Label Information Bottleneck for Label Enhancement

Qinghai Zheng[1], Jihua Zhu[2]*, Haoyu Tang[3]
[1]College of Computer and Data Science, Fuzhou University, China
[2]School of Software Engineering, Xi'an Jiaotong University, Xi'an, China
[3]School of Software, Shandong University, Jinan, China

Abstract

In this work, we focus on the challenging problem of Label Enhancement (LE), which aims to exactly recover label distributions from logical labels, and present a novel Label Information Bottleneck (LIB) method for LE. For the recovery process of label distributions, the label irrelevant information contained in the dataset may lead to unsatisfactory recovery performance. To address this limitation, we make efforts to excavate the essential label relevant information to improve the recovery performance. Our method formulates the LE problem as the following two joint processes: 1) learning the representation with the essential label relevant information, 2) recovering label distributions based on the learned representation. The label relevant information can be excavated based on the "bottleneck" formed by the learned representation. Significantly, both the label relevant information about the label assignments and the label relevant information about the label gaps can be explored in our method. Evaluation experiments conducted on several benchmark label distribution learning datasets verify the effectiveness and competitiveness of LIB. Our source codes are available at `https://github.com/qinghai-zheng/LIBLE`

1. Introduction

Learning with label ambiguity is important in computer vision and machine learning. Different from the traditional Multi-Label Learning (MLL), which employs multiple logical labels to annotate one instance to address the label ambiguity issue [20], Label Distribution Learning (LDL) considers the relative importance of different labels and draws much attention in recent years [6, 8, 14, 18, 26]. By distinguishing the description degrees of all labels, LDL annotates one instance with a label distribution. Therefore, LDL is a more general learning paradigm, MLL can be regarded as a special case of LDL [8, 10, 12].

Recently, many LDL methods are proposed and achieve great success in practice [3, 9, 14, 18]. Instances with exact label distributions are vital for the training process of LDL methods. Nevertheless, annotating instances with label distributions is time-consuming [24, 28]. We take the label distribution annotation process of SJAFFE dataset for example here. SJAFFE dataset is the facial expression dataset, which contains 213 grayscale images collected from 10 Japanese female models, each facial expression image is rated by 60 persons on 6 basic emotions, including happiness, surprise, sadness, fear, anger, and disgust, with a five-level scale from 1 - 5, the higher value indicates the higher emotion intensity. Consequently, the average score of each emotion is served as the emotion label distribution [14, 28]. Clearly, the above annotation process is costly and it is unpractical to annotate data with label distributions manually, especially when the number of data is large. Fortunately, most existing datasets in the field of computer vision and machine learning are annotated by single-label or multi-labels [7, 29], therefore, a highly recommended promising solution is Label Enhancement (LE), which attempts to recover the desired label distributions exactly from existing logical labels [24, 28, 32].

Driven by the urgent requirement of obtaining label distributions and the convenience of LE, some LE methods are proposed in recent years [5, 7, 11, 13, 15, 17, 21, 24, 28, 29]. Given a dataset $\boldsymbol{X} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_n\} \in \mathbb{R}^{q \times n}$, in which q and n denote the number of dimensions and the number of instances, the potential label set is $\{y_1, y_2, \cdots, y_c\}$. The available logical labels and the desired distribution labels of $\boldsymbol{X}$ are separately indicated by $\boldsymbol{L} = \{\boldsymbol{l}_1, \boldsymbol{l}_2, \cdots, \boldsymbol{l}_n\}$ and $\boldsymbol{D} = \{\boldsymbol{d}_1, \boldsymbol{d}_2, \cdots, \boldsymbol{d}_n\}$, where $\boldsymbol{l}_i$ and $\boldsymbol{d}_i$ are:

$$\boldsymbol{l}_i = (l_i^{y_1}, l_i^{y_2}, \cdots, l_i^{y_c})^T, \boldsymbol{d}_i = (d_i^{y_1}, d_i^{y_2}, \cdots, d_i^{y_c})^T. \quad (1)$$

To be specific, LE aims to recover $\boldsymbol{D}$ based on the information provided by $\boldsymbol{X}$ and $\boldsymbol{L}$. For most existing LE methods, their objectives can be concisely summarized as follows:

$$\min_{\theta} \ \|f_\theta(\boldsymbol{X}) - \boldsymbol{L}\|_F^2 + \gamma reg(f_\theta(\boldsymbol{X})), \quad (2)$$

in which $\boldsymbol{D} = f_\theta(\boldsymbol{X})$, $f_\theta(\cdot)$ indicates the mapping from $\boldsymbol{X}$ to $\boldsymbol{D}$, $reg(\cdot)$ denotes the regularization function, and γ

*Corresponding author, E-mail: zhujh@xjtu.edu.cn

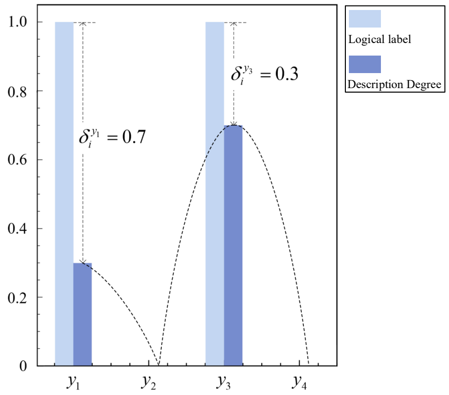


Figure 1. Illustration of label relevant information. Excavating the essential label relevant information directly is challenging and we adopt a indirect way here. We jointly investigate the information about the assignments of labels to the instance and the information about the label gaps between logical labels and label distributions. Given the i-th instance $\boldsymbol{x}_i$, the label gap of the y_j label is $\delta_i^{y_j} = l_i^{y_j} - d_i^{y_j}$. The information contained in $l_i^{y_j}$ and $\delta_i^{y_j}$ can be amalgamated to form the essential label relevant information. In other words, we employ $l_i^{y_j}$ to explore the label relevant information about the label assignments and $\delta_i^{y_j}$ to excavate the label relevant information about the label gaps. To a certain degree, the combination of $\delta_i^{y_j}$ and $l_i^{y_j}$ is equivalent to $d_i^{y_j}$. As depicted here, $l_i^{y_j}$ indicates that y_1 and y_3 are related labels and $\delta_i^{y_j}$ provides the importance of y_1 and y_3.

is the trade-off parameter. Most existing LE methods vary in $reg(\cdot)$. For example, GLLE [28] calculates the distance-based similarity matrix of data and employs the smoothness assumption [33] to construct $reg(\cdot)$; LESC [24] considers the global sample correlations and introduces the low-rank constraint as the regularization; PNLR [13] leverages $reg(\cdot)$ to maintain positive and negative label relations during the recovery process. Although a remarkable progress can be made by aforementioned methods, they ignore the label irrelevant information contained in $\boldsymbol{X}$, which prevents the further improvement of recovery results. For example, in the LE task of recovering facial age label distributions, the label irrelevant information, such as specularities information, cast shadows information, and occlusions information, may result in the incorrect mapping process of $f_\theta(\cdot)$ and the unsuitable regularization of $reg(\cdot)$, eventually leads to the unsatisfactory recovery performance.

To overcome the aforementioned limitation, we present a Label Information Bottleneck (LIB) method for LE. Concretely, the core idea of LIB is to learn the latent representation $\boldsymbol{H}$, which preserves the maximum label relevant information, from $\boldsymbol{X}$, and jointly recovers the label distributions based on the latent representation. For the LE problem, the label relevant information is the information that describes the description degrees of labels. It is tough to explore the label relevant information directly. As shown in Fig. 1, we decompose the label relevant information into two components, namely the assignments of labels to the instance and the label gaps between label distributions and logical labels. Inspired by Information Bottleneck (IB) [25], LIB utilizes the existing logical labels to explore the information about the assignments of labels to the instance. Unlike simply employing the original IB on the LE task, our method further considers the information about the label gaps between label distributions and logical labels. It is noteworthy that the above two components of the label relevant information are jointly explored in our method, and that is why we term the proposed method Label Information Bottleneck (LIB). The main contributions can be summarized as follows:

- We decompose the label relevant information into the information about the assignments of labels to instance and the information about the label gaps between logical labels, both of which can be jointly explored during the learning process of our method.
- We introduce a novel LE method, termed LIB, which excavates the label relevant information to exactly recover the label distributions. Based on the original IB, which explores the label assignments information for LE, LIB further explores the label gaps information.
- We verify the effectiveness of LIB by performing extensive experiments on several datasets. Experimental results show that the proposed method can achieve the competitive performance, compared to state-of-the-art LE methods.

2. Related Work

2.1. Label Enhancement

To recover the label distributions from the existing logical labels, many efforts are made recently [24, 28]. In general, most existing LE methods can be roughly divided into two categories, namely, algorithm adaptation and specialized algorithm [8, 24].

Algorithm adaptation extends some existing methods to achieve the goal of LE [5, 15]. For example, FCM [5] recovers the label distributions by utilizing the fuzzy clustering and fuzzy relabeling. To be specific, FCM utilizes the fuzzy C-means clustering to get different clusters and cluster prototypes, then obtains membership degrees of each instance with respect to different cluster prototypes, finally annotates all instances with label distributions by employing the fuzzy composition and softmax normalization. KM [15] leverages the fuzzy SVM to achieve the membership function. During the recovery process, KM separates instances into two clusters and employs the nonlinear function to get the radius and distances between centers and kernelized instances, and then gets the label distributions with the help of the softmax normalization.

Specialized algorithm is specially designed to deal with the LE problem. Most existing LE methods belong to the category of specialized algorithm and have the basic objective Eq. (2). By using different constraints, different methods adopt different $reg(\cdot)$ in Eq. (2). For example, based on the assumption that instances closed in the feature space are more likely to share the same label, GLLE [28] employs the following local graph information in the feature space to boost the recovery performance:

$$q_{i,j} = \begin{cases} \exp\left(-\frac{\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2}{2\varepsilon^2}\right), \text{if } \boldsymbol{x}_j \in k\left(i\right), \\ 0, \text{ otherwise}, \end{cases} \tag{3}$$

where $k\left(i\right)$ denotes the k-nearest neighbours of $\boldsymbol{x}_i$. $reg(\cdot)$ in GLLE is constructed as follows:

$$reg(f_\theta(\boldsymbol{X})) = \sum_{i,j} q_{i,j} \left\| f_\theta(\boldsymbol{x}_i) - f_\theta(\boldsymbol{x}_j) \right\|_2^2. \tag{4}$$

Unlike GLLE, LESC [24] considers the global graph information and uses the low-rank representation learning [19]:

$$\min_{\boldsymbol{G},\boldsymbol{E}} \ \|\boldsymbol{G}\|_* + \lambda_2 \|\boldsymbol{E}\|_{2,1}, \text{ s.t.}, \boldsymbol{X} = \boldsymbol{X}\boldsymbol{G} + \boldsymbol{E}, \tag{5}$$

where $\boldsymbol{G}$ indicates the low-rank representation of instances in the feature space. The regularization function in LESC is written as follows:

$$reg(f_\theta(\boldsymbol{X})) = \left\| f_\theta(\boldsymbol{X}) - f_\theta(\boldsymbol{X})\boldsymbol{G} \right\|_F^2. \tag{6}$$

For these aforementioned LE methods, they all neglect the label irrelevant information contained in $\boldsymbol{X}$, the negative effect of which can result in the unsatisfactory recovery results. Taking the recovery of facial emotion label distributions for example, the inaccurate graph information would be obtained in GLLE and LESC with the presence of label irrelevant information, such as the identity information, hindering the further improvement of recovery results.

2.2. Information Bottleneck

Information bottleneck (IB) [1, 25, 25] is an information theoretic principle, which describes the relevant information in data formally. To be concrete, IB has the following objective:

$$\min_{\boldsymbol{B}} \ -I(\boldsymbol{B},\boldsymbol{C}), \text{ s.t.}, I(\boldsymbol{A},\boldsymbol{B}) \leqslant I_c, \tag{7}$$

where $I(\cdot,\cdot)$ measures the mutual information and I_c is the information constraint. Clearly, IB aims to learn the representation $\boldsymbol{B}$, which preserves the relevant information about $\boldsymbol{C}$, from $\boldsymbol{A}$. Considering the scenario of LE, it is natural to get the following formula:

$$\min_{\boldsymbol{H}} \ -I(\boldsymbol{H},\boldsymbol{L}), \text{ s.t.}, I(\boldsymbol{X},\boldsymbol{H}) \leqslant I_c. \tag{8}$$

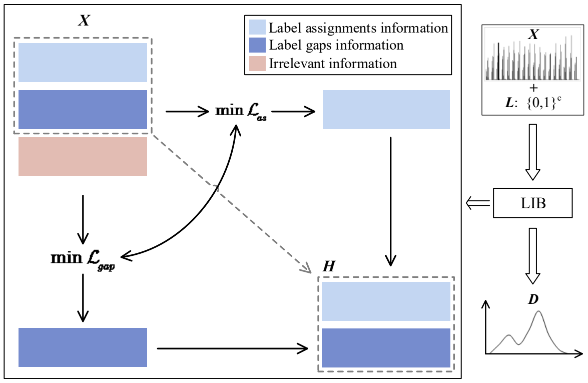


Figure 2. Framework of the proposed LIB. Since the way that directly explores the label relevant information is challenging, LIB decomposes the label relevant information into two components, namely the assignments of labels to the instance and the label gaps between label distributions and logical labels, and adopts an indirect path, which explores the information about label assignments and label gaps simultaneously. By minimizing $\mathcal{L}_{as}$, LIB explores the information about assignments of labels to the instance. Meanwhile, by minimizing $\mathcal{L}_{gap}$, LIB explores the information about the label gaps between logical labels and label distributions. Consequently, the label relevant information can be effectively explored and the label distributions can be exactly recovered.

As discussed in Section 1, the information merely about the assignments of labels to the instance can be explored based on Eq. (8), which neglects the vital information about the label gaps between logical labels and label distributions.

Recently, IB has been successfully utilized in many real-world applications [1, 2, 22, 27, 30, 31]. To the best of our knowledge, the method introduced in this paper is the first work that leverages IB to deal with the LE problem. More notably, rather than using IB simply (as shown in Eq. (8)), our method conducts more in-depth exploration to exactly recover label distributions based on IB.

3. The Proposed Method

3.1. The Objective Construction

Generally, the basic idea can be written as follows:

$$\min_{\boldsymbol{H}} \ \mathcal{L}_{as} + \alpha\mathcal{L}_{gap}, \text{ s.t.}, I(\boldsymbol{X},\boldsymbol{H}) \leqslant I_c, \tag{9}$$

where $\mathcal{L}_{as}$ excavates the information about the *as*signments of labels to the instance, $\mathcal{L}_{gap}$ investigates the information about the label *gap*s between the logical labels and distribution labels, α is the trade-off parameter, the constraint aims to remove the label irrelevant information. The framework of our method is depicted in Fig. 2. It's worth noting that employing the original IB for LE merely explores the information about the label assignments. While our LIB makes attempts to capture the information about both the label assignments and the description degrees of labels.

3.1.1 Label assignmens information modeling

For $\mathcal{L}_{as}$, inspired by IB, we have the following formula:

$$\mathcal{L}_{as} = -I(\boldsymbol{H}, \boldsymbol{L}). \tag{10}$$

According to the concept of mutual information, $\mathcal{L}_{as}$ can be rewritten out in full as follows:

$$\mathcal{L}_{as} = -\sum_{\boldsymbol{h}} \sum_{\boldsymbol{l}} p(\boldsymbol{h}, \boldsymbol{l}) \log \frac{p(\boldsymbol{l}|\boldsymbol{h})}{p(\boldsymbol{l})}. \tag{11}$$

For the convenience of optimization, we introduce the variational approximation $q(\boldsymbol{l}|\boldsymbol{h})$ to $p(\boldsymbol{l}|\boldsymbol{h})$. Since both the Kullback Leibler divergence and the entropy are positive:

$$\begin{aligned} &\mathrm{KL}(p(\boldsymbol{l}|\boldsymbol{h})||q(\boldsymbol{l}|\boldsymbol{h})) = \sum_{\boldsymbol{l}} p(\boldsymbol{l}|\boldsymbol{h}) \log \frac{p(\boldsymbol{l}|\boldsymbol{h})}{q(\boldsymbol{l}|\boldsymbol{h})} \geqslant 0 \\ &\Rightarrow \sum_{\boldsymbol{l}} p(\boldsymbol{l}|\boldsymbol{h}) \log p(\boldsymbol{l}|\boldsymbol{h}) \geqslant \sum_{\boldsymbol{l}} p(\boldsymbol{l}|\boldsymbol{h}) \log q(\boldsymbol{l}|\boldsymbol{h}), \end{aligned} \tag{12}$$

$$\mathbb{E}_{p(\boldsymbol{l})}[-\log p(\boldsymbol{l})] = -\sum_{\boldsymbol{l}} p(\boldsymbol{l}) \log p(\boldsymbol{l}) \geqslant 0, \tag{13}$$

based on Markov chain that $\boldsymbol{L} \leftarrow \boldsymbol{X} \rightarrow \boldsymbol{H}$, we can get:

$$L_{as} \leqslant -\sum_{\boldsymbol{x}} \sum_{\boldsymbol{l}} \sum_{\boldsymbol{h}} p(\boldsymbol{x}, \boldsymbol{l}) p(\boldsymbol{h}|\boldsymbol{x}) \log q(\boldsymbol{l}|\boldsymbol{h}). \tag{14}$$

3.1.2 Label gaps information modeling

To investigate the label-relevant information about the description degrees of labels, we introduce the label gaps between logical labels and label distributions $\boldsymbol{\Delta}$, and consider the conditional self-information, i.e., $I(\boldsymbol{\Delta}|\boldsymbol{H})$. Therefore, we construct $\mathcal{L}_{gap}$[1] as follows:

$$\begin{aligned} \mathcal{L}_{gap} &= I(\boldsymbol{\Delta}|\boldsymbol{H}) = -\log p(\boldsymbol{\Delta}|\boldsymbol{H}) \\ &= -\sum_{\boldsymbol{\delta}} \sum_{\boldsymbol{h}} \log p(\boldsymbol{\delta}|\boldsymbol{h}) \\ &= -\sum_{\boldsymbol{l}} \sum_{\boldsymbol{h}} \log p(\boldsymbol{l} - \hat{\boldsymbol{d}}|\boldsymbol{h}). \end{aligned} \tag{15}$$

where $\boldsymbol{\delta} = \boldsymbol{l} - \hat{\boldsymbol{d}}$, $\hat{\boldsymbol{d}}$ is the label distribution recoveried in our method.

3.1.3 Label irrelevant information modeling

Regarding the label irrelevant information, LIB employs the constraint in Eq. (9) to discard it during the learning process. $I(\boldsymbol{X}, \boldsymbol{H})$ can be formulated as follows:

$$I(\boldsymbol{X}, \boldsymbol{H}) = \sum_{\boldsymbol{x}} \sum_{\boldsymbol{h}} p(\boldsymbol{x}, \boldsymbol{h}) \log \frac{p(\boldsymbol{h}|\boldsymbol{x})}{p(\boldsymbol{h})}. \tag{16}$$

Since it is difficult to calculate $p(\boldsymbol{h})$ directly, we also introduce the variational approximation $q(\boldsymbol{h})$ to $p(\boldsymbol{h})$. Similar to Eq. (12), based on $\mathrm{KL}(p(\boldsymbol{h})||q(\boldsymbol{h})) \geqslant 0$, we have:

$$\sum_{\boldsymbol{h}} p(\boldsymbol{h}) \log p(\boldsymbol{h}) \geqslant \sum_{\boldsymbol{h}} p(\boldsymbol{h}) \log q(\boldsymbol{h}). \tag{17}$$

Subsequently, the following formula can be written:

$$\begin{aligned} I(\boldsymbol{X}, \boldsymbol{H}) &\leqslant \sum_{\boldsymbol{x}} \sum_{\boldsymbol{h}} p(\boldsymbol{x}, \boldsymbol{h}) \log \frac{p(\boldsymbol{h}|\boldsymbol{x})}{q(\boldsymbol{h})} \\ &= \sum_{\boldsymbol{x}} \sum_{\boldsymbol{l}} p(\boldsymbol{x}, \boldsymbol{l}) \mathrm{KL}(p(\boldsymbol{h}|\boldsymbol{x})||q(\boldsymbol{h})). \end{aligned} \tag{18}$$

3.1.4 Objective of LIB

By employing the Lagrange multiplier method and combining Eq. (9), (14), (15), and (18), we have:

$$\begin{aligned} \mathcal{L} &= \mathcal{L}_{as} + \alpha \mathcal{L}_{gap} + \beta I(\boldsymbol{X}, \boldsymbol{H}) \\ &\leqslant -\sum_{\boldsymbol{x}} \sum_{\boldsymbol{l}} \sum_{\boldsymbol{h}} p(\boldsymbol{x}, \boldsymbol{l}) p(\boldsymbol{h}|\boldsymbol{x}, \boldsymbol{l}) \log q(\boldsymbol{l}|\boldsymbol{h}) \\ &\quad - \alpha \sum_{\boldsymbol{l}} \sum_{\boldsymbol{h}} \log p(\boldsymbol{l} - \hat{\boldsymbol{d}}|\boldsymbol{h}) \\ &\quad + \beta \sum_{\boldsymbol{x}} \sum_{\boldsymbol{l}} p(\boldsymbol{x}, \boldsymbol{l}) \mathrm{KL}(p(\boldsymbol{h}|\boldsymbol{x})||q(\boldsymbol{h})). \end{aligned} \tag{19}$$

where β is the Lagrange multiplier. Considering the bound of $\mathcal{L}$ and using the empirical Monte Carlo approximation of sampling [23], we have the following objective of LIB:

$$\begin{aligned} \mathcal{L}_{LIB} = &\frac{1}{n} \sum_{i=1}^{n} [-\sum_{\boldsymbol{h}} p(\boldsymbol{h}|\boldsymbol{x}_i) \log q(\boldsymbol{l}_i|\boldsymbol{h}) \\ &+ \beta \mathrm{KL}(p(\boldsymbol{h}|\boldsymbol{x}_i)||q(\boldsymbol{h}))] - \alpha \sum_{\boldsymbol{l}} \sum_{\boldsymbol{h}} \log p(\boldsymbol{l} - \hat{\boldsymbol{d}}|\boldsymbol{h}). \end{aligned} \tag{20}$$

3.2. The Optimization of LIB

To minimize the objective of $\mathcal{L}_{LIB}$, we use the reparameterization trick [16, 23]. For $p(\boldsymbol{h}|\boldsymbol{x})$, we assume that:

$$p(\boldsymbol{h}|\boldsymbol{x}) \sim \mathcal{N}(\boldsymbol{\mu}_{\boldsymbol{h}|\boldsymbol{x}}, \boldsymbol{\sigma}^2_{\boldsymbol{h}|\boldsymbol{x}} \boldsymbol{I}), \tag{21}$$

where $\boldsymbol{\mu}_{\boldsymbol{h}|\boldsymbol{x}}$ and $\boldsymbol{\sigma}_{\boldsymbol{h}|\boldsymbol{x}}$ are obtained by using the encoder network $f_{\theta_{en}}(\cdot)$, i.e., $\boldsymbol{\mu}_{\boldsymbol{h}|\boldsymbol{x}} = f^{\boldsymbol{\mu}}_{\theta_{en}}(\boldsymbol{x})$ and $\boldsymbol{\sigma}_{\boldsymbol{h}|\boldsymbol{x}} = f^{\boldsymbol{\sigma}}_{\theta_{en}}(\boldsymbol{x})$. Subsequently, we have that:

$$\boldsymbol{h} = \boldsymbol{\mu}_{\boldsymbol{h}|\boldsymbol{x}} + \boldsymbol{\sigma}_{\boldsymbol{h}|\boldsymbol{x}} \odot \boldsymbol{\epsilon}, \tag{22}$$

where $\boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$ and $\odot$ is the element-wise product. For $q(\boldsymbol{l}|\boldsymbol{h})$, we assume:

$$q(\boldsymbol{l}|\boldsymbol{h}) \sim \mathcal{N}(\boldsymbol{\mu}_{\boldsymbol{l}|\boldsymbol{h}}, \boldsymbol{I}), \tag{23}$$

[1] It can be also interpreted and derived from the view of the probability distribution: $\max\limits_{\boldsymbol{\Delta}} \log p(\boldsymbol{H}, \boldsymbol{\Delta}) \Rightarrow \max\limits_{\boldsymbol{\Delta}} \log p(\boldsymbol{\Delta}|\boldsymbol{H}) + \log p(\boldsymbol{H}) \Rightarrow \max\limits_{\boldsymbol{\Delta}} \log p(\boldsymbol{\Delta}|\boldsymbol{H})$. We appreciate reviewers for their helpful comments.

where $\boldsymbol{\mu}_{\boldsymbol{l}|\boldsymbol{h}}$ is learned by using the decoder network $f_{\theta_{de}}(\cdot)$, namely, $\boldsymbol{\mu}_{\boldsymbol{l}|\boldsymbol{h}} = f_{\theta_{de}}(\boldsymbol{h})$. For $q(\boldsymbol{h})$, we assume that:

$$q(\boldsymbol{h}) \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}). \tag{24}$$

For $p(\boldsymbol{l} - \hat{\boldsymbol{d}}|\boldsymbol{h})$, the following assumption is adopted:

$$p(\boldsymbol{l} - \hat{\boldsymbol{d}}|\boldsymbol{h}) \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\sigma}^2_{\boldsymbol{\delta}|\boldsymbol{h}}\boldsymbol{I}), \tag{25}$$

where $\boldsymbol{\sigma}_{\boldsymbol{\delta}|\boldsymbol{h}}$ can be achieved by introducing the gap deviation network $f_{\theta_{gd}}(\cdot)$, i.e., $\boldsymbol{\sigma}_{\boldsymbol{\delta}|\boldsymbol{h}} = f_{\theta_{gd}}(\boldsymbol{h})$. For the recovered label distribution $\hat{\boldsymbol{d}}$, we introduce the label distribution network $f_{\theta_{ld}}(\cdot)$ and has the following formula:

$$\hat{\boldsymbol{d}} = f_{\theta_{ld}}(\boldsymbol{h}). \tag{26}$$

Consequently, based on Eq. (21)-(26), we have:

$$\begin{aligned}
&\min_{\theta_{en},\theta_{de},\theta_{gd},\theta_{ld}} \mathcal{L}_{LIB} \\
\Rightarrow &\min_{\theta_{en},\theta_{de},\theta_{gd},\theta_{ld}} \frac{1}{n}\sum_{\boldsymbol{l}}[\frac{1}{2}\|\boldsymbol{\mu}_{\boldsymbol{l}|\boldsymbol{h}} - \boldsymbol{l}\|_2^2 \\
&+\alpha(\frac{1}{2}(\boldsymbol{l} - \hat{\boldsymbol{d}})^T(\boldsymbol{\sigma}^{-2}_{\boldsymbol{\delta}|\boldsymbol{h}}\boldsymbol{I})(\boldsymbol{l} - \hat{\boldsymbol{d}}) + \log\det(\boldsymbol{\sigma}^2_{\boldsymbol{\delta}|\boldsymbol{h}}\boldsymbol{I}))] \\
&+\frac{\beta}{2}\sum_{\boldsymbol{x}}[\boldsymbol{\mu}^T_{\boldsymbol{h}|\boldsymbol{x}}\boldsymbol{\mu}_{\boldsymbol{h}|\boldsymbol{x}} + \mathrm{tr}(\boldsymbol{\sigma}^2_{\boldsymbol{h}|\boldsymbol{x}}\boldsymbol{I}) - \log\det(\boldsymbol{\sigma}^2_{\boldsymbol{h}|\boldsymbol{x}}\boldsymbol{I})].
\end{aligned} \tag{27}$$

When the problem of Eq. (27) is optimized, we can effectively recover the desired label distributions. To be specific, given $\{\boldsymbol{X}, \boldsymbol{L}\}$, we can obtain $\boldsymbol{H}$ according to Eq. (22) and achieve the recovery results based on Eq. (26), namely, $\hat{\boldsymbol{D}} = f_{\theta_{ld}}(\boldsymbol{H})$.

3.3. Comparison with Existing LE Methods

The main difference between LIB and existing methods is that our method deals with the problem of LE from the perspective of information bottleneck. Considering the first term in Eq. (2), it aims to minimize $\|\boldsymbol{d} - \boldsymbol{l}\|_2^2$ under the assumption that information in the label distributions is inherited from the initial logical labels [24, 28, 29]. For LIB, the more reasonable term:

$$\frac{1}{2}(\boldsymbol{l} - \hat{\boldsymbol{d}})^T(\boldsymbol{\sigma}^{-2}_{\boldsymbol{\delta}|\boldsymbol{h}}\boldsymbol{I})(\boldsymbol{l} - \hat{\boldsymbol{d}}) + \log\det(\boldsymbol{\sigma}^2_{\boldsymbol{\delta}|\boldsymbol{h}}\boldsymbol{I}) \tag{28}$$

which can be deduced by excavating the label relevant information about the label gaps between logical labels and label distributions.

Besides, we compare our method with the recently proposed LEVI [29] further. Although the objectives of LEVI and LIB are somewhat similar in form, they are essentially different as follows: 1) LIB makes attempts from the perspective of information bottleneck, while LEVI from the view of variational inference; 2) The formulas of LEVI and LIB are just partially similar in form, since the variational inference is employed as the optimization tool in LIB. The details of these two formulas are totally different; 3) LEVI requires an extra regularizer, i.e., $\|\boldsymbol{d} - \boldsymbol{l}\|_2^2$, to constrain the recovery process, while LIB achieves $\boldsymbol{d}$ based on the more reasonable term, i.e., Eq. (28).

Dataset	# dimension q	# instance n	# labels c
Artificial_toy	3	2601	3
Movie	1869	7755	5
SBU-3DFE	243	2500	6
SJAFFE	243	213	6
Yeast-alpha	24	2465	18
Yeast-cdc	24	2465	15
Yeast-cold	24	2465	4
Yeast-diau	24	2465	7
Yeast-dtt	24	2465	4
Yeast-elu	24	2465	14
Yeast-heat	24	2465	6
Yeast-spo	24	2465	6
Yeast-spo5	24	2465	3
Yeast-spoem	24	2465	2

Table 1. Details of datasets. The numbers of dimension q, instance n, and labels c are provided here.

4. Experiments

To verify the effectiveness and competitiveness of LIB, extensive experiments are conducted in this section.

4.1. Experimental Setup

As shown in Table 1, we use both one toy dataset and 13 real-world datasets for evaluation[2]. For the toy dataset, i.e., Artificial dataset, it is utilized to vividly show the recovery performance [28]. Movie dataset is collected from movies, SBU-3DFE and SJAFFE datasets are two facial expression datasets. Yeast datasets (alpha to spoem) are collected from 10 biological experiments on the budding yeast genes [4]. It is important to note that only the ground-truth label distributions are provided by these datasets. Therefore, we adopt the binarization strategy, which is also used in existing LE works [24, 28, 29], to ensure the consistency of evaluation.

We compare our method LIB to 7 LE methods, including FCM [5], KM [15], LP [17], ML [11], GLLE [28], LESC [24], and LEVI [29]. The first two methods belong to the algorithm adaption, and the rest methods are specialized algorithms. For the sake of fairness, we utilize the parameter settings recommended in their original works. Specifically, for FCM, we set the parameter $\beta = 2$. For KM, we leverage the Gaussian kernel. For LP, we set the parameter $\alpha = 0.5$. For ML, we set the number of neighbors $k = c{+}1$. For GLLE, we select λ from $\{0.01, 0.1, ..., 100\}$ and set the number of neighbors k to $c + 1$. For LESC, λ_1 and λ_2 are selected from $\{0.0001, 0.1, ..., 10\}$. For LEVI, MLPs with

[2] http://palm.seu.edu.cn/xgeng/LDL/index.htm

two hidden layers and softplus activation functions are utilized, and the results are reported after 150 training epochs. For LIB, we select α and β from $\{0.001, 0.01, ..., 10\}$, and the fully connected networks with 3 layers and sigmoid activation function are leveraged in the proposed method.

To evaluate the recovery performance, we adopt 6 metrics, namely Chebyshev, Canberra, Clark, Kullback-Leibler, Cosine, and Intersection [8, 24, 28]. Given the ground-truth label distribution $\boldsymbol{d}$ and the recovered label distribution $\hat{\boldsymbol{d}}$, the first four metrics and the rest two metrics respectively measure the distance and similarity between $\boldsymbol{d}$ and $\hat{\boldsymbol{d}}$:

$$D_{\text{Chebyshev}}(\boldsymbol{d}, \hat{\boldsymbol{d}}) = \max_i \left| d^{y_i} - \hat{d}^{y_i} \right|, \tag{29}$$

$$D_{\text{Canberra}}(\boldsymbol{d}, \hat{\boldsymbol{d}}) = \sum_{i=1}^{c} \frac{\left| d^{y_i} - \hat{d}^{y_i} \right|}{d^{y_i} + \hat{d}^{y_i}}, \tag{30}$$

$$D_{\text{Clark}}(\boldsymbol{d}, \hat{\boldsymbol{d}}) = \sqrt{\sum_{i=1}^{c} \frac{\left(d^{y_i} - \hat{d}^{y_i} \right)^2}{\left(d^{y_i} + \hat{d}^{y_i} \right)^2}}, \tag{31}$$

$$D_{\text{Kullback}-\text{Leibler}}(\boldsymbol{d}, \hat{\boldsymbol{d}}) = \sum_{i=1}^{c} d^{y_i} \ln \frac{d^{y_i}}{\hat{d}^{y_i}}, \tag{32}$$

$$S_{\text{Cosine}}(\boldsymbol{d}, \hat{\boldsymbol{d}}) = \frac{\sum_{i=1}^{c} d^{y_i} \hat{d}^{y_i}}{\sqrt{\sum_{i=1}^{c} \left(d^{y_i}\right)^2} \sqrt{\sum_{i=1}^{c} \left(\hat{d}^{y_i}\right)^2}}, \tag{33}$$

$$S_{\text{Intersection}}(\boldsymbol{d}, \hat{\boldsymbol{d}}) = \sum_{i=1}^{c} \min \left(d^{y_i}, \hat{d}^{y_i} \right). \tag{34}$$

The smaller values of distance metric and similarity metric indicate the better and the worse results, respectively.

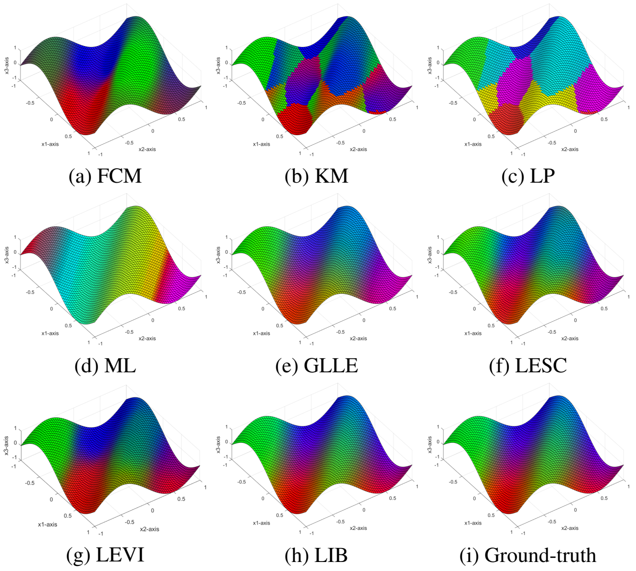
4.2. Visualization Results on Toy Dataset

The recovery results on the Artificial dataset are vividly presented in Fig. 3, which shows the three-dimensional label distributions by the RGB color channels separately. The more similar the color patterns of the recovered results and the ground-truth are, the better the recovery results are.

It can be seen that FCM, GLLE, LESC, LEVI, and LIB can obtain the similar color pattern, while KM, LP, and ML are incapable to obtain the promising recovery performance on Artificial dataset. Regarding the visualization results of FCM, GLLE, LESC, LEVI, and LIB, the color pattern that is most close to the ground-truth is achieved by our LIB.

4.3. Comparison Results on Real-world Datasets

We provide the detailed comparison results on 13 real-world datasets in Table 2. Overall, LIB has the competitive recovery performance. We have the following observations: 1) Compared with FCM and KM, which belong to the category of algorithm adaption, remarkable improvements can be achieved by our method; 2) Compared with the methods belonging to the category of specialized algorithm, LIB can also obtain better recovery results in most cases. For example, LIB obtains the best recovery results on Movie datasets in all metrics. Moreover, although LESC can obtain slightly favorable results in some cases, the corresponding recovery results of LIB are also promising and competitive. The underlying reason may be that LESC further considers the sample correlations during the recovery process; 3) The recovery performance of all methods can be roughly ranked as LIB>LESC≈LEVI>GLLE>LP≈FCM>ML>KM. We can conclude that LIB is suitable for the LE problem. The underlying reason for the significant improvement is that the label relevant information, including the information about label assignments and the information about label gaps, can be effectively investigated by LIB.



(a) FCM (b) KM (c) LP

(d) ML (e) GLLE (f) LESC

(g) LEVI (h) LIB (i) Ground-truth

Figure 3. Visualization results of the label distributions recovered by different methods ((a)-(h)) and the ground-truth ((i)) on Artificial dataset. (Best viewed in color.)

4.4. Analysis and Discussion of LIB

We analyze the parameter sensitivity of LIB firstly, and then we conduct the ablation studies as well.

4.4.1 Sensitivity of LIB

In the proposed method, we choose the values of α and β from $\{0.001, 0.01, ..., 10\}$. To show the parameter sensitivity of LIB, we conduct experiments on SBU-3DFE datasets with different values of α and β. Regarding the dimension of the learned latent representation, we set it to 256 for all datasets. The experimental results in metrics of Chebyshev distance, Cosine coefficient, and Intersection similarity are

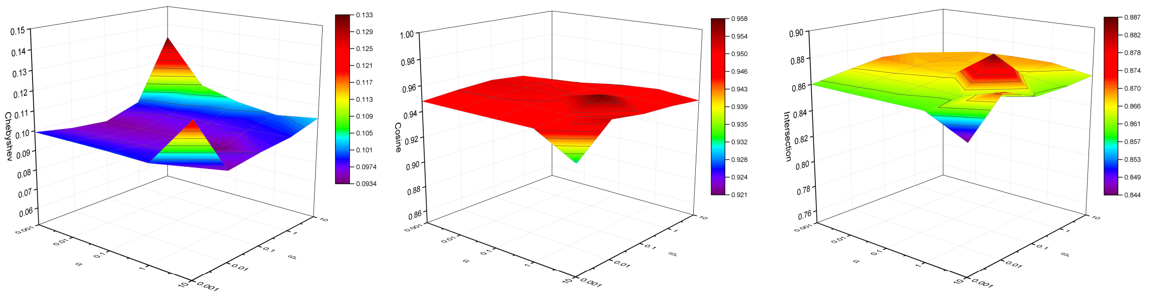
Metric	Chebyshev ↓								Clark ↓							
Method	FCM	KM	LP	ML	GLLE	LESC	LEVI	LIB	FCM	KM	LP	ML	GLLE	LESC	LEVI	LIB
Movie	0.230	0.234	0.161	0.164	0.122	0.121	0.110	**0.107**	0.859	1.766	0.913	1.140	0.569	0.564	0.551	**0.517**
SUB-3DFE	0.135	0.238	0.123	0.233	0.126	0.122	0.095	**0.094**	0.482	1.907	0.580	1.848	0.391	0.378	0.303	**0.297**
SJAFFE	0.132	0.214	0.107	0.186	0.087	**0.069**	0.075	0.071	0.522	1.874	0.502	1.519	0.377	0.276	0.290	**0.262**
Yeast-alpha	0.044	0.063	0.040	0.057	0.020	0.015	**0.012**	0.017	0.821	3.153	1.185	3.088	0.337	**0.253**	0.319	0.275
Yeast-cdc	0.051	0.076	0.042	0.071	0.022	0.019	**0.016**	0.017	0.739	2.885	1.014	2.825	0.306	0.251	0.323	**0.242**
Yeast-cold	0.141	0.252	0.137	0.242	0.066	0.056	0.082	**0.054**	0.433	1.472	0.503	1.440	0.176	0.152	0.269	**0.146**
Yeast-diau	0.124	0.152	0.099	0.148	0.053	**0.042**	0.044	0.049	0.838	1.886	0.788	1.844	0.296	**0.224**	0.295	0.273
Yeast-dtt	0.097	0.257	0.128	0.244	0.052	0.043	0.084	**0.034**	0.329	1.477	0.499	1.446	0.143	0.119	0.294	**0.092**
Yeast-elu	0.052	0.078	0.044	0.072	0.023	0.019	**0.017**	0.018	0.579	2.768	0.973	2.711	0.295	0.241	0.317	**0.224**
Yeast-heat	0.169	0.175	0.086	0.165	0.049	0.046	0.052	**0.039**	0.580	1.802	0.568	1.764	0.213	0.199	0.288	**0.165**
Yeast-spo	0.130	0.175	0.090	0.171	0.062	0.060	0.055	**0.053**	0.520	1.811	0.558	1.768	0.266	0.258	0.277	**0.224**
Yeast-spo5	0.162	0.277	0.114	0.273	0.099	0.092	0.091	**0.076**	0.395	1.059	0.274	1.036	0.197	0.185	0.209	**0.158**
Yeast-sopem	0.233	0.408	0.163	0.403	0.088	0.087	0.115	**0.069**	0.401	1.028	0.272	1.004	0.132	0.129	0.182	**0.104**
Avg.Rank	6.077	8.000	5.000	6.846	3.769	2.308	2.463	**1.538**	5.385	8.000	5.615	7.000	3.385	1.923	3.462	**1.231**
Metric	Canberra ↓								Kullback-Leibler ↓							
Method	FCM	KM	LP	ML	GLLE	LESC	LEVI	LIB	FCM	KM	LP	ML	GLLE	LESC	LEVI	LIB
Movie	1.664	3.444	1.720	1.934	1.045	1.034	0.974	**0.920**	0.381	0.452	0.177	0.218	0.123	0.120	0.082	**0.077**
SUB-3DFE	1.020	4.121	1.245	4.001	0.820	0.799	0.637	**0.611**	0.094	0.603	0.105	0.565	0.069	0.064	0.042	**0.041**
SJAFFE	1.081	4.010	1.064	3.138	0.781	0.561	0.600	**0.531**	0.107	0.558	0.077	0.391	0.050	0.029	0.032	**0.027**
Yeast-alpha	2.883	11.809	4.544	11.603	1.134	**0.846**	1.249	0.893	0.100	0.630	0.121	0.602	0.013	**0.008**	0.011	0.009
Yeast-cdc	2.415	9.875	3.644	9.695	0.959	0.765	1,148	**0.747**	0.091	0.630	0.111	0.601	0.014	0.010	0.014	**0.008**
Yeast-cold	0.734	2.566	0.924	2.519	0.305	0.263	0.501	**0.250**	0.113	0.586	0.103	0.556	0.019	0.015	0.035	**0.012**
Yeast-diau	1.895	4.261	1.748	4.180	0.671	**0.480**	0.689	0.621	0.159	0.538	0.127	0.509	0.027	**0.017**	0.023	0.022
Yeast-dtt	0.501	2.594	0.941	2.549	0.248	0.206	0.562	**0.158**	0.065	0.617	0.103	0.586	0.013	0.010	0.042	**0.005**
Yeast-elu	1.689	9.110	3.381	8.949	0.902	0.727	1.093	**0.670**	0.059	0.617	0.109	0.589	0.013	0.009	0.014	**0.008**
Yeast-heat	1.157	3.849	1.293	3.779	0.430	0.401	0.646	**0.327**	0.147	0.586	0.089	0.556	0.017	0.015	0.027	**0.011**
Yeast-spo	0.998	3.854	1.231	3.772	0.548	0.533	0.605	**0.454**	0.110	0.562	0.084	0.532	0.029	0.028	0.025	**0.019**
Yeast-spo5	0.563	1.382	0.401	1.355	0.305	0.284	0.311	**0.241**	0.123	0.334	0.042	0.317	0.034	0.031	0.028	**0.021**
Yeast-sopem	0.534	1.253	0.365	1.226	0.183	0.180	0.248	**0.144**	0.208	0.531	0.067	0.503	0.027	0.027	0.036	**0.018**
Avg.Rank	5.231	8.000	5.692	7.000	3.231	2.00	3.692	**1.154**	5.692	8.000	5.385	6.923	3.462	2.154	3.077	**1.154**
Metric	Cosine ↑								Intersection ↑							
Method	FCM	KM	LP	ML	GLLE	LESC	LEVI	LIB	FCM	KM	LP	ML	GLLE	LESC	LEVI	LIB
Movie	0.773	0.880	0.929	0.919	0.936	0.937	0.954	**0.955**	0.677	0.649	0.778	0.779	0.831	0.833	0.849	**0.859**
SUB-3DFE	0.912	0.812	0.922	0.815	0.927	0.932	0.956	**0.958**	0.827	0.579	0.810	0.587	0.850	0.855	0.882	**0.887**
SJAFFE	0.906	0.827	0.941	0.857	0.958	0.973	0.969	**0.974**	0.821	0.593	0.837	0.661	0.872	0.905	0.897	**0.909**
Yeast-alpha	0.922	0.751	0.911	0.756	0.987	**0.992**	0.989	**0.992**	0.844	0.532	0.774	0.537	0.938	**0.953**	0.932	0.951
Yeast-cdc	0.929	0.754	0.916	0.759	0.987	0.991	0.987	**0.992**	0.847	0.533	0.779	0.538	0.937	0.950	0.925	**0.951**
Yeast-cold	0.922	0.779	0.925	0.784	0.982	0.986	0.970	**0.988**	0.833	0.559	0.794	0.565	0.924	0.935	0.881	**0.938**
Yeast-diau	0.882	0.799	0.915	0.803	0.975	**0.985**	0.980	0.979	0.760	0.588	0.788	0.593	0.906	**0.933**	0.908	0.913
Yeast-dtt	0.959	0.759	0.921	0.763	0.988	0.991	0.965	**0.995**	0.894	0.541	0.786	0.546	0.939	0.949	0.866	**0.961**
Yeast-elu	0.950	0.758	0.918	0.763	0.987	0.991	0.987	**0.992**	0.883	0.539	0.782	0.544	0.936	0.949	0.924	**0.952**
Yeast-heat	0.883	0.779	0.932	0.783	0.984	0.986	0.977	**0.990**	0.807	0.559	0.805	0.564	0.929	0.934	0.897	**0.946**
Yeast-spo	0.909	0.800	0.939	0.803	0.974	0.975	0.978	**0.982**	0.836	0.575	0.819	0.580	0.909	0.912	0.903	**0.925**
Yeast-spo5	0.922	0.882	0.969	0.884	0.971	0.974	0.979	**0.983**	0.838	0.724	0.886	0.727	0.901	0.908	0.909	**0.924**
Yeast-sopem	0.878	0.812	0.950	0.815	0.978	0.978	0.972	**0.985**	0.767	0.592	0.837	0.597	0.912	0.913	0.885	**0.931**
Avg.Rank	5.846	7.923	5.308	6.923	3.462	2.154	2.923	**1.231**	5.385	8.000	5.692	6.846	3.385	2.007	3.462	**1.154**

Table 2. Recovery results on 13 real-world datasets. ↓ indicates that "the smaller the better" and ↑ means that "the larger the better". The average ranks (Avg.Rank) on all datasets are also reported for all methods. We highlight the best recovery results.

provided in Fig. 4. It can be observed that LIB method can get promising recovery results and is robust with respect to different values of α and β in a large range.

4.4.2 Ablation studies of LIB

The ablation studies are conducted to further verify the effectiveness of introducing the label information bottleneck framework for LE. In the proposed objective Eq. (9), $\mathcal{L}_{as}$ and $\mathcal{L}_{gap}$ explore the label assignments information and label gaps information during the recovery process. As can be observed from Eq. (14) and Eq. (15), only the latent representation $\boldsymbol{H}$ can be learned if we employ $\mathcal{L}_{as}$ merely during the recovery process. Consequently, considering the goal of LE, we compare the proposed LIB with the method termed LIB$_{gap}$, which only employs $\mathcal{L}_{gap}$ to achieve the recovery results. In other words, LIB$_{gap}$ investigates the label gaps information in the case of not considering the label assign-



(a) Chebyshev ↓

(b) Cosine ↑

(c) Intersection ↑

Figure 4. The recovery results in metrics of (a) Chebyshev distance, (b) Cosine coefficient, and (c) Intersection similarity with different values of α and β. ↓ indicates that "the smaller the better" and ↑ means that "the larger the better". The experimental results demonstrate that LIB is robust with respect to different values of λ and β. (Best viewed in color.)

Metric	Chebyshev ↓		Clark ↓		Canberra ↓		Kullback-Leibler ↓		Cosine ↑		Intersection ↑	
Method	LIB$_{gap}$	LIB	LIB$_{gap}$	LIB	LIB$_{gap}$	LIB	LIB$_{gap}$	LIB	LIB$_{gap}$	LIB	LIB$_{gap}$	LIB
Movie	0.120	**0.107**	0.563	**0.517**	1.029	**0.920**	0.099	**0.077**	0.938	**0.955**	0.834	**0.859**
SUB-3DFE	0.130	**0.094**	0.395	**0.297**	0.849	**0.611**	0.079	**0.041**	0.923	**0.958**	0.846	**0.887**
SJAFFE	0.113	**0.071**	0.391	**0.262**	0.816	**0.531**	0.066	**0.027**	0.938	**0.973**	0.860	**0.909**
Yeast-alpha	0.018	**0.017**	0.281	**0.275**	0.920	**0.893**	0.010	**0.009**	0.991	**0.992**	0.950	**0.951**
Yeast-cdc	0.019	**0.017**	0.254	**0.242**	0.782	**0.747**	0.009	**0.008**	0.991	**0.992**	0.948	**0.951**
Yeast-cold	0.061	**0.017**	0.162	**0.146**	0.280	**0.250**	0.016	**0.012**	0.985	**0.988**	0.930	**0.938**
Yeast-diau	0.050	**0.049**	0.288	**0.273**	0.659	**0.621**	0.025	**0.022**	0.977	**0.979**	0.908	**0.913**
Yeast-dtt	0.045	**0.034**	0.124	**0.092**	0.217	**0.158**	0.010	**0.005**	0.991	**0.995**	0.946	**0.961**
Yeast-elu	0.019	**0.018**	0.237	**0.224**	0.714	**0.670**	0.009	**0.008**	0.992	**0.992**	0.949	**0.952**
Yeast-heat	0.045	**0.039**	0.193	**0.165**	0.388	**0.327**	0.014	**0.011**	0.986	**0.990**	0.936	**0.946**
Yeast-spo	0.059	**0.053**	0.253	**0.224**	0.523	**0.454**	0.025	**0.019**	0.976	**0.982**	0.914	**0.925**
Yeast-spo5	0.097	**0.076**	0.193	**0.158**	0.300	**0.241**	0.032	**0.021**	0.971	**0.983**	0.903	**0.924**
Yeast-sopem	0.088	**0.069**	0.130	**0.104**	0.181	**0.144**	0.027	**0.018**	0.977	**0.985**	0.912	**0.931**

Table 3. Recovery results of LIB$_{gap}$ and LIB on 13 real-world datasets. ↓ indicates that "the smaller the better" and ↑ means that "the larger the better". We highlight the best recovery results.

ments information during the recovery process.

To be specific, LIB$_{gap}$ has with the following objective:

$$\min_{\theta_{gd},\theta_{ld}} \frac{1}{2}\sum_{\boldsymbol{x}}[(\boldsymbol{l}-\hat{\boldsymbol{d}})^T(\boldsymbol{\sigma}_{\boldsymbol{\delta}|\boldsymbol{x}}^{-2}\boldsymbol{I})(\boldsymbol{l}-\hat{\boldsymbol{d}})+\log\det(\boldsymbol{\sigma}_{\boldsymbol{\delta}|\boldsymbol{x}}^{2}\boldsymbol{I})]. \tag{35}$$

Notably, $\boldsymbol{\sigma}_{\boldsymbol{\delta}|\boldsymbol{x}} = f_{\theta_{gd}}(\boldsymbol{x})$ and $\hat{\boldsymbol{d}} = f_{\theta_{ld}}(\boldsymbol{x})$, which are different from the objective Eq. (27) utilized in LIB. Only the partial label relevant information, i.e., the information about the label gaps, is explored in LIB$_{gap}$.

Table 3 provides the recovery results of LIB$_{gap}$ and LIB. It can be observed that LIB outperforms LIB$_{gap}$ in all cases. Compared with LIB, LIB$_{gap}$ merely makes the effort to explore the label gap information to boost the recovery performance, while LIB excavates both the information about the label assignments and the information about the label gaps jointly. Therefore, the promising recovery performance can be achieved by LIB. Furthermore, according to the results provided in Table 2 and 3, the recovery results of LIB$_{gap}$ seem to be competitive, which also indicates that the exploration of information about label gaps is beneficial for LE.

5. Conclusion

In this paper, we present a new perspective to deal with the Label Enhancement (LE) problem and introduce the novel Label Information Bottleneck (LIB) method. The label relevant information is decomposed into the information about label assignments and the information about label gaps. Consequently, our method transform the LE problem into simultaneously learning the latent representation and modeling the label gaps. Extensive experiments carried on both the toy dataset and real-world datasets verify the competitiveness of LIB.

Acknowledgments

This work was supported by the National Key R&D Program of China under Grant 2020AAA0109602; the Education and Research Foundation for Middle-aged and Young Teacher of Fujian Province under Grant JAT220005; the National Natural Science Foundation of China (NSFC) under Grant 62206156; Alibaba Group through Alibaba Innovative Research Program, No.21169774.

References

[1] A Alemi, Ian Fischer, J Dillon, and Kevin Murphy. Deep variational information bottleneck int. In *Conf. on Learning Representations*, 2017. 3

[2] Rana Ali Amjad and Bernhard C Geiger. Learning representations for neural network-based classification using the information bottleneck principle. *IEEE transactions on pattern analysis and machine intelligence*, 42(9):2225–2239, 2019. 3

[3] Shikai Chen, Jianfeng Wang, Yuedong Chen, Zhongchao Shi, Xin Geng, and Yong Rui. Label distribution learning on auxiliary label space graphs for facial expression recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13984–13993, 2020. 1

[4] Michael B Eisen, Paul T Spellman, Patrick O Brown, and David Botstein. Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences*, 95(25):14863–14868, 1998. 5

[5] Neamat El Gayar, Friedhelm Schwenker, and Günther Palm. A study of the robustness of knn classifiers trained using soft labels. In *IAPR Workshop on Artificial Neural Networks in Pattern Recognition*, pages 67–80. Springer, 2006. 1, 2, 5

[6] Bin-Bin Gao, Hong-Yu Zhou, Jianxin Wu, and Xin Geng. Age estimation using expectation of label distribution learning. In *IJCAI*, pages 712–718, 2018. 1

[7] Yongbiao Gao, Yu Zhang, and Xin Geng. Label enhancement for label distribution learning via prior knowledge. In *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, pages 3223–3229, 2021. 1

[8] Xin Geng. Label distribution learning. *IEEE Transactions on Knowledge and Data Engineering*, 28(7):1734–1748, 2016. 1, 2, 6

[9] Xin Geng and Peng Hou. Pre-release prediction of crowd opinion on movies by label distribution learning. In *Twenty-fourth international joint conference on artificial intelligence*, 2015. 1

[10] Youngkyu Hong, Seungju Han, Kwanghee Choi, Seokjun Seo, Beomsu Kim, and Buru Chang. Disentangling label distribution for long-tailed visual recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6626–6636, 2021. 1

[11] Peng Hou, Xin Geng, and Min-Ling Zhang. Multi-label manifold learning. In *Thirtieth AAAI Conference on Artificial Intelligence*, 2016. 1, 5

[12] Xiuyi Jia, Zechao Li, Xiang Zheng, Weiwei Li, and Sheng-Jun Huang. Label distribution learning with label correlations on local samples. *IEEE Transactions on Knowledge and Data Engineering*, 33(4):1619–1631, 2019. 1

[13] Xiuyi Jia, Yunan Lu, and Fangwen Zhang. Label enhancement by maintaining positive and negative label relation. *IEEE Transactions on Knowledge and Data Engineering*, 2021. 1, 2

[14] 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 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9841–9850, 2019. 1

[15] Xiufeng Jiang, Zhang Yi, and Jian Cheng Lv. Fuzzy svm with a new fuzzy membership function. *Neural Computing & Applications*, 15(3-4):268–276, 2006. 1, 2, 5

[16] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. In *International Conference on Learning Representations*. 4

[17] Yu-Kun Li, Min-Ling Zhang, and Xin Geng. Leveraging implicit relative labeling-importance information for effective multi-label learning. In *IEEE International Conference on Data Mining*, pages 251–260. IEEE, 2015. 1, 5

[18] Miaogen Ling and Xin Geng. Indoor crowd counting by mixture of gaussians label distribution learning. *IEEE Transactions on Image Processing*, 28(11):5691–5701, 2019. 1

[19] Guangcan Liu, Zhouchen Lin, Shuicheng Yan, Ju Sun, Yong Yu, and Yi Ma. Robust recovery of subspace structures by low-rank representation. *IEEE transactions on pattern analysis and machine intelligence*, 35(1):171–184, 2012. 3

[20] Weiwei Liu, Haobo Wang, Xiaobo Shen, and Ivor W Tsang. The emerging trends of multi-label learning. *IEEE transactions on pattern analysis and machine intelligence*, 44(11):7955–7974, 2021. 1

[21] Jia-Qi Lv, Ning Xu, Ren-Yi Zheng, and Xin Geng. Weakly supervised multi-label learning via label enhancement. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, pages 3101–3107, Macao, China, 2019. 1

[22] Kaizhi Qian, Yang Zhang, Shiyu Chang, Mark Hasegawa-Johnson, and David Cox. Unsupervised speech decomposition via triple information bottleneck. In *International Conference on Machine Learning*, pages 7836–7846. PMLR, 2020. 3

[23] Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In *International conference on machine learning*, pages 1278–1286. PMLR, 2014. 4

[24] Haoyu Tang, Jihua Zhu, Qinghai Zheng, Jun Wang, Shanmin Pang, and Zhongyu Li. Label enhancement with sample correlations via low-rank representation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 5932–5939, 2020. 1, 2, 3, 5, 6

[25] Naftali Tishby, Fernando C Pereira, and William Bialek. The information bottleneck method. *arXiv preprint physics/0004057*, 2000. 2, 3

[26] Jing Wang and Xin Geng. Label distribution learning machine. In *International Conference on Machine Learning*, pages 10749–10759. PMLR, 2021. 1

[27] Tailin Wu, Hongyu Ren, Pan Li, and Jure Leskovec. Graph information bottleneck. *Advances in Neural Information Processing Systems*, 33:20437–20448, 2020. 3

[28] Ning Xu, Yun-Peng Liu, and Xin Geng. Label enhancement for label distribution learning. *IEEE Transactions on Knowledge and Data Engineering*, 33(4):1632–1643, 2021. 1, 2, 3, 5, 6

[29] Ning Xu, Jun Shu, Yun-Peng Liu, and Xin Geng. Variational label enhancement. In *International Conference on Machine Learning*, pages 10597–10606. PMLR, 2020. 1, 5

[30] Junchi Yu, Jie Cao, and Ran He. Improving subgraph recognition with variational graph information bottleneck. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19396–19405, 2022. 3

[31] Anguo Zhang, Yueming Gao, Yuzhen Niu, Wenxi Liu, and Yongcheng Zhou. Coarse-to-fine person re-identification with auxiliary-domain classification and second-order information bottleneck. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 598–607, 2021. 3

[32] Qinghai Zheng, Jihua Zhu, Haoyu Tang, Xinyuan Liu, Zhongyu Li, and Huimin Lu. Generalized label enhancement with sample correlations. *IEEE Transactions on Knowledge and Data Engineering*, 35(1):482–495, 2023. 1

[33] Xiaojin Zhu, John Lafferty, and Ronald Rosenfeld. *Semi-supervised learning with graphs*. PhD thesis, Carnegie Mellon University, language technologies institute, 2005. 2