ICANN19
28th International Conference on Artificial Neural Networks

Joint Dictionary Learning for Unsupervised Feature Selection

Yang Fan[1], Jianhua Dai[2] , Qilai Zhang[1] and Shuai Liu[2]

1. College of Intelligence and Computing, Tianjin University

2. College of Information Science and Engineering, Hunan Normal University

Email: fany@tju.edu.cn and jhdai@hunnu.edu.cn

Introduction

The selected features should model data distribution, preserve data reconstruction and maintain manifold structure. However, most UFS methods don't consider these three factors simultaneously. Motivated by this, we propose a novel joint dictionary learning method, which handles these three key factors simultaneously. In joint dictionary learning, an intrinsic space shared by feature space and pseudo label space is introduced, which can model cluster structure and reveal data reconstruction. To ensure the sparseness of intrinsic space, the l_1-norm regularization is imposed on the representation coefficients matrix. The joint learning of robust sparse regression model and spectral clustering can select features that maintain data distribution and manifold structure.

Approach

To select the most representative features, we consider the key factors. Under the framework of joint dictionary learning, the consistent intrinsic space of samples is adaptively learned by feature space and pseudo label space, and the accurate cluster labels are reconstructed by the intrinsic space. By leveraging the interactions between these two goals, we can preserve the data reconstruction well and capture accurate cluster structure.

$$\begin{aligned}\min_{\mathbf{D}_x,\mathbf{D}_u,\mathbf{A},\mathbf{U}} \quad & ||\mathbf{X}-\mathbf{D}_x\mathbf{A}||_F^2+||\mathbf{U}^T-\mathbf{D}_u\mathbf{A}||_F^2+\beta||\mathbf{A}||_1 \\ \text{s.t.} \quad & ||\mathbf{d}_{xi}||_2^2\le 1, ||\mathbf{d}_{ui}||_2^2\le 1, \forall i\end{aligned}$$

Since the cluster structure can reveal the data distribution of instances well, we introduce a feature selection matrix W to preserve the cluster structure via linear sparse regression, by which the original features can be projected into corresponding clusters. The feature selection framework based on data distribution is formulated as:

$$\min_{\mathbf{W}} \quad \alpha||\mathbf{U}-\mathbf{X}^T\mathbf{W}||_F^2+\delta||\mathbf{W}||_{2,1}$$

We expect that the obtained pseudo label space can also preserve the local geometric structure of the samples. In other words, similar samples should be grouped into the same cluster.

$$\min_{\mathbf{U}} \quad \gamma\mathrm{Tr}(\mathbf{U}^T\mathbf{L}\mathbf{U}), \quad \text{s.t.} \quad \mathbf{U}^T\mathbf{U}=\mathbf{I}, \mathbf{U}\ge\mathbf{0}$$

Putting (1), (2), and (3) together, the proposed approach JDLUFS is to solve the following optimization model:

$$\begin{aligned}\min_{\mathbf{D}_x,\mathbf{D}_u,\mathbf{A},\mathbf{W},\mathbf{U}} \quad & ||\mathbf{X}-\mathbf{D}_x\mathbf{A}||_F^2+||\mathbf{U}^T-\mathbf{D}_u\mathbf{A}||_F^2+\beta||\mathbf{A}||_1 \\ + \quad & \alpha||\mathbf{U}-\mathbf{X}^T\mathbf{W}||_F^2+\gamma\mathrm{Tr}(\mathbf{U}^T\mathbf{L}\mathbf{U})+\delta||\mathbf{W}||_{2,1} \\ \text{s.t.} \quad & ||\mathbf{d}_{xi}||_2^2\le 1, ||\mathbf{d}_{ui}||_2^2\le 1, \forall i, \mathbf{U}^T\mathbf{U}=\mathbf{I}, \mathbf{U}\ge\mathbf{0}\end{aligned}$$

We employ an alternating optimization strategy to solve the proposed optimization problem.

$$\begin{cases}\mathbf{D}_x^{t+1}=\arg\min_{\mathbf{D}_x}||\mathbf{X}-\mathbf{D}_x\mathbf{A}||_F^2+\mu||\mathbf{D}_x-\mathbf{H}^t+\mathbf{S}^t||_F^2 \\ \mathbf{H}^{t+1}=\arg\min_{\mathbf{H}}\mu||\mathbf{D}_x^{t+1}-\mathbf{H}^t+\mathbf{S}^t||_F^2, \quad \text{s.t.} \quad ||\mathbf{h}_i||_2^2\le 1, \forall i \\ \mathbf{S}^{t+1}=\mathbf{S}^t+\mathbf{D}_x^{t+1}-\mathbf{H}^{t+1}, \quad update\ \mu\ if\ appropriate\end{cases}$$

$$\hat{\mathbf{A}}=(\mathbf{D}_x^T\mathbf{D}_x+\mathbf{D}_u^T\mathbf{D}_u+\beta\mathbf{I})^{-1}(\mathbf{D}_x^T\mathbf{X}+\mathbf{D}_u^T\mathbf{U})$$

$$\mathbf{U}_{ij}=\frac{(\mathbf{A}^T\mathbf{D}_u^T+\alpha\mathbf{X}^T\mathbf{W}+2\lambda\mathbf{U})_{ij}}{(\mathbf{U}+\alpha\mathbf{U}+\gamma\mathbf{L}\mathbf{U}+2\lambda\mathbf{U}\mathbf{U}^T\mathbf{U})_{ij}}\mathbf{U}_{ij}$$

$$\mathbf{W}^{t+1}=\mathbf{G}^{t-1}\mathbf{X}(\mathbf{X}^T\mathbf{G}^{t-1}+\frac{\delta}{\alpha}\mathbf{I})^{-1}\mathbf{U}$$

Algorithm 1 The JDLUFS algorithm.

Input: Data matrix $\mathbf{X}\in\mathbb{R}^{d\times n}$; Parameters $p, c, \alpha, \beta, \gamma$ and δ; Dictionary size k; Loss variation ratio σ; Maximum number of iterations N;

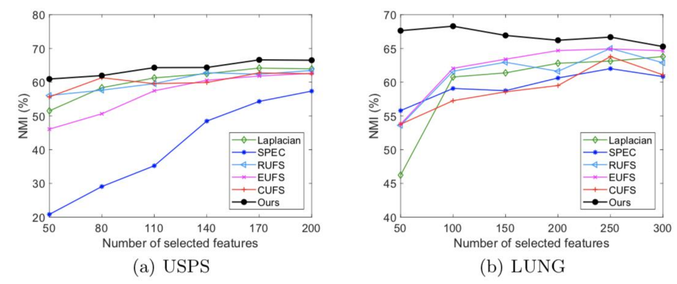
Output: t features from the dataset.

1: Initialize $\alpha=1$, $\lambda=10^6$, $\mathbf{D}_x$, $\mathbf{D}_u$, $\mathbf{W}$ and $\mathbf{U}$;
2: Compute affinity graph $\mathbf{S}$ and Laplacian matrix $\mathbf{L}$;
3: **repeat**
4: $\quad\hat{\mathbf{A}}=(\mathbf{D}_x^T\mathbf{D}_x+\mathbf{D}_u^T\mathbf{D}_u+\beta\mathbf{I})^{-1}(\mathbf{D}_x^T\mathbf{X}+\mathbf{D}_u^T\mathbf{U})$;
5: $\quad$Update $\mathbf{D_x}$ and $\mathbf{D_u}$ by calculating iteration problem (6);
6: $\quad\mathbf{U}_{ij}=\frac{(\mathbf{A}^T\mathbf{D}_u^T+\alpha\mathbf{X}^T\mathbf{W}+2\lambda\mathbf{U})_{ij}}{(\mathbf{U}+\alpha\mathbf{U}+\gamma\mathbf{L}\mathbf{U}+2\lambda\mathbf{U}\mathbf{U}^T\mathbf{U})_{ij}}\mathbf{U}_{ij}$
7: $\quad$Update $\mathbf{W}$ by solving (13) using IRLS;
8: **until** Up to the maximum number of iterations or loss variation ratio
9: Sort all d features according to $||\mathbf{w}_i||_2$ in descending order and select the top-t ranked features.

Experiments

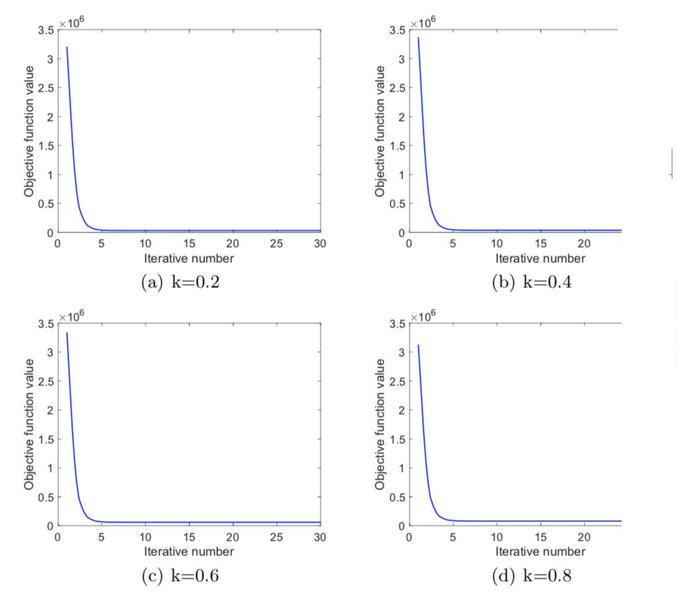
We conduct extensive comparative experiments to evaluate our algorithm in terms of both classification and clustering performance.

Dataset	USPS	COIL20	warpPIE10P	ISOLET	LUNG	Pixraw10P
	Classification results(Classification Accuracy %± std)					
Laplacian	94.08±3.65	85.94±6.05	94.28±3.56	76.56±4.82	91.78±1.17	87.50±15.24
SPEC	78.66±23.71	57.14±21.19	97.19±5.60	73.30±8.03	89.77±2.25	68.00±20.06
RUFS	95.41±5.70	94.31±11.44	99.89±3.56	84.04±8.15	93.61±1.65	97.83±1.78
EUFS	95.47±3.53	93.01±8.83	99.75±0.72	80.95±10.15	93.13±1.56	97.50±4.34
CUFS	95.86±1.82	93.49±4.52	99.91±0.44	84.16±3.00	91.13±1.40	99.00±3.53
Ours(0.2)	96.01±2.81	**95.22 ± 3.68**	99.92±1.90	85.63±3.81	94.14±1.53	99.00±2.80
Ours(0.4)	96.09±2.88	95.06±3.67	99.93±1.88	86.57±3.02	94.14±1.53	**99.17±2.99**
Ours(0.6)	**96.11±2.86**	95.01±3.61	**99.95±1.82**	**86.77±3.97**	**94.43±1.56**	99.00±2.89
Ours(0.8)	96.10±2.76	94.89±3.39	99.83±1.83	85.03±3.77	94.14±1.63	99.00±2.33
	Clustering results(NMI %± std)					
Laplacian	61.08±4.88	65.68±2.97	22.36±2.57	70.62±2.77	59.78±6.60	79.43±13.61
SPEC	47.85±23.71	50.54±21.19	52.77±5.60	66.51±8.03	59.09±2.25	66.13±20.06
RUFS	62.55±3.52	73.38±6.96	42.28±4.21	76.62±8.80	65.50±5.17	87.35±3.60
EUFS	62.27±2.52	67.61±2.16	**68.50±12.91**	70.47±6.37	59.11±7.55	79.76±4.88
CUFS	61.89±1.69	71.63±2.47	48.23±4.99	75.23±3.87	59.52±2.61	81.95±5.66
Ours(0.2)	62.98±1.63	**73.98±2.39**	44.50±3.30	78.14±4.44	67.15±2.08	90.88±6.22
Ours(0.4)	**63.03±1.73**	73.87±2.47	39.04±3.17	77.74±4.49	**67.38±2.10**	90.48±6.20
Ours(0.6)	62.94±1.75	73.97±2.49	41.23±3.25	**78.53±4.67**	66.90±2.20	**91.42±6.22**
Ours(0.8)	63.02±1.66	73.87±2.23	41.58±3.22	77.63±4.36	67.22±2.29	90.94±6.73
	Clustering results(ACC %± std)					
Laplacian	64.58±4.28	53.19±2.93	21.44±1.54	55.49±3.12	70.49±8.68	66.22±17.20
SPEC	50.38±12.99	35.72±10.04	42.98±5.90	52.01±4.06	64.85±3.29	56.95±10.41
RUFS	66.43±3.10	61.01±7.44	34.53±2.31	62.03±9.61	79.61±4.39	81.20±4.43
EUFS	66.46±2.98	54.46±1.92	**58.93±10.31**	56.86±6.66	71.71±6.53	74.00±5.04
CUFS	65.06±2.03	60.52±3.23	41.42±3.76	62.15±5.21	69.70±2.46	75.97±5.56
Ours(0.2)	**67.74±2.47**	61.21±3.75	37.10±3.04	64.26±4.92	82.34±3.65	83.00±5.35
Ours(0.4)	67.58±2.70	61.24±3.79	32,60±3.99	64.92±4.01	**83.15±3.97**	83.35±5.37
Ours(0.6)	67.49±2.61	**61.71±3.74**	34.37±3.92	**65.59±4.21**	79.38±3.89	**84.22±5.52**
Ours(0.8)	67.67±2.57	61.43±3.53	35.02±4.02	63.46±4.78	82.33±3.10	83.57±5.10



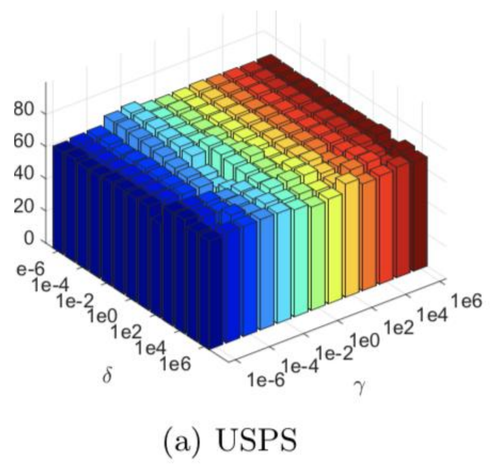
(a) USPS (b) LUNG

Fig. 1. Clustering performance (NMI) of all the methods on USPS and LUNG datasets.

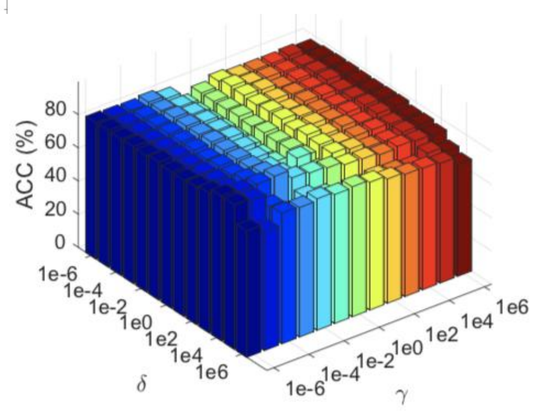


(a) k=0.2 (b) k=0.4 (c) k=0.6 (d) k=0.8

Fig. 3. Convergence analysis on Pixraw10P dataset with different diction



(a) USPS



(b) LUNG

We can see that JDLUFS achieves better Classification Accuracy, NMI and ACC than compared methods, which means our model tends to select representative features. We can observe that the clustering performance doesn't vary much, which indicates that our method is not very sensitive to the parameters γ and δ with wide ranges. The experimental results show that our algorithm converges within 30 iterations and is not sensitive to the dictionary size.

Conclusion

we propose a joint dictionary learning method for unsupervised feature selection, which considers data distribution, data reconstruction and data local structure simultaneously and seamlessly integrates these three key factors into a unified framework. Compared with the existing unsupervised feature selection methods preserve data reconstruction via matrix factorization, we learn an intrinsic space shared by feature space and pseudo label space which can better reconstruct cluster labels and reduce reconstruction error. Moreover, we adopt linear sparse regression term to maintain the cluster structure of samples and $l_{2,1}$-norm is imposed on feature selection matrix to select the most discriminative features.