Label Distribution Learning from Logical Label (Supplementary File)

Paper ID: 3314

A Proof of Theorem 1 and Theorem 2

A.1 Proof of Theorem 1

Theorem 1. *The KL-divergence loss function ℓ can be written as $\mathrm{KL}(\mathbf{D}, \mathbf{P})$, where $\mathbf{D} \in \mathbb{R}^{n\times c}$ and $\mathbf{P} \in \mathbb{R}^{n\times c}$ are the recovered LD matrix and the prediction matrix respectively, in which $\mathbf{P}$ can be expressed by the prediction weight matrix $\mathbf{W} \in \mathbb{R}^{m\times c}$. Let $\mathcal{H} = \mathbf{D} \times \mathbf{W}$ represent the family of functions for DLDL, with functions $(\mathbf{D}, \mathbf{W}) \in \mathcal{H}$. We assume the complexity of $\mathbf{W}$ and the rank of $\mathbf{D}$ are upper bounded by ϵ_1 and ϵ_2 respectively, i.e., $||\mathbf{W}||_F \leq \epsilon_1$ and $rank(\mathbf{D}) \leq \epsilon_2$. According to **Definition 1**, the Rademacher complexity of DLDL with KL-divergence loss ℓ is upper bounded as follows:*

$$\widehat{\mathcal{R}}_S(\ell \circ \mathcal{H}) \leq \frac{\epsilon_2\sqrt{cm \cdot \exp\left(m|X_{max}\epsilon_1|\right)/\exp\left(-m|X_{min}\epsilon_1|\right)}}{\sqrt{n}}, \tag{1}$$

in which X_{max}, X_{min} are the maximum and minimum element in the feature matrix $\mathbf{X}$ respectively.

Proof:

$$\begin{aligned}
\widehat{\mathcal{R}}_S(\ell \circ \mathcal{H}) &= \mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{h\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^{n} \sigma_i \ell(h\left(x_i\right))\right] \\
&\leq \frac{1}{n}\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{h\in\mathcal{H}} \sum_{i=1}^{n}\sum_{j=1}^{m} \sigma_{ij} D_{ij} \ln \frac{D_{ij}}{P_{ij}}\right] \\
&\leq \frac{1}{n}\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{h\in\mathcal{H}} \sum_{i=1}^{n}\sum_{j=1}^{m} \sigma_{ij} D_{ij} \frac{D_{ij}}{P_{ij}}\right].
\end{aligned} \tag{2}$$

Let $A_{ij} = \sigma_{ij}D_{ij}, B_{ij} = \frac{D_{ij}}{P_{ij}}$, it is easy to prove that

$$\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{h\in\mathcal{H}} \sum_{i=1}^{n}\sum_{j=1}^{m} \sigma_{ij} D_{ij} \frac{D_{ij}}{P_{ij}}\right] = \frac{1}{n}\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{h\in\mathcal{H}} < \mathbf{A}^T, \mathbf{B} >_F\right] \leq \frac{1}{n}\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{h\in\mathcal{H}} ||\mathbf{A}||_F||\mathbf{B}||_F\right]. \tag{3}$$

Since $\sigma_{ij} \leq 1$ [3], $D_{ij} \leq 1, rank(\mathbf{D}) \leq \epsilon_2$, we have

$$\frac{1}{n}\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{h\in\mathcal{H}} ||\mathbf{A}||_F||\mathbf{B}||_F\right] \leq \frac{1}{n}\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{h\in\mathcal{H}} ||\mathbf{D}||_* \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{c} \frac{1}{P_{ij}^2}}\right] \leq \frac{1}{n}\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{h\in\mathcal{H}} \epsilon_2 \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{c} \frac{1}{P_{ij}^2}}\right]. \tag{4}$$

Since $||\mathbf{W}||_F \leq \epsilon_1, P_{ij} = \frac{\exp\left(X_iW\right)_j}{\sum_{k=1}^{m}\exp\left(X_iW\right)_k}$, then $W_{min} \geq -\epsilon_1$, $W_{max} \leq \epsilon_1$, in which W_{max}, W_{min} are the maximum and minimum element in the matrix $\mathbf{W}$ respectively. Thus, we can prove that

$$\begin{aligned} &\tfrac{1}{n}\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{h\in\mathcal{H}} \epsilon_2 \sqrt{\sum_{i=1}^n\sum_{j=1}^c \tfrac{1}{P_{ij}^2}}\right] \\ &\le \tfrac{1}{n}\mathbb{E}_{\boldsymbol{\sigma}}\left[\epsilon_2\sqrt{ncm\cdot\exp\left(m\left|X_{\max}\epsilon_1\right|\right)/\exp\left(-m\left|X_{\min}\epsilon_1\right|\right)}\right]\\ &\le \tfrac{\epsilon_2\sqrt{cm\cdot\exp(m|X_{\max}\epsilon 1|)/\exp(-m|X_{\min}\epsilon 1|)}}{\sqrt{n}} \end{aligned}\tag{5}$$

According to Eqs. (2) to (5), we have **Theorem 1**, i.e., Eq. (1).

A.2 Proof of Theorem 2

Theorem 2. *Denote $\mathbf{D}\in\mathbb{R}^{n\times c}$ and $\mathbf{W}\in\mathbb{R}^{m\times c}$ as the recovered LD matrix and the prediction weight matrix, and we assume that the complexity of $\mathbf{W}$ and the rank of $\mathbf{D}$ are upper bounded by ϵ_1 and ϵ_2 respectively. Then we have the upper bound of Θ:*

$$\Theta \le \sum_{i=1}^n\sum_{j=1}^c \ln\left(m\exp\left(m|X_{max}\epsilon_1|\right)\right)/\exp\left(-m|X_{min}\epsilon_1|\right).\tag{6}$$

Proof:

$$\mathrm{KL}(\mathbf{D},\mathbf{P}) = \sum_{i=1}^n\sum_{j=1}^c D_{ij}\ln\frac{D_{ij}}{P_{ij}}.\tag{7}$$

Since $D_{ij}\le 1$, we have

$$\begin{aligned}\mathrm{KL}(\mathbf{D},\mathbf{P}) &\le \sum_{i=1}^n\sum_{j=1}^c \ln\frac{1}{P_{ij}}\\ &\le \sum_{i=1}^n\sum_{j=1}^c \ln\left(m\exp\left(m|X_{max}\epsilon_1|\right)\right)/\exp\left(-m|X_{min}\epsilon_1|\right).\end{aligned}\tag{8}$$

Because Θ is defined as the upper bound of the KL-divergence function, we finally have Theorem 2, i.e., Eq. (6).

B Proof of Theorem 3

By assuming that the number of zeros in the ground-truth label distribution matrix $\mathbf{D}_g$ is large enough, we have the following Lemma 1:

Lemma 1. *The ground-truth label distribution matrix $\mathbf{D}_g$ can be sorted as follows:*

$$\mathbf{D}_g = \begin{pmatrix} <\mathbf{SD}_1> \\ <\mathbf{SD}_2> \\ \cdots \\ <\mathbf{SD}_c> \\ \mathbf{WD}_1 \\ \mathbf{WD}_2 \\ \cdots \\ \mathbf{WD}_{n-c} \end{pmatrix},\tag{9}$$

in which $<\mathbf{SD}_l>$ is an a_l-tuple which consists of a_l rows of partial-zero vectors with only their l-th column is one ($1\le l\le c$), and $\mathbf{WD}_l$ equaling to a vector, which satisfies $\mathbf{1}^T\mathbf{WD}_i = 1$ ($c < i \le n$). Because the number of zeros is large enough, we assume that $\forall l\in[1,c], a_l$ is also large enough. ***In the following discussion, we refer to $<\mathbf{SD}_l>$ as a strong sample tuple, and $<\mathbf{WD}_i>$ as a weak sample.***

For the sake of simplicity, we assume that the local similarity matrix $\mathbf{A}$ is binary, in which $\mathbf{A}_{ij}=1$ means that the i-th row in $\mathbf{D}_g$ is related to the j-th row of it and vice versa.

Because a weak sample can be related to many strong samples in a specified strong sample tuple and all strong samples in a common tuple are the same, so we use a single strong sample to represent the tuple it belongs to:

$$\mathbf{D}_g = \begin{pmatrix} \mathbf{SD}_1 \\ \mathbf{SD}_2 \\ \cdots \\ \mathbf{SD}_c \\ \mathbf{WD}_1 \\ \mathbf{WD}_2 \\ \cdots \\ \mathbf{WD}_{n-c} \end{pmatrix}, \tag{10}$$

and we define the coefficient of correlation strength k_{il} for each $\mathbf{SD}_l$. It indicates that there are k_{il} rows of $\mathbf{SD}_l$ are related to the i-th weak sample $\mathbf{WD}_i$ $(1 \leq l \leq c, c < i \leq n)$. Because each a_i is large enough, so $\forall i, k_{il} < a_l$. In addition, we define the total number of connections to strong samples $\Sigma k_i = \sum_{l=1}^{c} k_{il}$ for convenience. If a weak sample $\mathbf{WD}_i$ is related to another weak sample $\mathbf{SD}_j$, $k_{ij} = 1$.

For the rows in the ground-truth label distribution matrix $\mathbf{D}_g$, we classify them into three cases according to their characteristics and their relationships to other rows, in which the probability of the i-th case is p_i and the total recovery difference is $\mathcal{L}_i$, i.e.

$$\mathcal{L}_i = \sum_{j \in case\ i} \sum_{l} |d_{jl} - d_{g(jl)}|, \tag{11}$$

where d_{jl} is the element of row j and column l in the recovered label distribution matrix $\mathbf{D}$ and $d_{g(jl)}$ is the element of row j and column l in the ground-truth label distribution matrix $\mathbf{D}_g$.

Case 1: The i-th row is a strong sample:

In this case, the logical label of this row $\mathbf{y}_i$ should be equal to the corresponding label distribution $\mathbf{d}_{g(i)}$. Due to the restriction of $\mathbf{0}_{m\times c} \leq \mathbf{D} \leq \mathbf{Y}$, our algorithm can precisely recover this row. So we have

$$\mathcal{L}_1 = n * c * p_1 * 0 = 0. \tag{12}$$

Case 2: The i-th row is a weak sample, and it is mainly connected to the strong samples:

The total number of connections to strong samples of this row is Σk_i, we further assume the total number of connections to weak samples is $\epsilon \Sigma k_i$, where ϵ is a small number. And we have Lemma 2 holds in this case:

Lemma 2. *For $\mathbf{D}_g$, when the i-th row is a weak sample and it is mainly connected to the strong samples, the coefficient of correlation strength plays a major role in determining the value of the l-th element of this row. So the coefficient of correlation strength of two rows is inversely proportional to the distance of their corresponding elements:*

$$\forall i, l, \quad k_{ij}|d_{g(il)} - d_{g(jl)}| = \Sigma_{m \neq j} k_{im} |d_{g(il)} - d_{g(ml)}| + \rho, \tag{13}$$

where ρ is a small number.

In Eq. (13), we take $j = l$, and accordingly have

$$\begin{aligned} k_{il}|d_{g(il)} - d_{g(ll)}| &= \Sigma_{m \neq j} k_{im} |d_{g(il)} - d_{g(ml)}| + \rho, \\ k_{il}(1 - d_{g(il)}) &= (\Sigma k - kl)\,(d_{g(il)} - 0) + \Sigma_{m=c+1}^{n} k_{im} |d_{g(il)} - d_{g(ml)}| + \rho, \\ k_{il}(1 - d_{g(il)}) &\leq (\Sigma k - kl)\,(d_{g(il)}) + \Sigma_{m=c+1}^{n} k_{im} + \rho, \\ k_{il}(1 - d_{g(il)}) &\leq (\Sigma k - kl)\,(d_{g(il)}) + \epsilon \Sigma k + \rho, \\ k_{il} - \rho - \epsilon \Sigma k &\leq \Sigma k d_{g(il)}, \\ d_{g(il)} &\geq \frac{k_{il} - \rho - \epsilon \Sigma k}{\Sigma k}, \end{aligned} \tag{14}$$

in which ρ, ϵ are small numbers.

And we also have

$$
\begin{aligned}
k_{il}|d_{g(il)} - d_{g(ll)}| &= \Sigma_{m\neq j} k_{im}|d_{g(il)} - d_{g(ml)}| + \rho, \\
k_{il}(1 - d_{g(il)}) &= (\Sigma k - kl)\,(d_{g(il)} - 0) + \Sigma_{m=c+1}^{n} k_{im}|d_{g(il)} - d_{g(ml)}| + \rho, \\
k_{il}(1 - d_{g(il)}) &\geq (\Sigma k - kl)\,(d_{g(il)}) - \Sigma_{m=c+1}^{n} k_{im} + \rho, \\
k_{il}(1 - d_{g(il)}) &\geq (\Sigma k - kl)\,(d_{g(il)}) - \epsilon\Sigma k + \rho, \\
k_{il} - \rho + \epsilon\Sigma k &\geq \Sigma k d_{g(il)}, \\
d_{g(il)} &\leq \frac{k_{il} - \rho + \epsilon\Sigma k}{\Sigma k}.
\end{aligned}
\tag{15}
$$

Finally, we can get

$$
\frac{k_{il} - \rho - \epsilon\Sigma k}{\Sigma k} \leq d_{g(il)} \leq \frac{k_{il} - \rho + \epsilon\Sigma k}{\Sigma k}. \tag{16}
$$

For our model, the optimization target of the label enhancement part is:

$$
\min_{\mathbf{D}} \operatorname{tr}\left(\mathbf{D}\mathbf{G}\mathbf{D}^{\mathrm{T}}\right) + \lambda||\mathbf{D}||_{\mathrm{F}}^{2} \qquad \text{s.t. } \mathbf{0}_{\mathrm{m}\times\mathrm{c}} \leq \mathbf{D} \leq \mathbf{Y}, \mathbf{D}\mathbf{1}_{\mathrm{c}} = \mathbf{1}_{\mathrm{n}}, \tag{17}
$$

where $\lambda = \frac{\beta}{\alpha}$.

Expanding the trace term, Eq. (17) becomes

$$
\min_{\mathbf{D}} \sum_{i,j} A_{ij}||\boldsymbol{d}_i - \boldsymbol{d}_j||_2^2 + \lambda||\mathbf{D}||_F^2 \qquad \text{s.t. } \mathbf{0}_{m\times c} \leq \mathbf{D} \leq \mathbf{Y}, \mathbf{D}\mathbf{1}_c = \mathbf{1}_n. \tag{18}
$$

Let $\mathcal{L}_D = \sum_{i,j} A_{ij}||\boldsymbol{d}_i - \boldsymbol{d}_j||_2^2 + \lambda||\mathbf{D}||_F^2$, the gradient of $\mathcal{L}_D$ for d_{il} is

$$
\frac{\partial \mathcal{L}_D}{\partial d_{il}} = \sum_j 2A_{ij}(d_{il} - d_{jl}) + 2\lambda d_{il}. \tag{19}
$$

Let the gradient $\frac{\partial \mathcal{L}_D}{\partial d_{il}}$ equal to 0, we have

$$
\begin{aligned}
\sum_j 2A_{ij}(d_{il} - d_{jl}) + 2\lambda d_{il} &= 0 \\
\sum_{j=1}^{c} A_{ij}(d_{il} - d_{jl}) + \sum_{j=c+1}^{n} A_{ij}(d_{il} - d_{jl}) + \lambda d_{il} &= 0 \\
k_{il}(d_{il} - 1) + (\Sigma k - k_{il})\,(d_{il} - 0) + \sum_{j=c+1}^{n} A_{ij}(d_{il} - d_{jl}) + \lambda d_{il} &= 0 \\
(1+\epsilon)\Sigma k d_{il} - k_{il} + \epsilon\Sigma k + \lambda d_{il} &\geq 0 \\
\frac{k_{il} - \epsilon\Sigma k}{(1+\epsilon)\Sigma k + \lambda} &\leq d_{il}.
\end{aligned}
\tag{20}
$$

Similar to Eq. (15), we also have

$$
d_{il} \leq \frac{k_{il} + \epsilon\Sigma k}{(1+\epsilon)\Sigma k + \lambda}, \tag{21}
$$

so finally we have

$$
\frac{k_{il} - \epsilon\Sigma k}{(1+\epsilon)\Sigma k + \lambda} \leq d_{il} \leq \frac{k_{il} + \epsilon\Sigma k}{(1+\epsilon)\Sigma k + \lambda}. \tag{22}
$$

Combining Eq. (16) with Eq. (22), we can get

$$
\begin{aligned}
|d_{g(il)} - d_{il}| &\leq \frac{k_{il} - \rho + \epsilon\Sigma k}{\Sigma k} - \frac{k_{il} - \epsilon\Sigma k}{(1+\epsilon)\Sigma k + \lambda} \\
&\leq \frac{k_{il} - \rho + \epsilon\Sigma k}{(1+\epsilon)\Sigma k + \lambda} - \frac{k_{il} - \epsilon\Sigma k}{(1+\epsilon)\Sigma k + \lambda} \\
&= \frac{2\epsilon\Sigma k + \rho}{(1+\epsilon)\Sigma k + \lambda} \\
&\leq \frac{2\epsilon}{(1+\epsilon) + \frac{\lambda}{\Sigma k}}.
\end{aligned} \tag{23}
$$

So in case 2, we have

$$
\mathcal{L}_2 \leq \frac{2ncp_2\epsilon}{(1+\epsilon) + \frac{\lambda}{\Sigma k}}. \tag{24}
$$

Case 3: The i-th row is a weak sample, and it's mainly related with weak samples:

Because the number of zeros in the ground-truth label distribution matrix $\mathbf{D}_g$ is large, which indicates the number of strong samples is much more than that of weak samples, so the probability of this case p_3 is very low. In this case, we have

$$
\mathcal{L}_3 \leq n * c * p_3 * 1 = ncp_3. \tag{25}
$$

Adding up $\mathcal{L}_1$, $\mathcal{L}_2$ and $\mathcal{L}_3$, we have the average total difference between $\mathbf{D}$ and $\mathbf{D}_g$, i.e., $\overline{\mathcal{L}} = \frac{\sum_{i,l} |d_{il} - d_{g(il)}|}{nc}$ is upper bounded by

$$
\begin{aligned}
\overline{\mathcal{L}} &= \frac{\mathcal{L}_1 + \mathcal{L}_2 + \mathcal{L}_3}{nc} \\
&\leq \frac{0 + \frac{2ncp_2\epsilon}{(1+\epsilon) + \frac{\lambda}{\Sigma k}} + ncp_3}{nc} \\
&= \frac{2p_2\epsilon}{(1+\epsilon) + \frac{\lambda}{\Sigma k}} + p_3.
\end{aligned} \tag{26}
$$

Because ϵ, p_3 are small numbers and $\lambda \ll \Sigma k$, the average total loss $\overline{\mathcal{L}}$ tends to zero, which proves that our algorithm can precisely recover the ground-truth label distribution under certain specific assumptions.

C Datasets and Evaluation Metrics

Table S1: Statistics of the six datasets

Dataset (abbr.)	# Instances	# Features	# Labels
Natural Scene [1] (NS)	1200	294	9
SCUT_FBP [7] (SCUT)	1500	300	5
RAF_ML [6] (RAF)	4908	200	6
SCUT-FBP5500 [2] (FBP)	5500	512	5
Ren-Cecps [4] (REN)	2000	100	8
Twitter_LDL [8] (Twitter)	6027	200	8

The basic statistics of the six datasets are shown in Table S1, and these datasets are publicly available at http://palm.seu.edu.cn/xgeng/LDL/index.htm#data, http://www.hcii-lab.net/data/SCUT-FBP/EN/download.html. Table S2 shows the formulas of the four evaluation metrics, where $Cor(i, j)$

Table S2: Formulas of the four evaluation metrics.

Measure	Formula
Chebyshev↓	$Dis_1(D, \hat{D}) = \max_j \lvert d_j - \hat{d}_j \rvert$
Clark↓	$Dis_2(D, \hat{D}) = \sqrt{\sum_{j=1}^{c} \frac{(d_j - \hat{d}_j)^2}{(d_j + \hat{d}_j)^2}}$
One-error↓	$Dis_3(D, \hat{D}) = \sum_{i=1}^{n} \sum_{j=1}^{c} Cor(i, j)$
Intersection↑	$Sim_1(D, \hat{D}) = \sum_{j=1}^{c} \min(d_j - \hat{d}_j)$

is formulated as

$$Cor(i, j) = \begin{cases} 1, & D(i, j) \leq \delta \text{ and } \hat{D}(i, j) \leq \delta, \\ 1, & D(i, j) > \delta \text{ and } \hat{D}(i, j) > \delta, \\ 0, & \text{otherwise.} \end{cases} \tag{27}$$

Here, δ is a threshold value fixed to 0.01.

D Full Result Tables

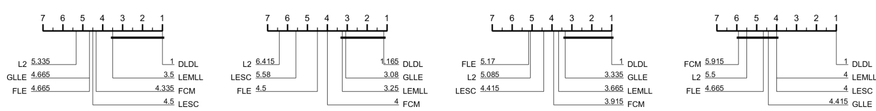
In this section, we perform ten-fold cross-validation and present the full table of average recovery and predictive results with standard deviation (std). The ordering rule is that the mean value takes precedence, and if the mean values are the same then the one with smaller standard deviation is ranked higher. From Table S3 and S4, we can see that our algorithm DLDL ranks first in all cases of the recovery results and in most cases of the predictive results, which clearly shows the superiority of our algorithm over other baseline algorithms.

Table S3: The full recovery results with standard deviation of testing instances on the six datasets and the best average rank (i.e., Avg.Rank) is shown in boldface.

Method	NS	SCUT	RAF	FBP	REN	Twitter	Avg.Rank
				Chebyshev↓			
DLDL	**0.0845±0.0324(1)**	**0.2821±0.0157(1)**	**0.3133±0.0115(1)**	**0.2783±0.0271(1)**	**0.0306±0.0040(1)**	**0.2989±0.0946(1)**	**1.00(1)**
L^2	0.3556±0.0157(6)	0.3818±0.0033(7)	0.3837±0.0017(5)	0.3966±0.0006(7)	0.6445±0.0168(3)	0.5415±0.0254(7)	5.83(6)
FLE	0.3496±0.0094(5)	0.3780±0.0068(6)	0.3901±0.0026(7)	0.3863±0.0006(5)	0.6637±0.0043(4)	0.3310±0.0012(3)	5.00(5)
GLLE	0.3257±0.0264(2)	0.3466±0.0119(2)	0.3801±0.0114(3)	0.3630±0.0007(2)	0.6686±0.0075(5)	0.4710±0.0051(4)	3.00(2)
LEMLL	0.3291±0.0332(3)	0.3527±0.0183(3)	0.3699±0.0048(2)	0.3897±0.0026(6)	0.6369±0.0063(2)	0.3271±0.0004(2)	3.00(2)
LESC	0.3601±0.0089(7)	0.3665±0.0013(5)	0.3845±0.0021(6)	0.3790±0.0007(4)	0.6746±0.0072(7)	0.5105±0.0030(6)	5.83(6)
FCM	0.3466±0.0084(4)	0.3596±0.0024(4)	0.3825±0.0013(4)	0.3638±0.0004(3)	0.6725±0.0024(6)	0.5064±0.0016(5)	4.33(4)
				Clark↓			
Method	NS	SCUT	RAF	FBP	REN	Twitter	Avg.Rank
DLDL	**2.3668±0.0173(1)**	**0.9538±0.0269(1)**	**1.0991±0.0702(1)**	**1.0588±0.0806(1)**	**0.8568±0.0010(1)**	**1.2227±0.0617(1)**	**1.00(1)**
L^2	2.4620±0.0112(5)	1.4968±0.0051(7)	1.5966±0.0064(2)	1.5060±0.0007(6)	2.6508±0.0068(3)	2.4016±0.0107(7)	5.00(5)
FLE	2.4682±0.0078(6)	1.4949±0.0049(6)	1.6153±0.0079(7)	1.5011±0.0015(5)	2.6574±0.0033(4)	2.3846±0.0026(6)	5.67(7)
GLLE	2.4285±0.0282(3)	1.4722±0.0147(2)	1.6100±0.0115(6)	1.4787±0.0010(3)	2.6598±0.0047(5)	2.3669±0.0035(3)	3.67(3)
LEMLL	2.4279±0.0316(2)	1.4855±0.0209(5)	1.6049±0.0117(3)	1.6214±0.0020(7)	2.6420±0.0034(2)	2.3647±0.0014(2)	3.50(2)
LESC	2.4751±0.0080(7)	1.4847±0.0032(4)	1.6063±0.0007(5)	1.4904±0.0022(4)	2.6650±0.0048(7)	2.3844±0.0046(5)	5.33(6)
FCM	2.4540±0.0123(4)	1.4778±0.0029(3)	1.6055±0.0035(4)	1.4770±0.0016(2)	2.6642±0.0020(6)	2.3832±0.0020(4)	3.83(4)
				One-error↓			
Method	NS	SCUT	RAF	FBP	REN	Twitter	Avg.Rank
DLDL	**0.0000±0.0592(1)**	**0.0037±0.0111(1)**	**0.0189±0.0335(1)**	**0.0912±0.0892(1)**	**0.0000±0.0000(1)**	**0.0859±0.0875(1)**	**1.00(1)**
L^2	0.6119±0.0197(3)	0.2743±0.0107(6)	0.2810±0.0044(2)	0.2765±0.0012(4)	0.8157±0.0063(7)	0.6578±0.0066(6)	4.67(5)
FLE	0.6364±0.0044(6)	0.2693±0.0032(5)	0.2904±0.0040(7)	0.2754±0.0040(3)	0.8130±0.0020(6)	0.6569±0.0014(3)	5.00(7)
GLLE	0.6259±0.0693(4)	0.2663±0.0153(3)	0.2879±0.0117(5)	0.2774±0.0006(5)	0.8109±0.0027(4)	0.6575±0.0015(5)	4.33(3)
LEMLL	0.6348±0.0188(5)	0.2679±0.0035(4)	0.2883±0.0146(6)	0.3187±0.0016(7)	0.7377±0.0023(2)	0.6161±0.0004(2)	4.33(3)
LESC	0.6386±0.0043(7)	0.2894±0.0012(7)	0.2860±0.0005(3)	0.2752±0.0028(2)	0.8117±0.0031(5)	0.6569±0.0020(4)	4.67(5)
FCM	0.5968±0.0062(2)	0.2635±0.0020(2)	0.2874±0.0017(4)	0.2780±0.0011(6)	0.7943±0.0004(3)	0.6595±0.0012(7)	4.00(2)
				Intersection↑			
Method	NS	SCUT	RAF	FBP	REN	Twitter	Avg.Rank
DLDL	**0.9082±0.0387(1)**	**0.6894±0.0180(1)**	**0.6298±0.0119(1)**	**0.6987±0.0274(1)**	**0.9694±0.0041(1)**	**0.6985±0.0913(1)**	**1.00(1)**
L^2	0.4123±0.0097(4)	0.5068±0.0029(5)	0.4976±0.0103(6)	0.5014±0.0017(7)	0.2430±0.0045(2)	0.3444±0.0042(7)	5.17(6)
FLE	0.3907±0.0044(5)	0.5197±0.0104(4	0.4949±0.0034(7)	0.5195±0.0080(6)	0.2133±0.0021(3)	0.5023±0.0123(3)	4.67(4)
GLLE	0.4586±0.0548(3)	0.5623±0.0151(2)	0.5077±0.0146(3)	0.5466±0.0010(2)	0.2046±0.0039(4)	0.4264±0.0062(4)	3.00(2)
LEMLL	0.4739±0.0323(2)	0.4635±0.0051(6)	0.5280±0.0106(2)	0.5368±0.0011(4)	0.1826±0.0028(7)	0.5866±0.0006(2)	3.83(3)
LESC	0.3885±0.0050(6)	0.5350±0.0018(3)	0.5005±0.0004(5)	0.5228±0.0018(5)	0.1948±0.0177(5)	0.3726±0.0086(6)	5.33(7)
FCM	0.3724±0.0062(7)	0.4319±0.0028(7)	0.5011±0.0013(4)	0.5437±0.0007(3)	0.1852±0.0008(6)	0.3826±0.0089(5)	4.67(4)

Table S4: The full predictive results with standard deviation of testing instances on the six datasets and the best average rank (i.e., Avg.Rank) is shown in boldface.

Method	NS	SCUT	RAF	FBP	REN	Twitter	Avg.Rank
			Chebyshev↓				
DLDL	0.4071±0.0121(2)	0.4086±0.0205(2)	**0.3911±0.2550(1)**	**0.2972±0.0273(1)**	**0.6164±0.0217(1)**	**0.3657±0.0505(1)**	**1.33(1)**
L^2	0.4967±0.0129(7)	0.4468±0.0902(7)	0.4064±0.0047(7)	0.4118±0.0067(7)	0.6913±0.0366(7)	0.5444±0.0323(7)	7.00(7)
FLE	**0.4011±0.0309(1)**	0.4254±0.0121(6)	0.3934±0.0075(3)	0.3915±0.0017(6)	0.6897±0.0106(6)	0.4330±0.0014(2)	4.00(5)
GLLE	0.4257±0.0264(3)	**0.4006±0.0105(1)**	0.3968±0.0334(4)	0.3641±0.0069(3)	0.6833±0.0036(4)	0.4807±0.0011(4)	3.16(2)
LEMLL	0.4490±0.0237(4)	0.4148±0.0178(5)	0.3975±0.0106(5)	0.3286±0.0537(2)	0.6321±0.0188(2)	0.4706±0.0032(3)	3.50(3)
LESC	0.4743±0.0147(6)	0.4143±0.0111(4)	0.4008±0.0025(6)	0.3795±0.0010(5)	0.6873±0.0408(5)	0.5138±0.0105(6)	5.33(6)
FCM	0.4692±0.0298(5)	0.4132±0.0019(3)	0.3931±0.0022(2)	0.3689±0.0020(4)	0.6736±0.0132(3)	0.5071±0.0044(5)	3.67(4)
			Clark↓				
DLDL	**2.5059±0.0132(1)**	**1.5321±0.0264(1)**	**1.6165±0.3860(1)**	**1.1913±0.1544(1)**	**2.6403±0.0089(1)**	**2.5688±0.0870(1)**	**1.00(1)**
L^2	2.5111±0.0224(4)	1.6470±0.1481(7)	1.6356±0.0034(7)	1.5491±0.0110(7)	2.6723±0.0145(6)	2.4084±0.0165(3)	5.67(7)
FLE	2.5062±0.0319(2)	1.5426±0.0069(3)	1.6246±0.0130(5)	1.5013±0.0040(5)	2.6735±0.0076(7)	2.3951±0.0017(4)	4.33(4)
GLLE	2.5203±0.0212(6)	1.6169±0.0226(6)	1.6242±0.2832(4)	1.4785±0.0044(2)	2.6686±0.0011(4)	2.3764±0.0027(7)	4.83(6)
LEMLL	2.5278±0.0241(7)	1.5548±0.0387(5)	1.6334±0.2243(6)	1.5188±0.0895(6)	2.6486±0.0118(2)	2.4233±0.0104(2)	4.67(5)
LESC	2.5065±0.0226(3)	1.5353±0.0075(2)	1.6172±0.0052(2)	1.4915±0.0054(4)	2.6714±0.0253(5)	2.3856±0.0056(5)	3.50(2)
FCM	2.5192±0.0189(5)	1.5452±0.0025(4)	1.6216±0.0020(3)	1.4828±0.0013(3)	2.6610±0.0062(3)	2.3832±0.0041(6)	4.00(3)
			One-error↓				
DLDL	**0.5852±0.0122(1)**	**0.2924±0.0224(1)**	**0.2853±0.2453(1)**	**0.1134±0.0731(1)**	**0.1939±0.0066(1)**	**0.3015±0.1135(1)**	**1.00(1)**
L^2	0.6549±0.0211(7)	0.3404±0.0828(7)	0.2909±0.0080(4)	0.3171±0.0086(7)	0.8149±0.0429(4)	0.6605±0.0066(7)	6.00(7)
FLE	0.6525±0.0173(6)	0.3093±0.0053(4)	0.2931±0.0041(6)	0.2754±0.0035(2)	0.8159±0.0044(5)	0.5791±0.0103(3)	4.33(3)
GLLE	0.6401±0.0158(3)	0.3384±0.0206(6)	0.2924±0.1658(5)	0.2772±0.0026(4)	0.8162±0.0009(6))	0.6602±0.0014(6)	5.00(6)
LEMLL	0.6170±0.0240(2)	0.3030±0.0228(2)	0.2870±0.1646(2)	0.2919±0.0498(6)	0.8090±0.0101(2)	0.5539±0.0201(2)	2.67(2)
LESC	0.6431±0.0111(5)	0.3073±0.0047(3)	0.2902±0.0026(3)	0.2770±0.0057(3)	0.8214±0.0042(7)	0.6589±0.0082(5)	4.33(3)
FCM	0.6417±0.0114(4)	0.3151±0.0024(5)	0.2941±0.0015(7)	0.2775±0.0039(5)	0.8132±0.0044(3)	0.6177±0.0020(4)	4.67(5)
			Intersection↑				
DLDL	0.4347±0.0127(2)	**0.4825±0.0145(1)**	**0.4864±0.2367(1)**	**0.6523±0.0356(1)**	**0.2719±0.0078(1)**	**0.6025±0.0215(1)**	**1.17(1)**
L^2	0.3695±0.0155(5)	0.3871±0.0924(7)	0.4689±0.0133(6)	0.5482±0.0135(3)	0.1852±0.0019(7)	0.3387±0.0023(7)	5.83(7)
FLE	0.3530±0.0175(7)	0.4810±0.0091(3)	0.4802±0.0068(3)	0.5087±0.0015(7)	0.1864±0.0048(6)	0.3523±0.0012(4)	5.00(6)
GLLE	0.4002±0.0170(3)	0.4250±0.0179(6)	0.4643±0.0119(5)	0.5455±0.0025(4)	0.1974±0.0010(4)	0.4117±0.0016(2)	4.00(4)
LEMLL	**0.4412±0.0173(1)**	0.4514±0.0150(5)	0.4776±0.1775(4)	0.5929±0.0232(2)	0.2636±0.0100(2)	0.3392±0.0023(6)	3.33(2)
LESC	0.3772±0.0108(4)	0.4679±0.0111(4)	0.4703±0.0029(5)	0.5238±0.0009(6)	0.1910±0.0075(5)	0.3770±0.0102(3)	4.50(5)
FCM	0.3583±0.0127(6)	0.4818±0.0045(2)	0.4824±0.0005(2)	0.5389±0.0012(5)	0.2070±0.0037(3)	0.3408±0.0151(5)	3.83(3)



(a) Chebyshev (b) Clark (c) One-error (d) Intersection

Figure S1: Comparison of DLDL against other six methods with the Bonferroni-Dunn test (CD = 2.6002 at 0.05 significance level).

E Significance Test

In this subsection, we use the Bonferroni–Dunn test at the 0.05 significance level to test whether DLDL achieves significantly better performance compared to other algorithms. Specifically, we combine the recovery results with the predictive results to conduct the Bonferroni-Dunn test, that is, the number of algorithms is 7 and the number of datasets is considered as 12 (2 groups of experiments × 6 datasets). Then, we use DLDL as the control algorithm with a critical difference (CD) to correct for the mean level difference with the comparison algorithms.

The results are shown in Fig. S1, where the algorithms not connected with DLDL are considered to have significantly different performance from the control algorithms. It is impressive that DLDL achieves the lowest rank in terms of all evaluation metrics and the effectiveness of it is also more significant than L^2, LESC and FCM based on Chebyshev, Clark, One-error and Intersection.

F Predictive Results based on An Additional LDL Model

In the main body of the paper, we use SA-BFGS as the LDL model to generate the LDs of the testing samples. In Table S5, we apply an additional LDL model named LDLSF [5] to predict the LDs of testing instances based on the recovered LDs by performing the five baseline LE methods. Here DLDL and L^2 can directly predict the LDs without an external LDL model. From the table, we can clearly see that although we apply a different LDL model, the predictive results of DLDL is still relatively good in average rank. Specifically, out of the 24 statistical comparisons, DLDL ranks 1st in 70.8% cases and ranks 2nd in 20.8% cases. In addition, DLDL performs much better than L^2 in terms of the three evaluation metrics (i.e., Clark, Canberra and Intersection). These results clearly validate the effectiveness and superiority of DLDL in directly training an LDL model from the logical labels.

Table S5: The new predictive results of testing instances on the six datasets and the best average rank (i.e., Avg.Rank) is shown in boldface.

Method	Chebyshev↓						Avg.Rank
	NS	SCUT	RAF	FBP	REN	Twitter	
LDLSF	0.3555	0.4149	0.1583	0.1563	0.5871	0.3532	-
DLDL	**0.3474(1)**	**0.4021(1)**	**0.3009(1)**	0.3113(2)	0.6451(2)	**0.2622(1)**	**1.33(1)**
L^2	0.3903(7)	0.4223(6)	0.3774(6)	**0.2142(1)**	0.6861(7)	0.5332(7)	5.67(7)
FLE	0.3483(2)	0.4133(2)	0.3948(5)	0.3660(3)	0.6719(3)	0.2968(2)	2.83(2)
GLLE	0.3579(5)	0.4155(5)	0.3387(3)	0.3500(4)	0.6707(5)	0.4650(3)	4.17(4)
LEMLL	0.3502(3)	0.4152(3)	0.3043(4)	0.3173(6)	0.6554(4)	0.3767(4)	4.00(3)
LESC	0.3603(4)	0.4111(4)	0.3491(2)	0.3688(7)	0.6629(6)	0.5043(5)	4.67(5)
FCM	0.3758(6)	0.4124(6)	0.3892(7)	0.3698(5)	**0.3975(1)**	0.5082(6)	5.17(6)
Method	Clark↓						Avg.Rank
	NS	SCUT	RAF	FBP	REN	Twitter	
LDLSF	2.4272	1.5641	1.4523	1.3402	2.6407	2.5577	-
DLDL	**2.4740(1)**	**1.5400(1)**	**1.5174(1)**	1.4690(2)	**1.3231(1)**	2.4274(7)	**2.00(1)**
L^2	2.5042(7)	1.5856(7)	1.6100(6)	**1.3910(1)**	2.6680(7)	2.3995(6)	5.67(7)
FLE	2.4935(6)	1.5451(5)	1.6344(5)	1.4783(3)	2.6630(4)	2.3845(5)	4.67(5)
GLLE	2.4734(3)	1.5411(2)	1.5618(3)	1.4646(4)	2.6623(5)	**2.3658(1)**	3.00(2)
LEMLL	2.4765(2)	1.5477(2)	1.5192(4)	1.4329(6)	2.6560(3)	2.3544(2)	3.17(3)
LESC	2.4776(4)	1.5366(4)	1.5752(2)	1.4825(7)	2.6556(6)	2.3822(3)	4.33(4)
FCM	2.4863(5)	1.5374(6)	1.6272(7)	1.4782(5)	1.6440(2)	2.3850(4)	4.83(6)
Method	One-error↓						Avg.Rank
	NS	SCUT	RAF	FBP	REN	Twitter	
LDLSF	0.4764	0.2940	0.2461	0.2718	0.7898	0.6075	-
DLDL	**0.5352(1)**	**0.2940(1)**	**0.2899(1)**	0.2727(2)	**0.1994(1)**	**0.5012(1)**	**1.17(1)**
L^2	0.6449(3)	**0.2940(1)**	0.2948(2)	**0.2609(1)**	0.8134(2)	0.6604(3)	2.00(2)
FLE	0.5931(2)	0.2960(5)	0.2948(2)	0.2727(2)	0.8144(3)	0.5676(2)	2.67(5)
GLLE	0.6449(3)	0.2960(5)	0.2948(2)	0.2727(2)	0.8144(3)	0.6604(3)	3.00(6)
LEMLL	0.6454(7)	0.3020(7)	0.2950(7)	0.2727(2)	0.8144(3)	0.6604(3)	4.83(7)
LESC	0.6449(3)	**0.2940(1)**	0.2948(2)	0.2727(2)	0.8144(3)	0.6604(3)	2.33(3)
FCM	0.6449(3)	**0.2940(1)**	0.2948(2)	0.2727(2)	0.8144(3)	0.6604(3)	2.33(3)
Method	Intersection↑						Avg.Rank
	NS	SCUT	RAF	FBP	REN	Twitter	
LDLSF	0.5140	0.4823	0.8030	0.1233	0.3269	0.6035	-
DLDL	**0.5178(1)**	**0.5011(1)**	**0.6341(1)**	0.6367(3)	0.3542(2)	**0.6775(1)**	**1.50(1)**
L^2	0.4021(6)	0.4790(7)	0.5129(6)	**0.7593(1)**	0.1907(7)	0.3515(7)	5.67(7)
FLE	0.5031(2)	0.4840(4)	0.5398(5)	0.5615(5)	0.2063(5)	0.6324(2)	3.83(3)
GLLE	0.4260(4)	0.4832(5)	0.5727(3)	0.5859(4)	0.2081(4)	0.4366(4)	4.00(4)
LEMLL	0.4700(3)	0.4820(6)	0.6255(2)	0.6385(2)	0.2292(3)	0.5363(3)	3.17(2)
LESC	0.4094(5)	0.4881(2)	0.5571(4)	0.5600(6)	0.2055(6)	0.3911(5)	4.67(5)
FCM	0.3849(7)	0.4870(3)	0.4915(7)	0.5405(7)	**0.4853(1)**	0.3820(6)	5.17(6)

References

[1] Xin Geng, Renyi Zheng, Jiaqi Lv, and Yu Zhang. Multilabel ranking with inconsistent rankers. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 44(09):5211–5224, 2022.

[2] Lingyu Liang, Luojun Lin, Lianwen Jin, Duorui Xie, and Mengru Li. SCUT-FBP5500: A diverse benchmark dataset for multi-paradigm facial beauty prediction. In *24th International Conference on Pattern Recognition*, pages 1598–1603. IEEE Computer Society, 2018.

[3] A. Talwalkar M. Mohri, A. Rostamizadeh. *Foundations of Machine Learning*. MIT Press, Cambridge, Massachusetts, USA, 2018.

[4] Changqin Quan and Fuji Ren. Construction of a blog emotion corpus for chinese emotional expression analysis. In *Proceedings of the 2009 conference on empirical methods in natural language processing*, pages 1446–1454, 2009.

[5] Tingting Ren, Xiuyi Jia, Weiwei Li, Lei Chen, and Zechao Li. Label distribution learning with label-specific features. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*, pages 3318–3324, 2019.

[6] Li Shang and Weihong Deng. Blended emotion in-the-wild: Multi-label facial expression recognition using crowdsourced annotations and deep locality feature learning. *International Journal of Computer Vision*, 127(6-7):884–906, 2019.

[7] Duorui Xie, Lingyu Liang, Lianwen Jin, Jie Xu, and Mengru Li. Scut-fbp: A benchmark dataset for facial beauty perception. In *2015 IEEE International Conference on Systems, Man, and Cybernetics*, pages 1821–1826, 2015.

[8] Jufeng Yang, Ming Sun, and Xiaoxiao Sun. Learning visual sentiment distributions via augmented conditional probability neural network. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, pages 224–230. AAAI Press, 2017.