IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received February 24, 2020, accepted April 9, 2020, date of publication April 14, 2020, date of current version April 29, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2987922

A Structure-Induced Framework for Multi-Label Feature Selection With Highly Incomplete Labels

TIANTIAN XU AND LONG ZHAO
School of Computer Science and Technology, Qilu University of Technology (Shandong Academy of Sciences), Jinan 250353, China
Corresponding authors: Tiantian Xu (xtt-ok@163.com) and Long Zhao (zxcvbnm9515@163.com)

This work was supported in part by the National Natural Science Foundation of China under Grant 61906104 and Grant 61806105, and in part by the Natural Science Foundation of Shandong province, China, under Grant ZR2019BF018.

ABSTRACT Feature selection has shown significant promise in improving the effectiveness of multi-label learning by constructing a reduced feature space. Previous studies typically assume that label assignment is complete or partially complete; however, missing-label and unlabeled data are commonplace and accompanying occurrences in real applications due to the high expense of manual annotation and label ambiguity. We call this "highly incomplete labels" problem. Such label incompleteness severely damages the inherent label structures and masks true label correlations. In this paper, we propose a novel structure-induced feature selection model to simultaneously identify the most discriminative features and recover the highly incomplete labels. To our best knowledge, it is the first attempt to explore the local density structure of data to capture the intricate feature-label dependency in the highly incomplete learning scenarios. Feature selection is guided by the label structure reconstruction, and highly incomplete labels are recovered via the structure transferred from feature space. In this elegant manner, the processes of selecting discriminative features and recovering incomplete labels are coupled in a unified optimization framework. Comprehensive experiments on public benchmark datasets demonstrate the superiority of the proposed approach.

INDEX TERMS Feature selection, multi-label learning, weakly-supervised learning, label correlation.

I. INTRODUCTION

Feature selection is essential for analyzing high-dimensional data. It aims to select a subset of relevant and discriminative features from the original high-dimensional feature space for a compact yet accurate representation of the data. With the increasing availability of multi-label data wherein one instance is related to multiple labels, dozens of feature selection approaches specific to multi-label learning have been proposed to reduce dimensionality and improve learning performance [1]–[3].

Traditional multi-label feature selection algorithms are devised under the assumption that the label matrix of training data is complete. In real-world applications, however, this assumption is tough to satisfy because complete labels tend to become increasingly difficult to fetch (in consideration of the explosive increase of data size). A common learning scenario in image annotation is demonstrated in Fig. 1. Partially labeled and unlabeled instances simultaneously exist in the multi-label image dataset. The partially labeled instance is annotated with *penguin*, *pine tree*, and *snow*, while other labels are missing (e.g., *building*, *sky*, *window*, and *handrail*). Furthermore, some instances are completely unlabeled due to limited resources, thereby yielding empty label sets. Such label incompleteness masks the inherent label structure

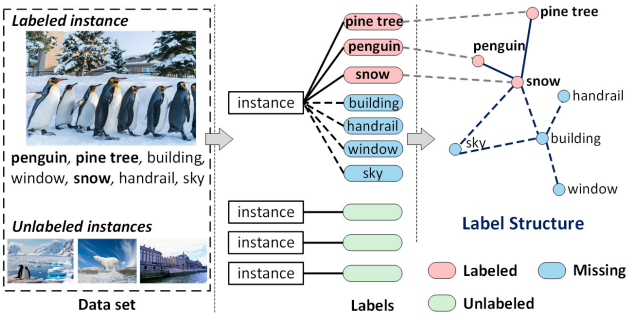


FIGURE 1. A learning scenario for multi-label feature selection with highly incomplete labels: partial labels are annotated (marked as the red parts), and other labels are missing (blue parts) or completely unannotated (green parts).

The associate editor coordinating the review of this manuscript and approving it for publication was Pasquale De Meo.

and deteriorates the performance of feature selection. For instance, the correlation between *building* and *window* cannot be extracted due to missing labels; correspondingly, the discriminative features for both labels (or even other closely related labels, such as *snow*) can hardly be selected. In another example, high-dimensional proteomic data are easy to collect because of high-throughput biotechnologies, but the process is limited by the applied experimental protocols. That is, highly confidential proteins bear labels (protein function), and other proteins (sometimes most of them) are not annotated.

The existing multi-label feature selection approaches address either the missing label issue or the semi-supervised issue (unlabeled and completely labeled instances). A pioneer study on multi-label feature selection with missing labels [4] jointly recovered label space and removed irrelevant features while ignoring the latent guide information given by the unlabeled instances. By contrast, some semi-supervised multi-label feature selection approaches exploit manifold learning based on local geometry structure to capture label correlations, which are then used to steer feature selection [5]–[7]. These works derive benefits from a moderate number of completely labeled instances; in the learning scenario of this work, the traditional semi-supervised learning loses its effectiveness because of lacking the sufficient support from completely labeled instances [8]. Holistically handling the missing-label instances and unlabeled instances is the focus of this work.

Furthermore, the aforementioned works generally yield correlation information by constructing local geometry structures based on distance metrics; nevertheless, distance metrics are inappropriate in some cases to measure the intricate dependencies between features and labels. On the one hand, the pairwise distances measured in the high-dimensional feature space may be not qualitatively meaningful due to the curse of dimensionality [9]; on the other hand, distance metrics are typically irrespective of the data distribution information and thus fail to capture the intrinsic structure especially when data are not uniform density distributions [10]. A well-known example is that two instances in the sparse region may be more similar than two instances with equal inter-distance in the dense region [11]. Thus we argue that, local geometry structures are weak to capture the intrinsic correlation information in some complex learning scenarios.

These observations inspire us to approach incomplete label.[1] information-oriented multi-label feature selection. We tackle three major challenges in this work: **(1)** how to extract underlying local structure from the original feature space (filled with a considerable number of irrelevant and redundant features); **(2)** how to incorporate feature-label dependency based on the extracted local structure to recover label structure; and **(3)** how to imitate the annotator in selecting discriminative features based on the recovered label structure.

Inspired by the chicken-and-egg relationship, we endeavor to accomplish three tasks in a unified framework and adopt a mutual optimization mechanism to solve the latter two tasks. A novel feature selection framework is proposed by employing the weakly supervised setting specific to the highly incomplete learning cases, called **S**tructure-induced **M**ulti-label **F**eature **S**election (SMFS). Concretely, the intrinsic local density structure of the feature space is extracted through the probability mass estimation, and subsequently transferred to estimate missing labels. Under the *Smoothness Assumption* that *the points close to each other are more likely to share the same label*, label structure reconstruction and feature selection are alternately performed in SMFS, that is, the recovered labels contribute to finding the discriminative features and the selected features facilitate the recovery of labels.

In summary, the main contributions are as follows:

- We pioneerly attempt to conduct multi-label feature selection with highly incomplete labels, making sense to practical applications. Our tasks are more complex and challenging than existing ones due to the considerable lack of prior assignment information within instances and between instances.
- To our best knowledge, we firstly construct the local density structure by considering the concrete data distributions in multi-label feature selection and follow the *Smoothness Assumption* to explore feature-label dependency, both of which facilitate effective feature selection and label recovery in a unified framework.
- Extensive experimental comparisons with the state-of-the-art feature selection approaches are conducted on various multi-label benchmark datasets to validate the performance of SMFS from the empirical point of view.

II. RELATED WORK

A. MULTI-LABEL LEARNING

Multi-label learning, which assumes one instance is associated with several labels simultaneously, has been diversely applied in real-world applications, e.g., text classification [12], [13], bioinformatics [14], [15], image tagging [16], [17], action recognition [18], *et al.* Generally, in terms of label correlations, exiting multi-label classification approaches can be classified into three categories: First-order approach, second-order approach and high-order approach.

The first group is that binary methods separate the multi-label problem into multiple single-label subproblems, and employ single-label classifier for each subproblems [12], [16]. An inevitable drawback is that these approaches commonly ignore label correlations, which have revealed important roles in many state-of-the-art works [19]. And, when the size of label set is large or even enormous, the number of binary classifiers would be larger, which leads to class-imbalance problems. The second group is that

[1]Incomplete labels comprise two parts in our work, i.e., missing labels and unannotated labels. Missing labels denote those missing annotations in partially labeled instances, and unannotated labels correlate with the instances without any annotations (i.e., the learning targets in common semi-supervised cases)

multi-label learning approaches generally consider pairwise label correlations. For instance, Fürnkranz *et al.* [20] utilized calibrated label ranking (CLR) to transfer the multi-label learning problem into a pairwise label ranking problem. These approaches use label ranking technologies to score each instance-label pair with minimizing label ranking loss functions [21], [22]. Nevertheless, label correlations in real-world applications could be complex and exceed pairwise relationship. Thus, the last group is that global label correlation is taken into consideration in multi-label learning. For example, the work in [23] utilized the classifier chain (CC) that transformed the multi-label learning problem into a chain of binary classification problems. Another solutions to extract high-order label correlations were construct based on shared subspaces or latent label space [24]. For instance, the work in [25] captured the low-rank label matrix by exploiting dictionary learning. Recently, deep learning method [26] was also employed to simultaneously learn feature subspace and label subspace. These approaches endeavor to extract a common subspace shared among multiple labels, thereby suffering high costs for estimating complex label correlations.

B. MULTI-LABEL FEATURE SELECTION

On the basis of label information, existing multi-label feature selection approaches can be roughly categorized into three families.

The first family is typically implemented over complete label information. Some multi-label feature selection approaches [27], [28] divide the multi-label learning problem into multiple subproblems, which fails to take the label interdependency into account. A majority of approaches try to incorporate label correlations into the process of model construction to help select discriminative features [2], [3], [29], [30]. To remove the irrelevant and noisy features, sparse regularization [31] is also imposed on the feature selection matrices. These approaches perform under the condition that the training data are equipped with complete label assignments, which is hard to satisfy in reality.

The second family comprises the semi-supervised approaches, which focus on dealing with a certain volume of unlabeled data. For example, the work in [32] employs a simple linear regression model to learn the label matrix. A popular strategy is to utilize the manifold learning or shared subspace learning to capture label correlations [5]–[7], [17], [33]. Although the aforementioned approaches can process unlabeled data, they need to be enlightened with abundant complete assignments of labeled data.

The last family lays emphasis on tackling the missing label issues that are rarely touched. An example is the work in [4] that employ robust linear regression to missing label recovery and feature selection simultaneously. However, these studies neglect the importance of a large number of unlabeled instances for feature selection.

Another series of works that relate to ours are local structure-based approaches. Most multi-label feature selection algorithms [4]–[7], [34] explore correlation information based on the local structure to guide the selection of relevant features. Despite their widespread applications, most approaches capture the local geometry structure through distance metrics that completely rely on the geometric positions of instances and thus hardly capture their intricate relations for the following reasons. First, many approaches [3], [6], [33], [35] resort to the Gaussian function to construct local geometry structure, which makes these models sensitive to the parameter tuning. Next, the reconstruction or adaptive weights are measured based on nearest neighbors and employed to characterize geometric properties [7], [34], which may be not stable and may produce distorted structure [36]. Furthermore, local geometry structure completely ignores the data distributions, which are beneficial for observing the nature of data and construct effective feature selection models. Previous researchers have pointed out two instances in a sparse region are more similar than two instances with the same inter-distance in a dense region [10], [11]. This property motivates us to capture the local information by resorting to the density structure of data.

We must emphasize the difference between the traditional manifold learning and SMFS. Traditional manifold learning is mostly used to induce a low-dimensional embedding of the latent manifold. The recovery of the label manifold in SMFS is similar to the popular manifold learning strategy called locally linear embedding [37]. However, SMFS performs this recovery not by embedding the feature manifold into the label manifold; that is, in SMFS, two manifolds exist in different spaces.

In contrast to existing multi-label feature selection approaches, the proposed SMFS firstly approaches the highly incomplete label learning task, and derives benefits in the following aspects: (**1**) the intrinsic local density structure is effectively captured by probability mass estimation in the feature space, which can be used to measure the reliable feature-label dependencies; (**2**) the local label structure is reconstructed via the local feature structure information according to the *Smoothness Assumption*, and contributes to the complete label recovery; and (**3**) feature selection and label recovery are mutually boosted in a unified framework.

III. THE PROPOSED FRAMEWORK

A. PROBLEM FORMULATION

Here, we provide the basic notations used in this study. Let $\boldsymbol{X} = [\boldsymbol{x}_1; \ldots; \boldsymbol{x}_n] \in \mathbb{R}^{n\times d}$ be a data matrix for n instances, where $\boldsymbol{x}_i \in \mathbb{R}^d$ is the i-th instance, and $\mathcal{H}(\boldsymbol{X}) = \{H_t\}_{t=1}^{\tau}$ denotes the set of all partitions available in $\boldsymbol{X}$. $\boldsymbol{Y} = \begin{bmatrix} \boldsymbol{Y}_l \\ \boldsymbol{Y}_u \end{bmatrix} \in \mathbb{R}^{n\times c}$ is the label matrix, where c is the number of labels. $\boldsymbol{Y}_l$ represents the partially labeled matrix with some missing labels, where $y_{ij} = 1$ if the i-th instance is associated with the j-th label, and $y_{ij} = -1$ indicates that the i-th instance is not assigned or omitted with the j-th label [8]. $\boldsymbol{Y}_u$ denotes the unlabeled matrix, where y_{ij} is initially set to -1 and iteratively learned in the subsequent optimization process.

B. STRUCTURE-INDUCED MULTI-LABEL FEATURE SELECTION

When label matrix is highly incomplete, inherent label structure is inevitably damaged and the label correlation based on this incomplete label information hinders the selection of relevant features and the recovery of labels. To address this problem, we recover the underlying true label matrix $\boldsymbol{Y}$ based on local structure reconstruction via transferring the local structure from feature space to label space. Hence, the extraction of intrinsic local structure in the feature space that encodes the intricate feature-label dependencies is significant for effective model construction. Motivated by the work in [11], we capture the local density structure of the feature space solely dependent on data distribution via probability mass estimation. We formalize the concepts as follows. Its underlying intuition is that the affinity relation between two instances primarily depends on the amount of probability mass in the partition that covers two instances. For any two instances $\boldsymbol{x}_i, \boldsymbol{x}_j$, $\rho\left(\boldsymbol{x}_i, \boldsymbol{x}_j | H; \boldsymbol{X}\right)$ is the smallest partition containing $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ w.r.t. $\boldsymbol{X}$. In practice, ρ estimates the probability of the partition by counting the number of instances in that partition and the probability mass would be estimated from τ partitions as follows:

$$\rho\left(\boldsymbol{x}_i, \boldsymbol{x}_j | H; \boldsymbol{X}\right) = \arg\min_{r} \sum_{\boldsymbol{x}_q \in \boldsymbol{X}} \mathbb{I}\left(\boldsymbol{x}_q \in r | r \in H\right), \quad (1)$$

$$\xi\left(\boldsymbol{x}_i, \boldsymbol{x}_j\right) = \frac{1}{\tau} \sum_{t=1}^{\tau} \frac{|\rho\left(\boldsymbol{x}_i, \boldsymbol{x}_j | H_t\right)|}{|\boldsymbol{X}|}. \quad (2)$$

where $\boldsymbol{x}_i, \boldsymbol{x}_j \in r$, r is the partition, and $\mathbb{I}(\cdot)$ is an indicator function. The probability mass of $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ is defined as $\xi\left(\boldsymbol{x}_i, \boldsymbol{x}_j\right) \in (0, 1]$, where a small $\xi\left(\boldsymbol{x}_i, \boldsymbol{x}_j\right)$ indicates a highly relevant relation between two instances. Consequently, based on $\xi\left(\boldsymbol{x}_i, \boldsymbol{x}_j\right)$, the local density structure in the feature space can be constructed through the k-lowest probability mass neighborhoods, which preserves inherent local correlation information. To ensure the local density structure validity in recovering labels structure, the weight matrix $\boldsymbol{W}$ of local density structure is defined as

$$w_{ij} = \frac{exp(-\xi\left(\boldsymbol{x}_i, \boldsymbol{x}_j\right))}{\sum_q exp\left(-\xi\left(\boldsymbol{x}_i, \boldsymbol{x}_q\right)\right)}, \quad \text{if } j \in \mathcal{N}_k(i) \quad (3)$$

where $\boldsymbol{W} \in \mathbb{R}^{n \times n}$ and $\mathcal{N}_k(i)$ denotes the index set for the local density structure of i-th instance measured by $\xi\left(\boldsymbol{x}_i, \boldsymbol{x}_j\right)$. This strategy promises weights smoothly dependent on the probability mass estimation [38]. We constrain $\mathbf{1}^T \boldsymbol{w}_i = 1$ to ensure the local structure that transfers from feature space to label space remain invariant [39], where $\mathbf{1}$ is the vector of all ones.

Here H can produce larger partitions in sparse region than those in dense region, which means that two instances in sparse region are more similar than two instances of equal geometry distance in dense region [10]. Fig. 2 compares the contour of the two structure under non-uniform distribution

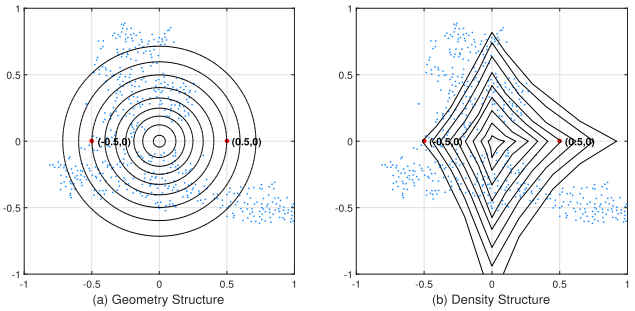


FIGURE 2. The geometry structure versus the density structure: Contours w.r.t. (0, 0). A 2-dimensional visualization of the multi-dimensional Forest dataset, transformed by t-SNE [40].

Forest dataset taken from website[2]. This example shows that the geometry structure (i.e., resorting distance metrics) has the same contour irrespective of the data distribution. In contrast, the density structure (i.e., resorting probability mass estimation) can adapts its contour to the local data distribution, i.e., from the peak (0, 0), the contour decreases at a slower rate in the sparsest region than in dense regions. Accordingly, local density structure has a higher value at $(0.5, 0)$ than that at $(-0.5, 0)$. The ability to adapt to the density structure of a dataset is the key advantage of the local density structure in feature selection, which naturally induces from *Smoothness Assumption* that, the local density structure can be transferred from feature space to label space to help better recover label structure.

According to the local density structure in the feature space, we can estimate the confidence of the missing label j for the i-th ($1 \leq i \leq l$) instance using its already known labels via the *Smoothness Assumption* as follows:

$$y_{ij} = \begin{cases} \boldsymbol{w}_i \boldsymbol{y}_{\cdot j}, & \text{if } y_{ij} = -1 \\ y_{ij}, & \text{otherwise.} \end{cases} \quad (4)$$

In Eq. (4), $\boldsymbol{y}_i$ is the label vector for the i-th ($1 \leq i \leq l$) instance with known and estimated labels. If $y_{ij} = -1$ and the instances correlated with the i-th instance also correlate with the j-th label, $y_{ij} = w_{i1}y_{1j} + w_{i2}y_{2j} + \cdots + w_{in}y_{nj}$, then the missing label is assigned with a large value (high confidence). Take the electronic medical record diagnosis for example. If the patient records similar to the i-th patient record are annotated with cardiopathy (j-th label), the i-th patient has a high risk of cardiopathy.

Accordingly, local density structure is captured and missing labels are estimated, with the transferred local density structure, the reconstruction of the complete label matrix can be infer to the minimization of

$$\min_{\boldsymbol{Y}} \sum_i \|\boldsymbol{y}_i - \sum_j w_{ij}\boldsymbol{y}_j\|_2^2. \quad (5)$$

Since the local density structure in feature space adapts to the density structure of a dataset, we naturally induce the consistency of the local density structure between feature space and

[2]http://archive.ics.uci.edu/ml/datasets.php.

label space based on *Smoothness Assumption*, that is, label space and feature space share same local density structure, which is fundamentally different with label semantic enrichment that employs reconstruction error [39]. Hence, label density structure is effectively recovered through transferring the local density structure from feature space.

Finally, we jointly select features and recover incomplete labels based on the *Consistency Assumption*, *Smoothness Assumption*, and local structure reconstruction, to overcome the challenging issue of highly incomplete labels. The *Consistency Assumption* ensures that the predicted label matrix is consistent with the initial label matrix. The learning objective of the proposed SMFS is defined as

$$\min_{\boldsymbol{P},\boldsymbol{Y}} \sum_i s_{ii} \left\|\boldsymbol{x}_i\boldsymbol{P}-\boldsymbol{y}_i\right\|_2^2+\alpha \sum_i \|\boldsymbol{y}_i-\sum_j w_{ij}\boldsymbol{y}_j\|_2^2+\beta \|\boldsymbol{P}\|_{2,1}\,,$$
$$s.t.\ \mathbf{1}^T\boldsymbol{w}_i=1, \quad 0\le w_{ij}\le 1, \quad -1\le y_{ij}\le 1, \tag{6}$$

where $\boldsymbol{P} \in \mathbb{R}^{d\times c}$ is a map or feature selection matrix in which p_{ij} records the discriminative ability of the i-th feature to the j-label. α and β are trade-off parameters, and $\boldsymbol{S} \in \mathbb{R}^{n\times n}$ is a diagonal matrix to distinguish labeled and unlabeled instances, $s_{ii}=1$ if $\boldsymbol{x}_i$ is labeled and 0 otherwise. SMFS is added a $\|\boldsymbol{P}\|_{2,1}$ penalty to encourage sparseness. An optimal feature subspace is constructed according to the feature selection matrix $\boldsymbol{P}$.

Eq. (6) solves feature selection and label recovery simultaneously. Highly missing labels corrupt the estimation of the true label distribution, which is generally non-uniform. Hence, the model in Eq. (6) wherein the loss for predicting each label is equally weighed should be adjusted, to prevent the overall loss being dominated by the frequent labels, sacrificing the prediction accuracy of rare labels [41]. Thus, we weigh labeled instance in a tf-idf-like fashion so that losses from rare labels are given more weights during training. Specifically, each label on the i-th instance is assigned with a cost $\varsigma_i=\frac{1}{c_i}$, where c_i is the number of times the label i appears in the training set. Thus, the matrix $\boldsymbol{S}$ is re-defined as

$$s_{ii}=\sum_{j\in\mathcal{N}_y(i)}\varsigma_j, \quad \text{if } \boldsymbol{x}_i \text{ is labeled} \tag{7}$$

where $\mathcal{N}_y(i)$ represents the indexes set of the positive labels of the i-th instance (i.e., positive label index in $\boldsymbol{y}_i$). By incorporating Eq. (7) into Eq. (6), the contributions of rare labels in label reconstruction are emphasized, and this is more consistent with the true label distributions to some extent.

In summary, SMFS makes the first attempt to extract the intrinsic local structure of feature space solely based on data distribution, which consequently help encode the intricate feature-label dependencies and recover labels based on local structure reconstruction in multi-label feature selection. The recovered label matrix can be considered as a regularization of selecting diverse features to discriminate different labels, considering that these labels have distinct characteristics. Consequently, the learning processes of selecting discriminative features and recovering incomplete labels are mutually optimized until convergence.

IV. ALTERNATING OPTIMIZATION

In this section, we present an optimization algorithm for the proposed SMFS to solve its non-smooth objective function involving $l_{2,1}$-norm. Detailed convergence analysis and computational complexity analysis are available in Section V.

A. CONSTRUCT LOCAL STRUCTURE

we use a recursive partition scheme called *iForest* [42] to implement probability mass estimation, as used in [11], which can be divided into two steps. The first step is to build an *iForest* consisting of τ *iTrees* as the partition structure. The second step is the evaluation step. Instances $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ are passed through each *iTree* to find the mass of the deepest node that contains $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$. $\xi\left(\boldsymbol{x}_i,\boldsymbol{x}_j\right)$ is the mean of these mass values over τ *iTrees* based on Eq. (2). Finally, local density structure in feature space is constructed based on the k lowest probability mass neighbors that are derived directly from data, and the weight matrix of local density structure $\boldsymbol{W}$ is computed by Eq. (3).

We give Proposition 1 to prove that the optimization problem in Eq. (14) is jointly convex with respect to $\boldsymbol{P}$ and $\boldsymbol{Y}$.

Proposition 1: Denote $\boldsymbol{P} \in \mathbb{R}^{d\times c}$, $\boldsymbol{Y} \in \mathbb{R}^{n\times c}$, $\boldsymbol{W}_Y \in \mathbb{R}^{n\times n}$. The minimization of $\psi\left(\boldsymbol{P},\boldsymbol{Y}\right)$ is jointly convex with respect to $\boldsymbol{P}$ and $\boldsymbol{Y}$.

Proof: (Proof Sketch): Each term in $\psi\left(\boldsymbol{P},\boldsymbol{Y}\right)$ is positive semi-definite. Thus, minimizing $\psi\left(\boldsymbol{P},\boldsymbol{Y}\right)$ is a convex problem.

Hence, the objective function in Eq. (6) can achieve its optimal value by optimizing variables $\boldsymbol{P}$ and $\boldsymbol{Y}$. Due to the high complexity of optimizing multiple variables simultaneously, we convert the problem into a series of sub-solution processes in which only one variable is involved.

B. UPDATE P

when $\boldsymbol{Y}$ is fixed, optimizing Eq. (6) becomes the following problem w.r.t variable $\boldsymbol{P}$:

$$\min_{\boldsymbol{P}} Tr\left(\left(\boldsymbol{XP}-\boldsymbol{Y}\right)^T\boldsymbol{S}\left(\boldsymbol{XP}-\boldsymbol{Y}\right)\right)+\beta Tr\left(\boldsymbol{P}^T\boldsymbol{DP}\right), \tag{8}$$

where $\boldsymbol{D}$ is a diagonal matrix where D_{ii} is defined as

$$\boldsymbol{D}=\begin{pmatrix} \frac{1}{2\|\boldsymbol{P}^1\|_2} & & \\ & \ddots & \\ & & \frac{1}{2\|\boldsymbol{P}^d\|_2} \end{pmatrix}. \tag{9}$$

Since $\boldsymbol{D}$ is related to $\boldsymbol{P}$, it is difficult to solve this problem. Hence, we can obtain $\boldsymbol{D}$ with initialized $\boldsymbol{P}$ in an iterative manner. Then, we have

$$\boldsymbol{P}=\left(\boldsymbol{X}^T\boldsymbol{SX}+\beta\boldsymbol{D}\right)^{-1}\left(\boldsymbol{X}^T\boldsymbol{SY}\right). \tag{10}$$

C. UPDATE Y

when $\boldsymbol{P}$ is fixed, the optimization problem in Eq. (6) becomes the following problem w.r.t. variable $\boldsymbol{Y}$:

$$\min_{\boldsymbol{Y}} Tr\left(\left(\boldsymbol{XP}-\boldsymbol{Y}\right)^T\boldsymbol{S}\left(\boldsymbol{XP}-\boldsymbol{Y}\right)\right)+\alpha\sum_i\|\boldsymbol{y}_i-\sum_j w_{ij}\boldsymbol{y}_j\|_2^2. \tag{11}$$

Due to the constraint $\mathbf{1}^T \boldsymbol{w}_i = 1$ $(\forall i)$, the reconstruction error in terms of label structure is rearranged, and thus Eq. (11) is rewritten as

$$\min_{\boldsymbol{Y}} Tr\left((\boldsymbol{XP}-\boldsymbol{Y})^T \boldsymbol{S} (\boldsymbol{XP}-\boldsymbol{Y}) \right) + \alpha Tr\left(\boldsymbol{Y}^T \boldsymbol{W}_Y \boldsymbol{Y}\right), \quad (12)$$

where $\boldsymbol{W}_Y = (\boldsymbol{I}-\boldsymbol{W})^T (\boldsymbol{I}-\boldsymbol{W}) \in \mathbb{R}^{n\times n}$ and $\boldsymbol{I}$ is the identity matrix. Then, the optimal label matrix is calculated as

$$\boldsymbol{Y} = (\boldsymbol{SI} + \alpha \boldsymbol{W}_Y)^{-1} \boldsymbol{SXP}. \quad (13)$$

In each iteration, y_{ij} needs to be set in the range of $[-1, 1]$ to avoid a trivial solution. Hence, we set $y_{ij} = 1$ if $y_{ij} > 1$ and $y_{ij} = -1$ if $y_{ij} < -1$. By computing optimal $\boldsymbol{W}$ and iteratively updating $\boldsymbol{P}$ and $\boldsymbol{Y}$, Eq. (6) monotonically decreases and finally converges to a fixed point demonstrated in Section V. To obtain the accurate labels, we set $y_{ij} = 1$ if $y_{ij} > 0$ and $y_{ij} = -1$ if $y_{ij} < 0$ at the end of the iteration.

V. THEORETICAL ANALYSIS

In this section, we prove the the convergence of our proposed SMFS and analyse the computational complexity.

For convenience, we reformulate the objective function as

$$\begin{aligned} \psi(\boldsymbol{P}, \boldsymbol{Y}) = & \left\| \sqrt{\boldsymbol{S}}(\boldsymbol{XP}-\boldsymbol{Y}) \right\|_F^2 + \beta \left\| \sqrt{\boldsymbol{D}}\boldsymbol{P} \right\|_F^2 \\ & + \alpha Tr\left(\boldsymbol{Y}^T \boldsymbol{W}_Y \boldsymbol{Y}\right), \\ s.t. \quad & \mathbf{1}^T \boldsymbol{w}_i = 1, \quad 0 \le w_{ij} \le 1, \quad -1 \le y_{ij} \le 1, \end{aligned} \quad (14)$$

where $\boldsymbol{W}_Y = (\boldsymbol{I}-\boldsymbol{W})^T (\boldsymbol{I}-\boldsymbol{W}) \in \mathbb{R}^{n\times n}$, $\boldsymbol{D}$ is a diagonal matrix, and $\boldsymbol{I}$ is the identity matrix.

A. CONVERGENCE ANALYSIS

Lemma 1: Given an optimization problem:

$$\min_{\boldsymbol{Z}} f(\boldsymbol{Z}) + \|\sqrt{\boldsymbol{R}}\Phi(\boldsymbol{Z})\|_F^2, \quad s.t. \boldsymbol{Z} \in \mathcal{F}, \quad (15)$$

where $f(\boldsymbol{Z})$ and $\Phi(\boldsymbol{Z})$ are matrix functions of $\boldsymbol{Z}$, $\mathcal{F}$ is the feasible region, and $\boldsymbol{R}$ is a diagonal matrix where R_{ii} is defined as $\frac{1}{2\|\Phi(\boldsymbol{Z}_0)\|_2}$. If $\boldsymbol{Z}^\star$ is the optimal solution of the above optimization problem Eq. (15), we have

$$f(\boldsymbol{Z}^\star) + \|\Phi(\boldsymbol{Z}^\star)\|_{2,1} \le f(\boldsymbol{Z}_0) + \|\Phi(\boldsymbol{Z}_0)\|_{2,1}. \quad (16)$$

Proof: Since $\boldsymbol{Z}^\star$ is the optimal solution of Eq. (16), we have

$$f(\boldsymbol{Z}^\star) + \|\sqrt{\boldsymbol{R}}\Phi(\boldsymbol{Z}^\star)\|_F^2 \le f(\boldsymbol{Z}_0) + \|\sqrt{\boldsymbol{R}}\Phi(\boldsymbol{Z}_0)\|_F^2. \quad (17)$$

Therefore

$$f(\boldsymbol{Z}^\star) + \sum_i \frac{1}{2}\frac{\|\Phi(\boldsymbol{Z}^\star)_i\|_2^2}{\|\Phi(\boldsymbol{Z}_0)\|_2} \le f(\boldsymbol{Z}_0) + \sum_i \frac{1}{2}\|\Phi(\boldsymbol{Z}_0)_i\|_2. \quad (18)$$

We have

$$\sum_i (\|\Phi(\boldsymbol{Z}^\star)_i\|_2 - \frac{1}{2}\frac{\|\Phi(\boldsymbol{Z}^\star)_i\|_2^2}{\|\Phi(\boldsymbol{Z}_0)\|_2}) \le \sum_i (1 - \frac{1}{2}\|\Phi(\boldsymbol{Z}_0)_i\|_2). \quad (19)$$

Summing Eq. (18) and Eq. (19), we obtain

$$f(\boldsymbol{Z}^\star) + \|\Phi(\boldsymbol{Z}^\star)\|_2 \le f(\boldsymbol{Z}_0) + \|\Phi(\boldsymbol{Z}_0)\|_2. \quad (20)$$

Finally, we obtain

$$f(\boldsymbol{Z}^\star) + \|\Phi(\boldsymbol{Z}^\star)\|_{2,1}^2 \le f(\boldsymbol{Z}_0) + \|\Phi(\boldsymbol{Z}_0)\|_{2,1}^2, \quad (21)$$

where the equality holds if and only if $\Phi(\boldsymbol{Z}^\star) = \Phi(\boldsymbol{Z}_0)$.

Algorithm 1: SMFS: Structure-Induced Multi-Label Feature Selection

Input: Training data $\boldsymbol{X}$, highly incomplete label matrix $\boldsymbol{Y}$, scale of the local structure k;
Output: Feature selection matrix $\boldsymbol{P}$, recovered label matrix $\boldsymbol{Y}$;

```
Set t = 0 and initialize P_0 with Gaussian distribution;
Estimate probability mass w.r.t X according to Eq. (2);
Construct local density structure of the feature space
  based on k lowest probability mass neighbors;
Compute the weight matrix W according to Eq. (3);
Estimate missing labels of labeled instances via Eq. (4);
repeat
    Compute D_t according to Eq. (9);
    Compute P_{t+1} according to Eq. (10);
    Compute Y_{t+1} according to Eq. (11);
    t = t + 1;
until convergence;
return P* and Y*.
```

Theorem 1: The alternate updating rules in Algorithm 1 monotonically decreases the objective function value in each iteration until convergence.

Proof: First, we define

$$\boldsymbol{P}_t = \arg\min_{\boldsymbol{P}} \left\| \sqrt{\boldsymbol{S}}(\boldsymbol{XP}_{t-1}-\boldsymbol{Y}) \right\|_F^2 + \beta \left\| \sqrt{\boldsymbol{D}}\boldsymbol{P}_{t-1} \right\|_F^2. \quad (22)$$

According to Eq. (22), we have

$$\begin{aligned} & \left\| \sqrt{\boldsymbol{S}}(\boldsymbol{XP}_t - \boldsymbol{Y}) \right\|_F^2 + \alpha Tr\left(\boldsymbol{Y}^T \boldsymbol{W}_Y \boldsymbol{Y}\right) \\ & + \beta \left\| \sqrt{\boldsymbol{D}_{t-1}}\boldsymbol{P}_t \right\|_F^2 \le \left\| \sqrt{\boldsymbol{S}}(\boldsymbol{XP}_{t-1} - \boldsymbol{Y}) \right\|_F^2 \\ & + \alpha Tr\left(\boldsymbol{Y}^T \boldsymbol{W}_Y \boldsymbol{Y}\right) + \beta \left\| \sqrt{\boldsymbol{D}_{t-1}}\boldsymbol{P}_{t-1} \right\|_F^2. \end{aligned} \quad (23)$$

According to lemma 1, we have

$$\begin{aligned} & \left\| \sqrt{\boldsymbol{S}}(\boldsymbol{XP}_t - \boldsymbol{Y}) \right\|_F^2 + \alpha Tr\left(\boldsymbol{Y}^T \boldsymbol{W}_Y \boldsymbol{Y}\right) + \beta \left\| \sqrt{\boldsymbol{D}_t}\boldsymbol{P}_t \right\|_F^2 \\ & \le \left\| \sqrt{\boldsymbol{S}}(\boldsymbol{XP}_{t-1} - \boldsymbol{Y}) \right\|_F^2 + \alpha Tr\left(\boldsymbol{Y}^T \boldsymbol{W}_Y \boldsymbol{Y}\right) \\ & + \beta \left\| \sqrt{\boldsymbol{D}_{t-1}}\boldsymbol{P}_{t-1} \right\|_F^2. \end{aligned} \quad (24)$$

By combining Eq. (23) and Eq. (24), this implies that

$$\psi(\boldsymbol{P}_t, \boldsymbol{Y}_{t-1},) \le \psi(\boldsymbol{P}_{t-1}, \boldsymbol{Y}_{t-1}). \quad (25)$$

Second, we fix $\boldsymbol{P}$ as $\boldsymbol{P}_t$ to optimize $\boldsymbol{Y}$, we have

$$\begin{aligned} \boldsymbol{Y}_t = arg\min_{\boldsymbol{Y}} & Tr\left((\boldsymbol{XP}_t - \boldsymbol{Y}_{t-1})^T \boldsymbol{S} (\boldsymbol{XP}_t - \boldsymbol{Y}_{t-1}) \right) \\ & + \alpha Tr\left(\boldsymbol{Y}_{t-1}^T \boldsymbol{W}_Y \boldsymbol{Y}_{t-1}\right). \end{aligned} \quad (26)$$

For local label structure reconstruction, the weight of local density structure in the feature space is constrained to be positive, thus it is easy to prove that the solution of Eq. (26) w.r.t $\boldsymbol{Y}$ is monotonically decreasing [37]. Thus we have

$$\psi\left(\boldsymbol{P}_t, \boldsymbol{Y}_t, \right) \leq \psi\left(\boldsymbol{P}_{t-1}, \boldsymbol{Y}_{t-1}\right). \tag{27}$$

Based on Eq. (25) and Eq. (27), the objective function value of Eq. (6) decreases monotonically in each iteration until the algorithm convergence. The proof is completed.

Theorem 2: Sequence $\{\boldsymbol{P}_t\}$ and $\{\boldsymbol{Y}_t\}$ produced in Algorithm 1 converges, and the limit point is a stationary point of Eq. (14).

Proof: Eq. (14) is a convex optimization problem, hence its solution obtained by Algorithm 1 is globally optimal.

In summary, the objective function value of Eq. (14) decreases monotonically in each iteration until the algorithm converges to a stationary point. The proof is completed.

B. TIME COMPLEXITY ANALYSIS

The time consumption of SMFS mainly lies in three parts. In local structure construction, structure is built by *iForest* and the weight matrix $\boldsymbol{W}$ is computed, which costs $\mathcal{O}\left(\tau\varphi\log\varphi + n^2\tau\log\varphi\right)$. In large datasets, $\varphi \ll n$; thus the cost is approximately equal to $\mathcal{O}\left(n^2\right)$. Then, the major computational cost lies in updating of variable $\boldsymbol{P}$ and $\boldsymbol{Y}$, the computational cost of matrix inverse can be avoided by iterative optimization [43]. In updating $\boldsymbol{P}$, the complexity is $\mathcal{O}(Td^2c + nd^2 + ndc)$, where T is the number of iterations. As c is smaller than d and n in real situations, the complexity of updating $\boldsymbol{P}$ can be calculated as $\mathcal{O}(Td^2 + nd^2)$, which is linear to the number of instances n. The cost of updating $\boldsymbol{Y}$ becomes $\mathcal{O}(Tn^2c + ndc)$. In fact, due to the k-sparsity of matrix $\boldsymbol{W}_Y$ (i.e., each column of $\boldsymbol{W}_Y$ has k nonzero elements and $k \ll n$), the complexity drops to $\mathcal{O}(Tnk + nd)$ in practice. Thus, the complexity of SMFS is summarized as $\mathcal{O}(Td^2 + nd^2 + Tnk)$ with updated $\boldsymbol{P}$ and $\boldsymbol{Y}$ in local structure construction with additional $\mathcal{O}(n^2)$.

VI. EMPIRICAL STUDY

Six groups of multi-label data sets fetched from Mulan library [3] are tested, as shown in Table 1. The benchmarks cover various domains, including audio, music, text, image, biology, and all of the numerical features are normalized with zero mean and unit variance in the experiments.

TABLE 1. Data sets description.

Data sets	Instances	Features	Labels	Domain
Birds	645	260	19	audio
Emotions	593	72	6	music
Enron	1702	1001	53	text
Language log	1459	1004	75	text
Scene	2407	294	6	image
Yeast	2417	103	14	biology

[3]http://mulan.sourceforge.net/datasets.html.

A. EXPERIMENTAL SETTINGS

In this study, we pioneerly deal with the learning problem of highly incomplete labels in multi-label feature selection. We investigate the performance of the proposed SMFS by comparing with the following baselines, including semi-supervised multi-label feature selection and multi-label feature selection with missing labels approaches (half of them were proposed within 3 years):

- All-Fea: all of the original features without any selection are tested.
- SFSS [5]: a state-of-the-art semi-supervised multimedia data analysis approach incorporating feature selection.
- CSFS [32]: a popular semi-supervised multi-label feature selection suitable for large-scale datasets without graph Laplacian matrix construction.
- FSCLA [6]: a recently proposed semi-supervised multi-label learning approach combining manifold learning with shared subspace construction to select discriminative features.
- SCFS [7]: a latest semi-supervised multi-label feature selection that extracts label correlations by maintaining feature-label space consistency.
- MLMLFS [4]: a supervised multi-label feature selection approach that firstly handles the missing labels through robust linear regression model.

For a fair comparison, the trade-off parameters are tuned for the baselines by five-fold cross validations with a "grid-search" strategy from $\left\{10^{-3}, 10^{-2}, 10^{-1}, 1, 10^{1}, 10^{2}\right\}$. We use *iForest* with the default setting (i.e., $\tau = 100$) [11] to accomplish the partition for the probability mass estimation. ML-KNN [44] with default parameter is chosen as the classifier due to its effectiveness that has been verified in many state-of-the-art works [6], [19]. Each approach selects $\{\frac{d}{6}, \frac{2d}{6}, \frac{3d}{6}, \frac{4d}{6}, \frac{5d}{6}\}$ features to build the ML-KNN classifier, where d is the number of original features. Here, the missing label ratio is set to 20% and 40% by randomly dropping the observed labels from the labeled training data [8], [45]; the semi-supervised label ratio is set to 10% and 30%. We report the mean average precision (MAP) and standard deviation averaged across ten independent runs on each dataset with different size of features, where each run randomly splits the original dataset into training and testing subsets with the 80/20 ratio.

B. CLASSIFICATION PERFORMANCE

We compare SMFS with the baselines in different learning scenarios respective with 10% and 30% labeled instances, in which the missing labels occupy 20% and 40%, and the results are recorded in Table 2 and Table 3.

The results indicate the following. (1) The compared feature selection approaches outperform ALL-Fea in most cases. In particular, SMFS achieves approximately $2\% - 8\%$ improvement across all benchmarks. This result shows that a discriminative reduced space constructed by selecting relevant features is beneficial for promoting

TABLE 2. MAP score(±standard deviation) when labeled data ratio (LR) is 10%, missing label ratio (MLR) is 20% and 40%. The best result and those not significantly worse than it are highlighted in bold (pairwise t-test at 5% significance level).

Data sets		Approaches						
	MLR	All−Fea	SFSS	CSFS	FSCLA	SCFS	MLMLFS	SMFS
Birds	20%	0.388±0.032	**0.431±0.042**	0.420±0.049	0.425±0.039	0.384±0.019	0.385±0.023	**0.437±0.035**
	40%	0.387±0.016	0.395±0.038	0.398±0.033	0.405±0.041	0.371±0.031	0.377±0.018	**0.422±0.027**
Emotions	20%	0.667±0.013	0.650±0.030	0.673±0.028	0.664±0.037	0.676±0.024	0.678±0.024	**0.710±0.009**
	40%	0.656±0.014	0.622±0.038	0.667±0.041	0.625±0.041	0.663±0.039	0.628±0.036	**0.682±0.022**
Enron	20%	0.547±0.020	0.542±0.002	0.543±0.009	0.555±0.013	0.571±0.014	0.575±0.023	**0.586±0.018**
	40%	0.500±0.010	0.528±0.014	0.532±0.009	0.535±0.013	0.529±0.036	0.533±0.025	**0.563±0.014**
Language log	20%	0.619±0.019	0.595±0.021	**0.631±0.009**	0.590±0.051	0.598±0.040	0.622±0.016	**0.637±0.009**
	40%	0.592±0.035	0.556±0.046	0.617±0.016	0.560±0.014	0.590±0.048	0.565±0.024	**0.627±0.010**
Scene	20%	0.789±0.011	0.622±0.019	0.718±0.030	0.658±0.024	0.778±0.018	0.765±0.026	**0.811±0.008**
	40%	0.770±0.017	0.581±0.028	0.669±0.018	0.609±0.020	0.730±0.030	0.713±0.025	**0.799±0.008**
Yeast	20%	0.663±0.015	0.704±0.015	0.712±0.013	0.712±0.016	0.707±0.017	**0.740±0.013**	**0.737±0.010**
	40%	0.645±0.012	0.698±0.011	0.707±0.012	0.698±0.012	0.679±0.015	**0.732±0.015**	**0.728±0.013**

TABLE 3. MAP score(±standard deviation) when LR is 30%, MLR is 20% and 40%.The best result and those not significantly worse than it are highlighted in bold (pairwise t-test at 5% significance level).

Data sets		Approaches						
	MLR	All−Fea	SFSS	CSFS	FSCLA	SCFS	MLMLFS	SMFS
Birds	20%	0.400±0.038	0.501±0.011	0.481±0.053	0.532±0.039	0.493±0.037	0.521±0.020	**0.545±0.033**
	40%	0.397±0.035	0.473±0.023	0.438±0.036	0.497±0.024	0.462±0.034	0.478±0.026	**0.536±0.036**
Emotions	20%	0.696±0.040	0.687±0.028	0.728±0.028	0.707±0.021	0.733±0.016	0.735±0.033	**0.759±0.019**
	40%	0.685±0.019	0.688±0.025	0.708±0.033	0.676±0.018	0.691±0.041	0.703±0.018	**0.736±0.024**
Enron	20%	0.560±0.048	0.579±0.017	0.577±0.015	0.570±0.015	0.554±0.032	0.605±0.016	**0.614±0.011**
	40%	0.559±0.014	0.553±0.019	0.576±0.005	0.548±0.018	0.545±0.055	0.580±0.014	**0.592±0.014**
Language log	20%	0.623±0.014	0.617±0.012	**0.643±0.002**	0.617±0.014	0.627±0.015	0.614±0.014	**0.646±0.009**
	40%	0.606±0.007	0.618±0.020	0.621±0.008	0.601±0.020	0.578±0.018	0.607±0.014	**0.638±0.015**
Scene	20%	0.797±0.012	0.774±0.015	0.801±0.011	0.792±0.016	0.815±0.013	0.816±0.001	**0.833±0.011**
	40%	0.816±0.009	0.739±0.019	0.773±0.018	0.757±0.015	0.795±0.014	0.747±0.009	**0.826±0.014**
Yeast	20%	0.682±0.011	0.732±0.008	**0.744±0.013**	0.734±0.013	0.717±0.021	0.731±0.015	**0.748±0.008**
	40%	0.660±0.007	0.722±0.019	**0.735±0.007**	0.712±0.010	0.695±0.021	**0.737±0.009**	**0.738±0.009**

learning performance. (2) SMFS yields better performance than four semi-supervised multi-label approaches (i.e., SFSS, CSFS, FSCLA, and SCFS) on the benchmarks of Emotions, Enron, and Scene. This result indicates that label structure reconstruction based on local density structure in the feature space gains a competitive advantage in recovering the highly incomplete label matrix. (3) SMFS generally outperforms MLMLFS, indicating that the unlabeled data possess potential guide information that is beneficial for selecting excellent features. (4) When the densities between labels tend to be equal in the overlap region, local density structure depending on data distributions approximates local geometry structure, thus our approach exhibits the same remarkable performance as CSFS and MLMLFS do on the benchmark Yeast. (5) SMFS is a comparatively safe approach as it maintains a comparable or slightly decline performance when the missing label ratio rises from 20% to 40%. With many labels missing, the label structure is destroyed to a great extent, which will be further evaluated in Section VI-D.

Generally, SMFS is superior to the benchmarks under the evaluation of MAP across different highly incomplete label learning scenarios. This finding validates the effectiveness of SMFS in selecting discriminative features under the intricate recognition situation with highly incomplete labels. One of the major factors that account for the superiority of SMFS is its strategy of reconstructing incomplete labels via transferring the local density structure in the feature space, as evaluated in Section VI-C.

C. LOCAL STRUCTURE ANALYSIS

In this paper, the local density structure that adapts to the density structure of a datasets is the key advantage to preserve reliable local correlation information in the highly incomplete label learning scenarios. Hence, we evaluate the effects of local density structure w.r.t. preserving local correlations by varying the number of lowest probability mass neighbors k from 2 to 25, utilizing Scene as the benchmark with 40% missing labels of 30% labeled data, and using four metrics, namely, MAP, Hamming Loss, One Error, and Ranking Loss [5], [11].

Table 4 shows that the scale of local structure (i.e., closest matching neighbors k) makes effect on the performance of the proposed SMFS, where the best results on each metric averaged across ten independent runs are shown in bold face.

TABLE 4. Experimental results of SMFS on Scene benchmark with different scale of local structure considered.

Evaluation metric	Scale of local structure					
	k=2	k=5	k=10	k=15	k=20	k=25
MAP	0.818±0.014	0.822±0.013	0.825±0.007	**0.826±0.014**	0.808±0.017	0.810±0.011
Hamming Loss	0.137±0.004	0.128±0.007	0.128±0.006	**0.126±0.005**	0.135±0.010	0.132±0.010
One Error	0.112±0.005	0.111±0.007	**0.109±0.006**	**0.109±0.008**	0.122±0.012	0.121±0.006
Ranking Loss	0.296±0.028	**0.287±0.023**	**0.287±0.009**	0.289±0.019	0.311±0.028	0.306±0.018

The results also reflect that a small or large scale of the local structure may degrade the learning performance, because of missing useful neighbor information or incorporating noisy information. The results of SMFS reported throughout the paper are obtained with a moderate value of k as 15.

To assess the performance of SMFS in constructing the local density structure in the feature space, we compare SMFS with two local geometry structure-based approaches that have been widely applied in multi-label learning:

- LGS-D: The local geometry structure is constructed by determining the k nearest neighbors through a distance metric; the weights of structure are also defined via a distance metric [4]–[7];
- LGS-L: A distance-based k nearest neighbor strategy is adopted in local geometry structure construction; the weights are defined based on linear combination of neighbors [34], [37], [39].

We perform a simple comparison between the local density structure and the local geometry structure. We replace the strategy of constructing the feature structure in SMFS with LGS-D and LGS-L. Their selection performance in the cases of 20%, 40%, and 60% missing labels of 30% labeled instances on Scene benchmark is demonstrated in Fig. 3. In the figure, SMFS consistently outperforms the other two approaches under the evaluation of MAP. This result reveals that the local density structure in feature space that is dependent on data distribution that contributes to the improvement of recovering highly incomplete labels and selecting discriminative features in the subsequent learning process. In other words, this kind of structure can encode the intrinsic feature-label dependencies to help effective model construction in the wild and complex multi-label learning scenarios.

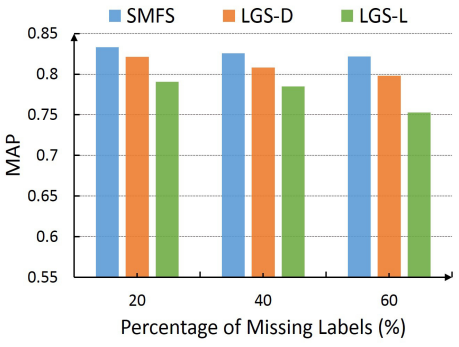


FIGURE 3. Comparisons of three local structure construction approaches.

We provide a simple comparison between our probabilistic mass strategy and the distance strategy in Fig. 4. The structure of the majority label (i.e., the label that possesses the largest number of instances) in the Birds benchmark is respectively constructed by Euclidean distance and probability mass estimation, and visualized by the multidimensional scaling (MDS) [4]. Positive and negative instances are scatteredly distributed in the structure built by probability mass, indicating easier to be separated than in the structure built by instance distances.

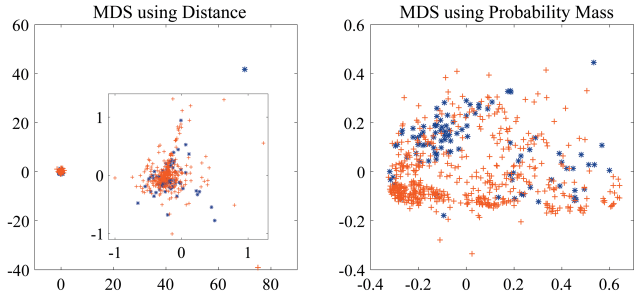
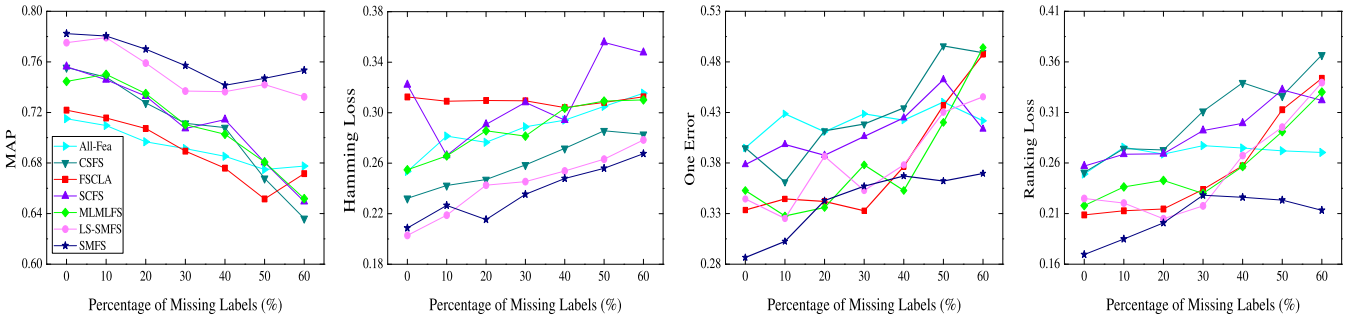


FIGURE 4. MDS demonstration for the affinity matrix of majority-label instances in Birds: Euclidean distance and probability mass estimation are used to construct the affinity matrix, and the blue and red points respectively represent the positive and negative instances.

D. EFFECTS OF THE SIZE OF MISSING LABELS

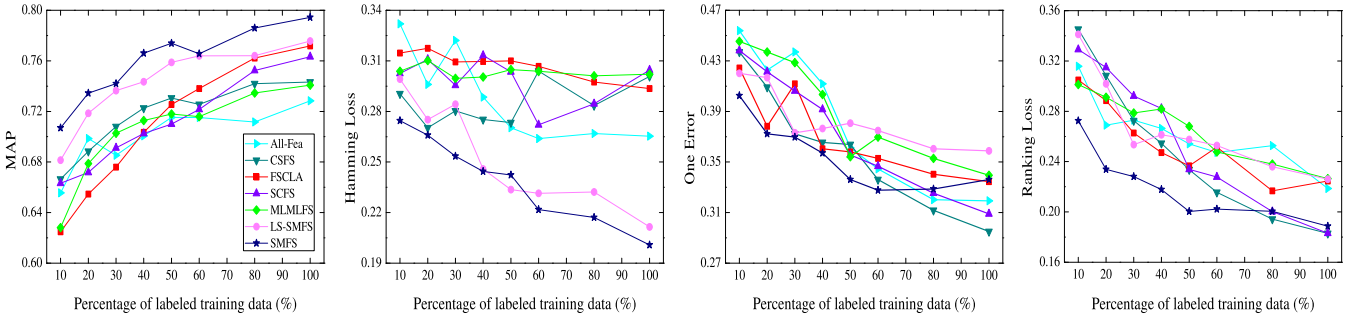
To evaluate the effects of the damaged label structure on multi-label feature selection, we vary the missing label ratio of the 30% labeled training data on Emotions as {0%, 10%, 20%, 30%, 40%, 50%, 60%} and demonstrate the performance of the compared approaches under for metrics, namely, MAP, Hamming Loss, One Error and Ranking Loss, as shown in Fig. 5. The followings can be observed. (1) With the scale of missing labels increasing, the selection performance of the compared approaches tends to deteriorate, and SMFS presents a relatively slower decline trend than the contrast groups, indicating its effectiveness in different complex recognition situations. (2) When the missing label ratio is 0%, SMFS degenerates to the semi-supervised multi-label approach, and still yields a relatively excellent performance.

[4]MDS is a popular technique for visualizing the information contained in an affinity relation matrix [46].



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

FIGURE 5. Variations of MAP, Hamming Loss, One Error, and Ranking Loss with increasing the percentage of missing labels on Emotions. In terms of the MAP metric, the higher score indicates the better performance, which the others are contrary.



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

FIGURE 6. Variations of MAP, Hamming Loss, One Error, and Ranking Loss with increasing the percentage of labeled data on Emotions. In terms of the MAP metric, the higher score indicates the better performance, which the others are contrary.

Hence, SMFS is applicable to semi-supervised learning tasks and can be expected to accomplish dimension reduction tasks in a broad variety of multi-label learning scenarios. (3) The performance of the approaches that directly employ label correlations to select features, such as SFSS, FSCLA, and SCFS, is significantly influenced when a large number of labels are missing because of the damaged label structure and correlations.

E. EFFECTS OF THE SIZE OF LABELED DATA

To further assess the performance of SMFS in different learning scenarios, we vary the labeled data ratio with 40% missing labels on Emotions as {10%, 20%, 30%, 40%, 50%, 60%, 80%, 100%} under four metrics, namely, MAP, Hamming Loss, One Error and Ranking Loss, as shown in Fig. 6. The followings can be observed. (1) SMFS performs superior to MLMLFS, which indicates that the underlying guide information in unlabeled data can help select informative features and effectively recover highly incomplete labels. (2) When label information is relatively mostly available (i.e., labeled data ratio are above 80%), SMFS is comparative with SFSS, FSCLA, and SCFS, illustrating that inherent label correlations benefit for the selection of discriminative feature. (3) SMFS considerably beats the baselines under various metrics in the severe learning situations wherein large portions of data are unannotated (i.e., labeled data ratio is below 60%), mostly available in terms of MAP and Ranking Loss, indicating that SMFS provides a deep insight into the abundant information possessed by the unlabeled data and effectively utilizes this information into model construction.

F. EFFECTS OF THE NUMBER OF SELECTED FEATURES

In this section, we conduct experiments to analyse the performance of SMFS with different size of selected features, employing Language log dataset with 20% missing labels of 30% labeled training data. Fig. 7 shows that the MAP score yields by SMFS on the benchmark when the number of selected features increases from 167 to 1004 (i.e., the original

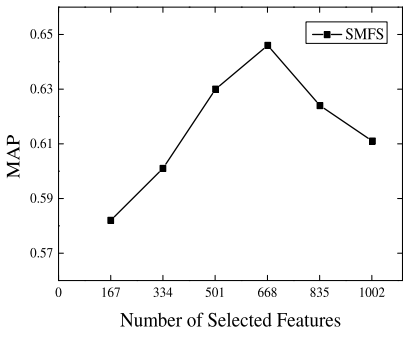
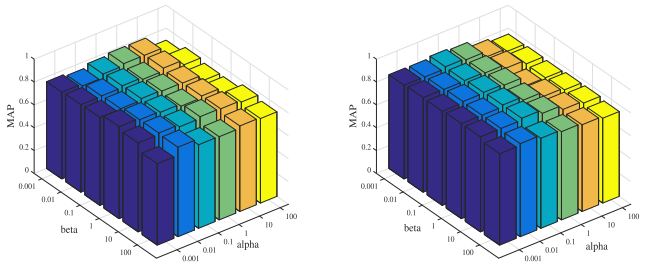
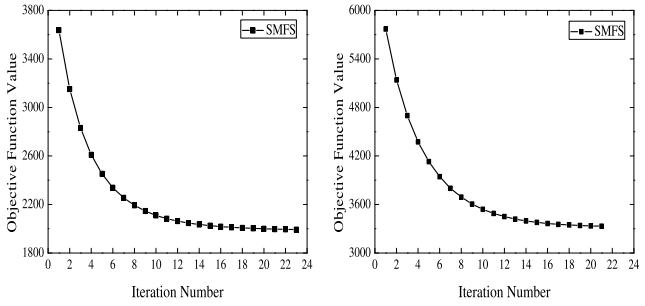


FIGURE 7. Variations of MAP with increasing the number of selected features.



(a) Emotions (b) Scene

FIGURE 8. Variations of MAP with different α and β.



(a) Emotions (b) Scene

FIGURE 9. Convergence curves of SMFS.

feature figure). We can observe that (1) the performance of SMFS is improved with increasing the size of selected features from 167 to 669, duo to more discriminative features are extracted into feature subset, and the highest MAP score is yielded when the selected number is equal to 669; and (2) the performance of SMFS gradually declines as the number of selected features continuously increased, where irrelevant and noise features exist in the feature set. Hence, there is no definite pattern about the best selected number across different benchmarks. In this paper, we vary the number of selected features and report the average selection performance for each baseline.

G. PARAMETER SENSITIVITY AND CONVERGENCE ANALYSIS

In this section, we conduct parameter sensitivity analysis for the proposed SMFS with 20% missing labels of 30% labeled data on Emotions and Scene benchmarks over the trade-off parameter α and β in terms of MAP. As reported in Fig. 8, we can see, within the considered range of values, SMFS is relatively nonsensitive to the variations of α and β and remains a stable performance over different parameter configurations.

We experimentally study the speed of convergence of SMFS with 20% missing labels of 30% labeled data on Emotions and Scene benchmarks, as shown in Fig. 9. Trade-off parameters (i.e., α and β) are set to 1, which is a median value in the tuned range. We can see that the proposed approach converges within 30 iterations on two benchmarks, validating that SMFS is efficient in tackling the complex learning cases with highly incomplete prior knowledge.

VII. CONCLUSION

In this work, structure-induced multi-label feature selection approach is proposed to handle highly incomplete labels, which integrates two strategies, i.e., multi-label feature selection and label structure reconstruction, in a mutually beneficial manner. First, local density structure is captured, which facilitates better extracting intricate dependencies between features and labels. Then, local label structure is effectively reconstructed by the structure transferred from the feature space and provides complete label information to guide feature selection. A seamless fusion of both terms, i.e., reliable local information and complete label structure, contributes to the selection of discriminative features for multi-label recognition.

REFERENCES

[1] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *J. Mach. Learn. Res.*, vol. 3, pp. 1157–1182, Jan. 2003.
[2] Q. Gu, Z. Li, and J. Han, "Correlated multi-label feature selection," in *Proc. 20th ACM Int. Conf. Inf. Knowl. Manage. (CIKM)*, 2011, pp. 1087–1096.
[3] L. Jian, J. Li, K. Shu, and H. Liu, "Multi-label informed feature selection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2016, pp. 1627–1633.
[4] P. Zhu, Q. Xu, Q. Hu, C. Zhang, and H. Zhao, "Multi-label feature selection with missing labels," *Pattern Recognit.*, vol. 74, pp. 488–502, Feb. 2018.
[5] Z. Ma, F. Nie, Y. Yang, J. R. R. Uijlings, N. Sebe, and A. G. Hauptmann, "Discriminating joint feature analysis for multimedia data understanding," *IEEE Trans. Multimedia*, vol. 14, no. 6, pp. 1662–1672, Dec. 2012.
[6] X.-D. Wang, R.-C. Chen, C.-Q. Hong, Z.-Q. Zeng, and Z.-L. Zhou, "Semi-supervised multi-label feature selection via label correlation analysis with l_1-norm graph embedding," *Image Vis. Comput.*, vol. 63, pp. 10–23, 2017.
[7] 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. Manage. (CIKM)*, 2018, pp. 783–792.
[8] F. Zhao and Y. Guo, "Semi-supervised multi-label learning with incomplete labels," in *Proc. Int. Joint Conf. Artif. Intell.*, 2015, pp. 4062–4068.
[9] C. C. Aggarwal, A. Hinneburg, and D. A. Keim, "On the surprising behavior of distance metrics in high dimensional space," in *Proc. Int. Conf. Database Theory*, 2001, pp. 420–434.
[10] K. M. Ting, Y. Zhu, and Z.-H. Zhou, "Isolation kernel and its effect on SVM," in *Proc. 24th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Jul. 2018, pp. 2329–2337.
[11] K. M. Ting, Y. Zhu, M. Carman, Y. Zhu, and Z.-H. Zhou, "Overcoming key weaknesses of distance-based neighbourhood methods using a data dependent dissimilarity measure," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD)*, 2016, pp. 1205–1214.
[12] T. Joachims, "Text categorization with support vector machines: Learning with many relevant features," in *Proc. Eur. Conf. Mach. Learn.* Berlin, Germany: Springer, 1998, pp. 137–142.
[13] M.-L. Zhang and Z.-H. Zhou, "Multilabel neural networks with applications to functional genomics and text categorization," *IEEE Trans. Knowl. Data Eng.*, vol. 18, no. 10, pp. 1338–1351, Oct. 2006.
[14] L. J. Jensen, R. Gupta, H.-H. Staerfeldt, and S. Brunak, "Prediction of human protein function according to gene ontology categories," *Bioinformatics*, vol. 19, no. 5, pp. 635–642, Mar. 2003.
[15] H. Wang, L. Yan, H. Huang, and C. Ding, "From protein sequence to protein function via multi-label linear discriminant analysis," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 14, no. 3, pp. 503–513, May 2017.
[16] 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, Sep. 2004.
[17] X. Chang, H. Shen, S. Wang, J. Liu, and X. Li, "Semi-supervised feature analysis for multimedia annotation by mining label correlation," in *Proc. Pacific-Asia Conf. Knowl. Discovery Data Mining*, 2014, pp. 74–85.

[18] W. Liu, I. W. Tsang, and K.-R. Müller, "An easy-to-hard learning paradigm for multiple classes and multiple labels," *J. Mach. Learn. Res.*, vol. 18, no. 1, pp. 3300–3337, 2017.
[19] M.-L. Zhang and Z.-H. Zhou, "A review on multi-label learning algorithms," *IEEE Trans. Knowl. Data Eng.*, vol. 26, no. 8, pp. 1819–1837, Aug. 2014.
[20] J. Fürnkranz, E. Hüllermeier, E. L. Mencía, and K. Brinker, "Multilabel classification via calibrated label ranking," *Mach. Learn.*, vol. 73, no. 2, pp. 133–153, Nov. 2008.
[21] O. Dekel, Y. Singer, and C. D. Manning, "Log-linear models for label ranking," in *Proc. Adv. Neural Inf. Process. Syst.*, 2004, pp. 497–504.
[22] A. Elisseeff and J. Weston, "A kernel method for multi-labelled classification," in *Proc. Adv. Neural Inf. Process. Syst.*, 2002, pp. 681–687.
[23] J. Read, B. Pfahringer, G. Holmes, and E. Frank, "Classifier chains for multi-label classification," *Mach. Learn.*, vol. 85, no. 3, pp. 333–359, Dec. 2011.
[24] K.-H. Huang and H.-T. Lin, "Cost-sensitive label embedding for multi-label classification," *Mach. Learn.*, vol. 106, nos. 9–10, pp. 1725–1746, Oct. 2017.
[25] X.-Y. Jing, F. Wu, Z. Li, R. Hu, and D. Zhang, "Multi-label dictionary learning for image annotation," *IEEE Trans. Image Process.*, vol. 25, no. 6, pp. 2712–2725, Jun. 2016.
[26] C.-K. Yeh, W.-C. Wu, W.-J. Ko, and Y.-C. F. Wang, "Learning deep latent space for multi-label classification," in *Proc. AAAI Conf. Artif. Intell.*, 2017, pp. 1–7.
[27] N. Spolaôr, E. A. Cherman, M. C. Monard, and H. D. Lee, "Filter approach feature selection methods to support multi-label learning based on relieff and information gain," in *Proc. Brazilian Symp. Artif. Intell.*, 2012, pp. 72–81.
[28] A. Alalga, K. Benabdeslem, and N. Taleb, "Soft-constrained Laplacian score for semi-supervised multi-label feature selection," *Knowl. Inf. Syst.*, vol. 47, no. 1, pp. 75–98, Apr. 2016.
[29] A. Braytee, W. Liu, D. R. Catchpoole, and P. J. Kennedy, "Multi-label feature selection using correlation information," in *Proc. ACM Conf. Inf. Knowl. Manage. (CIKM)*, 2017, pp. 1649–1656.
[30] 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.
[31] F. Nie, H. Huang, X. Cai, and C. H. Ding, "Efficient and robust feature selection via joint $\ell_{2,1}$-norms minimization," in *Proc. Annu. Conf. Neural Inf. Process. Syst.*, 2010, pp. 1813–1821.
[32] X. Chang, F. Nie, Y. Yang, and H. Huang, "A convex formulation for semi-supervised multi-label feature selection," in *Proc. AAAI Conf. Artif. Intell.*, 2014, pp. 1171–1177.
[33] F. Nie, D. Xu, I. W.-H. Tsang, and C. Zhang, "Flexible manifold embedding: A framework for semi-supervised and unsupervised dimension reduction," *IEEE Trans. Image Process.*, vol. 19, no. 7, pp. 1921–1932, Jul. 2010.
[34] M. Luo, F. Nie, X. Chang, Y. Yang, A. G. Hauptmann, and Q. Zheng, "Adaptive unsupervised feature selection with structure regularization," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 4, pp. 944–956, Apr. 2018.
[35] Z. Ma, F. Nie, Y. Yang, J. R. R. Uijlings, and N. Sebe, "Web image annotation via subspace-sparsity collaborated feature selection," *IEEE Trans. Multimedia*, vol. 14, no. 4, pp. 1021–1030, Aug. 2012.
[36] Z. Zhang and J. Wang, "MLLE: Modified locally linear embedding using multiple weights," in *Proc. Annu. Conf. Neural Inf. Process. Syst.*, 2007, pp. 1593–1600.
[37] S. T. Roweis, "Nonlinear dimensionality reduction by locally linear embedding," *Science*, vol. 290, no. 5500, pp. 2323–2326, Dec. 2000.
[38] M. Guillaumin, T. Mensink, J. Verbeek, and C. Schmid, "TagProp: Discriminative metric learning in nearest neighbor models for image auto-annotation," in *Proc. IEEE 12th Int. Conf. Comput. Vis.*, Sep. 2009, pp. 309–316.
[39] P. Hou, X. Geng, and M. L. Zhang, "Multi-label manifold learning," in *Proc. AAAI Conf. Artif. Intell.*, 2016, pp. 1680–1686.
[40] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, pp. 2579–2605, Nov. 2008.
[41] M. Chen, A. Zheng, and K. Weinberger, "Fast image tagging," in *Proc. Annu. Conf. Neural Inf. Process. Syst.*, 2013, pp. 1274–1282.
[42] F. T. Liu, K. M. Ting, and Z.-H. Zhou, "Isolation forest," in *Proc. 8th IEEE Int. Conf. Data Mining*, 2008, pp. 413–422.
[43] D. Wang, F. Nie, and H. Huang, "Large-scale adaptive semi-supervised learning via unified inductive and transductive model," in *Proc. 20th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD)*, 2014, pp. 482–491.
[44] M.-L. Zhang and Z.-H. Zhou, "ML-KNN: A lazy learning approach to multi-label learning," *Pattern Recognit.*, vol. 40, no. 7, pp. 2038–2048, Jul. 2007.
[45] H.-C. Dong, Y.-F. Li, and Z.-H. Zhou, "Learning from semi-supervised weak-label data," in *Proc. AAAI Conf. Artif. Intell.*, 2018, pp. 2926–2933.
[46] F. Wickelmaier, "An introduction to MDS," Sound Qual. Res. Unit, Aalborg Univ., Aalborg, Denmarks, Tech. Rep., 2003, vol. 46, no. 5, pp. 1–26.

TIANTIAN XU received the B.E. and M.E. degrees in computer applications from the Qilu University of Technology, in 2012 and 2015, respectively, and the Ph.D. degree in software engineering from the Ocean University of China, in 2018. She is currently a Lecturer with the School of Computer Science and Technology, Qilu University of Technology (Shandong Academy of Sciences). She has published research articles at international journals. Her research interests include pattern recognition, association rules, sequential pattern mining, machine learning, and feature selection.

LONG ZHAO received the M.S. degree in computer science and technology from Shandong Polytechnic University, in 2009, and the Ph.D. degree from Wuhan University, in 2016. He is currently a Lecturer with the School of Computer Science and Technology, Qilu University of Technology (Shandong Academy of Sciences). His research interests include image processing, machine learning, and knowledge discovery.