partial multi-label learning is a very challenging problem. In this paper, we present a novel PML method based on sparse asymmetric label correlations (PML-SALC). This method utilizes the sparse asymmetric label correlation matrix to better learn label confidence and models the relationship between the feature and label confidence to make the model smoother. Extensive experiments show that PML-SALC can effectively alleviate the negative influence of noisy labels and learn a more robust model.

CRediT authorship contribution statement

Peng Zhao: Conceptualization, Methodology, Writing – review & editing, Funding acquisition. **Shiyi Zhao:** Software, Validation, Formal analysis, Data curation, Writing – original draft. **Xuyang Zhao:** Data curation. **Huiting Liu:** Investigation. **Xia Ji:** Resources.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported by the Natural Science Foundation of Anhui Province, China (grant number 2108085MF212); the Key Natural Science Project of Anhui Provincial Education Department, China (grant numbers KJ2021A0043); the National Natural Science Foundation of China (grant number 61602004); and the Natural Science Foundation of Anhui Province, China (grant numbers 2008085MF219, 1908085MF188).

References

[1] Min-Ling Zhang, Zhi-Hua Zhou, A review on multi-label learning algorithms, IEEE Trans. Knowl. Data Eng. 26 (2013) 1819–1837, https://ieeexplore.ieee.org/document/6471714.

[2] Zhao-Min Chen, Xiu-Shen Wei, Peng Wang, Yanwen Guo, Multi-label image recognition with graph convolutional networks, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 5177–5186, http://dx.doi.org/10.1109/CVPR.2019.00532.

[3] Siddharth Gopal, Yiming Yang, Multilabel classification with meta-level features, in: Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval, 2010, pp. 315–322, http://dx.doi.org/10.1145/1835449.1835503.

[4] Ming-Kun Xie, Sheng-Jun Huang, Partial multi-label learning, in: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 32, 2018, https://ojs.aaai.org/index.php/AAAI/article/view/11644.

[5] Eva Gibaja, Sebastián Ventura, A tutorial on multilabel learning, ACM Comput. Surv. 47 (2015) 1–38, http://dx.doi.org/10.1145/2716262.

[6] Z.-H. Zhou, M.-L. Zhang, Multi-label learning, 2017, http://dx.doi.org/10.1007/978-1-4899-7687-1_910.

[7] Matthew R. Boutell, Jiebo Luo, Xipeng Shen, Christopher M Brown, Learning multi-label scene classification, Pattern Recognit. 37 (2004) 1757–1771, http://dx.doi.org/10.1016/j.patcog.2004.03.009.

[8] Thorsten Joachims, Text categorization with support vector machines: Learning with many relevant features, in: European Conference on Machine Learning, Springer, 1998, pp. 137–142, http://dx.doi.org/10.1007/BFb0026683.

[9] André Elisseeff, Jason Weston, et al., A kernel method for multi-labelled classification, in: NIPS, vol. 14, 2001, pp. 681–687, https://proceedings.neurips.cc/paper/2001/file/39dcaf7a053dc372fbc391d4e6b5d693-Paper.pdf.

[10] Johannes Fürnkranz, Eyke Hüllermeier, Eneldo Loza Mencía, Klaus Brinker, Multilabel classification via calibrated label ranking, Mach. Learn. 73 (2008) 133–153, http://dx.doi.org/10.1007/s10994-008-5064-8.

[11] Grigorios Tsoumakas, Ioannis Katakis, Ioannis Vlahavas, Random k-labelsets for multilabel classification, IEEE Trans. Knowl. Data Eng. 23 (2010) 1079–1089, http://dx.doi.org/10.1109/TKDE.2010.164.

[12] Jesse Read, Bernhard Pfahringer, Geoff Holmes, Eibe Frank, Classifier chains for multi-label classification, Mach. Learn. 85 (2011) 333, http://dx.doi.org/10.1007/s10994-011-5256-5.

[13] N. Xu, J. Shu, Y.-P. Liu, X. Geng, Variational label enhancement, in: International Conference on Machine Learning, ICML'20, 2020, pp. 10597–10606, https://proceedings.mlr.press/v119/xu20g.html.

[14] N. Xu, Y.P. Liu, X. Geng, Label enhancement for label distribution learning, IEEE Trans. Knowl. Data Eng. 33 (2021) 1632–1643, http://dx.doi.org/10.1109/TKDE.2019.2947040, https://ieeexplore.ieee.org/document/8868206.

[15] Timothee Cour, Ben Sapp, Ben Taskar, Learning from partial labels, J. Mach. Learn. Res. 12 (2011) 1501–1536, https://dl.acm.org/doi/10.5555/1953048.2021049.

[16] Eyke Hüllermeier, Jürgen Beringer, Learning from ambiguously labeled examples, Intell. Data Anal. 10 (2006) 419–439, http://dx.doi.org/10.3233/IDA-2006-10503.

[17] Liping Liu, Thomas G. Dietterich, A conditional multinomial mixture model for superset label learning, in: Advances in Neural Information Processing Systems, Citeseer, 2012, pp. 548–556, https://proceedings.neurips.cc/paper/2012/file/aaebdb8bb6b0e73f6c3c54a0ab0c6415-Paper.pdf.

[18] Lei Feng, Bo An, Partial label learning with self-guided retraining, in: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, 2019, pp. 3542–3549, http://dx.doi.org/10.1609/aaai.v33i01.33013542.

[19] Rong Jin, Zoubin Ghahramani, Learning with multiple labels, in: NIPS, vol. 2, Citeseer, 2002, pp. 897–904, https://proceedings.neurips.cc/paper/2002/file/653ac11ca60b3e021a8c609c7198acfc-Paper.pdf.

[20] Yi-Chen Chen, Vishal M. Patel, Rama Chellappa, P. Jonathon Phillips, Ambiguously labeled learning using dictionaries, IEEE Trans. Inf. Forensics Secur. 9 (2014) 2076–2088, http://dx.doi.org/10.1109/TIFS.2014.2359642.

[21] Fei Yu, Min-Ling Zhang, Maximum margin partial label learning, in: Asian Conference on Machine Learning, PMLR, 2016, pp. 96–111, http://proceedings.mlr.press/v45/Yu15.html.

[22] Cai-Zhi Tang, Min-Ling Zhang, Confidence-rated discriminative partial label learning, in: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 31, 2017, https://dl.acm.org/doi/10.5555/3298483.3298615.

[23] N. Xu, C. Qiao, X. Geng, M.-L. Zhang, Instance-dependent partial label learning, Adv. Neural Inf. Process. Syst. 34 (2021) https://proceedings.neurips.cc/paper/2021/hash/e38e37a99f7de1f45d169efcdb288dd1-Abstract.html.

[24] Min-Ling Zhang, Fei Yu, Cai-Zhi Tang, Disambiguation-free partial label learning, IEEE Trans. Knowl. Data Eng. 29 (2017) 2155–2167, http://dx.doi.org/10.1109/TKDE.2017.2721942.

[25] Min-Ling Zhang, Zhi-Hua Zhou, ML-KNN: A lazy learning approach to multi-label learning, Pattern Recognit. 40 (2007) 2038–2048, http://dx.doi.org/10.1016/j.patcog.2006.12.019.

[26] Min-Ling Zhang, Lei Wu, Lift: Multi-label learning with label-specific features, IEEE Trans. Pattern Anal. Mach. Intell. 37 (2014) 107–120, https://doi.ieeecomputersociety.org/10.1109/TPAMI.2014.2339815.

[27] Jun-Peng Fang, Min-Ling Zhang, Partial multi-label learning via credible label elicitation, in: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, 2019, pp. 3518–3525, http://dx.doi.org/10.1609/aaai.v33i01.33013518.

[28] Lijuan Sun, Songhe Feng, Tao Wang, Congyan Lang, Yi Jin, Partial multi-label learning by low-rank and sparse decomposition, in: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, 2019, pp. 5016–5023, http://dx.doi.org/10.1609/aaai.v33i01.33015016.

[29] Ming-Kun Xie, Sheng-Jun Huang, Partial multi-label learning with noisy label identification, in: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, 2020, pp. 6454–6461, http://dx.doi.org/10.1609/aaai.v34i04.6117.

[30] Z. Li, J. Liu, J. Tang, H. Lu, Robust structured subspace learning for data representation, IEEE Trans. Pattern Anal. Mach. Intell. 37 (2015) 2085–2098, https://ieeexplore.ieee.org/document/7031960.

[31] Z. Li, J. Tang, Weakly supervised deep matrix factorization for social image understanding, IEEE Trans. Image Process. 26 (2016) 276–288, https://ieeexplore.ieee.org/document/7728069.

[32] Z. Li, J. Tang, T. Mei, Deep collaborative embedding for social image understanding, IEEE Trans. Pattern Anal. Mach. Intell. 41 (2018) 2070–2083, https://ieeexplore.ieee.org/document/8403294.

[33] Z. Li, J. Tang, L. Zhang, J. Yang, Weakly-supervised semantic guided hashing for social image retrieval, Int. J. Comput. Vis. 128 (2020) https://link.springer.com/article/10.1007/s11263-020-01331-0.

[34] X. Jia, X. Zheng, W. Li, C. Zhang, Z. Li, Facial emotion distribution learning by exploiting low-rank label correlations locally, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 9841–9850, https://ieeexplore.ieee.org/document/8953790.

[35] D. Zhao, Q. Gao, Y. Lu, D. Sun, Y. Cheng, Consistency and diversity neural network multi-view multi-label learning, Knowl.-Based Syst. 218 (2021) 106841, https://www.sciencedirect.com/science/article/pii/S0950705121001040.

[36] C. Zhu, P. Wang, L. Ma, R. Zhou, L. Wei, Global and local multi-view multi-label learning with incomplete views and labels, Neural Comput. Appl. 32 (2020) 15007–15028, https://link.springer.com/content/pdf/10.1007/s00521-020-04854-2.pdf.

[37] C. Wang, X. Chen, B. Chen, F. Nie, B. Wang, Z. Ming, Learning unsupervised node representation from multi-view network, Inform. Sci. 579 (2021) 700–716, https://www.sciencedirect.com/science/article/pii/S0020025521007775.

[38] K. Jia, J. Lin, M. Tan, D. Tao, Deep multi-view learning using neuron-wise correlation-maximizing regularizers, IEEE Trans. Image Process. 28 (2019) 5121–5134, https://ieeexplore.ieee.org/document/8708947.

[39] Weiwei Liu, Xiaobo Shen, Haobo Wang, Ivor W. Tsang, The emerging trends of multi-label learning, 2020, arXiv preprint arXiv:2011.11197.

[40] Haobo Wang, Weiwei Liu, Yang Zhao, Chen Zhang, Tianlei Hu, Gang Chen, Discriminative and correlative partial multi-label learning, in: IJCAI, 2019, pp. 3691–3697, http://dx.doi.org/10.24963/ijcai.2019/512.

[41] Ziwei Li, Gengyu Lyu, Songhe Feng, Partial multi-label learning via multi-subspace representation, in: International Joint Conference on Artificial Intelligence, 2020, http://dx.doi.org/10.24963/ijcai.2020/362.

[42] Guoxian Yu, Xia Chen, Carlotta Domeniconi, Jun Wang, Zhao Li, Zili Zhang, Xindong Wu, Feature-induced partial multi-label learning, in: 2018 IEEE International Conference on Data Mining, ICDM, IEEE, 2018, pp. 1398–1403, http://dx.doi.org/10.1109/ICDM.2018.00192.

[43] Gengyu Lyu, Songhe Feng, Yidong Li, Noisy label tolerance: A new perspective of partial multi-label learning, Inform. Sci. 543 (2021) 454–466, http://dx.doi.org/10.1016/j.ins.2020.09.019.

[44] Yan Yan, Shining Li, Lei Feng, Partial multi-label learning with mutual teaching, Knowl.-Based Syst. 212 (2021) 106624, https://www.sciencedirect.com/science/article/pii/S095070512030753X.

[45] Ning Xu, Yun-Peng Liu, Xin Geng, Partial multi-label learning with label distribution, in: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, 2020, pp. 6510–6517, http://dx.doi.org/10.1609/aaai.v34i04.6124.

[46] Sheng-Jun Huang, Miao Xu, Ming-Kun Xie, Masashi Sugiyama, Gang Niu, Songcan Chen, Active feature acquisition with supervised matrix completion, in: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2018, pp. 1571–1579, http://dx.doi.org/10.1145/3219819.3220084.

[47] Jian-Feng Cai, Emmanuel J. Candès, Zuowei Shen, A singular value thresholding algorithm for matrix completion, SIAM J. Optim. 20 (2010) 1956–1982, http://dx.doi.org/10.1137/080738970.
[48] Zhouchen Lin, Minming Chen, Yi Ma, The augmented lagrange multiplier method for exact recovery of corrupted low-rank matrices, 2010, arXiv preprint arXiv:1009.5055.
[49] Patrick L. Combettes, Valérie R. Wajs, Signal recovery by proximal forward-backward splitting, Multiscale Model. Simul. 4 (2005) 1168–1200, http://dx.doi.org/10.1137/050626090.
[50] Sotiris Diplaris, Grigorios Tsoumakas, Pericles A Mitkas, Ioannis Vlahavas, Protein classification with multiple algorithms, in: Panhellenic Conference on Informatics, Springer, 2005, pp. 448–456, http://dx.doi.org/10.1007/11573036_42.
[51] D. Turnbull, L. Barrington, D. Torres, G. Lanckriet, Semantic annotation and retrieval of music and sound effects, IEEE/ACM Trans. Audio Speech Lang. Process. 16 (2008) 467–476, http://dx.doi.org/10.1109/TASL.2007.913750.
[52] Ioannis Katakis, Grigorios Tsoumakas, Ioannis Vlahavas, Multilabel text classification for automated tag suggestion, in: Proceedings of the ECML/PKDD, vol. 18, Citeseer, 2008, p. 5, https://www.kde.cs.uni-kassel.de/wp-content/uploads/ws/rsdc08/pdf/9.pdf.
[53] Cees G.M. Snoek, Marcel Worring, Jan Van Gemert, Jan-Mark Geusebroek, Arnold W.M. Smeulders, The challenge problem for automated detection of 101 semantic concepts in multimedia, in: Proceedings of the 14th ACM International Conference on Multimedia, 2006, pp. 421–430, https://dl.acm.org/doi/10.1145/1180639.1180727.
[54] Forrest Briggs, Yonghong Huang, Raviv Raich, Konstantinos Eftaxias, Zhong Lei, William Cukierski, Sarah Frey Hadley, Adam Hadley, Matthew Betts, Xiaoli Z. Fern, et al., The 9th annual MLSP competition: New methods for acoustic classification of multiple simultaneous bird species in a noisy environment, in: 2013 IEEE International Workshop on Machine Learning for Signal Processing, MLSP, IEEE, 2013, pp. 1–8, http://dx.doi.org/10.1109/MLSP.2013.6661934.
[55] Konstantinos Trohidis, Grigorios Tsoumakas, George Kalliris, Ioannis P. Vlahavas, Multi-label classification of music into emotions, in: ISMIR, vol. 8, 2008, pp. 325–330, https://ismir2008.ismir.net/papers/ISMIR2008_275.pdf.
[56] Min-Ling Zhang, Fei Yu, Solving the partial label learning problem: An instance-based approach, in: IJCAI, 2015, pp. 4048–4054, https://dl.acm.org/doi/10.5555/2832747.2832813.
[57] Janez Demšar, Statistical comparisons of classifiers over multiple data sets, J. Mach. Learn. Res. 7 (2006) 1–30, https://jmlr.org/papers/v7/demsar06a.html.