Latent Topic-aware Multi-Label Classification

Jianghong Ma[1][0000−0002−0524−3584] and Yang Liu[2][0000−0003−2252−3665]

[1] City University of Hong Kong, Hong Kong
jianghma2@cityu.edu.hk
[2] The Hong Kong University of Science and Technology, Hong Kong

Abstract. In real-world applications, data are often associated with different labels. Although most extant multi-label learning algorithms consider the label correlations, they rarely consider the topic information hidden in the labels, where each topic is a group of related labels and different topics have different groups of labels. In our study, we assume that there exists a common feature representation for labels in each topic. Then, feature-label correlation can be exploited in the latent topic space. This paper shows that the sample and feature exaction, which are two important procedures for removing noisy and redundant information encoded in training samples in both sample and feature perspectives, can be effectively and efficiently performed in the latent topic space by considering topic-based feature-label correlation. Empirical studies on several benchmarks demonstrate the effectiveness and efficiency of the proposed topic-aware framework.

Keywords: Multi-label learning · Sample and feature extraction · Feature-label correlation · Topic

1 Introduction

Multi-label classification (MLC) is a task of predicting labels of new samples based on training sample-label pairs [24, 6]. Usually, the number of training samples is high and the number of features for each sample is also high. As involving irrelevant samples or features can negatively impact model performance, the academia has seen many efforts for extracting an informative subset of samples or features for classification. For sample extraction, some works select a subset of common training instances shared by all testing instances [5, 25], while some works select a subset of different training instances for each testing instance [27, 23]. For feature extraction, some works select the same features for all labels [29, 20], while some works select different features for each label [10, 9]. In our study, we assume each testing instances should have its own specific training instances. However, most of instance-specific sample extraction methods overlook the gap [27, 1] between features and labels. The correlation between instances based on features cannot be assumed to be the same as that based on labels. Although the method in [23] has discovered this problem, it is still not clear why the input-output correlation can be well captured in the learned latent subspace. We also

assume each label should have its own specific features. However, most of label-specific feature extraction methods overlook the relationship between features and labels. They often select discriminative features for each label based on label correlations rather than based on feature-label correlations.

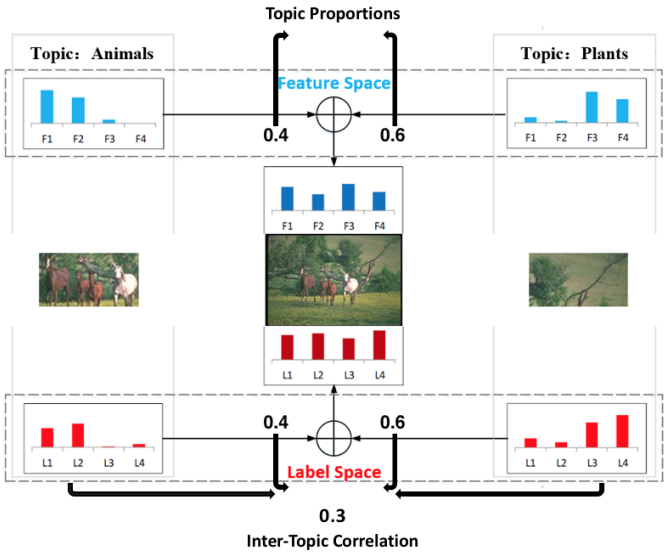

Fig. 1. A simple example of topic-aware data factorization

Based on the above discussion, this paper focuses on bridging the input-output gap and exploiting input-output correlation for sample and feature extraction respectively. Here, we propose a novel topic-aware framework by assuming each sample can be seen as a combination of topics with different proportions. The input and output share the same topic proportions, but they have different feature and label distributions for different topics. For showing a simple example, an image in the corel5k dataset [4] in Fig.1 can be seen as a weighted combination of topic animals and plants with proportion coefficient being 0.4 and 0.6 respectively, where each topic has its own feature and label distributions. The important labels for topic animals are 'horses' and 'foals' while the important labels for topic plants are 'trees' and 'grass' in the given example. The important features for each topic should also be different. It should be noted that we assume some topics are correlated to each other. Two topics are assumed to be highly correlated to each other if they share similar label distributions or they often co-occur in samples. The topic proportions and feature/label proportions can be mined by non-negative matrix factorization on both input and output spaces. As features and labels share the same latent topic space, there is no gap between features and labels in this space. We can exploit the inter-instance relationship

in the latent topic space. This kind of relationship can be directly applied to the output space. For example, if an image and the image in Fig.1 share the similar topic distribution instead of feature distribution, we assume these two images have similar label sets. Because a shared structure between features and labels is extracted in the latent topic space, the correlation between features and labels is mined in this space. This topic distribution can be seen as new features for each instance. We then exploit the topic-label relationship for label-specific new feature extraction instead of original feature extraction. For example, the label 'grass' is only related to several topics, such as topic plants in Fig.1. These topics can be seen as discriminative new features for the label 'grass'.

The major contributions of this paper are summarized as:

- Introducing a novel concept, topic, where each instance is combined by multiple topics with different proportions and each topic has its corresponding feature/label distributions;
- Proposing a label-specific feature extraction algorithm in the learned topic space by considering the relationship between features and labels;
- Proposing an instance-specific sample extraction algorithm in the learned topic space by considering the gap between features and labels;
- Conducting intensive experiments on multiple benchmark datasets to demonstrate the effectiveness of the proposed topic-aware framework.

2 Topic-aware Multi-Label Classification-TMLC

In this section, we introduce the formulation of the proposed framework.

2.1 Preliminaries

Here, the matrix and the vector are denoted by the uppercase character (e.g., Γ) and lowercase character (e.g., γ) respectively. For matrix Γ, its $(i,j)^{th}$ entry is represented as $\Gamma_{i,j}$; its i^{th} row and j^{th} column is represented as $\Gamma_{i,:}$ and $\Gamma_{:,j}$ respectively. The column vector e is a vector with all entries being 1.

Suppose in each multi-labeled dataset, the input data is represented by $X = [X^t, X^s] \in \mathbb{R}^{d\times n}$ with $X^t = [x_1^t, ..., x_{n_t}^t]$ and $X^s = [x_1^s, ..., x_{n_s}^s]$ as training and testing input matrices respectively; the output data is represented by $Y^t = [y_1^t, ..., y_{n_t}^t] \in \{0,1\}^{k\times n_t}$, where n_t, n_s, d and k is the number of training samples, testing samples, features and labels respectively.

2.2 The overview of TMLC

In order to perform topic-aware multi-label classification, we consider two-level mapping. The commonly used mappings in the traditional MLC algorithms is illustrated in Fig.2(a). First, the predictive model h of the mapping between X^t and Y^t in the training data can be applied to the testing data [17]. Second, the predictive model g of the mapping between X^t and X^s in the input data can be

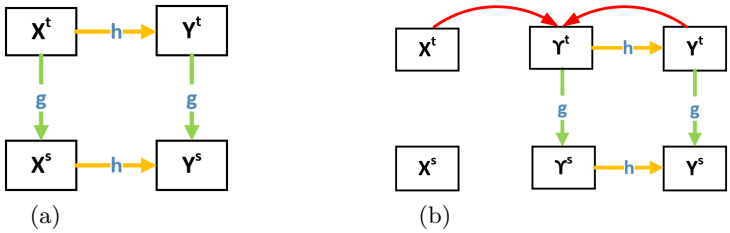

Fig. 2. (a)The mappings in the traditional MLC methods, (b)The mappings in the proposed MLC methods.

applied to the output data [14]. In fact, the first and second kind of mapping usually overlooks the input-output correlation and input-output gap respectively. In the proposed topic-aware framework shown in Fig.2(b), we learn the mapping between Υ^t and Y^t which can be applied to the corresponding testing data in the latent topic space, where Υ^t encodes the input-output correlation by learning it as a shared structure between inputs and outputs. We also learn the mapping between Υ^t and Υ^s which can be applied to the output data, as there exists no gap between Υ^t and Y^t because inputs and outputs share the same Υ.

In our study, Υ^t can be learned from the original X^t and Y^t by assuming each instance can be seen as a combination of different topics where each topic has different feature and label distributions. The detailed exploration of Υ^t and Υ^s in the training and testing spaces respectively can be found in Section 2.3. We assume that some topics are often correlated to each other. The extraction of inter-topic correlation which can be used to guide the following topic-aware feature and sample exactions can be found in Section 2.4. For the first kind of mapping in Fig.2(b), we perform label-specific new feature extraction by exploiting the topic-label relationship. The corresponding technical details can be found in Section 2.5. For the second kind of mapping in Fig.2(b), we perform instance-specific sample with new representations extraction by exploiting the inter-instance relationship in the latent topic space. The corresponding technical details can be found in Section 2.5.

2.3 Topic-aware Data Factorization

In our study, we assume each instance can be reconstructed by some topics with different weights. Each topic combines some related labels (objects), where these related labels share a common feature space. Then, the topic space can be seen as a common subspace shared by feature and labels. The common subspace establishes the correlations between the input and output. Specifically, X^t (Y^t) can be decomposed into a matrix that describes the feature (label) distributions of different topics and a matrix that describes the topic proportions of different instances, which is called topic-aware factorization. Matrix factorization is widely used for classification or data representation [30, 31, 22]. The mathematic formulation of topic-aware data factorization for feature and label structure in

training data can be found as follows:

$$\begin{aligned}\min_{F,L,\Upsilon^t} \sum_{i}^{n_t} \left\|x_i^t - F\upsilon_i^t\right\|_F^2 + \sum_{i}^{n_t} \left\|y_i^t - L\upsilon_i^t\right\|_F^2 \\ + 2\lambda(\|F\|_1 + \|L\|_1),\ s.t.\ F, L, \upsilon_i^t \geq 0,\end{aligned} \tag{1}$$

where $F = [f_1, ..., f_r] \in \mathbb{R}^{d\times r}$ and each f_i is a vector to indicate the feature distribution in topic i with r being the number of topics. The large (small) value in f_i denotes that the corresponding feature is highly (weakly) related to topic i. $L = [l_1, ..., l_r] \in \mathbb{R}^{k\times r}$ and each l_i is a vector to indicate the label distribution in topic i with r being the number of topics. The feature and label spaces share the same latent space $\Upsilon^t = [\upsilon_1^t, ..., \upsilon_{n_t}^t] \in \mathbb{R}^{r\times n_t}$ with each υ_m^t is used to automatically weight different topics for training sample m, as features and labels are two parallel views to represent each topic but with different distributions. As each topic is only related to a few number of features (labels), the corresponding F (L) should be sparse. The parameter λ is used to control the sparsity.

It should be noted that the i^{th} row of L ($L_{i,:}$) indicates the topic distribution for label i. Then, $L_{i,:}$ can be seen as the representation of label i in the topic space. If label i and label j are highly correlated, the corresponding $\|L_{i,:} - L_{j,:}\|_F^2$ should be small. Then, labels in the latent topic space can be jointly learned by involving label correlation as follows:

$$\min_{L} \sum_{i,j}^{k,k} C_{i,j} \left\|L_{i,:} - L_{j,:}\right\|_F^2, s.t.\ L \geq 0, \tag{2}$$

where we utilize cosine similarity to calculate C with $C_{i,j} = \cos(Y_{i,:}^t, Y_{j,:}^t)$.

After obtaining the feature structure of topics in the training data (F), the topic proportions of testing samples can be found as follows:

$$\min_{\Upsilon^s} \sum_{j}^{n_s} \left\|x_j^s - F\upsilon_j^s\right\|_F^2,\ s.t.\ \upsilon_j^s \geq 0, \tag{3}$$

where $\Upsilon^s = [\upsilon_1^s, ..., \upsilon_{n_s}^s] \in \mathbb{R}^{r\times n_s}$ and each υ_j^s indicates the topic proportions of testing sample j.

The values in feature/label distributions in each topic and topic proportions in each sample are all assumed to be nonnegative, as nonnegativity is consistent with the biological modeling of data [18, 8, 13], especially for image data.

2.4 Inter-topic correlation

In our study, we assume topics are correlated to each other, which is always neglected in many works. Because some topics may share similar label distributions and some topics may often co-occur in some samples, we can learn inter-topic correlation based on L which shows the label distributions of different topics and Υ^t which shows the topic combinations of different samples. As topic information

can be represented by labels and labels can be inferred by each other, we can learn the inter-label correlation by considering the inter-topic correlation. Thus, the inter-topic correlation can be exploited by the following objective function

$$\begin{aligned} \min_{\Xi} \left\|Y^t - L\Xi\Upsilon^t\right\|_F^2 + \left\|C - L\Xi L^T\right\|_F^2, \\ s.t.\ \Xi = \Xi^T, \Xi e = e, \Xi \geq 0, diag(\Xi) = 0, \end{aligned} \tag{4}$$

where $\Xi \in \mathbb{R}^{r\times r}$ is the matrix to denote the relational coefficients between any pair of topics. In the next sections, we will show that inter-topic correlation can be directly used to guide feature and sample extractions in the topic space.

2.5 Topic-aware Label-specific Feature Extraction

One of the objective of our task is to learn the predictive mapping L from latent topics to labels. Each $L_{i,j}$ represents the selection weight for label i to topic j. The higher the relationship between label i and topic j, the higher weight should be given to $L_{i,j}$.

In order to learn the importance of each label in each topic, we add another term to Eq.(1) as follows:

$$\min_{L,\Upsilon^t} \sum_{i,j}^{k,r} L_{i,j} \left\|Y_{i,:}^t - \Upsilon_{j,:}^t\right\|_F^2, s.t.\ L, \Upsilon^t \geq 0, \tag{5}$$

where $Y_{i,:}^t$ indicates the distribution of label i in different samples and $\Upsilon_{j,:}^t$ indicates the distribution of topic j in different samples. The more similar $Y_{i,:}^t$ and $\Upsilon_{j,:}^t$, the more likely the topic j can be characterized by the label i.

After obtaining the predictive mapping L, the label prediction based on the label-specific feature extraction in the topic space is

$$Y^s = L\Psi\Upsilon^s, \tag{6}$$

when considering inter-topic correlation, where

$$\Psi = \iota I + (1-\iota)\Xi, \tag{7}$$

where ι is a parameter to balance the weight between a topic and other topics.

2.6 Topic-aware Instance-specific Sample Extraction

One of the objective of our task is to learn the predictive mapping $\Theta \in \mathbb{R}^{n_t\times n_s}$ from training to testing samples in the latent topic space. Each $\Theta_{i,j}$ represents the selection weight for training sample i to testing sample j. The higher the relationship between two samples, the higher weight should be given to $\Theta_{i,j}$.

Here, we use Pearson correlation to measure the correlation between each training and testing samples in a topic level as follows:

$$\Phi_{i,j} = \frac{\sum_{p=1}^{r}(\Pi_{p,i}^t - \overline{\pi_i^t})(\Pi_{p,j}^s - \overline{\pi_j^s})}{(\sqrt{\sum_{p=1}^{r}(\Pi_{p,i}^t - \overline{\pi_i^t})^2})(\sqrt{\sum_{p=1}^{r}(\Pi_{p,j}^s - \overline{\pi_j^s})^2})}, \tag{8}$$

where

$$\Pi^t = \Psi\Upsilon^t, \Pi^s = \Psi\Upsilon^s, \tag{9}$$

by considering inter-topic correlation in each sample. $\overline{\pi_i^t}$ and $\overline{\pi_j^s}$ denotes the mean value of π_i^t and π_j^s, respectively.

The value of Φ, in the range of $[-1, 1]$, can show the positive and negative relationship between two samples. Then, each $\Theta_{i,j}$ can be learned by considering the sign consistency between Θ and Φ. The closer to 0 the absolute value of $\Phi_{i,j}$ is, the more independent of training sample i and testing sample j is, the less contribution of training sample i when predicting sample j, the smaller absolute value of the corresponding $\Theta_{i,j}$ should be. Then, each $\Theta_{i,j}$ can also be learned by considering the sparsity regularization between Θ and Φ. After combining the above analysis, Θ can be solved by the following objective function

$$\min_{\Theta} \left\|\Pi^t\Theta - \Pi^s\right\|_F^2 + \sum_{i,j}^{n_t,n_s} (-\alpha\Theta_{i,j}\Phi_{i,j} + \beta(1 - |\Phi_{i,j}|)|\Theta_{i,j}|), \tag{10}$$

where α and β are two parameters to balance above three terms.

After obtaining the predictive mapping Θ, the label prediction based on the instance-specific sample extraction is

$$Y^s = Y^t\Theta. \tag{11}$$

2.7 Optimization

Update $\boldsymbol{F, L, \Upsilon^t}$: Objective functions in Eq.(1), Eq.(2) and Eq.(5) can be combined to form an inequality and non-negative constrained quadratic optimization problem, which can be solved by introducing Lagrangian multipliers. The combined objective function can be extended to

$$\begin{aligned}\min_{F,L} &\left\|X^t - F\Upsilon^t\right\|_F^2 + \left\|Y^t - L\Upsilon^t\right\|_F^2 + tr(Z^1F + Z^2L + Z^3\Upsilon^t) \\ &+ \sigma tr((E^1A + AE^1 - 2LL^T)C) + \sigma tr((E^2B + SE^2 - 2\Upsilon^tY^{tT})L) \\ &+ 2\lambda tr(E^1F + E^2L),\end{aligned} \tag{12}$$

where σ is a parameter to balance terms of label-topic relationship and topic-based label-label relationship. E^i is an all-one matrix and Z^i is a Lagrange multiplier for nonnegative constraint. $A = diag(LL^T)$, $B = diag(Y^tY^{tT})$ and $S = diag(\Upsilon^t\Upsilon^{tT})$, where $diag(A)$ indicates the diagonal entries in A. The KKT conditions of $Z_{p,q}^i H_{p,q} = 0$, where H is used to represent any variable from $\{F, L, \Upsilon^t\}$, to the derivative of the above function w.r.t. H can be applied to update one variable while fixing the other two variables, because the objective function is convex when any two variables are fixed. Specifically, each variable can be updated as follows:

$$F_{i,j} \leftarrow F_{i,j}\frac{(X^t\Upsilon^{tT})_{i,j}}{(F\Upsilon^t\Upsilon^{t^T} + \lambda)_{i,j}}, \tag{13}$$

$$L_{i,j} \leftarrow L_{i,j} \frac{((1+\sigma)(Y^t \Upsilon^{tT}) + \sigma CL)_{i,j}}{(L\Upsilon^t \Upsilon^{t^T} + \sigma(G + (E^2C)^T \odot L) + \lambda)_{i,j}}, \tag{14}$$

$$\Upsilon^t_{i,j} \leftarrow \Upsilon^t_{i,j} \frac{(F^T X^t + (1+\sigma) L^T Y^t)_{i,j}}{((F^T F + L^T L)\Upsilon^t + (1/2)\sigma(E^3 L)^T \odot \Upsilon^t)_{i,j}}, \tag{15}$$

where $G = 1/2(BE^{2T} + E^{2T}S)$ and $\odot$ indicates the Hadamart product.

Update $\boldsymbol{\Upsilon^s}$: Similarly, the objective function in Eq.(3) in the testing phase can be extended to

$$\min_{\Upsilon^s} \|X^s - F\Upsilon^s\|_F^2 + Z^4 \Upsilon^s, \tag{16}$$

where ϱ_i is a Lagrange multiplier for sum-one constraint. Then, Υ^s can be updated as follows:

$$\Upsilon^s_{i,j} \leftarrow \Upsilon^s_{i,j} \frac{(F^T X^s)_{i,j}}{(F^T F \Upsilon^s)_{i,j}}. \tag{17}$$

Update $\boldsymbol{\Xi}$: The problem (4) can be effectively solved by dividing into two subproblems [26] as follows:

$$\Xi = \arg\min_{\Xi} \|\Xi - \Xi_r^p\|_F^2, s.t.\ diag(\Xi) = 0, \Xi e = e, \Xi^T e = e, \tag{18}$$

and

$$\Xi = \arg\min_{\Xi} \|\Xi - \Xi_r^p\|_F^2, s.t.\ \Xi \geq 0, \tag{19}$$

where

$$\Xi_r^p = \frac{(\Xi^p - 1/\eta \nabla_\Xi \mathcal{L}(\Xi^p)) + (\Xi^p - 1/\eta \nabla_\Xi \mathcal{L}(\Xi^p))^T}{2}, \tag{20}$$

which is obtained from the problem (4) with its first order approximation at the previous point Ξ^p by considering the symmetric constraint $\Xi = \Xi^T$. The parameter η is the Lipschitz parameter, which is calculated according to $\nabla_\Xi \mathcal{L}(\Xi)$, is $r\sqrt{2(\sum_{i,j}^{r,r} \max^2((L^T L)_{:,i}(L^T L)_{j,:}) + \sum_{i,j}^{r,r} \max^2((L^T L)_{:,i}(\Upsilon^t \Upsilon^{tT})_{j,:}))}$.

By using the Lagrange multipliers for three constraints in Eq.(18), the first subproblem can be solved by

$$\Xi = \Xi_r^p - \frac{1}{r} tr(\Xi_r^p) I - \frac{r - e^T \Xi_r^p e + tr(\Xi_r^p)}{r-1} I + R + R^T, \tag{21}$$

where $R = (\frac{I}{r} + \frac{2-r}{2r^2(r-1)} ee^T)(e - \Xi_r^p e + \frac{tr(\Xi_r^p)e}{r})e^T$.

The second subproblem can be solved by

$$\Xi = \lceil \Xi_r^p \rceil_{\geq 0}, \tag{22}$$

where $\lceil \Xi_r^p \rceil_{\geq 0}$ let all negative values in Ξ_r^p change to 0.

Thus, we solve the problem (4) by successively alternating between above two subproblems.

Update Θ: The predictive mapping Θ, which can be easily solved by using proximal gradient descend method. First, the gradient w.r.t. Θ in the problem (10) without considering the sparsity term is

$$\nabla_{\Theta}\mathcal{L} = \Pi^{tT}\Pi^{t}\Theta - \Pi^{tT}\Pi^{s} - \alpha\Phi. \tag{23}$$

The sparsity term can be solved by applying element-wise soft-thresholding operator [10]. Then, Θ can be updated as follows:

$$\Theta_{t+1}(i,j) \leftarrow \text{prox}_{\frac{\beta(1-|\Phi_{i,j}|)}{L_f}}(\Theta^{t}(i,j) - \frac{1}{L_f}\nabla_{\Theta}\mathcal{L}(\Theta^{t})(i,j)), \tag{24}$$

where $\Theta^{t}(i,j) = \Theta_{t}(i,j) + \frac{b_{t-1}-1}{b_t}(\Theta_{t}(i,j) - \Theta_{t-1}(i,j))$. L_f is the Lipschitz parameter which is treated as the trace of the second differential of $F(\Theta)$ ($L_f = \left\|\Pi^{t}\Pi^{tT}\right\|_F$). The sequence b_t should satisfy the condition of $b_t^2 - b_t \leq b_{t-1}^2$. The prox is defined as

$$\text{prox}_{\varepsilon}(a) = \text{sign}(a)\max(|a| - \varepsilon, 0). \tag{25}$$

The procedure of the proposed TMLC is summarized in Algorithm 1. The update of Υ^t, L and F requires $O(n_t dr + n_t kr + r^2 d + r^2 k)$, $O(n_t kr + k^2 r + r^2 k)$ and $O(n_t dr)$ respectively. The update of Ξ requires $O(n_t kr + k^2 r + r^2 k + r^2 n_t)$. For each new instance, the update of Υ^s and Θ requires $O(r^2 d + r^2 + rd)$ and $O(n_t r^2 + n_t r)$ respectively. As the proposed method scales linearly with the number of instances, making it suitable for the large-scale datasets.

3 Relations to Previous Works and Discussions

Our work is related to sparse feature extraction methods, because Υ^t can be seen as new features in the proposed topic-aware framework and only several new features are extracted for each label. Feature extraction has been studied over decades, as its corresponding algorithms can be used to select the most informative features for enhancing the classification accuracy. Among these approaches, sparse learning strategies are widely used for their good performance. For example, some works focus on selecting features shared by all labels [20, 12, 7, 3] by imposing $l_{2,1}$-norm regularizer. Although these works can deliver favorable results, some researchers prefer to use $l_{2,0}$-norm regularizer [2, 21] to solve the original $l_{2,0}$-norm constrained feature selection problem. However, each label may be related to different features. Recently, certain works impose lasso to select label-specific features [19, 10, 9, 15]. The above works are different from the proposed label-specific feature extraction in the latent topic space algorithm, as they target on selecting a subset of original features for each label whereas we focus on selecting a subset of new features for each label where the correlation between the original features and labels is exploited in the new feature space.

Our work is also related to methods that use kNN technique between training and testing samples. The typical works are lazy kNN [27], CoE [23], LM-kNN [16] and SLEEC [1]. These works select highly correlated k training samples for

Algorithm 1 Topic-aware Multi-label Classification (TMLC)

Initialize: $f_i = 1/|\Omega_i| \sum_{x_m^t \in \Omega_i} x_m^t$, $l_j = 1/|\Omega_i| \sum_{y_m^t \in \Omega_i} y_m^t$, where $\{\Omega_1, ..., \Omega_r\}$ are r clusters generated from the training data [11], $|\Omega_i|$ is the number of samples in Ω_i, Υ^t and Υ^s as all one matrices, Θ as a random matrix.

Compute C with $C_{i,j} = \cos(Y_{i,:}^t, Y_{j,:}^t)$;
Repeat
update Υ^t according to Eq. (13);
update L according to Eq. (14);
update F according to Eq. (15);
Until Convergence;
Repeat
update Ξ according to Eq.(21) an Eq.(22);
Until Convergence;
Compute Ψ according to Eq.(7);
Repeat
update Υ^s according to Eq. (17);
Until Convergence;
compute Π^t and Π^s according to Eq.(9);
compute Φ according to Eq.(8);
Repeat
update Θ according to Eq. (24);
Until Convergence;
Predict $Y^s = L\Psi\Upsilon^s + Y^t\Theta$.

each testing sample in the original or projected spaces. The labels of testing samples are selected from the labels of these k training samples. The above works are different from the proposed instance-specific sample extraction in the latent topic space algorithm, because they only consider the positive relationship between training sets and testing sets while we consider the positive and negative relationships between these two sets. Specifically, the positively (negatively) related training samples are given positive (negative) weights when predicting the corresponding testing sample. We deem that the classification accuracy can be improved by combining the positively and negatively related training samples, as certain testing samples may share the similar positively related training samples but have different negatively related training samples.

4 Experiments

In this section, results of intensive experiments on real-world multi-labeled datasets are used to demonstrate the effectiveness of the proposed TMLC.

4.1 Datasets

In our study, we conduct experiments on various benchmarks from different domains, where features of image datasets and other datasets are obtained from

LEAR[3] and MULAN[4], respectively. Details of these datasets are listed in Table 1. All features of each dataset are normalized in the range of $[0, 1]$.

Table 1. Details of seven benchmarks

Dataset	# samples	# Features	# Labels	Domain
corel5k	4999	1000	260	image
pascal07	9963	1000	20	image
iaprtc12	19627	1000	291	image
espgame	20770	1000	268	image
mirflickr	25000	1000	457	image
yeast	2417	103	14	biology
cal500	502	68	174	music

Table 2. Performance in terms of Macro-F_1

Methods	Macro-F1				
	TMLC	CoE	LM-kNN	JFSC	LLSFDL
cal500	**0.240±0.009**	0.111±0.015	0.106±0.004	0.123±0.005	0.153±0.020
corel5k	**0.083±0.007**	0.075±0.016	0.069±0.003	0.044±0.008	0.023±0.000
yeast	**0.473±0.014**	0.391±0.002	0.372±0.014	0.440±0.005	0.434±0.005
iaprtc12	0.135±0.019	**0.150±0.003**	0.132±0.008	0.047±0.002	0.020±0.001
pascal07	**0.347±0.004**	0.323±0.007	0.289±0.006	0.254±0.008	0.215±0.003
espgame	0.086±0.012	**0.141±0.012**	0.137±0.005	0.055±0.014	0.024±0.001
mirflickr	**0.010±0.016**	0.005±0.020	0.002±0.001	0.002±0.019	0.001±0.000

4.2 Evaluation Metrics

In our experiments, three widely adopted F_1 measures including Macro-F_1, Micro-F_1 and Example-F_1 [28] are used to evaluate the multi-label classification performance.

4.3 Methods

We compared the following state-of-the-art related multi-label methods for classification in the experiments.

1. LM-kNN: it proposes a large margin distance metric learning with k nearest neighbors constraints for multi-label classification [16].

[3] https://lear.inrialpes.fr/people/guillaumin/data.php
[4] http://mulan.sourceforge.net/datasets-mlc.html

Table 3. Performance in terms of Micro-F_1

Methods	Micro-F1				
	TMLC	CoE	LM-kNN	JFSC	LLSFDL
cal500	**0.475±0.009**	0.412±0.007	0.369±0.002	0.472±0.010	0.468±0.017
corel5k	**0.312±0.013**	0.301±0.001	0.287±0.010	0.109±0.003	0.124±0.006
yeast	**0.651±0.007**	0.620±0.006	0.613±0.011	0.487±0.009	0.482±0.006
iaprtc12	**0.323±0.005**	0.309±0.009	0.296±0.015	0.112±0.003	0.096±0.023
pascal07	0.451±0.009	**0.454±0.006**	0.450±0.010	0.342±0.012	0.335±0.020
espgame	**0.218±0.025**	0.215±0.005	0.212±0.003	0.064±0.013	0.042±0.005
mirflickr	**0.022±0.008**	0.007±0.000	0.005±0.001	0.003±0.010	0.002±0.000

Table 4. Performance in terms of Example-F_1

Methods	Example-F1				
	TMLC	CoE	LM-kNN	JFSC	LLSFDL
cal500	**0.470±0.017**	0.410±0.010	0.361±0.004	0.468±0.009	0.464±0.017
corel5k	**0.296±0.012**	0.255±0.020	0.245±0.009	0.149±0.006	0.112±0.001
yeast	**0.638±0.002**	0.598±0.001	0.583±0.010	0.473±0.008	0.469±0.007
iaprtc12	**0.276±0.004**	0.257±0.010	0.234±0.013	0.093±0.003	0.065±0.005
pascal07	**0.413±0.003**	0.403±0.021	0.388±0.010	0.322±0.018	0.320±0.011
espgame	**0.185±0.018**	0.169±0.016	0.153±0.000	0.048±0.013	0.033±0.003
mirflickr	**0.011±0.011**	0.004±0.017	0.002±0.001	0.001±0.019	0.001±0.000

2. CoE: it conducts multi-label classification through the cross-view k nearest neighbor search among learned embeddings [23].
3. JFSC: it learns label-specific features and shared features for the discrimination of each label by exploiting two-order label correlations [10].
4. LLSFDL: it learns label-specific features and class-dependent labels for multi-label classification by mining high-order label correlations [9] .

The former two works are instance-specific sample extraction methods, while the latter two works are label-specific sample extraction methods.

4.4 Experimental Results

These are some parameters that need to be tuned for all the compared methods. In TMLC, the number of topics r is set to 50 and 100 for datasets with the number of instances smaller and larger than $15,000$ respectively. The parameter ι is set to 0.5 for equally weighting of a topic and other topics. The parameters σ and λ are selected from $\{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}\}$. The parameters α and β are selected from $\{10^{-3}, 10^{-2}, 10^{-1}, 1, 10^{1}, 10^{2}, 10^{3}\}$. These parameters are tuned by 5-cross validation on the training set. The parameters of other compared methods are tuned according to their corresponding papers.

In the experiment, each dataset is divided into 5 equal-sized subsets. In each run, one subset is used as the testing set and the remaining 4 sets are used as the training set. Each of these subsets is used in turn to be the testing set. Then,

there are total 5 runs. Table 2, Table 3 and Table 4 illustrate the average results (mean ± std) of different multi-label algorithms in terms of three different kinds of F_1 measures. Based on the experimental results, the following observations can be made.

1. In the two instance-specific sample extraction algorithms (LM-kNN and CoE), the latter method always delivers the better results. It may due to the fact that like the proposed TMLC, the latter method also exploits the feature-label correlation. CoE mines the feature-label correlation in a cross-view perspective while TMLC mines the feature-label correlation in a topic-view perspective.
2. In the two label-specific feature extraction algorithms (JFSC and LLSFDL), the former method always delivers the better results. It may due to the fact that like the proposed TMLC, the former method also exploits the sample-label correlation. JFSC mines the sample-label correlation based on a discriminant model while TMLC mines the sample-label correlation based on a topic model.
3. Compared with other methods, TMLC always gets better performances, because the proposed method has considered both the gap and correlation between inputs and outputs.

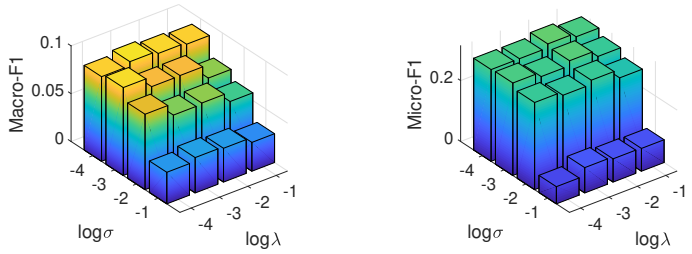

Fig. 3. The results with different σ and λ in the corel5k dataset based on (a) Macro-F_1; (b) Micro-F_1.

4.5 Parameter Analysis

The proposed label-specific feature extraction has two parameters including σ and λ. To study how these parameters affect the classification results, the performance variances with different values of these two parameters on the corel5k dataset are illustrated in Fig.3. During this process, the parameters α and β are set to their optimal values. Obviously, the performance is good when λ is larger compared with σ, it indicates that each topic is only related to several labels.

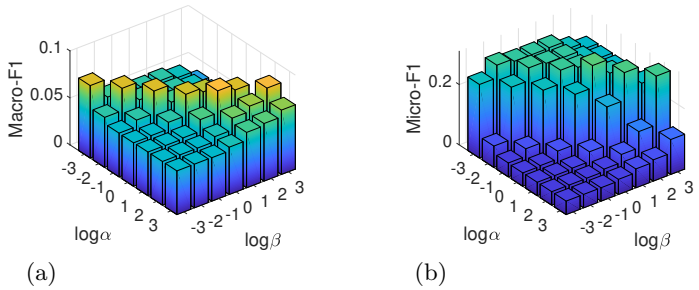

Fig. 4. The results with different α and β in the corel5k dataset based on (a) Macro-F_1; (b) Micro-F_1.

The proposed instance-specific sample extraction also has two parameters including α and β. To study how these parameters affect the classification results, the performance variances with different values of these two parameters on the corel5k dataset are illustrated in Fig.4. During this process, the parameters σ and λ are set to their optimal values. Clearly, the classification results are always favorable when α and β have equal values. It indicates that sign consistency and sparsity regularization are equally important for instance-specific sample extraction by exploiting inter-sample correlation in the latent topic space.

5 Conclusion

In this study, we propose a novel topic-aware multi-label classification framework where the informative training samples are extracted for each testing sample in the latent topic space and the informative features are also extracted for each label in the latent topic space. Compared with the existing instance-specific sample extraction methods, the proposed TMLC method has bridged the feature-label gap by aligning features and labels in a latent topic space. The mapping between training and testing inputs in the latent topic space can be directly applied to the corresponding outputs. It is worth noting that the input and output have physical meaning in the latent topic space, because they can both illustrate topic proportions of each sample in the latent space. Each of them can also illustrate its corresponding distribution for each topic. Different with the current label-specific feature extraction methods, the proposed TMLC method has considered the feature-label correlation by selecting features for each label in the latent topic space where the feature-label correlation is captured in this space. The intensive empirical studies on real-world benchmarks demonstrate the efficacy of the proposed framework.

References

1. Bhatia, K., Jain, H., Kar, P., Varma, M., Jain, P.: Sparse local embeddings for extreme multi-label classification. In: Advances in neural information processing systems. pp. 730–738 (2015)
2. Cai, X., Nie, F., Huang, H.: Exact top-k feature selection via l2, 0-norm constraint. In: Twenty-third international joint conference on artificial intelligence (2013)
3. Chang, X., Nie, F., Yang, Y., Huang, H.: A convex formulation for semi-supervised multi-label feature selection. In: Twenty-eighth AAAI conference on artificial intelligence (2014)
4. Duygulu, P., Barnard, K., de Freitas, J.F., Forsyth, D.A.: Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. In: European conference on computer vision. pp. 97–112. Springer (2002)
5. Elhamifar, E., Sapiro, G., Sastry, S.S.: Dissimilarity-based sparse subset selection. IEEE Transactions on Pattern Analysis and Machine Intelligence **38**(11), 2182–2197 (Nov 2016). https://doi.org/10.1109/TPAMI.2015.2511748
6. Gibaja, E., Ventura, S.: A tutorial on multilabel learning. ACM Computing Surveys (CSUR) **47**(3), 52 (2015)
7. Guo, Y., Xue, W.: Probabilistic multi-label classification with sparse feature learning. In: IJCAI. pp. 1373–1379 (2013)
8. Hoyer, P.O.: Modeling receptive fields with non-negative sparse coding. Neurocomputing **52**, 547–552 (2003)
9. Huang, J., Li, G., Huang, Q., Wu, X.: Learning label-specific features and class-dependent labels for multi-label classification. IEEE Transactions on Knowledge and Data Engineering **28**(12), 3309–3323 (2016)
10. Huang, J., Li, G., Huang, Q., Wu, X.: Joint feature selection and classification for multilabel learning. IEEE transactions on cybernetics (2017)
11. Huang, S.J., Zhou, Z.H.: Multi-label learning by exploiting label correlations locally. In: Twenty-sixth AAAI conference on artificial intelligence (2012)
12. Jian, L., Li, J., Shu, K., Liu, H.: Multi-label informed feature selection. In: IJCAI. pp. 1627–1633 (2016)
13. Lee, D.D., Seung, H.S.: Learning the parts of objects by non-negative matrix factorization. Nature **401**(6755), 788 (1999)
14. Liu, H., Li, X., Zhang, S.: Learning instance correlation functions for multilabel classification. IEEE transactions on cybernetics **47**(2), 499–510 (2017)
15. Liu, H., Zhang, S., Wu, X.: Mlslr: Multilabel learning via sparse logistic regression. Information Sciences **281**, 310–320 (2014)
16. Liu, W., Tsang, I.W.: Large margin metric learning for multi-label prediction. In: Twenty-Ninth AAAI Conference on Artificial Intelligence (2015)
17. Ma, J., Zhang, H., Chow, T.W.S.: Multilabel classification with label-specific features and classifiers: A coarse- and fine-tuned framework. IEEE Transactions on Cybernetics pp. 1–15 (2019). https://doi.org/10.1109/TCYB.2019.2932439
18. Ma, J., Chow, T.W.: Topic-based algorithm for multilabel learning with missing labels. IEEE transactions on neural networks and learning systems **30**(7), 2138–2152 (2018)
19. Ma, J., Chow, T.W., Zhang, H.: Semantic-gap-oriented feature selection and classifier construction in multilabel learning. IEEE Transactions on Cybernetics (2020)
20. Nie, F., Huang, H., Cai, X., Ding, C.H.: Efficient and robust feature selection via joint $l_{2,1}$-norms minimization. In: Advances in neural information processing systems. pp. 1813–1821 (2010)

21. Pang, T., Nie, F., Han, J., Li, X.: Efficient feature selection via $l_{\{}2, 0\}$-norm constrained sparse regression. IEEE Transactions on Knowledge and Data Engineering (2018)
22. Ren, J., Zhang, Z., Li, S., Wang, Y., Liu, G., Yan, S., Wang, M.: Learning hybrid representation by robust dictionary learning in factorized compressed space. IEEE Transactions on Image Processing **29**, 3941–3956 (2020)
23. Shen, X., Liu, W., Tsang, I.W., Sun, Q.S., Ong, Y.S.: Multilabel prediction via cross-view search. IEEE transactions on neural networks and learning systems **29**(9), 4324–4338 (2017)
24. Tsoumakas, G., Katakis, I.: Multi-label classification: An overview. International Journal of Data Warehousing and Mining **3**(3) (2006)
25. Wei, K., Iyer, R., Bilmes, J.: Submodularity in data subset selection and active learning. In: International Conference on Machine Learning. pp. 1954–1963 (2015)
26. Zhang, H., Sun, Y., Zhao, M., Chow, T.W., Wu, Q.J.: Bridging user interest to item content for recommender systems: An optimization model. IEEE transactions on cybernetics (2019)
27. Zhang, M.L., Zhou, Z.H.: Ml-knn: A lazy learning approach to multi-label learning. Pattern recognition **40**(7), 2038–2048 (2007)
28. Zhang, M.L., Zhou, Z.H.: A review on multi-label learning algorithms. IEEE transactions on knowledge and data engineering **26**(8), 1819–1837 (2014)
29. Zhang, R., Nie, F., Li, X.: Self-weighted supervised discriminative feature selection. IEEE transactions on neural networks and learning systems **29**(8), 3913–3918 (2018)
30. Zhang, Z., Jiang, W., Qin, J., Zhang, L., Li, F., Zhang, M., Yan, S.: Jointly learning structured analysis discriminative dictionary and analysis multiclass classifier. IEEE transactions on neural networks and learning systems **29**(8), 3798–3814 (2017)
31. Zhang, Z., Zhang, Y., Liu, G., Tang, J., Yan, S., Wang, M.: Joint label prediction based semi-supervised adaptive concept factorization for robust data representation. IEEE Transactions on Knowledge and Data Engineering **32**(5), 952–970 (2019)