Pseudo Label Guided Subspace Learning for Multi-view Data

Shudong Hou(✉), Heng Liu, and Xiujun Wang

Anhui University of Technology, Maanshan 340122, China
{shudonghou,xjwang}@ahut.edu.cn, hengliusky@aliyun.com

Abstract. Multi-view spectral clustering methods could utilize the complementary information from different views to increase the robustness of clustering performances. Graph structures are usually revealed as affinity matrices. A pseudo label guided spectral embedding algorithm (PLGS) is proposed in this paper to enhance the consistence between graph matrices and spectral clustering results. Through iteratively estimating the pseudo labels of all samples and similarity matrices, the cluster assignment vector could be calculated with more confidence. Extensive experimental results on several benchmark datasets show promising performance and verify the effectiveness of our method.

Keywords: Spectral clustering · Pseudo label · Multi-view data · Unsupervised learning

1 Introduction

In recent years, multi-view learning methods [1] have received increasing attentions by exploring the consistency and complementary information of different views. It is difficult to fuse the heterogeneous properties from various representations together. If the relationships among different views are not modeled appropriately, the performance may be degraded compared to the best single one. The widely popular methods for multi-view learning are grouped into three main categories: co-training, multiple kernel learning and subspace learning.

As a classical representative paradigm, co-training method [2] utilizes the labeled data to train two classifiers, then categorizes unlabeled data separately. Then the predictive samples with great confidence are added to the labeled data to train the other classifier, and this procedure repeats. Multiple kernel learning [3] is originally proposed to learn a kernel matrix through optimizing a linear combination of kernel matrices. And it can be naturally extended to fuse heterogeneous data sources. Subspace learning-based approaches [4,5] aim to learn a common subspace shared by multiple views. The most typical method should be canonical correlation analysis (CCA) [6] that finds a latent space where the correlations of two projections are mutually maximized.

Different from pattern classification, data clustering aims at grouping vast unlabeled samples into several clusters in such a way that samples in the same

Z. Lin et al. (Eds.): PRCV 2019, LNCS 11859, pp. 75–83, 2019.
https://doi.org/10.1007/978-3-030-31726-3_7

cluster are more similar to each other and k-means clustering is the most well-known method [7]. Spectral clustering (SC) [8] constructs a graph similarity matrix and solves a relaxation of the normalized min-cut problem on this graph. It has gained lots of attention because of its robust performance. For multi-view data, it is assumed that all samples from different views share the same graph structure. Multi-view spectral clustering aims at discovering this intrinsic graph structure information exhibited by various data from several different views. Each view of the same object includes special features may not be described by other views. It is important to utilize the complementary information maximally and enhance the robustness of the final mixed clustering results. Hence co-regularized spectral clustering [9] find the consistent clusterings across the views through co-regularizing the clustering hypotheses. One challenging problems in spectral clustering methods is how affinity matrices are constructed.

Graph matrices are appeared in various methods when the local and global structure information is needed. The graph structure is described by encoding the pairwise similarities among all samples. However it is not reasonable to compute all distances between any two samples if they are far apart from each other. It is assumed that the data are satisfied with the local manifold structure. It is more robust to only compute the nearby several samples to construct the similarity matrix. Through a sparse similarity matrix, the local manifold structure could be better exhibited without lots of unnecessary links. The k-nearest and ε neighbours are widely adopted to compute similarity matrices since their simplicity and effectiveness. However it is hard to select the best k and ε values. Recently the graph matrix is optimized as a sub-problem when optimizing a unified global objective function instead of the original pre-computed similarity matrix [10]. Sparse [11,12] and low rank representation [13,14] could select the local samples by self-expressive abilities. It formulates the graph matrix automatically once the threshold is given. Multi-view low-rank sparse subspace clustering (MLRSSC) [15] learns a joint subspace representation imposing both sparse and low-rank constraint conditions. Kernel trick is utilized when the nonlinear extension is developed [16]. Multi-view learning with adaptive neighbors (MLAN) [17] performs clustering and learns the graph matrix simultaneously. The obtained optimized graph can be partitioned into the intrinsic clusters directly without a back-end processing. In [18], the common consensus information is leveraged instead of the weighted sum of different graphs. It is often happened that some values or views of one object are missing in practice. For traditional multi-view learning, this object is abandoned. By setting the connected weights corresponding to missing instances as 0, incomplete multi-view spectral clustering with adaptive graph learning (IMSC_AGL) [19] could flexibly handle kinds of incomplete cases and prove its effectiveness in incomplete multi-view learning.

Clustering with adaptive neighbors (CAN) [20] tries to acquire a fixed k-rank graph matrix and finish clustering using graph connected components without a back-end k-means method. It is said that the initialization of k-means is a big problem. However after the graph matrices is optimized, the initialization problem can be solved by repeating several times independently. And the restricted

k-rank constraint on graph matrices needs more iterations to balance the weighted factor. It is hard to judge which one costs more resources.

Inspired by learning to learn, a pseudo label guided multi-view spectral clustering method is proposed in this work. The consistence between data and models is maximally remained. If two samples do not share the same cluster assignment, the neighborhood relationship is not reliable. At the first step, we assume all paired samples have the same class label when the distance between them is more close than others. This operation may create some misleading linking edges. We hope to correct them by the following iterations. In each loop, only first k nearest samples with the same cluster assignment are selected as reliable neighbors. Then the similarity matrix is updated according to former spectral clustering results. The true label is approximately estimated after several repeats.

2 Related Work

In this section, we will first review the basic principles of multi-view spectral clustering. Then the CAN is revisited.

2.1 Spectral Clustering

Given a data set $X = \{x_1, \cdots, x_n\} \in \mathbf{R}^{d \times n}$, spectral clustering methods need to construct the graph matrix W first. Then the Laplace matrix is defined as

$$L = D - W \tag{1}$$

Thus the objective function of spectral clustering can be defined as

$$\begin{aligned} &\min_{F} Tr(F^T L F) \\ &s.t.\ \ F^T F = I \end{aligned} \tag{2}$$

The optimal F is solved by eigen-decomposition. Then the final clustering is performed by using the formulated F as the low dimensional embedding of the raw data X.

2.2 Clustering with Adaptive Neighbors

Spectral clustering actually is a graph theory-based method. Thus the clustering task can be viewed as a graph cut problem. The ideal graph has exact c connected components for c-class clustering. Usually this strong constraint is difficult to satisfy due to the noisy and complex data distribution. For the sake of achieving the ideal graph cut, a reasonable low-rank constraint is added when constructing the similarity matrix S:

$$\begin{aligned} &\min_{s_i \in \mathbf{R}^{n \times 1}} \sum_{i,j}^{n} \|x_i - x_j\|_2^2 s_{ij} + \alpha \|S\|_F^2 \\ &s.t. \quad \forall i, s_i^T \mathbf{1} = 1, 0 \le s_{ij} \le 1, rank(L_S) = n - c \end{aligned} \tag{3}$$

where s_i is a column vector with j-th element as s_{ij} and L_S is the Laplace matrix of S [17]. In each iterative loop, the value of α is adjusted to automatically select the local samples.

3 Methodology

The intuitive objective of multi-view clustering is mining the local common structure information. Since the unavoidable noise existed in each view, it requires more focus to balance the weights when fusing all graph matrices together. Different from CAN, our proposed PLGS iteratively estimates the pseudo label of all samples.

3.1 Model

For multi-view data, let $\mathcal{X} = \{\mathbf{X}^1, \mathbf{X}^2, \cdots, \mathbf{X}^V\}$ denotes the V-view feature sets where $\mathbf{X}^v \in \mathbf{R}^{n \times d_v}$ means the v-th feature set. In each feature set, the nearest k neighbors are selected to construct the similarity matrix S_i. Then the global similarity matrix S_g is calculated by integrating all S_i together. The classical spectral clustering method is performed based on S_g. Lastly, the nearest neighbors are corrected according to previous clustering results. Only the neighbors that are in the same cluster are remained in the similarity matrix, otherwise this pseudo label is not reliable and deleted. After the similarity matrix is updated, a new clustering result is generated again.

The integrated objective function is defined as:

$$\begin{aligned} \min_{S,Q} \quad & Tr(Q^T L_g Q) \\ s.t. \quad \forall i \neq j, \quad & cluster(q_i) = cluster(q_j) \\ \&\& S_{ij} \geq 0 \&\& & (x_j \in K(x_i) || x_i \in K(x_j)) \end{aligned} \tag{4}$$

where q_i is the i-th column of Q, L_g is the Laplace matrix of S_g, $K(x_i)$ represents the nearest neighbors for sample x_i and *cluster* is the assignment vector calculated by SC methods. The main idea is finding a global optimized similarity matrix that is consistent with the spectral clustering result.

3.2 Optimization

To solve this challenging problem, an alternative iterative solution is adopted. The initial similarity matrix is constructed as follows:

$$S_{ij}^0 = \begin{cases} exp(-dist(x_i, x_j)^2/(2\sigma^2)), & x_j \in K(x_i) || x_i \in K(x_j) \\ 0, & otherwise \end{cases} \quad (i, j = 1, \cdots, n) \tag{5}$$

where $dist(x_i, x_j)$ means the distance between sample x_i and x_j. It is measured by the weighted average of all views. For simplicity the weights of all views are set to be the same $1/v$. Then the assignment vector *cluster* is acquired by spectral clustering (2).

Instead of the strictly matrix rank constraints, the k-means method is utilized to get the cluster assignment vector. Since its randomly initialization, the clustering results are different from each other for individual replicates. So the k-means are repeated for t times and the cluster assignment vector is computed as follows:

$$\begin{cases} cluster_f(x_i) = cluster_f(x_j), & if\ \#(cluster_k(x_i) == cluster_k(x_j)) \geq \theta \\ cluster_f(x_i) \neq cluster_f(x_j), & if\ \#(cluster_k(x_i) == cluster_k(x_j)) < \theta \end{cases}$$
$$(i, j = 1, \cdots, n,\ k = 1, 2, \cdots, t) \quad (6)$$

where the function # means "the number of". It records how many times these two samples are in the same cluster. If this value is larger than the predefined threshold θ, we let them share the same cluster in the final assignment vector $cluster_f$. According to the new generated *cluster* assignment vector, the similarity matrix is corrected by deleting the inconsistent values.

$$S_{ij}^{(t+1)} = \begin{cases} S_{ij}^t, & if\ cluster(x_i) = cluster(x_j) \\ 0, & otherwise \end{cases} \quad (i, j = 1, \cdots, n) \qquad (7)$$

Based on the above analysis, the overall algorithm for solving (4) is summarized in Algorithm 1.

Algorithm 1. PLGS Algorithm

```
Require:
    Multi-view data $\mathcal{X}$, the neighborhood size $k$ and the repeat number $t$.
Ensure:
    The graph $S_g$ and cluster assignment vector $label$.
Construct the initial similarity matrix $S$ according to (5).
Formulate the global similarity matrix $S_g$ by integrating all similarity matrices $\{S\}_{v=1}^V$ from all $V$ views.
while not convergence
    Apply spectral clustering on $S_g$ and get $F$ by (2).
    Perform $t$ times $k$-means clustering and get assignment vector $label$ using (6).
    Update $S_g$ by (7).
end while
```

4 Experiments

In order to evaluate the effectiveness of the proposed method, extensive experiments are performed on several real-world multi-view datasets.

4.1 Datasets and Settings

The experimental results are reported on four real-world datasets: UCI Digits[1], Reuters, 3-sources[2] and Prokayotic. The detailed information of these datasets are listed in Table 1.

In the experiments, two evaluation metrics are used to verify the effectiveness of the proposed method. They are the accuracy and normalized mutual information (NMI). The clustering accuracy is defined as

$$accuracy = \frac{\#correct\ decisions}{\#total\ decisions} \tag{8}$$

And the NMI is defined as

$$NMI(\mu,\nu) = \frac{2\sum_{i=1}^{c}\sum_{j=1}^{\hat{c}}\frac{n_{ij}}{n} log\frac{n_{ij}n}{\sum_{i=1}^{c}n_i\sum_{j=1}^{\hat{c}}n_j}}{-\sum_{i=1}^{c}\frac{n_i}{n}log\frac{n_i}{n}-\sum_{j=1}^{\hat{c}}\frac{n_j}{n}log\frac{n_j}{n}} \tag{9}$$

where n_{ij} denotes the number of data in cluster i and class j, n_i and n_j denotes the data number belonging to the ground-truth (μ_i) and clustering result ν_j respectively.

Table 1. Statistics of the multi-view datasets

Dataset	Samples	Views	Clusters
UCI Digits	2000	6	10
Reuters	600	5	6
3-sources	169	3	6
Prokayotic	551	3	4

4.2 Experimental Results

Five methods, including spectral clustering, CAN, MLAN [17][3], MLRSSC and its kernel extension [15][4], are used for comparison. All parameters of these algorithms are set to values based on the respective source codes provided by their authors. The experimental results are shown in Table 2. For SC and CAN, the best single view result is reported.

Compare with spectral clustering, the performance of CAN is much better. This shows that the adaptive neighbors are more reliable than nearest neighbors. It is hard to adjust the parameter values of MLRSSC and its results are

[1] http://archive.ics.uci.edu/ml/datasets/Multiple+Features.

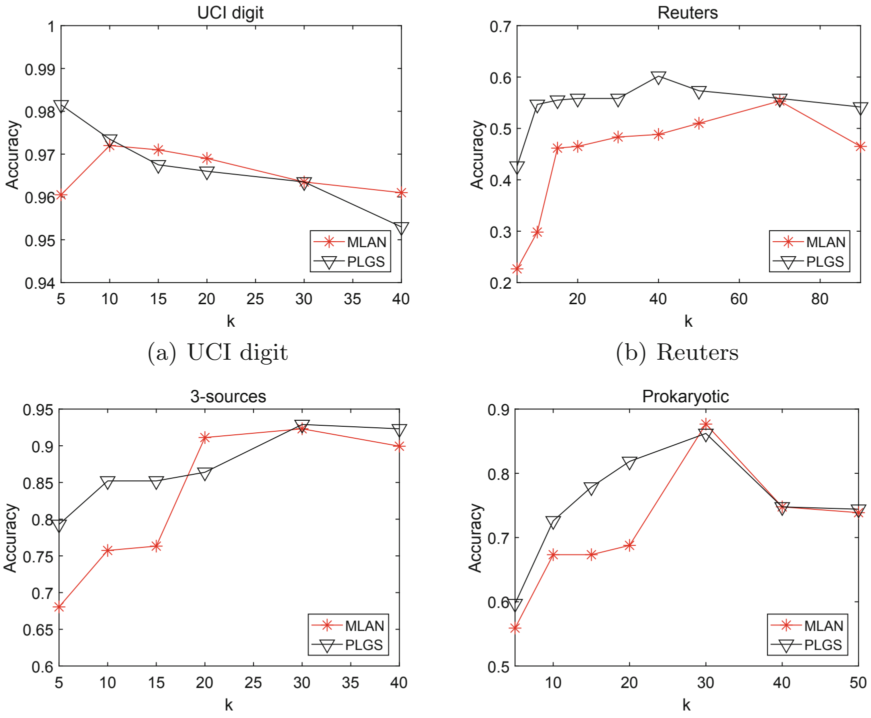
[2] http://mlg.ucd.ie/datasets/3sources.html.

[3] http://www.escience.cn/people/fpnie/papers.html.

[4] https://github.com/mbrbic/MultiViewLRSSC.

Table 2. Performance of different methods on four multi-view datasets.

Dataset	UCI digit		Reuters		3-sources		Prokaryotic	
	Accuray	NMI	Accuray	NMI	Accuray	NMI	Accuray	NMI
SC-best	85.45	88.44	56.00	42.54	88.76	75.06	60.25	9.26
CAN-best	86.65	89.61	37.33	32.82	70.41	61.69	75.32	47.96
MLRSSC	88.22	87.21	46.27	33.88	69.08	59.59	65.93	32.62
KMLRSSC	81.70	77.08	45.17	32.67	60.65	52.13	65.05	40.80
MLAN	97.20	93.60	55.33	41.35	92.31	81.87	**87.66**	58.21
PLGS	**98.15**	**95.72**	**60.17**	**43.86**	**92.90**	**83.84**	86.21	**59.92**



(a) UCI digit

(b) Reuters

(c) 3-sources

(d) Prokaryotic

Fig. 1. The sensitivity analysis of k

not satisfying. The rank-constraint is remained during the whole processing in MLAN while MLRSSC aims at optimizing a trace-norm minimization problem actually. Our proposed PLGS utilizes the k-nearest neighbors and pseudo labels of all samples to enhance the sparse and discriminative abilities of feature representations. Its promising clustering results are presented to demonstrate the effectiveness of PLGS.

4.3 Parameter Sensitivity

A predefined k value needs to be determined for MLAN and PLGS. To further verify the effectiveness of our proposed method, the sensitivity of k is analyzed in Fig. 1.

Although the neighbors in MLAN are selected adaptively, its clustering results are more sensitive compared with our PLGS. When the k value is large, two methods almost have the same performances. If the k value is small, PLGS usually performs better than MLAN.

5 Conclusion

In this paper, a pseudo label-guided clustering method is proposed to solve the multi-view clustering problem. Instead of solving a rank constraint optimization, we utilize a very simple idea to increase the sparse and discriminative abilities of feature representations. For PLGS, the global similarity matrix is calculated by average with the same weights. If the weights are carefully designed and iterative estimated, a better performance will be reached.

Acknowledgements. This work is partly supported in part by Natural Science Foundation of Anhui Province under Grant 1808085QF210 and Grant 1608085MF129. And in part by the Major and Key Project of Natural Science of Anhui Provincial Department of Education under Grant KJ2015ZD09 and Grant KJ2018A0043.

References

1. Xu, C., Tao, D., Xu, C.: A survey on multi-view learning. arXiv preprint arXiv:1304.5634 (2013)
2. Blum, A., Mitchell, T.: Combining labeled and unlabeled data with co-training. In: Proceedings of the Eleventh Annual Conference on Computational Learning Theory, pp. 92–100 (1998)
3. Lanckriet, G.R.G., Cristianini, N., Bartlett, P., El Ghaoui, L., Jordan, M.I.: Learning the kernel matrix with semidefinite programming. J. Mach. Learn. Res. **5**, 27–72 (2004)
4. Xu, C., Tao, D.: Multi-view intact space learning. IEEE Trans. Pattern Anal. Mach. Intell. **37**(12), 2531–2544 (2015)
5. Kan, M., Shan, S., Zhang, H., Lao, S., Chen, X.: Multi-view discriminant analysis. IEEE Trans. Pattern Anal. Mach. Intell. **38**(1), 188–194 (2016)
6. Hotelling, H.: Relations between two sets of variates. Biometrika **28**(3/4), 321–377 (1936)
7. MachQueen J.: Some methods for classification and analysis of multi-variate observations. In: Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability, pp. 281–297 (1965)
8. von Luxburg, U.: A tutorial on spectral clustering. Stat. Comput. **17**(4), 395–416 (2007)
9. Kumar A., Rai P., Daume H.: Co-regularized multi-view spectral clustering. In: Proceedings of the 24th International Conference on Neural Information Processing Systems (NIPS 2011), pp. 1413–1421 (2011)

10. Pang, Y., Zhou, B., Nie, F.: Simultaneously learning neighborship and projection matrix for supervised dimensionality reduction. IEEE Trans. Neural Netw. Learn. Syst. (in press)
11. Elhamifar, E., Vidal, R.: Sparse subspace clustering: algorithm, theroy, and applications. IEEE Trans. Pattern Anal. Mach. Intell. **35**(11), 2765–2781 (2013)
12. Lu, C., Yan, S., Lin, Z.: Convex sparse spectral clustering: single-view to multi-view. IEEE Trans. Image Process. **25**(6), 2833–2843 (2016)
13. Liu, G., Lin, Z., Yan, S., Sun, J., Yu, Y., Ma, Y.: Robust recovery of subspace structures by low-rank representation. IEEE Trans. Pattern Anal. Mach. Intell. **35**(1), 171–184 (2013)
14. Lu, C., Feng, J., Yan, S., Lin, Z.: A unified alternating direction method of multipliers by majorization minimization. IEEE Trans. Pattern Anal. Mach. Intell. **40**(3), 527–541 (2018)
15. Brbić, M., Kopriva, I.: Multi-view low-rank sparse subspace clustering. Pattern Recogn. **73**, 247–256 (2018)
16. Houthuys, L., Langone, R., Suykens, J.A.K.: Multi-view kernel spectral clustering. Inf. Fusion **44**, 46–56 (2018)
17. Nie, Y., Cai, G., Li, J., Li, X.: Auto-wighted multi-view learning for image clustering and semi-supervised classification. IEEE Trans. Image Process. **27**(3), 1501–1511 (2018)
18. Zhan, K., Nie, N., Wang, J., Yang, Y.: Multiview consensus graph clustering. IEEE Trans. Image Process. **28**(3), 1261–1270 (2019)
19. Wen, J., Xu, Y., Liu, H.: Incomplete multiview spectral clustering with adaptive graph learning. IEEE Trans. Cybern. (in press)
20. Wang, Q., Qin, Z., Nie, F., Li, X.: Spectral embedded adaptive neighbors clustering. IEEE Trans. Neural Netw. Learn. Syst. **40**(3), 1265–1271 (2019)