Privileged Label Enhancement with Multi-Label Learning

Wenfang Zhu[1,2] , **Xiuyi Jia**[1,2,3*] , **Weiwei Li**[4]
[1]Key Laboratory of Information Perception and Systems for Public Security of MIIT, Nanjing University of Science and Technology, China
[2]Jiangsu Key Lab of Image and Video Understanding for Social Security, Nanjing University of Science and Technology, China
[3]State Key Laboratory for Novel Software Technology, Nanjing University, China
[4]College of Astronautics, Nanjing University of Aeronautics and Astronautics, China
{zwf.ang, jiaxy}@njust.edu.cn, liweiwei@nuaa.edu.cn

Abstract

Label distribution learning has attracted more and more attention in view of its more generalized ability to express the label ambiguity. However, it is much more expensive to obtain the label distribution information of the data rather than the logical labels. Thus, label enhancement is proposed to recover the label distributions from the logical labels. In this paper, we propose a novel label enhancement method by using privileged information. We first apply a multi-label learning model to implicitly capture the complex structural information between instances and generate the privileged information. Second, we adopt LUPI (learning with privileged information) paradigm to utilize the privileged information and employ RSVM+ as the prediction model. Finally, comparison experiments on 12 datasets demonstrate that our proposal can better fit the ground-truth label distributions.

1 Introduction

In recent years, the multi-label learning (MLL) framework has been studied for fitting multi-semantic problems successfully [Zhou and Zhang, 2017]. However, MLL can not express the relative importance of each label (i.e., label importance) to an instance, that is a more general label ambiguity problem. It is worth noting that label distribution learning (LDL) [Jia *et al.*, 2018; Ren *et al.*, 2019a; Ren *et al.*, 2019b; Jia *et al.*, 2019a] has been an active research area for the past few years due to its potential to address the label ambiguity problem as well as successful applications in many real-world tasks [Ling and Geng, 2019; Jia *et al.*, 2019b]. In LDL, an instance x is assigned a real number d_x^y to each possible label y, representing the degree to which y describes x. The LDL model has a stronger expression ability but also puts forward higher requirements for the presentation of training data, resulting in a high cost for labelling the datasets with distribution information. Fortunately, we have many multi-label datasets containing simple logical labels. To utilize existing mutli-label datasets, Xu et al. [2018] proposed a learning paradigm called label enhancement (LE), which aims to recover the hidden label distribution value from the logical labels of the datasets.

Label enhancement is to recover the label distributions from the logical labels in the training set via leveraging the topological information of the feature space and the correlation among the labels. One recent attempt is the GLLE [Xu *et al.*, 2018] algorithm which uses a common measurement method to construct a local similarity matrix based on the manifold hypothesis and adds it to the model training as prior knowledge. Another typical algorithm named ML [Hou *et al.*, 2016] assumes that each instance can be optimally reconstructed by using a linear combination of its neighbors, and according to the smoothness assumption, the topological structure of the feature space can be transferred to the label space local by local. In summary, how to mine and leverage additional information is one of the key problems that label enhancement concerns.

In this paper, we propose a privileged label enhancement method with multi-label learning (PLEML). Firstly, we apply a multi-label learning model to generate additional information for LE, which has two advantages: one is that the multi-label learning model itself can capture the mapping relationship between the feature space and the label space, and the predictions given on the instances imply the complex structure information of the data. Another one is that it is an implicit way to leverage additional information, which can avoid introducing noise by constructing structured information manually and lose the structured information of the data itself. Besides, in the multi-label learning procedure, we also adopt the low-rank structure to implicitly exploit the correlation of the labels. Secondly, inspired by [Vapnik and Vashist, 2009], we use LUPI (learning with privileged information) paradigm to make reasonable use of additional information. LUPI focuses on improving the learning with the auxiliary information which is supplied by a teacher about instances at

*Corresponding author: Xiuyi Jia. This work is jointly supported by the National Key R&D Program of China (2018YFB1003902), the National Natural Science Foundation of China (61773208, 61906090), the Natural Science Foundation of Jiangsu Province (BK20170809, BK20191287), the Fundamental Research Funds for the Central Universities (30920021131), and the China Postdoctoral Science Foundation (2018M632304).

the training stage. Since this auxiliary information will not be available at the test stage, it is referred to as privileged information. Finally, we apply the RSVM+ model as the final prediction model, which is a support vector machine discriminative model implementing the LUPI paradigm. We also compare our proposed method with existing LE methods on 12 real-world datasets. In summary, the major contributions of this paper are:

1) We propose a novel privileged label enhancement method with multi-label learning, named PLEML, which can fully and accurately utilize the additional information to recover label distributions from logical labels.

2) We apply the multi-label learning model to generate the privileged information, which can implicitly capture the complex structural information of the data itself.

3) We introduce the LUPI paradigm to make reasonable use of the generated privileged information. To the best of our knowledge, this is the first try of applying LUPI into the field of LE.

The rest of this paper is organized as follows. Firstly, related works are briefly reviewed. Secondly, the technical details of the proposed approach is presented. Thirdly, the comparative experimental results on different tasks are reported. Finally, we conclude this paper.

2 Related Work

The existing LE algorithms are mainly divided into two categories. The first category is fuzzy-based label enhancement, which utilizes the idea of fuzzy mathematics, and uses methods such as fuzzy clustering, fuzzy operation, and nuclear membership to dig out relevant information between labels and convert logical labels into label distributions. For example, the fuzzy clustering-based LE algorithm FCM [Gayar *et al.*, 2006] and kernel-based LE algorithm KM [Jiang *et al.*, 2006]. Another category is graph-based label enhancement, which uses graph models to represent topological structures between instances, and enhances logical labels of an instance to its corresponding label distribution by establishing the relationship between instance correlations and label correlations. Typical graph-based label enhancement methods include label propagation-based LE algorithm (LP) [Li *et al.*, 2015], manifold learning-based LE algorithm (ML) [Hou *et al.*, 2016], and graph laplacian-based label enhancement algorithm (GLLE) [Xu *et al.*, 2018].

Learning Using Privileged Information (LUPI) was first introduced by Vapnik and Vashist [2009], which assumes a teacher-student learning scenario, where a teacher can provide descriptive information (privileged information) about a course (primary data) to assist a student (model) to learn through the guidance of similarity control and knowledge transfer. And the privileged information stands for the information which is only available in the training stage but not available in the testing stage. Besides classification [Yao *et al.*, 2019], privileged information has also been used for clustering [Marcacini *et al.*, 2014], hashing [Zhou *et al.*, 2016], and etc. This paper is the first work to use multi-label learning to generate privileged information for label enhancement.

3 Proposed Method

3.1 Formulation of Label Enhancement

The main notations used in this paper are listed as follows. Let $\mathcal{X} = \{x_1, x_2, \ldots, x_n\} \in \mathcal{R}^{n\times d}$ denote the feature space, where n denotes the number of instances, d denotes the dimension of the feature. $\mathcal{Y} = \{0, 1\}^c$ represents the complete set of logical labels where c is the number of all possible labels. For each instance $x_i \in \mathcal{X}$, the logical label vector of x_i is denoted by $y_i = \{y_{i1}, y_{i2}, \ldots, y_{ic}\}$, each element $y_{ik} = 1$ if the label y_k is related to x_i, otherwise $y_{ik} = 0$. The description degree of y to x is denoted by d_x^y, and the ground-truth label distribution of x_i is denoted by $d_i = \{d_{x_i}^{y_1}, d_{x_i}^{y_2}, \ldots, d_{x_i}^{y_c}\}$.

Given a training set $\mathcal{M} = \{(x_i, y_i)|1 \le i \le n\}$, LE recovers the label distribution $\hat{d}_i$ of x_i from the logical label vector y_i, converting $\mathcal{M}$ to label distribution data $\hat{\mathcal{D}} = \{(x_i, d_i)|1 \le i \le n\}$, satisfying $\hat{d}_{x_i}^{y_k} \in [0, 1]$ and $\sum_{k=1}^{c} \hat{d}_{x_i}^{y_k} = 1$, making the prediction label distribution $\bar{\mathcal{D}}$ as close as possible to the the true label distribution $\mathcal{D}$.

3.2 Privileged Information Learning

With the previous discussion, our goal is to get the prediction value $Y^* = [y_1^*, y_2^*, \cdots, y_n^*]$ by utilizing the complex structure information between instances implicitly through multi-label learning, and add Y^* to the training of the label enhancement model as privileged information. Therefore, this privileged information should have the following characteristics: 1) Y^* is the instance's prediction value, which is not inherited from its original logical labels; 2) The generation process of Y^* utilizes structural information transferred from the instance feature space implicitly; 3) Y^* has a low-rank property since the labels are related.

To solve this problem, we consider using a linear model for prediction.

$$y_{ik}^* = x_i \bar{W}_k, \tag{1}$$

here, we add an additional dimension with a constant value of 1 for each data $x_i (1 \le i \le n)$, so $x_i = [x_{i1}, x_{i2}, \cdots, x_{id}, 1]$. The offset term $\bar{b}_k$ has been expanded into $\bar{W}_k$, and $\bar{W}_k = [\bar{W}_{1k}, \bar{W}_{2k}, \cdots, \bar{W}_{dk}, \bar{b}_k]^T$ represents the weight parameters of the linear model corresponding to the k-th label.

Accordingly, the goal of our method is to determine the best parameter $\bar{W}$ that can generate the privileged information Y^* given the instance x_i. Thus our goal becomes to find the optimal model $\bar{W}$ which minimizes:

$$\bar{W} = \arg\min_{\bar{W}} L(\bar{W}) + \lambda_1 \Omega(\bar{W}) + \lambda_2 Z(\bar{W}), \tag{2}$$

where L is the loss function defined on the training data, Ω is a regularizer to control the complexity of the output model, Z is a regularizer to enforce the characteristic of label correlations, and λ_1 and λ_2 are two parameters to balance the three terms. Here, for easy computation, we use the square of the Euclidean distance as the loss function defined by:

$$L(\bar{W}) = \frac{1}{2}\|Y^* - Y\|_F^2, \tag{3}$$

where $\|\cdot\|_F^2$ denotes the Frobenius norm of a matrix, Y^* and Y denote the predicted value and the logical value of the

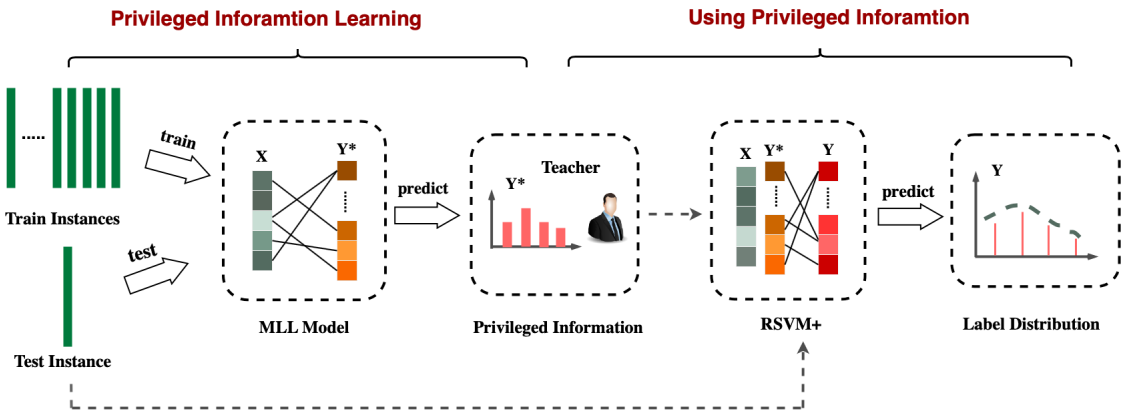

Figure 1: The framework of PLEML. The multi-label learning model captures the injective relationship between the feature space and the label space. When predicting for the test instance, the prediction result implicitly uses the structural relationship between the instances, so the prediction results Y^* can be used as a kind of privileged information, such as an *Oracle teacher* providing additional information in the process of LE. Finally, we use the RSVM+ model for label enhancement, and use feature information and privileged information to obtain the final label distribution value.

training set, respectively. For the second term of Eq. (2), we simply implement it as follows:

$$\Omega(\bar{W}) = \|\bar{W}\|_F^2. \tag{4}$$

The third term of Eq. (2) is employed to enforce the low-rank structure of the predicted label value, which implicitly exploits the label correlations. However, the rank of a matrix is difficult to optimize, therefore, the trace norm $\|\cdot\|_{tr}$ is utilized in this paper as a convex approximation of the rank of a matrix. The trace norm $\|\cdot\|_{tr}$ is defined as the sum of singular values, i.e., $\|\cdot\|_{tr} = \sum_i \sigma_i(\cdot)$, where σ_i is the i-th singular value of the matrix. Thus, the final term of Eq. (2) based on low-rank label correlations is derived as follows:

$$Z(\bar{W}) = \|Y^*\|_{tr}. \tag{5}$$

Formulating the multi-label problem into an optimization framework over Eq. (3), Eq. (4) and Eq. (5), the following optimization problem is obtained:

$$\min_{\bar{W}} \frac{1}{2}\|Y^* - Y\|_F^2 + \lambda_1\|\bar{W}\|_F^2 + \lambda_2\|Y^*\|_{tr}. \tag{6}$$

If the best parameter $\bar{W}$ is determined, the prediction value of the test instance y_i^* can be generated through Eq. (1). For simplicity, we divide the datasets into 4/5 training set and 1/5 testing set. Therefore, after 5-fold cross-validation, we can get the prediction value of each instance.

Eq. (6) can be optimized by using ADMM [Boyd *et al.*, 2011], thus, we first rewrite our objective into the following equivalent form:

$$\begin{aligned} &\min_{\bar{W},Z} \frac{1}{2}\|Y^* - Y\|_F^2 + \lambda_1\|\bar{W}\|_F^2 + \lambda_2\|Z\|_{tr} \\ &s.t. \quad Y^* - Z = 0. \end{aligned} \tag{7}$$

The augmented Lagrangian function of Eq. (7) is:

$$\begin{aligned} &\min_{\bar{W},Z,\Lambda} \frac{1}{2}\|Y^* - Y\|_F^2 + \lambda_1\|\bar{W}\|_F^2 + \lambda_2\|Z\|_{tr} \\ &+ < \Lambda, Y^* - Z > + \frac{\rho}{2}\|Y^* - Z\|_F^2, \end{aligned} \tag{8}$$

where Λ is the Lagrange multiplier, ρ is the penalty parameter, and $< \cdot, \cdot >$ is the Frobenius dot-product. The optimization problem of Eq. (8) can be solved using the alternating solution method.

$$\begin{aligned} \bar{W}^{t+1} = &\arg\min_{\bar{W}} \frac{1}{2}\|Y^* - Y\|_F^2 + \lambda_1\|\bar{W}\|_F^2 + \lambda_2\|Z^t\|_{tr} \\ &+ < \Lambda^t, Y^* - Z^t > + \frac{\rho}{2}\|Y^* - Z^t\|_F^2, \end{aligned} \tag{9}$$

$$\begin{aligned} Z^{t+1} = &\arg\min_{Z} \lambda_2\|Z\|_{tr} + < \Lambda^t, Y^{*t+1} - Z > \\ &+ \frac{\rho}{2}\|Y^{*t+1} - Z\|_F^2, \end{aligned} \tag{10}$$

$$\Lambda^{t+1} = \Lambda^t + \rho(\|Y^{*t+1} - Z^{t+1}\|_F^2). \tag{11}$$

L-BFGS [Yuan, 1991] can be applied to optimize Eq. (9), and the closed-form solution of Eq. (10) can be solved by the singular value thresholding (SVT) algorithm.

3.3 Label Enhancement using Privileged Information

Following the first stage of privileged information learning, the original datasets can be transformed into its essential counterpart: $\widetilde{\mathcal{D}} = \{(x_i, y_i^*, y_i)|1 \le i \le n\}$, where x_i represents available information, y_i^* represents privileged information, and y_i represents logical label values, n represents the number of instances. Here we use the LUPI (learning with privileged information) paradigm to deal with the problem of label enhancement with privileged information. This paradigm focuses on using privileged information about the instances provided by the teacher during the training phase to improve learning. In [Vapnik and Vashist, 2009], RSVM+ model based on the LUPI paradigm for regression is introduced. The main idea of RSVM+ is to define a linear or nonlinear correction function in the privileged space, and use this

Algorithm 1 The PLEML algorithm

Input: Feature matrix $\mathcal{X}$; logical label matrix $\mathcal{Y}$.
Parameter: $\lambda_1, \lambda_2, \gamma, C$.
Output: The regression parameters matrix W.

1: Initialize Λ, Z, ρ and $\bar{W}$;
2: solve $\bar{W}$ in Eq. (8) using ADMM method;
3: obtain Y^* by the Eq. (1).
4: Initialize $W, W^*, \bar{W}^*$;
5: solve W in Eq. (12) using penalty function method;
6: **return** W.

correction function to estimate slack variables in support vector regression methods. Therefore, by converting it to the LUPI paradigm based on RSVM+, we develop a novel label enhancement model. The decision rule and correcting (slack) function are defined as linear functions $\hat{d}_{x_i}^{y_k} = x_i W_k + b_k$ and $\xi_i = y_i^* W_k^* + b_k^*$, $\hat{\xi}_i = y_i^* \hat{W}_k^* + \hat{b}_k^*$, respectively. So the objective function of PLEML can be expressed as:

$$\begin{aligned} \min_{\theta} \quad & \frac{1}{2}\sum_{k=1}^{c}(\|W_k\|_2^2 + \gamma(\|W_k^*\|_2^2 + \|\hat{W}_k^*\|_2^2)) + \\ & + C\sum_{k=1}^{c}\sum_{i=1}^{n}(y_i^* W_k^*) + C\sum_{k=1}^{c}\sum_{i=1}^{n}(y_i^* \hat{W}_k^*) \\ s.t. \quad & x_i W_k - y_{ik} \le \varepsilon + y_i^* W_k^* \\ & y_{ik} - x_i W_k \le \varepsilon + y_i^* \hat{W}_k^* \\ & y_i^* W_k^* \ge 0 \\ & y_i^* \hat{W}_k^* \ge 0, \end{aligned} \tag{12}$$

where $\theta = \{W_k, W_k^*, \hat{W}_k^*\}$ are the parameters to be optimized, and $\{\gamma, C\}$ are the weighted coefficients. Particularly, we absorb the bias term to obtain a compact variant of the original RSVM+, because it is turned out to have a simpler form in the dual space and can be solved more efficiently. Specifically, for each x_i and y_i^*, an additional dimension with a constant value of 1 is added, so the $W_k = \{W_{1k}, W_{2k}, \cdots, W_{dk}, b_k\}$, $W_k^* = \{W_{1k}^*, W_{2k}^*, \cdots, W_{ck}^*, b_k^*\}$ and $\hat{W}_k^* = \{\hat{W}_{1k}^*, \hat{W}_{2k}^*, \cdots, \hat{W}_{ck}^*, \hat{b}_k^*\}$.

Eq. (12) is an optimization problem with inequality constraints. Therefore, we can choose the penalty function method to solve it [Sun and Yuan, 2006]. The basic idea of the penalty function method is to use a penalty function to convert a constraint problem into an unconstrained problem, and then use the unconstrained optimal method to solve it. The framework of our PLEML is shown in Figure 1 and the detailed algorithm of the proposed method is shown in Algorithm 1.

4 Experiments

4.1 Datasets

There are 12 real-world label distribution datasets in our experiments, including two facial expression datasets SJAFFE [Lyons *et al.*, 1998] and SBU_3DFE [Yin *et al.*, 2006], ten biological experiments datasets Yeast [Eisen *et al.*, 1998]. Some basic statistics about these 12 datasets are given in Table 1.

No.	Datasets	#Instances	#Features	#Labels
1	SJAFFE	213	243	6
2	SBU_3DFE	2500	243	6
3	Yeast-spoem	2465	24	2
4	Yeast-spo5	2465	24	3
5	Yeast-dtt	2465	24	4
6	Yeast-cold	2645	24	4
7	Yeast-heat	2465	24	6
8	Yeast-spo	2465	24	6
9	Yeast-diau	2465	24	7
10	Yeast-elu	2465	24	14
11	Yeast-cdc	2465	24	15
12	Yeast-alpha	2465	24	18

Table 1: Statistics of the 12 datasets.

4.2 Evaluation Measures

To measure the distance or similairy between the recovered label distributions and the ground-truth label distributions, according to Geng's suggestion [Geng, 2016], six LDL measures are adopted, i.e., Chebyshev distance (Cheb), Clark distance(Clark), Canberra metric (Canber), Kullback-Leibler divergence (KL), Cosine coefficient (Cosine) and Intersection similarity (Intersec). The former four are distance measures and the last two are similarity measures. For Cheb, Clark, Canberra and KL, the smaller the value, the better the generalization performance. For Cosine and Intersec, the larger the value, the better the performance.

4.3 Methodology

We implemented two groups of experiments. In the first group, we use the same binarization method [Xu *et al.*, 2018] to generate the logical labels from the ground-truth label distributions for the regular-scale datasets. Then, we recover the label distributions from the logical labels via the LE algorithms. Finally, we compare the recovered label distributions with the ground-truth label distributions. In the second group, in order to further test the effectiveness of LDL after the LE pre-process on the logical-labeled datasets, we first recover the label distributions from the logical labels via the LE algorithms and then use the recovered label distributions for LDL training. We choose LDL-SCL [Zheng *et al.*, 2018] as the LDL algorithm in this experiment. Finally, the label distributions predicted by LDL-SCL on the recovered data are compared with those predictions made on the original data with ground-truth label distributions.

4.4 Experimental Setting

The performance of PLEML is compared against five label enhancement learning algorithms, including FCM [Gayar *et al.*, 2006], KM [Jiang *et al.*, 2006], LP [Li *et al.*, 2015], ML [Hou *et al.*, 2016], and GLLE [Xu *et al.*, 2018].

With the previous discussion, FCM, KM and LP are fuzzy-based label enhancement, and ML, GLLE are graph-based label enhancement. For the comparison algorithms, parameter configurations suggested in corresponding literatures are used, for GLLE, the parameter Λ is chosen among

data	algorithm	SJAFFE	SBU_3DEF	Yeast-spoem	Yeast-spo5	Yeast-dtt	Yeast-cold	Yeast-heat	Yeast-spo	Yeast-diau	Yeast-elu	Yeast-cdc	Yeast-alpha
Cheb ↓	PLEML	0.0972	**0.1209**	**0.0891**	**0.0921**	**0.0373**	**0.0540**	**0.0435**	0.0603	**0.0416**	**0.0165**	**0.0166**	**0.0137**
	GLLE	**0.0906**	0.1315	0.0891	0.1009	0.0407	0.0583	0.0449	**0.0590**	0.0503	0.0191	0.0185	0.0168
	FCM	0.1341	0.1352	0.2354	0.1639	0.0942	0.1403	0.1566	0.1252	0.0865	0.0531	0.0309	0.0353
	KM	0.2170	0.2368	0.4081	0.2764	0.2490	0.2470	0.1705	0.1747	0.1536	0.0758	0.0758	0.0616
	LP	0.1071	0.1228	0.1633	0.1146	0.1286	0.1371	0.0863	0.0902	0.0989	0.0437	0.0415	0.0400
	ML	0.2120	0.2341	0.4053	0.2746	0.2448	0.2432	0.1654	0.1715	0.1483	0.0721	0.0713	0.0568
Clark ↓	PLEML	0.4312	**0.3640**	**0.1318**	**0.1855**	**0.1014**	**0.1465**	**0.1880**	0.2559	**0.2224**	**0.2043**	**0.2191**	**0.2124**
	GLLE	**0.3081**	0.3933	0.1321	0.1991	0.1112	0.1552	0.1949	**0.2543**	0.2816	0.2381	0.2540	0.2808
	FCM	0.5121	0.4246	0.3812	0.3559	0.3053	0.4092	0.5097	0.4429	0.6660	0.6177	0.6342	0.8742
	KM	1.8740	1.9062	1.0283	1.0590	1.4763	1.4714	1.8021	1.8110	1.8856	2.7673	2.8849	3.1521
	LP	0.5050	0.5810	0.2718	0.2741	0.1286	0.2864	0.5683	0.5585	0.7879	0.9735	1.0143	1.1864
	ML	1.8444	1.8761	1.0150	1.0469	1.4603	1.4546	1.7820	1.7882	1.8636	2.7377	2.8531	3.1175
Canber ↓	PLEML	0.8937	**0.7801**	**0.1837**	**0.2849**	**0.1741**	**0.2528**	**0.3741**	0.5284	**0.4778**	**0.6014**	**0.6546**	**0.6879**
	GLLE	**0.6267**	0.8409	0.1840	0.3100	0.1919	0.2679	0.3920	**0.5269**	0.6411	0.7171	0.7818	0.9202
	FCM	1.0706	0.9051	0.5169	0.5312	0.5226	0.7010	1.0603	0.9043	1.4674	1.9762	2.2488	2.6927
	KM	4.0083	4.1209	1.2529	1.3820	2.5961	2.5674	3.8514	3.8548	4.2576	9.1129	9.8760	11.8116
	LP	1.0708	1.2463	0.3655	0.4013	0.9434	0.5297	1.2939	1.2341	1.7490	3.3835	3.6460	4.5494
	ML	0.7839	4.0593	1.2382	1.3680	2.5714	2.5415	3.8141	3.8116	4.2192	9.0292	9.7836	11.7027
KL ↓	PLEML	0.0658	**0.0644**	**0.0272**	**0.0299**	**0.0066**	**0.0135**	**0.0134**	0.0271	**0.0159**	**0.0065**	**0.0073**	**0.0056**
	GLLE	**0.0382**	0.0730	0.0280	0.0345	0.0079	0.0148	0.0146	**0.0269**	0.0242	0.0088	0.0095	0.0090
	FCM	0.1060	0.0811	0.2089	0.1169	0.0571	0.1003	0.1345	0.0980	0.1608	0.0773	0.0989	0.1128
	KM	0.5613	0.6034	0.5318	0.3350	0.6166	0.5861	0.5865	0.5636	0.5399	0.6201	0.6341	0.6340
	LP	0.0776	0.1054	0.0672	0.0427	0.1041	0.1035	0.0891	0.0846	0.1274	0.1094	0.1115	0.0492
	ML	0.5398	0.5820	0.5148	0.3245	0.5986	0.5682	0.5689	0.5452	0.5211	0.6017	0.6146	0.6145
Cosine ↑	PLEML	0.9477	**0.9361**	0.9768	**0.9736**	**0.9937**	**0.9873**	**0.9872**	0.9747	**0.9853**	**0.9938**	**0.9930**	**0.9945**
	GLLE	**0.9613**	0.9212	**0.9777**	0.9698	0.9926	0.9859	0.9863	**0.9758**	0.9775	0.9914	0.9912	0.9912
	FCM	0.9022	0.9148	0.8815	0.9193	0.9569	0.9234	0.8893	0.9132	0.9225	0.9329	0.9410	0.9507
	KM	0.8261	0.8130	0.8123	0.8816	0.7600	0.7803	0.7801	0.7995	0.7982	0.7590	0.7543	0.7518
	LP	0.9410	0.9220	0.9503	0.9686	0.7855	0.9660	0.9323	0.9386	0.9146	0.9176	0.9158	0.9108
	ML	0.8282	0.8139	0.8138	0.8828	0.7619	0.7822	0.7819	0.8015	0.8018	0.7613	0.7570	0.7546
Intersec ↑	PLEML	0.8581	**0.8587**	**0.9109**	**0.9079**	**0.9570**	**0.9736**	**0.9385**	0.9129	**0.9334**	**0.9576**	**0.9569**	**0.9620**
	GLLE	**0.8926**	0.8472	**0.9109**	0.8991	0.9527	0.9332	0.9356	**0.9133**	0.9099	0.9489	0.9484	0.9496
	FCM	0.8161	0.8361	0.7647	0.8361	0.8802	0.8355	0.8185	0.8437	0.8223	0.8590	0.8600	0.8752
	KM	0.5927	0.5793	0.5919	0.7236	0.5411	0.5595	0.5591	0.5749	0.5883	0.5395	0.5327	0.5323
	LP	0.8361	0.8096	0.8367	0.8855	0.9210	0.8756	0.8048	0.8184	0.7875	0.7814	0.7791	0.7733
	ML	0.5961	0.5832	0.5947	0.7254	0.5437	0.5622	0.5617	0.5776	0.5902	0.5421	0.5534	0.5350

Table 2: Comparison results of label enhancement methods on real-world datasets. The best performance on each measure is marked in bold.

data	Ground-truth	PLEML	GLLE	FCM	KM	LP	ML
SJAFFE	0.0968 ±0.0069	**0.1205 ±0.0063**	0.1229 ±0.0084	0.1229 ±0.0081	0.1225±0.0084	0.1189±0.0083	0.1193±0.0081
SBU_3DFE	0.1235 ±0.0016	**0.1291 ±0.0016**	0.1389 ±0.0016	0.1382 ±0.0016	0.1375±0.0015	0.1376±0.0016	0.1374 ±0.0016
Yeast-spoem	0.0886 ±0.0036	**0.0903 ±0.0036**	0.0908 ±0.0035	0.0916 ±0.0045	0.0990 ±0.0015	0.0912 ±0.0035	0.0904 ±0.0040
Yeast-spo5	0.0921 ±0.0019	**0.0930 ±0.0022**	0.1014 ±0.0025	0.1229 ±0.0021	0.1107 ±0.0009	0.0963 ±0.0024	0.1010 ±0.0028
Yeast-dtt	0.0361 ±0.0009	**0.0373 ±0.0006**	0.0405 ±0.0005	0.0641 ±0.0014	0.0790 ±0.0006	0.0416 ±0.0016	0.0566 ±0.0034
Yeast-cold	0.0510±0.0009	**0.0541 ±0.0011**	0.0582±0.0009	0.0846 ±0.0007	0.0835±0.00012	0.0897 ±0.0009	0.0941±0.0042
Yeast-heat	0.0421±0.0006	**0.0435 ±0.0006**	0.0436±0.0007	0.0553 ±0.0017	0.0689±0.0004	0.0491 ±0.0007	0.0511±0.0025
Yeast-spo	0.0585±0.0010	0.0603 ±0.0013	**0.0590±0.0010**	0.0867 ±0.0015	0.0684±0.0013	0.0665 ±0.0012	0.0692±0.0032
Yeast-diau	0.0371±0.0006	**0.0416 ±0.0006**	0.0502±0.0004	0.0494 ±0.0006	0.0689±0.0004	0.0712 ±0.0004	0.0863±0.0012
Yeast-elu	0.0164±0.0001	**0.0166 ±0.0002**	0.0191±0.0001	0.0262 ±0.0003	0.0249±0.0001	0.0266 ±0.0002	0.0283±0.0004
Yeast-cdc	0.0161±0.0003	**0.0165 ±0.0003**	0.0182±0.0002	0.0248 ±0.0001	0.0249±0.0001	0.0252 ±0.0001	0.0267±0.0006
Yeast-aplha	0.0135±0.0002	**0.0137 ±0.0002**	0.0168±0.0002	0.0248 ±0.0001	0.0249±0.0001	0.0252 ±0.0001	0.0308±0.0007

Table 3: Comparison of the LDL after the LE pre-process against the direct LDL measured by Cheb ↓. The best results on each row are highlighted in boldface.

data	Ground-truth	PLEML	GLLE	FCM	KM	LP	ML
SJAFFE	0.8514 ±0.0070	0.8369 ±0.0063	**0.8442 ±0.0010**	0.8256 ±0.0069	0.8283±0.0076	0.8332±0.0073	0.8327±0.0070
SBU_3DFE	0.8541 ±0.0011	**0.8508 ±0.0016**	0.8442 ±0.0010	0.8406 ±0.0012	0.8483±0.0017	0.8409±0.0014	0.8472 ±0.0017
Yeast-spoem	0.9114 ±0.0036	**0.9097 ±0.0036**	0.9092 ±0.0035	0.9085 ±0.0045	0.9046 ±0.0010	0.9019 ±0.0036	0.9090 ±0.0008
Yeast-spo5	0.9079 ±0.0019	**0.9070 ±0.0022**	0.8986 ±0.0025	0.8771 ±0.0021	0.8893 ±0.0009	0.9037 ±0.0025	0.8990 ±0.0028
Yeast-dtt	0.9583 ±0.0009	**0.9570 ±0.0006**	0.9528 ±0.0012	0.9232 ±0.0017	0.8668 ±0.0006	0.9512 ±0.0019	0.9329 ±0.0010
Yeast-cold	0.9410±0.0009	**0.9378 ±0.0011**	0.9333±0.0009	0.8889 ±0.0039	0.8704±0.0007	0.9326 ±0.0014	0.8975±0.0038
Yeast-heat	0.9405±0.0006	**0.9387 ±0.0006**	0.9384±0.0007	0.8744 ±0.0026	0.8650±0.0008	0.9378 ±0.0008	0.9287±0.0029
Yeast-spo	0.9153±0.0010	**0.9131 ±0.0013**	0.9129±0.0010	0.8787 ±0.0024	0.9127±0.0013	0.9120 ±0.0015	0.8918±0.0053
Yeast-diau	0.9404±0.0006	**0.9335 ±0.0006**	0.9098±0.0004	0.8811 ±0.0013	0.8898±0.0005	0.9156 ±0.0012	0.8371±0.0027
Yeast-elu	0.9587±0.0001	**0.9575 ±0.0002**	0.9491±0.0001	0.9084 ±0.0019	0.8670±0.0006	0.9479 ±0.0009	0.9198±0.0017
Yeast-cdc	0.9579±0.0003	**0.9575 ±0.0003**	0.9491±0.0002	0.8837 ±0.0026	0.8656±0.0003	0.9486 ±0.0014	0.9180±0.0027
Yeast-aplha	0.9624±0.0002	**0.9620 ±0.0002**	0.9500±0.0002	0.8794 ±0.0019	0.8666±0.0002	0.9127 ±0.0014	0.9124±0.0030

Table 4: Comparison of the LDL after the LE pre-process against the direct LDL measured by Cosine ↑. The best results on each row are highlighted in boldface.

data	Measurements	PLEML	random PI
Yeast-cdc	Cheb ↓	**0.0166**	0.0172
	Canber ↓	**0.6546**	0.6803
	Cosine ↑	**0.9930**	0.9925
Yeast-diau	Cheb ↓	**0.0416**	0.0455
	Canber ↓	**0.4778**	0.5317
	Cosine ↑	**0.9853**	0.9824
SJAFFE	Cheb ↓	**0.0972**	0.1001
	Canber ↓	0.8937	**0.8598**
	Cosine ↓	**0.9477**	0.9456
SBU_3DFE	Cheb ↓	**0.1209**	0.1364
	Canber ↓	**0.7801**	0.8772
	Cosine ↑	**0.9361**	0.9203

Table 5: The comparison of PLEML and random privileged information.

$\{10^{-2}, 10^{-1}, \cdots, 10^{2}\}$ and the number of neighbors k is set to c+1. The kernel function in GLLE is Gaussian kernel. The parameter α in LP is set to 0.5. The number of neighbors k for ML is set to c+1. The parameter β in FCM is set to 2. The kernel function in KM is Gaussian kernel. For PLEML, the values of the parameters λ_1 and λ_2 are selected among $\{2^{-4}, 2^{-3}, \cdots, 2^{8}\}$, and $\gamma = 0.1, C = 0.1$. For LDL-SCL algorithm, the parameters are set to $\lambda_1 = 0.001$, $\lambda_2 = 0.001$, and m = 5.

It is worth mentioning that, for the first group of experiments, it is not necessary to divide the dataset into training set and testing set because LE is a kind of unsupervised learning procedure. Therefore, we can directly apply the privilege information to the whole data. For the second group of experiments, 10 times 10 fold cross-validation is employed for each dataset and the average results are recorded.

4.5 Experimental Results

The experimental result of the first group is summarized in Table 2. For the second group of experiments, due to page limitation and refer to [Xu *et al.*, 2018], we only show the effect on the two evaluation measures of Chebyshev and Cosine in Table 3 and Table 4, respectively. The results of other measures are similar.

From Table 2, we can see that PLEML significantly outperforms FCM, ML, LP, KM on the most measures. Compared with the GLLE algorithm, PLEML performs slightly worse on the SJAFFE and Yeast-spo datasets on all the measures, and performed slightly worse on the Yeast-spoem dataset in terms of Cosine. However, in other cases, the PLEML algorithm is better than the GLLE algorithm. From Table 3 and Table 4, we can find that the PLEML algorithm significantly outperforms other comparison algorithms in terms of Cheb and Cosine. Therefore, we can obtain the following three conclusions: 1) The specialized LE algorithms generally perform better than those algorithms obtained from algorithm adaptation; 2) The predicted label distribution obtained by PLEML fits best with the ground-true label distribution; 3) The LDL model trained with the label distribution dataset obtained by PLEML has the best performance. In summary, the proposed PLEML algorithm has obvious advantages over other well-established label enhancement algorithms.

4.6 Validation

In order to verify that the privileged information obtained through multi-label learning can effectively improve model performance. We conducted an experiment using random values in the range (0, 1) as privileged information combined with the RSVM+ model. The comparison results are shown in Table 5, and "random PI" indicates that only random values in the range (0, 1) are used as the privilege information to learn. From the comparison results of the two experiments, the privileged information obtained through multi-label learning can significantly improve the performance of the label enhancement model. However, the result of PLEML is slightly worse than "random PI" on the dataset SJAFFE in terms of Canberra, because the number of instances is 213, which is relatively small. This may cause the multi-label model to not fully capture the mapping relationship between features and labels, and the resulting privileged information will be inaccurate.

5 Conclusion

In this paper, we studied LE problem and proposed a novel PLEML method by using privileged information, which mainly solves the problem of how to fully and accurately use additional information to enhance model performance in label enhancement. Different from existing methods, PLEML utilizes complex structure information to generate additional information through a multi-label model. RSVM+ based on LUPI paradigm is first introduced to LE problem to utilize the privileged information. The experimental results on several real-world datasets demonstrate the effectiveness of PLEML.

References

[Boyd *et al.*, 2011] Stephen P. Boyd, Neal Parikh, Eric Chu, Borja Peleato, and Jonathan Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning*, 3(1):1–122, 2011.

[Eisen *et al.*, 1998] M. B. Eisen, P. T. Spellman, P. O. Brown, and D. Botstein. Cluster analysis and display of genome-wide expression patterns. *National Academy of Sciences of the United States of America*, 95(25):14863–14868, 1998.

[Gayar *et al.*, 2006] Neamat El Gayar, Friedhelm Schwenker, and Günther Palm. A study of the robustness of KNN classifiers trained using soft labels. In *Proceedings of the Artificial Neural Networks in Pattern Recognition*, pages 67–80, 2006.

[Geng, 2016] Xin Geng. Label distribution learning. *IEEE Transactions on Knowledge and Data Engineering*, 28(7):1734–1748, 2016.

[Hou *et al.*, 2016] Peng Hou, Xin Geng, and Min-Ling Zhang. Multi-label manifold learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1680–1686, 2016.

[Jia *et al.*, 2018] Xiuyi Jia, Weiwei Li, Junyu Liu, and Yu Zhang. Label distribution learning by exploiting label correlations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3310–3317, 2018.

[Jia *et al.*, 2019a] Xiuyi Jia, Tingting Ren, Lei Chen, Jun Wang, Jihua Zhu, and Xianzhong Long. Weakly supervised label distribution learning based on transductive matrix completion with sample correlations. *Pattern Recognition Letters*, 125:453–462, 2019.

[Jia *et al.*, 2019b] Xiuyi Jia, Xiang Zheng, Weiwei Li, Changqing Zhang, and Zechao Li. Facial emotion distribution learning by exploiting low-rank label correlations locally. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, pages 9841–9850, 2019.

[Jiang *et al.*, 2006] Xiufeng Jiang, Zhang Yi, and Jiancheng Lv. Fuzzy SVM with a new fuzzy membership function. *Neural Computing and Applications*, 15(3-4):268–276, 2006.

[Li *et al.*, 2015] Yu-Kun Li, Min-Ling Zhang, and Xin Geng. Leveraging implicit relative labeling-importance information for effective multi-label learning. In *Proceedings of the International Conference on Data Mining*, pages 251–260, 2015.

[Ling and Geng, 2019] Miaogen Ling and Xin Geng. Soft video parsing by label distribution learning. *Frontiers of Computer Science*, 13(2):302–317, 2019.

[Lyons *et al.*, 1998] Michael J. Lyons, Shigeru Akamatsu, Miyuki Kamachi, and Jiro Gyoba. Coding facial expressions with gabor wavelets. In *Proceedings of the International Conference on Face & Gesture Recognition*, pages 200–205, 1998.

[Marcacini *et al.*, 2014] Ricardo Marcondes Marcacini, Marcos Aurélio Domingues, Eduardo R. Hruschka, and Solange Oliveira Rezende. Privileged information for hierarchical document clustering: A metric learning approach. In *Proceedings of the International Conference on Pattern Recognition*, pages 3636–3641, 2014.

[Ren *et al.*, 2019a] Tingting Ren, Xiuyi Jia, Weiwei Li, Lei Chen, and Zechao Li. Label distribution learning with label-specific features. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pages 3318–3324, 2019.

[Ren *et al.*, 2019b] Tingting Ren, Xiuyi Jia, Weiwei Li, and Shu Zhao. Label distribution learning with label correlations via low-rank approximation. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pages 3325–3331, 2019.

[Sun and Yuan, 2006] Wenyu Sun and Ya-Xiang Yuan. *Optimization theory and methods: nonlinear programming*, volume 1. Springer Science & Business Media, 2006.

[Vapnik and Vashist, 2009] Vladimir Vapnik and Akshay Vashist. A new learning paradigm: Learning using privileged information. *Neural Networks*, 22(5-6):544–557, 2009.

[Xu *et al.*, 2018] Ning Xu, An Tao, and Xin Geng. Label enhancement for label distribution learning. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pages 2926–2932, 2018.

[Yao *et al.*, 2019] Yazhou Yao, Fumin Shen, Jian Zhang, Li Liu, Zhenmin Tang, and Ling Shao. Extracting privileged information for enhancing classifier learning. *IEEE Transactions on Image Processing*, 28(1):436–450, 2019.

[Yin *et al.*, 2006] Lijun Yin, Xiaozhou Wei, Yi Sun, Jun Wang, and Matthew J. Rosato. A 3d facial expression database for facial behavior research. In *Proceedings of the International Conference on Automatic Face and Gesture Recognition*, pages 211–216, 2006.

[Yuan, 1991] Yaxiang Yuan. A modified bfgs algorithm for unconstrained optimization. *Ima Journal of Numerical Analysis*, 11(3):325–332, 1991.

[Zheng *et al.*, 2018] Xiang Zheng, Xiuyi Jia, and Weiwei Li. Label distribution learning by exploiting sample correlations locally. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 4556–4563, 2018.

[Zhou and Zhang, 2017] Zhi-Hua Zhou and Min-Ling Zhang. Multi-label learning. In *Sammut, C., and Webb, G. I., eds., Encyclopedia of Machine Learning and Data Mining*, pages 875–881. Berlin: Springer, 2017.

[Zhou *et al.*, 2016] Joey Tianyi Zhou, Xinxing Xu, Sinno Jialin Pan, Ivor W. Tsang, Zheng Qin, and Rick Siow Mong Goh. Transfer hashing with privileged information. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pages 2414–2420, 2016.