VALSE 2016 武汉

Discriminative Analysis Dictionary Learning

Jun Guo[1], Yanqing Guo[1], Xiangwei Kong[1], Man Zhang[2], and Ran He[2]*
1: Dalian University of Technology, 2: National Laboratory of Pattern Recognition
Email: guojun@mail.dlut.edu.cn

Introduction

- Two major branches of Dictionary Learning (DL):
  - –Synthesis DL: $\min_{\mathbf{D},\mathbf{X}} \sum_{i=1}^{n} dist(\mathbf{y}_i, \mathbf{D}\mathbf{x}_i)$ $s.t.$ $\mathbf{D} \in \mathcal{D},\ \|\mathbf{x}_i\|_0 \le T_0,\ i = 1, 2, \cdots, n$
  - –Analysis DL: $\min_{\mathbf{\Omega},\mathbf{X}} \sum_{i=1}^{n} dist(\mathbf{x}_i, \mathbf{\Omega}\mathbf{y}_i)$ $s.t.$ $\mathbf{\Omega} \in \mathcal{W},\ \|\mathbf{x}_i\|_0 \le T_0,\ i = 1, 2, \cdots, n$
- Dictionary Learning (DL) in pattern classification:
  - –Synthesis DL: $\min_{\mathbf{D},\mathbf{X}} \sum_{i=1}^{n} dist(\mathbf{y}_i, \mathbf{D}\mathbf{x}_i) + \lambda\mathcal{F}(\mathbf{D},\mathbf{X},\text{label}(\mathbf{Y}),\text{structure}(\mathbf{Y}))$ $s.t.$ $\mathbf{D} \in \mathcal{D},\ \|\mathbf{x}_i\|_0 \le T_0,\ i = 1, 2, \cdots, n$
  - –Analysis DL: There are few works. Our paper focuses on this.

Motivation

- Goal: Improve the classification performance of Analysis DL.
  - –Analysis DL + a simple classifier (i.e. kNN)
  - –In the coding space:
    - ➢same-label neighbors are orderly preserved
    - ➢neighbors with different labels are repelled
- Approach: **Discriminative Analysis Dictionary Learning (DADL)**
  - –Integrate two significant characters into Analysis DL
    - ➢① strengthen discriminability
    - ➢② preserve local topology structure
  - –Better control outliers and noise for classification
    - ➢③ employ Correntropy Induced Metric (CIM) instead of Mean Square Error (MSE)

The Proposed Method

- ① Strengthen discriminability
  - –Integrate a code consistent term: $\min_{\mathbf{\Omega},\mathbf{X}} \sum_{i=1}^{n} dist(\mathbf{x}_i, \mathbf{\Omega}\mathbf{y}_i) + \lambda_1 \sum_{i=1}^{n} dist(\mathbf{x}_i, \mathbf{h}_i)$ $s.t.$ $\mathbf{\Omega} \in \mathcal{W},\ \|\mathbf{x}_i\|_0 \le T_0,\ i = 1, 2, \cdots, n$
  - –Generate target codes (e.g., Hadamard code).
- ② Preserve local topology structure
  - –Definition 1: A coding process is called *local topology preserving* when the following condition holds: if $dist(\mathbf{y}_i, \mathbf{y}_u) \le dist(\mathbf{y}_i, \mathbf{y}_v)$, then $dist(\mathbf{x}_i, \mathbf{x}_u) \le dist(\mathbf{x}_i, \mathbf{x}_v)$.
  - –Therefore, determining appropriate $\{\mathbf{x}_u, \mathbf{x}_v\}$ for $\mathbf{x}_i$:
    $$\max_{\mathbf{x}_u,\mathbf{x}_v} \mathbf{A}_i(u,v)\left[dist(\mathbf{x}_i,\mathbf{x}_u) - dist(\mathbf{x}_i,\mathbf{x}_v)\right]$$
    - ➢$\mathbf{A}_i$ is antisymmetric with $\mathbf{A}_i(u,v) = dist(\mathbf{y}_i,\mathbf{y}_u) - dist(\mathbf{y}_i,\mathbf{y}_v)$.
    - ➢However, this loss is an unsupervised type, neglecting labels.
  - –Considering each sample's label:
    $$\mathbf{A}'_i(u,v) \triangleq \begin{cases} -\mathbf{A}_i(u,v)\, sign\left[\mathbf{A}_i(u,v)\right] & , label(\mathbf{y}_i) = label(\mathbf{y}_u) \ne label(\mathbf{y}_v) \\ \mathbf{A}_i(u,v)\, sign\left[\mathbf{A}_i(u,v)\right] & , label(\mathbf{y}_i) = label(\mathbf{y}_v) \ne label(\mathbf{y}_u) \\ \mathbf{A}_i(u,v) & , otherwise \end{cases}$$
  - –Replace $\mathbf{A}_i$ with $\mathbf{A}'_i$ that is also antisymmetric:
    $$\max_{\mathbf{x}_u,\mathbf{x}_v} \mathbf{A}'_i(u,v)\left[dist(\mathbf{x}_i,\mathbf{x}_u) - dist(\mathbf{x}_i,\mathbf{x}_v)\right]$$
  - –So that in the coding space:
    - ➢same-label neighbors are orderly preserved
    - ➢neighbors with different labels are repelled
  - –Then, we obtain a supervised loss function:
    $$\max_{\mathbf{X}} \sum_{i=1}^{n}\sum_{u=1}^{n}\sum_{v=1}^{n} \mathbf{A}'_i(u,v)\left[dist(\mathbf{x}_i,\mathbf{x}_u) - dist(\mathbf{x}_i,\mathbf{x}_v)\right].$$
  - –Reformulate: (please refer to Proposition 1 in our paper)
    $$\min_{\mathbf{X}} \sum_{i=1}^{n}\sum_{j=1}^{n} \mathbf{W}_{ij} dist(\mathbf{x}_i,\mathbf{x}_j), \text{ where } \mathbf{W}_{ij} = \sum_{u=1}^{n} \mathbf{A}'_i(u,j).$$
    - ➢to simultaneously preserve neighborhood ranking information as well as neighborhood relationship:
    $$\mathbf{W}_{ij} = \begin{cases} \sum_{\mathbf{y}_u \in \mathcal{N}_i} \mathbf{A}'_i(u,j) & , \mathbf{y}_j \in \mathcal{N}_i \\ 0 & , otherwise \end{cases}$$
    ($\mathcal{N}_i$ is a set containing the k nearest neighbors of $\mathbf{y}_i$)
    - ➢to directly learn the analysis dictionary : $\min_{\mathbf{\Omega}} \sum_{i=1}^{n}\sum_{j=1}^{n} \mathbf{W}_{ij} dist(\mathbf{\Omega}\mathbf{y}_i, \mathbf{\Omega}\mathbf{y}_j)$
- ③ Employ CIM instead of MSE
  - –Correntropy Induced Metric (CIM): $dist(\mathbf{u},\mathbf{v}) = \left[1 - \exp\left(-\|\mathbf{u}-\mathbf{v}\|_2^2/\sigma^2\right)\right]^{1/2}$
    - ➢more robust to outliers and noise
  - –Final objective function:
    $$\begin{aligned} \min_{\mathbf{\Omega},\mathbf{X}} \quad & J = J_0 + \lambda_1 J_1 + \lambda_2 J_2 \\ s.t. \quad & \mathbf{\Omega} \in \mathcal{W}, \\ & \|\mathbf{x}_i\|_0 \le T_0,\ \forall i \end{aligned} \qquad \begin{cases} J_0 = \sum_{i=1}^{n}\left\{1 - \exp\left(-\frac{\|\mathbf{x}_i - \mathbf{\Omega}\mathbf{y}_i\|_2^2}{\sigma^2}\right)\right\} \\ J_1 = \sum_{i=1}^{n}\left\{1 - \exp\left(-\frac{\|\mathbf{x}_i - \mathbf{h}_i\|_2^2}{\sigma^2}\right)\right\} \\ J_2 = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\left\{\mathbf{W}_{ij}\left[1 - \exp\left(-\frac{\|\mathbf{\Omega}\mathbf{y}_i - \mathbf{\Omega}\mathbf{y}_j\|_2^2}{\sigma^2}\right)\right]\right\} \end{cases}$$

Optimization

- Half-quadratic (HQ) technique
  - –For a fixed z, there exists a dual potential function $\varphi(\cdot)$, such that:
    $$1 - \exp\left(-\frac{z^2}{\sigma^2}\right) = \inf_{p\in\mathbb{R}}\left\{pz^2 + \varphi(p)\right\}.$$
    - ➢The infimum can be reached at $p = \exp\left(-\frac{z^2}{\sigma^2}\right)$.
  - –Therefore, the augmented function of our objective function based on the half-quadratic (HQ) technique:
    $$\begin{aligned} \min_{\mathbf{\Omega},\mathbf{X},\mathbf{P},\mathbf{Q},\mathbf{R}} \quad & \hat{J} = \hat{J}_0 + \lambda_1 \hat{J}_1 + \lambda_2 \hat{J}_2 \\ s.t. \quad & \mathbf{\Omega} \in \mathcal{W}, \\ & \|\mathbf{x}_i\|_0 \le T_0,\ \forall i \end{aligned} \qquad \begin{cases} \hat{J}_0 = \sum_{i=1}^{n}\left\{\mathbf{P}_{ii}\frac{\|\mathbf{x}_i - \mathbf{\Omega}\mathbf{y}_i\|_2^2}{\sigma^2} + \phi_i(\mathbf{P}_{ii})\right\} \\ \hat{J}_1 = \sum_{i=1}^{n}\left\{\mathbf{Q}_{ii}\frac{\|\mathbf{x}_i - \mathbf{h}_i\|_2^2}{\sigma^2} + \varphi_i(\mathbf{Q}_{ii})\right\} \\ \hat{J}_2 = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\left\{\mathbf{W}_{ij}\mathbf{R}_{ij}\frac{\|\mathbf{\Omega}\mathbf{y}_i - \mathbf{\Omega}\mathbf{y}_j\|_2^2}{\sigma^2} + \mathbf{W}_{ij}\psi_{ij}(\mathbf{R}_{ij})\right\} \end{cases}$$
- Optimize the augmented function: (please refer to our paper)
  - –Update the analysis dictionary and sparse codes.
  - –Update auxiliary variables introduced by HQ.
  - –Alternatively minimized until convergence.

Experiments

- Comparing Algorithms
  - –the baseline Analysis DL + SVM [ICIP 2014]
  - –the classical SRC [PAMI 2009] and CRC [ICCV 2011]
  - –other famous DL methods: DLSI [CVPR 2010], FDDL [ICCV 2011], LC-KSVD [PAMI 2013], DPL [NIPS 2014]
- Results

Classification accuracies (%) on five datasets.

	YaleB	AR	Caltech 101	Scene 15	UCF 50
ADL+SVM	95.4	96.1	64.5	90.1	72.3
SRC	96.5	97.5	70.7	91.8	75.0
CRC	97.0	98.0	68.2	92.0	75.6
DLSI	97.0	97.5	73.1	91.7	75.4
FDDL	96.7	97.5	73.2	92.3	76.5
LC-KSVD	96.7	97.8	73.6	92.9	70.1
DPL	97.5	98.3	73.9	97.7	77.4
DADL	**97.7**	**98.7**	**74.6**	**98.3**	**78.0**

Training time (s) on five datasets.

	YaleB	AR	Caltech 101	Scene 15	UCF 50
DPL	5.92	15.21	180.54	56.84	652.03
DADL	4.23	11.16	121.47	36.52	330.23

Testing time (ms) on five datasets.

	YaleB	AR	Caltech 101	Scene 15	UCF 50
DPL	0.19	0.42	1.45	1.36	1.62
DADL	0.16	0.39	1.39	1.31	1.48

- Analysis
  - –Our proposed DADL method achieves higher accuracies than other dictionary learning methods.
  - –DPL outperforms state-of-the-art DL methods in terms of running time. Our proposed DADL method runs faster than DPL.