GUIDED CO-TRAINING FOR MULTI-VIEW SPECTRAL CLUSTERING

Chung-Kuei Lee *Tyng-Luh Liu*

Institute of Information Science, Academia Sinica, Taiwan

ABSTRACT

We address the problem of how to design a more effective co-training scheme to tackle the multi-view spectral clustering. The conventional co-training procedure treats information from all views equally and often converges to a compromised consensus view that does not fully utilize the multi-view information. We instead propose to learn an augmented view and construct its corresponding affinity matrix from a spectral decomposition of an information-rich matrix formed by the eigenvectors of the Laplacian matrices from all views. As the augmented view is expected to be more favorable for carrying out spectral clustering, we design a new pairwise co-training procedure to guide the improvements of the given multiple views separately and iteratively. Our experimental results on three popular benchmark datasets support that the convergent augmented view by the guided co-training process is useful to multi-view spectral clustering, and can yield state-of-the-art performance.

Index Terms**—** Spectral clustering, graph Laplacian

1. INTRODUCTION

Clustering is one of the most important unsupervised learning problems, which partitions a set of objects into disjoint groups. There are two main factors that dictate the outcome of this process, namely, data representation and similarity measure. In *spectral clustering*, *e.g.*, [1], an affinity matrix is used to integrate the two aspects of information. Spectral analysis is then applied to further encode the data in a simpler but more structured form so that k-means clustering can be used to accomplish the task. Still, real-world applications involving clustering are often too complicated and require data to be characterized by more than one type of feature. Take, for example, that in natural language processing we may have parallel texts in two or more languages, and in computer vision we could use several descriptors to better describe each image. For such applications, it is necessary to extend spectral clustering to satisfactorily take account of multi-view affinity matrices. Essentially, how to more effectively accomplish such information fusion is the crux of our method.

We develop a new co-training learning framework to iteratively carry out multi-view spectral clustering. Previous attempts relevant to using co-training or co-regularization for spectral clustering can be found in [2, 3, 4]. These approaches rely mostly on maximizing the mutual agreement of different views. For example, the spectral clustering algorithm in [2] uses co-training to iteratively modify the affinity matrix of each view by projecting it to the union of the subspaces spanned by, say, the top k eigenvectors of the other views. However, since we generally do not know in advance which views are more appropriate in clustering the underlying data, the co-training procedure may converge to a unified but compromised one, and consequently fails to significantly improve the clustering performance. To avoid such a dilemma, we instead propose to first generate an *augmented* view by constructing an additional affinity matrix that is obtained from a spectral decomposition of a rectangular matrix formed by stacking in columns the eigenvectors of the Laplacian matrices from all views. As the new affinity matrix is generated by considering all the information from the various feature descriptors, it is reasonable to assume that the derived augmented view is more efficacious for clustering the underlying data than any of the original views. Based on this assumption, we introduce a new pairwise co-training procedure. At each iteration, rather than consider all the views simultaneously, we use the augmented view to separately "guide" the improvement of each view. Once all the original views have been adjusted, we can then update the augmented view and repeat the iterative process. This way the proposed multi-view spectral clustering is more likely to overcome the inefficiency of conventional co-training caused by modifying an affinity matrix according to information from inferior views, and yield better clustering performance.

2. RELATED WORK

As the motivation behind our method is to more adequately utilize the multi-view information to boost the performance of spectral clustering, the literature review below emphasizes relevant techniques for both issues.

Most of the multi-view learning approaches can be classified into three categories: (1) *co-training*, (2) *multiple kernel learning* and (3) *subspace learning*, [5]. The co-training paradigm is first introduced by Blum and Mitchell [6]. Its main idea is to train separate learners and enforces the agreement of the learners. Kumar and Daumé [2] develop a multi-view spectral clustering in the co-training fashion. The co-

978-1-4673-9961-6/16/$31.00 ©2016 IEEE

regularization framework, proposed by Sindhwani *et al.* [7], can be regarded as a regularized version of the co-training algorithm, and subsequently motivates the formulation of the co-regularized multi-view spectral clustering by Kumar *et al.* [3]. Two specific schemes are described in [3]: The first is to pairwise regularize the features, and the second is to find a centroid which regularizes each view toward a consensus answer. Other algorithms in this category include [8] and [4], where the former learns a shared graph Laplacian from various feature types and the latter determines the weights among different regularization costs automatically.

Multiple kernel learning (MKL) works with a set of predefined kernels that naturally correspond to different views. The framework is formulated to learn an optimal linear or nonlinear combination of base kernels, and thus accomplish the view fusion. It is convenient to leverage with MKL to design multi-view clustering *e.g.*, [9, 10]. The algorithms of this category typically compute weighted kernels and combine accordingly to accomplish clustering. Relevant approaches by including a probabilistic model for each feature type or learning a mixture of them can be found in [11, 12].

Subspace learning is popular in designing algorithms for multi-view clustering. It assumes that the input views are generated from a latent subspace. Therefore, the learning task aims to obtain an appropriate subspace from the input views and then produces a shared representation for all views. This strategy has been used to generate many different types of representation, such as *canonical correlation analysis* [13] and *universal similarity matrix* [14], for clustering.

3. OUR METHOD

There are two key components in the proposed multi-view spectral clustering algorithm. First, we introduce a new technique to generate an augmented view and its affinity matrix. Second, we explore favorable properties of the new view to establish a guided co-training procedure to improve the given views, and consequently converge to a more effective affinity matrix of the augmented view for spectral clustering.

3.1. Augmented view

Given a collection $\mathcal{X}$ of n data points, assume that there are m different feature descriptors to describe the data. We can express the multi-view representation by $\mathcal{X} = \{\mathbf{x}_1^v, \ldots, \mathbf{x}_n^v\}_{v=1}^m$ where $\mathbf{x}_i^v$ denotes the feature vector of data point i in the vth view. Then, the affinity matrix A^v associated with view v $(1 \leq v \leq m)$ can be defined by

$$A_{ij}^v = \begin{cases} \exp(-\|\mathbf{x}_i^v - \mathbf{x}_j^v\|^2/2\sigma^2), & \text{if } i \neq j, \\ 0, & \text{otherwise}, \end{cases} \tag{1}$$

where $1 \leq i, j \leq n$ and we have used Gaussian similarity function to model the local neighborhood relations. In our implementation, we set the bandwidth σ to the median value of the set $\{\|\mathbf{x}_i^v - \mathbf{x}_j^v\| \, : \, 1 \leq i, j \leq n, i \neq j\}$. Also note that we adopt the spectral clustering formulation described in [1], and express the *normalized* graph Laplacian of view v as

$$L^v = (D^v)^{-\frac{1}{2}} A^v (D^v)^{-\frac{1}{2}} \tag{2}$$

where D^v is the corresponding degree matrix. Though, with (1), all the multi-view information of $\mathcal{X}$ can be encoded into the m affinity matrices, we indeed have no clue about their usefulness in carrying out spectral clustering. Such uncertainty could cause a co-training procedure like [2] to compromise on treating all the views equally and converge to a less satisfactory consensus view for clustering. We instead propose to iteratively generate an augmented view and its affinity matrix A^* from a spectral approximation of all the encoded information in $\{A^v\}$, and use the resulting A^* to guide the improvement of each A^v. Before we describe how we derive A^*, let us first state the Eckart-Young-Mirsky theorem for matrix approximation, which will be used in the derivation.

Theorem 1. *(Eckart-Young-Mirsky) For a matrix $M \in \mathbb{R}^{n \times q}$ with $rank(M) = r$ and singular value decomposition $M = U\Sigma V^T$, if $k < r$, we have*

$$\arg \min_{\substack{\tilde{M} \in \mathbb{R}^{n \times q} \\ rank(\tilde{M}) = k}} \|M - \tilde{M}\|_F = U\tilde{\Sigma}V^T, \tag{3}$$

where $\|\cdot\|_F$ denotes the Frobenius norm and $\tilde{\Sigma}$ is constructed from Σ by keeping the k largest singular values and replacing the rest by 0.

Assume now that we are to divide $\mathcal{X}$ into k clusters with multi-view spectral clustering. To derive A^* of the augmented view, we first acquire the corresponding Laplacian matrices $L^1, \ldots, L^m$ as in (2). Next, we compute the k "largest" eigenvectors of each L^v and stack these mk eigenvectors in columns to form a matrix $M \in \mathbb{R}^{n \times mk}$ as follow:

$$M = \left[\mathbf{w}_1^1, \mathbf{w}_2^1, \ldots, \mathbf{w}_k^1, \mathbf{w}_1^2, \ldots, \mathbf{w}_1^m, \ldots, \mathbf{w}_k^m\right], \tag{4}$$

where $\mathbf{w}_i^v$ is the ith largest eigenvector of L^v. From the Eckart-Young-Mirsky theorem, we know that M can be used to approximately reconstruct $A^1, \ldots, A^m$. Furthermore, M includes most of the information (under the rank-k constraint) from the m different views. Thus, it makes sense to construct A^* from M by retaining as much information about M as possible, and consequently exploring all the m views. To this end, we modify M by normalizing each row to unit norm and then compute the SVD decomposition, $M = U\Sigma V^T$. The affinity matrix A^* of the augmented view is given by

$$A^* = (U\Sigma^{\frac{1}{2}})(U\Sigma^{\frac{1}{2}})^T \in \mathbb{R}^{n \times n}. \tag{5}$$

It is clear that A^* of the augmented view is symmetric and inherits most of the (column-space) structure from M. Later in our experiments, we will show that the improvements in clustering performance owing to A^* are significantly better than those by any of the original views.

3.2. Guided co-training

Algorithm 1: *Multi-view Spectral Clustering via Guided Co-training*

```
Input  : $m$-view affinity matrices: $A^1, \dots, A^m$.
         Number of clusters: $k$.
         Number of iterations: $iter$.
Output: Clustering result.
$t \leftarrow 0$;
while $t < iter$ do
    Compute $A^*$ using (5);
    for $v \leftarrow 1$ to $m$ do
        $A^v \leftarrow A^* \odot A^v$;
    $t \leftarrow t + 1$;
Perform spectral clustering with $A^*$;
```

Once we know how to generate the augmented view and its affinity matrix A^*, we are ready to establish a guided co-training procedure where its iterative improvements are more stable and appreciable. (See Figure 1.)

The guided m-view co-training with $A^1, \dots, A^m$ proceeds as follows. We begin by constructing the affinity matrix A^* of the augmented view as in (5). Then, we replace A^v of view $v \in \{1, \dots, m\}$ with $A^* \odot A^v$, where $\odot$ denotes the *Hadamard product*. Having processed all the m affinity matrices, we need to update A^* and repeat the iterative procedure until changes in A^* are not significant (or reaching a pre-specified number of iterations). Detailed steps of the guided co-training for multi-view spectral clustering are listed in Algorithm 1.

4. EXPERIMENTAL RESULTS

We demonstrate the usefulness of the proposed method for data clustering, and compare its effectiveness with several popular multi-view spectral clustering techniques.

4.1. Evaluation protocols

Datasets. Three benchmark datasets for clustering are used in the evaluation.

Oxford Flower 17. The dataset is introduced by Nilsback and Zisserman [15]. It comprises 17 categories of flowers with 80 images for each class. Each image is characterized by three different features: color, shape and texture. To construct the affinity matrix, we use the Gaussian kernel with χ^2-distance as the norm.

UCI Multiple Features. The collection includes features of handwritten numerals extracted from a collection of Dutch utility maps [16]. There are 10 classes of images with 200 instances each. The images are described by six different features: (1) Fourier coefficients of the character shapes; (2) profile correlations; (3) Karhunen-Love coefficients; (4) pixel averages in 2 x 3 windows; (5) Zernike moments; and (6) morphological features. The affinity matrix is established by Gaussian kernel with Euclidean norm.

UC Merced Land Use. This dataset is extracted from large optical images (RGB color space) of the US Geological Survey, taken over various regions of the United States [17]. It contains 2100 images belonging to 21 categories with 100 images each. The feature we used includes GIST [18], pHOG [19], Color Histogram [20] and LLCs [21] with 1×1, 2×2 and 4×4 partitions. The similarity between images is computed by Gaussian kernel with Euclidean norm.

Baseline algorithms. We compare our method with six competing techniques for multi-view spectral clustering. We briefly describe them and their respective abbreviation below.

Single$_{best}$: It carries out spectral clustering [1] on the most informative view.

Pairwise co-Regularized Spectral Clustering (PRSC): Proposed in [3], this method computes the eigenvector matrix U^v for each view and then encourages pairwise similarity across all the views under the new representation (U^v). After all U^v's are learned, perform k-means clustering on the embedding concatenation.

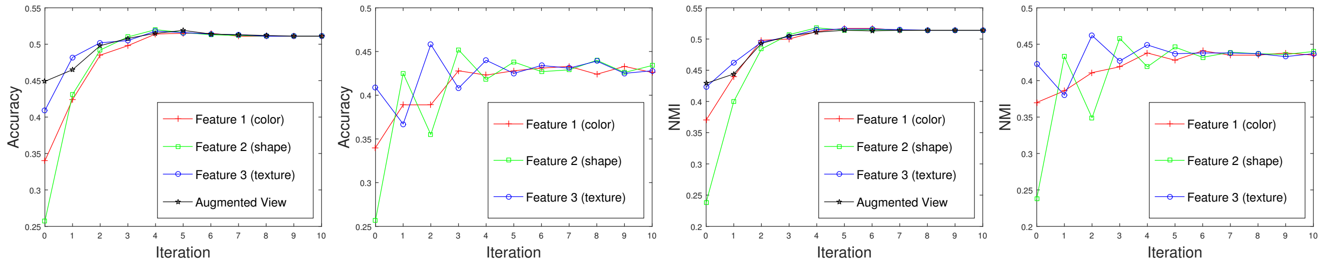
Centroid based co-Regularization Spectral Clustering (CRSC): The algorithm is modified from PRSC. It is also proposed in [3]. In contrast to PRSC which has $\binom{m}{2}$ pairwise regularization terms (m is the number of views), CRSC has m regularization terms that regularize each view-specific set of eigenvectors U^v's toward a common centroid U^*. Then, apply k-means clustering to U^* for the final result.

Affinity Aggregation Spectral Clustering (AASC): For affinity matrices generated from different views, this method [22] derives an optimal weight for each view and then aggregates a single affinity matrix accordingly. The clustering result is achieved by performing spectral clustering on the aggregated affinity matrix.

Co-training Multi-view Spectral Clustering (CMSC): To the best of our knowledge, the technique [2] is the first multi-view spectral clustering algorithm designed in the co-training scheme. This algorithm uses the spectral embedding from one view to constrain the similarity graphs used for the other views. We run this algorithm 30 iterations and report the score of the best view.

Multi-feature Spectral Clustering with Minimax Optimization (MSCMO): Similar to CRSC, the method [4] tries to find a common embedding U^* for all views and then performs k-means clustering on it. Moreover, this algorithm reaches a more harmonic consensus by weighting the cost terms differently during the minimax optimization.

Evaluation metrics. We evaluate clustering results by two standard measures: *clustering accuracy* (ACC) and *normalized mutual information* (NMI). For both measures, a higher



(a) Ours (b) CMSC [2] (c) Ours (d) CMSC [2]

Fig. 1: Convergence study on Oxford Flower 17. The plots show the accuracy score (ACC) and NMI score versus number of iterations. Unlike in the cases of CMSC [2], the guided co-training by our method yields more stable improvements iteratively.

value indicates a better clustering result. Like in [4], the reported value is obtained by averaging the results of 10 tests.

4.2. Comparisons and results

We compare our method with competing clustering techniques and report the results in Table 1. For the Oxford Flower 17 dataset and UCI Multiple Features dataset, our method outperforms all the other algorithms. It is interesting that the gap between our approach and the compared algorithms on UCI Multiple Features dataset is significantly larger than the gap on Oxford Flower 17 dataset. Moreover, for the UC Merced Land Use dataset, while our method achieves the best score on the NMI metric, it falls slightly behind MSCMO on the accuracy metric. Because our method relies on the affinity matrix constructed from the spectral embedding of all views, we speculate that the performance is more "sensitive" to how well the spectral embedding of each view complements others. Developing a method to generate a *better* affinity matrix which contains information from all views would be a promising direction for future research.

Figure 1 displays the variation in NMI and accuracy metric on the Oxford Flower 17 dataset as the number of iterations increases. The results in Figures 1(a) and 1(c) are produced by our method, while the other two are by CMSC [2]. For our algorithm, we present the scores of three features (color, shape and texture) and the augmented view. For the comparison algorithm, we plot the scores of the three features. From the line chart, we can see that the variations in improvement from [2] are more substantial than ours. In fact, the authors of [2] also observe the perturbation phenomenon and consider seeking for new ideas to dig more information from it. For our method, the NMI and accuracy scores are more stable as they converge after roughly nine iterations. However, both measures reach maximum around the fourth and fifth iteration rather than the final convergence value. As suggested in [2], it would be helpful to establish additional heuristic clustering performance measures, *e.g.*, cluster compactness measure, to decide the stopping criterion and further improve a co-training algorithm.

Table 1: Comparisons among Single$_{best}$, PRSC [3], CRSC [3], AASC [22], CMSC [2], MSCMO [4], `Concat`, and our method. `Concat` is a naive scheme that performs k-means clustering over the rows of the matrix M defined in (4).

Method	Oxford Flower 17		UCI Multiple Features		UC Merced Land Use	
	ACC	NMI	ACC	NMI	ACC	NMI
Single$_{best}$	0.404	0.425	0.710	0.660	0.387	0.449
PRSC	0.419	0.435	0.769	0.728	0.368	0.447
CRSC	0.449	0.461	0.770	0.713	0.395	0.468
AASC	0.410	0.422	0.683	0.649	0.226	0.291
CMSC	0.428	0.439	0.779	0.745	0.397	0.458
MSCMO	0.493	0.484	0.800	0.755	**0.404**	0.483
Concat	0.497	0.506	0.938	0.877	0.392	0.467
Ours	**0.512**	**0.514**	**0.949**	**0.896**	0.401	**0.491**

5. CONCLUSIONS

We have introduced a novel multi-view spectral clustering algorithm that explores a guided co-training procedure to ensure iteratively stable improvements on the clustering result. We test our method on three benchmark datasets, including UCI Multiple Features, Oxford Flower 17 and UC Merced Land Use, and achieve state-of-the-art performance on the NMI measure for all the three datasets.

Compared with the competing multi-view spectral clustering approaches, our method can boost the performance significantly on the UCI Multiple Features dataset, moderately on the Oxford Flower 17 dataset, while the improvement on the UC Merced Land Use dataset is comparable to the state-of-the-art. It would be interesting to study the underlying reasons for this phenomenon. In the future, we plan to exploit the hidden properties which decide how well different views *complement* each other. Following this direction, we expect to uncover more interesting and subtle properties of multi-view spectral clustering and advance the research of a wide spectrum of clustering problems.

Acknowledgement: We thank the reviewers for their valuable comments. This work was supported in part by MOST grant 102-2221-E-001-021-MY3.

6. REFERENCES

[1] Andrew Y Ng, Michael I Jordan, Yair Weiss, et al., "On spectral clustering: Analysis and an algorithm," *Advances in neural information processing systems*, vol. 2, pp. 849–856, 2002.

[2] Abhishek Kumar and Hal Daumé, "A co-training approach for multi-view spectral clustering," in *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, 2011, pp. 393–400.

[3] Abhishek Kumar, Piyush Rai, and Hal Daume, "Co-regularized multi-view spectral clustering," in *Advances in Neural Information Processing Systems*, 2011, pp. 1413–1421.

[4] Hongxing Wang, Chaoqun Weng, and Junsong Yuan, "Multi-feature spectral clustering with minimax optimization," in *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*. IEEE, 2014, pp. 4106–4113.

[5] Chang Xu, Dacheng Tao, and Chao Xu, "A survey on multi-view learning," *arXiv preprint arXiv:1304.5634*, 2013.

[6] Avrim Blum and Tom Mitchell, "Combining labeled and unlabeled data with co-training," in *Proceedings of the eleventh annual conference on Computational learning theory*. ACM, 1998, pp. 92–100.

[7] Vikas Sindhwani, Partha Niyogi, and Mikhail Belkin, "A co-regularization approach to semi-supervised learning with multiple views," in *Proceedings of ICML workshop on learning with multiple views*. Citeseer, 2005, pp. 74–79.

[8] Xiao Cai, Feiping Nie, Heng Huang, and Farhad Kamangar, "Heterogeneous image feature integration via multi-modal spectral clustering," in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*. IEEE, 2011, pp. 1977–1984.

[9] Yen-Yu Lin, Tyng-Luh Liu, and Chiou-Shann Fuh, "Multiple kernel learning for dimensionality reduction," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 33, no. 6, pp. 1147–1160, 2011.

[10] Dongyan Guo, Jian Zhang, Xinwang Liu, Ying Cui, and Chunxia Zhao, "Multiple kernel learning based multi-view spectral clustering," in *2014 22nd International Conference on Pattern Recognition (ICPR)*. IEEE, 2014, pp. 3774–3779.

[11] Steffen Bickel and Tobias Scheffer, "Multi-view clustering.," in *ICDM*, 2004, vol. 4, pp. 19–26.

[12] Grigorios F Tzortzis and Aristidis C Likas, "Multiple view clustering using a weighted combination of exemplar-based mixture models," *Neural Networks, IEEE Transactions on*, vol. 21, no. 12, pp. 1925–1938, 2010.

[13] Kamalika Chaudhuri, Sham M Kakade, Karen Livescu, and Karthik Sridharan, "Multi-view clustering via canonical correlation analysis," in *Proceedings of the 26th annual international conference on machine learning*. ACM, 2009, pp. 129–136.

[14] Xiaochun Cao, Changqing Zhang, Huazhu Fu, Si Liu, and Hua Zhang, "Diversity-induced multi-view subspace clustering," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 586–594.

[15] Maria-Elena Nilsback and Andrew Zisserman, "A visual vocabulary for flower classification," in *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*. IEEE, 2006, vol. 2, pp. 1447–1454.

[16] M. Lichman, "UCI machine learning repository," 2013.

[17] Yi Yang and Shawn Newsam, "Spatial pyramid co-occurrence for image classification," in *Computer Vision (ICCV), 2011 IEEE International Conference on*. IEEE, 2011, pp. 1465–1472.

[18] Aude Oliva and Antonio Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," *International journal of computer vision*, vol. 42, no. 3, pp. 145–175, 2001.

[19] Anna Bosch, Andrew Zisserman, and Xavier Munoz, "Representing shape with a spatial pyramid kernel," in *Proceedings of the 6th ACM international conference on Image and video retrieval*. ACM, 2007, pp. 401–408.

[20] Frank Palermo, James Hays, and Alexei A Efros, "Dating historical color images," in *Computer Vision–ECCV 2012*, pp. 499–512. Springer, 2012.

[21] Jinjun Wang, Jianchao Yang, Kai Yu, Fengjun Lv, Thomas Huang, and Yihong Gong, "Locality-constrained linear coding for image classification," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*. IEEE, 2010, pp. 3360–3367.

[22] Hsin-Chien Huang, Yung-Yu Chuang, and Chu-Song Chen, "Affinity aggregation for spectral clustering," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2012, pp. 773–780.