IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received July 1, 2021, accepted August 17, 2021, date of publication August 23, 2021, date of current version August 30, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3106680

Robust Cross-View Embedding With Discriminant Structure for Multi-Label Classification

KAIXIANG WANG
School of Mathematical Sciences, Nanjing Normal University, Nanjing 210023, China
e-mail: nnuwangkaixiang@outlook.com

ABSTRACT Label embedding is an important family of multi-label classification algorithms which can jointly extract the information of all labels for better performance. However, few works have been done to develop the multi-label embedding methods that can effectively deal with the interference of noisy data during training process. The noise often makes the labels of a few samples incorrect (*i.e.*, missing or mislabeled), which could lead to a poor learning performance. To address this issue, we propose a novel cross-view based model. It performs a robust and discriminant embedding, namely Robust Cross-view Embedding with Discriminant Structure for Multi-label Classification (RCEDS). In RCEDS, a novel hypergraph fusion technique is designed to explore and utilize the complementary between the feature space and the label space to make the proposed RCEDS robust. Meanwhile, we use double-side metric learning to mine the consistency between the feature space and the label space to effectively improve the discriminative ability of our proposed RCEDS. Furthermore, we conduct a deep extension of RCEDS and effectively apply it to image annotation. Extensive experimental results on data sets with many labels demonstrate that our proposed approach can attain better classification performance than the existing label embedding algorithms.

INDEX TERMS Multi-label classification, label embedding, cross view, hypergraph fusion, double-side metric learning.

I. INTRODUCTION

Multi-label learning is an active research topic in the field of machine learning and pattern recognition. In the multi-label learning framework, each sample is represented by a feature vector, while it may belong to multiple categories. The goal is to induce a function that is able to assign multiple proper labels (from a given label set) to unseen instances [1], [2]. With the introduction of the concept of multi-label learning, many scholars have carried out research on this topic and put forward a lot of efficient algorithms.

However, increasing number of labels will result in the exponential increase of the number of label sets. As a consequence, the standard multi-label classification methods that work in original label space can easily become computationally impractical in training multi-label classifiers. Fortunately, there is usually some redundant information in the label space and the labels are universally correlated with each other. For this reason, some researchers began to study the method of dimensionality reduction in label space by using the label relationship. The expectation was to improve the classification accuracy while reducing the training and testing time of the whole model.

As an important family of the multi-label classification algorithms, a variety of label space reduction methods (LSDR) have been developed in the literature to address multi-label classification with many labels. LSDR algorithms [3]–[5] consider a low-dimensional embedded label space for digesting the information between labels and conducting more effective learning. However, as observed, a certain level of noise often exists in the real multi-label datasets that may degrade the performance of label embedding. The noise makes the labels of a few samples incorrect, which can be divided into two cases: (1) several essential labels are missing, and (2) several other labels are mislabeled. As shown in Figure 1, it is obvious that there is no 'road' in the picture located at row 1, column 1, so 'road' is the mis-assigned label while 'beach' is the missing label. These kinds of noisy data will increase the training error of classifiers and adversely affect the effectiveness of the multi-label classification.

To this end, we propose Robust Cross-view Embedding with Discriminant Structure for Multi-label Classification (RCEDS), which is illustrated in Figure 1. In DCERMC, the basic cross-view embedding is adopted to explore the

The associate editor coordinating the review of this manuscript and approving it for publication was Jolanta Mizera-Pietraszko.

correlations between the feature space and the label space. Meanwhile, in order to make our embedding method more robust and discriminative, we effectively utilize both the complementarity and consistency between the feature space and the label space. Specifically, in terms of complementarity, we construct corresponding hypergraphs in the feature space and the label space, and then fuse the two hypergraphs from the perspective of random walk. The high-order correlation of the instances in the feature space can effectively correct the inaccurate correlation in the label space caused by noisy data, and such an anti-noise ability can make our method robust. In terms of consensus, we use double-side metric learning to reduce the distances among a group of samples which are close in both feature space and label space consistently. Similar samples in both original feature space and label space will be closer in the latent space, which dramatically enhances the discriminant ability of our proposed embedding method.

Therefore, the main contributions of this paper are highlighted as follows:

(1) In this paper, we propose a novel cross-view based embedding method for multi-label classification, which is robust and discriminative by utilizing the complementarity and consistency between the feature space and the label space;

(2) We develop a novel hypergraph fusion technique to complete the complementarity between the feature space and the label space. The high-order correlation of the instances in the feature space can effectively deal with noise in the label space;

(3) We adopt double-side metric learning to mine the consistency between the feature space and the label space and it can effectively improve the discriminative ability of our embedding method;

(4) We conduct a deep extension for our proposed RCEDS, which makes the method achieve outstanding performance in the application of image annotation.

The rest of this paper is organized as follows. Section 2 gives a review of related work. Then we formulate the problem and present the proposed approach in Section 3. We discuss the experimental results in Section 4 and conclude in Section 5.

II. RELATED WORK

The existing multi-label learning algorithms [1], [2] can be divided into two broad categories: *problem transformation methods* (PTMs) and *algorithm adaptation methods* (AAMs). For both PTMs and AAMs, a common challenging issue exists in the multi-label learning tasks, *i.e.,* the dimension of the output label space will increase exponentially as the number of labels increases. It is not difficult to find that the relevant information between labels may provide additional useful information for multi-label learning, which is beneficial to the performance of multi-label learning system.

Multi-label embedding learning aims to transform the original label space into a latent space by a series of means, which reduces the size of output space and the computational complexity by a large margin. It also can effectively exploit the hidden structure of the original space and make full use of the correlation between labels. A label embedding method based on compressed sensing was proposed by Hse *et al.* [6]. *Firstly*, the label space is projected into a low dimensional space by a compression sensing method. *Secondly*, a regression model is trained for each dimension in the low-dimensional label space. Principal label space transformation (PLST) which is based on classical dimensionality reduction algorithm PCA was proposed by Tai and Lin [7]. PLST preserved the reservation of label space information by minimizing the square loss between original label space and latent space. Conditional principal label space transformation (CPLST) method based on canonical correlation analysis theory was proposed by Chen and Lin [8], which takes both label space embedding loss and regression loss in the latent space into account, and achieves the effect of reducing the dimension of label space by using the feature space information. An end-to-end label space embedding method called Feature-aware implicit label space encoding (FAIE) was proposed by Lin *et al.* [9]. FAIE can directly learn a better hidden space by maximizing the recovery of the latent space and the prediction performance of the latent space. The end-to-end mode breaks the limitation of the latent space. Sparse local embedding for extreme classification (SLEEC) by using local correlations among instances was proposed by Bhatia *et al.* [10]. The tail labels attached only to a small number of instances which makes the label matrix sparse but not low-rank. SLEEC can cover this kind of shortage effectively when solving practical problems. A multi-label embedding method call Canonical-correlated autoencoder (C2AE) was proposed by Yeh *et al.* [3]. C2AE is based on deep learning and canonical correlation analysis, and it can deal with the multi-label classification problem of large-scale data well by using the deep neural network for spatial transformation. Cost-sensitive label embedding with multidimensional scaling (CLEMS) was a cost-sensitive multi-label embedding method proposed by Huang and Lin [4]. In the field of multi-label learning, different evaluation criteria can make totally different comments on the same result. It is the CLEMS that takes the lead to consider evaluation criteria in the process of learning. Co-Embedding (CoE) method was proposed by Sheng *et al.* [5] from a cross-view perspective. Co-Embedding (CoE) learns a common latent space where input and output are jointly embedded and well aligned. It is based on the linear mapping and with an extension named Co-Hashing (CoH) to deal with large-scale data.

We can find that the existing multi-label embedding methods [3]–[5] fail to effectively deal with noisy data, while some of them [3], [9], [10] don't take discriminative ability into consideration. In order to address those two limitations mentioned above simultaneously, we propose our Robust Cross-view Embedding with Discriminant Structure for multi-label classification (RCEDS). This method exploits and utilizes the complementarity and consistency between the feature space and the label space to learn a robust and discriminative embedding model, which is detailed in the following section.

III. PROPOSED APPROACH

For multi-label classification, let $D = \{(x_i, y_i)\}_{i=1}^{N} = \{X, Y\}$ denote a set of d dimensional training instances $X \in R^{d\times N}$ and the associated labels $Y \in \{0, 1\}^{K\times N}$, where N and K are the number of instances and label attributes, respectively. The goal of multi-label classification algorithms is to train a predictor $f : X \to Y$ from D in the training stage, so that the label $\hat{y}$ of a test instance $\hat{x}$ can be predicted accordingly.

A. HYPERGRAPH PRELIMINARIES

Let V denote a finite set of samples, and let E be a family of subsets e of V such that $U_{e\in E} = V$. $G = (V, E)$ is then called a hypergraph with the vertex set V and the hyperedge set E. A hyperedge which contains two vertices is just a simple graph edge. A weighted hypergraph is a hypergraph that has a positive number $\omega(e)$ associated with each hyperedge e, called the weight of hyperedge e. We denote a weighted hypergraph by $G = (V, E, \omega)$. A hyperedge e is said to be incident with a vertex v when $v \in e$. The degree of each vertex $v \in V$ is defined as:

$$d(v) = \sum_{\{e\in E|v\in e\}} \omega(e) \tag{1}$$

Let $|S|$ denote the cardinality of a given arbitrary S. The degree of a hyperedge $e \in E$ is defined as: $\delta(e) =|e|$. A hypergraph G can be represented by a $|V|\times|E|$ matrix H with entries $h(v, e) = 1$, if $v \in e$ and 0 otherwise, called the incidence matrix of G. Based on matrix H, the degree of each vertex and each hyperedge can be calculated as:

$$d(v) = \sum_{e\in E} \omega(e)h(v, e) \tag{2}$$

$$\delta(e) = \sum_{v\in V} h(v, e) \tag{3}$$

Let D_v and D_e denote the diagonal matrices containing the vertex and hyperedge degrees respectively, and let W denote the diagonal matrix containing the weights of hyperedges [11].

B. BASIC MODEL

Our proposed RCEDS learns a novel cross-view based model which is robust and has a strong discriminative ability. The main ideas are detailed in Figure 1. In Figure 1, we use different colors to represent different samples. We use squares to represent the characteristics of the samples, and circles to represent the set of samples. From the feature space, we can find that the four images with black, yellow, red, and green are all about *surfing*, and the blue and orange samples are about *running*. In the label space, we can find that the label set of the yellow and black samples has missing labels and incorrect labels. Our work is mainly focused on two aspects. In the first aspect, we hope to use the high-order relationship of the sample feature space to correct the noise labels in the label space by using the hypergraph structure. The higher-order relationships of the four samples in red and green in the feature space can correct errors and redundant labels. In the second aspect, we hope that the information in the two space can form complementarity, thereby enhancing the discriminativeness of the model. With the help of metric learning, we make the samples that are dissimilar in the two spaces separate more thoroughly.

Then the basic model of our proposed RCEDS is detailed as follows. It is commonly known that in cross-view learning data from the same objects described in different views share a certain common subspace [12], [13] [5]. This is consistent with the purpose of multi-label embedding which hopes to learn a great latent space. The latent space should have a strong correlation with both feature space and label space. To explore and utilize the complementation and consensus between feature space and label space, we develop two regularization terms $R_p(Z, L^p)$ and $R_s(W_p, W_e)$ corresponding to them respectively, which will be detailed in the following two sections.

The basic cross-view embedding can conduct dimension reduction while correlating a feature space and a label space at the same time. We denote the latent space as Z, then the objective function of our proposed RCEDS can be written as follows:

$$\begin{aligned} \min_{W_p,W_e,Z,L^p} \quad & \|W_p^T X - Z\|_F^2 + \|W_e^T Y - Z\|_F^2 \\ & + \lambda_1 R_p(Z, L^p) + \lambda_2 R_s(W_p, W_e) \\ s.t. \; & ZZ^T = I \end{aligned} \tag{4}$$

where $W_p^T X$ and $W_e^T Y$ denote the transformed feature and label data in the latent space respectively. An orthogonal constraint is imposed on Z to make the latent features uncorrelated. $R_p(Z, L^p)$ is the regularization term for complementary principle (robust component), and $R_s(W_p, W_e)$ is the regularization term for consensus principle (discriminant component).

C. COMPLEMENTARY PRINCIPLE (ROBUST COMPONENT)

We develop a novel hypergraph fusion technique to explore and utilize the complementarity between feature space and label space, which has also been effectively verified in our published paper [14]. We construct corresponding hypergraphs in both feature space and label space, then fuse the two hypergraphs from the perspective of random walks. The high-order correlation of the instances in the feature space can effectively correct inaccurate correlation in the label space caused by noisy data (refer to the bottom in Figure 1). The specific practices are detailed as follows:

Firstly, we construct hypergraphs in both feature space and label space. In the label space, a hypergraph is built where each vertex corresponds to one training instance and each hyperedge for one label includes all the training instances relevant to the same label. In the feature space, different from the simple graph where each edge represents the vertex-to-vertex relation, the incidence matrix H of a hypergraph describes the

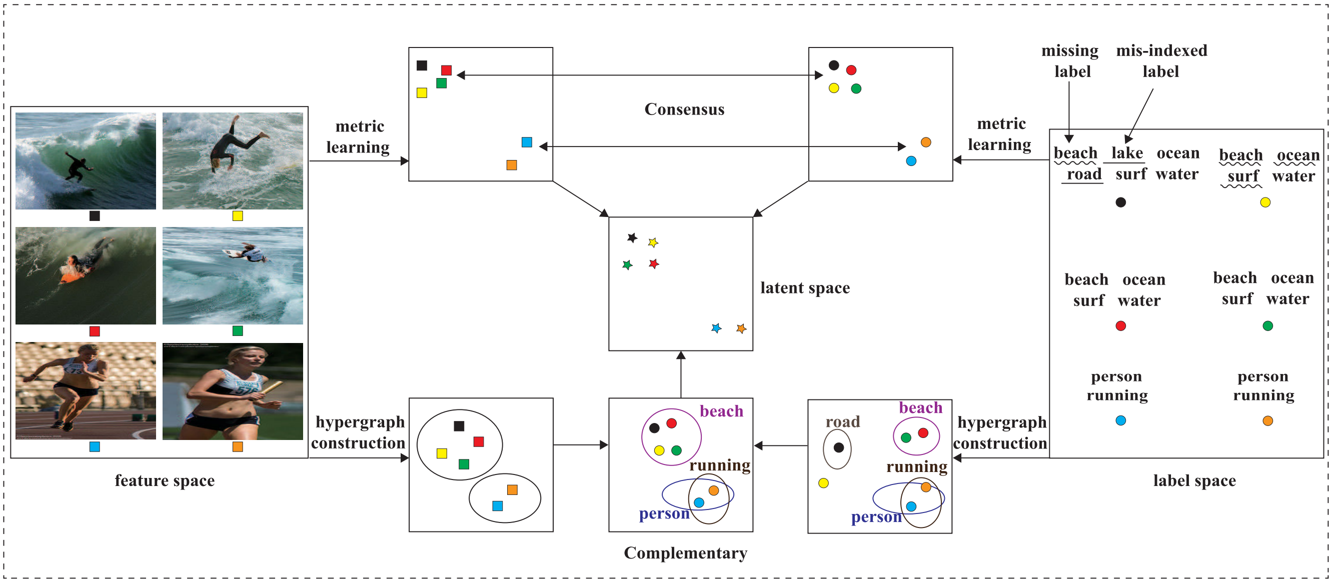


FIGURE 1. Illustration of the proposed RCEDS. The labels with wavy underline are the true labels but they are missing in the ground-truth, and the labels with straight underline are mis-indexed labels in ground-truth annotation. Hypergraph construction can capture the high-order correlation of instances in a feature space; as a result, several mislabled/missing labels can be well tackled and corrected (refer to the bottom in the figure); The double-side metric learning directed by the consistency between a feature space and a label space can help improve the discriminative ability of the proposed RCEDS (refer to the top in the figure).

vertex-to-hyperedge relation. To achieve this, we regard each sample as one vertex and try to generate a hyperedge for each vertex by following the method in [11]. More specifically, we generate the hyperedge e_i by the following formulation:

$$e_i = \{v_j | \theta(x_i, x_j) \leq 0.1\sigma_i\}, \quad i, j = 1, \ldots, n \tag{5}$$

where $\theta(x_i, x_j)$ indicates a similarity measurement between x_i and x_j while σ_i is the average similarity between x_i and each of the other samples.

Secondly, we develop a method to combine the information of the two hypergraphs. Specifically, we associate each hypergraph with a natural random walk [11]. Then the transition probabilities and stationary distribution of the hypergraphs can be fused. We can conduct the laplacian matrix of the fused hypergraph with the fused transition probabilities and stationary distribution.

Let T and Π denote the transition probability and stationary distribution matrix of this hypergraph random walk, respectively. $t(u, v)$ and $\pi(v)$ are each entry of T and Π, which can be denoted as:

$$t(u, v) = \sum_{e \in E} w(e) \frac{h(u, e)}{d(u)} \frac{h(v, e)}{\delta(e)} \tag{6}$$

$$\pi(v) = \frac{d(v)}{vol(V)} \tag{7}$$

where $vol(S)$ of S is the sum of the degrees of the vertices in V, that is, $vol(S) = \sum_{v \in V} d(v)$. Then we can explain the multiple hypergraph cut in terms of a random walk as follows:

$$\beta_1(u) = \frac{\alpha\pi_1(u)}{\alpha\pi_1(u) + (1-\alpha)\pi_2(u)} \tag{8}$$

$$\beta_2(u) = \frac{(1-\alpha)\pi_2(u)}{\alpha\pi_1(u) + (1-\alpha)\pi_2(u)} \tag{9}$$

Here $\beta_i(u)$ is the weight coefficient of the $i-th$ hypergraph. The parameter α is used to specify the relative importance of each hypergraph during fusion. Then we can define fused transition probabilities and stationary distribution as:

$$\begin{aligned} t(u, v) &= \beta_1(u)t_1(u, v) + \beta_2(u)t_2(u, v) \\ \pi(v) &= \alpha\pi_1(v) + (1-\alpha)\pi_2(v) \end{aligned} \tag{10}$$

Now we can clearly find that the above formulation of transition probability and stationary distribution are not simply a linear combination on each hypergraph. Then we can get the laplacian matrix of the fused hypergraph by the following equation:

$$\begin{aligned} L^p &= \Pi - \frac{\Pi P + P^T \Pi}{2} \\ &= \alpha\Pi_1(I - T_1) + (1-\alpha)\Pi_2(I - T_2) \end{aligned} \tag{11}$$

The information of the feature space and the label space can be integrated by the fusion of corresponding hypergraphs. We apply the fused information of feature space and label space which is encoded in the laplacian matrix to directly design the embedding model. Formally, this smoothness assumption can be expressed using the fused hypergraph Laplacian regularizer:

$$R_p(Z, L^p) = tr(ZL^pZ^T) \tag{12}$$

D. CONSENSUS PRINCIPLE (DISCRIMINANT COMPONENT)

To capture the consensus between feature space and label space, we design the consensus regularization term

$R_s(W_p, W_e)$ from the two following parts. The first part is to apply the classical metric learning method NCA [15], [16] to both feature space and label space. In the second part, we project the samples of the feature space and label space to their corresponding low-dimensional space with the transformation matrix W_p and W_e, respectively. Then we constrain the correlation between instances in the feature space and label space to be consistent.

Through the above operations, the similar samples in both original feature space and label space will be closer in the latent space while dissimilar ones are far apart, which effectively enhances the discriminant ability of our embedding method.

Formally, the consensus regularization can be expressed as follows:

$$R_s(W_p, W_e) = \sum_{k=1}^{K}\sum_{i=1}^{N}\sum_{j\in k_i}(p_{ij} + e_{ij}) + \|P - \beta E\|_F^2 \quad (13)$$

where the each entry of P and E are defined as follows:

$$p_{ij} = \frac{exp(-\|W_p^T x_i - W_p^T x_j\|_2^2)}{\sum_{l\neq i} exp(-\|W_p^T x_i - W_p^T x_l\|_2^2)} \quad (14)$$

$$e_{ij} = \frac{exp(-\|W_e^T y_i - W_e^T y_j\|_2^2)}{\sum_{l\neq i} exp(-\|W_e^T y_i - W_e^T y_l\|_2^2)} \quad (15)$$

Following the above consideration, the overall objective function for our multi-label classification model is obtained as follows:

$$\begin{aligned} \min_{W_p, W_e, Z, L^p} \quad & \|W_p^T X - Z\|_F^2 + \|W_e^T Y - Z\|_F^2 \\ & + \lambda_1 tr(ZL^pZ) \\ & + \lambda_2(\sum_{k=1}^{n}\sum_{i=1}^{N}\sum_{j\in k_i}(p_{ij} + e_{ij}) \\ & + \|P - \beta E\|_F^2) \\ s.t.\ & Z^T Z = I \end{aligned} \quad (16)$$

where L^p is the normalized Laplacian matrix of the fused hypergraph.

E. OPTIMIZATION

Note that directly minimizing the objective function in Eq.(16) is intractable. The orthogonal constraint is nonconvex, which makes the problem more challenging to be solved. Accordingly, we provide an iterative algorithm with Cayley transformation [17] to update these variables to reach a local minimum.

1) UPDATE L^p

The hyperedges generated from the original training data may result in an inaccurate hypergraph. To deal with this, we design to learn the hyperedges from low-dimensional training data, whose redundant and irrelevant label space information have been removed as much as possible [18]. Next we will describe in detail how to update L^p:

More specifically, we generate the hyperedge e_i by the following formulation [11]:

$$e_i = \{v_j | \theta(W_e^T y_i, W_e^T y_j) \le 0.1\tilde{\sigma_i}\}, \quad i, j = 1, \dots, n \quad (17)$$

where σ_i is the average similarity between $W_e^T y_i$ and each of the other low-dimensional samples.

For L^p, we use the low-dimension training data to conduct the hypergraph on label space. Then the hypergraph based on labels H^l will be updated correspondingly. We conduct the transition probability and stationary distribution on hypergraph H^l, then the confused laplacian matrix L^p will be updated correspondingly.

2) UPDATE W_p, W_e

With W_e, Z and L^p fixed, the problem in Eq.(16) reduces to:

$$\begin{aligned} \min_{W_p} \ \|W_p^T X - Z\|_F^2 + \lambda_2(&\sum_{k=1}^{K}\sum_{i=1}^{N}\sum_{j\in k_i} p_{ij} \\ &+ \sum_{i=1}^{N}\sum_{j=1}^{N}(p_{ij} - \beta e_{ij})^2) \end{aligned} \quad (18)$$

We define G_{W_p} as the gradient with respect to W_p. It can be computed as follows (denote $x_{ij} = x_i - x_j$):

$$\begin{aligned} G_{W_p} = & \ 2W_p^T XX^T - 2ZX^T \\ & -2\lambda_2 W_p^T(\sum_{k=1}^{K}\sum_{i=1}^{N}\sum_{j\in k_i} m_{ij} \\ & -\sum_{i=1}^{N}\sum_{j=1}^{N} m_{ij}(p_{ij} - \beta e_{ij})) \end{aligned} \quad (19)$$

where:

$$m_{ij} = p_{ij}(x_{ij}x_{ij}^T - \sum_{l\neq i} x_{il}x_{il}^T p_{il}) \quad (20)$$

The gradient corresponding to W_e can be similarly obtained as follows:

$$\begin{aligned} G_{W_e} = & \ 2W_e^T YY^T - 2ZY^T \\ & -2\lambda_2 W_e^T(\sum_{k=1}^{K}\sum_{i=1}^{N}\sum_{j\in k_i} n_{ij} \\ & -\sum_{i=1}^{N}\sum_{j=1}^{N} n_{ij}(p_{ij} - \beta e_{ij})) \end{aligned} \quad (21)$$

where:

$$n_{ij} = e_{ij}(y_{ij}y_{ij}^T - \sum_{l\neq i} y_{il}y_{il}^T e_{il}) \quad (22)$$

We use a stochastic batch update per element and a non-linear conjugate gradient update.

3) UPDATE Z

With W_p, W_e and L^p fixed, the problem in Eq.(16) reduces to:

$$\begin{aligned} \min_{Z} \ & \|W_p^T X - Z\|_F^2 + \|W_e^T Y - Z\|_F^2 \\ & + \lambda_1 (Z L^p Z^T) \\ s.t. \ & Z Z^T = I \end{aligned} \tag{23}$$

Motivated by COE [5], we adopt the Optimization with Orthogonality Constraints [19] to get a local optimal solution of Z. Specifically, the details of this step can be adapted from [5], [19].

With the above derivations, we can learn RCEDS by gradient optimization, and the pseudo code of training is summarized in Algorithm 1.

Once the learning of RCEDS is complete, label prediction of a test input can be easily achieved by the nearest neighbor algorithm and the pseudo code of prediction is summarized in Algorithm 2.

Algorithm 1 Training Process of RCEDS

```
**Input:** Feature matrix $X$, label matrix $Y$, parameter $\lambda 1$, $\lambda 2$
and dimension $M$ of the latent space
**Output:** $W_p$, $W_e$
Randomly initialize $W_p$, $W_e$, $Z$
**repeat**
    **repeat**
        Construct hypergraphs based on $X$ and $Y$, obtain $L^p$
        Update $Z$ by Curvilinear Search Algorithm based on
        Cayley Transformation
        Perform gradient descent on $W_p$ by Eq.(19)
        Perform gradient descent on $W_e$ by Eq.(21)
        Update $L^p$ with the new label hypergraph $H^l$ generated by Eq.(17)
    **until** Converge
**until** Converge
```

F. COMPUTATIONAL COMPLEXITY ANALYSIS

The computational cost of the proposed RCEDS is analyzed in this section. We apply mini-batch gradient descent to each loss term for updating the parameters. We denote r and k as the number of samples in each random block and the number of iterations. For simplicity, we assume that $N \gg d > r$ and $K \geqslant M$ holds in the real-world applications.

The computation of training RCEDS includes three main parts, updating Z, updating W_p and W_e, updating L^p.

(1) In the process of updating Z, calculating L^p requires $\mathcal{O}(kr^2K)$, then the computational cost of G_Z requires $\mathcal{O}(kr^3)$.

(2) In the process of updating W_p and W_e, calculating G_{W_p} and G_{W_p} requires $\mathcal{O}(krMK)$ and $\mathcal{O}(Mdkr)$. Then the computational cost of G requires $\mathcal{O}(krMd)$.

(3) In the process of updating L^p, the most time-consuming part is calculating transition probability T corresponding to the hypergraph of the feature space and it requires $\mathcal{O}(krd^2)$.

Thus, the overall training cost of RCEDS is $\mathcal{O}(krd^2)$. The computational complexity is acceptable and our proposed RCEDS is able to handle multi-label applications with many labels.

Algorithm 2 Predicting Process of RCEDS

```
**Input:** Label matrix $Y$, $W_p$, $W_e$, testing example $\hat{x}$
**Output:** Prediction $\hat{y}$
Computing the embedding of $Y$ as $Z = W_e^T(Y)$
Obtain the predicted vector $\hat{z} = W_p^T(\hat{x})$
Find $z_q \in Z$ such that $d(z_q, \hat{z})$ is the smallest
Attach the $y_q$ to $\hat{x}$ as its label vector
```

G. DEEP EXTENSION FOR RCEDS

With the development of deep learning, the deep features of the image are significantly better than other features. We hope to effectively use the deep features of the image to better improve the performance of the model. Benefiting from the strong representation ability of deep neural networks, we can extend the learning process of RCEDS to a nonlinear scenario [20].

Assume there is a neural network $g : R^d \rightarrow R^{d_m}$, embedding a d-dimensional object to a d_m-dimensional middle space. Based on the transformed representation, linear weights $W_p \in R^{d_m \times M}$ is constructed to project the embedding to the latent space. Specifically, for a pair of instances (x_i, x_j), we compute their distance in the latent space as $Dis^2(x_i, x_j) = \|W_p g(x_i) - W_p g(x_j)\|_F^2$. The deep feature describes the images more accurately and can more effectively characterize the distance between instances. In the implementation, we use VGG network [21] pretrained on ImageNet 2012 classification challenge dataset [22] to instantiate the function $g(\cdot)$.

IV. EXPERIMENTS

A. DATASETS AND SETTINGS

To validate the proposed Robust Cross-view Embedding with Discriminant Structure for multi-label classification (RCEDS), we download eight benchmark datasets in different domains with a relatively large number of labels for experiments, *i.e.*, cal500, delicious, EUR-Lex (subject matters), mediamill, Corel5k, iaprtc12, ESPGame and NUSWIDE from Mulan [23]. The statistics of the eight real world datasets are summarized in Table 1. For the datasets of text and video (cal500, delicious, EUR-Lex (subject matters), mediamill), we use traditional features from Mulan [23]. For the datasets of image (Corel5k, iaprtc12, ESPGame, NUSWIDE), we extract 4096-dimensional deep features by using the 16 layers VGG network [21] pretrained on ImageNet 2012 classification challenge dataset [22] with MatConvNet. We didn't perform any fine-tuning for the sake of fairness and computational efficiency.

In the multi-label learning problem, since each sample may have multiple category labels at the same time, the single-label evaluation metrics cannot be directly used for the performance evaluation of the multi-label learning system. Therefore, researchers have successively proposed a series of multi-label evaluation metrics. Here we consider four evaluation metrics, *i.e.*, Macro-F1, Micro-F1, One-error and Ranking loss, which are widely used in multi-label learning to evaluate the prediction performance. Based on the symbolic representation in the problem definition, we denote y_i as the set of related labels belonging to the sample x_i, then in order to characterize the binary classification performance of the predictors on each label, four basic quantities related to the test sample are usually used: TP_j (true positive), FP_j (false positive), TN_j (true negative) and FN_j (false negative).

$$\text{macro-F1} = \frac{1}{n}\sum_{j=1}^{n}\frac{2TP_j}{2TP_j + FN_j + FP_j} \tag{24}$$

$$\text{micro-F1} = \frac{2\sum_{j=1}^{n} TP_j}{2\sum_{j=1}^{n} TP_j + \sum_{j=1}^{n} FN_j + \sum_{j=1}^{n} FP_j} \tag{25}$$

$$\text{one-error} = \frac{1}{m}\sum_{i=1}^{m}\{[argmax_{y\in Y} f(x_i, y)] \notin y_i\} \tag{26}$$

$$\begin{aligned}\text{ranking loss} = \frac{1}{m}\sum_{i=1}^{m}\frac{1}{|y_i||\bar{y}_i|}|(y', y'')|\\ \times f(x_i, y') \leqslant f(x_i, y''), (y', y'') \in y_i \times \bar{y}_i|\end{aligned} \tag{27}$$

Here, $\bar{y}_i$ is the complementary set of y_i in Y. For Macro-F1 and Micro-F1, larger the values, better the performance. For One-error and Ranking loss, smaller values indicate better performance. The definitions of the four metrics can be found in [1].

TABLE 1. Datasets properties.

Dataset	Domain	Instances	Feature	Labels	Cardinality
cal500	music	502	68	174	26.0
delicious	text	16105	500	983	19.0
EUR-Lex(sm)	text	19348	5000	201	2.2
mediamill	video	43907	120	101	4.4
Corel5k	image	5000	4096	374	3.5
iaprtc12	image	19627	4096	291	5.7
ESPGame	image	23641	4096	268	4.7
NUSWIDE	image	269648	4096	81	1.9

In our experiments, we compare our proposed approach to the following state-of-art multi-label classification methods: Feature-aware Implicit label space Encoding (FaIE) [9], Sparse Local Embeddings for Extreme Classification (SLEEC) [10], Canonical-Correlated Autoencoder (C2AE) [3], Cost-sensitive Label Embedding with Multidimensional Scaling (CLEMS) [4] and Co-Embedding (CoE) [5]. We also report the results of some baseline algorithms, such as Binary Relevance (BR) [24], Classifier Chain (CC) [25] and Deep Canonical Correlation Analysis (DCCA) [26].

Our experiment consists of two main parts. The first part is Prediction Performance on Datasets of Music, Text and Video with Traditional Feature, which mainly proposed to verify that our method has great classification performance and anti-noise property. The second part is Prediction Performance on Datasets of Image with Deep Extension. which proposed to verify the deep expansion of the proposed RCEDS and verify that our method can also achieve good results in image multi-label classification. In addition, we also give Prediction Performance under Varying Degrees of Noise and Efficiency Analysis.

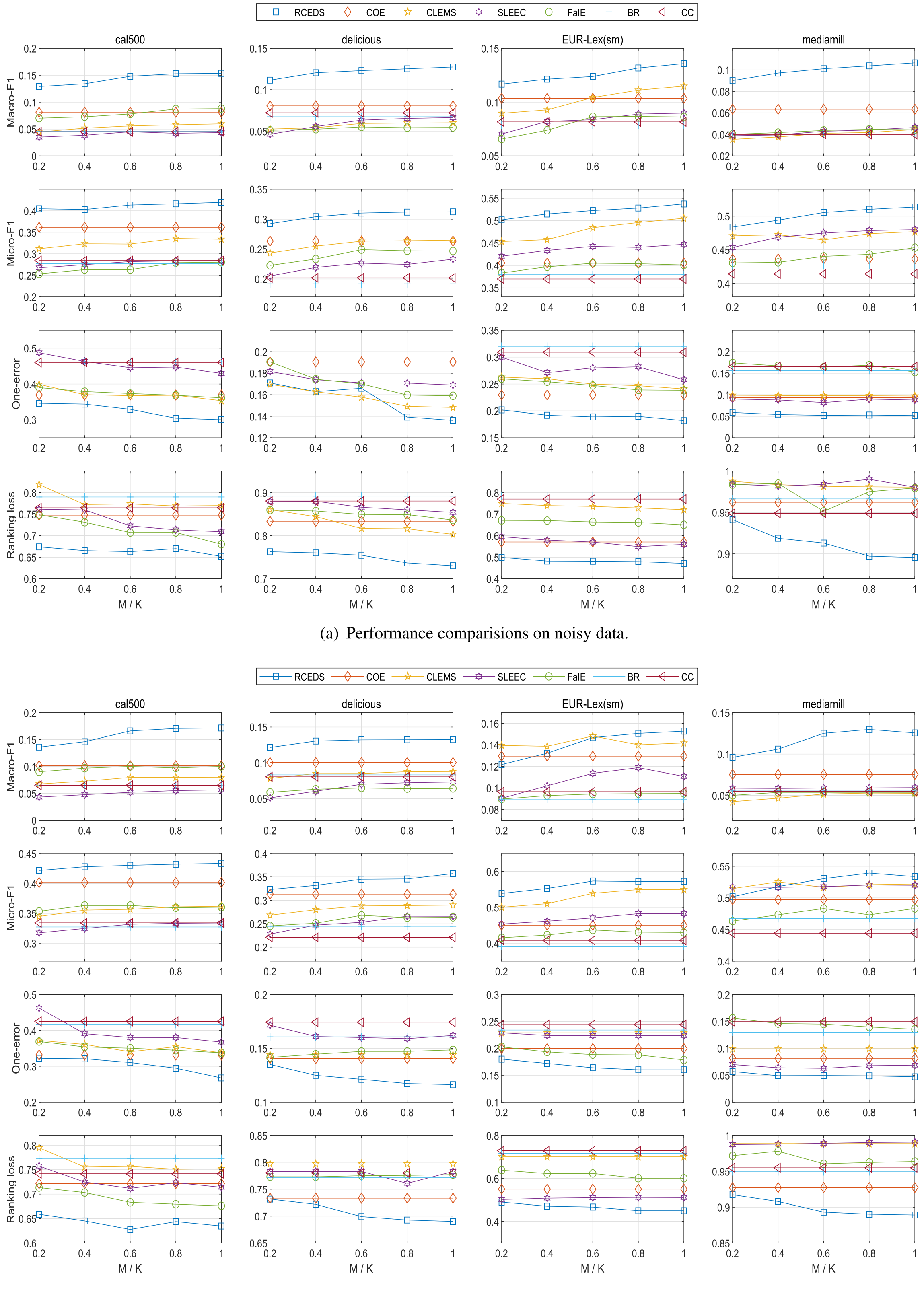
To select the parameters for these methods, we randomly hold one-fifth of training data in every dataset for validation. We choose linear classification/regression package LIBLINEAR with l_2-regularized logistic regression as the classifier for BR. Specifically, α in FaIE is selected from $[10^{-1}, 10^0, \cdots, 10^4]$, and we use linear ridge regression to learn predictive models from instance to code vectors. Following the experimental setting, we set the number of clusters as $\lfloor n/6000 \rfloor$ and the number of learners as 15 for SLEEC. For the method CLEMS, we set the critical as F1-score because our two metrics are both connected with F1-score. For the architecture of DCCA, DCVE, we follow the [3] to set F_p composed of 2 layers of fully connected layer structures while the embedding function F_e is a single fully connected function. For each fully connected layer, a total of 512 neurons are deployed. A leaky ReLU activation function is considered, while the batch size is fixed at 200. We select λ_1 and λ_2 from $\{10^{-10}, 10^{-9}, \cdots, 10^{10}\}$. We combine the analysis of specific data sets and local grid search to select these two parameters. All the experiments are performed on a 64-Bit Linux workstation with an Intel E5-2650 CPU, an NVIDIA Titan X Pascal card and 256GB memory.

B. PREDICTION PERFORMANCE ON DATASETS OF MUSIC, TEXT AND VIDEO WITH TRADITIONAL FEATURE

We perform RCEDS, CLEMS, SLEEC, FAIE, BR and CC on the datasets of text and video with different values of M/K (from 20% to 100%) where M and K are respectively the dimensions of the latent space and the original label space. Figure 2(a) illustrates and compares the performances of the above methods on noisy datasets. To generate the noisy datasets, we randomly choose 10% elements from each sample and replace them with random values [27]. Figure 2(b) illustrates and compares the performances of the above methods on clean datasets. From Figure 2, we can see that our RCEDS method performed favorably against most label embedding methods in most cases on both noisy and clean datasets, which well demonstrates its effectiveness. From the experimental results, we can draw the following interesting observations:

(1) The proposed RCEDS significantly outperforms most of the baselines on the four datasets. For example, on cal500 in Figure 2(a), our method improves the best results of the baselines by 6.2% (Macro-F1), 7.8% (Micro-F1), which validates our theoretical results.

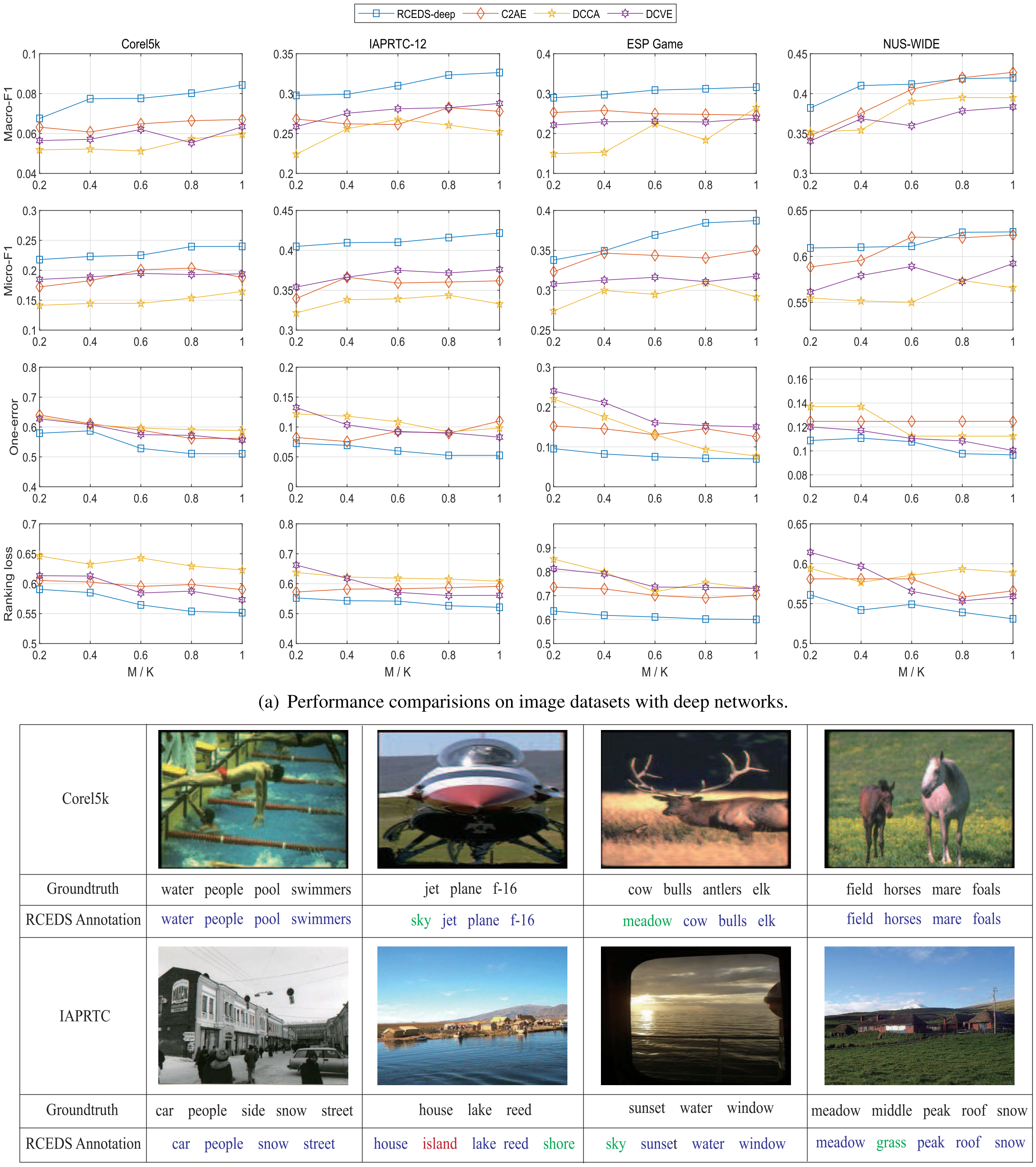
(2) From Figure 2(b), we can see that the proposed method is significantly better than the current multi-label



(a) Performance comparisions on noisy data.

(b) Performance comparisions on clean data.

FIGURE 2. Performance comparisons in terms of Macro-F1, Micro-F1, One-error and Ranking loss with different latent space dimension ratios (M/K) on datasets of music, text and video with traditional feature.



(a) Performance comparisions on image datasets with deep networks.

Corel5k				
Groundtruth	water people pool swimmers	jet plane f-16	cow bulls antlers elk	field horses mare foals
RCEDS Annotation	water people pool swimmers	sky jet plane f-16	meadow cow bulls elk	field horses mare foals
IAPRTC				
Groundtruth	car people side snow street	house lake reed	sunset water window	meadow middle peak roof snow
RCEDS Annotation	car people snow street	house island lake reed shore	sky sunset water window	meadow grass peak roof snow

(b) Several image annotation examples on IAPRTC datasets. For each image, the labels in black are those that match with ground-truth annotation. The blue labels denote the correctly predicted ones, while the red labels are those that are wrongly predicted. Besides, we use green labels to represent the labels that are correctly predicted but missing in the ground-truth annotations.

FIGURE 3. Performance comparisons in terms of Macro-F1, Micro-F1, One-error and Ranking loss with different latent space dimension ratios (M/K) on datasets of image with deep extension.

embedding methods when dealing with noisy data, which shows that our proposed RCEDS has strong anti-noise ability.

(3) When the ratio of M/K is above 40%, the result is basically stable. We can use a lower-dimensional space to preserve the information of the original space effectively by extracting the information of all labels. It reflects that our method can effectively mine the hidden structure of the original label space.

(4) The CLEMS method does not take the feature space into consideration when conducting label embedding, which makes the latent space learned by it fail to have enough predictability. The COE method has relatively weak discriminability and limited ability to be against noise. Compared with them, with a capability of learning a more robust and discriminative latent space which has stronger predictability, our RCEDS can solve the problems mentioned above very well.

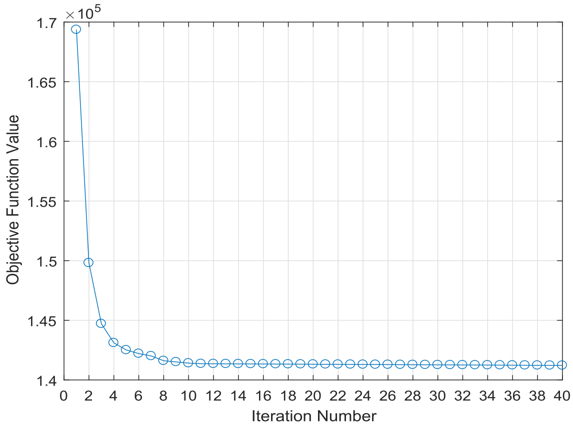


FIGURE 4. Convergence curce of the proposed RCEDS on Corel5k.

We can clearly see that our method performs well on the datasets in different fields because our decoding function is robust and matches well with the encoding function.

C. PREDICTION PERFORMANCE ON DATASETS OF IMAGE WITH DEEP EXTENSION

We perform RCEDS-deep, C2AE, DCVE and DCCA on the datasets of image with different values of M/K (from 20% to 100%) and the results are illustrated in Figure 3(a). Besides, we present a case study in which the proposed method is applied to a multi-label image annotation application. RCEDS is applied on the famous IAPRTC-12 dataset and the annotation results of several randomly selected images are illustrated in Figure 3(b). From the Figure 3, we can draw the following interesting observations:

(1) The proposed RCEDS with deep extension, which can effectively explore and utilize consistency and complementarity between a feature space and a label space, is obviously superior to other methods based on deep learning from many aspects, one of which, for example, is the outperformance in preprocessing image datasets.

(2) The proposed RCEDS correctly predicts most labels for these images and our method can even find the labels missing in the ground truth annotations. For example, our method tags the image in Row 2, Column 4 with the label 'grass' missed in the ground truth. The performance on multi-label image annotation applications suggests that our methods can work well in the image annotation applications.

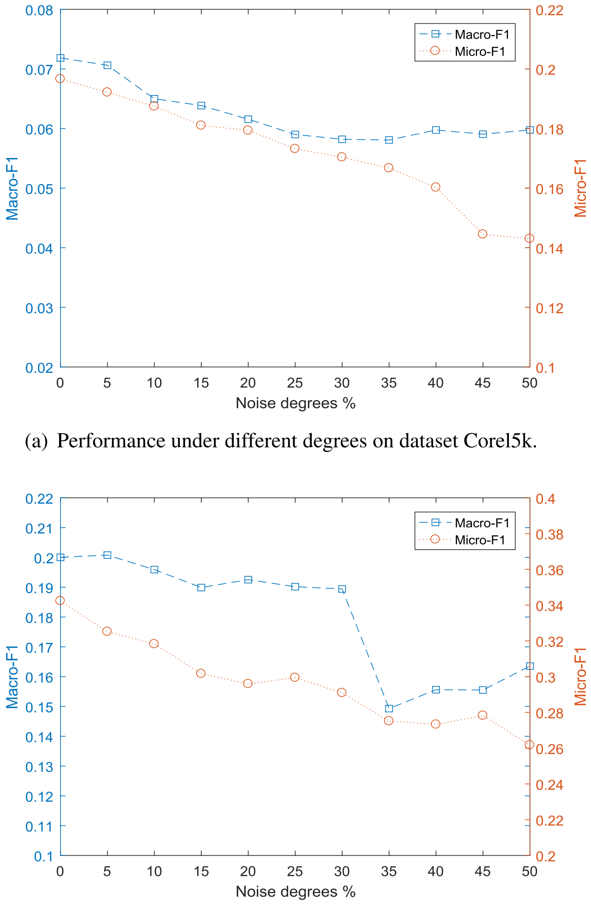
To further verify the effectiveness of our derived deep latent space, we consider several example labels from ESPGame and list their corresponding neighboring ones in Table 2. From this table, we see that the neighboring labels observed in the latent space exhibit highly correlated semantic information. This confirms our RCEDS is sufficiently exploiting label dependency during the learning process.

D. CONVERGENCE ANALYSIS

In this section, we empirically study the convergence of the proposed RCEDS. The convergence curves of RCEDS on Corel5k dataset with $M/K = 0.8$ are plotted in Figure 4. As can be seen in the figure, the objective converges quickly in a few iterations. We omit the results generated on other datasets since they are similar with the observation in Figure 4.

TABLE 2. Visualization of embedded labels for ESPGame.

Label	Nearest neighbors
sea	ocean, cloud, lake, boat, fish
swim	beach, legs, feet, ship, wave
animal	dog, ear, nose, horse, tail
face	eye, smile, hat, hand, picture
army	soldier, sword, war, coat, triangle, finger
airplane	plane, wing, model, war



(a) Performance under different degrees on dataset Corel5k.

(b) Performance under different degrees on dataset cal500.

FIGURE 5. Prediction Performance under varying degrees of noise.

E. PREDICTION PERFORMANCE UNDER VARYING DEGREES OF NOISE

In order to better reflect the efficient performance of our method being against noise, we conduct experiments with varying degrees of noise interference. To generate the noisy datasets, we randomly choose different proportion (from 0% to 50%) elements from each sample and replace them with random values [27]. Here we have chosen the cal500 and Corel5k datasets to conduct experiments with

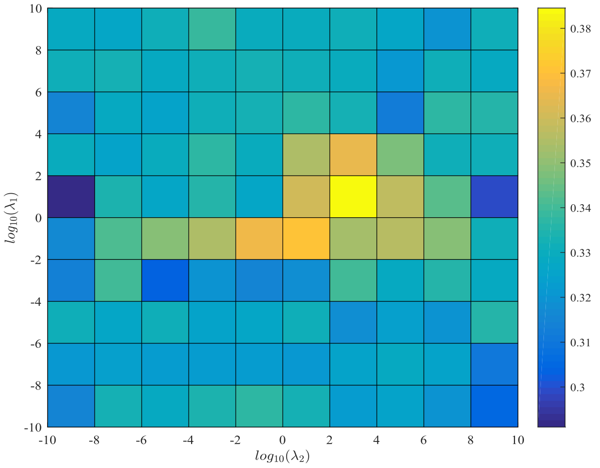


FIGURE 6. Effects of λ_1, λ_2 in RCEDS on the performance of multi-label classification on cal500.

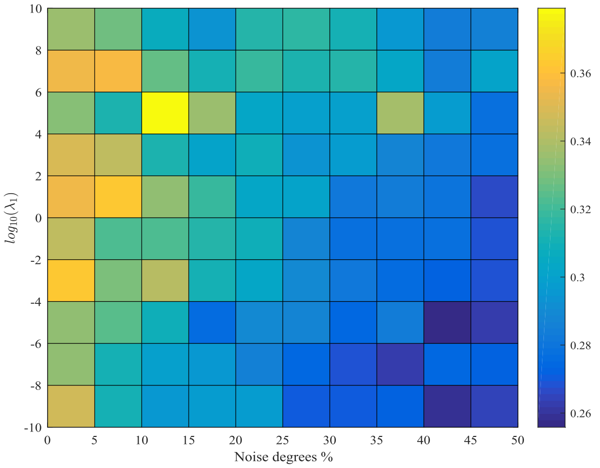


FIGURE 7. Effects of λ_1 on varying degrees of noise in RCEDS with a fixed λ_2 on the performance of multi-label classification on cal500.

different levels of noise interference and show the result in Figure 5.

F. PARAMETER ANALYSIS

Furthermore, we conduct experiments to see the effects of two parameters λ_1, λ_2 in the proposed RCEDS. Figure 6 gives an illustration of the variances of multi-label classification performance (Micro-F1) as λ_1, λ_2 vary in $\{10^{-10}, 10^{-8}, \cdots, 10^{10}\}$ in a run on cal500 with $M/K = 0.2$.

We conduct experiments to see the effects of complementary principle regularizer in dealing with noise. Figure 7 gives an illustration of the variances of multi-label classification performance (Micro-F1) with a fixed λ_2. These two regularizers enhance the learned latent space from two different perspectives: discriminability and robustness. Thus, with the absence of any one of them, the performance will drop. When analyzing the individual effect of these two regularizers, specially, from the experimental results in Figure 6 and Figure 7, we can draw the following interesting observations: when the noise is relatively weak, the complementarity regularizer has a weak influence, and the consensus regularizer has a better performance in alleviating the negative effect brought by it; when the noise is relatively strong, the complementarity regularizer will play a dominant role in dealing with noise through the high-order relationship. Nevertheless, whatever the scene is, the consensus regularizer can dramatically improve the discriminability of the model by the metric learning embedded in it.

V. CONCLUSION

In this paper, we propose Robust Cross-view Embedding with Discriminant Structure for multi-label classification (RCEDS) to solve the problem of multi-label classification. RCEDS deals with noisy datasets by using the hypergraph fusion technique which explores and utilizes the complementary between a feature space and a label space. The accurate information in the feature space can correct the label space noise to improve the robustness of our model effectively. Meanwhile, the double-side metric learning is conducted to explore the consistency between the feature space and the label space, utilizing consensus effectively to improve the discriminative ability of our embedding method. Furthermore, we conduct a deep extension for our proposed RCEDS, which proves that in the method have an outstanding performance in the application of image annotation. The experiment results demonstrate that RCEDS is superior to state-of-the-art label embedding algorithms in many cases.

There are many interesting future works. For example, our proposed RCEDS may be extended to deal with multi-modal datasets. Dealing with multi-modal multi-label data in the process of learning semantic latent space is an interesting issue to be studied in the future.

REFERENCES

[1] 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.

[2] Z.-H. Zhou and M.-L. Zhang, "Multi-label learning," in *Encyclopedia of Machine Learning and Data Mining*, C. Sammut and G. I. Webb, Eds. Berlin, Germany: Springer, 2017, pp. 875–881.

[3] C. Yeh, W. Wu, W. Ko, and Y. F. Wang, "Learning deep latent spaces for multi-label classification," in *Proc. AAAI Conf. Artif. Intell.*, 2017, pp. 2838–2844.

[4] 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.

[5] X. Shen, W. Liu, I. W. Tsang, Q.-S. Sun, and Y.-S. Ong, "Multilabel prediction via cross-view search," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 9, pp. 4324–4338, Sep. 2018.

[6] D. Hsu, S. M. Kakade, J. Langford, and T. Zhang, "Multi-label prediction via compressed sensing," in *Proc. Annu. Conf. Neural Inf. Process. Syst.*, 2009, pp. 772–780.

[7] F. Tai and H.-T. Lin, "Multilabel classification with principal label space transformation," *Neural Comput.*, vol. 24, no. 9, pp. 2508–2542, 2012.

[8] Y. Chen and H. Lin, "Feature-aware label space dimension reduction for multi-label classification," in *Proc. Annu. Conf. Neural Inf. Process. Syst.*, 2012, pp. 1529–1537.

[9] Z. Lin, G. Ding, M. Hu, and J. Wang, "Multi-label classification via feature-aware implicit label space encoding," in *Proc. Int. Conf. Mach. Learn.*, 2014, pp. 325–333.

[10] K. Bhatia, H. Jain, P. Kar, M. Varma, and P. Jain, "Sparse local embeddings for extreme multi-label classification," in *Proc. Annu. Conf. Neural Inf. Process. Syst.*, 2015, pp. 730–738.

[11] J. H. Deng-yong Zhou and B. Schölkopf, "Learning with hypergraphs: Clustering, classification, and embedding," in *Proc. Annu. Conf. Neural Inf. Process. Syst.*, 2006, pp. 1601–1608.

[12] X. Shen, F. Shen, Q.-S. Sun, Y. Yang, Y.-H. Yuan, and H. T. Shen, "Semi-paired discrete hashing: Learning latent hash codes for Semi-paired cross-view retrieval," *IEEE Trans. Cybern.*, vol. 47, no. 12, pp. 4275–4288, Dec. 2017.

[13] Y. Guo and M. Xiao, "Cross language text classification via subspace co-regularized multi-view learning," in *Proc. Int. Conf. Mach. Learn.*, 2012, pp. 1615–1622.

[14] K. Wang, "Robust embedding framework with dynamic hypergraph fusion for multi-label classification," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2019, pp. 982–987.

[15] J. Goldberger, S. Roweis, G. Hinton, and R. Salakhutdinov, "Neighbourhood components analysis," in *Proc. Annu. Conf. Neural Inf. Process. Syst.*, 2005, pp. 513–520.

[16] R. Jin, S. Wang, and Z.-H. Zhou, "Learning a distance metric from multi-instance multi-label data," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 896–902.

[17] S. Friedland, J. Nocedal, and M. L. Overton, "The formulation and analysis of numerical methods for inverse eigenvalue problems," *SIAM J. Numer. Anal.*, vol. 24, no. 3, pp. 634–667, Jun. 1987.

[18] X. Zhu, Y. Zhu, S. Zhang, R. Hu, and W. He, "Adaptive hypergraph learning for unsupervised feature selection," in *Proc. 26th Int. Joint Conf. Artif. Intell.*, Aug. 2017, pp. 3581–3587.

[19] Z. Wen and W. Yin, "A feasible method for optimization with orthogonality constraints," *Math. Program.*, vol. 142, no. 1, pp. 397–434, 2013.

[20] H.-J. Ye, D.-C. Zhan, Y. Jiang, and Z.-H. Zhou, "What makes objects similar: A unified multi-metric learning approach," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 5, pp. 1257–1270, May 2019.

[21] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Represent.*, 2015, pp. 1–14.

[22] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.

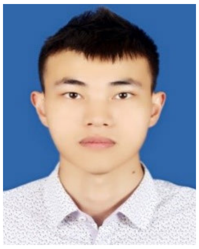
[23] G. Tsoumakas, E. Spyromitros-Xioufis, J. Vilcek, and I. Vlahavas, "Mulan: A Java library for multi-label learning," *J. Mach. Learn. Res.*, vol. 12, pp. 2411–2414, Jun. 2011.

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

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

[26] G. Andrew, R. Arora, J. Bilmes, and K. Livescu, "Deep canonical correlation analysis," in *Proc. Int. Conf. Mach. Learn.*, 2013, pp. 1247–1255.

[27] S. Li and Y. Fu, "Robust multi-label semi-supervised classification," in *Proc. IEEE Int. Conf. Big Data (Big Data)*, Dec. 2017, pp. 27–36.



KAIXIANG WANG received the B.S. degree from the School of Mathematical Sciences, Nanjing Normal University, Nanjing, China, in 2016, and the M.S. degree from the School of Computer Science and Technology, Nanjing Normal University, in 2019. He is currently pursuing the Ph.D. degree with the School of Mathematical Sciences, Nanjing Normal University. His research interests include machine learning, computer vision and data mining.