Preserving Ordinal Consensus: Towards Feature Selection for Unlabeled Data

Jun Guo, Heng Chang, and Wenwu Zhu

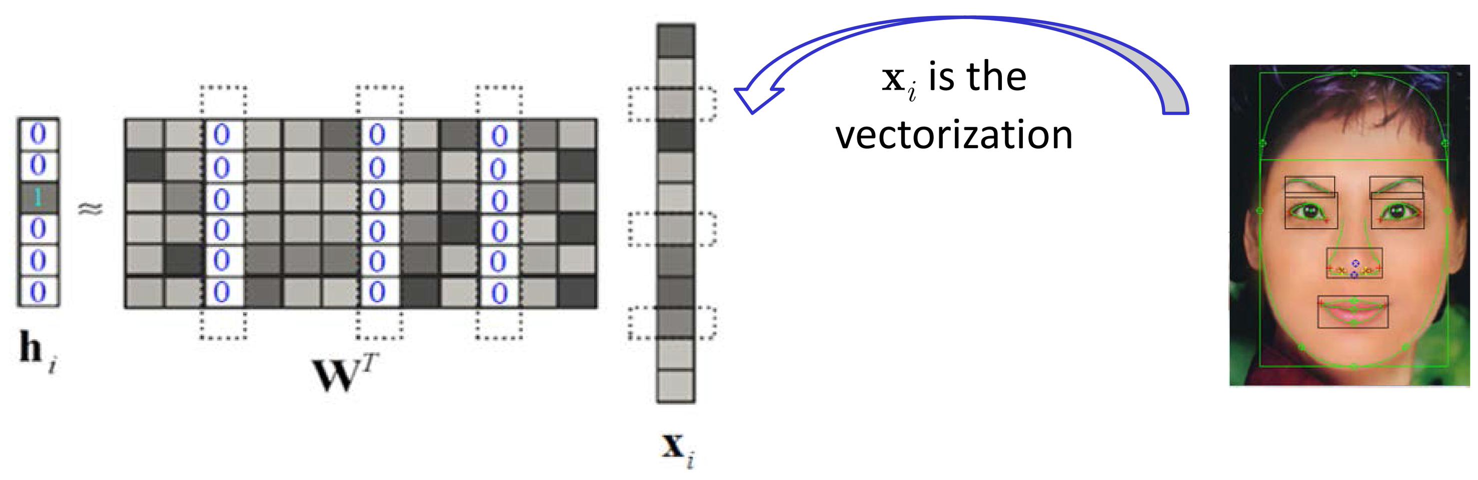
Tsinghua University, China

TBSI Tsinghua-Berkeley Shenzhen Institute

GitHub Homepage

Introduction

- Regression-based feature selection is frequently-used to address the above practical problem.



$$\min_{\mathbf{W}} \begin{Bmatrix} \left\|\mathbf{W}^T\mathbf{X}-\mathbf{H}\right\|_F^2 \\ +\beta\left\|\mathbf{W}\right\|_{2,1} \\ +\alpha\Upsilon(\mathbf{W},\mathbf{X},\mathbf{H}) \end{Bmatrix}$$

Original data matrix: $\mathbf{X} = [\mathbf{x}_1, \cdots, \mathbf{x}_n] \in \mathbb{R}^{d_1 \times n}$

Target matrix: $\mathbf{H} = [\mathbf{h}_1, \cdots, \mathbf{h}_n] \in \mathbb{R}^{d_2 \times n}$

Feature selection matrix: $\mathbf{W} \in \mathbb{R}^{d_1 \times d_2}$ $(d_1 > d_2)$

Regularization term: $\Upsilon(\mathbf{W},\mathbf{X},\mathbf{H})$

- In unsupervised scenarios:
  - $\mathbf{H} = [\mathbf{h}_1, \cdots, \mathbf{h}_n] \in \mathbb{R}^{d_2 \times n}$ is frequently determined by learning pseudo labels through classical machine learning algorithms, such as
    - ➢ linear regression [UDFS, IJCAI'11]
    - ➢ K-means clustering [RUFS, IJCAI'13]
    - ➢ spectral clustering [NDFS, AAAI'12]
    - ➢ bi-orthogonal semi-NMF [SOCFS, CVPR'15]
  - For the regularization term $\Upsilon(\mathbf{W},\mathbf{X},\mathbf{H})$, most existing methods mainly focus on preserving sample-level relations (e.g., locality). The **feature-level** relationship especially **ordinal information** is totally neglected.

Experiment

- Datasets:

Table 1: Description of Benchmark Datasets.

Dataset	# of Samples	# of Features	# of Classes	Type
LUNG	203	3312	5	cancer
COIL20	1440	1024	20	object
Isolet1	1560	617	26	spoken letter
USPS	9298	256	10	written digit
AT&T	400	644	40	human face
UMIST	575	644	20	human face

- Comparing Algorithms:
  - the baseline that uses all the feature dimensions (All Features)
  - the classical Laplacian Score [NIPS'05]
  - other famous unsupervised feature selection methods: MCFS [KDD'10], UDFS [IJCAI'11], NDFS [AAAI'12], RUFS [IJCAI'13], SOCFS [CVPR'15]

- Results:

Table 2: Clustering results (NMI%±STD). The best results are in boldface.

	LUNG	COIL20	Isolet1	USPS	AT&T	UMIST
All Features	51.7±5.4	76.3±1.8	75.9±1.6	60.9±0.8	80.5±1.8	42.1±2.3
Laplacian Score	42.9±5.0 (300)	71.8±2.0 (300)	73.1±1.5 (300)	59.5±2.1 (200)	80.4±1.8 (300)	45.1±3.4 (200)
MCFS	45.6±4.5 (300)	74.9±2.2 (150)	74.4±1.9 (200)	61.2±1.8 (200)	80.2±1.9 (200)	45.1±3.2 (150)
UDFS	49.6±5.1 (300)	74.7±1.6 (300)	73.6±1.6 (300)	56.8±1.4 (200)	80.6±1.8 (150)	44.9±2.7 (300)
UDFS + doublet	49.9±5.0 (300)	75.0±1.8 (300)	73.9±1.9 (250)	57.0±1.5 (200)	81.2±1.9 (200)	45.1±2.9 (250)
UDFS + triplet	51.7±5.1 (250)	75.4±1.7 (250)	74.4±1.7 (250)	57.5±1.5 (170)	82.5±1.8 (150)	45.6±2.7 (300)
NDFS	48.3±5.2 (250)	76.0±1.6 (300)	78.4±1.8 (250)	60.7±1.3 (140)	80.3±1.8 (300)	47.8±3.1 (150)
NDFS + doublet	48.8±5.0 (300)	76.2±1.7 (250)	78.8±1.8 (300)	62.7±1.5 (170)	80.9±2.0 (300)	48.0±2.9 (150)
NDFS + triplet	49.9±5.0 (250)	76.9±1.9 (250)	79.1±1.7 (250)	63.5±1.3 (140)	82.2±1.9 (300)	48.5±2.8 (200)
RUFS	49.1±5.1 (250)	77.0±2.2 (150)	78.9±1.1 (300)	61.5±1.4 (170)	80.9±1.7 (300)	46.4±3.0 (150)
RUFS + doublet	49.7±5.2 (250)	77.3±2.4 (200)	79.2±1.3 (250)	61.9±1.7 (200)	81.1±1.7 (300)	46.9±3.1 (200)
RUFS + triplet	51.0±5.0 (250)	77.8±2.1 (150)	79.7±1.2 (250)	62.5±1.6 (170)	82.3±1.7 (300)	47.2±3.0 (200)
SOCFS	55.7±6.2 (250)	74.8±2.3 (300)	78.3±1.9 (300)	61.6±1.4 (110)	81.1±1.6 (100)	49.4±3.2 (50)
SOCFS + doublet	55.9±6.0 (300)	75.0±2.2 (250)	79.2±2.0 (300)	61.9±1.1 (110)	81.4±1.3 (200)	50.0±3.0 (100)
SOCFS + triplet	56.6±5.9 (250)	75.3±2.1 (250)	80.0±2.0 (250)	62.2±1.0 (110)	82.3±1.2 (100)	50.3±3.0 (100)
Ours ($\alpha = 0$)	52.3±6.3 (300)	74.7±2.6 (250)	77.3±2.1 (250)	62.1±1.7 (200)	79.8±1.9 (150)	48.3±3.5 (50)
Ours (doublet)	56.8±6.1 (250)	77.5±2.3 (250)	78.9±2.0 (300)	62.9±1.5 (200)	83.6±1.6 (200)	51.5±3.3 (100)
Ours (triplet)	**60.2±5.8** (250)	**80.1±2.2** (200)	**82.2±1.6** (200)	**64.5±1.0** (200)	**86.2±1.6** (200)	**52.6±3.1** (100)

- Our proposed method achieves **higher** accuracies than other methods.
- Triplet is more **effective** than doublet for feature-selection-based clustering.

Our Proposed Method

Definition

Given distance function $dis(\cdot,\cdot)$ and projection function $\Phi(\cdot)$. A data point $\mathbf{z}_i$ and its neighbors $\mathbf{z}_u$ and $\mathbf{z}_v$ form a triplet. The projection of this triplet is defined as an ***Ordinal Consensus Preserving*** process when the following condition holds: If $dis(\mathbf{z}_i,\mathbf{z}_u) \le dis(\mathbf{z}_i,\mathbf{z}_v)$, then $dis(\Phi(\mathbf{z}_i),\Phi(\mathbf{z}_u)) \le dis(\Phi(\mathbf{z}_i),\Phi(\mathbf{z}_v))$.

- For $\mathbf{z}_i$, we optimize: $\max_{\Phi(\cdot)} \mathbf{A}^i_{uv}\left[dis\left(\Phi(\mathbf{z}_i),\Phi(\mathbf{z}_u)\right) - dis\left(\Phi(\mathbf{z}_i),\Phi(\mathbf{z}_v)\right)\right]$
  - ➢ $\mathbf{A}^i$ is an antisymmetric matrix.
  - ➢ The $(u,v)^{th}$ element of $\mathbf{A}^i$ is defined as $\mathbf{A}^i_{uv} = dis\left(\mathbf{z}_i,\mathbf{z}_u\right) - dis\left(\mathbf{z}_i,\mathbf{z}_v\right)$.

Rearrangement Inequality

If real numbers $s_1 \le s_2 \le \cdots \le s_n$ and $t_1 \le t_2 \le \cdots \le t_n$, then

$s_n t_1 + s_{n-1} t_2 + \cdots + s_1 t_n \le s_{\sigma(1)} t_1 + s_{\sigma(2)} t_2 + \cdots + s_{\sigma(n)} t_n \le s_1 t_1 + s_2 t_2 + \cdots + s_n t_n$,

where $\sigma(1), \sigma(2), \ldots, \sigma(n)$ is the permutation of $\{1, 2, \cdots, n\}$.

- For all data points: $\max_{\Phi(\cdot)} \sum_{i=1}^{n} \sum_{u,v\in\mathcal{N}_i} \mathbf{A}^i_{uv}\left[dis\left(\Phi(\mathbf{z}_i),\Phi(\mathbf{z}_u)\right) - dis\left(\Phi(\mathbf{z}_i),\Phi(\mathbf{z}_v)\right)\right]$
  - ➢ $\mathcal{N}_i$ indicates the k nearest neighbors of $\mathbf{z}_i$.

expand

$$\max_{\Phi(\cdot)} -\sum_{i=1}^{n}\sum_{u,v\in\mathcal{N}_i} \mathbf{A}^i_{vu} dis\left(\Phi(\mathbf{z}_i),\Phi(\mathbf{z}_u)\right) - \sum_{i=1}^{n}\sum_{u,v\in\mathcal{N}_i} \mathbf{A}^i_{uv} dis\left(\Phi(\mathbf{z}_i),\Phi(\mathbf{z}_v)\right)$$

$$\mathbf{C}_{ij} = \begin{cases} \sum_{u\in\mathcal{N}_i} \mathbf{A}^i_{uj} & , j \in \mathcal{N}_i \\ 0 & , j \notin \mathcal{N}_i \end{cases}$$

$$\max_{\Phi(\cdot)} -\sum_{i=1}^{n}\sum_{u=1}^{n} \mathbf{C}_{iu} dis\left(\Phi(\mathbf{z}_i),\Phi(\mathbf{z}_u)\right) - \sum_{i=1}^{n}\sum_{v=1}^{n} \mathbf{C}_{iv} dis\left(\Phi(\mathbf{z}_i),\Phi(\mathbf{z}_v)\right)$$

- The equivalent form: $\min_{\Phi(\cdot)} \sum_{i=1}^{n}\sum_{j=1}^{n} \mathbf{C}_{ij} dis\left(\Phi(\mathbf{z}_i),\Phi(\mathbf{z}_j)\right)$
- Note that the i^{th} projection vector and the i^{th} feature dimension share the one-to-one correspondence:

$$\mathbf{W}^T\mathbf{X} = (\mathbf{w}^1)^T\mathbf{x}^1 + \ldots + (\mathbf{w}^i)^T\mathbf{x}^i + \ldots + (\mathbf{w}^j)^T\mathbf{x}^j + \ldots$$

  - ➢ This one-to-one correspondence can be regarded as an implicit function $\Phi(\cdot)$.
  - ➢ For simplicity, we use squared Euclidean distance, then we obtain the Laplacian matrix $\mathbf{L}$ of the above-defined weighting matrix $\mathbf{C}$.
  - ➢ To enforce $\mathbf{W}$ to preserve the feature-level ordinal consensus: $\min_{\mathbf{W}} Tr\left(\mathbf{W}^T\mathbf{L}\mathbf{W}\right)$.
- The overall objective function:

$$\min_{\mathbf{W},\mathbf{U},\mathbf{V}} \left\|\mathbf{W}^T\mathbf{X}-\mathbf{U}\mathbf{V}\right\|_F^2 + \beta\left\|\mathbf{W}\right\|_{2,1} + \alpha Tr\left(\mathbf{W}^T\mathbf{L}\mathbf{W}\right)$$
$$s.t.\quad \mathbf{W}^T\mathbf{W} = \mathbf{I},\ \mathbf{V}_{\cdot i} \in \{0,1\}^c,\ \|\mathbf{V}_{\cdot i}\|_0 = 1,\ \forall i.$$

- Considering the existence of noise and outliers increasing the risk of involving in bad local minima, we incorporate Self-Paced Learning into our **final objective function**:

$$\min_{\mathbf{w},\mathbf{r}} \sum_i r_i \ell_i\left(\mathbf{x}_i,\mathbf{w}\right) + f\left(\lambda,\mathbf{r}\right) \quad s.t.\quad r_i \in [0,1],\forall i,$$

$$\min_{\mathbf{W},\mathbf{U},\mathbf{V},\mathbf{r}\in[0,1]^n} \sum_{i=1}^{n} r_i\left\|\mathbf{W}^T\boldsymbol{x}_i - \mathbf{U}\boldsymbol{v}_i\right\|_2^2 + f\left(\lambda,\mathbf{r}\right) + \beta\left\|\mathbf{W}\right\|_{2,1} + \alpha Tr\left(\mathbf{W}^T\mathbf{L}\mathbf{W}\right)$$
$$s.t.\quad \mathbf{W}^T\mathbf{W} = \mathbf{I},\ \mathbf{V}_{\cdot i} \in \{0,1\}^c,\ \|\mathbf{V}_{\cdot i}\|_0 = 1,\ \forall i.$$

  - ➢ Please refer to our paper for the optimization.

Key References

[1] "Laplacian score for feature selection," NIPS'05.
[2] "Unsupervised feature selection for multi-cluster data," KDD'10.
[3] "$l_{2,1}$-norm regularized discriminative feature selection for unsupervised learning," IJCAI'11.
[4] "Unsupervised feature selection using nonnegative spectral analysis," AAAI'12.
[5] "Robust unsupervised feature selection," IJCAI'13.
[6] "Unsupervised simultaneous orthogonal basis clustering feature selection," CVPR'15.
[7] "Unsupervised Feature Selection with Ordinal Locality," ICME'17.