Label Enhancement for Label Distribution Learning

Ning Xu, An Tao and Xin Geng
Southeast University,
xning@seu.edu.cn

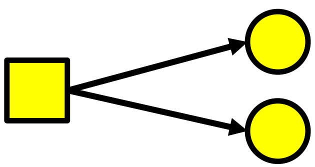
Outline

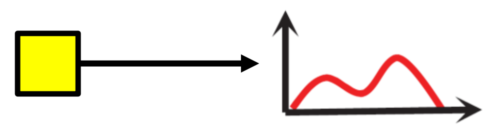
- **Introduction**
- **Label Enhancement**
- **Experiments**
- **Conclusion**

Introduction

Logical labels: answers the essential question "which label can describe the instance".

Label distribution: involves the explicit relative importance of each label.



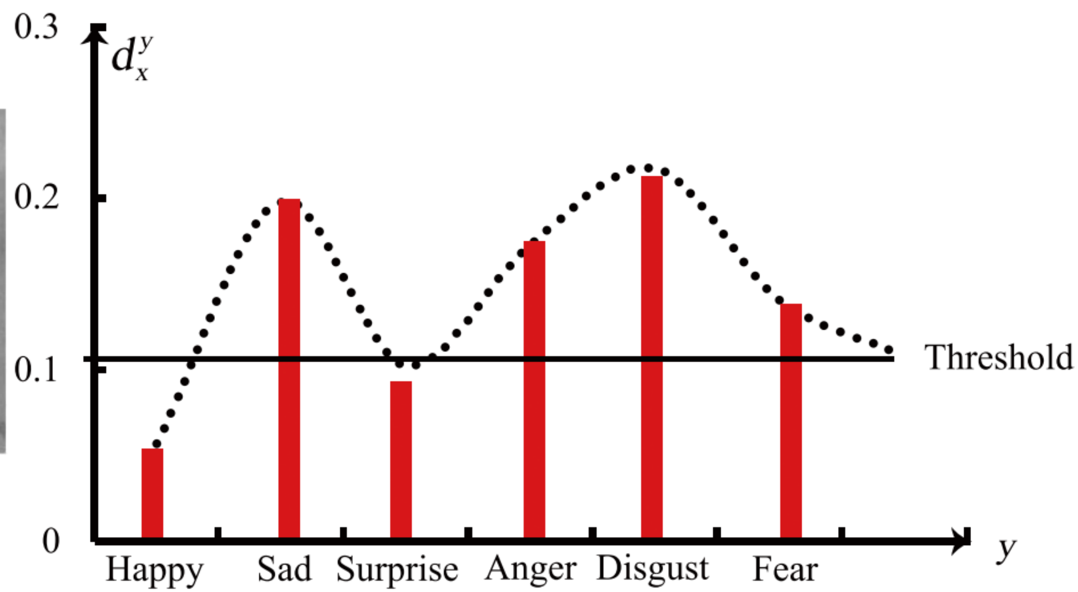


Multi-label learning

Label distribution learning

Introduction

- Label distribution is more general than logical labels.
  - The differentiation between the relevant and irrelevant labels is relative.



Introduction

- Label distribution is more general than logical labels.
  - When multiple labels are associated with an instance, the relative importance among them is more likely to be different rather than exactly equal.

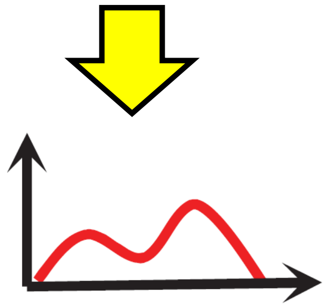
Introduction

- Label distribution is more general than logical labels.
  - The “irrelevance” of each irrelevant label may be very

Introduction

- The process of quantifying the description degree is:
  - **Costly**
  - **Difficult**
- A bipartite partition of the label set into relevant and irrelevant labels with respect to an instance is actually a simplification of the real problem.
  - 1 : relevant label
  - 0 : irrelevant label

{0, 1, 0, 1, 0}



We need a way to recover the label distributions from the logical labels in the training set

Label enhancement (LE)

Formulation of Label Enhancement

The logical label vector of $\boldsymbol{x}_i$ is denoted by $\boldsymbol{l}_i = \left(l_{\boldsymbol{x}_i}^{y_1}, l_{\boldsymbol{x}_i}^{y_2}, \ldots, l_{\boldsymbol{x}_i}^{y_c}\right)^{\mathrm{T}}$, where

$l_{\boldsymbol{x}_i}^{y_j} \in \{0,1\}$ represents whether y_j describes $\boldsymbol{x}_i$.

The label distribution of $\boldsymbol{x}_i$ is denoted by $\boldsymbol{d}_i = \left(d_{\boldsymbol{x}_i}^{y_1}, d_{\boldsymbol{x}_i}^{y_2}, \ldots, d_{\boldsymbol{x}_i}^{y_c}\right)^{\mathrm{T}}$, where

$d_{\boldsymbol{x}_i}^{y_j} \in [0,1]$ represents the description degree of y_j to $\boldsymbol{x}_i$.

LE can be defined as follows.

Given a training set $S = \{(\boldsymbol{x}_i, \boldsymbol{l}_i) | 1 \leq i \leq n\}$, LE recovers the label distribution $\boldsymbol{d}_i$ of $\boldsymbol{x}_i$ from the logical label vector $\boldsymbol{l}_i$, and thus transforms S into a LDL training set $E = \{(\boldsymbol{x}_i, \boldsymbol{d}_i) | 1 \leq i \leq n\}$.

Existing Label Enhancement Algorithms

- **Fuzzy Label Enhancement**
  - The LE algorithm based on fuzzy clustering (FCM)
    [Gayar et al., ANNPR'06]
  - The LE algorithm based on kernel method (KM)
    [Jiang et al., NCA, 2006]
- **Graph-based Label Enhancement**
  - The LE algorithm based on label propagation (LP)
    [Li et al., ICDM’15]
  - The LE algorithm based on manifold learning (ML)
    [Hou et al., AAAI’16]

The LE algorithm based on fuzzy clustering（FCM）

[Gayar et al., ANNPR'06]

Fuzzy C-means clustering

$$m_{\boldsymbol{x}_i}^k = \frac{1}{\sum_{j=1}^{p}\left(\frac{Dist(\boldsymbol{x}_i,\boldsymbol{\mu}_k)}{Dist(\boldsymbol{x}_i,\boldsymbol{\mu}_j)}\right)^{\frac{1}{\beta-1}}}$$

Connection matrix *A*

$$\boldsymbol{A}_j = \boldsymbol{A}_j + \boldsymbol{m}_{\boldsymbol{x}_i}, if\, l_{\boldsymbol{x}_i}^{y_j} = 1$$

Fuzzy composition

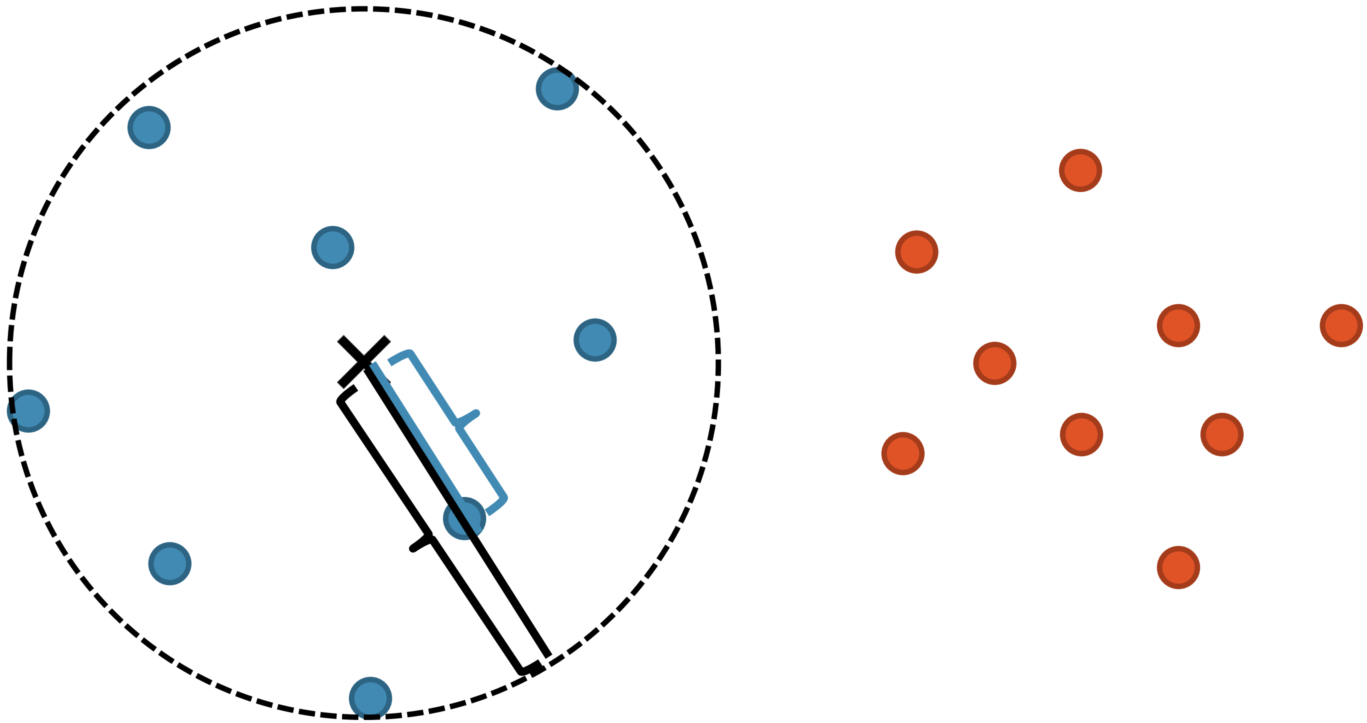
$$\tilde{\boldsymbol{d}}_i = \boldsymbol{A} \circ \boldsymbol{m}_{\boldsymbol{x}_i}^{\top}$$

Softmax normalization

$$d_{\boldsymbol{x}_i}^{y} = \frac{e^{\tilde{d}_{\boldsymbol{x}_i}^{y}}}{\sum_y e^{\tilde{d}_{\boldsymbol{x}_i}^{y}}}$$

The LE algorithm based on kernel method（KM）

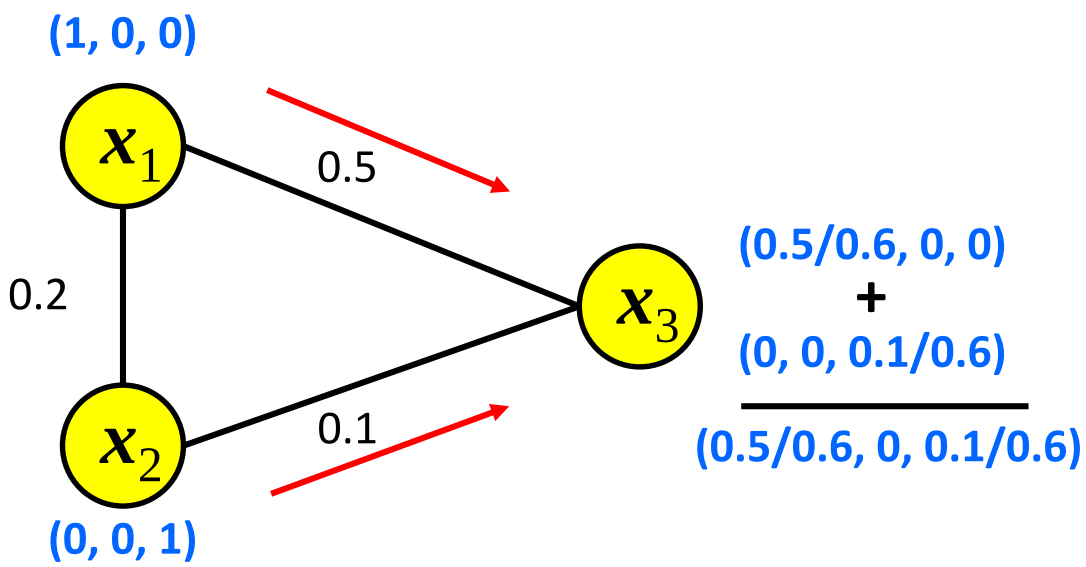
[Jiang et al., NCA, 2006]



Inducing nonlinear relationship via kernel method

The LE algorithm based on label propagation （LP）

[Li, Zhang and Geng, ICDM’15]

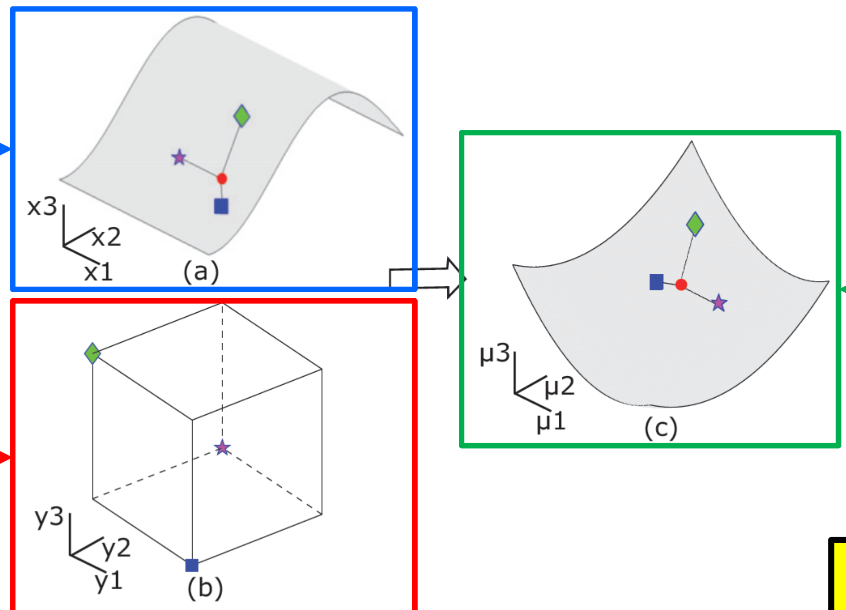


Label Propagation in training set

The LE algorithm based on manifold learning （ML）

[Hou, Geng and Zhang, AAAI’16]

- Feature space：continuous Euclidean space
- Label space：discrete logical space



smoothness assumption

The local topological structure is transferred from the feature space to the label space.

Graph Laplacian Label Enhancement (GLLE)

- Model

a nonlinear transformation

$$\boldsymbol{D}_i = \boldsymbol{W}^{\top}\varphi(\boldsymbol{x}_i) + \boldsymbol{b} = \widehat{\boldsymbol{W}}\boldsymbol{\phi}_i$$

$\widehat{\boldsymbol{W}} = [\boldsymbol{W}^{\mathbf{T}}, \boldsymbol{b}], \boldsymbol{\phi}_i = [\boldsymbol{\varphi}(\boldsymbol{x}_i); \mathbf{1}]$

Goal Determining the best parameter $\widehat{\boldsymbol{W}}^*$

information

came from

- Original logical labels
- Feature space

- Target function

Original logical labels

$$\min_{\widehat{\boldsymbol{W}}} L\left(\widehat{\boldsymbol{W}}\right) + \lambda\Omega(\widehat{\boldsymbol{W}})$$

Leveraging the topological information
of the feature space

Graph Laplacian Label Enhancement (GLLE)

- The first part of the target function

$$L(\widehat{\boldsymbol{W}}) = \sum_{i=1}^{n} \left\| \widehat{\boldsymbol{W}} \boldsymbol{\phi}_i - \boldsymbol{L}_i \right\|^2$$

least squares (LS)

- The second part of the target function

smoothness assumption

$$\Omega(\widehat{\boldsymbol{W}}) = \sum_{i,j} a_{ij} \left\| \boldsymbol{D}_i - \boldsymbol{D}_j \right\|^2$$

$$= \boldsymbol{tr}(\widehat{\boldsymbol{W}} \boldsymbol{\phi} \boldsymbol{G} \boldsymbol{D}^\top \boldsymbol{\phi}^{\mathrm{T}} \widehat{\boldsymbol{W}}^{\mathrm{T}})$$

$$a_{ij} = \begin{cases} \exp\left(-\frac{\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2}{2\sigma^2}\right) & \text{if } \boldsymbol{x}_j \in N(i) \\ 0 & \text{otherwise} \end{cases}$$

Graph Laplacian

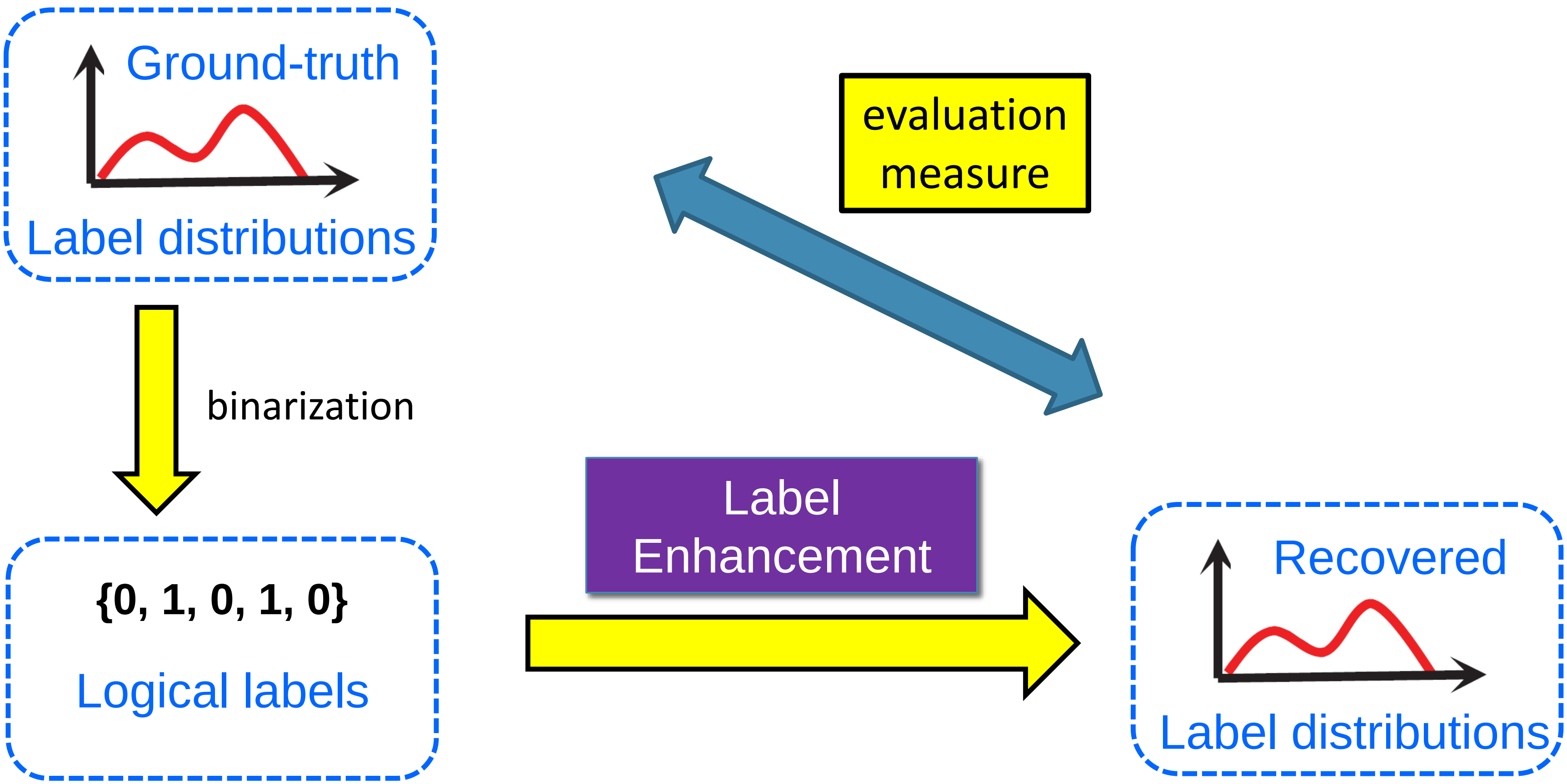
$$\boldsymbol{G} = \widehat{\boldsymbol{A}} - \boldsymbol{A},\ \hat{a}_{ij} = \sum_{j=1}^{n} a_{ij}$$

Experiments

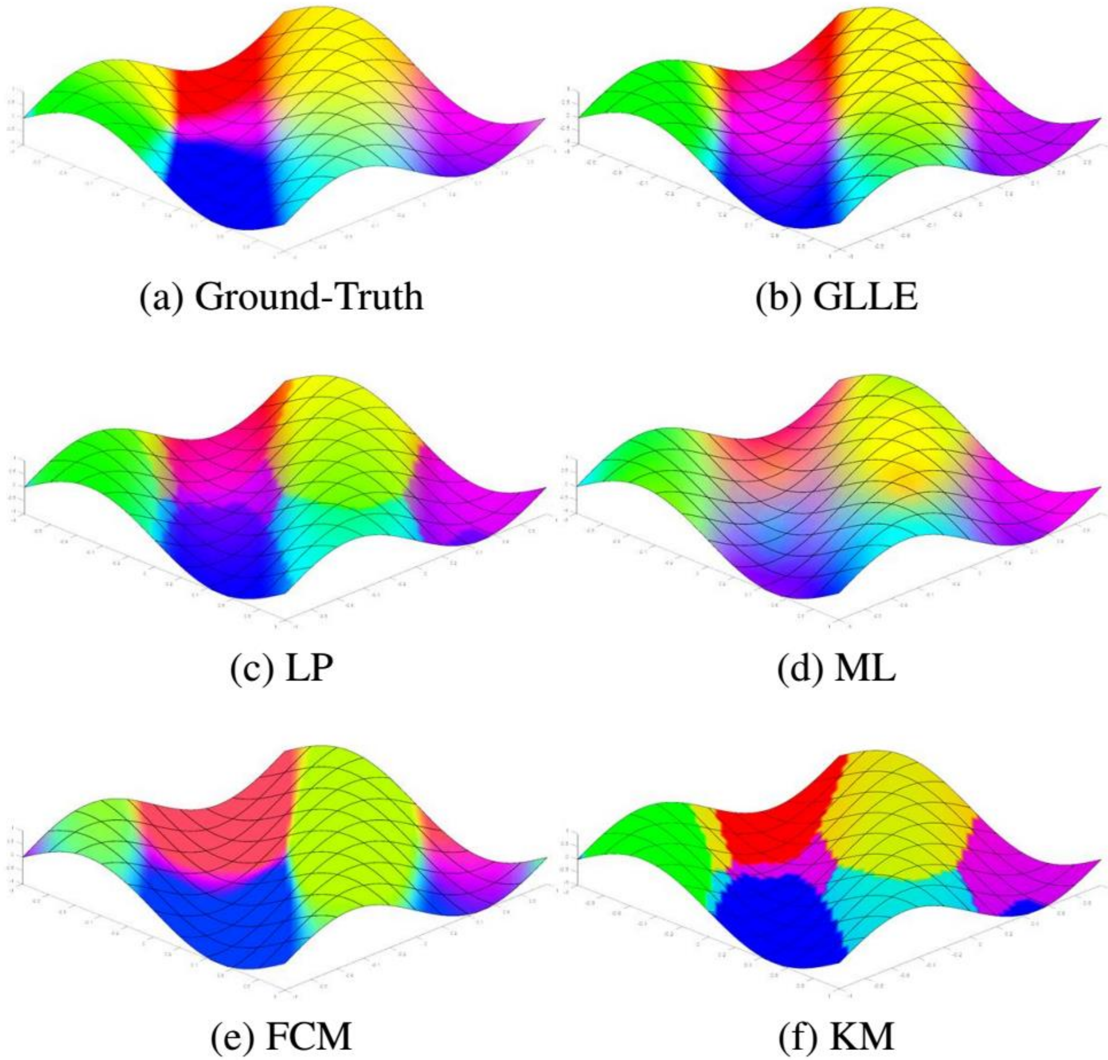
No.	Dataset	#Examples	#Features	#Labels
1	Artificial	2601	3	3
2	SJAFFE	213	243	6
3	Natural Scene	2,000	294	9
4	Yeast-spoem	2,465	24	2
5	Yeast-spo5	2,465	24	3
6	Yeast-dtt	2,465	24	4
7	Yeast-cold	2,465	24	4
8	Yeast-heat	2,465	24	6
9	Yeast-spo	2,465	24	6
10	Yeast-diau	2,465	24	7
11	Yeast-elu	2,465	24	14
12	Yeast-cdc	2,465	24	15
13	Yeast-alpha	2,465	24	18
14	SBU_3DFE	2,500	243	6
15	Movie	7,755	1,869	5

Table 1: Statistics of the 15 Datasets Used in the Experiments

Recovery Performance



Recoverv Performance



(a) Ground-Truth (b) GLLE

(c) LP (d) ML

(e) FCM (f) KM

Figure 2: Comparison between the ground-truth and recovered label distributions (regarded as RGB colors) on the artificial manifold.

Recovery Performance

Datasets	FCM	KM	LP	ML	GLLE
Artificial	0.188(3)	0.260(5)	0.130(2)	0.227(4)	**0.108(1)**
SJAFFE	0.132(3)	0.214(5)	0.107(2)	0.190(4)	**0.100(1)**
Natural Scene	0.368(5)	0.306(4)	**0.275(1)**	0.295(2)	0.296(3)
Yeast-spoem	0.233(3)	0.408(5)	0.163(2)	0.400(4)	**0.108(1)**
Yeast-spo5	0.162(3)	0.277(5)	0.114(2)	0.273(4)	**0.092(1)**
Yeast-dtt	0.097(2)	0.257(5)	0.128(3)	0.244(4)	**0.065(1)**
Yeast-cold	0.141(3)	0.252(5)	0.137(2)	0.242(4)	**0.093(1)**
Yeast-heat	0.169(4)	0.175(5)	0.086(2)	0.165(3)	**0.056(1)**
Yeast-spo	0.130(3)	0.175(5)	0.090(2)	0.171(4)	**0.067(1)**
Yeast-diau	0.124(3)	0.152(5)	0.099(2)	0.148(4)	**0.084(1)**
Yeast-elu	0.052(3)	0.078(5)	0.044(2)	0.072(4)	**0.030(1)**
Yeast-cdc	0.051(3)	0.076(5)	0.042(2)	0.071(4)	**0.038(1)**
Yeast-alpha	0.044(3)	0.063(5)	0.040(2)	0.057(4)	**0.033(1)**
SBU_3DFE	0.135(2)	0.238(5)	**0.123(1)**	0.233(4)	0.141(3)
Movie	0.230(4)	0.234(5)	0.161(2)	0.164(3)	**0.160(1)**
Avg. Rank	3.13	4.93	1.93	3.73	1.27

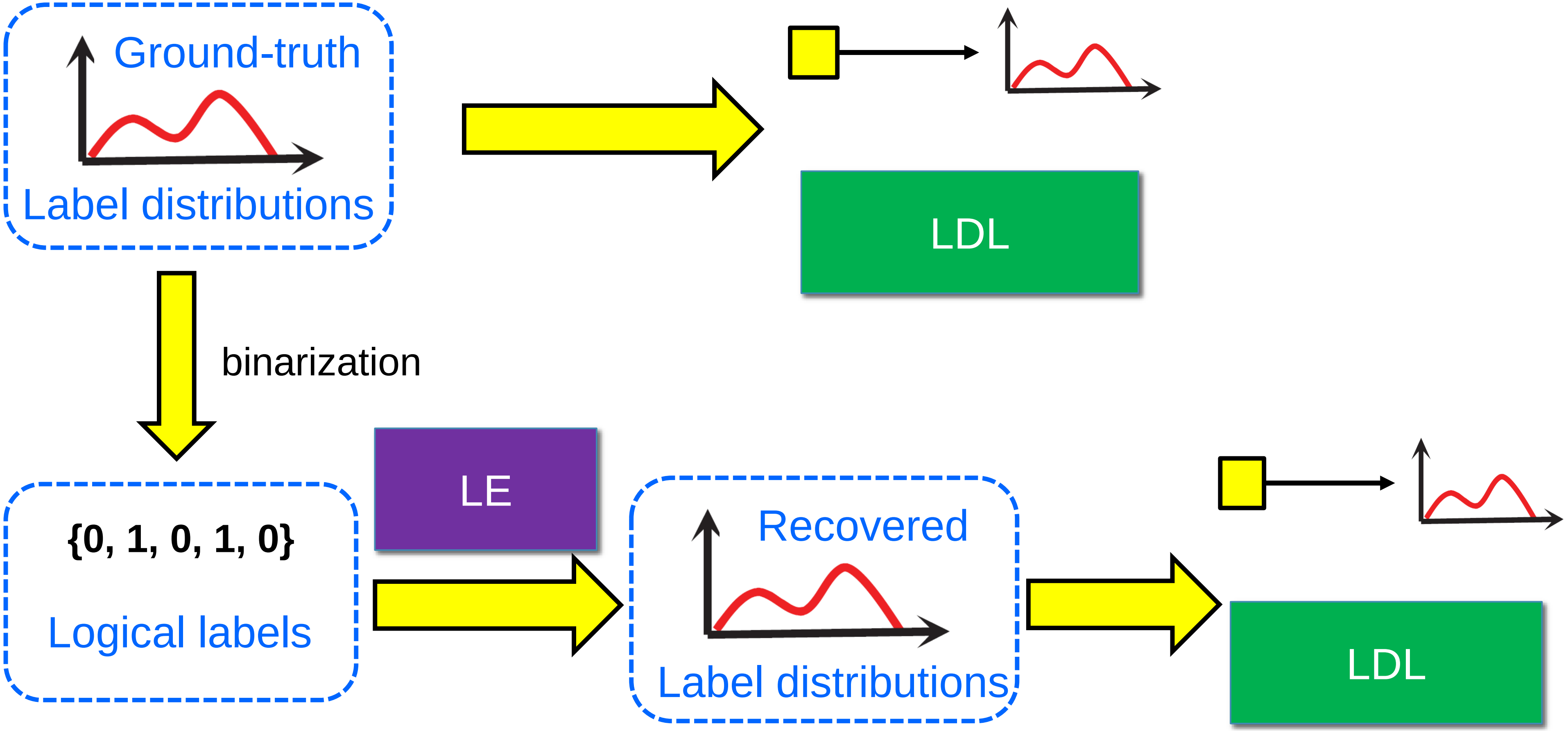
Table 2: Recovery Results (value(rank)) Measured by Cheb $\downarrow$

Recovery Performance

Datasets	FCM	KM	LP	ML	GLLE
Artificial	0.933(3)	0.918(5)	0.974(2)	0.925(4)	**0.980(1)**
SJAFFE	0.906(3)	0.827(5)	0.941(2)	0.857(4)	**0.946(1)**
Natural Scene	0.593(5)	0.748(4)	**0.860(1)**	0.818(2)	0.769(3)
Yeast-spoem	0.878(3)	0.812(5)	0.950(2)	0.815(4)	**0.968(1)**
Yeast-spo5	0.922(3)	0.882(5)	0.969(2)	0.884(4)	**0.974(1)**
Yeast-dtt	0.959(2)	0.759(5)	0.921(3)	0.763(4)	**0.983(1)**
Yeast-cold	0.922(3)	0.779(5)	0.925(2)	0.784(4)	**0.969(1)**
Yeast-heat	0.883(3)	0.779(5)	0.932(2)	0.783(4)	**0.980(1)**
Yeast-spo	0.909(3)	0.800(5)	0.939(2)	0.803(4)	**0.968(1)**
Yeast-diau	0.882(3)	0.799(5)	0.915(2)	0.803(4)	**0.939(1)**
Yeast-elu	0.950(2)	0.758(5)	0.918(3)	0.763(4)	**0.978(1)**
Yeast-cdc	0.929(2)	0.754(5)	0.916(3)	0.759(4)	**0.959(1)**
Yeast-alpha	0.922(2)	0.751(5)	0.911(3)	0.756(4)	**0.973(1)**
SBU_3DFE	0.912(2)	0.812(5)	**0.922(1)**	0.815(4)	0.900(3)
Movie	0.773(5)	0.880(4)	**0.929(1)**	0.919(2)	0.900(3)
Avg. Rank	2.93	4.87	2.07	3.73	1.40

Table 3: Recovery Results (value(rank)) Measured by Cosine ↑

LDL Predictive Performance



LDL Predictive Performance

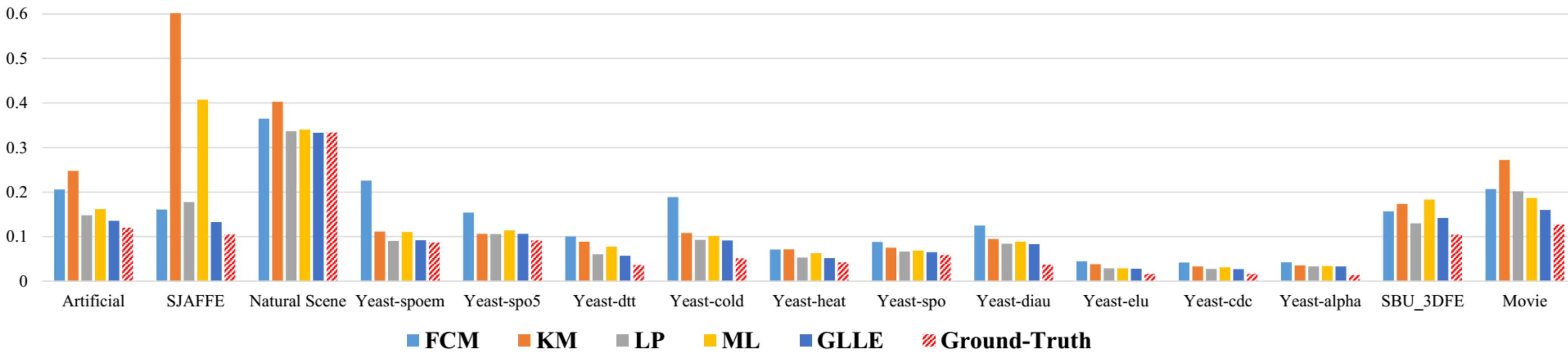


Figure 3: Comparison of the LDL after the LE pre-process against the direct LDL measured by Cheb ↓.

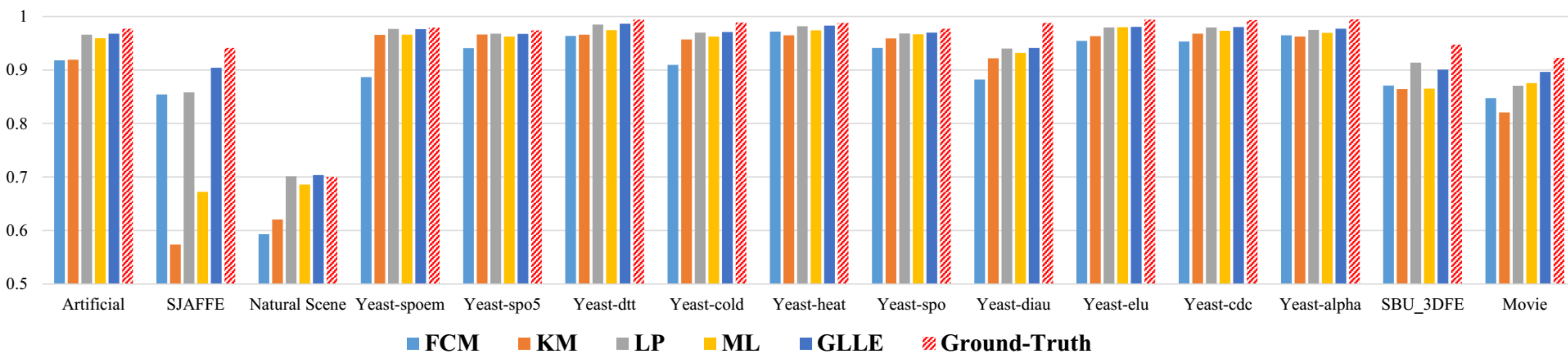


Figure 4: Comparison of the LDL after the LE pre-process against the direct LDL measured by Cosine ↑.

Conclusion

Major contribution: This paper shows *Label Enhancement*, which can recover the label distributions from the logical labels in the training sets via leveraging the topological information of the feature space and the correlation among the labels.

More information

http://palm.seu.edu.cn/
http://cse.seu.edu.cn/people/xgeng/index.htm/

THANK YOU！

http:// palm.seu.edu.cn