Label Embedding Based on Multi-Scale Locality Preservation

Cheng-Lun Peng, An Tao, Xin Geng

Reporter: Cheng-Lun Peng

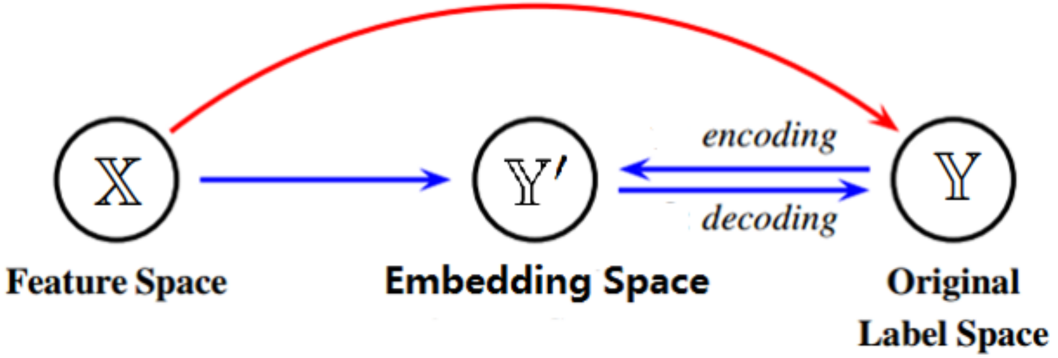
Date: July 17, 2018

Outline

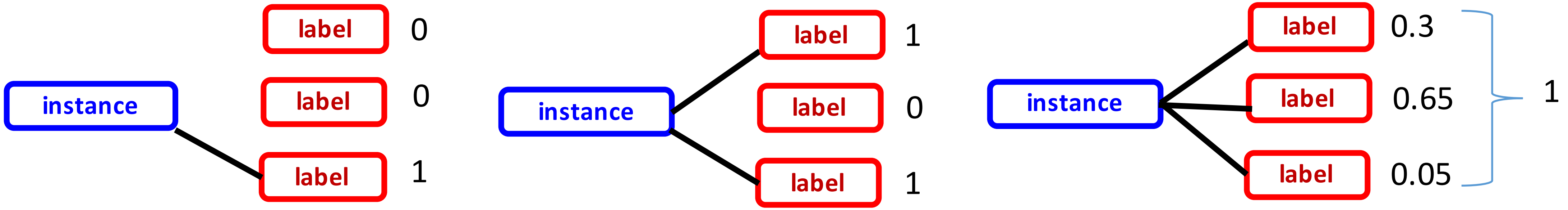
1. Background
2. Proposed Method: MSLP
3. Experiment
4. Conclusion

1 Background: LE & LDL

- **Label Embedding** (LE): A Learning Strategy
  - **Usual Steps**
    - encoding process (encoder)
    - learning process (predictor)
    - decoding process (decoder)



- **Label Distribution Learning** (LDL): A Learning Paradigm



Single-label learning

Multi-label learning

Label Distribution learning

- **Our Work**

  Propose a specially designed LE Method named MSLP for LDL, which is the first attempt of applying LE in LDL

1 Background: The Meaning of Our Work

❑ Why Apply LE in LDL

- The labels in LDL may encounter problems (e.g., redundancy, noise, …)
- Effective exploitation of the label correlations is crucial for the success for LDL.
- LE owns advantages in addressing problematic labels and capturing latent correlation between labels.

❑ What's The Challenges of Applying LE in LDL

- There are no LE method for LDL proposed yet. Most existing LE methods are designed for SLL and MLL, i.e., focusing on the binary labels (0/1).
- Two main issues
  a) How to exploit the information of label distributions efficiently.
  b) How to design a decoder that restricts the recovered label vector to satisfy the constraints of the label distribution.

1 Background: Symbol Definition

❑ Symbol Definition

$S = \{\boldsymbol{x}_i, \boldsymbol{y}_i\}_{i=1}^N$: Dataset

$\boldsymbol{x}_i \in \mathbb{X} = R^M$: i-th instance

$\boldsymbol{y}_i \in \mathbb{Y} = R^L$: i-th label vector $\quad y_i^c \in [0, 1] \quad \sum_{c=1}^L y_i^c = 1$

$\boldsymbol{X}_{N*M} = [\boldsymbol{x}_1, ..., \boldsymbol{x}_N]^t$ and $\boldsymbol{Y}_{N*L} = [\boldsymbol{y}_1, ..., \boldsymbol{y}_N]^t$

$\boldsymbol{y}_i' \in \mathbb{Y}' = \mathbb{R}^l$: i-th embedded label vector

$\boldsymbol{Y}_{N*l}' = [\boldsymbol{y}_1', ..., \boldsymbol{y}_N']^t$

2 MSLP: Motivation

❑ Motivation

◆ Locality Preserving Embedding for The Label Space

Inspired by Laplacian Eigenmaps [Belkin and Niyogi, 2002], MSLP aims to make the data points with similar label distributions close to each other in the embedding space.

$$\min_{\boldsymbol{Y}'} \ \frac{1}{2}\sum_{i,j} \| \boldsymbol{y}'_i - \boldsymbol{y}'_j \|^2 \boldsymbol{W}^{+}_{\boldsymbol{y},ij}$$

$$s.t. \ \boldsymbol{Y}'^{T}\boldsymbol{D}^{+}\boldsymbol{Y}' = \boldsymbol{I}$$

$$Nei_{\boldsymbol{y}}(i) = \psi_{\boldsymbol{y}}(\boldsymbol{p}_i,\ k^{+},\ \{\boldsymbol{p}_j \,|\, \boldsymbol{p}_j \neq \boldsymbol{p}_i \wedge \boldsymbol{p}_j \in S\}).$$

$$\boldsymbol{D}^{+}_{ii} = \textstyle\sum_j \boldsymbol{W}^{+}_{\boldsymbol{y},ij},\ \boldsymbol{Y}'_{N*l} = [\boldsymbol{y}'_1, \dots, \boldsymbol{y}'_N]^t$$

Find k^{+} nearest neighbors for data point $\boldsymbol{p_i}$ in the label space among the given point set

$$\boldsymbol{W}^{+}_{\boldsymbol{y},ij} = \begin{cases} exp(-\frac{dis(\boldsymbol{y}_i, \boldsymbol{y}_j)}{\sigma}), i \in Nei_{\boldsymbol{y}}(j) \ or \ j \in Nei_{\boldsymbol{y}}(i) \\ 0, \qquad otherwise \end{cases}$$

2 MSLP: Explicit Assumption

❑ Explicit Assumption

Assume an explicit mapping from the features to the embedded labels

$$\min_{\mathbf{Y}'} \frac{1}{2} \sum_{i,j} \| \boldsymbol{y}_i' - \boldsymbol{y}_j' \|^2 \boldsymbol{W}_{\boldsymbol{y},ij}^{+}$$

$$s.t. \ \ \mathbf{Y}'^T \boldsymbol{D}^{+} \mathbf{Y}' = \boldsymbol{I}$$

$$\boldsymbol{y}' = \boldsymbol{V}^T \boldsymbol{x}$$

$$\min_{\mathbf{V}} \frac{1}{2} \sum_{ij} \| \boldsymbol{V}^T \boldsymbol{x}_i - \boldsymbol{V}^T \boldsymbol{x}_j \|^2 \boldsymbol{W}_{\boldsymbol{y},ij}^{+} + \lambda \| \boldsymbol{V} \|_F^2$$

$$s.t. \ \ \boldsymbol{V}^T \boldsymbol{X}^T \boldsymbol{D}^{+} \boldsymbol{X} \boldsymbol{V} = \boldsymbol{I}$$

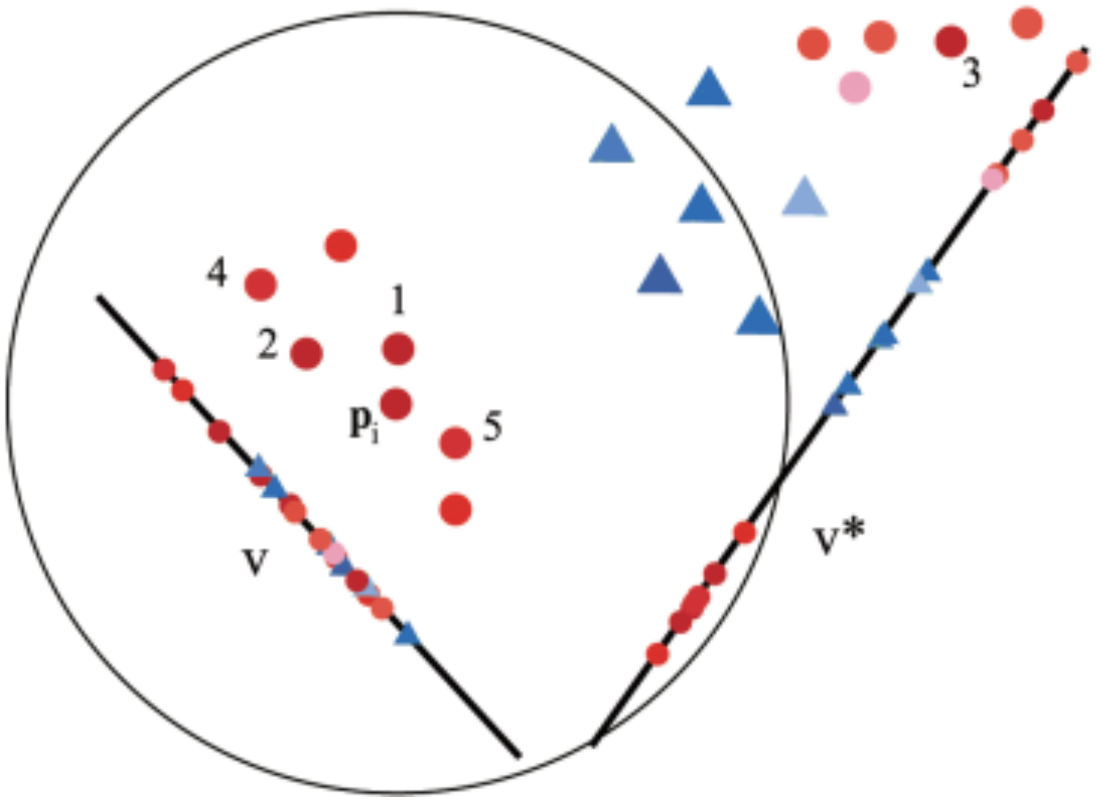
L2 Regularization

Advantage:

- Makes the process of label embedding feature-aware
- Omits the additional learning process from $\mathbb{X}$ to $\mathbb{Y}'$ after completing embedding.

2 MSLP: Explicit Assumption

❑ Problem of Explicit Linear Assumption



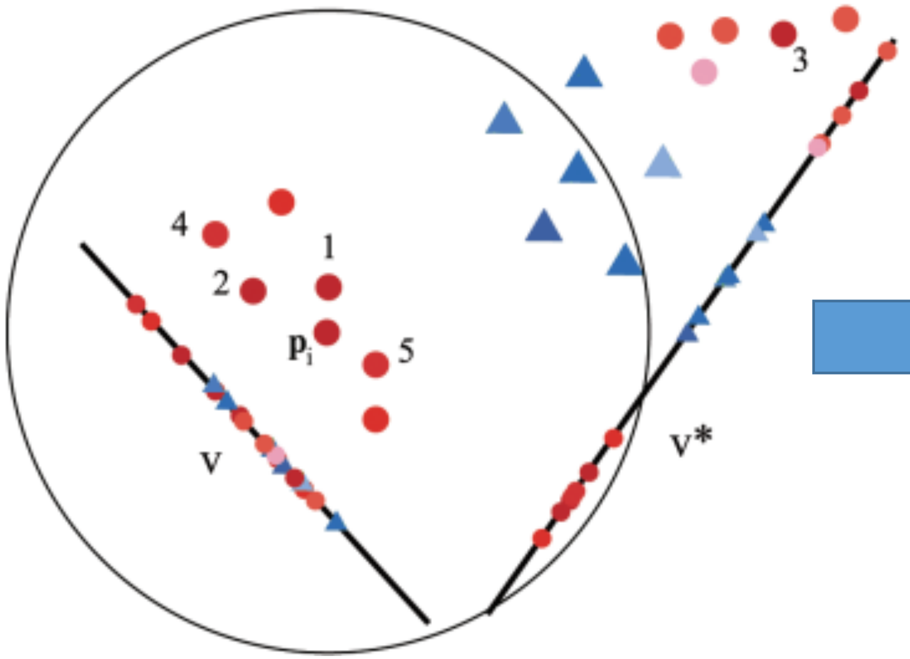
$$\min_{\boldsymbol{V}} \ \frac{1}{2}\sum_{ij} \| \boldsymbol{V}^T \boldsymbol{x}_i - \boldsymbol{V}^T \boldsymbol{x}_j \|^2 \, \boldsymbol{W}^{+}_{\boldsymbol{y},ij} + \lambda \| \boldsymbol{V} \|_F^2$$

The solution for ***V*** will tend to be dominated by the large feature distances of data pairs where $\boldsymbol{W}^{+}_{\boldsymbol{y},ij} \neq 0.$

Data pairs which keep very close in the label space, but keep far away from each other in the feature space.

2 MSLP: Restriction

❑ Multi-Scale Locality Preservation



$$Nei_{\boldsymbol{y}}(i) = \psi_{\boldsymbol{y}}(\boldsymbol{p}_i,\ k^+,\ Nei_{\boldsymbol{x}}(i))$$

$$Nei_{\boldsymbol{x}}(i) = \psi_{\boldsymbol{x}}(\boldsymbol{p}_i,\ \alpha k^+,\ \{\boldsymbol{p}_j \mid \boldsymbol{p}_j \neq \boldsymbol{p}_i \wedge \boldsymbol{p}_j \in S\}) \quad \alpha \geq 1$$

Restriction:

The k^+ nearest neighbors of one data point in label space should be found

within

its αk^+ nearest neighbors in feature space.

That is, utilizing different locality granularity in the label space and the feature space, the locality information of data points in both spaces are integrated.

2 MSLP: Robust to Noise

- **Smoothness Assumption** [Chapelle *et al.*,2006]
  Neighboring data points in feature space are more likely to share the similar labels.

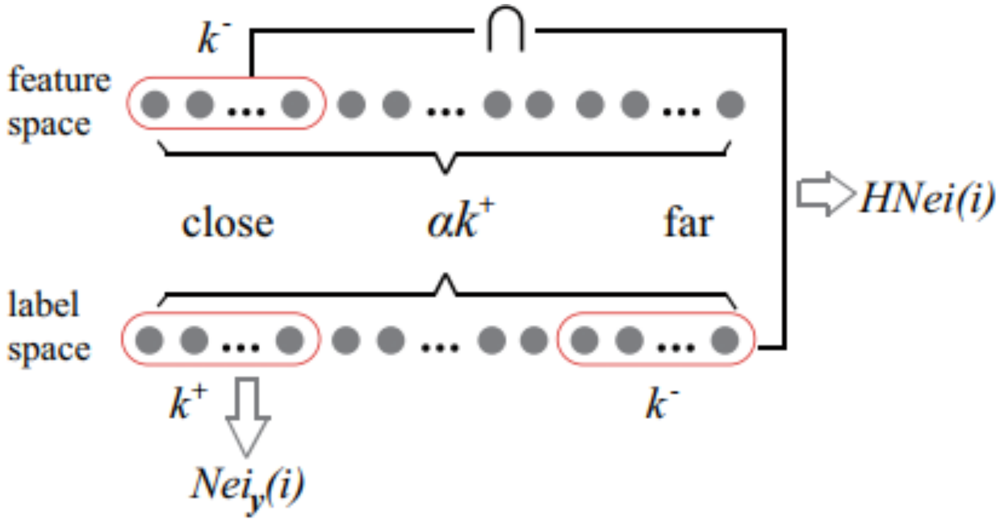
- **Hetero-neighbors**
  Data pairs which keep very close in the feature space, but keep far away from each other in the label space.

$$HNei(i) = \psi_{\boldsymbol{x}}(\boldsymbol{p}_i, k^-, Nei_{\boldsymbol{x}}(i)) \cap \psi_{\boldsymbol{y}}(\boldsymbol{p}_i, -k^-, Nei_{\boldsymbol{x}}(i)),$$

$$Nei_{\boldsymbol{x}}(i) = \psi_{\boldsymbol{x}}(\boldsymbol{p}_i,\ \alpha k^+,\ \{\boldsymbol{p}_j \mid \boldsymbol{p}_j \neq \boldsymbol{p}_i \wedge \boldsymbol{p}_j \in S\})$$

$$max \sum_{ij} \| \boldsymbol{V}^T \boldsymbol{x}_i - \boldsymbol{V}^T \boldsymbol{x}_j \|^2 \boldsymbol{W}_{ij}^-$$

$$\boldsymbol{W}_{ij}^- = \begin{cases} 1, & i \in HNei(j) \ or \ j \in HNei(i) \\ 0, & otherwise \end{cases}.$$



2 MSLP: Objective

- **The objective of MSLP**

$$\min_{\boldsymbol{V}} \frac{\beta}{2}\sum_{ij} \| \boldsymbol{V}^T\boldsymbol{x}_i - \boldsymbol{V}^T\boldsymbol{x}_j \|^2 \boldsymbol{W}^{+}_{\boldsymbol{y},ij} - \frac{(1-\beta)}{2}\sum_{ij} \| \boldsymbol{V}^T\boldsymbol{x}_i - \boldsymbol{V}^T\boldsymbol{x}_j \|^2 \boldsymbol{W}^{-}_{ij} + \lambda \| \boldsymbol{V} \|_F^2$$

$$s.t.\ \boldsymbol{V}^T\boldsymbol{X}^T\boldsymbol{D}^{+}\boldsymbol{X}\boldsymbol{V} = \boldsymbol{I}$$

$\beta \in [0,1]$ balances the importance of the first two terms

2 MSLP: Solution

$$\min_{\boldsymbol{V}} \ \frac{\beta}{2}\sum_{ij} \| \boldsymbol{V}^T\boldsymbol{x}_i - \boldsymbol{V}^T\boldsymbol{x}_j \|^2 \boldsymbol{W}^+_{\boldsymbol{y},ij} - \frac{(1-\beta)}{2}\sum_{ij} \| \boldsymbol{V}^T\boldsymbol{x}_i - \boldsymbol{V}^T\boldsymbol{x}_j \|^2 \boldsymbol{W}^-_{ij} + \lambda \| \boldsymbol{V} \|_F^2$$

$$\begin{aligned}
&\frac{\beta}{2}\sum_{ij} tr[\boldsymbol{V}^T(\boldsymbol{x}_i-\boldsymbol{x}_j)(\boldsymbol{x}_i-\boldsymbol{x}_j)^T\boldsymbol{V}]\boldsymbol{W}^+_{\boldsymbol{y},ij} \\
&= \beta\sum_{i} tr[\boldsymbol{V}^T\boldsymbol{x}_i\boldsymbol{D}^+_{ii}\boldsymbol{x}_i^T\boldsymbol{V}] - \beta\sum_{ij} tr[\boldsymbol{V}^T\boldsymbol{x}_i\boldsymbol{W}^+_{\boldsymbol{y},ij}\boldsymbol{x}_j^T\boldsymbol{V}] \\
&= \beta\, tr[\boldsymbol{V}^T(\sum_i \boldsymbol{x}_i\boldsymbol{D}^+_{ii}\boldsymbol{x}_i^T - \sum_{ij}\boldsymbol{x}_i\boldsymbol{W}^+_{\boldsymbol{y},ij}\boldsymbol{x}_j^T)\boldsymbol{V}] \\
&= \beta\, tr[\boldsymbol{V}^T\boldsymbol{X}^T(\boldsymbol{D}^+ - \boldsymbol{W}^+)\boldsymbol{X}\boldsymbol{V}]
\end{aligned}$$

Through similar computing,

$(1-\beta)tr[\boldsymbol{V}^T\boldsymbol{X}^T(\boldsymbol{D}^- - \boldsymbol{W}^-)\boldsymbol{X}\boldsymbol{V}]$, where $\boldsymbol{D}^-_{ii} = \sum_j \boldsymbol{W}^-_{ij}$.

$$\min_{\boldsymbol{V}} \ tr[\boldsymbol{V}^T(\boldsymbol{X}^T(\beta\boldsymbol{M}^+ - (1-\beta)\boldsymbol{M}^-)\boldsymbol{X} + \lambda\boldsymbol{I})\boldsymbol{V}]$$

where $\boldsymbol{M}^+ = \boldsymbol{D}^+ - \boldsymbol{W}^+$ and $\boldsymbol{M}^- = \boldsymbol{D}^- - \boldsymbol{W}^-$.

2 MSLP: Solution

$$\min_{\boldsymbol{V}} \; tr[\boldsymbol{V}^T(\boldsymbol{X}^T(\beta \boldsymbol{M}^+ - (1-\beta)\boldsymbol{M}^-)\boldsymbol{X} + \lambda \boldsymbol{I})\boldsymbol{V}]$$
$$s.t. \;\; \boldsymbol{V}^T\boldsymbol{X}^T\boldsymbol{D}^+\boldsymbol{X}\boldsymbol{V} = \boldsymbol{I}$$

Applying the Lagrangian method, the problem can be transfered into a general eigen-decomposition problem.

$$(\boldsymbol{X}^T\boldsymbol{M}\boldsymbol{X} + \lambda \boldsymbol{I})\boldsymbol{v} = \eta(\boldsymbol{X}^T\boldsymbol{D}^+\boldsymbol{X})\boldsymbol{v}, \;\text{where } \boldsymbol{M} = \beta \boldsymbol{M}^+ - (1-\beta)\boldsymbol{M}^-$$

The optimal ***V*** consists of the first l normalized eigenvectors corresponding to the top l smallest eigenvalues.

2 MSLP: Decoder

- **Testing Phrase**

For an unseen instance $\boldsymbol{x}_u$, we first compute its corresponding embedded label vector $\widehat{\boldsymbol{y}}'_u = \boldsymbol{V}^T \boldsymbol{x}_u$. Then, the *knn*-based decoder recovers the predicted label distribution $\widehat{\boldsymbol{y}}_u$ by averaging the label distributions of k nearest neighbors of $\widehat{\boldsymbol{y}}'_u$ among the embedded label matrix $\boldsymbol{Y}'$.

3 Experiment: Configuration

❑ Compared Methods

- **Eight** popular LDL methods:
  IIS-LDL, CPNN, BFGS-LDL, LDSVR, AA-BP, AA-KNN, PT-SVM, PT-Bayes
- Four typical Feature Embedding methods:
  CCA, NPE, PCA, **LPP** **(The Linear version of Laplacian Eigenmaps)**
  The compared FE methods are allowed to be extended to their kernel version with the rbf kernel, which gives them full chances to beat MSLP.

❑ Widely-used Metrics in LDL

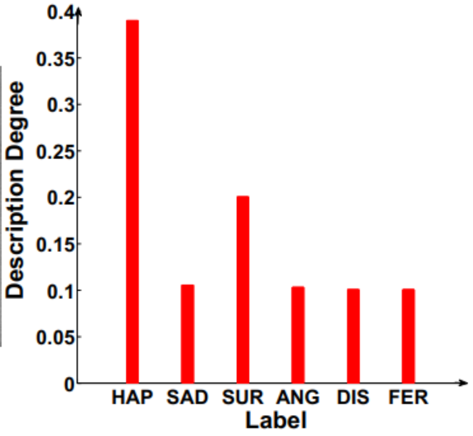
- Four distance metrics: Chebyshev, Clark, Kullback-Leibler, Canberra
- Two similarity metrics: Cosine and Intersection

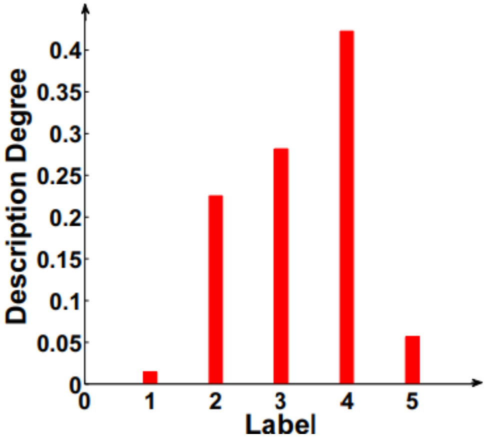
❑ Other Settings

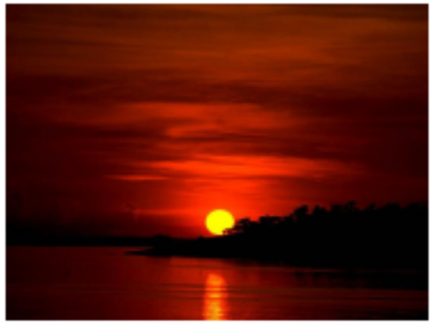
- the embedding ratio of the dimensionality ranges over {10%, 20%, …, 100%}
- Running each method with the best tuned parameters
- 10-fold cross validation
- Pairwise t-tests at 90% significance level

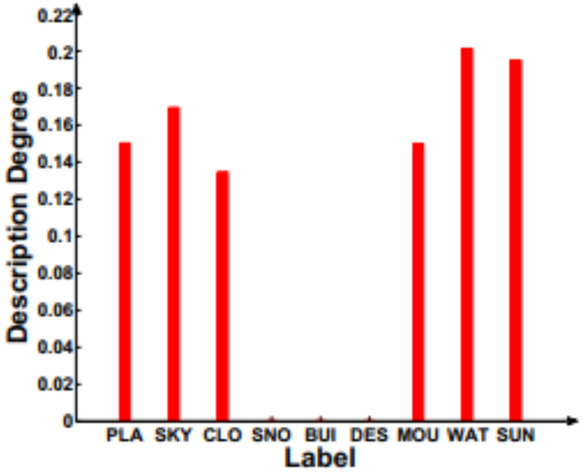
3 Experiment: Datasets

Datasets	#S	#Label	#Feature	Domain
s-JAFEE	213	6	243	facial expression recognition
s-BU-3DEF	2500	6	243	facial expression recognition
SCUT-FBP	1500	5	300	facial beauty sense
M^2B	1240	5	250	facial beauty sense
Nature_Scene	2000	9	294	natural scene annotation

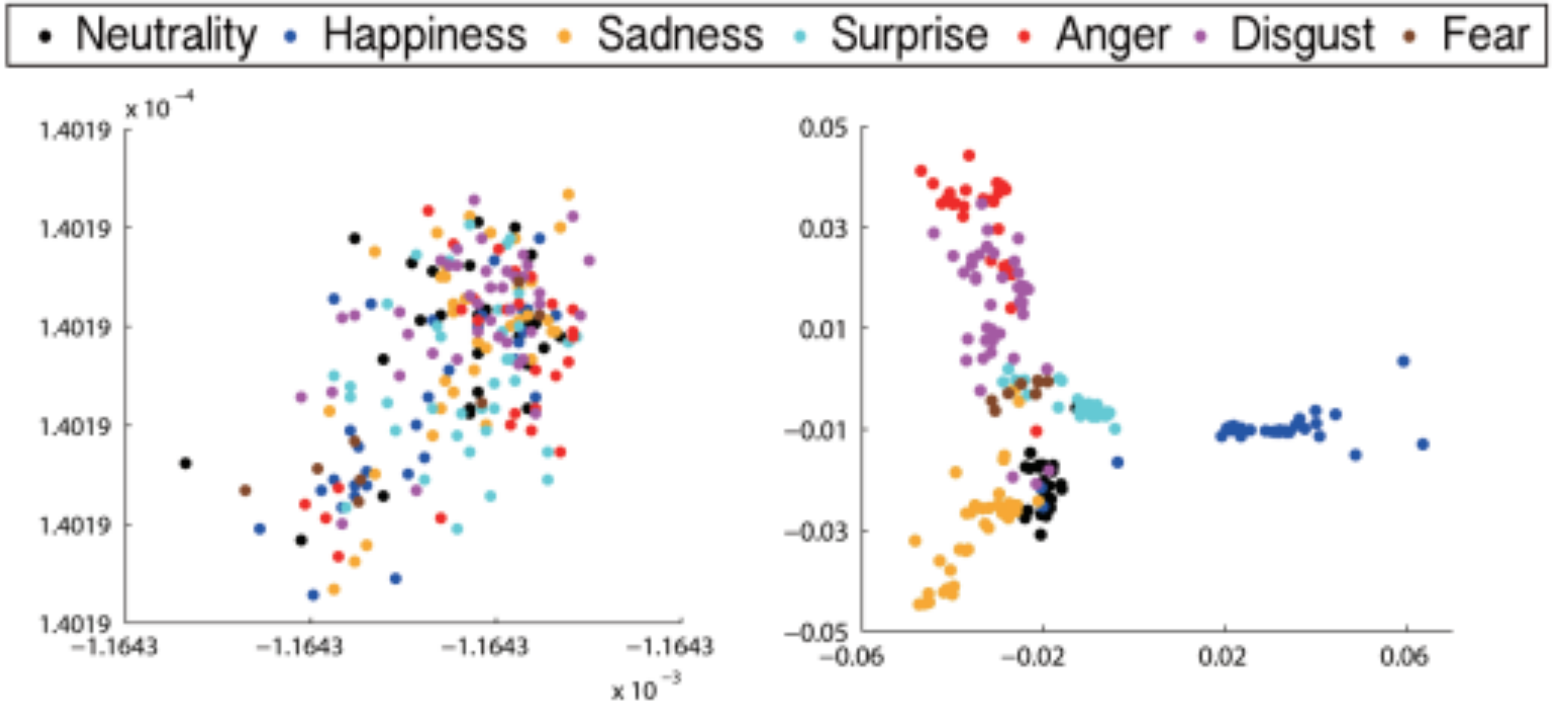








3 Experiment: Visualization



(a) LPP (b) MSLP

Figure 7: Embedding results on the s-JAFFE.

- Different colors are used to display images according to the highest description degree of the basic emotions

3 Experiment: Quantitative Results

Datasets	SCUT-FBP						Multi-Modality Beauty (M^2B)					
Metrics	Cheb ↓	Cla ↓	Can ↓	KL ↓	Cos ↑	Inter ↑	Cheb ↓	Cla ↓	Can ↓	KL ↓	Cos ↑	Inter ↑
CPNN	0.4436	1.5943	3.2116	1.1801	0.5759	0.4918	0.3726	1.3012	2.6129	0.5649	0.7082	0.5584
LDSVR	0.2694	1.4280	2.7575	0.5684	0.8341	0.7121	0.3611	1.2704	2.5457	**0.5363**	**0.7235**	0.5781
BFGS-LDL	0.3517	1.5321	2.9985	0.8572	0.6905	0.5557	0.3720	**1.2180**	**2.4166**	0.6898	0.6759	0.5698
IIS-LDL	0.6493	1.8615	3.9359	3.2270	0.3593	0.2985	0.3790	1.3136	2.6368	0.5864	0.6970	0.5522
AA-BP	0.2538	1.4002	2.5981	0.4157	0.8372	0.6948	0.3781	1.3142	2.6356	0.5851	0.6992	0.5529
AA-KNN	0.2148	1.2761	2.2953	0.3602	0.8691	0.7435	0.3754	1.2204	2.4190	0.6884	0.6711	0.5600
PT-SVM	0.4184	1.5736	3.1111	1.1354	0.5784	0.5080	0.4139	1.3576	2.7366	0.8057	0.5990	0.5004
PT-Bayes	0.3836	1.5376	3.0516	1.1328	0.6779	0.5156	0.6905	2.0668	4.4951	11.832	0.4474	0.3044
LPP	0.2202	1.3219	2.4098	0.3256	0.8667	0.7358	0.3675	1.2557	2.5167	0.5893	0.6987	0.5662
NPE	0.2133	1.3057	2.3630	**0.2854**	0.8784	0.7446	0.3645	1.2688	2.5295	0.5643	0.7095	0.5695
PCA	0.2144	1.3082	2.3701	0.3144	0.8717	0.7435	0.3688	1.2433	2.4858	0.6151	0.6902	0.5666
CCA	0.2141	1.2938	2.3404	**0.2927**	0.8724	0.7473	**0.3559**	1.2382	2.4690	0.5476	0.7192	**0.5811**
MSLP	**0.2046**	**1.2602**	**2.2477**	**0.2813**	**0.8823**	**0.7608**	**0.3549**	**1.2095**	**2.4058**	0.5684	0.7127	**0.5844**

Table 2: Experimental results on facial beauty sense.

3 Experiment: Quantitative Results

Across all metrics, MSLP ranks 1st in 93.3% cases.

Datasets	Natural_Scene (NS)					
Metrics	Cheb ↓	Cla ↓	Can ↓	KL ↓	Cos ↑	Inter ↑
CPNN	0.3136	2.4720	6.8613	0.9022	0.6847	0.4908
LDSVR	0.4082	2.3884	6.7650	1.1158	0.6372	0.5093
BFGS-LDL	0.3342	2.3956	6.5829	0.9310	0.6979	0.5416
IIS-LDL	0.3569	2.4737	6.8221	0.9437	0.6649	0.4618
AA-BP	0.3387	2.4593	6.7762	0.8925	0.6898	0.4907
AA-KNN	0.3055	2.2548	5.8358	**0.7919**	0.7309	0.5567
PT-SVM	0.4282	2.5696	7.2756	1.5533	0.4583	0.3497
PT-Bayes	0.4047	2.5206	7.1398	2.2363	0.5601	0.3305
LPP	0.3113	2.2959	6.0166	0.8307	0.7178	0.5390
NPE	0.3021	2.2324	5.7596	**0.7910**	0.7359	0.5619
PCA	0.3048	2.2512	5.8199	**0.7913**	0.7310	0.5577
CCA	0.3188	2.3139	6.1598	0.8361	0.7243	0.5319
MSLP	**0.2927**	**2.2198**	**5.7214**	**0.7849**	**0.7406**	**0.5696**

Table 3: Experimental results on natural scene annotation.

4 Conclusion

❑ Conclusion

- The **first attempt** of embedding LE into LDL.
- MSLP is **insensitive** to the presence of hetero-neighbors and **integrates** the locality structure of points in **both spaces** with different granularity.
- Experiments reveal the **effectiveness** of MSLP **in** gathering points with similar label distributions in the embedding space.

❑ Future Work

- Explore if there exist better ways to utilize the structure information described by the label distributions.
- Shift MSLP to some other learning paradigms (e.g., multi-output regression) which own numerical labels.

Thank you

Q & A