Dependence-Guided Multi-view Clustering

Xia Dong[1,3,4], Danyang Wu[1,3,4], Feiping Nie[1,3,4], Rong Wang[2,3,4], and Xuelong Li[1,3,4], *Fellow, IEEE*

[1] School of Computer Science [2] School of Cybersecurity

[3] Center for OPTical IMagery Analysis and Learning (OPTIMAL) [4] Northwestern Polytechnical University (NWPU), Xi'an, China

xiadongpgh@gmail.com, danyangwu41x@mail.nwpu.edu.cn, feipingnie@gmail.com, wangrong07@tsinghua.org.cn, li@nwpu.edu.cn

Poster Number: 3547

Problem Formulation

1. The dependence between unified embedding $\boldsymbol{P}$ and embedding of each view $\boldsymbol{F}^{(v)}$ is measured by **Hilbert Schmidt Independence Criterion (HSIC)** [1]. **2.** A cluster indicator matrix $\boldsymbol{Y}$ is recovered from the unified embedding $\boldsymbol{P}$ via estimating a rotation matrix $\boldsymbol{R} \in \mathbb{R}^{C\times C}$. Considering the orthogonality of $\boldsymbol{P}$, we transform $\boldsymbol{Y}$ into its orthogonal counterpart $\boldsymbol{Y}(\boldsymbol{Y}^\top\boldsymbol{Y})^{-\frac{1}{2}}$. Then cluster indicator matrix $\boldsymbol{Y}$ can be recovered by finding a rotation matrix $\boldsymbol{R}$ to minimize the squared Euclidean distance between $\boldsymbol{PR}$ and $\boldsymbol{Y}(\boldsymbol{Y}^\top\boldsymbol{Y})^{-\frac{1}{2}}$. **3.** To guarantee the diversity of different views, a weight is introduced for each view. Overall, the objective function of the proposed DGMC is as follows:

$$\max_{\boldsymbol{P},\boldsymbol{Y},\boldsymbol{R}} \sum_{v=1}^{V} \alpha^{(v)} \mathrm{Tr}(\boldsymbol{HPP}^\top\boldsymbol{HF}^{(v)}\boldsymbol{F}^{(v)\top}) - \lambda\|\boldsymbol{Y}(\boldsymbol{Y}^\top\boldsymbol{Y})^{-\frac{1}{2}} - \boldsymbol{PR}\|_F^2 \quad s.t.\ \boldsymbol{P}^\top\boldsymbol{P} = \boldsymbol{I}, \boldsymbol{Y} \in \mathbf{Ind}, \boldsymbol{R}^\top\boldsymbol{R} = \boldsymbol{I}, \tag{1}$$

where **Ind** defines the set of cluster indicator matrices, λ is a balance parameter, $\alpha^{(v)}$ is the pre-weight of the v-th view. Nevertheless, it is hard to determine weights without prior knowledge.

Contributions

A novel approach named Dependence-Guided Multi-view Clustering (DGMC) is proposed. The main contributions of this paper are summarized as follows.

- The proposed model enhances the dependence between unified embedding learning and clustering, as well as increases the dependence between unified embedding and embedding of each view.
- A joint framework for unified embedding learning and clustering is proposed.
- A unified embedding can be learned from different views in Reproducing Kernel Hilbert Spaces (RKHSs) to capture the high-order and non-linear dependence among these views.
- Implicit-weight learning mechanism enhances the diversity of different views.

An Equivalent Model

In this paper, an implicit-weight learning mechanism is introduced to smartly learn $\alpha^{(v)}$. To this end, we give the Remark that problem (1) is equivalent to the following problem (2) with implicit weights:

$$\max_{\boldsymbol{P},\boldsymbol{Y},\boldsymbol{R}} \sum_{v=1}^{V} (\mathrm{Tr}(\boldsymbol{HPP}^\top\boldsymbol{HF}^{(v)}\boldsymbol{F}^{(v)\top}))^r - \lambda\|\boldsymbol{Y}(\boldsymbol{Y}^\top\boldsymbol{Y})^{-\frac{1}{2}} - \boldsymbol{PR}\|_F^2 \quad s.t.\ \boldsymbol{P}^\top\boldsymbol{P} = \boldsymbol{I}, \boldsymbol{Y} \in \mathbf{Ind}, \boldsymbol{R}^\top\boldsymbol{R} = \boldsymbol{I}, \tag{2}$$

where $r > 1$ controls the curvature of weighted-learning curve. Suppose the solutions of problems (1) and (2) are $\langle \boldsymbol{P}_*, \boldsymbol{Y}_*, \boldsymbol{R}_*\rangle$ and $\langle \boldsymbol{P}_0, \boldsymbol{Y}_0, \boldsymbol{R}_0\rangle$, respectively. According to the KKT condition *w.r.t* $\boldsymbol{P}$ of problem (1), it can be easily verified that $\langle \boldsymbol{P}_0, \boldsymbol{Y}_0, \boldsymbol{R}_0\rangle = \langle \boldsymbol{P}_*, \boldsymbol{Y}_*, \boldsymbol{R}_*\rangle$ if $\alpha^{(v)}$ is calculated as $\alpha^{(v)} = r\cdot(\mathrm{Tr}(\boldsymbol{HP}_0\boldsymbol{P}_0^\top\boldsymbol{HF}^{(v)}\boldsymbol{F}^{(v)\top}))^{r-1}$.

Optimization

1. **When $\boldsymbol{P}$ and $\boldsymbol{R}$ are fixed**, problem (1) becomes

$$\min_{\boldsymbol{Y}\in\mathbf{Ind}} \|\boldsymbol{Y}(\boldsymbol{Y}^\top\boldsymbol{Y})^{-\frac{1}{2}} - \boldsymbol{PR}\|_F^2. \tag{3}$$

Further, by simply deriving problem (3), we have

$$\max_{\boldsymbol{Y}\in\mathbf{Ind},\boldsymbol{G}=\boldsymbol{PR}} \mathrm{Tr}((\boldsymbol{Y}^\top\boldsymbol{Y})^{-\frac{1}{2}}\boldsymbol{Y}^\top\boldsymbol{G}). \tag{4}$$

Let $\boldsymbol{Q} = \boldsymbol{Y}^\top\boldsymbol{Y}$, then $q_{lj} = \sum_{i=1}^N y_{il}y_{ij}$. Since $\boldsymbol{Y} \in \mathbf{Ind}$, $y_{il}y_{ij} = 0$ and $q_{lj} = 0$ hold if $l \neq j$. Thus $\boldsymbol{Q}^{-\frac{1}{2}}$ is a diagonal matrix with the l-th diagonal element as $(\boldsymbol{y}_l^\top\boldsymbol{y}_l)^{-\frac{1}{2}}$. Then problem (4) becomes

$$\max_{\boldsymbol{Y}\in\mathbf{Ind}} \sum_{l=1}^{C} \boldsymbol{y}_l^\top\boldsymbol{g}_l(\boldsymbol{y}_l^\top\boldsymbol{y}_l)^{-\frac{1}{2}}. \tag{5}$$

Since this problem is independent among different rows, we can solve $\boldsymbol{Y}$ row by row. Given an optimal $\widetilde{\boldsymbol{Y}}$, to update the i-th row $\boldsymbol{y}^i$, all we need to consider is the increment of the objective function value from $y_{il} = 0$ to $y_{il} = 1$. Therefore, Problem (5) can be solved by **coordinate descent method**.

2. **When $\boldsymbol{P}$ and $\boldsymbol{Y}$ are fixed**, problem (1) becomes

$$\min_{\boldsymbol{R}^\top\boldsymbol{R}=\boldsymbol{I}} \|\boldsymbol{Y}(\boldsymbol{Y}^\top\boldsymbol{Y})^{-\frac{1}{2}} - \boldsymbol{PR}\|_F^2, \tag{6}$$

which is the **classical orthogonal procrustes problem**. Then a closed-form solution to $\boldsymbol{R}$ is $\boldsymbol{R} = \boldsymbol{UV}^\top$, where $\boldsymbol{USV}^\top$ is the SVD of $(\boldsymbol{P}^\top\boldsymbol{Y}(\boldsymbol{Y}^\top\boldsymbol{Y})^{-\frac{1}{2}})$.

3. **When $\boldsymbol{Y}$ and $\boldsymbol{R}$ are fixed**, problem (1) becomes

$$\max_{\boldsymbol{P}^\top\boldsymbol{P}=\boldsymbol{I}} \mathrm{Tr}(\boldsymbol{P}^\top\boldsymbol{AP}) - \lambda\mathrm{Tr}(\boldsymbol{P}^\top\boldsymbol{B}), \tag{7}$$

where $\boldsymbol{A} = \boldsymbol{H}\sum_{v=1}^V w^{(v)}\boldsymbol{F}^{(v)}\boldsymbol{F}^{(v)\top}\boldsymbol{H}$, $\boldsymbol{Z} = \boldsymbol{Y}(\boldsymbol{Y}^\top\boldsymbol{Y})^{-\frac{1}{2}}$, $\boldsymbol{B} = \boldsymbol{ZR}^\top$. Problem (7) is the typical **Quadratic Problem on the Stiefel Manifold (QPSM)**, which can be solved by an efficient algorithm [2].

Theoretical Support

The Lagrangian function of problem (2) is as follows:

$$\mathcal{L} = \sum_{v=1}^{V} (\mathrm{Tr}(\boldsymbol{HPP}^\top\boldsymbol{HF}^{(v)}\boldsymbol{F}^{(v)\top}))^r - \lambda\|\boldsymbol{Y}(\boldsymbol{Y}^\top\boldsymbol{Y})^{-\frac{1}{2}} - \boldsymbol{PR}\|_F^2 - \Delta(\boldsymbol{P},\boldsymbol{Y},\boldsymbol{R},\boldsymbol{\Lambda}), \tag{8}$$

where $\Delta(\boldsymbol{P},\boldsymbol{Y},\boldsymbol{R},\boldsymbol{\Lambda})$ is the penalty term for the constraints in problem (2). Let the derivative of problem (8) *w.r.t* $\boldsymbol{P}$ be zero, and according to the chain-rule, we have

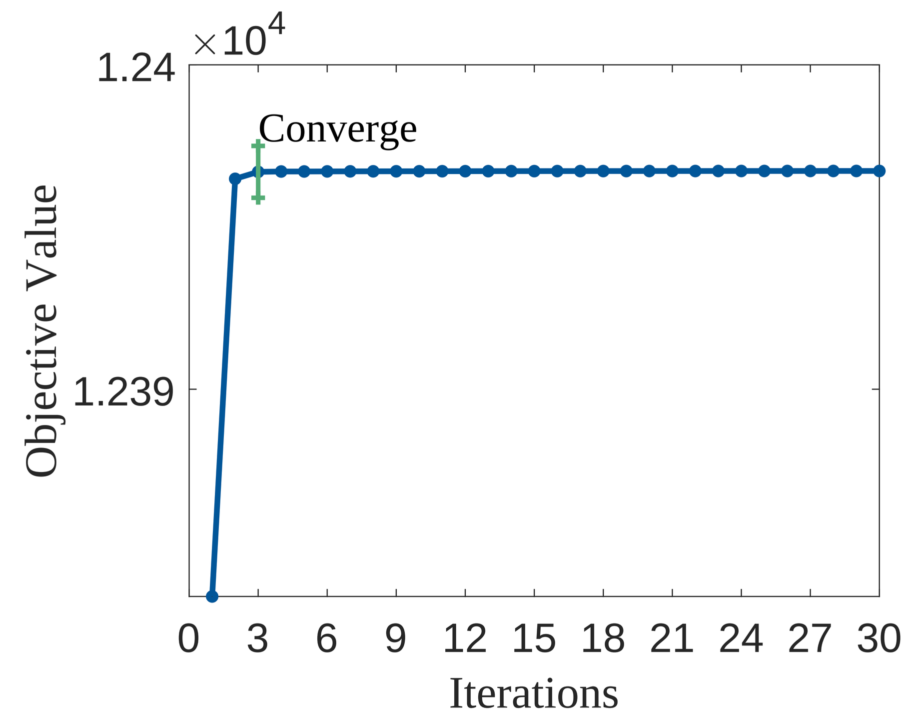
$$\sum_{v=1}^{V} \frac{\partial(\mathrm{Tr}(\boldsymbol{HPP}^\top\boldsymbol{HF}^{(v)}\boldsymbol{F}^{(v)\top}))^r}{\partial\mathrm{Tr}(\boldsymbol{HPP}^\top\boldsymbol{HF}^{(v)}\boldsymbol{F}^{(v)\top})}\frac{\partial\mathrm{Tr}(\boldsymbol{HPP}^\top\boldsymbol{HF}^{(v)}\boldsymbol{F}^{(v)\top})}{\partial\boldsymbol{P}} - \lambda\frac{\partial\|\boldsymbol{Y}(\boldsymbol{Y}^\top\boldsymbol{Y})^{-\frac{1}{2}} - \boldsymbol{PR}\|_F^2}{\partial\boldsymbol{P}} - \frac{\partial\Delta(\boldsymbol{P},\boldsymbol{Y},\boldsymbol{R},\boldsymbol{\Lambda})}{\partial\boldsymbol{P}} = 0. \tag{9}$$

By denoting $w^{(v)} \overset{\text{def}}{=} \frac{\partial(\mathrm{Tr}(\boldsymbol{HPP}^\top\boldsymbol{HF}^{(v)}\boldsymbol{F}^{(v)\top}))^r}{\partial\mathrm{Tr}(\boldsymbol{HPP}^\top\boldsymbol{HF}^{(v)}\boldsymbol{F}^{(v)\top})} = r\cdot(\mathrm{Tr}(\boldsymbol{HPP}^\top\boldsymbol{HF}^{(v)}\boldsymbol{F}^{(v)\top}))^{r-1}$ (**Eq. (10)**), then Eq. (9) becomes $\sum_{v=1}^{V} w^{(v)}\frac{\partial\mathrm{Tr}(\boldsymbol{HPP}^\top\boldsymbol{HF}^{(v)}\boldsymbol{F}^{(v)\top})}{\partial\boldsymbol{P}} - \lambda\frac{\partial\|\boldsymbol{Y}(\boldsymbol{Y}^\top\boldsymbol{Y})^{-\frac{1}{2}}-\boldsymbol{PR}\|_F^2}{\partial\boldsymbol{P}} - \frac{\partial\Delta(\boldsymbol{P},\boldsymbol{Y},\boldsymbol{R},\boldsymbol{\Lambda})}{\partial\boldsymbol{P}} = 0$, which is the KKT condition *w.r.t* $\boldsymbol{P}$ of problem (1) by letting $\alpha^{(v)} = w^{(v)}$. Overall, problem (2) can be optimized by iteratively performing the two steps: **Step 1.** update $w^{(v)}$ by Eq. (10); **Step 2.** update $\langle\boldsymbol{P},\boldsymbol{Y},\boldsymbol{R}\rangle$ by letting $\alpha^{(v)} = w^{(v)}$ and solving problem (1). Since there are three optimization variables in problem (1), we adopt alternative iterative optimization strategy to optimize them.

Experiment Results

The following Table gives the clustering results of the proposed DGMC and related four methods in terms of four metrics on three real-world datasets. The proposed DGMC shows large advantages over other methods in all the cases. Moreover, The following Figure shows the convergence curve of the proposed Algorithm on SCENE dataset. We can be see that the proposed Algorithm can converge within 3 iterations.

Dataset	Metric	AMGL	MEA	MLAN	MVGL	DGMC
MSRC	ACC	0.8571	0.8714	0.6952	0.8714	**0.9048**
	NMI	0.7623	0.7834	0.6565	0.7731	**0.8102**
	ARI	0.7081	0.7199	0.5332	0.7152	**0.7861**
	F	0.7494	0.7597	0.6089	0.7560	**0.8160**
BBCSport	ACC	0.7206	0.4963	0.7279	0.7169	**0.9062**
	NMI	0.6867	0.2347	0.7146	0.6858	**0.8230**
	ARI	0.5832	0.1492	0.6069	0.5857	**0.8510**
	F	0.7088	0.4546	0.7244	0.7098	**0.8881**
SCENE	ACC	0.5856	0.6031	0.5335	0.3251	**0.6763**
	NMI	0.5017	0.5042	0.4596	0.2093	**0.5282**
	ARI	0.3821	0.3907	0.3158	0.0728	**0.4538**
	F	0.4820	0.4901	0.4377	0.2680	**0.5243**



Algorithm Description

Algorithm to solve problem (2).

Input: Initialized $\mathbf{P}, \mathbf{R}$. $\boldsymbol{X}^{(v)}$, $\boldsymbol{F}^{(v)}$, V, λ, r

while *not converge* **do**

1. Update $\boldsymbol{w}^{(v)}$ via Eq. (10).
2. Update $\boldsymbol{Y}$ via solving problem (3).
3. Update $\boldsymbol{R}$ via solving problem (6).
4. Update $\boldsymbol{P}$ via solving problem (7).

Output: $\boldsymbol{Y}$

Reference

[1] Arthur Gretton, Olivier Bousquet, Alex Smola, and Bernhard Schölkopf. Measuring statistical dependence with hilbert-schmidt norms. In *Proceedings of the International Conference on Algorithmic Learning Theory*, pages 63–77. Springer, 2005.

[2] Feiping Nie, Rui Zhang, and Xuelong Li. A generalized power iteration method for solving quadratic problem on the stiefel manifold. *Science China Information Sciences*, 60(11):112101, 2017.