Deep Correlation Structure Preserved Label Space Embedding for Multi-label Classification

Kaixiang Wang nnuwangkaixiang@outlook.com
Ming Yang myang@njnu.edu.cn
Wanqi Yang yangwq@njnu.edu.cn
School of computer science and technology, Nanjing Normal University, Nanjing 210046, China
YiLong Yin ylyin@sdu.edu.cn
School of computer science and technology, Shandong University, Jinan 250000, China

Editors: Jun Zhu and Ichiro Takeuchi

Abstract

Label embedding is an effective and efficient method which can jointly extract the information of all labels for better performance of multi-label classification. However, most existing embedding methods ignore information of feature space or intrinsic structure of previous label space, such that their learned latent space will not have strong predictability and discriminant ability. We propose a novel deep neural network (DNN) based model, namely Deep Correlation Structure Preserved Label Space Embedding (DCSPE). Specifically, DCSPE derives a deep latent space by performing feature-aware label space embedding with deep canonical correlation analysis (DCCA) and preserving the intrinsic structure of the previous label space with proposed deep multidimensional scaling (DMDS). Our DCSPE is achieved by integrating the DNN architectures of the two DNN based models and can learn a feature-aware structure preserved deep latent space. Furthermore, extensive experimental results on datasets with many labels demonstrate that our proposed approach is significantly better than the existing label embedding algorithms.

Keywords: multi-label classification, label embedding, deep multidimensional scaling, deep canonical-correlation analysis

1. Introduction

Multi-label learning is one of the hot topics in the field of machine learning and pattern recognition. In the multi-label learning framework, each sample is represented by a feature vector, meanwhile it may belong to multiple category labels. The goal is to induce a function which is able to assign multiple proper labels (from a given label set) to unseen instances [Zhou and Zhang (2017), Zhang and Zhou (2014)]. With the introduction of the concept of multi-label learning, many scholars have carried out research on this basis. At present, in terms of multi-label learning, there are two main directions for research: one is to convert the multi-label problem into multiple single-label problems firstly, and then to solve each single-label problem, which is called *problem transformation*. Another method is to solve the multi-label problem by proposing or improving an existing single-label problem algorithm to adapt to the multi-label problem itself, Such a solution is called *algorithm adaptation*.

However, with the increase of the number of labels, these standard multi-label classification methods that work in the original label space can easily become computationally impractical in training multi-label classifiers. For example, when the number of labels is large, Binary Relevance (BR) will produce many kinds of sub-problems in the second category, which requires a lot of training and testing time, and thus it is difficult to build an efficient classification system Bi and Kwok (2013). However, 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 and reduce the training and forecasting time of the whole model.

As an important family of multi-label classification algorithm, a number of label space reduction methods (LSDR) have been developed in the literature to address multi-label classification with many labels. LSDR algorithms [Tai and Lin (2012), Lin et al. (2014), Zhou et al. (2017), Yeh et al. (2017), Huang and Lin (2017), Shen et al. (2018)] consider a low dimensional embedded label space for digesting the information between labels and conducting more effective learning. Most label space reduction methods(LSDR) rely on the assumption that the correlation of the labels can be captured via dimension reduction, making the learning part more effective (physically more meaningful) and efficient Huang and Lin (2017). However, some existing approaches [Tai and Lin (2012), Bhatia et al. (2015), Huang and Lin (2017)] perform label space embedding without considering the feature information, simultaneously few approaches can effectively maintain the intrinsic structure of previous label space, such that their learned latent space will not have strong predictability and discriminant ability. Such latent space will increase the error of space transformation and the training error of the classifiers in the latent space. The increase of these two errors will seriously affect the effectiveness of the multi-label classification.

We proposed a novel DNN based framework, Deep Correlation Structure Preserved Label Space Embedding (DCSPE) to deal with the above challenges. Our proposed DCSPE is motivated by C2AE Yeh et al. (2017) initially, both C2AE and our proposed DCSPE design the network to measure the correlations between feature and label space, as well as label correlations. Different from C2AE, our method takes the preservation of the intrinsic structural information of label space into consideration. The structural information enhances the latent space with discriminability which plays an important role in classification. C2AE performed label-correlation aware prediction by the introduced loss functions for the decoding outputs. However, our propose embedding method DCSPE has already exploited the global label correlations implicitly by learning the latent space Zhu et al. (2018).

Our proposed DCSPE utilizes the proposed deep multidimensional scaling (DMDS) to preserve the intrinsic structure of the label space while enables the embedding feature-aware with the deep canonical correlation analysis (DCCA). DMDS based on classical multidimensional scaling (MDS) maintains the similarity between samples in the label space. Similar samples in previous label space will still be very close in the latent space, which effectively enhances the discriminant ability of the latent space. DCCA maximizes the nonlinear correlations in the latent space between the feature vectors of the samples and their corresponding label vectors. The label space transformation directed by feature information will make the learned latent space more predictable. The main contributions of this paper are highlighted as follows: (1) DCSPE preserves the intrinsic structure of the previous label space during the embedding process which effectively reduces the loss of discriminatory information. (2) DCSPE is able to perform feature-aware label embedding by exploiting the nonlinear relationship between feature space and label space. (3) By utilizing and integrating the architectures of deep canonical correlation analysis and deep multidimensional scaling, DCSPE can mine a deep latent space with the architecture of DNNs.

The rest of this paper is organized as follows. Section 2 gives a brief 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.

2. Related Work

The proposed multi-label learning algorithms [Zhang and Zhou (2014), Zhou and Zhang (2017)] 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 some of the relevant information between labels may provide additional useful information for multi-label learning, which is beneficial to generalize 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 effectively reduces the size of output space and greatly reduces the computational complexity. It also can effectively exploit the hidden structure of the original space and make full use of the correlation between labels. Hsu et al. (2009) proposed a label embedding method based on compressed sensing. *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. Tai and Lin (2012) proposed a label space embedding algorithm PLST which is based on classical dimensionality reduction algorithm PCA. PLST preserved the reservation of label space information by minimizing the square loss between original label space and latent space. Chen and Lin (2012) proposed a CPLST method based on canonical correlation analysis theory, which takes both label space embedding loss and regression loss in latent space into account, and achieves the effect of reducing the dimensionality of label space by using the feature space information. Lin et al. (2014) proposed an end-to-end label space embedding method, which can directly learn a better hidden space by maximizing the recovery of latent space and the prediction performance of latent space. The end-to-end mode breaks the limitation of latent space. Bhatia et al. (2015) proposed a SLEEC algorithm by using local correlations among instances. The tail labels attached only to a small number of instances which make the label matrix sparse but not low-rank. SLEEC can solve the problem effectively in practical problems. Yeh et al. (2017) proposed a multi-label embedding method based on deep learning and canonical correlation analysis. It can deal with the multi-label classification problem of large-scale data well by using the deep neural network for spatial transformation. Huang and Lin (2017) proposed a cost-sensitive multi-label embedding method CLEMS. In view of the problem that the multi-label evaluation criteria differ greatly in the same result, CLEMS takes evaluation criteria into consideration in multi-label learning for the first time.

Preserving intrinsic structure during embedding learning has already been explored, some existing methods are devoted to maintaining the between-view similarity structure Wang et al. (2016), or maintaining the structure of both views simultaneously [Pan et al. (2014), Quadrianto and Lampert (2011)]. However, in multi-label embedding methods, we would prefer to only preserve the original supervised information (label space) to enhance the discriminability of latent space. The information of the feature space is used as auxiliary information to improve the predictability of the latent space. In order to achieve both of the above goals at the same time, we proposed our Deep Correlation Structure Preserved Label Space Embedding (DCSPE) to maintain the structure of original label space and make the latent space feature-aware, which will be detailed in the following section.

3. Proposed Approach

For multi-label classification, let $D = \{(x_i, y_i)\}_{i=1}^{N} = \{X, Y\}$ denotes a set of d dimensional training instances $X \in R^{N\times d}$ and the associated labels $Y \in \{0,1\}^{N\times K}$, 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.

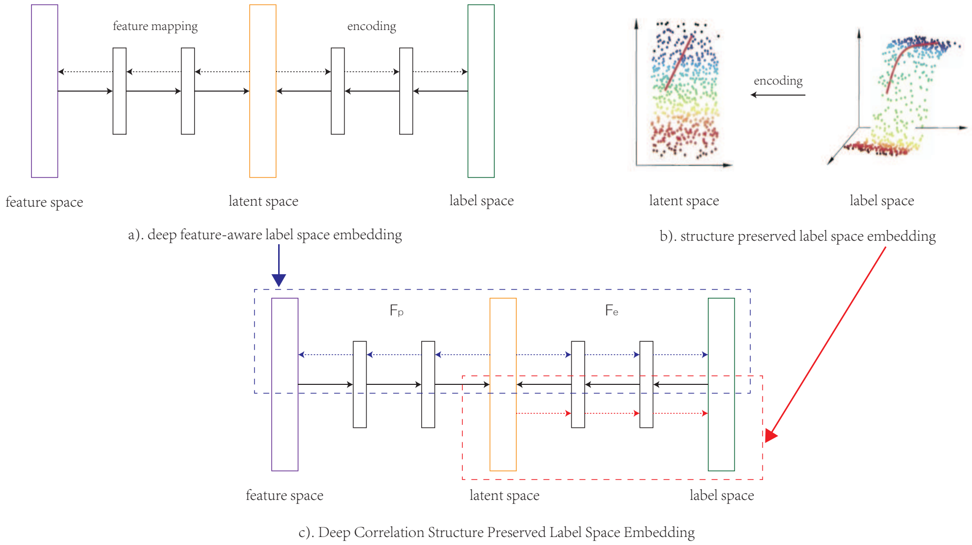

feature mapping
encoding
feature space
latent space
label space
a). deep feature-aware label space embedding
encoding
latent space
label space
b). structure preserved label space embedding
Fp
Fe
feature space
latent space
label space
c). Deep Correlation Structure Preserved Label Space Embedding

Figure 1: a) Deep feature-aware label space embedding proposed by C2AE Yeh et al. (2017)
b) Structure preserved label space embedding conduct by proposed DMDS
c) The proposed Deep Correlation Structure Preserved Label Space Embedding, which learns a deep latent space Z. We use the black lines to represent the DNNs (F_p and F_e). The blue and red dotted lines correspond to the error back propagation of DCCA and DMDS, respectively.

3.1. Deep Correlation Structure Preserved Label Space Embedding

As mentioned previously, we propose Deep Correlation Structure Preserved Label Space Embedding (DCSPE) which learns a deep feature-aware latent space that can preserve the intrinsic structure of the original label space. Specifically, the main steps of label embedding can be divided into two parts: encoding and decoding. Motivated by the developments of deep learning, we proposed a DMDS model based on the traditional MDS model to perform encoding. We denote the low-dimension label space in latent space as Z, for a test instance $\hat{x}$, we get the corresponding predicted embedded vector as $\hat{z} = g(\hat{x}) \in Z$. Our decoding function is to compute the nearest neighbor $z_q \in Z$ of $\hat{z}$ and return the corresponding $y_q \in Y$ as $\hat{y}$. DMDS maintains the similarity between samples in the label space. Similar samples in previous label space are still very close in the latent space, which effectively enhances the discriminant ability of the latent space. To improve the predictability of the latent space, we take the DCCA to correlate the feature space and label space. The DCCA can learn complex nonlinear transformations of feature space and label space such that the two view of multi-label data can be highly correlated in the common latent space. With the help of DCCA, we can effectively utilize the information of feature space to assist the label space embedding process, which makes the latent space we learned more predictive.

The training error of LSDR is mainly divided into two parts: one is the error of space transformation, and the other is the training error of the classifiers in the latent space. Our Deep Correlation Structure Preserved Label Space Embedding (DCSPE) can reduce the training error and space transformation error simultaneously and effectively. Firstly, the two transform functions (encoding and decoding) must be well matched and are able to carry out the space transformation and recovery succinctly and efficiently. In this paper, we effectively maintain the geometry of the original space by deep multidimensional scaling (DMDS) when we perform encoding and use the simple and effective nearest neighbor method to decode. Secondly, the learned latent space should have great predictability, hence we take deep canonical correlation analysis (DCCA) to make the encoding feature-aware, then the learned latent space can be closely correlated to the feature space. Through the above theoretical analysis of the training error, we can see that our model will effectively reduce the error and improve the prediction accuracy.

As shown in Figure 1, DCSPE contains two effective DNN models (*i.e.*, DMDS and DCCA) with two mapping functions to be determined: feature mapping F_p, encoding function F_e. The blue and red dotted lines correspond to the error back propagation of DCCA and DMDS respectively. We use DMDS which associates with Y to preserve the structure of the original label space and DCCA which associates with X and Y to make the latent space feature-aware. Thus, the objective function of DCSPE can be formulated as follows:

$$\min_{F_p, F_e} \quad \Phi(F_e(Y)) + \lambda \Psi(F_p(X), F_e(Y)) \tag{1}$$

where $\Phi(F_e(Y))$ and $\Psi(F_p(X), F_e(Y))$ denote the losses of structure preserved and feature-aware. And, we have the parameter λ balances between the above two types of loss functions.

Once the training process of our DCSPE is complete, we should apply the two DNNs to predict the labels of test inputs. For a test input $\hat{x}$, *firstly* we transform it into the derived latent space by F_p, $\hat{z} = F_p(\hat{x})$; *secondly* we select the nearest neighbor of $\hat{z}$ from Z which donates as z_q; *thirdly* we attach the y_q to $\hat{x}$ as its label vector.

3.2. Deep Multidimensional Scaling Analysis

Multidimensional scaling (MDS) describes a method of analyzing a dataset to reveal its intrinsic structure Kruskal (1964). It is a traditional dimensionality reduction approach that seeks to preserve the variability between instances, then similar points in the original dataset still lie together while the dissimilar points are still far away from each other. The basic idea of MDS is to keep the distance between any two points in a low dimensional space the same as their distance in the original space after dimensionality reduction. Traditional MDS is a classical linear dimension reduction method that only can reflect the linear relationship between sample points, but the relationship between sample points is nonlinear in many cases. With the introduction of deep neural networks, the proposed DMDS can conduct the non-linear dimension reduction while maintaining the original distance in low-dimensional latent space. To determine the $\Phi(F_e(Y))$ in Eq.(1), the objective function of deep multidimensional scaling (DMDS) method can be written as follows:

$$\min_{F_e} \quad \sum_{i,j}^{N} W_{ij}(d(F_e(y_i), F_e(y_j)) - \Delta_{ij})^2 \tag{2}$$

where Δ_{ij} denotes the dissimilarity between y_i and y_j.

Then we should construct Δ and W for solving DMDS. Considering that every label vector may have different importance, we set the importance weight W_{ij} to be the product of f_i and f_j, f_i is the frequency of y_i in D, and the symmetric W is constructed by the following equation:

$$W_{ij} = f_i \cdot f_j \tag{3}$$

We define a sparse symmetric $n \times n$ matrix Δ, indicating the dissimilarity among neighboring instances:

$$\Delta_{ij} = \begin{cases} \parallel y_i - y_j \parallel^2 & if \quad y_j \in \mathcal{N}_i \\ m & if \quad otherwise \end{cases} \tag{4}$$

where $\mathcal{N}_i$ is the instance set of i-th instance's k nearest neighbors.

3.3. Deep Canonical Correlation Analysis

Canonical correlation analysis (CCA) is a standard statistical technique for finding linear projections of two random vectors that are maximally correlated. Deep canonical correlation analysis (DCCA) is an extension of CCA in which maximally correlated nonlinear projections. DCCA computes representations of the two views by passing them through multiple stacked layers of nonlinear transformation. As two linear projections replaced by DNNs, DCCA solves the same objective function with the DNN model by a gradient descent technique Andrew et al. (2013). To determine $\Psi(F_p(X), F_e(Y))$ in Eq.(1), we adopt the idea of Kettenring (1971) and rewrite the correlation-based objective function as the following deep version: Yeh et al. (2017):

$$\begin{aligned} \min_{F_p, F_e} \quad & \|F_p(X) - F_e(Y)\|_F^2 \\ s.t. \quad & F_p(X)^T F_p(X) = F_e(Y)^T F_e(Y) = I \end{aligned} \tag{5}$$

where $F_p(X)$ and $F_e(Y)$ denote the transformed feature and label data in the derived latent space L.

As explained in Kettenring (1971), the above formulation is equivalent to the traditional CCA objection function of correlation maximization, but using the above formulation is more conducive for us to calculate the corresponding neural network loss and gradient of the two functions F_p and F_e efficiently.

3.4. Optimization

To learn the model of DCSPE, we need to solve the optimization problem of Eq.(1). Similar to the derivation of existing DNN models, we apply the technique of random gradient descent for each loss term to update the corresponding network parameters. The gradient of $\Phi(F_e)$ updates the encoding function F_e while that of $\Psi(F_p, F_e)$ updates both feature mapping function F_p and encoding function F_e.

There are several algorithms available in the literature for solving MDS. A representative algorithm is Scaling by MAjorizing a COmplicated Function (SMACOF) Kruskal (1964).

The objection of $\Phi(F_e)$ can be transformed to:

$$\min_{Z} \quad \eta^2(Z) - 2\rho(Z) + \eta^2_{\Delta} \tag{6}$$

with the following definitions:

$$Z = F_e(Y)$$
$$\eta^2(Z) = \sum_{i,j}^{N} W_{ij} d(z_i, z_j)^2, \quad \rho(Z) = \sum_{i,j}^{N} W_{ij} \Delta_{ij} d(z_i, z_j), \quad \eta^2_{\Delta} = \sum_{i,j}^{N} W_{ij} \Delta^2_{ij} \tag{7}$$

Let us denote the matrix $A_{ij} = (e_i - e_j)(e_i - e_j)'$ whose elements equal 1 at $a_{ii} = a_{jj} = 1, -1$ at $a_{ij} = a_{ji}$, and 0 otherwise. Furthermore, we define:

$$V = \sum_{i=1}^{N} \sum_{i=1}^{N} w_{ij} A_{ij} \tag{8}$$

as the weighted sum of row and column centered matrices A_{ij}. Hence, we can rewrite $\eta^2(Z) = trZ'VZ$. For a similar representation of $\rho(Z)$, we define the matrix:

$$B(Z) = \sum_{i=1}^{N} \sum_{i=1}^{N} w_{ij} s_{ij}(Z) A_{ij} \tag{9}$$

where:

$$s_{ij}(Z) = \begin{cases} \Delta_{ij}/d_{ij}(Z) & if \quad d_{ij}(Z) > 0 \\ 0 & if \quad d_{ij}(Z) = 0 \end{cases} \tag{10}$$

By using $B(Z)$, we can rewrite $\rho(Z)$ as $\rho(Z) = trZ'B(Z)Z$, consequently, the stress decomposition becomes:

$$\Phi(Z) = \eta^2_{\Delta} + trZ'V(Z) - 2trZ'B(Z)Z \tag{11}$$

Algorithm 1 Training process of DCSPE

```
**Input:** Feature matrix $X$, label matrix $Y$, parameter $\lambda$, and dimension $M$ of the latent space
**Output:** $F_p$, $F_e$
Randomly initialize $F_p$, $F_e$
Calculate $W$ and $\Delta$ by Eqs.(3) (4)
Calculate $V$ and $B$ by Eqs.(8) (9)
repeat
  Randomly select a batch of data $S[X]$ and $S[Y]$
  Perform gradient descent on $F_p$ by Eq.(15)
  Perform gradient descent on $F_e$ by Eqs.(12) (16)
until Converge
```

Algorithm 2 Predicting process of DCSPE

```
**Input:** Label matrix $Y$, $F_p$, $F_e$, testing example $\hat{x}$
**Output:** Prediction $\hat{y}$
Computing the embedding of $Y$, $Z = F_e(Y)$
Obtain the predicted vector $\hat{z} = F_p(\hat{x})$
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
```

The gradient of $\Phi(F_e)$ with respect to F_e can be derived as:

$$\frac{\partial \Phi(F_e)}{\partial F_e(Y)} = 2VF_e(Y) - 4B(F_e(Y))F_e(Y) \tag{12}$$

We find that the DMDS model can be well optimized by random gradient descent and we perform the same method to optimize the DCCA model. The DCCA is a constrained optimization problem, so we perform Augmented Lagrange method on it. With the method of Lagrange multipliers, the objection of $\Psi(F_p, F_e)$ can be transformed to:

$$\Psi(F_p, F_e) = Tr(C_1^T C_1) + \alpha Tr(C_2^T C_2 + C_3^T C_3) \tag{13}$$

where:

$$C_1 = F_p(X) - F_e(Y), \quad C_2 = F_p(X)^T F_p(X) - I, \quad C_3 = F_e(Y)^T F_e(Y) - I \tag{14}$$

The gradient of $\Psi(F_p, F_e)$ respecting to $F_p(X)$ and $F_e(Y)$ can be derived as:

$$\frac{\partial \Psi(F_p, F_e)}{\partial F_p(X)} = 2C_1 + 4\alpha F_p(X) C_2 \tag{15}$$

$$\frac{\partial \Psi(F_p, F_e)}{\partial F_e(Y)} = 2C_1 + 4\alpha F_e(Y) C_3 \tag{16}$$

With the above derivations, we can learn our DCSPE by gradient descent, and the pseudo code of training is summarized in Algorithm 1. Once the learning of DCSPE is complete, label prediction of a test input can be easily achieved by nearest neighbor algorithm and the pseudo code of prediction is summarized in Algorithm 2.

3.5. Computational Complexity Analysis

The computational cost of the proposed DCSPE is analyzed in this section. We denote h_p, h_e, q as the number of hidden layers in F_p, F_e and the hidden units in each hidden layer. Let W be the total number of weights and biases of the network in DCSPE, *i.e.* $W = (d+1) \times q + h_p \times (q+1) \times q + (q+1) \times M$. We apply mini-batch gradient descent for each loss term for updating the corresponding network parameters. We denote r, k as the number of samples in each random block and the number of iterations. For simplicity, we assume that $N >> d > r, K \geqslant M$ holds in the real-world applications.

The computational complexity of training DCSPE includes four main parts, the forward propagation phase, the calculation of gradient, the backward propagation phase and the weights and biases update phase. (1) In the forward propagation phase, leading to an overall computational cost which is $\mathcal{O}(krW)$; (2) In the gradient calculation phase, computing Eq.(11) requires $\mathcal{O}(kr^2d)$ and computing Eq.(14), (15) requires $\mathcal{O}(krd^2)$. Then the overall computational cost of this phase is $\mathcal{O}(krd^2)$. (3) In the backward phase, the overall computational cost for one sample in the backward propagation phase is $\mathcal{O}(M^2) + \mathcal{O}(Mq)$, which is at most $\mathcal{O}(W)$, Then the overall computational cost of this phase is $\mathcal{O}(krW)$; (4) In the weights and biases update phase, it is evident that the overall computational cost is again $\mathcal{O}(krW)$ Zhang and Zhou (2006). Thus, the overall training cost of DCSPE is $\mathcal{O}(krd^2)$, the computational complexity is acceptable and our proposed DCSPE is able to handle multi-label applications with many labels.

4. Experiments

4.1. Datasets and Settings

To validate the proposed Deep Correlation Structure Preserved Label Space Embedding (DCSPE), we download eight benchmark datasets in different domains with relatively large vocabularies for experiments, *i.e.*, EUR-Lex (directory), mediamill, EUR-Lex (subject matters), delicious, Corel5k, iaprtc12, ESPGame, NUSWIDE.

The statistics of the eight real world datasets are summarized in Table 1. For the datasets of text and video (EUR-Lex (directory), mediamill, EUR-Lex (subject matters), delicious), we use traditional features from Mulan Tsoumakas et al. (2011). For datasets of image (Corel5k, iaprtc12, ESPGame, NUSWIDE), we extract 4096-dimensional deep features by using the 16 layers VGG network Simonyan and Zisserman (2015) pretrained on ImageNet 2012 classification challenge dataset Deng et al. (2009) with MatConvNet, we didn't perform the fine-tuning for fairness and computational efficiency.

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 of all the methods. Note that for the multi-label metrics, their values vary between [0,1]. For Macro-F1 and Micro-F1, larger the values better the performance, but for One-error and Ranking loss, smaller the values better the performance. The definitions of the four metrics can be found in Zhang and Zhou (2014).

In our experiments, we compared the following state-of-art multi-label classification methods: Feature-aware Implicit label space Encoding (FaIE) Lin et al. (2014), Sparse Local Embeddings for Extreme Classification (SLEEC) Bhatia et al. (2015), Bayesian Non-

Table 1: Datasets properties

Dataset	Domain	Instances	Feature	Labels	Cardinality
delicious	text	16105	500	983	19.0
mediamill	video	43907	120	101	4.4
EUR-Lex(dc)	text	19348	5000	412	1.3
EUR-Lex(sm)	text	19348	5000	201	2.2
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

parametric Multi-label Classification (BNMC) Nguyen et al. (2016), Canonical-Correlated Autoencoder (C2AE) Yeh et al. (2017), Cost-sensitive Label Embedding with Multidimensional Scaling (CLEMS) Huang and Lin (2017). We also report the results of some baseline algorithms, such as Binary Relevance (BR) Fürnkranz et al. (2008), Classifier Chain (CC) Read et al. (2011), Deep Canonical Correlation Analysis (DCCA) Andrew et al. (2013).

To select the parameters for these methods, we randomly hold one fifth of training data in every datasets for validation. For the architecture of our DCSPE, we have F_p composed of 2 layers of fully connected layer structures while the embedding function F_e is 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 as 200. We select λ from $\{10^{-5}, 10^{-4}, \cdots, 10^{5}\}$ and follow the Yeh et al. (2017) fixed α as 0.5. All the experiments are performed on a 64-Bit Linux workstation with an Intel E5-2650 CPU, a NVIDIA Titan X Pascal card and 256GB memory.

4.2. Experiments Results of LSDR

We perform all algorithms on the datasets with different values of M/K (from 20% to 100%) where M and K are respectively the dimension of latent space and the original label space.

Figure 2 illustrates and compares the performances of the above methods, in which the horizontal axis denotes the ratio of the latent space dimension (M/K). From the figure, we can see that our DCSPE method performed favorably against most label embedding methods in most cases, which well demonstrates its effectiveness. From the experimental results, we can draw the following interesting observations.

The proposed DCSPE significantly outperform most of the baselines on the eight data sets. For example, on EUR-Lex (subject matters), our method improves the best results of the baselines by 4.5% (Micro-F1), 5.6% (Micro-F1) performance of the proposed methods, which validates our theoretical results. (1) In order to reflect the importance of considering the structure of label space, we set the λ as zero, then our model degenerated into DCCA. From the experimental results, We can clearly find that the result of our proposed DCSPE is significantly better than DCCA, which effectively reflects the importance of preserving intrinsic structural information of label space in multi-label embedding learning. (2) We can find that on all datasets, as M increases, the performance of FaIE varies a little, which is due to the orthonormality constraint on latent space for enabling FaIE to compactly

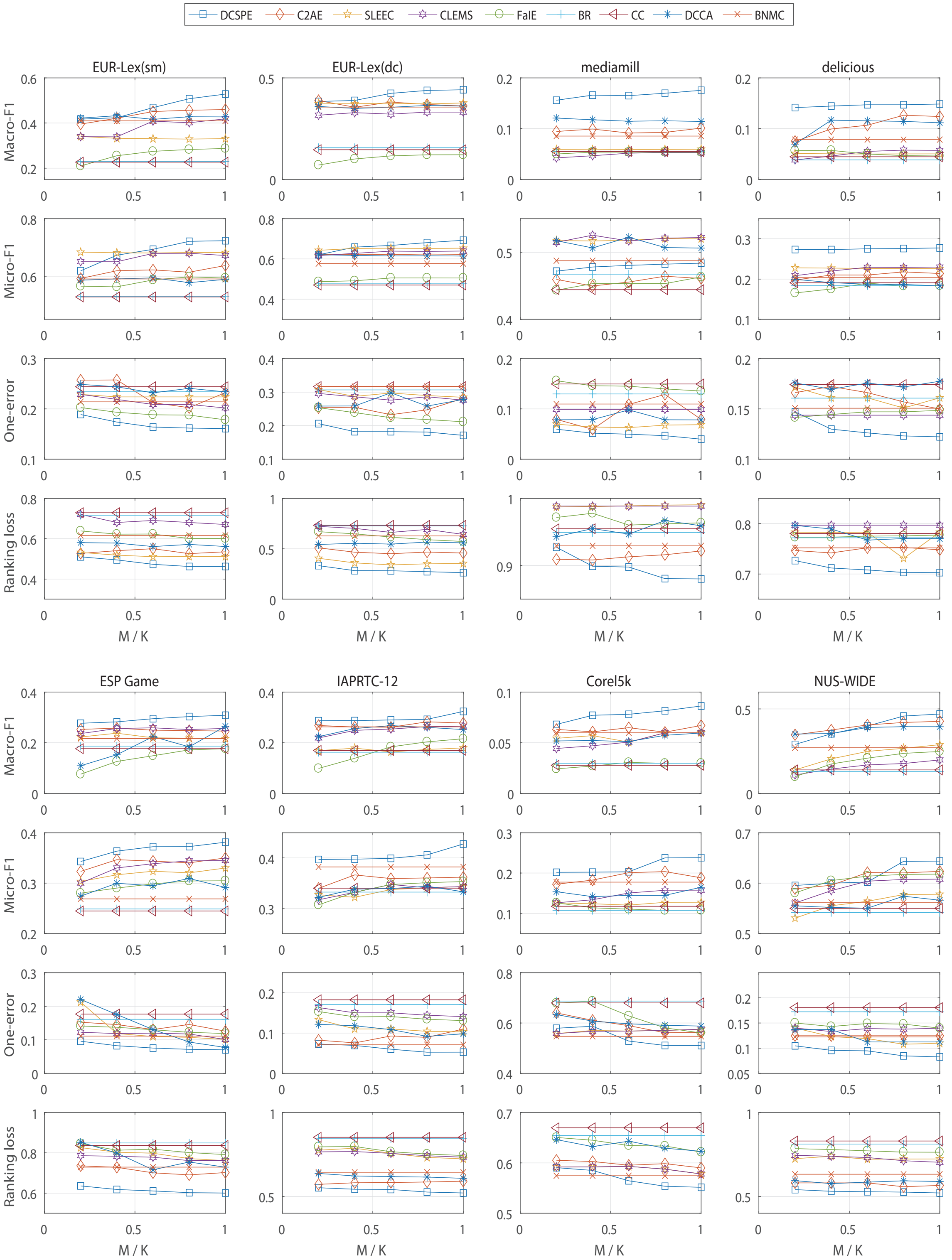


Figure 2: Performance comparisons with different latent space dimension ratios (M/K)

encode the original label space with a smaller M Lin et al. (2014). But our method and C2AE may decrease slightly sometimes when the M increases, because our orthonormality only conducts on the DCCA model and the DCCA is just a part of our method. When the ratio is above 40%, the result is basically stable. (3) 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 excavate the hidden structure of the original label space. FaIE is based on CCA, which describes the linear correlations between the feature space and label space. We adapt the DCCA to analyze the nonlinear correlations between feature space and label space. We found that the nonlinear method can describe the correlation between the feature and label space better. (4) Compared to the CLEMS method which adapts the traditional DMS model, we can clearly find that the CLEMS method didn't consider the feature space when conducting label embedding, which may lead to the result that the latent space learned may not have enough predictability. (5) In our method, the neural network F_p have two roles, one is the nonlinear projection from feature space to latent space, the other is the prediction function in testing phase. We can clearly see that our method performs well on datasets in different fields because our decoding function is robust and matches well with the encoding function. The experimental results on the datasets also verify the stability and generalization of our algorithm.

Table 2: Visualization of embedded labels for IAPRTC-12.

Label	Nearest Neighbors
sea	beach, coast, wave, bay, island, boat
table	room, chair, wood, bed, curtain
grandstand	stadium, seat, uniform, team, player, spectator
cliff	waterfall, canyon, face, rope, shrub
grave	mummy, skull, bone, corridor, couple
church	cathedral, ornament, bell, pinnacle, clock

To further verify the effectiveness of our derived deep latent space, we consider several example labels from IAPRTC-12 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 DCSPE in sufficiently exploiting label dependency during the learning process.

4.3. Case Study of Multi-label Image Annotation

We present a case study in which the proposed method is applied to a multi-label image annotation application. DCSPE is applied on the famous NUS-WIDE data sets. The annotation results of several randomly selected images are illustrated in Figure 3. The proposed DCSPE 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 2 with the label 'sunset' missed in the ground truth. The performance on multi-label image annotation applications suggests that our methods can work well in the real-world image annotation applications.

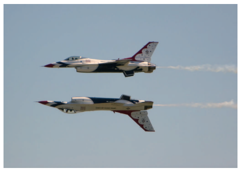
NUSWIDE				
Groundtruth	reflection sun tree water	clouds lake ocean sky water	beach clouds lake ocean sky water	clouds ocean person sky water
DCSPE Annotation	plants reflection tree water	clouds lake ocean sky sun sunset water	boats clouds lake ocean sky water	clouds person sky water
NUSWIDE				
Groundtruth	boats clouds sky vehicle	clouds lake ocean water	clouds road sky tree	clouds military plane sky
DCSPE Annotation	beach boats clouds sky vehicle water	clouds lake ocean sky sunset water	clouds plants road sky	clouds military plane sky

Figure 3: Several multi-label image annotation examples on NUS-WIDE data sets. 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.

4.4. Parameter Analysis

Furthermore, we conduct experiments to see the effects of the only parameter λ in the proposed DCSPE. Figure 4 gives an illustration of the variances of multi-label classification performance as λ varies in $\{10^{-10}, 10^{-9}, \cdots, 10^{10}\}$ in a run on Corel5k with M/K = 5%. From the illustration, we can draw the observations that the performance of DCSPE, in terms of Macro-F1 and Micro-F1, firstly increases and then decreases as λ varies from 10^{-10} to 10^{10}.

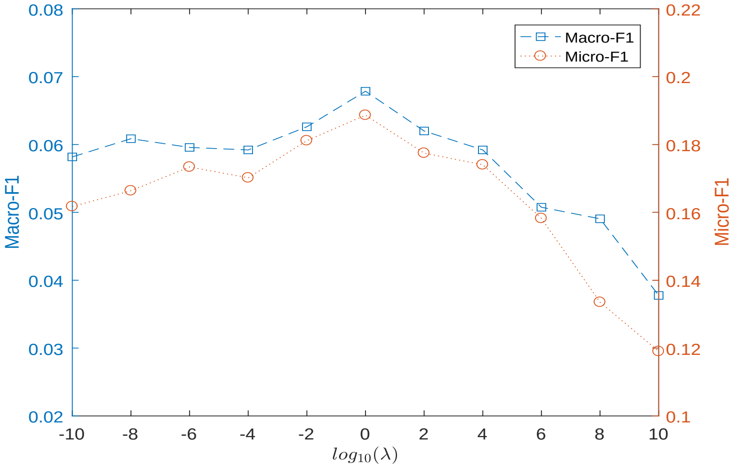


Figure 4: Effects of λ in DCSPE on the performance of multi-label classification on Corel5k

4.5. Convergence Analysis

In this section, we empirically study the convergence of the proposed DCSPE. The convergence curves of DCSPE on Corel5k dataset with $M/K = 0.8$ are plotted in Figure 5. As can be seen in the figure, the objective converges quickly in a few iterations. We omit the results of the other datasets since they are similar with the observation in Figure 5.

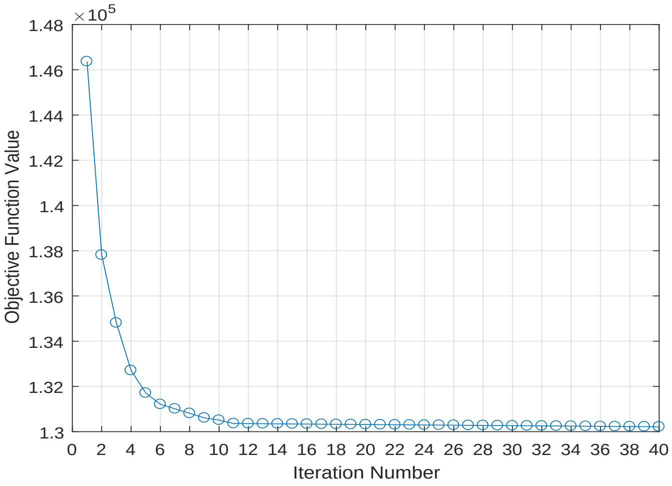


Figure 5: Convergence curce of the proposed DCSPE on Corel5k

5. Conclusion

In this paper, we proposed Deep Correlation Structure Preserved Label Space Embedding (DCSPE) for solving the problem of multi-label classification with many labels. DCSPE can successfully learn a feature-aware deep latent space which preserves the intrinsic structure of the original label space. The deep latent space we learned has strong predictability and discriminant ability. DCSPE makes the prediction more precise by reducing the error of space transformation and the training error of classifiers in the latent space. The experiment results demonstrate that DCSPE is superior to state-of-the-art label embedding algorithms.

There are many interesting future works. For example, our current proposal does not consider the problem of missing labels. Dealing with weak label data in the process of learning deep semantic latent space is an interesting issue in the future.

Acknowledgments

This work is supported by National Natural Science Foundation of China (61876087, 61432008), CERNET Innovation Project (NGII20170524) and Postgraduate Research & Practice Innovation Program of Jiangsu Province (KYCX18_1238).

References

Galen Andrew, Raman Arora, Jeff Bilmes, and Karen Livescu. Deep canonical correlation analysis. In *Proceedings of the 30th International Conference on Machine Learning*, pages 1247–1255, 2013.

Kush Bhatia, Himanshu Jain, Purushottam Kar, Manik Varma, and Prateek Jain. Sparse local embeddings for extreme multi-label classification. In *Advances in Neural Information Processing Systems 28*, pages 730–738, 2015.

Wei Bi and James T. Kwok. Efficient multi-label classification with many labels. In *Proceedings of the 30th International Conference on Machine Learning*, pages 405–413, 2013.

Yao-Nan Chen and Hsuan-Tien Lin. Feature-aware label space dimension reduction for multi-label classification. In *Advances in Neural Information Processing Systems 25*, pages 1529–1537, 2012.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009.

Johannes Fürnkranz, Eyke Hüllermeier, Eneldo LozaMencía, and Klaus Brinker. Multilabel classification via calibrated label ranking. *Machine Learning*, 73(2):133–153, 2008.

Daniel Hsu, Sham M. Kakade, John Langford, and Tong Zhang. Multi-label prediction via compressed sensing. In *Advances in Neural Information Processing Systems 22*, pages 772–780, 2009.

Kuan-Hao Huang and Hsuan-Tien Lin. Cost-sensitive label embedding for multi-label classification. *Machine Learning*, 106(9-10):1725–1746, 2017.

Jon Roberts Kettenring. Canonical analysis of several sets of variables. *Biometrika*, 58(3): 433–451, 1971.

Joseph Bernard Kruskal. Multidimensional scaling by optimizing goodness of fit to a non-metric hypothesis. *Psychometrika*, 29(1):1–27, 1964.

Zi-jia Lin, Gui-guang Ding, Ming-qing Hu, and Jian-min Wang. Multi-label classification via feature-aware implicit label space encoding. In *Proceedings of the 31th International Conference on Machine Learning*, pages 325–333, 2014.

Vu Nguyen, Sunil Gupta, Santu Rana, Cheng Li, and Svetha Venkatesh. A bayesian non-parametric approach for multi-label classification. In *Proceedings of the 8th Asian Conference on Machine Learning*, pages 254–269, 2016.

Ying-wei Pan, Ting Yao, Tao Mei, Hou-qiang Li, Chong-Wah Ngo, and Yong Rui. Click-through-based cross-view learning for image search. In *Proceedings of the 37th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 717–726, 2014.

Novi Quadrianto and Christoph H. Lampert. Learning multi-view neighborhood preserving projections. In *Proceedings of the 28th International Conference on Machine Learning*, pages 425–432, 2011.

Jesse Read, Bernhard Pfahringer, Geoff Holmes, and Eibe Frank. Classifier chains for multi-label classification. *Machine Learning*, 85(3):333–359, 2011.

Xiao-bo Shen, Wei-wei Liu, Ivor W. Tsang, Quan-Sen Sun, and Yew-Soon Ong. Multilabel prediction via cross-view search. *IEEE Transactions on Neural Networks and Learning Systems*, 29(9):4324–4338, 2018.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, 2015.

Farbound Tai and Hsuan-Tien Lin. Multi-label classification with principle label space transformation. *Neural Computation*, 24(9):2508–2542, 2012.

Grigorios Tsoumakas, Eleftherios Spyromitros-Xioufis, Jozef Vilcek, and Ioannis Vlahavas. Mulan: A java library for multi-label learning. *Journal of Machine Learning Research*, 12(7):2411–2414, 2011.

Li-wei Wang, Yin Li, and Svetlana Lazebnik. Learning deep structure-preserving image-text embeddings. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 5005–5013, 2016.

Chih-Kuan Yeh, Wei-Chieh Wu, Wei-Jen Ko, and Yu-Chiang Frank Wang. Learning deep latent spaces for multi-label classification. In *Proceedings of the 31th AAAI Conference on Artificial Intelligence*, pages 2838–2844, 2017.

Min-Ling Zhang and Zhi-Hua Zhou. Multilabel neural networks with applications to functional genomics and text categorization. *IEEE Transactions on Knowledge and Data Engineering*, 18(10):1338–1351, 2006.

Min-Ling Zhang and Zhi-Hua Zhou. A review on multi-label learning algorithms. *IEEE Transactions on Knowledge and Data Engineering*, 26(8):1819–1837, 2014.

Wen-Ji Zhou, Yang Yu, and Min-Ling Zhang. Binary linear compression for multi-label classification. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, pages 3546–3552, 2017.

Zhi-Hua Zhou and Min-Ling Zhang. Multi-label learning. *Encyclopedia of Machine Learning and Data Mining*, pages 875–881, 2017.

Yue Zhu, James T. Kwok, and Zhi-Hua Zhou. Multi-label learning with global and local correlation. *IEEE Transactions on Knowledge and Data Engineering*, 30(6):1081–1094, 2018.