Learning From Weakly Labeled Data Based on Manifold Regularized Sparse Model

Jia Zhang, Shaozi Li, *Senior Member, IEEE*, Min Jiang, *Senior Member, IEEE*, and Kay Chen Tan, *Fellow, IEEE*

***Abstract*—In multilabel learning, each training example is represented by a single instance, which is relevant to multiple class labels simultaneously. Generally, all relevant labels are considered to be available for labeled data. However, instances with a full label set are difficult to obtain in real-world applications, thus leading to the weakly multilabel learning problem, that is, relevant labels of training data are partially known and many relevant labels are missing, and even abundant training data are associated with an empty label set. To address the problem, we propose a new multilabel method to learn from weakly labeled data. To be specific, an optimization framework is constructed based on the manifold regularized sparse model, in which the correlations among labels and feature structure are considered to model global and local label correlations, thereby achieving discriminative feature analysis for mapping training data to ground-truth label space. Moreover, the proposed method has an excellent mechanism to conduct semisupervised multilabel learning by exploiting training data with the predicted label set of the unlabeled. Experiments on various real-world tasks reveal that the proposed method outperforms some state-of-the-art methods.**

***Index Terms*—Label correlations, missing labels, multilabel learning, semisupervised learning.**

I. Introduction

CONVENTIONAL supervised learning mainly focuses on labeling an unseen instance with an associated label. This formulation entails the restriction of each instance that is related to only one label. Actually, instances might be associated with multiple class labels simultaneously, such as automatic image annotation [1], text categorization [2], and protein function prediction [3]. Extending the conventional mechanism to train a classifier for each label independently is unwise for a large set of labels, and is not capable of handling label correlations [4], which is deemed to be important while class labels are highly correlated. To alleviate the problem, the paradigm of multilabel learning emerges and is developed to deal with multilabel data.

In the previous research for multilabel learning [5]–[9], a basic assumption is that all relevant labels of instances are provided for training. However, labels in multilabel data are often obtained via crowdsourcing or crawling webpages [10], [11]. Therefore, it is difficult to make all the relevant labels available. Instead, generally, training data are partially labeled, which leads to the challenging problem of multilabel learning with missing labels (MLMLs) [12], [13]. For example, in image annotation, labelers may only provide a few key labels to describe semantic or visual contents while the contents that are ambiguous or rare are simply neglected. As a result, the incomplete label assignment may cause errors in judgment and severely degrade the learning performance. Although a large amount of previous algorithms is well established and effective for multilabel learning, they are unable to deal with this situation well. Furthermore, in real-world applications, it is a high-cost process to obtain labeled training data, especially training data attributed to a large set of labels, thus causing only a small amount of labeled data that can be used for training [14], [15]. In the light of that labeled data may be insufficient for multilabel learning, exploiting unlabeled data appear to be significant for improving the learning performance. Semisupervised multilabel learning is a feasible solution to learn from both labeled and unlabeled data [16], [17], nevertheless, which is based on the assumption that all labeled data are associated with a full label set other than the incomplete label assignment with missing labels. Recently, many methods have been developed to conduct MLMLs [18], [19], or in a semisupervised setting [20], [21], but they may be inefficient for the joint learning from weakly labeled data.

In this article, we tackle the practical yet challenging problem, that is, semisupervised MLMLs, and propose a new multilabel method, namely, MSWL, which is designed to learn from weakly labeled data based on the manifold regularized sparse model. In order to achieve the goal, global and local label correlations (GLOCALs) are fully utilized

Manuscript received March 26, 2020; revised July 1, 2020; accepted August 5, 2020. This work was supported in part by the National Key Research and Development Program of China under Grant 2018YFC0831402; in part by the National Nature Science Foundation of China under Grant 61876159, Grant 61806172, Grant U1705286, and Grant 61876162; in part by the Fundamental Research Funds for the Central Universities, Xiamen University under Grant 20720200030; in part by the Shenzhen Scientific Research and Development Funding Program under Grant JCYJ20180307123637294; and in part by the Research Grants Council of the Hong Kong SAR under Grant CityU11202418 and Grant CityU11209219. This article was recommended by Associate Editor S. Ventura. *(Corresponding authors: Shaozi Li; Kay Chen Tan.)*

Jia Zhang and Shaozi Li are with the Department of Artificial Intelligence, Xiamen University, Xiamen 361005, China (e-mail: j.zhang@stu.xmu.edu.cn; szlig@xmu.edu.cn).

Min Jiang is with the Department of Artificial Intelligence, Xiamen University, Xiamen 361005, China, and also with the Fujian Key Laboratory of Machine Intelligence and Robotics, Xiamen University, Xiamen 361005, China (e-mail: minjiang@xmu.edu.cn).

Kay Chen Tan is with the Department of Computer Science, City University of Hong Kong, Hong Kong, and also with the City University of Hong Kong Shenzhen Research Institute, Shenzhen 518057, China (e-mail: kaytan@cityu.edu.hk).

Color versions of one or more of the figures in this article are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TCYB.2020.3015269

to improve the performance. Especially, a sparse representation method is used to reconstruct label manifold. In this way, the relationship between each label and all the other labels can be captured for global label correlations exploitation, which is also used to conduct the complement of missing labels. Moreover, by analyzing both labeled and unlabeled data, we exploit the underlying feature structure via manifold learning and preserve the similarity among different instances for local label correlations exploitation. Based on this, we design an optimization framework via joint $l_{2,1}$-norm regularization and consider learning the model in a semisupervised manner by making the prediction for unlabeled data available. Thus, discriminative features across multiple labels can be obtained by solving the proposed objective function with iterative optimization. Finally, extensive experiments on various multilabel datasets demonstrate that the proposed method is superior to some state-of-the-art methods.

In short, the major contributions of our work are threefold: 1) we propose a unified learning framework to handle weakly labeled data, which is not only capable to deal with the problem of missing labels but also can utilize unlabeled data for semisupervised learning; 2) we exploit global label correlations in a one-to-all manner via sparse representation, which are combined with local correlations to enhance the generalization performance; and 3) we design the learning model with sparsity, which helps generate a robust mapping result using discriminative features shared by multiple labels.

The remainder of this article is organized as follows. In Section II, we give a brief review of the related work on multilabel learning, MLMLs, and semisupervised multilabel learning. In Section III, we describe our proposed method in detail and give the optimization solution. Experimental results are reported in Section IV. Finally, the conclusion is given in Section V.

II. Related Work

A. Multilabel Learning

Multilabel learning has drawn a lot of attention in recent years, and comprehensive reviews on this topic are readily available in some excellent surveys [4], [22]–[24]. For multilabel learning, exploiting label correlations is a useful way to improve the learning performance [25]–[29]. According to the degree of label correlations, multilabel methods can be roughly classified into three categories, that is, first order, second order, and high order. First-order methods design one binary classifier to predict all class labels in a one-by-one style. A representative first-order example is a binary relevance (BR) approach [30], which is simple and effective, whereas it does not take label correlations into account. Second-order methods focus on modeling the relationship of label pairs. For achieving the purpose, one strategy is to model the interaction between class labels. For example, Fürnkranz *et al.* [31] proposed a calibrated label ranking (CLR) method by transforming the multilabel learning problem into a pairwise label ranking problem. The other one is to optimize the multilabel learning model involving *ranking loss*. For example, Xie *et al.* [32] proposed a multilabel consensus maximization (MLCM) method for ranking, which minimized the empirical ranking loss to generate the prediction from a set of base models. High-order methods utilize the relationship among all labels (or subsets of labels) to tackle the learning problem. Based on this, Read *et al.* [33] presented a classifier chain (CC) method for multilabel learning. By using the label vector as additional features, this method obtained a chain of binary classification for prediction. Tsoumakas and Vlahavas [34] proposed a novel problem transformation approach, that is, random *k*-labelsets (RA*k*EL), which learned from multilabel data via an ensemble of multiclass classification.

In general, the above methods are designed for global label correlations exploitation. In the light of that a label may be shared by a data subset, some researchers also consider to capture label correlations locally. For example, Huang and Zhou [35] proposed a multilabel learning method using local label correlations (ML-LOCs). The method applied clustering to the feature information enrichment with coding, and the generated code can capture labeling information of each instance to label correlations locally. Hou *et al.* [36] proposed a multilabel manifold learning (ML^2) method to model local label correlations. Based on the *smooth* assumption, the method reconstructed and exploited label manifold for multilabel learning. For a similar purpose, Zhang *et al.* [37] employed both of label space and feature space and proposed a weighted similarity method to measure the similarity between instances.

B. Multilabel Learning With Missing Labels

MLMLs focus on the issue that only a partial label set is available for learning, and many methods have been developed to solve this issue [12], [19], [38], [39]. In some of the early work, Sun *et al.* [13] presented a weak label learning (WELL) method to alleviate performance degeneration. In this method, the similarity matrix for each class label is first derived based on feature similarity. Next, following the principle that similar examples have similar class labels, these similarity matrices are utilized to guide the label recovery process. Chen *et al.* [40] proposed a FastTag method, which first tried to reconstruct incomplete label set from a few label assignment, and then learned a linear mapping to relate original features with the reconstructed label set. Xu *et al.* [41] proposed a novel Maxide method to enrich incomplete label set by learning the low-rank label correlations. Yu *et al.* [19] addressed the problem of missing labels via the low-rank empirical risk minimization for multilabel learning (LEML). Recently, Huang *et al.* [38] proposed a multilabel method with missing labels with label-specific features (LSMLs). First, the authors augmented a new supplementary label matrix by exploiting high-order label correlations and then learned a label-specific data representation for each label to construct the classifier for multilabel learning. He *et al.* [42] designed an optimization framework that can automatically capture label correlations and the shared feature subsets by all class labels, called MLMF, thus achieving the joint learning of label correlations, missing labels, and feature selection. Tan *et al.* [43] proposed a matrix completion-based

weak-label learning (McWL) method, which integrated complementary representations from multiple views to achieve the learning from weakly labeled data. Wu *et al.* [44] proposed a multilabel method using a mixed graph (ML-MG) to handle missing labels. Especially, this method has considered the label consistency between the prediction and the real value of provided labels with the constraint of the semantic hierarchy.

Furthermore, there are some other well-established manifold regularized multilabel methods to address the problem of missing labels. One typical strategy is to preserve both label correlations and instance similarity. For example, Zhu *et al.* [29] designed a multilabel method with GLOCAL. This method utilized the low rank of label space to generate a classifier output, and designed global and local manifold regularizers to exploit label correlations, thus minimizing the reconstruction error. Wu *et al.* [45] achieved MLMLs by integrating label smoothness and instance smoothness into the learning framework. In a similar way, Liu *et al.* [46] designed a modified SVM-based optimization framework with manifold regularization. Differently, the method designed a mapping function for reducing the number of samples which live in a margin area. These methods can obtain good performance, whereas they suffer from the issue of label correlations exploitation in a one-to-one manner, such as the correlations simply estimated by cosine similarity, thus limiting their effectiveness. More examples following the strategy with manifold regularization include the method based on class imbalance learning [47], [48] or graph-based method [49].

C. Semisupervised Multilabel Learning

For semisupervised multilabel learning, the modification and enhancement are mainly embodied with two strategies [50]. One is pure semisupervised multilabel learning that predicts test data by a trained model. For example, Chang *et al.* [16] proposed a convex formulation for semisupervised multilabel feature selection (CSFS). The authors designed an optimization method to predict the label set of unlabeled data and utilized the prediction with lower weight assignments for supervised multilabel learning. Liu *et al.* [51] proposed two nuclear-norm-based semisupervised multilabel methods by integrating multilabel local and global consistency. The other one is transductive multilabel (TRAM) learning, which follows the assumption that test data are from unlabeled data, and the goal is to obtain the optimal solution on the (test) data. A well-known example is the TRAM [17] method, which is one early work to complement the label set of unlabeled data via label set propagation. These methods focus on exploiting unlabeled data while labeled data are in a complete label assignment and treat missing labels indiscriminately as irrelevant ones, which results in powerless in the problem of missing labels.

To address the problem, several promising semisupervised multilabel methods have been proposed to provide the solution with missing labels. For example, Akbarnejad and Baghshah [10] presented an efficient semisupervised multilabel classifier (ESMC) that mapped feature space to label space by two sets of stochastic transformations and designed an effective probabilistic model to address the problem. Dong *et al.* [52] proposed a semisupervised weak-label (SSWL) method, which first considered to apply label similarity and instance similarity to the complement of missing labels, and then constructed an optimization framework to achieve weak-label classification by an ensemble of multiple models. However, the method is not suitable to handle large-scale data due to the high complexity of obtaining an optimal solution. Moreover, Zhao and Guo [11] proposed a semisupervised low-rank mapping (SLRM) method to the joint learning of label correlations and missing labels. Wei *et al.* [53] achieved semisupervised MLMLs by optimizing multilabel evaluation metrics, such as F_1 score and Top-k precision, and Ma and Chow [54] proposed a new multilabel method based on robust non-negative sparse graph for this task.

III. MSWL Algorithm

To formulate the problem for learning from weakly labeled data, some important notations used in this article are listed as follows. Let $\mathbf{X} = [\mathbf{x}_1; \mathbf{x}_2; \cdots; \mathbf{x}_n]$ be a training data matrix, and $\mathbf{x}_i \in \mathbb{R}^d$ $(1 \leq i \leq n)$ is the ith instance associated with a finite set of q possible labels $L = \{l_1, l_2, \ldots, l_q\}$. $\mathbf{Y} = [\mathbf{y}_1; \mathbf{y}_2; \cdots; \mathbf{y}_n] \in \{-1, 1\}^{n \times q}$ is the corresponding label matrix, where $\mathbf{y}_i = \{y_{i1}, y_{i2}, \ldots, y_{iq}\}$ is the set of the ground-truth labels of the ith instance. Assume that l_j is the jth label in L, $y_{ij} = 1$ in case that l_j is relevant to the ith instance; otherwise, $y_{ij} = -1$. Following the preset, we define the multilabel data in a weak-label learning setting. To be specific, we have a same feature representation $\mathbf{X}$ in weakly labeled data. However, the label assignment in $\mathbf{Y}$ is not complete. In this case, we define a corresponding incomplete label matrix $\mathbf{C} = [\mathbf{c}_1; \mathbf{c}_2; \cdots; \mathbf{c}_n] \in \{0, 1\}^{n \times q}$. $c_{ij} = 1$ represents that the ith instance has the jth label, and the underlying label y_{ij} must be 1. While $c_{ij} = 0$, y_{ij} is either 1 or -1. Explanatorily, $c_{ij} = 0$ and $y_{ij} = 1$ mean that the ith instance is related to the jth label but it is missing. The other is $c_{ij} = 0$ and $y_{ij} = 0$, meaning that the ith instance is not related to the jth label and it is unknown.

Based on this, we describe the proposed method for learning from weakly labeled data, as shown in Fig. 1. To be specific, training data are first input to model global and local manifold for label correlations exploitation. With the help of label correlations, the classifier is trained based on the manifold regularized sparse model under the supervision of weakly labeling information, finally, we can use the classifier to predict any unseen instance. For generating an efficient classifier, we take the aim at learning a coefficient matrix $\mathbf{W} \in \mathbb{R}^{d \times q}$. To guarantee that $\mathbf{W}$ helps to obtain excellent performance on multilabel prediction, we expect it to possess the following properties: 1) $\mathbf{W}$ can map data matrix $\mathbf{X}$ to original label matrix $\mathbf{Y}$ well; 2) the designed model has strong generalization performance to achieve MLMLs, and label correlations can be modeled to enhance the learning ability; 3) we expect that the designed model is adaptive to semisupervised multilabel learning; and 4) $\mathbf{W}$ can reveal discriminative features for mapping.

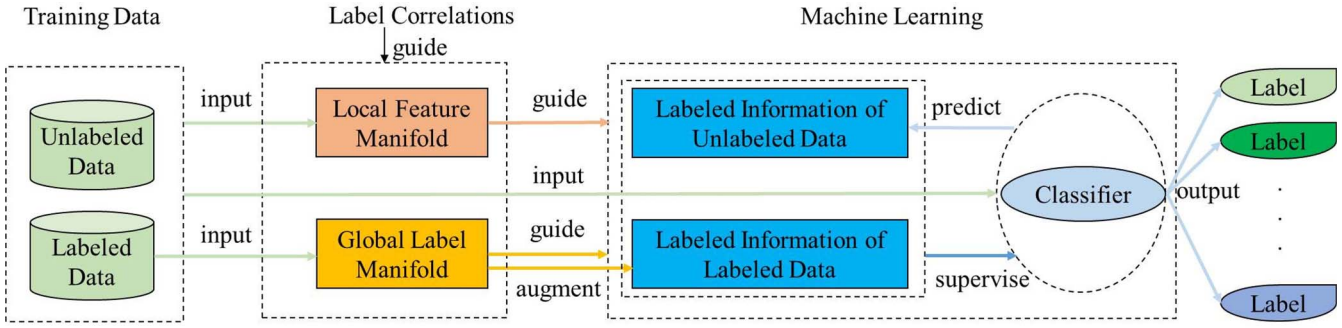

Fig. 1. Illustration of the proposed method.

A. Basic Model

For obtaining such a model for multilabel learning based on weakly labeled data, we generalize the learning problem as the following optimization formulation:

$$\min_{\mathbf{W}} \; V(\mathbf{X}, \mathbf{C}, \mathbf{W}) + \gamma \Omega(\mathbf{W}) + \mu Z(\mathbf{X}, \mathbf{C}, \mathbf{W}) \tag{1}$$

where V is defined as a loss function, Ω is to control the complexity of the model, and Z is employed to enforce the weak-label learning using label correlations information. In addition, γ and μ are parameters trading off the terms.

As we know, in weakly multilabel setting, each instance is associated with multiple class labels, whereas class labels of training instances are partially labeled. Therefore, it is impertinent to explore incomplete label space directly for multilabel learning. Following this principle, we first define the first term V, and choose the least-square loss function for training considering its simplicity and efficiency in various practical applications, as follows:

$$V(\mathbf{X}, \mathbf{C}, \mathbf{W}) = \left\| \mathbf{X}\mathbf{W} - \widetilde{\mathbf{Y}} \right\|_2^2 \tag{2}$$

where $\widetilde{\mathbf{Y}}$ is induced from incomplete label matrix $\mathbf{C}$ to approximate original label matrix $\mathbf{Y}$. For obtaining $\widetilde{\mathbf{Y}}$, we consider to enrich the existing incomplete label assignment by its neighboring labels and utilize label correlations to augment incomplete label matrix $\mathbf{C}$, as follows:

$$\widetilde{c}_{ij} = \sum_{p \in \mathcal{N}_j} c_{ip} b_{pj} \tag{3}$$

where $\widetilde{c}_{ij} \in \widetilde{\mathbf{C}}$ denotes the estimate of c_{ij} while c_{ij} is missing (or unknown), and $\widetilde{\mathbf{C}}$ is the supplementary label matrix of $\mathbf{C}$. $b_{pj} \in \mathbf{B} \in \mathbb{R}^{q \times q}$ denotes the similarity between the pth label and the jth label, and $\mathcal{N}_j$ denotes the neighbor set of the jth label. Suppose $\mathbf{B}$ is used to preserve the information of label correlations and calculated with sparsity to achieve the adaptive learning of neighboring labels (will be discussed later), we can define the supplementary label assignment in the form of matrix, that is, $\widetilde{\mathbf{C}} = \mathbf{C}(\mathbf{B} + \mathbf{I})$, in which $\mathbf{I}$ denotes the identity matrix. Based on this, an arbitrary element $\widetilde{y}_{ij} \in \widetilde{\mathbf{Y}}$ can be received, as follows:

$$\widetilde{y}_{ij} = \begin{cases} 1, & \widetilde{c}_{ij} \geq 1 \\ \widetilde{c}_{ij}, & 0 < \widetilde{c}_{ij} < 1 \\ 0, & \widetilde{c}_{ij} \leq 0. \end{cases} \tag{4}$$

In (4), an arbitrary element which is equal to 1 in $\mathbf{C}$ is reserved in $\widetilde{\mathbf{Y}}$, in the meanwhile, the underlying labels in $\mathbf{C}$ are explored for label complement.

The second term of (1) is set for reducing the model complexity. More important, in this article, we also utilize the regularizer to conduct a discriminative feature analysis. Obviously, it is beneficial with the help of sparse learning-based models imposed on $\mathbf{W}$ [55]. In consideration that $l_{2,1}$-norm regularization is robust to outliers and efficient to reveal discriminative features across multiple labels [56], we simply implement it into our optimization framework

$$\Omega(\mathbf{W}) = \|\mathbf{W}\|_{2,1}. \tag{5}$$

Equation (5) makes the optimal $\mathbf{W}$ sparse. Thus, discriminative features are noticed with large weights, while the weights of features, which are irrelevant and redundant for multilabel learning, are set to be 0 approximatively. Besides, the third term Z is the regularizer for facilitating label correlations exploitation, and the definition is described in the next section.

B. Learning Label Correlations

In this section, we focus on the exploitation and utilization of label correlations. Class labels are not completely independent with each other but have co-occurrence relationships in weakly labeled data [29], [40]. Therefore, mining such correlations may be efficient for weak-label learning, thus constructing the model with high generalization ability. Based on this, we explore the approach with manifold regularization to capture GLOCALs.

Global Label Correlations Exploitation: We conduct global label correlations exploitation based on label manifold reconstruction with a *one-to-all* style. As we know, each label theoretically can be recovered by that of its neighboring labels. For achieving the purpose, we model the relationship between one label and all the other labels with sparse representation, and employ l_1-norm regularization to obtain a sparse solution for the reconstruction of weak-label space, as follows:

$$\min_{\mathbf{b}_i} \quad \|\mathbf{C}_{-i}\mathbf{b}_i - \mathbf{c}_i\|_2^2 + \lambda \|\mathbf{b}_i\|_1 \tag{6}$$

where $\mathbf{C}_{-i}$ denotes the incomplete label matrix where the data matrix is without the ith datum, and λ is the parameter to balance the two terms. In addition, $\mathbf{B} = \{\mathbf{b}_i\}_{i=1}^q$ is the reconstruction coefficient matrix indicating the relationships among

all the labels, and each column of $\mathbf{B}$ contains the weights for reconstructing the corresponding label. Thus, global label correlations can be recovered by the reconstruction coefficient matrix $\mathbf{B}$. Moreover, as discussed before, it is note worthy that the generated matrix $\mathbf{B}$ is also suitable and helpful to the label complement while label matrix is incomplete.

Equation (6) is with a smooth convex loss function involving l_1-norm regularization. To solve the problem, the alternating direction method of multiplier (ADMM) method [57] is applied for optimization, and (6) can be solved by iteratively optimizing the following one:

$$\min_{\mathbf{b}_i,\mathbf{z}_i,\mathbf{b}_i=\mathbf{z}_i} \|\mathbf{C}_{-i}\mathbf{b}_i - \mathbf{c}_i\|_2^2 + \lambda\|\mathbf{z}_i\|_1. \tag{7}$$

According to the ADMM procedure, the optimization of (7) can be transformed into a series of unconstrained minimization problems with augmented Lagrangian function, and its Lagrangian arrives at

$$\begin{aligned} L_0(\mathbf{b}_i, \mathbf{z}_i, \mathbf{t}_i) = &\frac{1}{2}\|\mathbf{C}_{-i}\mathbf{b}_i - \mathbf{c}_i\|_2^2 + \lambda\|\mathbf{z}_i\|_1 \\ &+ \mathbf{t}_i^T(\mathbf{b}_i - \mathbf{z}_i) + \frac{\rho}{2}\|\mathbf{b}_i - \mathbf{z}_i\|_2^2 \end{aligned} \tag{8}$$

where $\mathbf{t}_i$ is the Lagrange multiplier and ρ is the penalty parameter. Minimizing (8) jointly with respect to $\mathbf{b}_i$, $\mathbf{z}_i$, and $\mathbf{t}_i$ is a challenging problem, so we alternately update one variable while the other variables are fixed

$$\begin{cases} \mathbf{b}_i^{(k+1)} = \left(\mathbf{C}_{-i}^T\mathbf{C}_{-i} + \rho\mathbf{I}\right)^{-1}\left(\mathbf{C}_{-i}^T\mathbf{c}_i + \rho\mathbf{z}_i^{(k)} - \mathbf{t}_i^{(k)}\right) \\ \mathbf{z}_i^{(k+1)} = \left(\mathbf{r}_i^{(k+1)} - \lambda/\rho\right)_+ - \left(-\mathbf{r}_i^{(k+1)} - \lambda/\rho\right)_+ \\ \mathbf{t}_i^{(k+1)} = \mathbf{t}_i^{(k)} + \rho\left(\mathbf{b}_i^{(k+1)} - \mathbf{z}_i^{(k+1)}\right) \end{cases} \tag{9}$$

where $\mathbf{r}_i^{(k+1)} = \mathbf{b}_i^{(k+1)} + \mathbf{t}_i^{(k)}/\rho$. The ADMM solution can be received by repeating the above three steps. By generating $\mathbf{b}_i$ $(1 \le i \le q)$, reconstruction coefficient matrix $\mathbf{B}$ can be obtained with zero diagonal elements.

With the help of the reconstruction coefficient matrix $\mathbf{B}$, we continue to learn from the underlying structure of label space. In the view that the model is trained under the supervision of $\widetilde{\mathbf{Y}}$, as shown in (2), which easily leads to that noisy data are involved in the supervision, thus causing a biased classifier induction. Therefore, we generalize the manifold regularizer with $\mathbf{W}$ other than $\mathbf{XW}$, and can model the approximation of label manifold by minimizing the following term:

$$\sum_{i=1}^{q}\left\|\mathbf{w}_{\cdot i} - \sum_{j\neq i}\mathbf{w}_{\cdot j}b_{ji}\right\|^2 \tag{10}$$

where $\mathbf{w}_{\cdot i}$ is the ith column of $\mathbf{W}$. In this way, the underlying structure in label space holds in $\mathbf{W}$ by exploiting the relationships among all the labels.

Local Label Correlations Exploitation: According to the assumption of locally linear embedding (LLE) [36], [58], each instance can be reconstructed by the features of its neighboring instances on feature manifold. Similarly, we characterize the manifold regularizer induced from feature space, thus exploiting local label correlations. Suppose the weight matrix is denoted as $\mathbf{S} \in \mathbb{R}^{n\times n}$, the reconstruction error of feature space can be represented by the following form:

$$\min_{\mathbf{S}} \quad \sum_{i=1}^{n}\left\|\mathbf{x}_i - \sum_{j=1}^{K}s_{ji}\mathbf{x}_j\right\|^2 \tag{11}$$

where K is the number of neighbors, and $s_{ji} \in \mathbf{S}$ is the similarity of $\mathbf{x}_i$ with its neighbor $\mathbf{x}_j$, and $\sum_{j=1}^{K} s_{ji} = 1$. Note that $s_{ji} = 0$ if $\mathbf{x}_j$ is not one of $\mathbf{x}_i$'s K-nearest neighbors. In consideration that similar instances may share a same label, we transform reconstruction coefficient matrix $\mathbf{S}$ from $\mathbf{X}$ to $\mathbf{XW}$, that is, the output of the designed model, we can infer that the output space also resides on feature manifold. Thus, it leads to the minimization of the term

$$\sum_{i=1}^{n}\left\|\left(\mathbf{x}_i - \sum_{j=1}^{K}s_{ji}\mathbf{x}_j\right)\mathbf{W}\right\|^2. \tag{12}$$

Equation (12) returns the approximation of feature manifold to build the regularizer of the output space. In this way, local label correlations can be learned to obtain a more robust predicted result.

Accordingly, the third term Z as shown in (1) can be well defined by the combination of (10) and (12). Considering that both of $\mathbf{W}$ and $\mathbf{S}$ are sparse matrices, we can rewrite Z in a more concise form, as follows:

$$Z(\mathbf{X}, \mathbf{C}, \mathbf{W}) = \alpha\|\mathbf{W} - \mathbf{WB}\|_F^2 + \beta\|\mathbf{XW} - \mathbf{SXW}\|_F^2 \tag{13}$$

where α and β are two tradeoff parameters. By using global and local manifold regulations, label correlations are exploited to enhance the ability of multilabel learning with weakly labeled data.

C. Extension to Semisupervised Learning

As discussed, by plugging (2), (5), and (13) into (1), we propose the following objective function that can achieve multilabel learning with weakly labeled data:

$$\begin{aligned} \min_{\mathbf{W}} \ L_1(\mathbf{W}) = &\left\|\mathbf{XW} - \widetilde{\mathbf{Y}}\right\|_2^2 + \alpha\|\mathbf{W} - \mathbf{WB}\|_F^2 \\ &+ \beta\|\mathbf{XW} - \mathbf{SXW}\|_F^2 + \gamma\|\mathbf{W}\|_{2,1}. \end{aligned} \tag{14}$$

Furthermore, we extend the proposed weak-label learning method applicable in semisupervised setting. Inspired by [16] and [59], we focus on the solution for generating a predicted label matrix for labeled and unlabeled data, and involving the predicted label matrix in the model training. Let the first m data be labeled $(m < n)$, and the remaining $u(= n-m)$ data be unlabeled, we define $\mathbf{F} = [\mathbf{F}_m; \mathbf{F}_u] \in \mathbb{R}^{n\times q}$, in which $\mathbf{F}_m = \widetilde{\mathbf{Y}}_m$ for labeled data, and $\mathbf{F}_u$ denotes the predicted label matrix for unlabeled data. For avoiding a trivial solution, we allow the value of any element in $\mathbf{F}$ to vary between 0 and 1. Then, our objective function can be transformed into the following one:

$$\begin{aligned} \min_{\mathbf{W},\mathbf{F},\mathbf{F}_m=\widetilde{\mathbf{Y}}_m} \ L_2(\mathbf{W}, \mathbf{F}) = &tr\left((\mathbf{XW} - \mathbf{F})^T\mathbf{U}(\mathbf{XW} - \mathbf{F})\right) \\ &+ \alpha\|\mathbf{W} - \mathbf{WB}\|_F^2 + \beta\|\mathbf{XW} - \mathbf{SXW}\|_F^2 \\ &+ \gamma\|\mathbf{W}\|_{2,1} \end{aligned} \tag{15}$$

where $tr(.)$ denotes the trace operator and $\mathbf{U}$ is the diagonal matrix. Empirically, we set $\mathbf{U}$ with the value of the first m diagonal elements is larger than the rest ones.

Compared with (14) and (15) has two significant properties as follows: 1) it includes an improved loss function to generate the predicted result for unlabeled data, which makes the proposed method capable to conduct semisupervised MLMLs and 2) it uses both unlabeled and labeled data to construct feature manifold regularizer for learning local label correlations. It is theoretically proved that the proposed method not only has the advantage in the case that relevant labels of training data are partially known but also is suitable to handle weakly labeled data in a semisupervised manner.

D. Solution

The proposed objective function in (15) involves two unknown variables, that is, $\mathbf{W}$ and $\mathbf{F}$. Here, we introduce an alternating iterative algorithm to obtain the optimal solution.

We first solve $\mathbf{W}$ while $\mathbf{F}$ is fixed. In the light that the optimization problem with respect to $\mathbf{W}$ is the combination of multiple convex functions, especially, the convergence of optimizing $\|\mathbf{W}\|_{2,1}$ has been well studied and proved in [56], then we can easily solve $\mathbf{W}$ by requiring $\partial L_2/\partial \mathbf{W}$ to 0. With some algebraic steps, we have

$$\frac{\partial L_2}{\partial \mathbf{W}} = \big(\mathbf{X}^T\mathbf{U}\mathbf{X} + \beta(\mathbf{X}-\mathbf{S}\mathbf{X})^T(\mathbf{X}-\mathbf{S}\mathbf{X}) + \gamma\mathbf{Q}\big)\mathbf{W} + \alpha\mathbf{W}(\mathbf{I}-\mathbf{B})(\mathbf{I}-\mathbf{B})^T - \mathbf{X}^T\mathbf{U}\mathbf{F} \tag{16}$$

where $\mathbf{Q}$ is the diagonal matrix which is defined as follows:

$$\mathbf{Q} = \begin{bmatrix} \frac{1}{2\|\mathbf{w}_1\|_2} & & \\ & \ddots & \\ & & \frac{1}{2\|\mathbf{w}_d\|_2} \end{bmatrix}. \tag{17}$$

In (17), we can observe that $\mathbf{Q}$ is related to $\mathbf{W}$, thus difficult to solve for $\mathbf{W}$ directly. To address the problem, an iterative method is presented to solve it. To be specific, we obtain $\mathbf{Q}$ with randomly initialized $\mathbf{W}$ and obtain the optimal solution until the iterative optimization converges. Suppose $\mathbf{Q}$ is prepared in the kth iteration, that is, $\mathbf{Q}^{(k)}$, $\mathbf{W}^{(k+1)}$ can be received with the following equation:

$$\mathbf{A}^{(k)}\mathbf{W}^{(k+1)} + \mathbf{W}^{(k+1)}\mathbf{D} = \mathbf{E} \tag{18}$$

where $\mathbf{A}^{(k)} = \mathbf{X}^T\mathbf{U}\mathbf{X} + \beta(\mathbf{X}-\mathbf{S}\mathbf{X})^T(\mathbf{X}-\mathbf{S}\mathbf{X}) + \gamma\mathbf{Q}^{(k)}$, $\mathbf{D} = \alpha(\mathbf{I}-\mathbf{B})(\mathbf{I}-\mathbf{B})^T$, and $\mathbf{E} = \mathbf{X}^T\mathbf{U}\mathbf{F}$. For solving this equation, several existing methods [45], [60], [61] can be employed to obtain $\mathbf{W}$, and we use the *Lyapunov* function[1] in MATLAB to solve the mathematical problem.

After updating the value of $\mathbf{W}$, another variable, that is, predicted label matrix $\mathbf{F}$, can be estimated by computing $\widetilde{\mathbf{F}} = \mathbf{X}\mathbf{W}$. Considering that the value of arbitrary element $f_{ij} \in \mathbf{F}$ is set in the range of 0–1 and $\mathbf{F}_m = \widetilde{\mathbf{Y}}_m$ for labeled data, we adjust the value of $\widetilde{\mathbf{F}}$ to obtain the update of $\mathbf{F}_u$ for unlabeled data, and define the formula as follows:

$$f_{ij} = \begin{cases} 1, & \widetilde{f}_{ij} \geq 1 \\ \widetilde{f}_{ij}, & 0 < \widetilde{f}_{ij} < 1 \\ 0, & \widetilde{f}_{ij} \leq 0. \end{cases} \tag{19}$$

In brief, the solution is designed by executing the iterative process for the unknown variables, and the pseudocode of the proposed algorithm is shown in Algorithm 1. As the objective value of (15) converges to a fixed point, the underlying relevant labels of unlabeled data can be learned to train the model (namely, the coefficient matrix W), thus facilitating the mapping of test data.

Algorithm 1 MSWL Algorithm

Input: Train data $\{(\mathbf{x}_i, \mathbf{c}_i)|1 \leq i \leq n\}$, parameters α, β, γ, λ, and the number of neighbors K;
Output: Mapping matrix $\mathbf{W}$, predicted label matrix $\mathbf{F}$.

Initialize $\mathbf{C}_m \leftarrow [\mathbf{c}_1; \mathbf{c}_2; ...; \mathbf{c}_m](m < n)$, and $\mathbf{C}_u \leftarrow [\mathbf{c}_{m+1}; \mathbf{c}_{m+2}; ...; \mathbf{c}_n]$ in which any element is set to 0;
for $i = 1 \rightarrow q$ **do**
  Compute $\mathbf{b}_i$ by optimizing (7);
end for
Compute supplementary label matrix $\widetilde{\mathbf{Y}}_m$ of the first m train data by (4);
Compute weight matrix $\mathbf{S}$ by optimizing (12);
$k \leftarrow 0$. Initialize $\mathbf{W}^{(k)}$ randomly, and $\mathbf{F}^{(k)} \leftarrow [\widetilde{\mathbf{Y}}_m; \mathbf{C}_u]$;
repeat
  Compute d-dimensional diagonal matrix $\mathbf{Q}^{(k)}$ as:

$$\mathbf{Q}^{(k)} \leftarrow \begin{bmatrix} \frac{1}{2\|\mathbf{w}_1^{(k)}\|_2} & & \\ & \ddots & \\ & & \frac{1}{2\|\mathbf{w}_d^{(k)}\|_2} \end{bmatrix};$$

  Update the optimal $\mathbf{W}^{(k+1)}$ by solving (18);
  Compute $\widetilde{\mathbf{F}}^{(k+1)} \leftarrow \mathbf{X}\mathbf{W}^{(k+1)}$;
  Adjust $\mathbf{F}_u^{(k+1)}$ by (19), and $\mathbf{F}^{(k+1)} \leftarrow [\widetilde{\mathbf{Y}}_m; \mathbf{F}_u^{(k+1)}]$;
  $k \leftarrow k+1$;
until *convergence*;
return $\mathbf{W} \leftarrow \mathbf{W}^{(k+1)}$, $\mathbf{F} \leftarrow \mathbf{F}^{(k+1)}$.

IV. Experiments

A. Experimental Datasets

We use a total of 11 benchmark datasets to evaluate the performance of the proposed method. Here, a brief introduction is given for these datasets. Corel5k and ESP Game are two widely used datasets for automatic image annotation.[2] As the literature suggested [46], the instances with few labels and the classes attributed few samples are deleted in our setting. Especially, the images, which have more than two positive labels with more than 100 samples per class, are selected from the original Corel5k dataset [62], and the images, which have more than five positive labels with more

[1] https://www.mathworks.com/help/control/ref/lyap.html

[2] http://lear.inrialpes.fr/people/guillaumin/data.php

TABLE I
EXPERIMENTAL DATASETS

Data set	#Instance	#Feature	#Label	Label card.	Label dens.
Corel5k	4396	1000	41	2.287	0.056
Corel16k1	13766	500	153	2.859	0.019
Corel16k2	13761	500	164	2.882	0.018
ESP Game	6775	1000	34	3.810	0.112
Bibtex	7395	1836	159	2.402	0.015
Arts	5000	462	26	1.636	0.063
Business	5000	438	30	1.588	0.053
Education	5000	550	33	1.461	0.044
Science	5000	743	40	1.451	0.036
Stackexchess	1675	585	227	2.411	0.011
Yeast	2417	103	14	4.238	0.303

than 300 samples per class, are selected from the original ESP Game dataset [63]. Moreover, we use 15 different visual descriptors, including one Gist descriptor, six global color histograms, and eight local bag-of-visual-words features, to extract features for these image datasets, and then apply a feature mapping method for nonlinear feature transformation as the literature suggested [40]. Finally, in the light of heavy computational burden, we use random projection [64] to reduce the dimensionality and characterize each image in these datasets by 1000-D feature vectors. In addition, Corel16k1 and Corel16k2[3] [1] are also used as benchmark datasets for image classification task. On text, Bibtex[4] [2], Stackexchess[4] [65], and four Yahoo datasets[5] [66], that is, Arts, Business, Education, and Science, are used in the experiments. On biology, the dataset of Yeast[4] [3] is used to recognize gene functional groups of yeast genes.

The characteristics of all the aforementioned datasets are summarized in Table I, where "#instance," "#feature," and "#label" denote the number of instances, the number of features, and the number of labels, respectively. Label cardinality (label card.) measures the average number of labels assigned to each instance, and label density (label dens.) is defined as cardinality divided by the number of labels.

B. Experimental Evaluation Metrics

We employ four widely used rank-based metrics in the experiment, including *area under the ROC curve*, *coverage*, *ranking loss*, and *average precision* [4], [12]. Let $\{(\mathbf{x}_i, \mathbf{y}_i)|1 \leq i \leq t\}$ be a test set, and $\{\mathbf{f}_1, \mathbf{f}_2, \ldots, \mathbf{f}_q\}$ be the classifier set to predict the relevant class labels of unseen instance $\mathbf{x}_i$, the definition of these metrics is shown as follows.

Area Under the ROC Curve (AUC): The metric evaluates the variation between true positive rate and false positive rate [12]. To be specific, we first rank all classes in the descending order of their scores, and then vary the number of predicted class labels from 1 to q and obtain the ROC curve [67] by calculating true positive rate and false positive rate for each number of predicted class labels. Finally, we estimate the area under the ROC curve as the final result.

Coverage: The metric evaluates how many steps are needed, on average, to go down the label ranking list so as to cover all the ground-truth labels of the instance. Suppose $\text{rank}(\mathbf{x}_i, l_k) = \sum_{j=1}^{q} [[\mathbf{f}_j(\mathbf{x}_i) \geq \mathbf{f}_k(\mathbf{x}_i)]]$ returns the rank of l_k when all labels in L are sorted in descending order based on q classifiers

$$\texttt{CV} = \frac{1}{t}\sum_{i=1}^{t} \max_{l_k \in \mathbf{y}_i} \text{rank}(\mathbf{x}_i, l_k) - 1. \tag{20}$$

Ranking Loss: The metric evaluates the fraction of reversely ordered label pairs. Suppose $D_i = \{(l_j, l_k)|\mathbf{f}_j(\mathbf{x}_i) \leq \mathbf{f}_k(\mathbf{x}_i), (l_j, l_k) \in \mathbf{y}_i \times \bar{\mathbf{y}}_i\}$, and $\bar{\mathbf{y}}_i$ is the complementary set of $\mathbf{y}_i$ in L

$$\texttt{RL} = \frac{1}{t}\sum_{i=1}^{t} \frac{|D_i|}{|\mathbf{y}_i||\bar{\mathbf{y}}_i|}. \tag{21}$$

Average Precision: The metric evaluates the average fraction of relevant labels ranked higher than a particular label $l_k \in \mathbf{y}_i$. Suppose $L_i = \{l_j|\text{rank}(\mathbf{x}_i, l_j) \leq \text{rank}(\mathbf{x}_i, l_k)\}$

$$\texttt{AP} = \frac{1}{t}\sum_{i=1}^{t} \frac{1}{|\mathbf{y}_i|} \sum_{l_j, l_k \in \mathbf{y}_i} \frac{|L_i|}{\text{rank}(\mathbf{x}_i, l_k)}. \tag{22}$$

These metrics evaluate the performance of multilabel algorithms from various aspects. For *AUC* and *average precision*, the larger the values the better the performance. For the other metrics, the smaller the values the better the performance.

C. Comparative Studies

We compare the proposed method MSWL with the following state-of-the-art algorithms.

MSWL (Proposed Method[6]): The parameters α and β are searched in $\{10^{-3}, 10^{-1}, \ldots, 10^{3}\}$, and γ is searched in $\{10^{-6}, 10^{-5}, \ldots, 10^{6}\}$. λ is simply set to 0.1, and the number of neighbors K is set to $q+1$, where q is the number of labels.

ML^{27} *[36]:* Promising multilabel learning approach by exploiting label manifold, which often outperforms other existing multilabel methods.

MLML [45]: State-of-the-art multilabel method with missing labels, which formulates the problem via considering label consistency and label smoothness. MLML, in this article, only denotes MLML-exact, and the parameters $\alpha_{\mathbf{X}}$ and $\alpha_{\mathbf{C}}$ are tuned in {0.01:0.1:0.81}.

ESMC[8] *[10]:* ESMC is capable of handling missing labels. The parameter $B > 0$, and λ, ρ, and σ_z are tuned in $\{10^1, 10^2, \ldots, 10^5\}$.

GLOCAL[9] *[29]:* Recently, superior approach to perform multilabel learning with GLOCALs. The parameters $\lambda = 1$ and $\lambda_2 = 10^{-3}$, λ_3 and λ_4 are tuned in $\{10^{-4}, 10^{-3}, \ldots, 1\}$, k is tuned in $\{5, 10, 15, 20, 30\}$, and g is varied from $\{4, 8, 16, 32, 64\}$.

CSFS [16]: A linear model is trained on both labeled and unlabeled data, and $l_{2,1}$-norm regularization is utilized for robust multilabel feature selection. The parameter μ is searched in $\{10^{-6}, 10^{-2}, \ldots, 10^{6}\}$.

[3]http://mulan.sourceforge.net/datasets-mlc.html
[4]http://www.uco.es/kdis/mllresources/
[5]http://www.lamda.nju.edu.cn/files/MDDM-expdata.rar
[6]https://jiazhang-ml.pub/MSWL-master.zip
[7]http://palm.seu.edu.cn/zhangml/files/ML2.zip
[8]https://github.com/Akbarnejad/ESMC_Implementation
[9]http://www.lamda.nju.edu.cn/files/Glocal.zip

TABLE II
COMPARISON RESULTS OF MSWL ON THE 11 DATASETS AGAINST OTHER COMPARING METHODS WHILE THE INCOMPLETE LABEL RATIO IS 30%

Method	*AUC*↑											Ave. Rank.
	Corel5k	Corel16k1	Corel16k2	ESP Game	Bibtex	Arts	Business	Education	Science	Stackexchess	Yeast	
MSWL	**0.880**	**0.858**	**0.865**	**0.791**	**0.920**	**0.845**	**0.944**	**0.905**	**0.857**	**0.892**	0.816	**1.091**
ML^2	0.874	0.818	0.827	0.759	0.913	0.818	0.935	0.874	0.847	0.855	0.792	3.818
MLML	0.838	0.724	0.730	0.741	0.753	0.799	0.883	0.861	0.803	0.725	**0.817**	4.636
GLOCAL	0.864	0.831	0.840	0.767	0.887	0.831	0.940	0.892	0.845	0.883	0.816	3.273
ESMC	0.879	0.845	0.852	0.783	0.918	0.842	0.942	0.902	0.852	0.884	0.810	2.182
Method	*Coverage*↓											Ave. Rank.
	Corel5k	Corel16k1	Corel16k2	ESP Game	Bibtex	Arts	Business	Education	Science	Stackexchess	Yeast	
MSWL	**8.599**	**41.424**	**43.164**	**14.703**	**22.354**	**4.756**	**2.196**	**3.692**	**6.320**	**47.342**	**6.244**	**1.000**
ML^2	9.082	53.155	57.176	16.650	24.359	5.528	2.546	4.931	6.850	58.640	6.736	3.864
MLML	11.262	74.964	79.779	17.332	59.926	6.130	3.101	5.418	8.751	100.848	6.340	4.727
GLOCAL	10.008	48.935	50.867	16.277	30.504	5.205	2.313	4.161	6.850	49.805	6.377	3.227
ESMC	8.672	45.549	47.728	14.912	22.696	4.837	2.228	3.789	6.361	48.971	6.412	2.182
Method	*Ranking loss*↓											Ave. Rank.
	Corel5k	Corel16k1	Corel16k2	ESP Game	Bibtex	Arts	Business	Education	Science	Stackexchess	Yeast	
MSWL	**0.115**	**0.140**	**0.136**	**0.204**	**0.075**	**0.119**	**0.035**	**0.080**	**0.117**	**0.109**	0.170	**1.273**
ML^2	0.120	0.179	0.174	0.233	0.079	0.142	0.042	0.103	0.123	0.137	0.194	3.818
MLML	0.157	0.272	0.268	0.248	0.240	0.169	0.069	0.122	0.168	0.268	**0.169**	4.636
GLOCAL	0.133	0.166	0.160	0.228	0.116	0.136	0.041	0.094	0.125	0.114	0.170	3.227
ESMC	**0.115**	0.152	0.148	0.208	0.076	0.126	0.038	0.084	0.117	**0.109**	0.176	2.046
Method	*Average precision*↑											Ave. Rank.
	Corel5k	Corel16k1	Corel16k2	ESP Game	Bibtex	Arts	Business	Education	Science	Stackexchess	Yeast	
MSWL	**0.554**	**0.322**	**0.319**	**0.475**	**0.595**	**0.619**	**0.884**	**0.622**	**0.593**	**0.466**	0.759	**1.273**
ML^2	**0.554**	0.308	0.331	0.458	0.593	0.594	0.879	0.619	0.591	0.436	0.722	2.909
MLML	0.521	0.239	0.233	0.443	0.351	0.565	0.832	0.583	0.522	0.231	**0.762**	4.682
GLOCAL	0.533	0.319	0.316	0.457	0.472	0.608	0.882	0.620	**0.593**	0.436	**0.762**	2.591
ESMC	0.545	0.306	0.301	0.454	0.555	0.608	0.880	0.619	0.584	0.406	0.744	3.546

TRAM[10] *[17]:* TRAM method via label set propagation. The parameter $k = 10$ as the default setting.

Among these comparing methods, MLML, ESMC, and GLOCAL are capable for MLMLs, in which ESMC and GLOCAL are recent studies to achieve the purpose. Note that ESMC not only can deal with missing labels but also performs well in semisupervised setting. CSFS and TRAM are two representative methods for semisupervised multilabel learning, which are widely used as comparing methods for performance evaluation [53], [59]. In addition, the parameter of each method is set according to the corresponding reference suggested. For parameter tuning, we adopt a grid-search strategy to search for the optimal parameter, which is determined by making the performance on test data best.

We employ five-fold cross-validation to conduct the experiment. Specifically, we randomly divide each dataset into five uniform folds, each fold is held-out in turn for test, while the remaining data are merged for training. As the validation is iterated five times, we obtain five results. As the reference suggested [46], for large-scale datasets,[11] we repeat the process two runs to obtain different partitions, and finally calculate the average result of ten results. For regular-scale datasets, we perform five-fold cross-validation six times and obtain 30 different results to calculate the average result.

D. Multilabel Learning With Missing Labels

In this section, the comparison experiment is conducted for MLMLs, and different incomplete label ratios are considered by varying portions of missing labels on training data, that is, 30% and 70%. For the setting of comparing methods, ESMC, GLOCAL, and MLML are selected in the light that they have the ability to deal with missing labels. In addition, ML^2 is a state-of-the-art multilabel learning method, and we treat ML^2 as a baseline. The detailed experimental result is shown in Tables II and III, where the best result among all the algorithms on each dataset is highlighted in boldface. Due to space limit, we report the standard deviation on the Web.[12]

Table II shows the comparison result of the proposed method with comparing algorithms, that is, ESMC, GLOCAL, MLML, and ML^2, on 11 datasets while the incomplete label ratio is 30%. From Table II, we can see that ESMC, GLOCAL, and MLML have their own strength in managing different real-world tasks while the proposed method has the best result regarding average ranking (Ave. Rank.), and can achieve better performance in most cases. To be specific, the proposed method is superior to these comparing methods on 10 out of 11 datasets except on Yeast. On the Yeast dataset, MLML has the best result except on *coverage*, and the proposed method and GLOCAL have good performance with the comparable result. In addition, we can see from Table II that the proposed method can obtain the best performance on some special datasets with tail labels, such as Bibtex and Stackexchess. Therefore, we can conclude that the proposed method is effective in the real scenario. Moreover, ML^2 has no explicit mechanism to tackle the problem of missing labels, but it can obtain good performance in some cases. For example, on Corel5k, Corel16k1, Corel16k2, ESP Game, and Bibtex

[10] http://www.lamda.nju.edu.cn/files/TRAM.zip

[11] In multilabel learning, a large-scale dataset usually means the data whose number of instances is not less than 5000 [36], [68].

[12] http://jiazhang-ml.pub/Supplement-MSWL.pdf

TABLE III
COMPARISON RESULTS OF MSWL ON THE 11 DATASETS AGAINST OTHER COMPARING METHODS WHILE THE INCOMPLETE LABEL RATIO IS 70%

Method	*AUC*↑											Ave. Rank.
	Corel5k	Corel16k1	Corel16k2	ESP Game	Bibtex	Arts	Business	Education	Science	Stackexchess	Yeast	
MSWL	**0.858**	**0.847**	**0.857**	**0.779**	**0.887**	**0.834**	**0.936**	**0.891**	**0.855**	**0.852**	**0.812**	**1.091**
ML^2	0.811	0.773	0.746	0.706	0.873	0.776	0.908	0.833	0.796	0.817	0.773	4.000
MLML	0.768	0.652	0.652	0.695	0.685	0.740	0.873	0.797	0.740	0.646	0.795	4.909
GLOCAL	0.840	0.823	0.829	0.749	0.848	0.829	0.934	0.890	0.840	0.850	**0.812**	2.591
ESMC	0.844	0.829	0.835	0.764	0.883	0.825	0.933	**0.891**	0.839	0.841	0.807	2.409
Method	*Coverage*↓											Ave. Rank.
	Corel5k	Corel16k1	Corel16k2	ESP Game	Bibtex	Arts	Business	Education	Science	Stackexchess	Yeast	
MSWL	**10.033**	**45.355**	**46.374**	**15.461**	**31.301**	**5.113**	**2.482**	**4.232**	**6.467**	**62.412**	**6.433**	**1.000**
ML^2	12.949	65.491	79.429	19.380	34.944	6.912	3.549	6.584	9.317	75.538	7.338	4.000
MLML	15.437	88.809	96.088	19.791	72.974	7.890	4.236	7.827	11.533	119.294	6.972	4.909
GLOCAL	11.442	51.313	54.665	17.195	40.052	5.295	2.492	4.296	7.083	63.141	6.446	2.636
ESMC	10.904	50.522	53.422	15.968	31.675	5.355	2.587	4.252	7.235	64.940	6.584	2.455
Method	*Ranking loss*↓											Ave. Rank.
	Corel5k	Corel16k1	Corel16k2	ESP Game	Bibtex	Arts	Business	Education	Science	Stackexchess	Yeast	
MSWL	**0.138**	**0.151**	**0.144**	**0.217**	**0.106**	**0.138**	**0.043**	**0.092**	**0.120**	0.152	0.176	**1.273**
ML^2	0.183	0.224	0.254	0.289	0.117	0.185	0.062	0.144	0.176	0.181	0.216	4.000
MLML	0.226	0.345	0.349	0.296	0.306	0.227	0.076	0.184	0.231	0.349	0.193	4.909
GLOCAL	0.158	0.173	0.169	0.246	0.155	0.140	**0.043**	0.095	0.133	**0.142**	**0.174**	2.455
ESMC	0.150	0.168	0.166	0.230	0.110	0.145	0.046	**0.092**	0.136	0.149	0.181	2.364
Method	*Average precision*↑											Ave. Rank.
	Corel5k	Corel16k1	Corel16k2	ESP Game	Bibtex	Arts	Business	Education	Science	Stackexchess	Yeast	
MSWL	**0.513**	**0.317**	**0.312**	**0.460**	**0.521**	0.575	**0.875**	**0.596**	**0.547**	**0.397**	0.748	**1.182**
ML^2	0.478	0.260	0.269	0.403	0.519	0.544	0.837	0.570	0.517	0.386	0.704	3.727
MLML	0.439	0.177	0.170	0.394	0.272	0.494	0.792	0.501	0.436	0.150	0.738	4.909
GLOCAL	0.493	0.298	0.292	0.429	0.418	**0.582**	0.874	0.594	**0.547**	0.371	**0.756**	2.455
ESMC	0.491	0.300	0.296	0.435	0.511	0.553	0.871	0.584	0.527	0.372	0.742	2.727

datasets, ML^2 compares favourably with ESMC, GLOCAL, and MLML regarding *average precision*.

We continue to compare all the algorithms while the incomplete label ratio is 70%, as shown in Table III. From Table III, we can come to a similar conclusion with Table II, and the proposed method has the advantage regarding MLMLs. Compared with these methods, the proposed method almost can obtain better performance on all the datasets, and GLOCAL has the best result on several datasets in terms of *average precision* and *ranking loss*, such as Arts, Stackexchess, and Yeast. Moreover, ML^2 has good performance in some cases while the incomplete label ratio is 30%, but it has an unsatisfactory result while the incomplete label ratio is larger. This suggests that it is crucial to deal with missing labels while the label assignment is incomplete.

Based on these observations, we can conclude that MSWL is effective in handling multilabel data with missing labels.

To further analyze the performance among all the methods, the performance of each method is recorded while the incomplete label ratio is 30%, and the test [69] is used as the favorable statistical significance test for method comparison on 11 datasets. Table IV illustrates the Friedman statistic F_F and the corresponding critical value on each metric, and we can see that the null hypothesis, which follows the principle that all the methods have equal performance, is clearly rejected in terms of each metric at significance level $\alpha = 0.05$. Thus, the Bonferroni–Dunn test [69] is utilized to complete the performance analysis. Here, MSWL is regarded as the control method whose average rank difference against the comparing method is calibrated with the critical difference (CD). Accordingly, MSWL is deemed to have significantly different performance to one comparing method if their average ranks differ by at least one CD ($CD = 1.6842$ in this article: # comparing algorithms $k = 5$, # datasets $N = 11$).

TABLE IV
SUMMARY OF THE FRIEDMAN STATISTICS F_F ($k = 5, N = 11$) AND THE CRITICAL VALUE ON EVALUATION METRICS (k: COMPARING METHODS AND N: DATASETS)

Evaluation metric	F_F	critical value($\alpha = 0.05$)
AUC	34.1606	2.61
Coverage	54.5333	
Ranking loss	26.9466	
Average precision	16.9188	

Fig. 2 shows the CD diagrams [69] with respect to each metric. Especially, any comparing method whose average rank is within one CD to that of MSWL is connected. Otherwise, the method, which is not connected with MSWL, is considered to have a significant different performance with the control method. From Fig. 2, we can see that MSWL significantly performs better compared with MLML. Compared with ESMC, MSWL has significantly better performance on *average precision*, which also significantly outperforms GLOCAL and ML^2 on multiple metrics respectively. Thus, we conclude that the proposed method can achieve highly competitive performance against these comparing methods.

E. Semisupervised Multilabel Learning

Next, we evaluate the performance of the proposed method of semisupervised multilabel learning. We randomly split 10% of training data as labeled data and other as unlabeled data. Note that ESMC, GLOCAL, TRAM, and CSFS are employed as comparing methods, in which ESMC, TRAM,

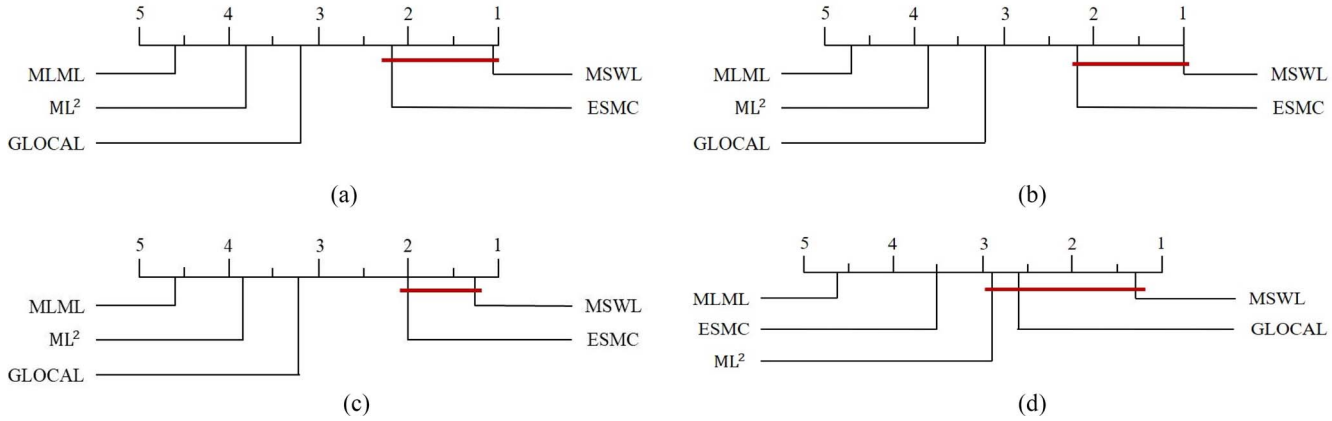

Fig. 2. Comparison of MSWL (control method) against other comparing methods in terms of MLMLs using the Bonferroni–Dunn test (*CD* = 1.6842 at 0.05 significance level). (a) *AUC*. (b) *Coverage*. (c) *Ranking loss*. (d) *Average precision*.

TABLE V
COMPARISON RESULTS OF MSWL ON THE 11 DATASETS AGAINST OTHER COMPARING METHODS IN SEMISUPERVISED SETTING

Method	*AUC*↓											Ave. Rank.
	Corel5k	Corel16k1	Corel16k2	ESP Game	Bibtex	Arts	Business	Education	Science	Stackexchess	Yeast	
MSWL	0.812	**0.835**	**0.837**	**0.766**	**0.859**	0.820	**0.941**	0.882	0.837	**0.826**	**0.798**	**1.727**
GLOCAL	0.784	0.803	0.808	0.723	0.789	0.785	0.933	0.867	0.817	0.796	0.792	4.546
ESMC	**0.822**	0.820	0.833	0.759	0.831	0.817	0.937	0.887	0.838	0.794	0.794	2.318
TRAM	0.800	0.806	0.813	0.749	0.810	0.814	0.933	**0.896**	0.838	0.784	0.784	3.591
CSFS	0.797	0.816	0.818	0.749	0.837	**0.829**	0.936	0.888	**0.842**	0.775	0.792	2.818
Method	*Coverage*↓											Ave. Rank.
	Corel5k	Corel16k1	Corel16k2	ESP Game	Bibtex	Arts	Business	Education	Science	Stackexchess	Yeast	
MSWL	12.482	**46.167**	**50.618**	**16.275**	**36.719**	5.422	**2.337**	4.438	7.251	**69.367**	**6.765**	**1.909**
GLOCAL	14.637	55.004	60.083	18.259	52.480	6.540	2.563	5.027	8.060	81.006	6.969	4.727
ESMC	**12.041**	51.205	53.220	16.421	43.856	5.383	2.403	4.365	7.154	80.869	6.841	2.273
TRAM	12.989	54.676	58.019	16.908	45.145	5.344	2.431	**3.948**	**6.786**	82.316	6.922	2.909
CSFS	13.497	52.350	57.557	17.357	42.367	**5.110**	2.452	4.310	6.958	85.146	7.002	3.182
Method	*Ranking loss*↓											Ave. Rank.
	Corel5k	Corel16k1	Corel16k2	ESP Game	Bibtex	Arts	Business	Education	Science	Stackexchess	Yeast	
MSWL	0.179	**0.159**	**0.162**	**0.233**	**0.134**	0.153	**0.044**	0.108	0.145	**0.194**	**0.192**	**1.909**
GLOCAL	0.214	0.191	0.188	0.273	0.211	0.184	0.050	0.123	0.163	0.202	0.199	4.591
ESMC	**0.175**	0.173	0.166	0.235	0.162	0.147	0.046	0.100	0.138	0.206	0.195	2.364
TRAM	0.189	0.188	0.182	0.243	0.180	0.148	0.045	**0.094**	**0.134**	0.246	0.204	3.136
CSFS	0.203	0.179	0.178	0.247	0.149	**0.139**	0.045	0.101	0.135	0.255	0.199	3.000
Method	*Average precision*↑											Ave. Rank.
	Corel5k	Corel16k1	Corel16k2	ESP Game	Bibtex	Arts	Business	Education	Science	Stackexchess	Yeast	
MSWL	**0.440**	**0.294**	**0.288**	**0.434**	0.457	**0.559**	0.867	**0.559**	**0.500**	0.290	**0.741**	**1.364**
GLOCAL	0.415	0.255	0.255	0.399	0.349	0.496	0.855	0.506	0.416	**0.301**	0.737	4.000
ESMC	0.429	0.286	0.282	**0.434**	0.402	0.528	0.862	0.537	0.473	0.253	0.736	3.091
TRAM	0.431	0.237	0.239	0.422	0.337	0.521	**0.872**	0.564	0.474	0.267	0.728	3.455
CSFS	0.405	0.267	0.256	0.428	**0.460**	0.543	0.866	0.552	0.485	0.223	0.730	3.091

and CSFS are able to work well in the semisupervised setting while GLOCAL has good performance for multilabel learning. Table V shows the performance of these multilabel methods on the 11 datasets.

According to the experimental result from Table V, we have a couple of observations: 1) in terms of average ranking, the proposed method performs the best on all the four evaluation metrics; 2) on all the 11 datasets, the proposed method achieves the best performance on seven datasets with respect to *AUC*, *ranking loss*, and *coverage*, and eight datasets on *average precision*; and 3) the selected comparing methods achieve the best performance on up to 2 out of 11 datasets. Thus, we can conclude that the proposed method benefits the performance with semisupervised multilabel learning, and has the advantage compared with some other methods.

Furthermore, we also use the Friedman test to conduct the statistical significance test, as shown in Table VI. We can observe that the null hypothesis is clearly rejected on each metric at a significance level of $\alpha = 0.05$. Then, we further use the Bonferroni–Dunn test for the analysis, and the result is shown in Fig. 3. From Fig. 3, we can see that MSWL has significantly better performance than GLOCAL. Compared with ESMC and CSFS, MSWL significantly performs better on *average precision*, which also significantly outperforms TRAM on *AUC* and *average precision*. Thus, in terms of semisupervised multilabel learning, the proposed method can achieve highly competitive performance against GLOCAL, TRAM, CSFS, and ESMC.

F. Sensitivity to Parameters

Three parameters are involved in the experiments. Among these parameters, α and β are used to reflect the influence of

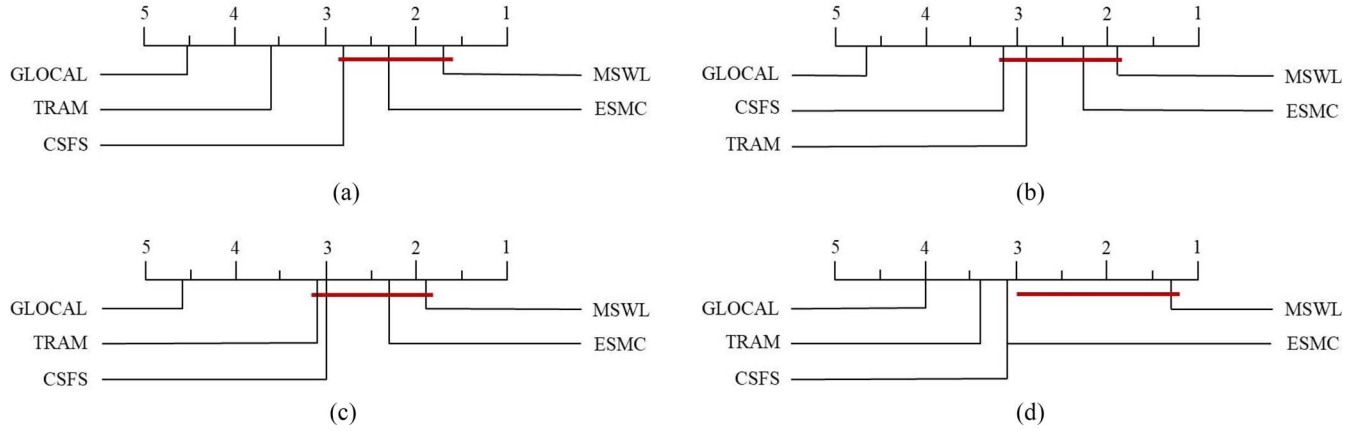

Fig. 3. Comparison of MSWL (control method) against other comparing methods in terms of semisupervised multilabel learning using the Bonferroni–Dunn test (CD = 1.6842 at 0.05 significance level). (a) *AUC*. (b) *Coverage*. (c) *Ranking loss*. (d) *Average precision*.

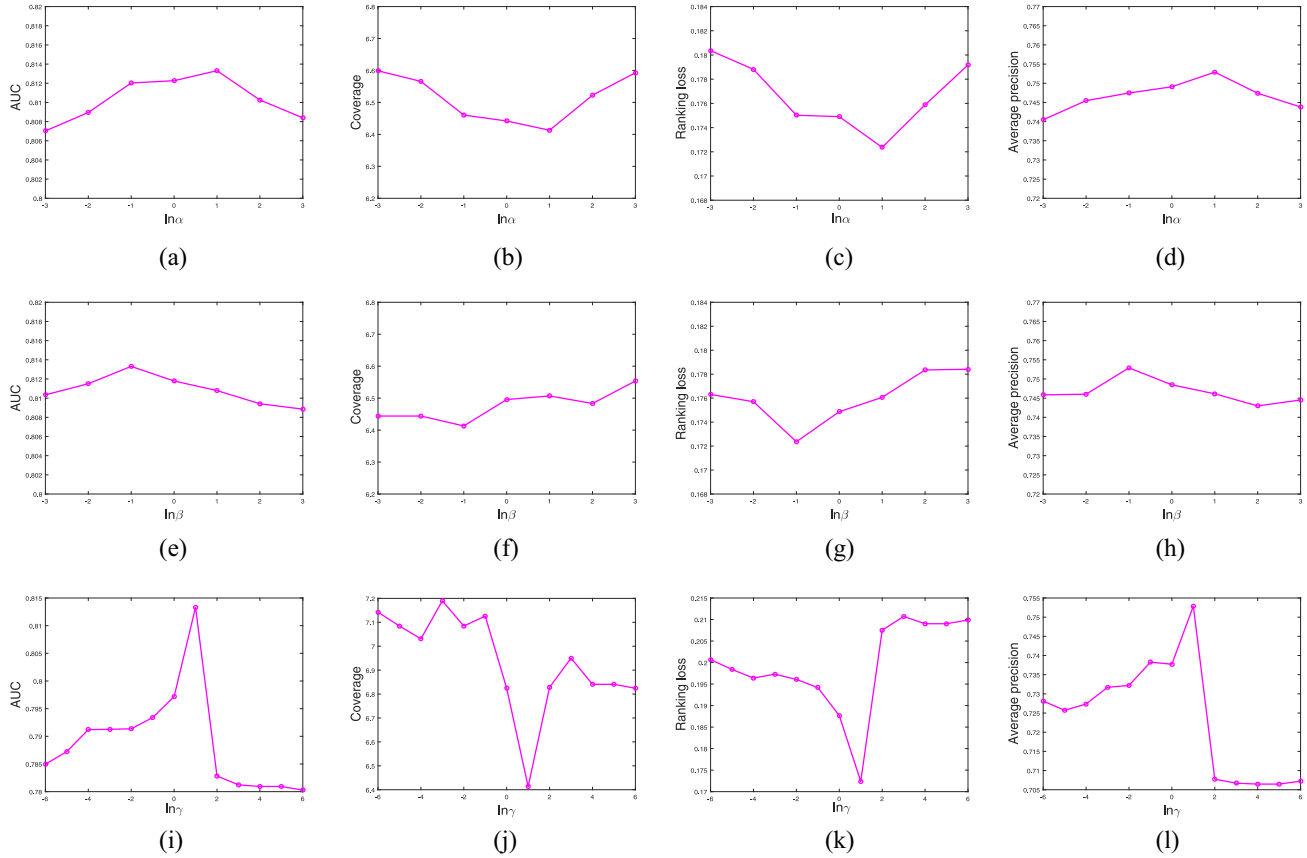

Fig. 4. Parameter sensitivity analysis of MSWL on the Yeast dataset. (a)–(d) Influence of parameter α with fixed β and γ. (e)–(h) Influence of parameter β with fixed α and γ. (i)–(l) Influence of parameter γ with fixed α and β.

GLOCALs, respectively, and γ is for the sparse-based regularizer. We analyze the sensitivity of α, β, and γ on the Yeast dataset, and vary one parameter while the others are fixed at their best setting.

We show the experimental result in Fig. 4, in which Fig. 4(a)–(d) demonstrates the influence of α, and we can see that the high performance of the proposed method is achieved at some intermediate values. This is because that the global label correlations are not fully utilized while α is small, and while α becomes larger, the performance is going to deteriorate in view of that the global label correlations dominate. Thus, the value of α tends to be a compromise. The performance changes in a similar way while β is varied, as shown in Fig. 4(e)–(h). Moreover, based on the range of performance variation, we can see that β is less sensitive than α, and β tends to be smaller than α for obtaining a good performance. Finally, Fig. 4(i)–(l) shows the influence of γ. From them, we can observe that the performance is going to change dramatically regarding each evaluation metric while varying γ. Thus, the proposed method is sensitive to this parameter.

TABLE VI
SUMMARY OF THE FRIEDMAN STATISTICS $F_F(k = 5, N = 11)$ AND THE CRITICAL VALUE ON EVALUATION METRICS

Evaluation metric	F_F	critical value($\alpha = 0.05$)
AUC	9.4378	
Coverage	9.0252	
Ranking loss	7.0783	2.61
Average precision	6.3957	

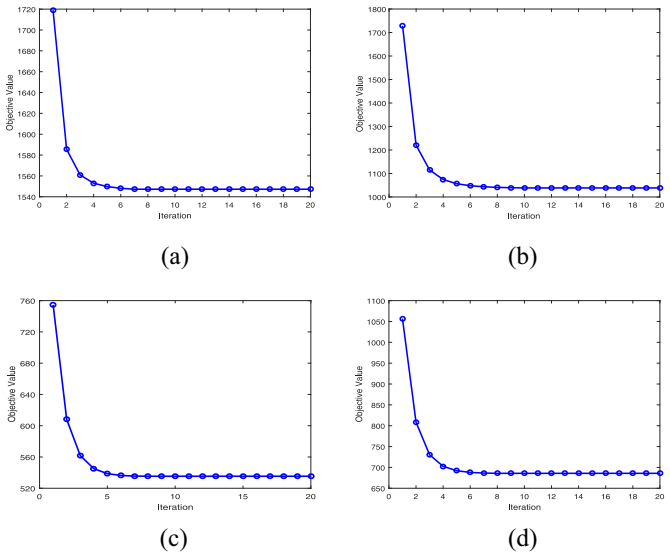

Fig. 5. Convergence analysis of MSWL on the (a) Corel5k, (b) ESP Game, (c) Business, and (d) Yeast datasets.

G. Convergence Analysis

We design the proposed method with an iterative optimization scheme, and the estimation of the two unknown variable sets, that is, **W** and **F**, is the crucial link to study the convergence rate. Once one of the unknown variable sets converges, the objective value reaches stable. *Remark:* The objective value is the output of the proposed optimization objective function, as shown in (15).

Fig. 5(a)–(d) demonstrates the change of the objective value with respect to each iteration on the Corel5k, ESP Game, Business, and Yeast datasets, respectively. From Fig. 5(a), we can observe that the objective value decreases dramatically, and then becomes stable after six iterations. Fig. 5(b)–(d) has a similar phenomenon, and the proposed method can converge within a few iterations. Thus, we can summarize that the proposed algorithm is feasible and effective in obtaining the optimal solution.

V. CONCLUSION

In this article, we proposed a new multilabel method MSWL, which not only helps to MLMLs but also can achieve semisupervised multilabel learning to exploit both labeled and unlabeled data. To achieve the goal, we designed a semisupervised optimization framework based on the manifold regularized sparse model. Furthermore, to alleviate the ambiguity induced by weak-label assignment, we analyzed labeling information to capture the correlations among labels and reconstructed the underly feature structure to preserve instance similarity. Based on this, GLOCALs can be taken into account to improve generalization performance. In short, one of the important contributions of this article is that a unified multilabel learning framework is proposed to the joint learning of label correlations, missing labels, and unlabeled data. By comparing with some state-of-the-art methods, we can conclude that the proposed method has the advantage of learning from weakly labeled data.

In future work, we have interest in further study of labeling information enrichment via label distribution learning, and will also pay attention to weakly multilabel learning via fusing multiple feature modalities.

REFERENCES

[1] K. Barnard, P. Duygulu, D. A. Forsyth, N. de Freitas, D. M. Blei, and M. I. Jordan, "Matching words and pictures," *J. Mach. Learn. Res.*, vol. 3, pp. 1107–1135, Feb. 2003.
[2] I. Katakis, G. Tsoumakas, and I. Vlahavas, "Multilabel text classification for automated tag suggestion," in *Proc. ECML/PKDD 2008 Discover. Challenge*, Antwerp, Belgium, 2008, pp. 75–83.
[3] A. Elisseeff and J. Weston, "A kernel method for multi-labelled classification," in *Proc. Adv. Conf. Neural Inf. Process. Syst.*, Vancouver, BC, Canada, 2001, pp. 681–687.
[4] M. Zhang and Z. Zhou, "A review on multi-label learning algorithms," *IEEE Trans. Knowl. Data Eng.*, vol. 26, no. 8, pp. 1819–1837, Aug. 2014.
[5] F. Herrera, F. Charte, A. J. Rivera, and M. J. del Jesus, *Multilabel Classification: Problem Analysis, Metrics and Techniques*. Cham, Switzerland: Springer, 2016.
[6] J. Liu, Y. Li, W. Weng, J. Zhang, B. Chen, and S. Wu, "Feature selection for multi-label learning with streaming label," *Neurocomputing*, vol. 387, pp. 268–278, Apr. 2020.
[7] H. Liu, X. Li, and S. Zhang, "Learning instance correlation functions for multilabel classification," *IEEE Trans. Cybern.*, vol. 47, no. 2, pp. 499–510, Feb. 2017.
[8] Z. Sun *et al.*, "Mutual information based multi-label feature selection via constrained convex optimization," *Neurocomputing*, vol. 329, pp. 447–456, Feb. 2019.
[9] J. Zhang, Y. Lin, M. Jiang, S. Li, Y. Tang, and K. C. Tan, "Multi-label feature selection via global relevance and redundancy optimization," in *Proc. 29th Int. Joint Conf. Artif. Intell.*, Yokohama, Japan, 2020, pp. 2512–2518.
[10] A. Akbarnejad and M. S. Baghshah, "An efficient semi-supervised multi-label classifier capable of handling missing labels," *IEEE Trans. Knowl. Data Eng.*, vol. 31, no. 2, pp. 229–242, Feb. 2019.
[11] F. Zhao and Y. Guo, "Semi-supervised multi-label learning with incomplete labels," in *Proc. 24th Int. Joint Conf. Artif. Intell.*, Buenos Aires, Argentina, 2015, pp. 4062–4068.
[12] S. S. Bucak, R. Jin, and A. K. Jain, "Multi-label learning with incomplete class assignments," in *Proc. 24th IEEE Conf. Comput. Vis. Pattern Recognit.*, Providence, RI, USA, 2011, pp. 2801–2808.
[13] Y. Sun, Y. Zhang, and Z. Zhou, "Multi-label learning with weak label," in *Proc. 24th AAAI Conf. Artif. Intell.*, Atlanta, GA, USA, 2010, pp. 593–598.
[14] J. Dai, Q. Hu, J. Zhang, H. Hu, and N. Zheng, "Attribute selection for partially labeled categorical data by rough set approach," *IEEE Trans. Cybern.*, vol. 47, no. 9, pp. 2460–2471, Sep. 2017.
[15] C. Zhang, J. Cheng, and Q. Tian, "Multiview, few-labeled object categorization by predicting labels with view consistency," *IEEE Trans. Cybern.*, vol. 49, no. 11, pp. 3834–3843, Nov. 2019.
[16] X. Chang, F. Nie, Y. Yang, and H. Huang, "A convex formulation for semi-supervised multi-label feature selection," in *Proc. 28th AAAI Conf. Artif. Intell.*, Quebec City, QC, Canada, 2014, pp. 1171–1177.
[17] X. Kong, M. K. Ng, and Z. Zhou, "Transductive multilabel learning via label set propagation," *IEEE Trans. Knowl. Data Eng.*, vol. 25, no. 3, pp. 704–719, Mar. 2013.
[18] B. Wu, S. Lyu, B. Hu, and Q. Ji, "Multi-label learning with missing labels for image annotation and facial action unit recognition," *Pattern Recognit.*, vol. 48, no. 7, pp. 2279–2289, 2015.

[19] H. Yu, P. Jain, P. Kar, and I. S. Dhillon, "Large-scale multi-label learning with missing labels," in *Proc. 31th Int. Conf. Mach. Learn.*, Beijing, China, 2014, pp. 593–601.

[20] S. Behpour, W. Xing, and B. D. Ziebart, "ARC: Adversarial robust cuts for semi-supervised and multi-label classification," in *Proc. 32nd AAAI Conf. Artif. Intell.*, New Orleans, LA, USA, 2018, pp. 2704–2711.

[21] L. Wu and M. Zhang, "Multi-label classification with unlabeled data: An inductive approach," in *Proc. Asian Conf. Mach. Learn.*, Canberra, ACT, Australia 2013, pp. 197–212.

[22] F. Charte, "A comprehensive and didactic review on multilabel learning software tools," *IEEE Access*, vol. 8, pp. 50330–50354, 2020.

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

[24] G. Tsoumakas, I. Katakis, and I. P. Vlahavas, "Mining multi-label data," in *Data Mining and Knowledge Discovery Handbook*, 2nd ed. Boston, MA, USA: Springer, 2010, pp. 667–685.

[25] J. Huang, G. Li, Q. Huang, and X. Wu, "Joint feature selection and classification for multilabel learning," *IEEE Trans. Cybern.*, vol. 48, no. 3, pp. 876–889, Mar. 2018.

[26] Y. Lin, Q. Hu, J. Zhang, and X. Wu, "Multi-label feature selection with streaming labels," *Inf. Sci.*, vol. 372, pp. 256–275, Dec. 2016.

[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.*, Macao, China, 2019, pp. 3691–3697.

[28] J. Zhang *et al.*, "Multi-label learning with label-specific features by resolving label correlations," *Knowl. Based Syst.*, vol. 159, pp. 148–157, Nov. 2018.

[29] Y. Zhu, J. T. Kwok, and Z. Zhou, "Multi-label learning with global and local label correlation," *IEEE Trans. Knowl. Data Eng.*, vol. 30, no. 6, pp. 1081–1094, Jun. 2018.

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

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

[32] S. Xie, X. Kong, J. Gao, W. Fan, and P. S. Yu, "Multilabel consensus classification," in *Proc. 13th IEEE Int. Conf. Data Min.*, Dallas, TX, USA, 2013, pp. 1241–1246.

[33] J. Read, B. Pfahringer, G. Holmes, and E. Frank, "Classifier chains for multi-label classification," *Mach. Learn.*, vol. 85, no. 3, pp. 333–359, 2011.

[34] G. Tsoumakas and I. P. Vlahavas, "Random *k*-labelsets: An ensemble method for multilabel classification," in *Proc. 18th Eur. Conf. Mach. Learn.*, Warsaw, Poland, 2007, pp. 406–417.

[35] S. Huang and Z. Zhou, "Multi-label learning by exploiting label correlations locally," in *Proc. 26th AAAI Conf. Artif. Intell.*, Toronto, ON, Canada, 2012, pp. 945–955.

[36] P. Hou, X. Geng, and M. Zhang, "Multi-label manifold learning," in *Proc. 30th AAAI Conf. Artif. Intell.*, Phoenix, AZ, USA, 2016, pp. 1680–1686.

[37] J. Zhang, C. Li, Z. Sun, Z. Luo, C. Zhou, and S. Li, "Towards a unified multi-source-based optimization framework for multi-label learning," *Appl. Soft Comput.*, vol. 76, pp. 425–435, Mar. 2019.

[38] J. Huang *et al.*, "Improving multi-label classification with missing labels by learning label-specific features," *Inf. Sci.*, vol. 492, pp. 124–146, Aug. 2019.

[39] X. Li, B. Shen, B. Liu, and Y. Zhang, "Ranking-preserving low-rank factorization for image annotation with missing labels," *IEEE Trans. Multimedia*, vol. 20, no. 5, pp. 1169–1178, May 2018.

[40] M. Chen, A. X. Zheng, and K. Q. Weinberger, "Fast image tagging," in *Proc. 30th Int. Conf. Mach. Learn.*, Atlanta, GA, USA, 2013, pp. 1274–1282.

[41] M. Xu, R. Jin, and Z. Zhou, "Speedup matrix completion with side information: Application to multi-label learning," in *Proc. Adv. Conf. Neural Inf. Process. Syst.*, 2013, pp. 2301–2309.

[42] Z. He, M. Yang, Y. Gao, H. Liu, and Y. Yin, "Joint multi-label classification and label correlations with missing labels and feature selection," *Knowl. Based Syst.*, vol. 163, pp. 145–158, Jan. 2019.

[43] 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 Min.*, San Diego, CA, USA, 2018, pp. 450–458.

[44] B. Wu, S. Lyu, and B. Ghanem, "ML-MG: Multi-label learning with missing labels using a mixed graph," in *Proc. IEEE Int. Conf. Comput. Vis.*, Santiago, Chile, 2015, pp. 4157–4165.

[45] B. Wu, Z. Liu, S. Wang, B. Hu, and Q. Ji, "Multi-label learning with missing labels," in *Proc. 22nd Int. Conf. Pattern Recognit.*, Stockholm, Sweden, 2014, pp. 1964–1968.

[46] Y. Liu, K. Wen, Q. Gao, X. Gao, and F. Nie, "SVM based multi-label learning with missing labels for image annotation," *Pattern Recognit.*, vol. 78, pp. 307–317, Jun. 2018.

[47] A. Braytee, W. Liu, A. Anaissi, and P. J. Kennedy, "Correlated multi-label classification with incomplete label space and class imbalance," *ACM Trans. Intell. Syst. Technol.*, vol. 10, no. 5, pp. 1–26, 2019.

[48] B. Wu, S. Lyu, and B. Ghanem, "Constrained submodular minimization for missing labels and class imbalance in multi-label learning," in *Proc. 30th AAAI Conf. Artif. Intell.*, Phoenix, AZ, USA, 2016, pp. 2229–2236.

[49] H. Yang, J. T. Zhou, and J. Cai, "Improving multi-label learning with missing labels by structured semantic correlations," in *Proc. Eur. Conf. Comput. Vis.*, Amsterdam, The Netherlands, 2016, pp. 835–851.

[50] Z. Zhou, "A brief introduction to weakly supervised learning," *Nat. Sci. Rev.*, vol. 5, no. 1, pp. 44–53, 2018.

[51] Y. Liu, F. Nie, and Q. Gao, "Nuclear-norm based semi-supervised multiple labels learning," *Neurocomputing*, vol. 275, pp. 940–947, Jan. 2018.

[52] H. Dong, Y. Li, and Z. Zhou, "Learning from semi-supervised weak-label data," in *Proc. 32nd AAAI Conf. Artif. Intell.*, New Orleans, LA, USA, 2018, pp. 2926–2933.

[53] T. Wei, L. Guo, Y. Li, and W. Gao, "Learning safe multi-label prediction for weakly labeled data," *Mach. Learn.*, vol. 107, no. 4, pp. 703–725, 2018.

[54] J. Ma and T. W. S. Chow, "Robust non-negative sparse graph for semi-supervised multi-label learning with missing labels," *Inf. Sci.*, vol. 422, pp. 336–351, Jan. 2018.

[55] Z. Zhang, Y. Xu, J. Yang, X. Li, and D. Zhang, "A survey of sparse representation: Algorithms and applications," *IEEE Access*, vol. 3, pp. 490–530, 2015.

[56] F. Nie, H. Huang, X. Cai, and C. H. Q. Ding, "Efficient and robust feature selection via joint 2, 1-norms minimization," in *Proc. Adv. Conf. Neural Inf. Process. Syst.*, Vancouver, BC, Canada, 2010, pp. 1813–1821.

[57] S. P. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Found. Trends Mach. Learn.*, vol. 3, no. 1, pp. 1–122, 2011.

[58] S. T. Roweis and L. K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science*, vol. 290, no. 5500, pp. 2323–2326, 2000.

[59] Y. Xu, J. Wang, S. An, J. Wei, and J. Ruan, "Semi-supervised multi-label feature selection by preserving feature-label space consistency," in *Proc. 27th ACM Int. Conf. Inf. Knowl. Manag.*, Torino, Italy, 2018, pp. 783–792.

[60] E. Anderson *et al.*, *LAPACK Users' Guide, 3rd ed*. Philadelphia, PA, USA: Soc. Ind. Appl. Math., 1999.

[61] J. Zhang, Z. Luo, C. Li, C. Zhou, and S. Li, "Manifold regularized discriminative feature selection for multi-label learning," *Pattern Recognit.*, vol. 95, pp. 136–150, Nov. 2019.

[62] P. Duygulu, K. Barnard, J. F. G. de Freitas, and D. A. Forsyth, "Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary," in *Proc. 7th Eur. Conf. Comput. Vis.*, Copenhagen, Denmark, 2002, pp. 97–112.

[63] L. von Ahn and L. Dabbish, "Labeling images with a computer game," in *Proc. Conf. Hum. Factors Comput. Syst.*, Vienna, Austria, 2004, pp. 319–326.

[64] S. S. Vempala, *The Random Projection Method* (DIMACS Series in Discrete Mathematics and Theoretical Computer Science), vol. 65. Providence, RI, USA: AMS, 2004.

[65] F. Charte and D. Charte, "Working with multilabel datasets in R: The mldr package," *R J.*, vol. 7, no. 2, p. 149, 2015.

[66] N. Ueda and K. Saito, "Parametric mixture models for multi-labeled text," in *Proc. Adv. Conf. Neural Inf. Process. Syst.*, Vancouver, BC, Canada, 2002, pp. 721–728.

[67] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognit. Lett.*, vol. 27, no. 8, pp. 861–874, 2006.

[68] M. Zhang and L. Wu, "LIFT: Multi-label learning with label-specific features," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 1, pp. 107–120, Jan. 2015.

[69] J. Demsar, "Statistical comparisons of classifiers over multiple data sets," *J. Mach. Learn. Res.*, vol. 7, pp. 1–30, Jan. 2006.

Jia Zhang received the M.S. degree from the School of Computer Science, Minnan Normal University, Zhangzhou, China, in 2016. He is currently pursuing the Ph.D. degree in artificial intelligence from Xiamen University, Xiamen, China.

He is broadly interested in machine learning, data mining, and artificial intelligence. He is currently working on multilabel learning, data fusion, feature selection, and weakly supervised learning.

Shaozi Li (Senior Member, IEEE) received the B.S. degree from Hunan University, Changsha, China, in 1983, the M.S. degree from Xi'an Jiaotong University, Xi'an, China, in 1988, and the Ph.D. degree from the National University of Defense Technology, Changsha, China, in 2009.

He is a Full Professor with the Artificial Intelligence Department, Xiamen University, Xiamen, China. He has directed and completed more than 20 research projects, including several national 863 Programs, National Nature Science Foundation of China, and the Ph.D. Programs Foundation of the Ministry of Education of China. Furthermore, he has authored nearly 300 papers in journals and international conferences. His research interests include artificial intelligence and its applications, moving objects detection and recognition, machine learning, computer vision, and multimedia information retrieval.

Prof. Li is the Vice Director of the Technical Committee on Collaborative Computing of CCF, and the Fujian Association of Artificial Intelligence. He is a Senior Member of ACM and the China Computer Federation.

Min Jiang (Senior Member, IEEE) received the bachelor's and Ph.D. degrees in computer science from Wuhan University, Wuhan, China, in 2001 and 2007, respectively.

Subsequently as a Postdoctoral Researcher with the Department of Mathematics, Xiamen University, Xiamen, China, where he is currently a Professor with the Department of Artificial Intelligence. His main research interests are machine learning, computational intelligence, and robotics. He has a special interest in dynamic multiobjective optimization, transfer learning, the software development and in the basic theories of robotics.

Dr. Jiang received the Outstanding Reviewer Award from the IEEE TRANSACTIONS ON CYBERNETICS in 2016. He is currently serving as an Associate Editor for the IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS and the IEEE TRANSACTIONS ON COGNITIVE AND DEVELOPMENTAL SYSTEMS. He is the Chair of IEEE CIS Xiamen Chapter.

Kay Chen Tan (Fellow, IEEE) received the B.Eng. (First Class Hons.) and Ph.D. degrees from the University of Glasgow, Glasgow, U.K., in 1994 and 1997, respectively.

He is a Full Professor with the Department of Computer Science, City University of Hong Kong, Hong Kong. He has published over 200 refereed articles and six books.

Prof. Tan is the Editor-in-Chief of the IEEE TRANSACTIONS ON EVOLUTIONARY COMPUTATION. He was the Editor-in-Chief of the *IEEE Computational Intelligence Magazine* from 2010 to 2013. He currently serves on the Editorial Board Member of over 20 journals. He was an Elected Member of IEEE CIS AdCom from 2017 to 2019.