MULTI-VIEW NETWORK-BASED SOCIAL-TAGGED LANDMARK IMAGE CLUSTERING

So Yeon Kim Kyung-Ah Sohn

Department of Computer Engineering, Ajou University, S. Korea

ABSTRACT

The multiple types of social media data have abundant information, but learning multi-modal social data is challenging due to data heterogeneity and noise in user-generated data. To address this problem, we propose a multi-view network-based clustering approach that is robust to noise and fully reflects the underlying structure of the comprehensive network. To demonstrate the proposed approach, we experimented with clustering challenging tagged images of landmarks. The results show that the proposed method outperforms other previously reported multi-view clustering algorithms and better utilizes the advantages of the network for each view. Furthermore, the tagged-image network constructed by the proposed method and the clustering results are extensively analyzed.

Index Terms— multi-view learning, network clustering, tagged-image clustering

1. INTRODUCTION

The emergence of multiple types of data have made it easier to collect user-generated data from many different sources, especially from social media. The data is not only provided as a single view, but also in the form of multiple views. For example, when using Flickr or Instagram, the user uploads an image with hashtags, descriptions, and geo-location. Though the massive data of individual views may have sufficient information for learning the structure of the data, multiple representations contribute to a better understanding of the data that results in improved performance. For instance, the features from natural images in social media can be noisy and have insufficient information. The user normally provides more information about the image in the form of social tags and descriptions. In this respect, multi-view approaches have attracted much attention in many machine learning studies.

As images are included in a social network, it is more informative to represent them as a network. For a social image network, it is useful to seek groups that have similar features. This can be accomplished with a state-of-the-art network clustering algorithm such as spectral clustering; it is effective at partitioning a graph that has meaningful sub-patterns and capturing the global structure of the graph [1]. To combine multiple types of information such as images or texts in a network, multi-view spectral clustering approaches have been studied that apply general spectral clustering algorithms for each view and then combine them to minimize information loss on each view [2-7].

In this paper, we propose a multi-view network-based clustering method for integrating multi-view data. It first fuses similarity networks from each view and then carries out spectral clustering on the fused network, as shown in Figure 1. One advantage of the proposed method is that it can analyze the multiple types of networks in a comprehensive view, and we can get insight into the multiple types of networks even from a small amount of data. In addition to fusing the data, a learning task can be applied that is robust to noise and to data heterogeneity.

We demonstrate the proposed method on a social tagged-image clustering problem for challenging landmark images, where it is hard to distinguish images only with visual features. Our experiments show that the proposed method achieves better performance than other previously reported multi-view spectral clustering algorithms and shows robustness to the noise. In addition, the clustering results of the tagged-image network are further analyzed and visualized to see how the proposed method combines different types of data.

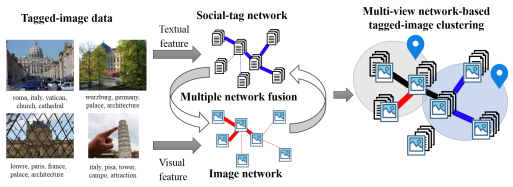

Figure 1. An overview of the proposed multi-view network-based clustering method

2. RELATED WORK

For multi-view data clustering, various spectral clustering algorithms have been studied. In [2], spectral clustering was applied in two views based on the minimizing-disagreement idea; a bipartite graph was created from each view and standard spectral clustering applied on the graph. Most recent multi-view spectral clustering approaches are based on the co-training process across different views [3, 5, 6].

Motivated by the general co-training approach, two co-regularized spectral clustering algorithms have been proposed in [4]. One is a pairwise co-regularization which maximizes the pairwise agreement on multiple views. This algorithm enforces the eigenvectors of a view pair (U^v, U^w) to have high pairwise similarity:

$$\max_{\substack{U^v\in\mathbb{R}^{n\times k}\\ U^w\in\mathbb{R}^{n\times k}}} \begin{aligned} &tr\left(U^{v^T}L^vU^v\right)+tr\left(U^{w^T}L^wU^w\right)\\ &+\lambda tr\left(\mathrm{U}^v\mathrm{U}^{v^T}U^w{U^w}^T\right)\\ &s.t.\,U^{v^T}U^v=I,U^{w^T}U^w=I. \end{aligned} \tag{1}$$

The other approach is a centroid-based co-regularization which enforces each view to look similarly towards a common centroid of the two views U*:

$$\max_{\substack{U^v\in\mathbb{R}^{n\times k}\\ U^w\in\mathbb{R}^{n\times k}}} \begin{aligned} &tr\left(U^{v^T}L^vU^v\right)+tr\left(U^{w^T}L^wU^w\right)+\\ &\lambda_v tr\left(U^vU^{v^T}U^*U^{*^T}\right)+\lambda_w tr\left(U^wU^{w^T}U^*U^{*^T}\right)\\ &s.t.\,U^{v^T}U^v=I,U^{w^T}U^w=I,U^{*^T}U^*=I\,. \end{aligned} \tag{2}$$

Both approaches need the hyper-parameter λ which trades-off the spectral clustering objectives and the spectral embedding agreement term. The authors in [4] pointed out that these two co-regularization methods are quite sensitive

to λ. Especially in the centroid based co-regularization, λ reflects the importance of each view, and noisy views could affect the optimum of U^*. Thus, it is important to select λ for all the views carefully, and it is necessary to have prior information on some of the noisy views.

Our proposed method is inspired by the similarity network fusion (SNF) approach [8] which iteratively fuses multiple networks into one similarity network representing the full spectrum of the underlying data. Our method has advantages in that we do not need to know the exact network structure of each view and can effectively combine complementary information from different types of data. It can be applied to any number of data types, which adds more flexibility to our framework. Our experimental results using social media data show that it is effective in capturing the structure of the multiple types of networks and improves the tagged-image clustering performance even from a small amount of challenging data.

3. METHODS

3.1. Dataset

We used the Flickr Div400 dataset [9], which contained social-tagged landmark images and metadata retrieved from Flickr using the name of the landmark as a query; it also provided the cluster ground truth and relevancy information generated by expert annotators.

To validate our proposed method, experiments were conducted with a development set in the Div400 dataset and only using images annotated as relevant to the location, which led to a total of 1,558 selected images for 25 landmarks.

As social landmark images share common visual features, they are hard to cluster using only visual features. Other challenges posed by this dataset is that the users may not have tagged some images or tagged some images with a duplicate or noisy information.

3.2. Construction of single-view networks

We first construct an undirected weighted graph $G = (V, E)$, where the vertex set $V = (v_1, v_2, v_3, \dots, v_n)$ is described with either the textual features or the visual descriptors on tagged images. The edge e_{ij} in E has the similarity value between two vertices v_i and v_j.

3.2.1. Text (Social-tag) features

A single image is annotated with a set of social tags, and the number of tags varies from image to image. Especially with social tags, the user tends to skip the word spacing. For example, both the tag 'irishdominicans' and the tag 'dominicans' may represent the Dominicans. However, if we simply match two tags, the similarity score becomes zero. Thus, we segment every tag set into their respective tags using the Viterbi algorithm [10], which seeks the most probable segmentation from a sequence of words based on the probability distribution of words appearing in the corpus; the one we used for segmentation was a collection of user-generated descriptions of images in the dataset.

Given the segmented tag set, we represent each tag by social term frequency-inverse document frequency (TF-IDF) [11] and Word2Vec model descriptors. The former is an adaptation of TF-IDF for the social space to reduce the effect of tagging a large number of images with the same words and assigns a social relevancy term through the use of user counts. The similarity between two tag sets X and Y is measured by the Jaccard index $J(X, Y) = \frac{|X \cap Y|}{|X \cup Y|}$.

The Word2Vec model is one of the word embedding models that represents a word with a semantic word vector trained on a large word corpus using a shallow neural network. The vector of a word which has semantic meaning can be generated with the pre-trained model of an extremely large corpus. We experimented with the skip-gram model [12] from the gensim framework [13] trained with 2 million words in the latest English Wikipedia dump. Additionally, we used a pre-trained model [14] with social tags of 4 million Flickr images. In each model, we set the maximum distance between current and predicted words within a collection of tags to 5, ignored words with a frequency lower than 5, and set the initial learning rate to be 0.025. Additionally, we skipped unseen words in the pre-trained model. Finally, each tag was represented by a vector with 400 elements. The similarity between tag sets was computed by cosine similarity.

3.2.2. Image features

First, the image was represented by the global histogram of oriented gradients (HOG) descriptor [15] computed on 3 by 3 image regions.

Besides the HOG descriptor, we considered the image features from the top-layer of a convolutional neural network (CNN) which shows human-level performance in image recognition. In our experiments, we used the top-layer of a pre-trained model of VGGnet with 19 layers [16], ResNet with 152 layers [17], and GoogLeNet [18] using the imagenet large scale visual recognition challenge (ILSVRC) dataset for feature extraction. The features of VGGnet, ResNet, and LeNet were 4096, 2048, and 1024 element vectors, respectively. The similarity between each image was measured by cosine similarity.

3.2.3. Network construction

Given pairwise similarity values, the similarity graph for each view can be defined with an affinity matrix W as

$$W(i,j) = exp\left(-\frac{\rho(v_i, v_j)}{2\alpha^2}\right). \quad (3)$$

where $\rho(v_i, v_j)$ is the distance between tagged-images v_i and v_j computed from the similarity value. The hyper parameter α is a scaling parameter that can be empirically set. We construct social tag and image networks by considering different features as described above.

3.3. Multi-view network-based tagged-image clustering

3.3.1. Multiple network fusion

To construct a tagged-image network, we combined social-tag and image networks using the SNF algorithm [8]. To compute the network fusion process, we need to compute a full and sparse kernel P with normalization on the similarity graph and the use of K nearest neighbors (KNN) to measure local affinity S as follows:

$$P(i,j) = \begin{cases} \frac{W(i,j)}{2\sum_{k \neq i} W(i,k)}, & j \neq i \\ \frac{1}{2}, & j = i \end{cases}, \quad (4)$$

$$S(i,j) = \begin{cases} \frac{W(i,j)}{\sum_{k \in N_i} W(i,k)}, & j \in N_i \\ 0, & otherwise \end{cases}, \quad (5)$$

where N_i represents a set of vertices' neighbors including vertices in the graph and W is an affinity matrix.

Note that P carries the full information about the similarity graph, and S only encodes the similarity to the K most similar vertices, which is a KNN graph of P. As this operation sets the similarities between non-neighboring points to zero, it now assigns similarities to non-neighbors through the graph diffusion process. For two different views v and w, this starts from P as an initial state using S as the kernel matrix, and iteratively updates the matrix by cross diffusion processes as follows:

$$P_{t+1}^{v}(i,j) = \sum_{k\in N_i}\sum_{l\in N_j} S^{v}(i,k) \times S^{v}(j,l) \times P_{t}^{w}(k,l),$$
$$P_{t+1}^{w}(i,j) = \sum_{k\in N_i}\sum_{l\in N_j} S^{w}(i,k) \times S^{w}(j,l) \times P_{t}^{v}(k,l). \quad (6)$$

After t iterations, the overall status matrix is computed as $P^c = \frac{P_t^v + P_t^w}{2}$. Finally, the general spectral clustering algorithm can be applied to the learned status matrix P^c.

Since the similarity information is only propagated through the common neighborhood as in equation (6), the proposed method is robust to noise. In addition, the graph diffusion process is guaranteed to converge [8].

3.3.2. Network clustering

From the fused network, we can apply any learning algorithm. In this study, we performed social-tagged landmark image clustering. The general spectral clustering algorithm was applied to the fused tagged-image network, then the tagged images were grouped into a cluster.

Here, we briefly describe the general spectral clustering algorithm [19]. An affinity matrix W for each view is defined as in Equation (3). Given the similarity graph of affinity matrix W from n images, we want to find a partition in the graph which corresponds to k clusters. The graph cut algorithm is used to minimize RatioCut [20] by solving the following objective function:

$$\max_{U\in\mathbb{R}^{n\times k}} tr(\mathrm{Q}^T\mathrm{LQ}), \; s.t. \; \mathrm{Q}^T\mathrm{Q} = I, \quad (7)$$

where $L = I - D^{-\frac{1}{2}}WD^{-\frac{1}{2}}$ is a normalized graph Laplacian, D is a network degree matrix, I is an identity matrix, $Q = U(U^TU)^{-\frac{1}{2}}$ is a scaled partition matrix and U is the embedding of the data points that can be given to the k-means algorithm to obtain cluster memberships.

3.4. Evaluation measures

The images in the dataset are grouped based on a specific landmark, which is used as a cluster ground truth. We clustered each single-view network and the fused network and evaluated the clustering results with two measures.

The adjusted Rand index (ARI) is an adjusted form of the Rand Index used to measure the similarity between two clustering results with a range between -1 and 1. For random clustering, the ARI is close to 0. Normalized Mutual Information (NMI) is a mutual information metric whose range is normalized to be between 0 and 1. Given two clustering results U and V, $NMI = \frac{I(U,V)}{\sqrt{H(U)H(V)}}$, where $I(U,V)$ is mutual information on U and V, which measures the mutual dependence between two clustering results, and $H(U)$ and $H(V)$ are the entropies of each clustering result. The higher the ARI or NMI, the higher the concordance with the cluster ground truth.

4. RESULTS

4.1. Comparison of different descriptors

We first compared the clustering accuracy using different descriptors on single-view networks. For each single-view network, the affinity matrix W in equation (3) was constructed by varying parameter α from 0.2 to 2.0.

For the social-tag network, social TF-IDF descriptors and Word2Vec models trained using an English Wikipedia dump and Flickr corpus were used. The NMIs score of the clustering results are shown in Figure 2. The best performance was obtained by the simple social TF-IDF descriptor. Although the social-tag set was quite noisy and some of the user-generated tags were not seen in the pre-trained Word2Vec model, the performance difference between the best-performing network and the next best was not very big considering that it was accomplished with single-view network clustering only.

Likewise, we compared the HOG feature and features from the three CNN models in image network clustering. As in many recent researches, CNN features outperformed the HOG feature, which shows that images in this dataset were identified much more accurately than using the HOG feature. Although the performance difference between CNN features was not big, the network with the ResNet feature showed the best performance. Due to the challenging features of the dataset, the overall performance of image network clustering was worse than the social-tag network.

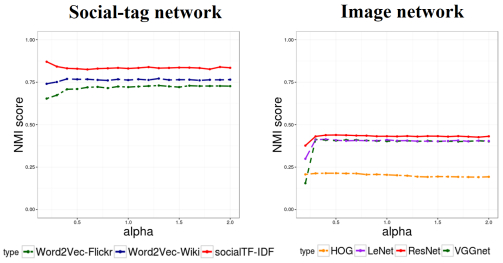

Figure 2. NMI score of the clustering results using different descriptors across different values of α in the similarity matrix W

4.2. Comparison between different algorithms

In this section, we compare the performance of the proposed multi-view network-based clustering method with other existing methods. To achieve this, we combined the best-performing networks for each view, the text network with social TF-IDF descriptors and the image network with ResNet features. Then, the proposed method was compared with minimizing-disagreement spectral clustering and co-regularized spectral clustering introduced in Section 2.

The scaling parameter α of the affinity matrix in equation (3) is empirically set to the optimal one which shows the best performance for each algorithm from the experiments with varying α from 0.2 to 1. In the co-regularized methods, the performance can be sensitive to the additional parameter λ in equation (1,2) and we chose the optimal λ by varying λ from 0 to 1 incremented by 0.1. The performance of pairwise co-regularization method

varied from 0.7997 to 0.8572 (NMI). For centroid-based co-regularization, we need to set two lambda values for each view. As we don't have prior information of which view is more noisy, we experimented with the same λ and the performance varied from 0.8415 to 0.853 (NMI). In the proposed multi-view network-based clustering algorithm, the number of neighbors N in equation (6) was set to 30. As the SNF and co-regularized methods converged after around 10-15 iterations, we set this to 15. The optimal parameters for each method are summarized in Table 1.

The proposed method showed the best performance (NMI = 0.897) as shown in Figure 3. Furthermore, this proves that insufficient information in the image network was greatly supplemented by combining the social tags, and the proposed method minimized the information loss when combining two features compared to other previously reported multi-view clustering algorithms.

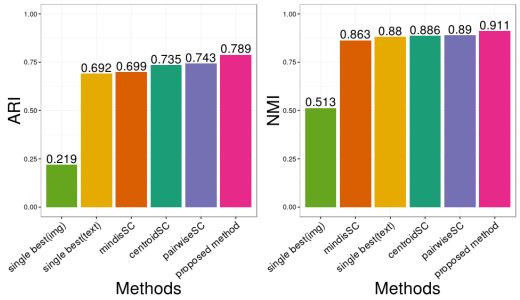

Figure 3. Performance comparison between different algorithms in terms of ARI and NMI scores of the clustering results

Table 1. The optimal parameter selection on each multi-view clustering algorithms (α_v and α_w are for the text network and image network respectively)

Method	α_v	α_w	λ	ARI	NMI
Min-disagreement	0.7	0.4	-	0.7318	0.8539
Pairwise co-reg.	0.2	0.7	0.2	0.7528	0.8666
Centroid co-reg.	0.2	1	0.1	0.7555	0.8734
Proposed method	0.3	0.2	-	**0.7972**	**0.8968**

4.3. Multi-view network based cluster analysis

The proposed multi-view network-based clustering method not only improved the clustering performance, but was also an effective tool to gain insight into multiple featured networks. As we first fused the multiple networks and then applied the network clustering algorithm, we were able to visualize how well the images are clustered in the fused network, as shown in Figure 4 in which the heat-map of the affinity matrix for the fused similarity graph using the proposed method is presented. The heat-map shows that some clusters with only single features which are hard to distinguish between became dissimilar to each other when combining two views using multi-view network-based clustering, which led to improved clustering performance as a result. The heatmap of the fused similarity graph in Figure 4 proves that the proposed multi-view network-based clustering method is able to reduce the noise and data heterogeneity when combining different types of data.

For example, the sample images clustered as 'Wurzburg Residence Germany' are shown using single-view and multi-view network-based clustering in Figure 5. As can be seen, the social images are very hard to cluster with visual features and can be noisy using the social-tag set only. The multi-view network-based clustering method was quite effective in combining the complementary information with two different views and minimizing the noise in the clustering at the same time.

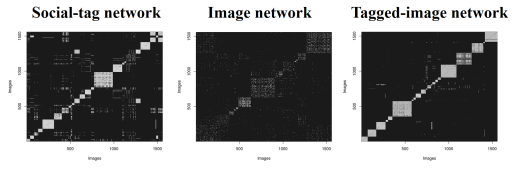

Figure 4. Heatmap of the proposed multi-view network-based clustering results

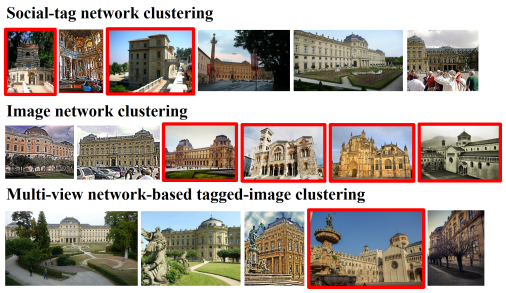

Figure 5. Sample images clustered as 'Wurzburg Residence Germany' with single-view network clustering and the proposed multi-view network-based clustering. The red boxes mark the images actually in different clusters

5. CONCLUSIONS

We proposed a multi-view network-based clustering framework for social-tagged landmark image clustering and demonstrated that the proposed method is effective in fusing networks with different data types. In our experiments, the proposed method was validated by applying it to the tagged-image clustering problem with data that made it quite hard to distinguish items with either textual or visual features only. The clustering results of the proposed method were competitive to other previously reported multi-view clustering algorithms. Furthermore, the experiments with the challenging landmark images showed that the proposed method was robust to the noise in the data when clustering different types of data.

As the multiple network fusion process is not scalable to the network with more than 10,000 nodes, we only validated with 1,558 images due to the computational problem. For the future work, the network fusion algorithm can be extended to be scalable for the larger data with parallelization.

6. ACKNOWLEDGEMENT

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education [NRF-2016R1D1A1B03933875] and by the Ministry of Science, ICT & Future Planning [2014R1A1A3051169].

7. REFERENCES

[1] U. Von Luxburg, "A tutorial on spectral clustering," *Statistics and computing,* vol. 17, pp. 395-416, 2007.

[2] V. R. De Sa, "Spectral clustering with two views," in *ICML workshop on learning with multiple views*, 2005, pp. 20-27.

[3] A. Kumar and H. 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.

[4] A. Kumar, P. Rai, and H. Daume, "Co-regularized multi-view spectral clustering," in *Advances in neural information processing systems*, 2011, pp. 1413-1421.

[5] C.-K. Lee and T.-L. Liu, "Guided co-training for multi-view spectral clustering," in *Image Processing (ICIP), 2016 IEEE International Conference on*, 2016, pp. 4042-4046.

[6] H. Wang, C. Weng, and J. Yuan, "Multi-feature spectral clustering with minimax optimization," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 4106-4113.

[7] R. Xia, Y. Pan, L. Du, and J. Yin, "Robust Multi-View Spectral Clustering via Low-Rank and Sparse Decomposition," in *AAAI*, 2014, pp. 2149-2155.

[8] B. Wang, A. M. Mezlini, F. Demir, M. Fiume, Z. Tu, M. Brudno*, et al.*, "Similarity network fusion for aggregating data types on a genomic scale," *Nature methods,* vol. 11, pp. 333-337, 2014.

[9] B. Ionescu, A.-L. Radu, M. Menéndez, H. Müller, A. Popescu, and B. Loni, "Div400: a social image retrieval result diversification dataset," in *Proceedings of the 5th ACM Multimedia Systems Conference*, 2014, pp. 29-34.

[10] G. D. Forney, "The viterbi algorithm," *Proceedings of the IEEE,* vol. 61, pp. 268-278, 1973.

[11] A. Popescu and G. Grefenstette, "Social media driven image retrieval," in *Proceedings of the 1st ACM International Conference on Multimedia Retrieval*, 2011, p. 33.

[12] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in neural information processing systems*, 2013, pp. 3111-3119.

[13] R. Rehurek and P. Sojka, "Software framework for topic modelling with large corpora," in *In Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, 2010.

[14] X. Li, S. Liao, W. Lan, X. Du, and G. Yang, "Zero-shot image tagging by hierarchical semantic embedding," in *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2015, pp. 879-882.

[15] O. L. Junior, D. Delgado, V. Gonçalves, and U. Nunes, "Trainable classifier-fusion schemes: An application to pedestrian detection," in *Intelligent Transportation Systems, 2009. ITSC'09. 12th International IEEE Conference on*, 2009, pp. 1-6.

[16] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556,* 2014.

[17] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770-778.

[18] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov*, et al.*, "Going deeper with convolutions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1-9.

[19] A. Y. Ng, M. I. Jordan, and Y. Weiss, "On spectral clustering: Analysis and an algorithm," in *NIPS*, 2001, pp. 849-856.

[20] Y.-C. Wei and C.-K. Cheng, "Towards efficient hierarchical designs by ratio cut partitioning," in *Computer-Aided Design, 1989. ICCAD-89. Digest of Technical Papers., 1989 IEEE International Conference on*, 1989, pp. 298-301.