Graph-Based Multi-View Learning: Joint Optimization of Structure Preservation, Feature Selection, and Regularization

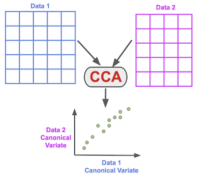
Wanjun Ning, Li Wang *Department of Mathematics, University of Texas at Arlington, Arlington, Texas*

Background

- In many real-world applications, data is collected from multiple sources or perspectives, forming multi-view datasets. Effectively processing and integrating these datasets is crucial for accurate analysis and decision-making.
- The images illustrate multi-view data in health informatics, combining vital signs, diagnostic tests, and psychological assessments for a comprehensive understanding of patient health. Effective integration of such data is crucial for advancing research and applications.

Canonical Correlation Analysis (CCA)

- A statistical method used to find linear relationships between two sets of variables.
- Aims to identify and measure the associations between these sets by finding the linear combinations that have the highest correlation.
- **Limitations**:
  - When dealing with high-dimensional data.
  - High computational complexity.
  - Doesn't enforce orthogonality, resulting in redundancy among the selected features.



Orthogonal Canonical Correlation Analysis (OCCA)

- The features extracted by OCCA are not only highly correlated but also independent of each other, reducing redundancy and improving interpretability.
- **Benefits**:
  - Reduces overfitting;
  - Good to used in multi-view datasets.
- **Comparison**: Analyzing Facial Features
  - CCA: find the linear combination of features from the eyes and nose regions that are most correlated.
  - OCCA:
    - extract the most correlated features from the eyes, ensuring they are orthogonal;
    - then do the same for the nose;
    - both highly informative and non-redundant.

Motivations

- **FSASL Limitation in Multi-View Learning:** Single-view
- **Challenges in Multi-View Learning:** orthogonality between views
- **OCCA's Optimization Challenge and the Advantages of UMvPLS:**
  - OCCA introduces useful orthogonality constraints but is not solvable with efficient methods like SVD;
  - UMvPLS maximizes variance within views and provides orthonormal projections, making it more suitable for multi-view scenarios.
- **Proposed Solution: GMvSF** – Combine FSASL's structural learning with UMvPLS's variance maximization and orthogonality to improve feature selection in high-dimensional, multi-view datasets.

Optimization Problems

Original Objectives

Let $\{\mathbf{X}_v \in \mathbb{R}^{d_v \times n}\}_{v=1}^{V}$ be the multi-view data sets.
The proposed optimization problem would be:

$$\begin{aligned} \min_{\mathbf{S},\mathbf{P}} \quad & \sum_{i,j}(\|\mathbf{P}^T\mathbf{x}_i - \mathbf{P}^T\mathbf{x}_j\|_2^2\mathbf{S}_{ij} + \mu\mathbf{S}_{ij}^2) + \gamma\|\mathbf{P}\|_{2,1} - \lambda\|\mathbf{P}^T\mathbf{X}\|_F^2 \\ \text{s.t.} \quad & \mathbf{P}^T\mathbf{P} = \mathbf{I}_m. \end{aligned} \tag{1}$$

This problem comprises two different variables with different regularization and constraints, it is hard to derive its closed solution directly.
Thus, use an alternative iterative algorithm to solve the problem, which converts the problem with a couple of variables into a series of sub problems where only one variable is involved.

Subproblem 1: Update S by fixing P

When P is fixed, we need to solve n decoupled sub problems in the following form:

$$\min_{\mathbf{s}_i} \sum_{j=1}^{n} \left(\|\mathbf{P}^T\mathbf{x}_i - \mathbf{P}^T\mathbf{x}_j\|_2^2\mathbf{S}_{ij} + \mu\mathbf{S}_{ij}^2\right), \quad \text{s.t.} \quad \mathbf{1}^T\mathbf{s}_i = 1, \quad \mathbf{S}_{ij} \geq 0. \tag{2}$$

Use the **(Complete the square)** to simplify this problem [1]:

$$\min_{\mathbf{s}_i^T} \|\mathbf{s}_i - \mathbf{q}_i\|_2^2, \quad \text{s.t.} \quad \mathbf{s}_i^T\mathbf{1}_n = 1, \quad \mathbf{s}_i \geq \mathbf{0}. \tag{3}$$

Subproblem 2: Update P by fixing S

When $\mathbf{S}$ is fixed, we need to solve the following problem:

$$\min_{\mathbf{P}} \sum_{i,j}\|\mathbf{P}^T\mathbf{x}_i - \mathbf{P}^T\mathbf{x}_j\|_2^2\mathbf{S}_{ij} + \gamma\|\mathbf{P}\|_{2,1} - \lambda\|\mathbf{P}^T\mathbf{X}\|_F^2, \quad \text{s.t.} \quad \mathbf{P}^T\mathbf{P} = \mathbf{I}_m. \tag{4}$$

Suppose $\mathbf{A} = \lambda\mathbf{X}\mathbf{X}^\top - 2\mathbf{X}\mathbf{L}_\mathbf{S}\mathbf{X}^\top - 2\gamma\mathbf{D}_\mathbf{P}$, this subproblem would be rewrote as:

$$\max_{\mathbf{P}} \quad \text{Tr}\left(\mathbf{P}^\top\mathbf{A}\mathbf{P}\right), \quad \text{s.t.} \quad \mathbf{P}^T\mathbf{P} = \mathbf{I}_m \tag{5}$$

This problem can refer to UMvPLS algorithm [2].

Improvement for Subproblem 2

But if we directly solve the Eq.5, calculating the corresponding eigenvalues and eigenvectors cost lots of computation time and storage. So we try to reduce the calculation of eigs. Let $\mathbf{B} = 2\mathbf{X}\mathbf{L}_\mathbf{S}\mathbf{X}^\top - \lambda\mathbf{X}\mathbf{X}^\top$, the problem now becomes:

$$\begin{aligned} \min_{\mathbf{P}} f(\mathbf{P}) &= \text{Tr}\left(\mathbf{P}^T\mathbf{B}\mathbf{P}\right) + 2\gamma\,\text{Tr}\left(\mathbf{P}^T\mathbf{D}\mathbf{P}\right) \\ \text{s.t.} \quad & \mathbf{P}^T\mathbf{P} = \mathbf{I}_\mathbf{m} \end{aligned} \tag{6}$$

The KKT condition for Eq.6 is $\mathbf{B}\mathbf{P} + \gamma\mathbf{D}\mathbf{P} = \mathbf{P}\mathbf{\Lambda}$, $\mathbf{P} \in \mathbb{O}^{d\times m}$, $\mathbf{\Lambda}^T = \mathbf{\Lambda} \in \mathbb{R}^{m\times m}$.

Let $E(\mathbf{P}) = \mathbf{B} + \gamma\mathbf{D}\mathbf{P}\mathbf{P}^T + \gamma\mathbf{P}\mathbf{P}^T\mathbf{D}$, then

$$E(\mathbf{P})\mathbf{P} = \mathbf{B}\mathbf{P} + \gamma\mathbf{D}\mathbf{P}\mathbf{P}^T\mathbf{P} + \gamma\mathbf{P}\mathbf{P}^T\mathbf{D}\mathbf{P} = \mathbf{P}(\mathbf{\Lambda} + \gamma\mathbf{P}^T\mathbf{D}\mathbf{P}).$$

Thus, $\mathbf{\Phi} = \mathbf{\Lambda} + \gamma\mathbf{P}^T\mathbf{D}\mathbf{P}$.

Key Observations

1. $E(\mathbf{P}) = \mathbf{B} + \gamma\mathbf{D}\mathbf{P}\mathbf{P}^T + \gamma\mathbf{P}\mathbf{P}^T\mathbf{D}$ is symmetric ✓
2. $\mathbf{\Phi} = \mathbf{\Lambda} + \gamma\mathbf{P}^T\mathbf{D}\mathbf{P} = \mathbf{P}^T\mathbf{B}\mathbf{P} + 2\gamma\mathbf{P}^T\mathbf{D}\mathbf{P}$ is symmetric ✓
3. $E(\mathbf{P}) = \mathbf{P}\mathbf{\Phi}$ ✓
4. $\text{Tr}(\mathbf{\Phi}) = \text{Tr}(\mathbf{\Lambda} + \gamma\mathbf{P}^T\mathbf{D}\mathbf{P}) = \text{Tr}(\mathbf{P}^T\mathbf{B}\mathbf{P} + 2\gamma\mathbf{P}^T\mathbf{D}\mathbf{P}) = f(\mathbf{P})$ ✓
   $E(\mathbf{P}) \in \mathbb{R}^{d\times d} \Rightarrow \mathbf{\Phi} \in \mathbb{R}^{m\times m}, \quad m \ll d$

Thus the problem tends to find m out of the smallest eigenvalues of $E(\mathbf{P})$.

Experiments

Datasets

Fig. Example of PIE and UCI Digit Dataset.

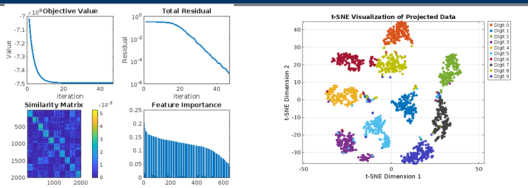
Results

Algorithm	Clustering Accuracy	NMI
AllFea	0.5609 ± 0.0546	0.5724 ± 0.0160
GLSPFS	0.6797 ± 0.0563	0.6746 ± 0.0345
JELSR_lpp	0.6761 ± 0.0660	0.6875 ± 0.0387
LapScore	0.6195 ± 0.0520	0.6016 ± 0.0265
LLCFS	0.5179 ± 0.0434	0.5448 ± 0.0141
MCFS	0.3850 ± 0.0166	0.3703 ± 0.0098
NDFS	0.6644 ± 0.0542	0.6425 ± 0.0228
SPFS	0.6353 ± 0.0527	0.6219 ± 0.0221
UDFS	0.6221 ± 0.0420	0.5931 ± 0.0172
FSASL	0.6850 ± 0.0590	0.6897 ± 0.0399
GMvSF	0.8510 ± 0.0001	0.7808 ± 0.0002

Table1: Clustering Accuracy and NMI for Different Algorithms

- FSASL [3]: simultaneously performs feature selection and structure learning;
- UDFS [4]: embeds discriminative analysis and the $l_{2,1}$-norm into the feature selection framework;
- JELSR [5]: combines embedding learning with sparse regression in an unsupervised setting;
- ...

Visualization



The convergence curves and residuals validate the optimization stability. The learned similarity matrix and feature importance indicate meaningful structure and feature selection. The t-SNE plot demonstrates the effectiveness of the learned representation in separating digit classes.

Conclusion

- Experiments on various multi-view datasets demonstrate that our framework effectively reconstructs the intrinsic data structure and outperforms state-of-the-art methods in both clustering and classification tasks.
- Visualization results further validate its ability to reveal meaningful patterns, highlighting its potential for a wide range of applications in machine learning and data analysis.

Reference

1. W. Wang, and M. A. Carreira-Perpinán, "Projection onto the probability simplex: An efficient algorithm with a simple proof, and an application." *arXiv preprint arXiv:1309.1541*, 2013.
2. L. Wang, L. Zhang, Z. Bai and R. Li, "Orthogonal canonical correlation analysis and applications," *Optimization Methods and Software*, 35(4), 787–807, 2020.
3. L. Du and Y. Shen, "Unsupervised feature selection with adaptive structure learning," *In Proceedings of the 21th ACM SIGKDD*, pp. 209–218, 2015.
4. Y. Yang, H. Shen, Z. Ma, Z. Huang and X. Zhou, "$L_{2,1}$-norm regularized discriminative feature selection for unsupervised learning," *In IJCAI International Joint Conference on Artificial Intelligence*, 2011.
5. C. Hou, F. Nie, X. Li, D. Yi and Y. Wu, "Joint Embedding Learning and Sparse Regression: A Framework for Unsupervised Feature Selection," *IEEE Trans. Cybern.*, 2014, 44, 793–804.

Contact: wxn2289@mavs.uta.edu