Learning to Select Features

Xianchao Xiu

Department of Automation

上海大学
Shanghai University

The Hong Kong Polytechnic University, March 20, 2025

Joint work with Anning Yang (SHU), Long Chen (SHU), Jianhao Li (SHU) and others

Outline

Introduction

Sparse Coding

Contrastive Learning

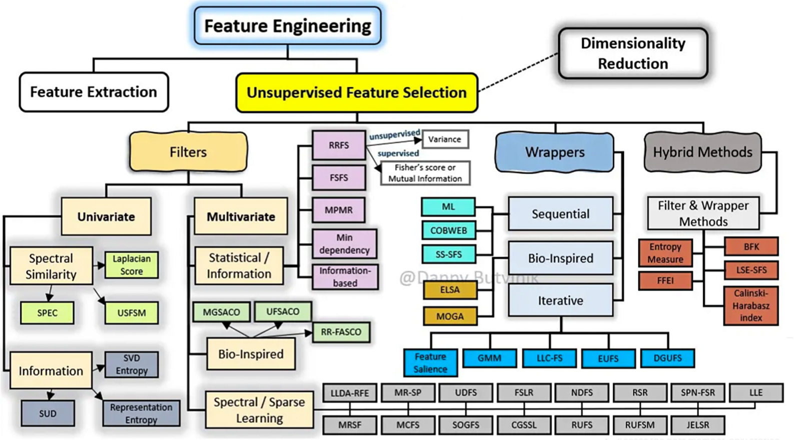
Deep Unfolding Networks

Large Language Models

Future Work

UFS

- Unsupervised feature selection *vs.* Feature extraction
- Select a subset of input features without labels



PCA

- Given $X = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n) \in \mathbb{R}^{d\times n}$, principal component analysis (PCA) is

$$\begin{aligned} \min_{W\in\mathbb{R}^{d\times p}} \quad & \frac{1}{2}\|X - WW^\top X\|_{\mathrm{F}}^2 \\ \mathrm{s.t.} \quad & W^\top W = I_p \end{aligned}$$

$$\Updownarrow$$

$$\begin{aligned} \min_{W\in\mathbb{R}^{d\times p}} \quad & -\mathrm{Tr}(W^\top XX^\top W) \\ \mathrm{s.t.} \quad & W^\top W = I_p \end{aligned}$$

- Unsupervised feature selection by sparse PCA

$$\begin{aligned} \min_{W\in\mathbb{R}^{d\times p}} \quad & -\mathrm{Tr}(W^\top XX^\top W) \\ \mathrm{s.t.} \quad & W^\top W = I_p, \ \|W\|_{2,0} \le s \end{aligned}$$

- The i-th feature can be measured by $\|\mathbf{w}^i\|$ since $\mathbf{z}_i = (\mathbf{w}^{1\top}, \ \mathbf{w}^{2\top}, \ \ldots \ , \mathbf{w}^{d\top})\mathbf{x}_i$
- The dimension number is often omitted when it does not cause ambiguity

SOTA

- Li-Nie-Bian et al, Sparse PCA via $\ell_{2,p}$-Norm Regularization for Unsupervised Feature Selection, IEEE TPAMI, 2023

$$\begin{aligned} \min_{W} \quad & -\mathrm{Tr}(W^\top X X^\top W) + \lambda \|W\|_{2,p}^p \ (0 < p < 1) \\ \text{s.t.} \quad & W^\top W = I \end{aligned}$$

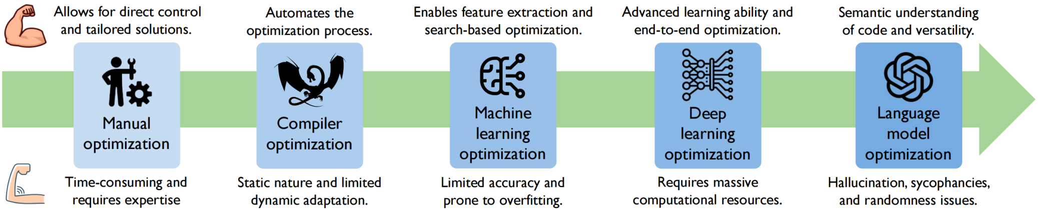
- Li-Sun-Zhang, Unsupervised Feature Selection via Nonnegative Orthogonal Constrained Regularized Minimization, arXiv:2403.16966

$$\begin{aligned} \min_{W,Y} \quad & \mathrm{Tr}(Y^\top L Y) + \alpha \|Y - X^\top W\|_{2,1} + \beta \|W\|_{2,1} + \gamma \|W\|_{\mathrm{F}}^2 \\ \text{s.t.} \quad & Y^\top Y = I, \ Y \geq 0 \end{aligned}$$

- Hu-Wang-Zhang et al, Bi-Level Spectral Feature Selection, IEEE TNNLS, 2025
- Jiao-Xue-Zhang, Sparse Learning-Based Feature Selection in Classification: A Multi-Objective Perspective, IEEE TETCI, 2025
- Li-Yu-Yang et al, Exploring Feature Selection With Limited Labels: A Comprehensive Survey of Semi-Supervised and Unsupervised Approaches, IEEE TKDE, 2024

Contribution

- Learning to select features
  - (Q1) How to learn feature structures $\Rightarrow$ Sparse coding
  - (Q2) How to learn data distributions $\Rightarrow$ Contrastive learning
  - (Q3) How to learn regularization parameters $\Rightarrow$ Deep unfolding networks
  - (Q4) How to learn feature selection $\Rightarrow$ Large language models



Outline

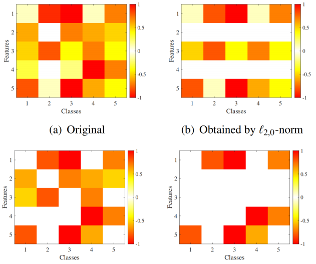
Model

- (Q1) How to learn feature structures

$$\begin{aligned} \min_{W} \quad & -\mathrm{Tr}(W^\top X X^\top W) \\ \text{s.t.} \quad & W^\top W = I, \; \|W\|_{2,0} \le s_1 \end{aligned}$$

$$\Downarrow$$

$$\begin{aligned} \min_{W} \quad & -\mathrm{Tr}(W^\top X X^\top W) \\ \text{s.t.} \quad & W^\top W = I, \; \|W\|_{2,0} \le s_1, \; \|W\|_0 \le s_2 \end{aligned}$$



(a) Original (b) Obtained by $\ell_{2,0}$-norm (c) Obtained by ℓ_0-norm (d) Obtained by double sparsity

- Double Sparsity Constrained Optimization for Feature Selection (DSCOFS)
  - $\|W\|_{2,0} \le s_1$: Global feature selection
  - $\|W\|_0 \le s_2$: Local feature selection

Algorithm

- Proximal alternating method (PAM)
- Model reformulation

$$\begin{aligned} \min_{W} \quad & -\mathrm{Tr}(W^\top X X^\top W) \\ \text{s.t.} \quad & W^\top W = I,\ \|W\|_{2,0} \le s_1,\ \|W\|_0 \le s_2 \end{aligned}$$

$$\Downarrow$$

$$\begin{aligned} \min_{W,Y,Z} \quad & -\mathrm{Tr}(W^\top X X^\top W) \\ \text{s.t.} \quad & W^\top W = I,\ \|Y\|_{2,0} \le s_1,\ \|Z\|_0 \le s_2 \\ & W = Y,\ W = Z \end{aligned}$$

$$\Downarrow$$

$$\begin{aligned} \min_{W,Y,Z} \quad & -\mathrm{Tr}(W^\top X X^\top W) + \mu_1 \|W - Y\|_{\mathrm{F}}^2 + \mu_2 \|W - Z\|_{\mathrm{F}}^2 \\ \text{s.t.} \quad & W^\top W = I,\ \|Y\|_{2,0} \le s_1,\ \|Z\|_0 \le s_2 \end{aligned}$$

Algorithm

- Input: $X, \mu_1, \mu_2, s_1, s_2, \tau_1, \tau_2, \tau_3$
- Initialize: (W^0, Y^0, Z^0)
- While not converged do
  - Update W^{k+1} by

$$\begin{aligned} \min_{W} \quad & -\mathrm{Tr}(W^\top X X^\top W) + \mu_1 \|W - Y^k\|_\mathrm{F}^2 + \mu_2 \|W - Z^k\|_\mathrm{F}^2 + \tau_1 \|W - W^k\|_\mathrm{F}^2 \\ \text{s.t.} \quad & W^\top W = I \end{aligned}$$

  - Update Y^{k+1} by

$$\begin{aligned} \min_{Y} \quad & \|W^{k+1} - Y\|_\mathrm{F}^2 + \tau_2 \|Y - Y^k\|_\mathrm{F}^2 \\ \text{s.t.} \quad & \|Y\|_{2,0} \le s_1 \end{aligned}$$

  - Update Z^{k+1} by

$$\begin{aligned} \min_{Z} \quad & \|W^{k+1} - Z\|_\mathrm{F}^2 + \tau_3 \|Z - Z^k\|_\mathrm{F}^2 \\ \text{s.t.} \quad & \|Z\|_0 \le s_2 \end{aligned}$$

Convergence

- Denote the objective function as

$$f(W, Y, Z) = -\operatorname{Tr}(W^\top X X^\top W X) + \mu_1 \|W - Y\|_{\mathrm{F}}^2 + \mu_2 \|W - Z\|_{\mathrm{F}}^2$$

- Suppose that $\beta \geq \max\{2(\lambda_0 + \lambda_1), 2m\lambda_2\}$
- (Theorem) Let $\{(W^k, Y^k, Z^k)\}$ be the generated sequence. Then the following properties hold:
  - $\{f(W^k, Y^k, Z^k)\}$ is strictly nonincreasing
  - The sequence $\{(W^k, Y^k, Z^k)\}$ is bounded
  - $\lim_{k\to\infty} \|(W^{k+1}, Y^{k+1}, Z^{k+1}) - (W^k, Y^k, Z^k)\|_{\mathrm{F}} = 0$
  - Any accumulation point (W^*, Y^*, Z^*) of the sequence $\{(W^k, Y^k, Z^k)\}$ is a stationary point in the sense that

$$0 \in \nabla f(W^*, Y^*, Z^*) + \mathrm{N}(W^*, Y^*, Z^*)$$

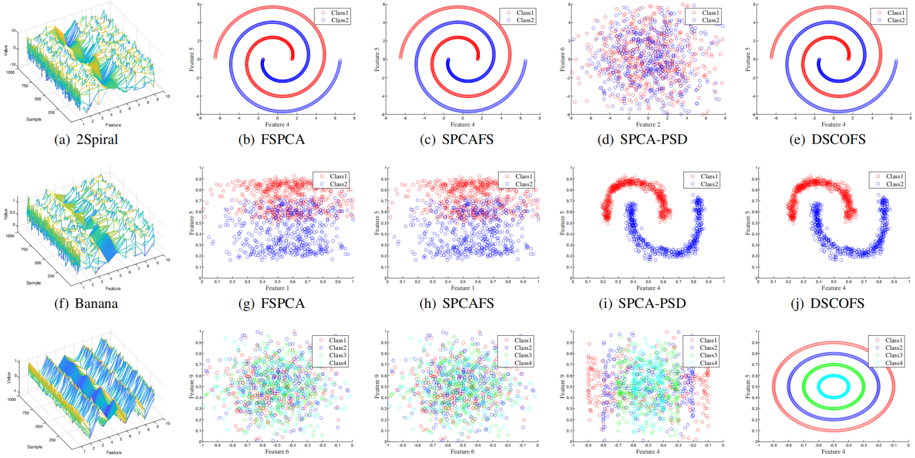
Experiments

- Compared methods
  - LapScore: He-Cai-Niyogi, NIPS, 2005
  - UDFS: Yang-Shen-Ma et al, IJCAI, 2011
  - SOGFS: Nie-Zhu-Li, IEEE TKDE, 2021
  - RNE: Liu-Ye-Li-Wang et al, KBS, 2020
  - SPCAFS: Li-Nie-Bian-Wu et al, IEEE TPAMI, 2023
  - FSPCA: Nie-Tian-Wang et al, IEEE TPAMI, 2023
  - SPCA-PSD: Zheng-Zhang-Liu et al, arXiv:2309.06202
- Implementation setups
  - Initialization: RandOrthhMat
  - Sparsity level: $s_1 \in \{10, 20, \ldots, 100\}$, $s_2 \in \{0.1, 0.2, \ldots, 0.9\}dp$
  - Stopping criteria:

$$\frac{|f(X^{k+1}, Y^{k+1}, Z^{k+1}) - f(X^k, Y^k, Z^k)|}{1 + |f(X^k, Y^k, Z^k)|} \leq 10^{-3}$$

Experiments

- Synthetic datasets



(a) 2Spiral (b) FSPCA (c) SPCAFS (d) SPCA-PSD (e) DSCOFS

(f) Banana (g) FSPCA (h) SPCAFS (i) SPCA-PSD (j) DSCOFS

(k) Dartboard (l) FSPCA (m) SPCAFS (n) SPCA-PSD (o) DSCOFS

Experiments

- Real datasets: Accuracy (ACC) $\uparrow$

Datasets	ALLfea	LapScore	UDFS	SOGFS	RNE	FSPCA	SPCAFS	SPCA-PSD	DSCOFS
COIL20	57.74±4.93	54.82±3.91 (100)	**58.71±3.47 (100)**	49.66±4.81 (100)	55.84±4.41 (90)	50.15±4.70 (100)	54.39±3.67 (100)	56.57±3.05 100)	**60.51±4.63 (100)**
USPS	65.12±4.95	62.02±4.09 (90)	59.52±2.97 (60)	55.58±3.07 (100)	46.04±2.69 (100)	**67.38±4.36 (60)**	67.34±4.49 (100)	65.38±4.26 (100)	**69.67±4.97 (100)**
lung_discrete	65.10±6.44	59.29±6.33 (70)	68.58±6.99 (100)	65.12±6.89 (100)	64.05±6.65 (100)	60.19±6.55 (40)	71.37±7.68 (100)	**72.22±8.02 (80)**	**73.12±8.48 (100)**
GLIOMA	56.84±5.24	58.88±3.96 (90)	56.80±4.85 (100)	57.44±6.16 (70)	58.32±7.31 (90)	47.92±4.61 (80)	50.60±5.02 (20)	**59.28±5.01 (90)**	**60.88±6.31 (80)**
UMIST	41.07±2.38	40.13±2.79 (100)	47.12±2.49 (40)	41.70±3.17 (100)	40.35±2.26 (90)	46.70±2.29 (100)	46.78±2.51 (90)	**47.98±2.91 (90)**	**48.10±3.01 (70)**
warpPIE10P	25.67±1.90	28.94±1.66 (100)	41.42±3.18 (20)	46.90±3.89 (20)	29.57±2.96 (90)	28.01±2.27 (50)	**48.76±3.86 (50)**	43.74±3.91 (70)	**49.00±3.88 (40)**
Isolet	57.89±3.82	52.21±2.76 (100)	41.95±2.07 (100)	49.31±2.32 (100)	47.12±2.06 (90)	**53.62±2.36 (100)**	53.04±2.33 (100)	51.91±2.15 (70)	**59.67±3.46 (100)**
MSTAR	77.04±7.98	67.87±3.49 (90)	78.15±5.80 (90)	73.74±5.89 (100)	69.16±6.03 (100)	75.52±6.22 (70)	**80.80±5.95 (100)**	79.70±6.43 (90)	**82.59±7.41 (100)**
Average	55.81±4.71	53.02±3.62	56.53±4.04	54.93±4.53	51.31±4.30	53.69±4.47	59.14±4.44	**59.60±4.47**	**62.94±5.27**

Experiments

- Real datasets: Normalized mutual information (NMI) $\uparrow$

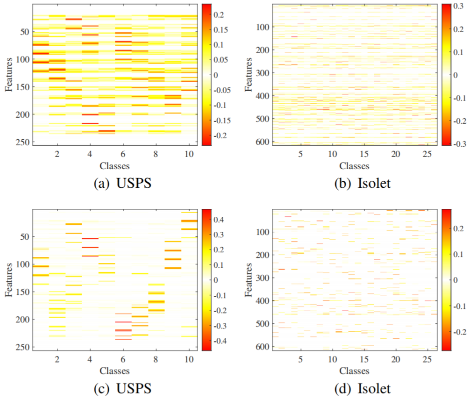
Datasets	ALLfea	LapScore	UDFS	SOGFS	RNE	FSPCA	SPCAFS	SPCA-PSD	DSCOFS
COIL20	75.37±1.96	69.59±1.48 (100)	**73.54±1.76 (100)**	68.92±1.84 (100)	70.43±1.92 (100)	68.50±1.56 (100)	69.98±1.45 (100)	69.85±1.41 (100)	**76.25±1.71 (100)**
USPS	61.12±2.01	59.46±1.80 (100)	54.69±2.11 (100)	52.96±1.54 (100)	45.36±1.93 (90)	**62.00±1.87 (60)**	60.98±2.37 (100)	60.90±2.02 (100)	**64.06±2.58 (100)**
lung_discrete	62.85±5.13	56.79±3.99 (100)	64.84±5.09 (100)	59.70±5.24 (100)	61.63±5.83 (70)	58.26±6.39 (40)	69.09±5.61 (100)	**70.93±5.46 (80)**	**70.98±7.00 (100)**
GLIOMA	48.86±5.72	51.03±2.48 (100)	47.22±3.53 (10)	48.67±10.98 (100)	48.62±6.32 (100)	21.94±5.28 (100)	24.14±6.97 (100)	**51.44±5.62 (90)**	**51.06±6.19 (80)**
UMIST	63.67±1.85	61.16±1.71 (100)	62.00±1.58 (100)	60.79±1.54 (100)	55.92±1.57 (70)	65.27±1.58 (100)	66.23±1.60 (90)	**66.25±1.72 (100)**	**67.24±1.85 (100)**
warpPIE10P	25.07±2.88	25.13±1.73 (90)	46.18±3.30 (20)	52.12±3.25 (20)	32.67±3.31 (90)	23.90±2.01 (50)	**52.63±3.33 (50)**	46.02±3.70 (70)	**52.65±3.29 (50)**
Isolet	75.72±1.70	69.77±1.20 (100)	56.29±1.11 (100)	67.40±1.44 (100)	64.27±0.95 (90)	**70.79±1.12 (100)**	67.71±1.33 (100)	69.69±0.80 (100)	**75.01±1.35 (100)**
MSTAR	82.42±3.31	74.10±1.76 (100)	76.45±2.47 (90)	76.39±1.70 (100)	66.87±1.99 (80)	78.39±2.17 (90)	**80.33±2.50 (100)**	79.17±2.77 (90)	**81.14±3.13 (100)**
Average	61.89±3.07	58.38±2.02	60.15±2.62	60.87±3.44	55.72±2.98	56.13±2.75	61.39±3.15	**64.28±2.94**	**67.30±3.39**

Experiments

- Ablation studies: Feature similarity rate (FSR)

$$\mathrm{FSR} = \frac{|\mathbb{T}_{\mathrm{our}} \cap \mathbb{T}_{2,0}|}{n}$$

Datasets	$\|W\|_0 \leq s_2$	ACC ↑	NMI ↑	FSR
COIL20	×	60.25±4.52	75.89±1.58	84
	√	60.51±4.42	76.25±1.71	
USPS	×	68.69±4.79	61.25±2.39	68
	√	69.67±4.97	64.06±2.58	
lung_discrete	×	71.42±7.95	69.74±6.11	92
	√	73.12±8.48	70.98±7.00	
GLIOMA	×	58.24±5.04	49.76±6.12	85
	√	60.88±6.31	51.06±6.19	
UMIST	×	47.33±3.05	67.44±1.88	95
	√	48.10±3.01	67.24±1.85	
warpPIE10P	×	47.91±4.99	51.19±3.79	89
	√	49.00±3.88	52.65±3.29	
Isolet	×	57.29±3.44	72.82±1.87	52
	√	59.67±3.46	75.01±1.35	
MSTAR	×	82.06±6.87	81.01±2.41	99
	√	82.59±7.41	81.14±3.13	



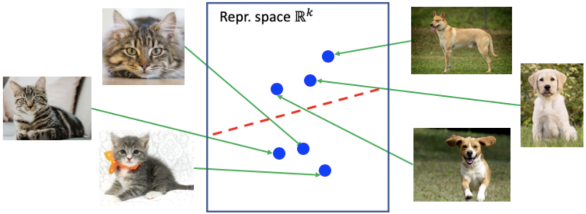
Outline

Motivation

- (Q2) How to learn data distributions

$$
\begin{aligned}
\min_{W} \quad & \frac{1}{2}\|X - WW^\top X\|_{\mathrm{F}}^2 \\
\text{s.t.} \quad & W^\top W = I, \ \|W\|_{2,0} \le s_1, \ \|W\|_0 \le s_2
\end{aligned}
$$

  - Convex loss: ℓ_1-norm, $\ell_{2,1}$-norm, quantile, Huber
  - Nonconvex loss: ℓ_p-norm, $\ell_{2,p}$-norm, SCAD, MCP, capped ℓ_1
- Contrastive learning: learn a discrimination model between positive and negative pairs



Motivation

- Let $X = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n)$ and $Y = (\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n)$ be two different pairs, the contrastive loss is defined as

$$L_c(X, Y) = \frac{1}{2n} \sum_{i=1}^{n} (L_c(\mathbf{x}_i) + L_c(\mathbf{y}_i))$$

$$L_c(\mathbf{x}_i) = -\log \frac{\exp(s(\mathbf{x}_i, \mathbf{y}_i)/\tau)}{\sum_{j=1, j\neq i}^{n} \exp(s(\mathbf{x}_i, \mathbf{x}_j)/\tau) + \sum_{j=1}^{n} \exp(s(\mathbf{x}_i, \mathbf{y}_j)/\tau)}$$

$$L_c(\mathbf{y}_i) = -\log \frac{\exp(s(\mathbf{y}_i, \mathbf{x}_i)/\tau)}{\sum_{j=1, j\neq i}^{n} \exp(s(\mathbf{y}_i, \mathbf{y}_j)/\tau) + \sum_{j=1}^{n} \exp(s(\mathbf{y}_i, \mathbf{x}_j)/\tau)}$$

- $s(\mathbf{x}, \mathbf{y}) = \mathbf{x}^\top \mathbf{y}$ is the similarity metric, τ is the temperature parameter

A simple framework for contrastive learning of visual representations

T Chen, S Kornblith, M Norouzi… - … on machine learning, 2020 - proceedings.mlr.press

… In our contrastive learning, as positive pairs are computed in the same device, the model can exploit the local information leakage to improve prediction accuracy without improving …

Model

- DSCOFS with contrastive learning (DSCOFS-CL)

$$\begin{aligned} \min_{W} \quad & \frac{1}{2}\|X - WW^\top X\|_{\mathrm{F}}^2 \\ \text{s.t.} \quad & W^\top W = I,\ \|W\|_{2,0} \le s_1,\ \|W\|_0 \le s_2 \end{aligned}$$

$$\Downarrow$$

$$\begin{aligned} \min_{W} \quad & L_c(X, WW^\top X) \\ \text{s.t.} \quad & W^\top W = I,\ \|W\|_{2,0} \le s_1,\ \|W\|_0 \le s_2 \end{aligned}$$

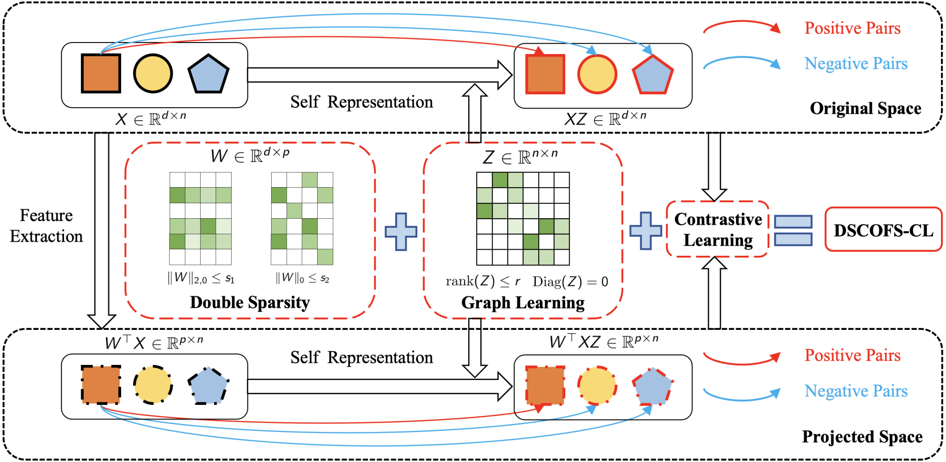
$$\Downarrow$$

$$\begin{aligned} \min_{W,Z} \quad & \lambda L_c(X, XZ) + (1-\lambda) L_c(W^\top X, W^\top XZ) \\ \text{s.t.} \quad & W^\top W = I,\ \|W\|_{2,0} \le s_1,\ \|W\|_0 \le s_2 \\ & \mathrm{rank}(Z) \le r,\ \mathrm{Diag}(Z) = 0 \end{aligned}$$

  - $\mathrm{rank}(Z) \le r$ represents the global structure
  - $\mathrm{Diag}(Z) = 0$ avoids the case where $Z = E$

Architecture

- DSCOFS-CL = Double Sparsity + Graph Learning + Contrastive Learning



Algorithm

- Proximal alternating method (PAM)

$$\begin{aligned}\min_{W,Z}\quad & \lambda L_c(X, XZ) + (1-\lambda) L_c(W^\top X, W^\top XZ)\\ \text{s.t.}\quad & W^\top W = I,\ \|W\|_{2,0} \le s_1,\ \|W\|_0 \le s_2\\ & \operatorname{rank}(Z) \le r,\ \operatorname{Diag}(Z) = 0\end{aligned}$$

$$\Downarrow$$

$$\begin{aligned}\min_{W,Z,Y,P,Q}\quad & \lambda L_c(X, XZ) + (1-\lambda) L_c(W^\top X, W^\top XZ)\\ \text{s.t.}\quad & \|P\|_{2,0} \le s_1,\ \|Q\|_0 \le s_2, \operatorname{rank}(Y) \le r,\ \operatorname{Diag}(Z) = 0\\ & W^\top W = I,\ Z = Y, W = P,\ W = Q\end{aligned}$$

$$\Downarrow$$

$$\begin{aligned}\min_{W,Z,Y,P,Q}\quad & \lambda L_c(X, XZ) + (1-\lambda) L_c(W^\top X, W^\top XZ) + \mu\|W^\top W - I\|_{\mathrm{F}}^2\\ & + \alpha\|Z - Y\|_{\mathrm{F}}^2 + \beta\|W - P\|_{\mathrm{F}}^2 + \gamma\|W - Q\|_{\mathrm{F}}^2\\ \text{s.t.}\quad & \|P\|_{2,0} \le s_1,\ \|Q\|_0 \le s_2,\ \operatorname{rank}(Y) \le r,\ \operatorname{Diag}(Z) = 0\end{aligned}$$

Algorithm

- Input: X, λ, μ, α, β, γ, s_1, s_2, r, τ_1, τ_2, τ_3, τ_4, τ_5
- Initialize: $(W^0, Z^0, Y^0, P^0, Q^0)$
- While not converged do
  - Update W^{k+1} by

$$
\begin{aligned}
\min_W \quad & (1-\lambda)L_c(W^\top X, W^\top XZ^k) + \mu\|W^\top W - I\|_{\mathrm{F}}^2 \\
& + \beta\|W - P^k\|_{\mathrm{F}}^2 + \gamma\|W - Q^k\|_{\mathrm{F}}^2 + \tau_1\|W - W^k\|_{\mathrm{F}}^2
\end{aligned}
$$

  - Update Z^{k+1} by

$$
\begin{aligned}
\min_Z \quad & \lambda L_c(X, XZ) + (1-\lambda)L_c(W^{k+1,\top}X, W^{k+1,\top}XZ) \\
& + \alpha\|Z - Y^k\|_{\mathrm{F}}^2 + \tau_2\|Z - Z^k\|_{\mathrm{F}}^2 \\
\text{s.t.} \quad & \mathrm{Diag}(Z) = 0
\end{aligned}
$$

  - Update Y^{k+1}
  - Update P^{k+1}
  - Update Q^{k+1}

Algorithm

- Define

$$
\begin{aligned}
f(W,Z,Y,P,Q) = {} & \lambda L_c(X,XZ) + (1-\lambda) L_c(W^\top X, W^\top XZ) \\
& + \mu \|W^\top W - I\|_{\mathrm{F}}^2 + \alpha \|Z - Y\|_{\mathrm{F}}^2 + \beta \|W - P\|_{\mathrm{F}}^2 + \gamma \|W - Q\|_{\mathrm{F}}^2 \\
& + \delta(Z) + \delta(Y) + \delta(P) + \delta(Q)
\end{aligned}
$$

- We call (W,Z,Y,P,Q) is a critical point if $0 \in \partial f(W,Z,Y,P,Q)$
- (Theorem) For each k, the sequence $\{(W^k, Z^k, Y^k, P^k, Q^k)\}$ generated by our PAM algorithm converges and $0 \in \partial f(W^*, Z^*, Y^*, P^*, Q^*)$ with

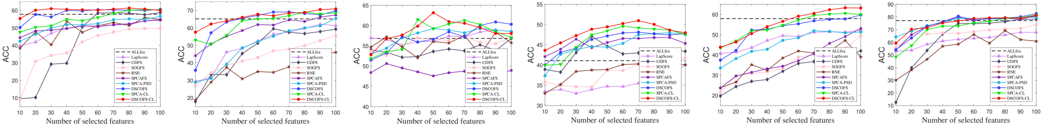
$$
\lim_{k \to +\infty} (W^k, Z^k, Y^k, P^k, Q^k) = (W^*, Z^*, Y^*, P^*, Q^*)
$$

  - Sufficient decreasing
  - Lower bounds for iterations
  - Kurdyka-Lojasiewicz properties

Experiments

- Real datasets: Accuracy (ACC) ↑

Datasets	ALLfea	LapScore	UDFS	SOGFS	RNE	SPCAFS	SPCA-PSD	DSCOFS	SPCA-CL	DSCOFS-CL
COIL20	57.74±4.93	54.82±3.91 (100)	58.71±3.47 (100)	49.66±4.81 (100)	55.84±4.41 (90)	54.39±3.67 (100)	56.57±3.05 (100)	**60.51±4.63 (100)**	60.31±3.49 (90)	**61.32±5.18 (80)**
USPS	65.12±4.95	62.02±4.09 (90)	59.52±2.97 (60)	55.58±3.07 (100)	46.04±2.69 (100)	67.34±4.49 (100)	65.38±4.26 (100)	**69.67±4.97 (100)**	68.88±4.05 (80)	**70.82±4.77 (100)**
GLIOMA	56.84±5.24	58.88±3.96 (90)	56.80±4.85 (100)	57.44±6.16 (70)	58.32±7.31 (90)	50.60±5.02 (20)	59.28±5.01 (90)	60.88±6.31 (80)	**61.48±6.20 (40)**	**63.16±7.46 (50)**
UMIST	41.07±2.38	40.13±2.79 (100)	47.12±2.49 (40)	41.70±3.17 (100)	40.35±2.26 (90)	46.78±2.51 (90)	47.98±2.91 (90)	48.10±3.01 (70)	**49.55±3.00 (60)**	**50.95±3.15 (70)**
Isolet	57.89±3.82	52.21±2.76 (100)	41.95±2.07 (100)	49.31±2.32 (100)	47.12±2.06 (90)	53.04±2.33 (100)	51.91±2.15 (70)	59.67±3.46 (100)	**60.53±3.75 (90)**	**63.22±3.50 (90)**
MSTAR	77.04±7.98	67.87±3.49 (90)	78.15±5.80 (90)	73.74±5.89 (100)	69.16±6.03 (100)	80.80±5.95 (100)	79.70±6.43 (90)	**82.59±7.41 (100)**	**81.57±6.28 (100)**	81.22±5.59 (100)
Average	59.28±4.88	55.99±3.50	57.04±3.69	54.57±4.24	52.81±4.13	58.83±4.00	60.14±3.97	63.57±4.96	**63.72±4.46**	**65.12±4.94**

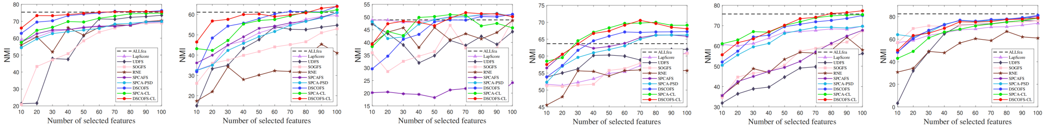


(a) COIL20 (b) USPS (c) GLIOMA (d) UMIST (e) Isolet (f) MSTAR

Experiments

- Real datasets: Normalized mutual information (NMI) ↑

Datasets	ALLfea	LapScore	UDFS	SOGFS	RNE	SPCAFS	SPCA-PSD	DSCOFS	SPCA-CL	DSCOFS-CL
COIL20	75.37±1.96	69.59±1.48 (100)	73.54±1.76 (100)	68.92±1.84 (100)	70.43±1.92 (100)	69.98±1.45 (100)	69.85±1.41 (100)	**76.25±1.71 (100)**	74.79±1.48 (100)	**75.76±1.76 (90)**
USPS	61.12±2.01	59.46±1.80 (100)	54.69±2.11 (100)	52.96±1.54 (100)	45.36±1.93 (90)	60.98±2.37 (100)	60.90±2.02 (100)	**64.06±2.58 (100)**	62.29±2.40 (100)	**63.95±2.67 (100)**
GLIOMA	48.86±5.72	51.03±2.48 (100)	47.22±3.53 (10)	48.67±10.98 (100)	48.62±6.32 (100)	24.14±6.97 (100)	**51.44±5.62 (90)**	51.06±6.19 (80)	50.95±4.10 (60)	**51.71±5.03 (70)**
UMIST	63.67±1.85	61.16±1.71 (100)	62.00±1.58 (100)	60.79±1.54 (100)	55.92±1.57 (70)	66.23±1.60 (90)	66.25±1.72 (100)	67.24±1.85 (100)	**69.98±1.84 (80)**	**70.54±1.70 (70)**
Isolet	75.72±1.70	69.77±1.20 (100)	56.29±1.11 (100)	67.40±1.44 (100)	64.27±0.95 (90)	67.71±1.33 (100)	69.69±0.80 (100)	75.01±1.35 (100)	**75.41±1.51 (100)**	**77.32±1.37 (100)**
MSTAR	82.42±3.31	74.10±1.76 (100)	76.45±2.47 (90)	76.39±1.70 (100)	66.87±1.99 (80)	**80.33±2.50 (100)**	79.17±2.77 (90)	**81.14±3.13 (100)**	78.63±2.50 (100)	78.88±1.60 (100)
Average	67.86±2.76	64.19±1.74	61.70±2.09	62.52±3.17	58.58±2.45	61.56±2.70	66.22±2.39	**69.13±2.80**	68.68±2.31	**69.69±2.36**



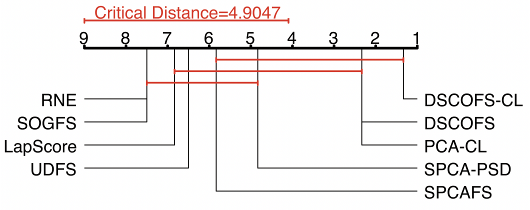
(a) COIL20 (b) USPS (c) GLIOMA (d) UMIST (e) Isolet (f) MSTAR

Experiments

- Friedman tests ($\mathcal{H}_0$: There is no significant difference of compared methods)

Methods	Ranking	P-value	Hypothesis
LapScore	6.83	0.00001	Reject
UDFS	6.50		
SOGFS	7.50		
RNE	7.50		
SPCAFS	5.83		
SPCA-PSD	4.83		
DSCOFS	2.33		
SPCA-CL	2.33		
DSCOFS-CL	1.33		

- Post-hoc Nemenyi tests



Outline

Motivation

- (Q3) How to learn regularization parameters

$$
\begin{aligned}
\min_{W,Z,Y,P,Q} \quad & \lambda L_c(X, XZ) + (1-\lambda) L_c(W^\top X, W^\top XZ) + \mu \|W^\top W - I\|_{\mathrm{F}}^2 \\
& + \alpha \|Z - Y\|_{\mathrm{F}}^2 + \beta \|W - P\|_{\mathrm{F}}^2 + \gamma \|W - Q\|_{\mathrm{F}}^2 \\
\text{s.t.} \quad & \|P\|_{2,0} \le s_1, \ \|Q\|_0 \le s_2, \ \mathrm{rank}(Y) \le r, \ \mathrm{Diag}(Z) = 0
\end{aligned}
$$

  - $\mu, \alpha, \beta, \gamma \in \{10^{-6}, 10^{-4}, 10^{-2}, 10^{0}, 10^{2}, 10^{4}, 10^{6}\}$
  - $s_1 \in \{10, 20, \ldots, 100\}$
  - $s_2 \in \{0.1, 0.2, \ldots, 0.5\} dp$
  - $r = 0.1d$
  - $\lambda = 0.5$
- From iterative optimization to deep unfolding networks
  - Gregor-LeCun, Learning Fast Approximations of Sparse Coding, ICML, 2010
  - Chen-Liu-Yin, Learning to optimize: A Tutorial for Continuous and Mixed-Integer Optimization, SCCM, 2024

Model

- Consider structured sparse PCA

$$\begin{aligned}\min_{W} \quad & \frac{1}{2}\|X - WW^\top X\|_{\mathrm{F}}^2 + \lambda\|W\|_{2,1} + \mu\|W\|_1 \\ \text{s.t.} \quad & W^\top W = I\end{aligned}$$

- Alternating direction method of multipliers (ADMM)

$$\begin{aligned}\min_{W} \quad & \frac{1}{2}\|X - WW^\top X\|_{\mathrm{F}}^2 + \lambda\|Y\|_{2,1} + \mu\|Z\|_1 \\ \text{s.t.} \quad & W^\top W = I, \; W = Y, \; W = Z\end{aligned}$$

$$\Downarrow$$

$$\begin{aligned}\mathcal{L}(W, Y, Z, \Lambda, \Pi) = {} & \frac{1}{2}\|X - WW^\top X\|_{\mathrm{F}}^2 + \lambda\|Y\|_{2,1} + \mu\|Z\|_1 \\ & + \langle \Lambda, W - Y\rangle + \frac{\alpha}{2}\|W - Y\|_{\mathrm{F}}^2 + \langle \Pi, W - Z\rangle + \frac{\beta}{2}\|W - Z\|_{\mathrm{F}}^2\end{aligned}$$

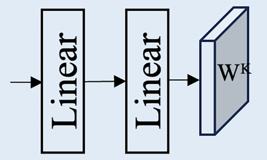
SPCA-Net

- Update W-block

$$\begin{aligned} \min_{W} \quad & f(W) := \frac{1}{2}\|X - WW^\top X\|_\mathrm{F}^2 + \frac{\alpha}{2}\|W - Y^k + \Lambda^k/\alpha\|_\mathrm{F}^2 + \frac{\beta}{2}\|W - Z^k + \Pi^k/\beta\|_\mathrm{F}^2 \\ \text{s.t.} \quad & W^\top W = I \end{aligned}$$

$$\Downarrow$$

$$\begin{aligned} \min_{W} \quad & f(W^k) + \langle \nabla f(W^k), W - W^k \rangle + \frac{1}{2\eta}\|W - W^k\|_\mathrm{F}^2 \\ \text{s.t.} \quad & W^\top W = I \end{aligned}$$



$$\Downarrow$$

$$W^{k+1} = UV^\top$$

$$\Downarrow$$

$$W^{k+1} = \textbf{LargNet}(U, V^\top)$$

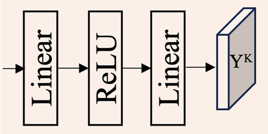
SPCA-Net

- Update Y-block

$$\min_{Y} \ \lambda\|Y\|_{2,1} + \frac{\alpha}{2}\|X^{k+1} - Y + \Lambda^k/\alpha\|_{\mathrm{F}}^2$$

$$\Downarrow$$

$$Y^{k+1} = \mathrm{sign}(\|X^{k+1} + \Lambda^k/\alpha\|_2) \circ \max(\|X^{k+1} + \Lambda^k/\alpha\|_2 - \lambda/\alpha, 0)$$



$$\Downarrow$$

$$Y^{k+1} = \frac{X^{k+1} + \Lambda^k/\alpha}{\|X^{k+1} + \Lambda^k/\alpha\|_2} \mathrm{ReLU}(\|X^{k+1} + \Lambda^k/\alpha\|_2 - \lambda/\alpha)$$

$$\Downarrow$$

$$Y^{k+1} = \textbf{GSoftNet}(X^{k+1} + \Lambda^k/\alpha, \lambda/\alpha)$$

SPCA-Net

- Update Z-block

$$\min_{Z} \ \mu\|Z\|_1 + \frac{\beta}{2}\|X^{k+1} - Z + \Pi^k/\beta\|_{\mathrm{F}}^2$$

$$\Downarrow$$

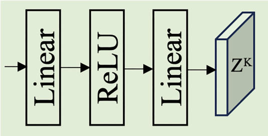
$$Z^{k+1} = \mathrm{sign}(X^{k+1} + \Pi^k/\beta) \circ \max(|X^{k+1} + \Pi^k/\beta| - \mu/\beta, 0)$$

$$\Downarrow$$

$$Z^{k+1} = \frac{X^{k+1} + \Pi^k/\beta}{|X^{k+1} + \Pi^k/\beta|}\mathrm{ReLU}(|X^{k+1} + \Pi^k/\beta| - \mu/\beta)$$

$$\Downarrow$$

$$Z^{k+1} = \textbf{SoftNet}(X^{k+1} + \Pi^k/\beta, \mu/\beta)$$



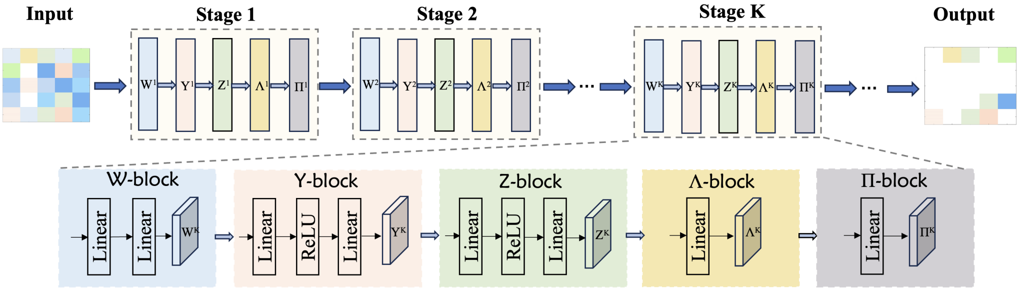
SPCA-Net

- Input: X, $\lambda, \mu, \alpha, \beta$
- Initialize: $(W^0,\ Y^0,\ Z^0,\ \Lambda^0,\ \Pi^0)$
- While $k = 1, \ldots, K$ do
  - Update W^{k+1} by
$$W^{k+1} = \textbf{LargNet}(U, V^\top)$$
  - Update Y^{k+1} by
$$Y^{k+1} = \textbf{GSoftNet}(X^{k+1} + \Lambda^k/\alpha, \lambda/\alpha)$$
  - Update Z^{k+1} by
$$Z^{k+1} = \textbf{SoftNet}(X^{k+1} + \Pi^k/\beta, \mu/\beta)$$
  - Update Λ^{k+1}, Π^{k+1} by
$$\Lambda^{k+1} = \textbf{Linear}(W^{k+1}, Y^{k+1}, \Lambda^k, \alpha),\ \ \Pi^{k+1} = \textbf{Linear}(W^{k+1}, Z^{k+1}, \Pi^k, \beta)$$
- Output: Trained W

Architecture

- All papameters $(\lambda, \mu, \alpha, \beta)$ are trained in an end-to-end manner
- The loss is defined as

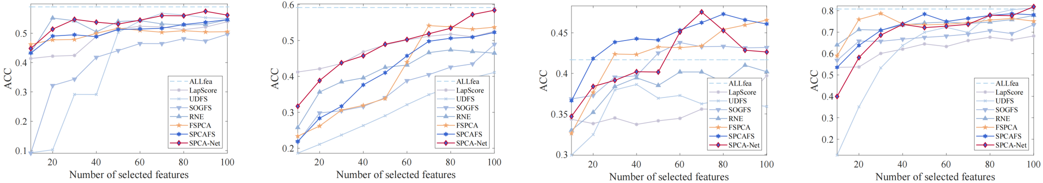
$$\mathrm{Loss} = \frac{1}{2}\|X - \bar{W}\bar{W}^{\top}X\|_{\mathrm{F}}^{2} + \lambda\|\bar{W}\|_{2,1} + \mu\|\bar{W}\|_{1}$$



Experiments

- Real datasets: Accuracy (ACC) ↑

Datasets	ALLfea	LapScore	UDFS	SOGFS	RNE	FSPCA	SPCAFS	SPCA-Net
COIL20	58.97±4.99 (10)	53.91±3.61 (100)	**56.70±3.09 (70)**	49.66±3.63 (100)	55.16±3.35 (20)	51.71±3.05 (50)	54.63±3.64 (100)	**57.46±2.76 (90)**
Isolet	59.18±3.19 (10)	52.55±2.83 (100)	41.11±1.71 (100)	48.93±2.69 (100)	47.39±2.91 (80)	**54.15±2.69 (70)**	52.26±2.81 (100)	**58.43±4.31 (100)**
UMIST	41.68±2.46 (10)	39.71±3.28 (100)	38.64±1.61 (40)	43.81±2.98 (80)	41.01±2.25 (90)	46.58±2.34 (100)	**47.32±3.48 (80)**	**47.58±4.97 (70)**
MSTAR	80.81±8.76 (10)	68.21±4.57 (100)	**81.25±7.48 (100)**	73.46±5.61 (100)	77.82±6.16 (100)	78.74±5.20 (30)	78.63±8.68 (90)	**81.90±6.87 (100)**

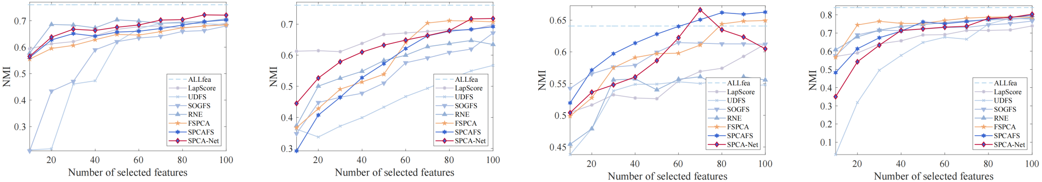


(a) COIL20 (b) Isolet (c) UMIST (d) MSTAR

Experiments

- Real datasets: Normalized mutual information (NMI) ↑

Datasets	ALLfea	LapScore	UDFS	SOGFS	RNE	FSPCA	SPCAFS	SPCA-Net
COIL20	76.04±1.69 (10)	69.01±1.53 (100)	69.12±1.17 (80)	68.03±1.59 (100)	**70.76±2.07 (100)**	68.41±1.60 (100)	70.29±1.31 (100)	**72.21±2.68 (90)**
Isolet	76.09±1.77 (10)	69.86±1.26 (100)	56.73±1.05 (100)	67.15±1.45 (100)	64.74±1.28 (90)	**71.12±1.11 (80)**	69.18±1.33 (100)	**71.80±1.59 (100)**
UMIST	64.07±1.76 (10)	61.23±2.15 (100)	55.43±1.50 (80)	61.46±2.03 (70)	56.08±1.80 (60)	64.94±1.65 (100)	**66.26±1.74 (100)**	**66.62±7.52 (70)**
MSTAR	83.96±3.14 (10)	73.90±1.62 (100)	78.18±3.64 (90)	76.56±1.54 (100)	78.26±2.51 (100)	78.87±2.52 (90)	**79.62±2.30 (100)**	**80.67±3.47 (90)**



(a) COIL20 (b) Isolet (c) UMIST (d) MSTAR

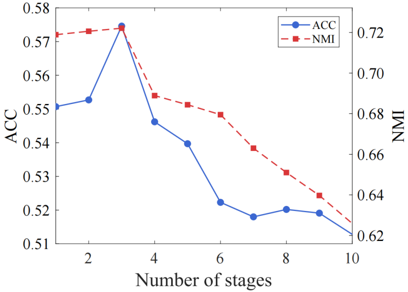
Experiments

- Ablation studies

Datasets	Network	ACC ↑	NMI ↑
COIL20	×	**55.12±2.67**	**70.44±1.37**
	√	**57.46±2.76**	**72.21±2.68**
Isolet	×	**51.84±2.82**	**67.02±1.43**
	√	**58.43±4.31**	**71.80±1.59**
UMIST	×	**40.65±2.29**	**55.88±1.62**
	√	**47.58±4.97**	**66.62±7.52**
MSTAR	×	**80.65±6.47**	**80.53±2.41**
	√	**81.90±6.87**	**80.67±3.47**

Datasets	Dynamic	ACC ↑	NMI ↑
COIL20	×	**56.71±3.83**	**71.49±3.67**
	√	**57.46±2.76**	**72.21±2.68**
Isolet	×	**52.06±3.71**	**68.91±2.36**
	√	**58.43±4.31**	**71.80±1.59**
UMIST	×	**42.63±2.78**	**60.12±1.69**
	√	**47.58±4.97**	**66.62±7.52**
MSTAR	×	**80.74±5.28**	**80.59±3.67**
	√	**81.90±6.87**	**80.67±3.47**

- Effect of deep unfolding stages



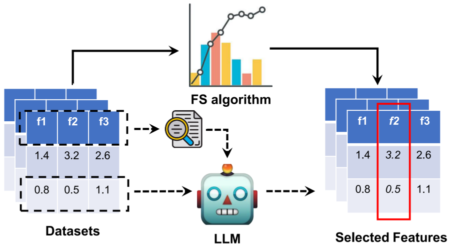
Outline

Motivation

- (Q4) How to learn feature selection

$$\min_{W} \quad -\mathrm{Tr}(W^\top X X^\top W)$$
$$\text{s.t.} \quad W^\top W = I, \ \|W\|_{2,0} \le s_1$$

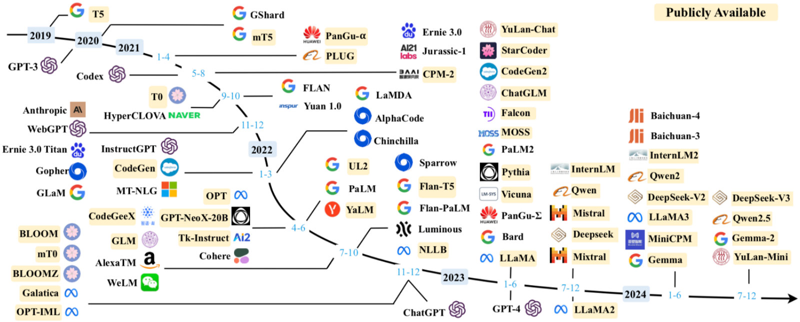
$\Rightarrow$



- From deep learning to large language models (LLMs)
  - Cho-Cund-Srivastava et al, LMPriors: Pre-Trained Language Models as Task-Specific Priors, NeurIPS, 2022
  - Han-Yoon-Arik et al, Large Language Models Can Automatically Engineer Features for Few-Shot Tabular Learning, ICML, 2024
  - Li-Tan-Liu, Exploring Large Language Models for Feature Selection: A Data-centric Perspective, SIGKDD, 2025

DeepSeek

- Guo-Yang-Zhang et al, DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning, arXiv:2501.12948
- Arrieta-Ugarte-Valle et al, o3-mini vs DeepSeek-R1: Which One is Safer? arXiv:2501.18438
- Muennighoff-Yang-Shi et al, S1: Simple Test-time Scaling, arXiv:2501.19393
- Gao-Jin-Ke et al, A Comparison of DeepSeek and Other LLMs, arXiv:2502.03688



Method

- **Dataset-specific Context**

  Using data collected via a telemarketing campaign at a Portuguese banking institution from 2008 to 2013, we wish to build a machine learning model that can predict whether a client will subscribe to a term deposit (target variable). The dataset contains a total of 16 features (e.g., age, marital status, whether the client has a housing loan). Prior to training the model, we first want to identify a subset of the 16 features that are most important for reliable prediction of the target variable.

- **Main System Prompt**

  For each feature input by the user, your task is to provide a feature importance score (between ⟨0.0⟩ and ⟨1.0⟩; larger value indicates greater importance) for predicting whether an individual will subscribe to a term deposit and a reasoning behind how the importance score was assigned. The results need to be written directly into a JSON file. Therefore, please do not include any extra text and return the results strictly in the given format. The scores for each feature should be different from one another.

- **Output Format Instruction**

  Here is an example output: "concept-1": "has credit in default ", "reasoning": "Clients with credits in default might be more hesitant to open new financial products due to their current financial situation and may be deemed a higher risk by the bank. Therefore, the score is 0.9.", "score": 0.9.

- **Main User Prompt**

  Provide a score and reasoning formatted according to the output schema above.

Method

- ## Dataset-specific Context

  Using data collected via a telemarketing campaign at a Portuguese banking institution from 2008 to 2013, we wish to build a machine learning model that can predict whether a client will subscribe to a term deposit (target variable). The dataset contains a total of 16 features (e.g., age, marital status, whether the client has a housing loan). Prior to training the model, we first want to identify a subset of the 16 features that are most important for reliable prediction of the target variable.

- ## Main System Prompt

  Please use the Random Forest (/ forward sequential selection / backward sequential selection / recursive feature elimination RFE / minimum redundancy maximum relevance selection MRMR / filtering by mutual information MI) model to directly analyze the dataset samples. This is a classification task, where "Class" represents the classification. Please analyze the importance scores of these features. The score range is [0.0, 1.0], and the score of each feature should be different. The output format is as follows, in JSON file format.

- ## Output Format Instruction

  Here is an example output: "concept-1": "has credit in default ", "reasoning": "Clients with credits in default might be more hesitant to open new financial products due to their current financial situation and may be deemed a higher risk by the bank. Therefore, the score is 0.9.", "score": 0.9.

- ## Main User Prompt

  Provide a score and reasoning formatted according to the output schema above.

Expriments

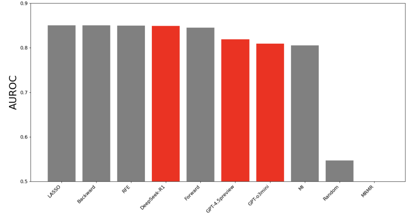
- Compared methods
  - DeepSeek-R1 (2025-01-20)
  - GPT-o3mini (2025-01-31)
  - GPT-4.5preview (2025-02-27)
  - LASSO
  - Forward sequential selection (Forward)
  - Backward sequential selection (Backward)
  - Recursive feature elimination (RFE)
  - Minimum redundancy maximum relevance selection (MRMR)
  - Mutual information (MI)
  - Random feature selection (Random)

- Statistics of datasets

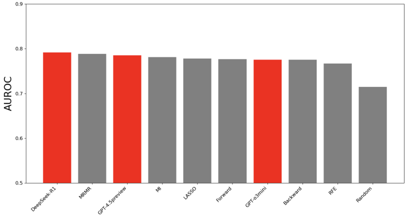
Datasets	Samples	Features
Bank	45211	16
Credit-G	1000	20
Pima Indians Diabetes	768	8
Give Me Some Credit	120269	10

Experiments

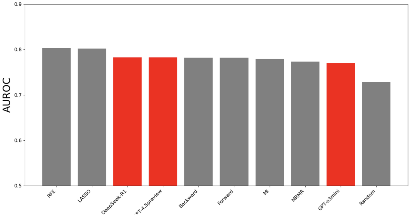
- LLMs vs. Data-driven methods



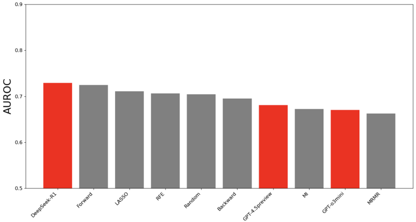
(a) Bank



(b) Credit-G



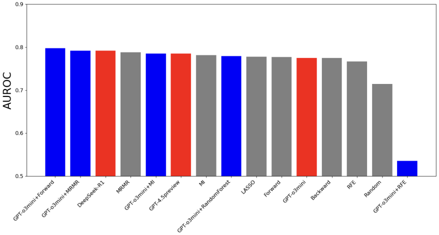
(c) Pima Indians Diabetes

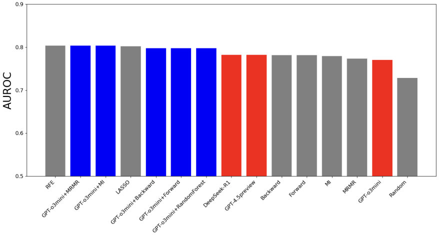


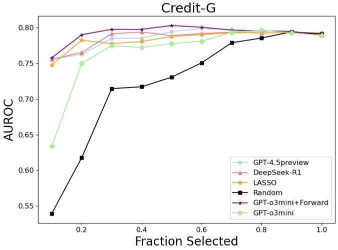
(d) Give Me Some Credit

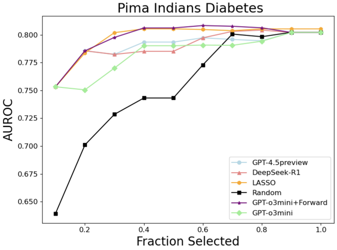
Experiments

- LLMs + Data-driven methods vs. LLMs vs. Data-driven methods



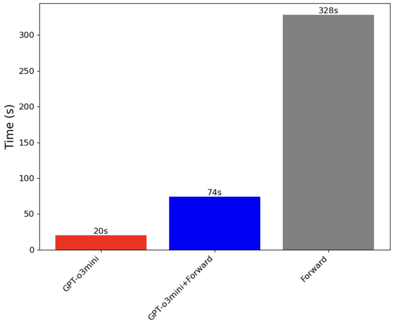


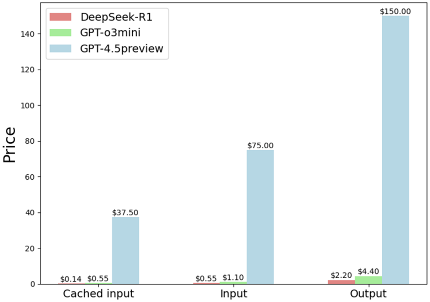




Experiments

- More interesting things should be investigated
  - Consider large datasets with more features, especially larger than thousands
  - Apply DeepSeek-R1 with different parameters, including 7B, 14B, 32B, 70B
  - Try RAG and fine-tuning to improve the stability and reliability
  - Expand to regression tasks, analyze feature correlation, etc





Outline

LLMs for Optimization

- Ramamonjison-Yu-Li et al, NL4Opt Competition: Formulating Optimization Problems Based on Their Natural Language Descriptions, NeurIPS, 2022
- Yang-Wang-Lu et al, Large Language Models as Optimizers, ICLR, 2024
- AhmadiTeshnizi-Gao-Udell, OptiMUS: Scalable Optimization Modeling with (MI)LP Solvers and Large Language Models, ICML, 2024
- Romera-Paredes-Barekatain et al, Mathematical Discoveries from Program Search with Large Language Models, Nature, 2024
- Jiang-Shu-Qian et al, LLMOPT: Learning to Define and Solve General Optimization Problems from Scratch, ICLR, 2025
- Fan-Ghaddar-Wang et al, Artificial Intelligence for Operations Research: Revolutionizing the Operations Research Process, arXiv:2401.03244
- Liu-Yao-Guo et al, A Systematic Survey on Large Language Models for Algorithm Design, arXiv:2410.14716
- Gong-Voskanyan-Brookes et al, Language Models for Code Optimization: Survey, Challenges and Future Directions, arXiv:2501.01277

Pruning for LLMs

- Frantar-Alistarh, SparseGPT: Massive Language Models Can be Accurately Pruned in One-Shot, ICML, 2023
- Sun-Liu-Bair et al, A Simple and Effective Pruning Approach for Large Language Models, ICLR, 2024
- Fang-Yin-Muralidharan et al, MaskLLM: Learnable Semi-Structured Sparsity for Large Language Models, NeurIPS, 2024
- Boza, Fast and Effective Weight Update for Pruned Large Language Models, TMLR, 2024
- Hu-Zhao-Li et al, FASP: Fast and Accurate Structured Pruning of Large Language Models, arXiv:2501.09412
- Cheng-Zhang-Shi, A Survey on Deep Neural Network Pruning: Taxonomy, Comparison, Analysis, and Recommendations, IEEE TPAMI, 2024
- Wan-Wang-Liu et al, Efficient Large Language Models: A Survey, TMLR, 2024

Thank you for your attention

xcxiu@shu.edu.cn