Multi-View Clustering Based on Belief Propagation

Chang-Dong Wang, *Member, IEEE*, Jian-Huang Lai, *Senior Member, IEEE*, and Philip S. Yu, *Fellow, IEEE*

Abstract—The availability of many heterogeneous but related views of data has arisen in numerous clustering problems. Different views encode distinct representations of the same data, which often admit the same underlying cluster structure. The goal of multi-view clustering is to properly combine information from multiple views so as to generate high quality clustering results that are consistent across different views. Based on max-product belief propagation, we propose a novel multi-view clustering algorithm termed multi-view affinity propagation (MVAP). The basic idea is to establish a multi-view clustering model consisting of two components, which measure the within-view clustering quality and the explicit clustering consistency across different views, respectively. Solving this model is NP-hard, and a multi-view affinity propagation is proposed, which works by passing messages both within individual views and across different views. However, the exemplar consistency constraint makes the optimization almost impossible. To this end, by using some previously designed mathematical techniques, the messages as well as the cluster assignment vector computations are simplified to get simple yet functionally equivalent computations. Experimental results on several real-world multi-view datasets show that MVAP outperforms existing multi-view clustering algorithms. It is especially suitable for clustering more than two views.

Index Terms—Clustering, multiple view, factor graph, max-product belief propagation

1 Introduction

Recently, many heterogeneous but related views of data have arisen in a number of fields, such as pattern recognition, social network mining, bioinformatics, natural language processing, computer vision, etc. [1], [2], [3], [4], [5]. For instance, in face categorization databases, to avoid the effect of uneven illumination distribution, not only visible light (VIS) but also near-IR (NIR) face images are captured in each instance, which are taken as two heterogeneous but related views [6]. In link-based document databases, the document similarity may not only be represented by the document content, which is taken as one view, but also be captured via their link relations such as citations in academic publications or hyperlink in webpages [7], which is taken as another view. Another similar example is in the case of multilingual documents where documents are available in more than one languages and each language is taken as a separate view.

Although it might be sufficient to learn from individual views, properly combining information from these related views will improve the learning performance. This leads to the emergence of a challenging machine learning problem termed multi-view learning [1], [8]. Recently, multi-view learning has been widely investigated from different perspectives, such as domain adaption [9], [10], transfer learning [11], active learning from multiple annotators [12], multiple kernel learning [13], [14], multi-view classification [2], [15], multi-view clustering [16], [17], [18], etc. From the viewpoint of both supervised classification and unsupervised clustering, the advantage of multi-view learning is that different views admit the same underlying class structure of the data [19]. Therefore, in multi-view clustering, the generated clusters should not only best capture the cluster structure in individual views but also be consistent across different views [3], [18], [20]. To this end, a multi-view model needs to be established to properly combine information from multiple views. However, directly combining multiple views by means of feature concatenation or similarity matrix addition would not help improve the clustering performance, especially when the multiple views are from different representation spaces.

In this paper, we propose a similarity-based clustering approach for multi-view clustering, termed multi-view affinity propagation (MVAP). A multi-view clustering model is firstly established, which generates high quality clustering results that are consistent across different views. The exemplar representation strategy is used, where each cluster is represented by one exemplar. To maximize the clustering quality in individual views, each data point should be assigned to the most suitable exemplar in each view; while to ensure the clustering consistency across views, the exemplars that each data point is assigned to should be as consistent as possible across different views. To this end, a two-term global objective function is proposed. The first term takes into account the sum of

- *C.-D. Wang is with the School of Mobile Information Engineering, Sun Yat-sen University, Zhuhai, P.R. China, and the SYSU-CMU Shunde International Joint Research Institute (JRI), Shunde, China. E-mail: changdongwang@hotmail.com.*
- *J.-H. Lai is with the School of Information Science and Technology, Sun Yat-sen University, Guangzhou, P.R. China, and Guangdong Key Laboratory of Information Security Technology, Guangzhou, China. E-mail: stsljh@mail.sysu.edu.cn.*
- *P.S. Yu is with the Department of Computer Science, University of Illinois at Chicago, Chicago, IL, and the Institute for Data Science, Tsinghua University, Beijing, China. E-mail: psyu@cs.uic.edu.*

Manuscript received 10 Jan. 2015; revised 9 Sept. 2015; accepted 20 Nov. 2015. Date of publication 25 Nov. 2015; date of current version 3 Mar. 2016. Recommended for acceptance by I. Davidson.
For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.
Digital Object Identifier no. 10.1109/TKDE.2015.2503743

similarities between each data point and its assigned exemplar in each view; while the second term takes into account the explicit clustering consistency across different views according to the item-wise exemplar consistency.

Directly searching for the optimal clusterings that maximize the objective function is NP-hard. To this end, based on the max-sum (the $\log$-domain max-product) belief propagation [21], we represent the objective function using factor graph and propose a multi-view affinity propagation to solve the model. It works by passing messages not only within individual views to maximize the within-view clustering quality but also across different views to ensure the cross-view clustering consistency. By using some previously designed mathematical techniques [22], [23], [24], these messages and the cluster assignment vector computations are simplified so as to get simple yet functionally equivalent computations.

To evaluate the effectiveness of the proposed approach, extensive experiments have been conducted on five real-world multi-view datasets. Experimental results show that the proposed MVAP approach outperforms the existing multi-view clustering algorithms. We also show that directly combining multiple views by means of feature concatenation or similarity matrix addition would not help improve the clustering performance. Comparison results on the triple-view datasets show that the proposed approach is especially suitable for clustering more than two views.

To summarize, this paper has the following contributions:

- A multi-view model is established to model the heterogeneous but related views, which simultaneously considers the clustering quality within individual view and the explicit clustering consistency across different views. A two-term objective function is consequently derived which is represented using factor graph.
- A multi-view affinity propagation is proposed to solve the model, which works by passing messages both within individual views and across different views.
- Extensive experiments have been conducted on several real-world multi-view datasets. The experimental results show that the proposed MVAP algorithm outperforms the existing multi-view clustering algorithms and is especially suitable for clustering more than two views.

The rest of this paper is organized as follows. Section 2 introduces the related works in multi-view clustering. In Section 3, we describe the proposed multi-view affinity propagation approach. In Section 4, extensive experiments have been conducted to show the effectiveness of the proposed approach. Section 5 concludes this paper.

2 Related Work

In the clustering literature, some efforts have been made to combine information from multiple views so as to generate better clustering results [3], [16], [17], [18], [20], [25], [26], [27], [28], [29], [30], [31]. In what follows, we will briefly introduce the related multi-view clustering methods in combining information from multiple views.

One early approach is to assess similarity matrix so as to incorporate disparate forms of information from different views. For instance, in [25], Bayesian network approach was used to construct a similarity matrix that fuses information from multiple views. In [26], a link matrix factorization was proposed to fuse information from multiple graph views, where multiple graphs represent relations of vertices from different views. Some other methods have been developed. In [17], a hierarchical non-parametric clustering model was designed for multi-view data. In [27], for clustering multilingual documents, a heuristic two-step approach was developed. At the first stage, the probabilistic latent semantic analysis (PLSA) is applied independently over each language (i.e., view) to obtain voting pattern, i.e., the topic pattern representing cluster signatures for each document. At stage two, the voting pattern is used to group documents such that documents belonging to the same group share similar voting patterns.

The view constrained strategy is another method that can be used to combine multiple views. In [3], Taralova et al. developed a view constrained clustering (SCC) that extends k-means by enforcing each cluster to contain samples from a minimum number of views. The SCC approach iterates between clustering and metric learning so as to minimize an objective function that takes into account grouping samples which are similar and representative of the views.

However, one common shortcoming of the aforementioned approaches is that they haven't taken into account the consistency of cluster structure among different views. In [19], Li et al. proposed a multi-view semi-supervised learning method, which trains a classifier in each view based on both labeled and unlabeled samples, where all classifiers are required to assign the same class label to each labeled and unlabeled sample by imposing a global constraint. Although this work does consider the consistency of cluster structure among different views, it is limited to semi-supervised learning which is not applicable without some labeling information.

Tensor methods have also been used for discovering latent pattern hidden in multi-view data [32], [33], [34], [35]. Selee et al. [32] reported a tensor decomposition called Implicit Slice Canonical Decomposition (IMSCAND) for clustering with multiple similarity matrixes, which stores each similarity as a slice in a tensor. Aiming to meet the need that multiple views can have different dimensions and weightings, Liu et al. [34] further proposed a tensor-based method for multi-view clustering, in which two new formulations are developed based on the Frobenius-norm objective function that model the multi-view data as a tensor and seek a joint latent optimal subspace by tensor analysis. In [35], a robust tensor clustering (RTC) was developed for face clustering with various noises. However, the tensor factorization methods only seek consistent eigenvectors across different views instead of directly enforcing consistent clustering labels across different views.

Due to the well-motivated mathematical framework, spectral clustering has been widely investigated for extending to multi-view clustering, which forms one major methodology in multi-view clustering [18], [20], [28], [29], [30]. Zhou and Burges [28] developed multi-view spectral clustering by generalizing the single-view normalized cut to the

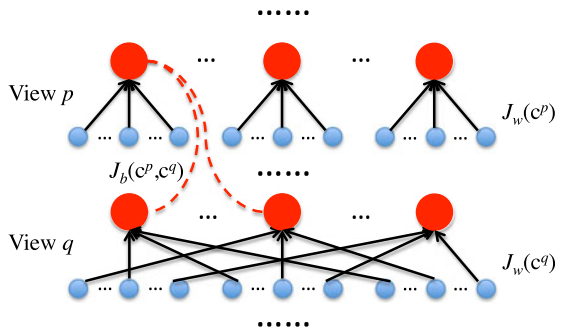


Fig. 1. Illustration of the multi-view model. The term $J_w(\mathbf{c}^p)$ measures the sum of the clustering quality within the pth view, and the term $J_b(\mathbf{c}^p, \mathbf{c}^q)$ measures the consistency of cluster assignment vectors $\mathbf{c}^p$ and $\mathbf{c}^q$.

multi-view case, where the goal is to find a cut that is good enough on average while it may not be the best for a single graph. Unlike the previous single-objective case, a multi-objective formulation was developed in [30], which simultaneously considers the quality of a single cut on multiple graphs. To solve this multi-objective function, Pareto optimization is utilized, which enforces multiple objectives to compete with each other in order to get a common single cut. In [18] and [20], two spectral clustering frameworks have been developed for multi-view clustering, which respectively use co-training and co-regularization to make the eigenvectors agree with each other across different views, and then apply the k-means method to the resulting eigenvectors to generate the final clustering results for each view respectively. Wang et al. [29] proposed a constrained spectral clustering, termed csp-p, which takes one view as the similarity matrix but encodes the other view as the constraints.

In the above spectral-based multi-view clustering methods, only [18] and [20] have considered the mechanism in trading-off the clustering quality within individual views with the clustering consistency across different views, which is critical in multi-view clustering, especially in the case of more than two views. However, similar to the tensor factorization methods, the two spectral-based methods achieve the goal by enforcing consistent eigenvectors across different views instead of directly enforcing consistent clustering labels across different views. That is, only implicit clustering consistency is achieved instead of explicit clustering consistency.

In this paper, we aim to address the above shortcomings by designing a novel model that directly takes into account the *balance* between the clustering quality within individual views *and* the explicit clustering consistency across different views, so as to generate high quality clustering results that are consistent across different views. A two-term global objective function is developed, which simultaneously maximizes the clustering quality in individual views while ensures the clustering consistency across different views.

3 Multi-View Affinity Propagation

3.1 The Multi-View Model

Assume that we are given a dataset consisting of n instances which are represented in m heterogeneous but related views, and in the pth view, $\forall p = 1, \ldots, m$, the normalized similarity matrix of these n instances is denoted as $\tilde{S}^p \in \mathbb{R}^{n \times n}$. By saying normalized similarity matrix, we mean that the values of the matrix are normalized to be in the range $[0, 1]$. The goal is to combine information embedded in these m similarity matrixes, so as to generate high quality clustering results that are consistent across different views. To this end, a multi-view model should be designed that considers both the clustering quality within each view and the clustering consistency across different views.

Using exemplar-based clustering methodology, the goal of multi-view clustering can be transferred to find m cluster assignment vectors $\{\mathbf{c}^1, \mathbf{c}^2, \ldots, \mathbf{c}^m\}$. The pth cluster assignment vector $\mathbf{c}^p = [c_1^p, c_2^p, \ldots, c_n^p]$ assigns each data point $\mathbf{x}_i$ to the most suitable exemplar (i.e., the c_i^pth data point) in the pth view, and these cluster assignment vectors should be as consistent as possible. To simultaneously consider the *clustering quality within individual views* and *explicit clustering consistency across different views*, we need to find the m cluster assignment vectors that maximize the following global objective function,

$$J(\{\mathbf{c}^1, \ldots, \mathbf{c}^m\}) = \sigma \sum_{p=1}^{m} J_w(\mathbf{c}^p) + (1-\sigma) \sum_{p=1,q=1,p\neq q}^{m} J_b(\mathbf{c}^p, \mathbf{c}^q), \tag{1}$$

where the first term $\sigma \sum_{p=1}^{m} J_w(\mathbf{c}^p)$ measures the sum of the clustering quality within individual views, and the second term $(1-\sigma) \sum_{p=1,q=1,p\neq q}^{m} J_b(\mathbf{c}^p, \mathbf{c}^q)$ measures the sum of the consistency of $m(m-1)$ pairs of cluster assignment vectors. And $\sigma \in (0, 1]$ is a parameter trading-off the within-view clustering quality against the cross-view clustering consistency. Fig. 1 illustrates the multi-view model. It is worthwhile mentioning that, in the first term, there is a constraint enforcing the exemplar consistency within individual views, as will be defined below. Therefore, to guarantee that the exemplars are consistent within individual views, i.e., to generate valid clustering results in individual views, the trade-off parameter σ must be larger than 0.

3.1.1 Within-View Clustering Quality

Aiming at maximizing the sum of similarities between each data point[1] i and its assigned exemplar $\mathbf{c}_i^p$, following the widely used affinity propagation (AP) clustering algorithm [22], [36], the clustering quality $J_w(\mathbf{c}^p)$ within the pth view is defined as

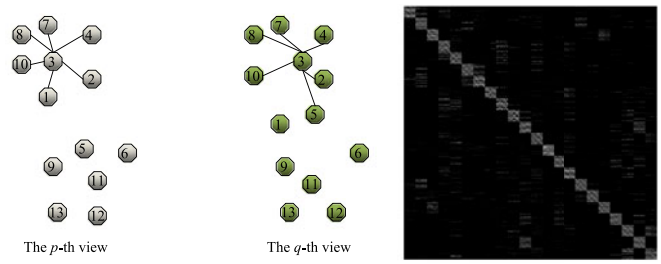
$$J_w(\mathbf{c}^p) = \sum_{i=1}^{n} \tilde{S}^p(i, \mathbf{c}_i^p) + \sum_{k=1}^{n} \tilde{\delta}_k^p(\mathbf{c}^p), \tag{2}$$

where $\tilde{\delta}_k^p(\mathbf{c}^p)$ is the exemplar consistency constraint in the pth view defined as follows,

$$\tilde{\delta}_k^p(\mathbf{c}^p) = \begin{cases} -\infty & \text{If } \mathbf{c}_k^p \neq k \text{ but } \exists i : \mathbf{c}_i^p = k \\ 0 & \text{Otherwise.} \end{cases}$$

That is, if some data point i selects data point k as its exemplar ($\mathbf{c}_i^p = k$), data point k must select itself as its exemplar ($\mathbf{c}_k^p = k$).

1. Throughout this paper, we will alternatively use subscript to refer to data point. That is, data point i refers to the ith data point, i.e., $\mathbf{x}_i$. Similarly, exemplar $\mathbf{c}_i^p$ refers to the exemplar $\mathbf{x}_{\mathbf{c}_i^p}$.



(a) $\tilde{L}^{pq}(3,3)$. (b) VIS/NIR dataset.

Fig. 2. Illustration of data view consistency: (a) Data view consistency across the pth and qth data views: $\tilde{L}^{pq}(3,3) = \frac{|\{1,2,3,4,7,8,10\} \bigcap \{2,3,4,5,7,8,10\}|}{|\{1,2,3,4,7,8,10\} \bigcup \{2,3,4,5,7,8,10\}|} = \frac{3}{4}$; (b) Data consistency matrix of a VIS/NIR dataset consisting of the VIS and NIR face images.

Please notice that this definition of the clustering quality is quite different from the clustering quality used by the classical k-means, and see [22], [24] for detail.

3.1.2 Cross-View Clustering Consistency

To measure the cross-view clustering consistency, we need to define the data consistency for the m views as follows.

For the pth view, let $\mathcal{N}_k^p(\mathbf{x})$ denote the k (e.g., $k = 30$ in our experiments, which generates the best results) nearest neighbors of $\mathbf{x}$ according to the similarity matrix $\tilde{S}^p$. The item-wise data consistency between the ith data point (i.e., $\mathbf{x}_i$) in the pth view and the jth data point (i.e. $\mathbf{x}_j$) in the qth view is defined as the Jaccard similarity of their k nearest neighbor sets. Let $\tilde{L}^{pq}(i,j)$ denote this data consistency, i.e.,

$$\begin{aligned}\tilde{L}^{pq}(i,j) &= JaccardSim(\mathcal{N}_k^p(\mathbf{x}_i), \mathcal{N}_k^q(\mathbf{x}_j)) \\ &= \frac{|\mathcal{N}_k^p(\mathbf{x}_i) \bigcap \mathcal{N}_k^q(\mathbf{x}_j)|}{|\mathcal{N}_k^p(\mathbf{x}_i) \bigcup \mathcal{N}_k^q(\mathbf{x}_j)|}.\end{aligned} \tag{4}$$

That is, the item-wise data consistency between two data points across two different views is defined according to their neighborhood consistency, as shown in Fig. 2a. The underlying rationale is that, if two data points across two different views share a larger degree of neighborhood information, they will be more consistent. For the entire dataset, this definition of consistency will well capture the relation of data distributions across different views. More specifically, we can use the mean value of the diagonal elements of $\tilde{L}^{pq}$ to measure the similarity of data distributions across views p and q, i.e., $\frac{1}{n}\sum_{i=1}^{n} \tilde{L}^{pq}(i,i)$. Based on this definition, the data distributions of two views p and q are said to be very similar if $\frac{1}{n}\sum_{i=1}^{n} \tilde{L}^{pq}(i,i)$ is large. In our experiments in Section 4.2, we have confirmed the effectiveness of this data consistency definition. That is, data views of high consistency will result in high cross-view clustering consistency measured by NMI [37]. Therefore, for the pth view and qth view, a data consistency matrix $\tilde{L}^{pq} \in \mathbb{R}^{n \times n}$ can be obtained to evaluate the data consistency across the two views. In this way, for all m views, $m(m-1)$ consistency matrices $\{\tilde{L}^{pq}, p = 1, \ldots, m, q = 1, \ldots, m, p \neq q\}$ can be obtained.

Fig. 2b plots the data consistency matrix of a VIS/NIR face image dataset, which is collected by Zou from University of Surrey [6]. The VIS/NIR dataset contains 1,056 face images belonging to 22 subjects, with 48 images for each subject. Each face image is in the form of both visible light and near-IR, which are taken as two views respectively. For clarity, the images are sorted in the order of the ground-truth class labels. From the figure, we can see that the defined data consistency has a clear property that data points belonging to the same class are more consistent across two views than data points belonging to different classes.

Using the definition of data view consistency, the clustering consistency $J_b(\mathbf{c}^p, \mathbf{c}^q)$ across the pth and qth views is defined as

$$J_b(\mathbf{c}^p, \mathbf{c}^q) = \sum_{i=1}^{n} \tilde{L}^{pq}(\mathbf{c}_i^p, \mathbf{c}_i^q). \tag{5}$$

In this definition, the clustering consistency across two views is defined according to the item-wise exemplar consistency. That is, the consistency of cluster assignments of data point $\mathbf{x}_i$ across the pth and qth views is defined as its exemplar consistency across two views.

The underlying rationale of using neighborhood similarity of exemplars as cross-view clustering consistency is as follows. It allows one data point to be assigned to slightly different exemplars in different views. This is because, although different views admit the same underlying cluster structure, they can have slightly different exemplars. If we use other principles (e.g., simply comparing the clusterings in different views via NMI [37] and using this value as a similarity score), it will enforce exactly the same partitioning across different views, which is too strict to be applicable. Based on the above analysis, we can see that in the case of high cross-view consistency, the same cluster structure across different views will be almost guaranteed, though slightly different exemplars are allowed across different views. Concerning the selection of nearest neighbor parameter k, by definition of (4) and (5), we can see that, when very small k is used, say $k = 1$, high cross-view consistency, i.e., high value of (5) would result in the clustering results having the same exemplar across different views. On the other hand, when very large k is used, say k approaching the number of data points, the cross-view consistency will take no effect on generating consistent clustering results across different views, i.e., leading to the other extreme. Therefore, a moderate k should be used. In our experiments, setting k in interval $[25, 35]$ would generate satisfactory clustering results on all datasets, with $k = 30$ being the best, i.e., high quality clustering results that are consistent across different views.

3.1.3 The Global Objective Function

Based on the above definitions, the objective function (1) can be expanded as

$$\begin{aligned}J(\{\mathbf{c}^1, \ldots, \mathbf{c}^m\}) = &\sum_{p=1}^{m}\left(\sum_{i=1}^{n} \sigma \cdot \tilde{S}^p(i, \mathbf{c}_i^p) + \sum_{k=1}^{n} \sigma \cdot \tilde{\delta}_k^p(\mathbf{c}^p)\right) \\ &+ \sum_{p=1,q=1,p\neq q}^{m}\left(\sum_{i=1}^{n} (1-\sigma) \cdot \tilde{L}^{pq}(\mathbf{c}_i^p, \mathbf{c}_i^q)\right).\end{aligned} \tag{6}$$

According to the definition of the exemplar consistency constraint function (3), if $\sigma > 0$, then a new constraint function

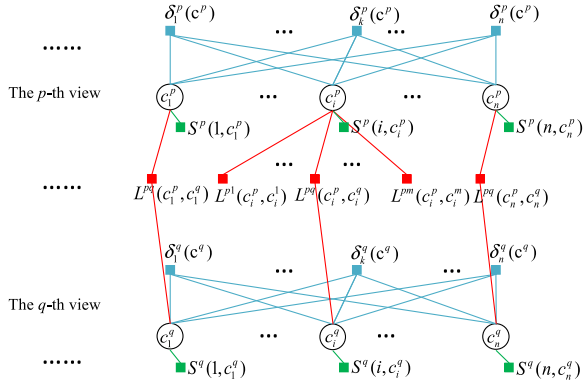


Fig. 3. Factor graph of multi-view affinity propagation.

$\delta_k^p = \sigma \cdot \tilde{\delta}_k^p$ also works. For notation simplicity, we denote $S^p = \sigma \cdot \tilde{S}^p$ and $L^{pq} = (1-\sigma) \cdot \tilde{L}^{pq}$. That is, the trade-off between the within-view clustering quality and the cross-view clustering consistency can be encoded in the input similarity matrices and data consistency matrices. In consequence, the expanded function (6) can be written as

$$J(\{\mathbf{c}^1, \ldots, \mathbf{c}^m\}) = \sum_{p=1}^{m} \left(\sum_{i=1}^{n} S^p(i, \mathbf{c}_i^p) + \sum_{k=1}^{n} \delta_k^p(\mathbf{c}^p) \right) + \sum_{p=1,q=1,p\neq q}^{m} \left(\sum_{i=1}^{n} L^{pq}(\mathbf{c}_i^p, \mathbf{c}_i^q) \right). \tag{7}$$

Please note that the proposed model can be viewed as a generalization of the single-view affinity propagation model [22]. First of all, when the number of views m reduces to one, the two-term global objective function (7) contains only the within-view clustering quality, which is exactly the same as the objective function of AP [22]. Secondly, when there are multiple views but the trade-off parameter σ is set to 1, the model degenerates into applying single-view affinity propagation model in individual views.

3.2 Multi-View Affinity Propagation

Directly searching for the optimal assignment vectors that maximize the objective function (7) is NP-hard. This is because the multi-view model is a generalization of the single-view one, the optimization of which is proven to be NP-hard [22]. To this end, a multi-view affinity propagation is proposed, which is based on the max-sum (the log-domain max-product) belief propagation. The max-sum belief propagation is a local-message-passing algorithm that is guaranteed to converge to the neighborhood maximum [38]. To apply max-sum belief propagation, the objective function (7) is represented by a factor graph, which is shown in Fig. 3, where the circle nodes represent the variables, i.e., cluster assignments, and the square nodes represent the functions in (7). In factor graph representation, if one function depends on one variable, then there is an edge linking the two corresponding nodes.

Following the max-sum belief propagation theory, each edge is associated with a pair of real-valued messages passing between the ending variable node and function node. Therefore, in our factor graph representation, there are four types of messages passing between variable nodes and

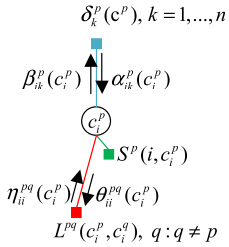


Fig. 4. Messages of multi-view affinity propagation.

function nodes, as shown in Fig. 4, which are denoted as $\alpha_{ik}^p(\mathbf{c}_i^p)$, $\beta_{ik}^p(\mathbf{c}_i^p)$, $\eta_{ii}^{pq}(\mathbf{c}_i^p)$ and $\theta_{ii}^{pq}(\mathbf{c}_i^p)$ respectively. The first two types are within-view messages, while the last two types are cross-view messages. Consequently, the proposed message passing scheme is termed multi-view affinity propagation. It should be noted that, there is another pair of messages passing between function $S^p(i, \mathbf{c}_i^p)$ and variable $\mathbf{c}_i^p$. However, since the value of the message passing from function $S^p(i, \mathbf{c}_i^p)$ to variable $\mathbf{c}_i^p$ always equals the similarity $S^p(i, \mathbf{c}_i^p)$, and the message passing from variable $\mathbf{c}_i^p$ to function $S^p(i, \mathbf{c}_i^p)$ is not used in other messages, we will use $S^p(i, \mathbf{c}_i^p)$ to denote the message passing from function $S^p(i, \mathbf{c}_i^p)$ to variable $\mathbf{c}_i^p$, and ignore the message passing from variable $\mathbf{c}_i^p$ to function $S^p(i, \mathbf{c}_i^p)$.

To obtain optimal variable assignments, all these messages are initialized as zero, and then they are updated via passing a message either from a variable to each adjacent function or from a function to each adjacent variable. The message from a variable to an adjacent function sums together the messages from all adjacent functions except the one receiving the message [21],

$$\mu_{x\to f}(x) \leftarrow \sum_{h\in \mathrm{ne}(x)\setminus\{f\}} \mu_{h\to x}(x), \tag{8}$$

where $\mathrm{ne}(x)$ denotes the set of functions adjacent to variable x. The message from a function to an adjacent variable involves a maximization over all arguments of the function except the variable receiving the message [21],

$$\mu_{f\to x}(x) \leftarrow \max_{X\setminus\{x\}} \left[f(X) + \sum_{y\in X\setminus\{x\}} \mu_{y\to f}(y) \right], \tag{9}$$

where $X = \mathrm{ne}(f)$ is the set of arguments of function f.

Therefore, the within-view messages $\alpha_{ik}^p(\mathbf{c}_i^p)$ and $\beta_{ik}^p(\mathbf{c}_i^p)$, and cross-view messages $\eta_{ii}^{pq}(\mathbf{c}_i^p)$ and $\theta_{ii}^{pq}(\mathbf{c}_i^p)$ are defined respectively as follows.

1) Within-view messages:

$$\begin{aligned} &\forall p = 1, \ldots, m, \quad \forall i = 1, \ldots, n, \quad \forall k = 1, \ldots, n \\ &\alpha_{ik}^p(\mathbf{c}_i^p) = \mu_{\delta_k^p \to \mathbf{c}_i^p}(\mathbf{c}_i^p) \\ &\qquad = \max_{\substack{j_1,\ldots,\\ j_{i-1},j_{i+1}\\ ,\ldots,j_n}} \left[\delta_k^p(j_1, \ldots, \mathbf{c}_i^p, \ldots, j_n) + \sum_{i':i'\neq i} \beta_{i'k}^p(j_{i'}) \right] \end{aligned} \tag{10}$$

$$\begin{aligned}\beta_{ik}^{p}(\mathbf{c}_i^p) &= \mu_{\mathbf{c}_i^p \to \delta_k^p(\mathbf{c}^p)}(\mathbf{c}_i^p) \\ &= S^p(i,\mathbf{c}_i^p) + \sum_{k':k'\neq k} \alpha_{ik'}^{p}(\mathbf{c}_i^p) + \sum_{q=1,q\neq p}^{m} \eta_{ii}^{pq}(\mathbf{c}_i^p).\end{aligned} \tag{11}$$

2) Cross-view messages:

$$\begin{aligned}&\forall p,q = 1,\ldots,m,\ p\neq q,\ \ \forall i = 1,\ldots,n \\ &\eta_{ii}^{pq}(\mathbf{c}_i^p) = \mu_{L^{pq}(\mathbf{c}_i^p,\mathbf{c}_i^q)\to \mathbf{c}_i^p} \\ &\qquad = \max_{\mathbf{c}_i^q}\left[L^{pq}(\mathbf{c}_i^p,\mathbf{c}_i^q) + \theta_{ii}^{qp}(\mathbf{c}_i^q)\right]\end{aligned} \tag{12}$$

$$\begin{aligned}\theta_{ii}^{pq}(\mathbf{c}_i^p) &= \mu_{\mathbf{c}_i^p \to L^{pq}(\mathbf{c}_i^p,\mathbf{c}_i^q)} \\ &= S^p(i,\mathbf{c}_i^p) + \sum_{k=1}^{n}\alpha_{ik}^{p}(\mathbf{c}_i^p) + \sum_{q'=1,q'\neq q,q'\neq p}^{m}\eta_{ii}^{pq'}(\mathbf{c}_i^p).\end{aligned} \tag{13}$$

To estimate the value of the cluster label $\mathbf{c}_i^p$ after any iteration, we sum together all incoming messages to $\mathbf{c}_i^p$ and take the value $\hat{\mathbf{c}}_i^p$ that maximizes the incoming messages, i.e.,

$$\hat{\mathbf{c}}_i^p = \arg\max_j \left[S^p(i,j) + \sum_{q=1,q\neq p}^{m}\eta_{ii}^{pq}(j) + \sum_{k=1}^{n}\alpha_{ik}^{p}(j)\right]. \tag{14}$$

Although the within-view messages and cross-view messages defined above ideally works well, the hard exemplar consistency constraint $\delta_k^p(\mathbf{c}^p)$ would make an optimization almost impossible (i.e., running quickly into local maxima). To this end, we need to simplify these messages as well as the cluster assignment vector computations, so as to derive simple yet functionally equivalent computations, as shown in the next section.

3.3 The Algorithm

To obtain the simplified messages, some mathematical techniques used in [22], [23], [24] are applied. The basic idea is as follows. First, the exemplar consistency constraint embedded in (10) is expanded. Then, each variable message is taken as the sum of a variable and a constant. By dexterously setting the constants and applying variable substitution, the computations are finally simplified to be simple yet functionally equivalent. The detailed derivation can be found in the supplementary material. The simplified messages are as follows.

1. Messages passing within the pth view

$$\tilde{\alpha}_{ik}^{p}(k) = \begin{cases}\sum_{i':i'\neq k}\max(0,\tilde{\beta}_{i'k}^{p}(k)) & k = i \\ \min\left[0,\tilde{\beta}_{kk}^{p}(k) + \sum_{i':i'\notin\{i,k\}}\max(0,\tilde{\beta}_{i'k}^{p}(k))\right] & k\neq i\end{cases} \tag{15}$$

$$\begin{aligned}\tilde{\beta}_{ik}^{p}(k) &= S^p(i,k) + \sum_{q=1,q\neq p}^{m}\tilde{\eta}_{ii}^{pq}(k) \\ &\quad - \max_{j:j\neq k}\left(S^p(i,j) + \tilde{\alpha}_{ij}^{p}(j) + \sum_{q=1,q\neq p}^{m}\tilde{\eta}_{ii}^{pq}(j)\right).\end{aligned} \tag{16}$$

2. Messages passing across any two views: from the qth view to the pth view

$$\begin{aligned}\tilde{\eta}_{ii}^{pq}(k) = \max_{k'}\Big[&L^{pq}(k,k') + S^q(i,k') + \tilde{\alpha}_{ik'}^{q}(k') \\ &+ \sum_{p'=1,p'\neq p,p'\neq q}^{m}\tilde{\eta}_{ii}^{qp'}(k')\Big].\end{aligned} \tag{17}$$

The message θ is eliminated.

The messages $\tilde{\alpha}_{ik}^{p}(k)$, $\tilde{\beta}_{ik}^{p}(k)$ and $\tilde{\eta}_{ii}^{pq}(k)$ are initialized as zero, which are then iteratively updated via equations (15), (16) and (17) respectively. The message-updating procedure may be terminated after the local decisions stay constant for some number of iterations t_{conv}, or after a maximum number of iterations t_{max}. In our experiments, t_{conv} is set to 100 and t_{max} is set to 1,000. After the termination of the message updating procedure, the optimal cluster assignment vectors $\{\hat{\mathbf{c}}^1,\hat{\mathbf{c}}^2,\ldots,\hat{\mathbf{c}}^m\}$ are computed as follows,

$$\hat{\mathbf{c}}_i^p = \arg\max_j\left[S^p(i,j) + \sum_{q=1,q\neq p}^{m}\tilde{\eta}_{ii}^{pq}(j) + \tilde{\alpha}_{ij}^{p}(j)\right]. \tag{18}$$

Note that the above formula outputs one cluster assignment vector for each view of the data, which well captures the cluster structure of that view. If it is a priori known that some view is better than the others, then the cluster assignment vector corresponding to that view is taken as the final clustering result of the data. However, this priori information is often unavailable in the unsupervised setting. To this end, we suggest to produce one common cluster assignment vector $\hat{\mathbf{c}} = [\hat{\mathbf{c}}_1,\hat{\mathbf{c}}_2,\ldots,\hat{\mathbf{c}}_n]$ as the final clustering result, in which each value $\hat{\mathbf{c}}_i$ maximizes the sum of all incoming messages associated with that object, i.e.,

$$\hat{\mathbf{c}}_i = \arg\max_j\sum_{p=1}^{m}\left[S^p(i,j) + \sum_{q=1,q\neq p}^{m}\tilde{\eta}_{ii}^{pq}(j) + \tilde{\alpha}_{ij}^{p}(j)\right]. \tag{19}$$

Like previous belief propagation based algorithms [22], [24], the main cause of failure mode of the proposed MVAP algorithm is that the objective function (7) has multiple optimum solutions with corresponding multiple fixed points of the update rules, which may prevent convergence. In this case, the message update may oscillate, with data points alternating between being exemplars and non-exemplars across different views. Even in the case of a single global optima, we may observe no convergence due to approximation introduced by belief propagation. One remedy is to introduce a damping to the message updating procedure, which could always avoid oscillations. Let μ denote any of the three messages on the lefthand side of equations (15) to (17), the damping is done as follows [36]:

$$\mu = \lambda\mu^{\text{old}} + (1-\lambda)\mu^{\text{new}}, \tag{20}$$

where μ^{old} is the message before update, and μ^{new} is the message after update. The higher values of the damping factor λ unsurprisingly lead to slower convergence rates but often more stable maximization (i.e., avoiding oscillations).

According to experiments, setting the damping factor λ to 0.9 is sufficient in most cases to ensure convergence [24], [36].

The proposed MVAP algorithm can be taken as a generalization of the single-view AP algorithm. This is because when the number of views reduces to one, the objective function (7) is equal to the objective function of AP as discussed before, and there is no cross-view message and the within-view messages become exactly the same as those of AP [22]. Setting the trade-off parameter σ to 1 can be analyzed similarly. Additionally, by comparing with AP, the computational complexity of the proposed MVAP algorithm becomes clear. That is, in each iteration, there are m copies of within-view messages and $m(m-1)$ copies of cross-view messages. The computational complexity of each within-view message of MVAP is the same as the message of AP. And the computational complexity of each cross-view message is only $\frac{1}{n}$ that of each within-view message. Because the number of views, i.e., m, is often far smaller than the number of objects, i.e., n, the computational complexity of MVAP is mainly determined by the within-view messages. Therefore, the time complexity of MVAP is m times of the time complexity of AP. As analyzed in [36], the computational complexity of the message passing in AP is $O(n^2)$. Therefore, the time complexity of MVAP is $O(mn^2)$. Similarly, the storage requirement of MVAP can be also analyzed. From the perspective of time and space complexities, the proposed MVAP algorithm can be viewed as a generalization of AP.

4 Experimental Evaluation

In this section, to evaluate the effectiveness of the proposed MVAP approach, extensive experiments are conducted on five real-world multi-view datasets. The datasets include one Visible Light/Near-IR face dataset, two link-based document datasets, one multilingual document dataset and one handwritten numeral multi-feature dataset. First, we analyze the effect of the trade-off parameter σ on the clustering performance, which reveals the importance of balancing the clustering quality within individual view and the clustering consistency across different views. Then the sensitivity to the nearest neighbor parameter k will be analyzed. Finally, comparison experiments are conducted to compare the proposed MVAP algorithm with not only conventional single-view clustering algorithms but also state-of-the-art multi-view clustering algorithms. Comparison results confirm that the proposed approach achieves significant performance improvements over the compared clustering algorithms. We also show that directly combining multiple views by means of feature concatenation or similarity matrix addition would not help improve the clustering performance. Moreover, experimental results on the two datasets consisting of more than two views reveal that the proposed MVAP approach is superior to the existing multi-view clustering algorithms in clustering more than two views.

All the experiments are implemented in Matlab7.8.0.347 (R2009a) 64-bit edition on a workstation (Windows 64 bit, eight Intel 2.00 GHz processors, 16 GB of RAM).

4.1 Testing Datasets and Evaluation Metric

In this section, we will introduce the five testing datasets used in the experiments and the evaluation metric.

4.1.1 Face Image Dataset

For face image clustering, the aforementioned VIS/NIR face dataset [6] is used. For each data view, the similarity matrix is computed as follows.

The SIFT features [39] are utilized, which are shown to be effective for image matching and category learning [24], [40], [41], [42], [43]. Each feature is represented by a 128-dimensional vector. A procedure suggested by Lowe is used to count the number of significant feature matches comparing image i with image k (denoted as $\mathcal{M}_{ik}$): for each local feature from image i, the nearest and second nearest features are sought in image k. The match is considered significant, if the distance ratio between the nearest and second-nearest neighbors is greater than a threshold ζ [39]. In our experiments, after trying a series of values ranging from 0.1 to 0.9 with step 0.05, setting $\zeta = 0.6$ generates the best clustering results for all the compared methods as well as the proposed method. Then the similarity matrix $[S(i,j)]_{N\times N}$ is defined to be the number of significant feature matches normalized by subtracting means across both dimensions [40]:

$$\tilde{S}(i,j) = \mathcal{M}_{ij} - \frac{1}{N}\sum_{k=1}^{N}\mathcal{M}_{ik} - \frac{1}{N}\sum_{k=1}^{N}\mathcal{M}_{kj}. \tag{21}$$

The similarity of data distributions across the two views is $\frac{1}{n}\sum_{i=1}^{n}\tilde{L}^{pq}(i,i) = 0.1851$.

4.1.2 Link-Based Document Datasets

Two link-based document datasets are used in our experiments, which are Cora and CiteSeer downloaded from http://www.cs.umd.edu/projects/linqs/projects/lbc/index.html. Each of these two datasets consists of two data views, namely, the document contents (abbr. *Cont*) and document citations (abbr. *Cite*).

The Cora dataset consists of 2,708 machine learning papers classified into one of seven classes, namely, Case_Based, Theory, Genetic_Algorithms, Neural_Networks, Probabilistic_Methods, Reinforcement_Learning and Rule_Learning. These 2,708 papers are linked via 5,429 citations, which are taken as one type of data view. Each publication on the dataset is described by a $0/1$-valued word vector indicating the absence/presence of the corresponding word from the dictionary of size 1,433, which is taken as another type of data view.

The CiteSeer dataset consists of 3,312 scientific publications classified into one of six classes, including Agents, AI, DB, IR, ML and HCI. These 3,312 publications are linked via 4,732 citations, which are taken as one type of data view. Each publication on the dataset is described by a $0/1$-valued word vector indicating the absence/presence of the corresponding word from the dictionary of size 3,703, which is taken as another type of data view.

The similarity matrix of the *Cont* view is computed via the cosine similarity of their $0/1$-valued word vector. On the other hand, the similarity matrix of *Cite* view is computed directly via the citation graph: the similarity of two documents is set to 1 if they are linked via citation; otherwise their similarity is set to 0.

On these two datasets, the similarities of data distributions across the two views (i.e., $\frac{1}{n}\sum_{i=1}^{n}\tilde{L}^{pq}(i,i)$) are 0.127 and 0.133 respectively, which are relatively small.

4.1.3 Reuters Multilingual Document Dataset

The Reuters multilingual document dataset is a collection of multilingual documents belonging to six large Reuters categories that are available from http://multilingreuters.iit.nrc.ca/ReutersMultiLingualMultiView.htm [15]. This dataset contains documents that are originally written in five different languages, namely English, French, German, Italian and Spanish. In order to produce multilingual versions of each document, each original document is translated into the other four languages using the Portage system [44]. After preprocessing, these documents are represented as a "bag of words" using a TFIDF-based weighting scheme. In our experiments, we use the RCV1 subset, namely, the documents original written in English as one view and the translated versions in French and German as another two views. A random subsample of 1,200 documents from this collection is used, with each category containing 200 documents. The similarity matrix of each view is computed via cosine similarity of its TFIDF representation.

The similarities of data distributions across English-French, English-German and French-German pair views are $\frac{1}{n}\sum_{i=1}^{n}\tilde{L}^{pq}(i,i) = 0.387, 0.385, 0.389$ respectively.

4.1.4 Handwritten Numeral Dataset

The multiple features (abbr. Mfeat) dataset is a handwritten numeral dataset downloaded from UCI machine learning repository [45]. This dataset consists of numeral 2,000 objects belonging to 10 categories, i.e., $'0'-'9'$, with 200 objects per category. Each object is represented by multiple feature sets. In our experiments, three feature sets are used, which are fou, pix and fac. The fou feature set contains 76 Fourier coefficients of the character shapes; the pix feature set contains 240 pixel averages in 2×3 windows; and the mfeat-fac feature set contains 216 profile correlations. The similarity matrix of each view is computed via the exponential of its minus normalized Euclidean distance, i.e., $\tilde{S}(i,j) = \exp(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|}{\max_{h,l}\|\mathbf{x}_h - \mathbf{x}_l\|})$.

The similarities of data distributions across fou-pix pair views, fou-fac pair views and pix-fac pair views are $\frac{1}{n}\sum_{i=1}^{n}\tilde{L}^{pq}(i,i) = 0.163, 0.133, 0.427$ respectively.

4.1.5 Evaluation Metric

Since on each testing dataset, ground-truth labeling is provided for evaluating the clustering accuracy, a widely adopted external evaluation metric, namely, normalized mutual information (NMI) [37], is used for measuring the clustering accuracy based on the underlying class labels. Although there exist many external clustering evaluation measurements, such as purity, entropy-based measures [46], clustering errors, and pair counting based indices [47], the mutual information provides a sound indication of the shared information between a pair of clusterings [24], [37], [48].

Given a dataset $\mathcal{X}$ of size n, the clustering labels π of c clusters and actual class labels ζ of $\hat{c}$ classes, a confusion matrix is formed first, where entry (i,j), $n_i^{(j)}$ gives the number of points in cluster i and class j. Then NMI can be computed from the confusion matrix [37]

$$NMI(\zeta,\pi) = \frac{2\sum_{l=1}^{c}\sum_{h=1}^{\hat{c}}\frac{n_l^{(h)}}{n}\log\frac{n_l^{(h)}n}{\sum_{i=1}^{c}n_i^{(h)}\sum_{i=1}^{\hat{c}}n_l^{(i)}}}{H(\pi)+H(\zeta)}, \quad (22)$$

where $H(\pi) = -\sum_{i=1}^{c}\frac{n_i}{n}\log\frac{n_i}{n}$ and $H(\zeta) = -\sum_{j=1}^{\hat{c}}\frac{n^{(j)}}{n}\log\frac{n^{(j)}}{n}$ are the Shannon entropy of cluster labels π and class labels ζ respectively, with n_i and $n^{(j)}$ denoting the number of points in cluster i and class j. A high NMI indicates the clustering and class labels match well.

Apart from NMI, in the comparison results, we also use purity as the evaluation metric. To compute the purity of a clustering result w.r.t. the ground-truth classes, each cluster is first assigned to the class which is the most frequent in the cluster. And then the purity is computed by counting the number of correctly assigned objects and dividing by the data set size n,

$$Purity(\Omega,\mathcal{C}) = \frac{1}{n}\sum_{k}\max_{j}|\Omega_k \cap \mathcal{C}_j|, \quad (23)$$

where $\Omega = \{\Omega_1, \Omega_2, \ldots, \Omega_k\}$ is the set of clusters and $\mathcal{C} = \{\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_j\}$ is the set of classes. Like the NMI evaluation metric, a higher purity indicates the better clustering.

4.2 Parameter Analysis

In this section, we will first analyze the effect of the trade-off parameter σ on the clustering performance. Then the sensitivity to the nearest neighbor parameter k will be analyzed.

First of all, we run the proposed MVAP algorithm on the five multi-view datasets with different σ, and analyze the clustering quality and the clustering consistency across different views on each dataset. The results are shown in Figs. 5 and 6 respectively. From Fig. 5, we can observe that, when σ is set between 0.4 and 0.6, which emphasizes a relative balance between the within-view clustering quality and the cross-view clustering consistency, the MVAP algorithm achieves the best clustering performances in all views. Especially, setting $\sigma = 0.5$ results in the highest NMI values. On the other hand, setting $\sigma = 1$, which ignores the consistency terms between views, leads to significant accuracy loss. In particular, on the Reuters dataset, the accuracy loss is as large as 25 percent. From these results, it is safe to draw the conclusion that it is very important to balance the within-view clustering quality with the cross-view clustering consistency. Therefore, in the following comparison experiments, except stated otherwise, the trade-off parameter σ is set to 0.5.

Fig. 6 shows the consistency of the clustering results across different views when different trade-off parameter σ is used. That is, y-axis reports the NMI value computed from the clustering labels of two different views (i.e., in (22), the clustering labels of one view are taken as ζ and the clustering labels of the other view are taken as π). From this figure, we can see that, as the trade-off parameter σ decreases, the cross-view clustering consistency increases. Especially, when σ is set to 0.1, the clustering results across

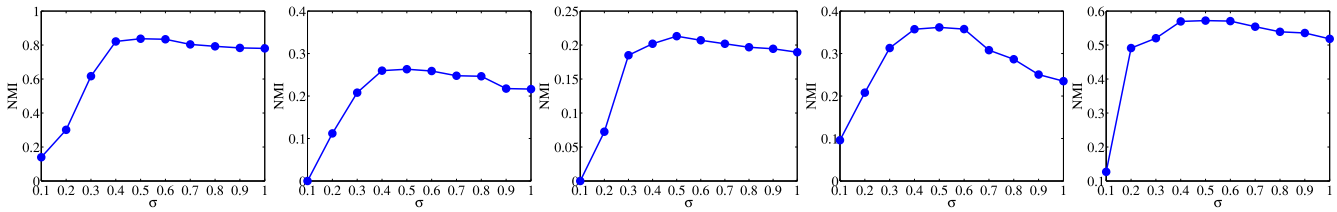


Fig. 5. Analysis on the trade-off parameter σ. On each dataset, the clustering quality, i.e., the NMI value of the common cluster assignment vector, is plotted as a function of the trade-off parameter σ.

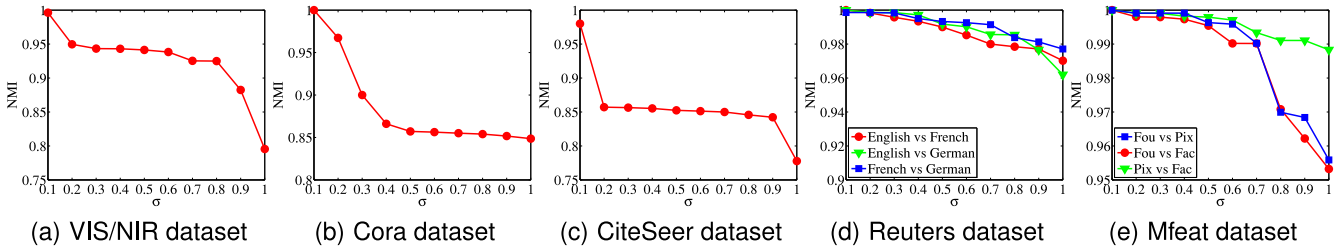


Fig. 6. Analysis on the trade-off parameter σ. On each dataset, the consistency of the clustering results across different views in terms of NMI is plotted as a function of the trade-off parameter σ.

different views become almost the same, that is, the NMI value between clustering assignments of different views is as large as 1. By observing Fig. 5, such low σ value leads to the lowest clustering performance. This is because, when setting $\sigma = 0.1$, the maximization of the proposed multi-view clustering objective function, i.e., (6), would almost ignore the within-view clustering quality but enforce $(\mathbf{c}_i^p, \mathbf{c}_i^q) = \arg\max \tilde{L}^{pq}, \forall i = 1, \ldots, n$. It means that all the data objects are assigned to almost the same exemplar across all views, i.e. $\mathbf{c}_i^p = \mathbf{c}_i^q$, regardless of the within-view clustering quality. Especially, on the Cora and CiteSeer datasets, all the data objects are assigned to exactly the same exemplar across all views, leading to a trivial partitioning (i.e., only one cluster in each view). According to the definition of the NMI metric, this leads to $NMI = 0$, as shown in Figs. 5b and 5c respectively.

On the other hand, when σ is set to 1, which implies that MVAP degenerates into applying AP in individual views, Fig. 6 shows that on most datasets, the clustering assignments from individual views still preserve high consistency. That is, the NMI value computed from the clustering assignments of disparate views is still as high as 0.8. In particular, on the Reuters and Mfeat datasets, this NMI value reaches 0.95. This phenomenon validates the fact that the different views often admit the same underlying cluster structure of the data. However, together with Fig. 5, the improved clustering quality obtained by using balanced trade-off parameter σ confirms the necessity of combining information from multiple views by means of trading-off the within-view clustering quality and the cross-view clustering consistency.

Another important result we can observe from Fig. 6 is that the trend of clustering consistency across different views coincides with the similarity of data distributions across the corresponding views. This fact confirms the effectiveness of the data consistency defined in (4). In particular, Fig. 6e shows that the clustering consistency across pix-fac views is larger than that across fou-pix views and fou-fac views, which coincides with their similarities of data distributions. Additionally, on some datasets like VIS/NIR, Cora and CiteSeer, although the similarities of data distributions across two views are relatively small (less than 0.2), the proposed MVAP algorithm can generate relative consistent clustering results across the corresponding views. The above parameter analysis reveals the importance of balancing the clustering quality within individual view and the clustering consistency across different views.

We also analyze the sensitivity to the nearest neighbor parameter k, as shown in Fig. 7. From the figure, we can see that, for all the datasets, setting k in interval $[25, 35]$ would generate satisfactory clustering results on all datasets, with $k = 30$ being the best, i.e., high quality clustering results that are consistent across different views. This analysis has proved the underlying rationale of the selection of the nearest neighbor parameter k to be 30.

4.3 Comparison Results

Two types of comparison methodologies are used. The first type is to compare the proposed MVAP approach with the conventional clustering algorithms designed for single-view clustering. Apart from performing the single-view clustering algorithms in individual views, we also perform these single-view clustering algorithms in combined views. The multiple views are combined by either directly concatenating features from different views if the features can be concatenated; otherwise they are combined by directly adding their similarity matrices. The second type of comparison methodology is to compare the proposed MVAP approach with six state-of-the-art multi-view clustering algorithms.

For the traditional single-view clustering algorithms, two representative approaches are selected as follows.

1) Affinity propagation [22], which is an exemplar-based clustering approach. The Matlab code is obtained from the authors' website.[2]

2. http://www.psi.toronto.edu/index.php?q=affinity%20propagation.

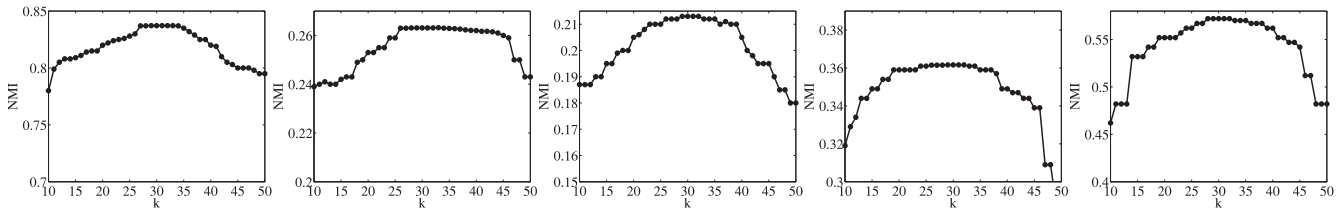


Fig. 7. Analysis on the parameter k. On each dataset, the clustering quality, i.e., the NMI value of the common cluster assignment vector, is plotted as a function of the parameter k.

2) Spectral clustering based on normalized cut (Ncut) [49], the Matlab code of which is obtained from the authors' website.[3]

The two single-view clustering algorithms are performed both in individual views and in combined views such as concatenated features or added similarity matrixes.

For the multi-view clustering algorithms, six recent algorithms are compared, which are listed as follows.

1) Co-trained multi-view spectral clustering (abbr. CoTrainSpectral) [18], which uses co-training to search eigenvectors that agree across different views and then apply k-means to the resulting eigenvectors to generate final clustering results. The Matlab code is obtained from the authors' website.[4]
2) Co-regularized multi-view spectral clustering (abbr. CoRegSpectral) [20], which uses co-regularization to make eigenvectors agree with each other across different views and then apply k-means to the resulting eigenvectors to generate final clustering results. The pairwise-based regularization is used. As suggested by the authors, the parameter λ is selected from the range $[0, 0.1]$ that generates the best clustering results. The Matlab code is obtained from the authors' website.[5]
3) Constrained multi-view spectral clustering (abbr. CSP-P) [29], which takes one view as the similarity matrix but encodes the other view as the constraints. It is limited to only two views. The Matlab code is obtained from the authors' website.[6]
4) Multi-view clustering via Canonical Correlation Analysis (abbr. CCA) and kernel CCA [31], which is a subspace multi-view clustering method based on (kernel) Canonical Correlation Analysis. We use LSCCA package[7] implementation of CCA and kernel CCA to extract latent representations, and then perform k-means.
5) MC-FR-OI [34], which is a tensor based multi-view clustering algorithm by integration of the Frobenius-norm objective function (MC-FR-OI), where higher order orthogonal iteration (HOOI) is used for optimization.

It should be noted that, since the MVAP approach is an exemplar-based clustering algorithm and most of the compared multi-view clustering approaches (namely, CoTrainSpectral, CoRegSpectral and CSP-P) are extensions of spectral clustering, therefore, the compared single-view clustering approaches are also selected from representative exemplar-based clustering algorithms (i.e., AP) and spectral clustering algorithms (i.e., Ncut). In the AP and MVAP algorithms, following the widely used strategy [22], [24], [36], the exemplar preferences are adjusted so as to generate the required number of clusters, and in the other compared methods, the number of clusters is given as input. That is, in all the clustering methods, the number of generated clusters is the same as the ground-truth. For comparison, we first report the single-view clustering performance of AP and Ncut on the five testing datasets. That is, Table 1 lists the mean and standard deviation of NMI and Purity over 100 runs generated by the two single-view clustering methods in individual views. These results will later be compared with the results generated by the two single-view clustering methods in directly combined views.

4.3.1 Face Image Dataset

Table 2 reports the performance of multi-view clustering on the VIS/NIR dataset, i.e., mean and standard deviation of NMI and Purity over 100 runs. For each single-view clustering algorithm (i.e., AP and Ncut), the corresponding column reports its clustering results generated in combined view via direct similarity matrix addition.

By comparing Table 2 with the rows associated with the VIS/NIR dataset in Table 1, it can be observed that, by directly adding similarity matrices from two views, the clustering performances of the two single-view clustering algorithms in the combined view degenerate seriously. That is, the clustering results generated by AP and Ncut in the added similarity matrix are worse than the worst clustering results generated separately in individual views. For instance, the AP algorithm obtains NMI and Purity in the NIR view as high as 0.747 and 0.760 respectively. However, in the directly combined view, the AP algorithm only obtains NMI and Purity as low as 0.653 and 0.677. This comparison result implies that improper combination of multiple views would not help improve but may even degenerate the clustering performance. However, when combining the two views properly, the clustering results generated jointly from multiple views are better than those generated in individual views, as shown by the results of the seven multi-view clustering algorithms. Among the seven multi-view clustering algorithms, the proposed MVAP algorithm generates the best clustering results, achieving on average 5 percent improvement over state-of-the-art multi-view clustering algorithms.

3. http://www.cis.upenn.edu/~jshi/software/Ncut_9.zip.
4. http://www.umiacs.umd.edu/~abhishek/code_cospectral.zip.
5. http://www.umiacs.umd.edu/~abhishek/code_coregspectral.zip.
6. http://bayou.cs.ucdavis.edu/papers/cikm12.demo.zip.
7. http://www.public.asu.edu/~jye02/Software/CCA/index.html.

TABLE 1
Single-View Clustering Performance in Terms of NMI and Purity on the Five Testing Datasets Over 100 Runs: Numbers in Parentheses are the Standard Deviations

Methods			AP	Ncut
VIS/NIR	VIS	NMI	0.662(0.000)	0.669(0.050)
		Purity	0.688(0.000)	0.695(0.050)
	NIR	NMI	0.747(0.000)	0.750(0.031)
		Purity	0.760(0.000)	0.766(0.029)
Cora	Cont	NMI	0.147(0.000)	0.150(0.013)
		Purity	0.153(0.000)	0.169(0.012)
	Cite	NMI	0.209(0.000)	0.208(0.017)
		Purity	0.230(0.000)	0.230(0.016)
CiteSeer	Cont	NMI	0.129(0.000)	0.132(0.023)
		Purity	0.134(0.000)	0.140(0.018)
	Cite	NMI	0.183(0.000)	0.185(0.033)
		Purity	0.199(0.000)	0.202(0.030)
Reuters	English	NMI	0.235(0.000)	0.247(0.014)
		Purity	0.249(0.000)	0.262(0.015)
	French	NMI	0.221(0.000)	0.244(0.013)
		Purity	0.241(0.000)	0.260(0.012)
	German	NMI	0.224(0.000)	0.245(0.019)
		Purity	0.243(0.000)	0.262(0.020)
Mfeat	Fou	NMI	0.465(0.000)	0.473(0.010)
		Purity	0.480(0.000)	0.491(0.009)
	Pix	NMI	0.510(0.000)	0.515(0.017)
		Purity	0.532(0.000)	0.539(0.015)
	Fac	NMI	0.513(0.000)	0.519(0.014)
		Purity	0.538(0.000)	0.546(0.013)

For this face clustering task, another state-of-the-art tensor based method, termed robust tensor clustering [35], is also compared. The RTC algorithm is robust to the noises in facial images, which first obtains a lower-rank approximation of the original tensor data (here two-view data) and then calculates high-order SVD of the approximate tensor to generate final clustering of the facial images. Experimental results show that RTC can get the average value of NMI and Purity as high as 0.830 and 0.842 respectively, a quite competitive result compared with the state-of-the-art multi-view clustering methods. This comparison confirms the effectiveness of the multi-view clustering models, in particular MVAP, in discovering latent cluster structure from multi-view data.

4.3.2 Link-Based Document Datasets

Table 3 reports the clustering performances on the Cora and CiteSeer datasets, i.e., mean and standard deviation of NMI and Purity over 100 runs. Note that because the CCA and kernel CCA algorithms are not applicable for the two link-based document datasets, the two methods are not performed on these two datasets. For the single-view clustering approaches, the combination of the two views are realized by directly adding the similarity matrices of the two views. Similar to the case of the VIS/NIR dataset, such direct combination of multiple views degenerates the clustering performance. According to the clustering results by AP and Ncut, although the average NMI and Purity values generated in the combined views are higher than those generated in the *Cont* view, they are not so good as those generated in the *Cite* view. This comparison again demonstrates that improper combination of multiple views would not help improve the clustering performance.

However, the multi-view clustering algorithms have achieved some clustering performance improvements over the single-view clustering algorithms. The similarities of data distributions across two views on these two datasets are relatively small (i.e., 0.127 and 0.133 respectively), but the multi-view clustering algorithms can generate far better results than applying single-view clustering methods in individual views. This result as well as the results on the other similar datasets show that the multi-view clustering methods are suitable for the case where the different views agree with each other on the clustering structure but their data distributions are not similar.

Compared with the state-of-the-art multi-view clustering algorithms, the proposed MVAP algorithm also generates better clustering results. For instance, on the Cora dataset, MVAP outperforms the four multi-view clustering algorithms by achieving about 8 percent performance improvement.

4.3.3 Reuters Multilingual Document Dataset

Since the Reuters dataset consists of three views, the experimental results are analyzed from two perspectives, namely

TABLE 2
Multi-View Clustering Performance in Terms of NMI and Purity on the VIS/NIR Dataset Over 100 Runs: Numbers in Parentheses are the Standard Deviations

Methods	AP	Ncut	CoTrainSpectral	CoRegSpectral	CSP-P	CCA	Kernel CCA	MC-FR-OI	MVAP
NMI	0.653(0.000)	0.663(0.038)	0.791(0.022)	0.783(0.012)	0.765(0.030)	0.790(0.009)	0.792(0.010)	0.805(0.008)	**0.837**(0.000)
Purity	0.677(0.000)	0.689(0.034)	0.822(0.029)	0.818(0.013)	0.771(0.025)	0.820(0.011)	0.823(0.011)	0.827(0.008)	**0.845**(0.000)

TABLE 3
Multi-View Clustering Performance in Terms of NMI and Purity on the Two Link-Based Document Datasets Over 100 Runs: Numbers in Parentheses are the Standard Deviations

Methods		AP	Ncut	CoTrainSpectral	CoRegSpectral	CSP-P	MC-FR-OI	MVAP
Cora	NMI	0.194(0.000)	0.199(0.019)	0.244(0.016)	0.243(0.021)	0.242(0.031)	0.245(0.019)	**0.263**(0.000)
	Purity	0.213(0.000)	0.220(0.015)	0.256(0.013)	0.253(0.019)	0.251(0.025)	0.259(0.020)	**0.271**(0.000)
CiteSeer	NMI	0.180(0.000)	0.179(0.020)	0.205(0.020)	0.206(0.020)	0.200(0.027)	0.209(0.021)	**0.213**(0.000)
	Purity	0.195(0.000)	0.193(0.017)	0.240(0.019)	0.241(0.019)	0.230(0.026)	0.244(0.023)	**0.259**(0.000)

TABLE 4
Multi-View Clustering Performance in Terms of NMI and Purity on the Reuters Dataset Over 100 Runs: Numbers in Parentheses are the Standard Deviations

Methods		AP	Ncut	CoTrainSpectral	CoRegSpectral	CSP-P	CCA	Kernel CCA	MC-FR-OI	MVAP
E-F Pair	NMI	0.237(0.000)	0.249(0.014)	0.299(0.015)	0.300(0.013)	0.295(0.012)	0.298(0.013)	0.300(0.012)	0.305(0.008)	**0.325**(0.000)
	Purity	0.252(0.000)	0.263(0.011)	0.325(0.014)	0.325(0.012)	0.320(0.010)	0.322(0.012)	0.325(0.011)	0.334(0.009)	**0.359**(0.000)
E-G Pair	NMI	0.236(0.000)	0.250(0.013)	0.300(0.017)	0.302(0.016)	0.299(0.016)	0.300(0.015)	0.301(0.014)	0.307(0.007)	**0.330**(0.000)
	Purity	0.250(0.000)	0.265(0.012)	0.326(0.015)	0.328(0.013)	0.323(0.011)	0.326(0.013)	0.328(0.010)	0.338(0.008)	**0.371**(0.000)
E-G Pair	NMI	0.229(0.000)	0.248(0.010)	0.299(0.014)	0.297(0.013)	0.287(0.011)	0.299(0.015)	0.305(0.014)	0.306(0.009)	**0.322**(0.000)
	Purity	0.247(0.000)	0.263(0.010)	0.325(0.013)	0.323(0.010)	0.310(0.012)	0.325(0.014)	0.336(0.013)	0.339(0.008)	**0.355**(0.000)
E-F-G	NMI	0.244(0.000)	0.251(0.017)	0.311(0.013)	0.310(0.010)	NA	0.311(0.012)	0.318(0.011)	0.320(0.010)	**0.361**(0.000)
	Purity	0.260(0.000)	0.267(0.009)	0.332(0.011)	0.330(0.010)	NA	0.332(0.011)	0.352(0.011)	0.354(0.010)	**0.379**(0.000)

E, F and G stand for English, French and German, respectively.

pairwise views and triple views, which are reported in Table 4. In the case of pairwise views, we try the combination of any two out of three views, which leads to the English-French pair, the English-German pair and the French-German pair respectively. And for each pair, a dataset is formed that consists of the two views. The clustering results are reported for such double-view dataset. In the case of triple views, the clustering results are reported for such triple-view dataset.

From the table, in the cases of both pairwise and triple views, by directly adding similarity matrices of multiple views, the single-view clustering algorithms (i.e., AP and Ncut) make a slight performance improvement over separately conducting clustering in individual views (by comparing with Table 1). This is unlike the previous two tables, where direct similarity matrix addition leads to degenerate performance for single-view clustering algorithms. The main reason may be that, on Reuters, the multiple views are generated by translation via statistical machine translation system, which are more consistent than the multiple views of the previous VIS/NIR dataset and link-based document datasets. However, compared with the performance improvement achieved by conducting multi-view clustering algorithms in multi-views, such improvement is marginal. This again confirms the advantage of properly combining different views for clustering multiple views.

Among the seven multi-view clustering algorithms, the proposed MVAP algorithm significantly outperforms the state-of-the-art multi-view clustering algorithms. For instance, in the case of the English-German pair, MVAP achieves a 7.4 percent improvement over the second best multi-view clustering algorithm, namely MC-FR-OI; while in the case of the triple views, MVAP achieves a much larger performance improvement, as large as 12.8 percent, which is very significant in such a challenging multilingual document dataset.

Another significant phenomenon can be observed that, the proposed MVAP algorithm has achieved a much larger performance gain when comparing the performance improvements in the triple views with the performance improvements in the pairwise views. For instance, by performing MVAP in the triple views, it obtains NMI as large as 0.361, which is at least 0.031 larger than by performing MVAP in the pairwise views. On the other hand, for the existing multi-view clustering algorithms such as CoTrainSpectral and CoRegSpectral, only at most 0.010 additional gain is obtained, which is far smaller than 0.031. In the case of both pairwise and triple views, the experimental results show that the proposed MVAP algorithm significantly outperforms the existing multi-view clustering approaches on the Reuters dataset.

4.3.4 Handwritten Numeral Dataset

Similar to the case on Reuters, the experimental results on the Mfeat dataset are analyzed from two perspectives, namely pairwise views and triple views, which are reported in Table 5. In the cases of both pairwise and triple views, by directly concatenating features of two or three views, the single-view clustering algorithms (i.e., AP and Ncut) make a slight performance improvement over conducting clustering in individual views (by comparing with Table 1). However, this clustering performance improvement is far less than that achieved by performing multi-view clustering algorithms in multiple views.

TABLE 5
Multi-View Clustering Performance in Terms of NMI (*abbr.* N) and Purity (*abbr.* P) on the Mfeat Dataset Over 100 Runs: Numbers in Parentheses are the Standard Deviations

Methods		AP	Ncut	CoTrainSpectral	CoRegSpectral	CSP-P	CCA	Kernel CCA	MC-FR-OI	MVAP
Fou-Pix Pair	N	0.512(0.000)	0.516(0.018)	0.535(0.010)	0.536(0.013)	0.530(0.029)	0.534(0.010)	0.535(0.013)	0.541(0.009)	**0.560**(0.000)
	P	0.537(0.000)	0.541(0.013)	0.563(0.009)	0.564(0.012)	0.557(0.023)	0.560(0.009)	0.562(0.012)	0.577(0.008)	**0.592**(0.000)
Fou-Fac Pair	N	0.515(0.000)	0.522(0.021)	0.534(0.018)	0.533(0.019)	0.531(0.016)	0.533(0.016)	0.534(0.017)	0.540(0.005)	**0.547**(0.000)
	P	0.539(0.000)	0.554(0.019)	0.562(0.017)	0.560(0.018)	0.558(0.015)	0.560(0.017)	0.562(0.017)	0.566(0.008)	**0.587**(0.000)
Pix-Fac Pair	N	0.517(0.000)	0.523(0.014)	0.541(0.012)	0.541(0.016)	0.535(0.025)	0.541(0.013)	0.541(0.015)	0.544(0.009)	**0.564**(0.000)
	P	0.543(0.000)	0.556(0.012)	0.570(0.010)	0.570(0.013)	0.562(0.024)	0.570(0.014)	0.570(0.016)	0.578(0.007)	**0.598**(0.000)
Fou-Pix-Fac	N	0.522(0.000)	0.526(0.020)	0.544(0.017)	0.544(0.016)	NA	0.543(0.017)	0.544(0.018)	0.550(0.010)	**0.572**(0.000)
	P	0.550(0.000)	0.563(0.019)	0.575(0.016)	0.575(0.013)	NA	0.573(0.018)	0.576(0.018)	0.582(0.011)	**0.609**(0.000)

TABLE 6
Comparison Results of the Average Time (in Seconds) Over 100 Runs on the Five Testing Datasets

Methods	AP	Ncut	CoTrainSpectral	CoRegSpectral	CSP-P	CCA	Kernel CCA	MC-FR-OI	MVAP
VIS/NIR	1.325	2.076	3.681	3.453	2.217	4.731	5.563	3.132	2.958
Cora	3.683	5.634	10.094	9.639	6.161	NA	NA	9.305	7.550
CiteSeer	4.377	7.027	12.491	11.629	7.484	NA	NA	9.582	8.975
Reuters	1.610	2.422	5.493	4.981	NA	5.307	6.631	4.970	5.182
Mfeat	2.759	4.296	7.478	7.234	NA	8.625	10.572	6.257	7.517

When comparing MVAP with the state-of-the-art multi-view clustering algorithms, MVAP outperforms its counterpart algorithms in the cases of both pairwise views and triple views. For instance, in Fou-Pix pair, MVAP generates the NMI value as large as 0.560, which is 0.019 larger than that generated by the second best MC-FR-OI; while in the triple views, MVAP generates the NMI value as large as 0.572, which is 0.022 larger than those generated by the second best MC-FR-OI. The comparison results on this dataset and the previous datasets show that explicitly enforcing clustering consistency across different views in MVAP outperforms the implicit way in CoTrainSpectral and CoRegSpectral.

Similar to the case on the Reuters dataset, when comparing the performance improvements in the triple views with the performance improvements in the pairwise views, the proposed MVAP algorithm also obtains an advantage over CoTrainSpectral, CoRegSpectral, CCA, kernel CCA, and MC-FR-OI. From the viewpoint of the performance gain using more than two views, the MVAP performs much better than the other multi-view clustering approaches.

In conclusion, the above comparison results show that the proposed MVAP algorithm significantly outperforms the existing multi-view clustering approaches on all the testing datasets. The main reason is that MVAP directly takes into account the balance between the clustering quality within individual views and the explicit clustering consistency across different views, which addresses the shortcomings of the existing methods.

4.3.5 Comparing Computational Efficiency

To illustrate the computational efficiency of the proposed MVAP algorithm, Table 6 reports the average time (in seconds) over 100 runs on the five testing datasets. From the table, it is clear that the proposed MVAP algorithm consumes about twice the computational time of AP on the three two-view datasets (i.e., VIS/NIR, Cora and CiteSeer) and three times the computational time of AP on the two three-view datasets (i.e., Reuters and Mfeat). On the other hand, among the seven multi-view clustering algorithms, although the proposed MVAP algorithm is not the fastest, yet its computational time is not the largest.

Additionally, to show the time complexity of MVAP, we generate a series of sub-datasets, named M400, M600, M800, M1000, M1200, M1400, M1600, M1800, M2000 respectively, from the original Mfeat dataset. The M400 sub-dataset consists of two randomly selected classes of Mfeat and hence 400 data points, and similarly for the other sub-datasets. Fig. 8 reports the average time (in seconds) consumed by MVAP over 100 runs on the nine sub-datasets of increasing sizes. The computational complexity shows that the proposed MVAP algorithm scales quadratically with the data size.

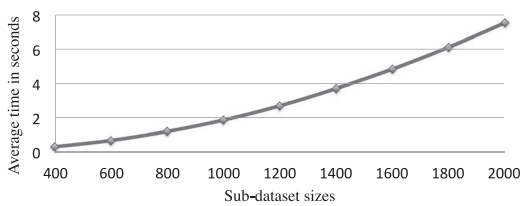


Fig. 8. Average time (in seconds) over 100 runs on the nine sub-datasets of increasing sizes.

5 Conclusions

In this paper, we have proposed a novel multi-view affinity propagation algorithm for multi-view clustering. A two-term multi-view model is established that is able to simultaneously maximize the clustering quality in individual views and ensure the explicit clustering consistency across views. The first term models the within-view clustering quality by summing together the similarities between data points and their assigned exemplars in each view; while the second term models the explicit cross-view clustering consistency by summing together the item-wise exemplar consistency across views. To solve this model, based on the max-sum belief propagation, a multi-view affinity propagation is proposed which works by passing messages not only within individual views but also across different views. For finding optimization efficiently, by using some previously designed mathematical techniques [22], [23], [24], these messages as well as the cluster assignment computations are simplified so as to get simple yet functionally equivalent computations. Extensive experiments have been conducted on five real-world multi-view datasets. Comparison results have validated the effectiveness of the proposed approach in clustering multiple views. Moreover, by comparing with existing multi-view clustering algorithms, it is shown that the proposed MVAP approach is especially suitable for clustering more than two views.

Acknowledgments

This work was supported by NSFC (61173084 and 61502543), GuangZhou Program (No. 201508010032), Guangdong Natural Science Funds for Distinguished Young Scholar, CCF-Tencent Open Research Fund, the PhD Start-up Fund of Natural Science Foundation of Guangdong Province, China (2014A030310180), the US National Science Foundation (NSF) through grants III-1526499, and CNS-1115234.

REFERENCES

[1] K. Crammer, M. Kearns, and J. Wortman, "Learning from multiple sources," *J. Mach. Learn. Res.*, vol. 9, pp. 1757–1774, 2008.

[2] M.-R. Amini and C. Goutte, "A co-classification approach to learning from multilingual corpora," *Mach. Learn.*, vol. 79, pp. 105–121, 2010.

[3] E. Taralova, F. D. la Torre, and M. Hebert, "Source constrained clustering," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2011, pp. 1927–1934.

[4] N. Chen, J. Zhu, F. Sun, and E. P. Xing, "Large-margin predictive latent subspace learning for multiview data analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 12, pp. 2365–2378, Dec. 2012.

[5] X. Chen, X. Xu, J. Z. Huang, and Y. Ye, "TW-*k*-means: Automated two-level variable weighting clustering algorithm for multiview data," *IEEE Trans. Knowl. Data Eng.*, vol. 25, no. 4, pp. 932–944, Apr. 2013.

[6] X. Zou, J. Kittler, and K. Messer, "Face recognition using active near-IR illumination," in *Proc. Brit. Mach. Vis. Conf.*, 2005, pp. 209–219.

[7] P. Sen and L. Getoor, "Link-based classification," Univ. Maryland, College Park, MD, Tech. Rep. CS-TR-4858, Feb. 2007.

[8] N. Cesa-Bianchi, D. R. Hardoon, and G. Leen, "Guest editorial: Learning from multiple sources," *Mach. Learn.*, vol. 79, pp. 1–3, 2010.

[9] Y. Mansour, M. Mohri, and A. Rostamizadeh, "Domain adaptation with multiple sources," in *Proc. Adv. Neural Inf. Process. Syst.*, 2008, pp. 1041–1048.

[10] M. Dredze, A. Kulesza, and K. Crammer, "Multi-domain learning by confidence-weighted parameter combination," *Mach. Learn.*, vol. 79, pp. 123–149, 2010.

[11] R. Luis, L. E. Sucar, and E. F. Morales, "Inductive transfer for learning Bayesian networks," *Mach. Learn.*, vol. 79, pp. 227–255, 2010.

[12] Y. Yan, R. Rosales, G. Fung, F. Farooq, B. Rao, and J. Dy, "Active learning from multiple knowledge sources," in *Proc. 15th Int. Conf. Artif. Intell. Stat.*, 2012, pp. 1350–1357.

[13] M. Szafranski, Y. Grandvalet, and A. Rakotomamonjy, "Composite kernel learning," *Mach. Learn.*, vol. 79, pp. 73–103, 2010.

[14] V. R. de Sa, P. W. Gallagher, J. M. Lewis, and V. L. Malave, "Multi-view kernel construction," *Mach. Learn.*, vol. 79, pp. 47–71, 2010.

[15] M.-R. Amini, N. Usunier, and C. Goutte, "Learning from multiple partially observed views – an application to multilingual text categorization," in *Proc. Adv. Neural Inf. Process. Syst.*, 2009, pp. 28–36.

[16] S. Bickel and T. Scheffer, "Multi-view clustering," in *Proc. IEEE Int. Conf. Data Mining*, 2004, pp. 19–26.

[17] S. Rogers, A. Klami, J. Sinkkonen, M. Girolami, and S. Kaski, "Infinite factorization of multiple non-parametric views," *Mach. Learn.*, vol. 79, pp. 201–226, 2010.

[18] A. Kumar and H. Daume III, "A co-training approach for multi-view spectral clustering," in *Proc. 28th Int. Conf. Mach. Learn.*, 2011, pp. 393–400.

[19] G. Li, K. Chang, and S. C. Hoi, "Multiview semi-supervised learning with consensus," *IEEE Trans. Knowl. Data Eng.*, vol. 24, no. 1, pp. 2040–2051, Nov. 2012.

[20] A. Kumar, P. Rai, and H. Daume III, "Co-regularized multi-view spectral clustering," in *Proc. Adv. Neural Inf. Process. Syst.*, 2011, pp. 1413–1421.

[21] F. R. Kschischang, B. J. Frey, and H.-A. Loeliger, "Factor graphs and the Sum-product algorithm," *IEEE Trans. Inf. Theory*, vol. 47, no. 2, pp. 498–519, Feb. 2001.

[22] B. J. Frey and D. Dueck. (2007). Clustering by passing messages between data points. *Science* [Online]. *315*, pp. 972–976, Available: http://www.psi.toronto.edu/index.php?q=affinity%20propagation

[23] I. E. Givoni and B. J. Frey, "A binary variable model for affinity propagation," *Neural Comput.*, vol. 21, no. 6, pp. 1589–1600, Jun. 2009.

[24] C.-D. Wang, J.-H. Lai, C. Y. Suen, and J.-Y. Zhu, "Multi-exemplar affinity propagation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 9, pp. 2223–2237, Sep. 2013.

[25] T. J. Leih, J. Harmse, and E. Giannopowlos, "Multiple source clustering: A probabilistic reasoning approach," in *Proc. 1st Australian Data Fusion Symp.*, 1996, pp. 141–146.

[26] W. Tang, Z. Lu, and I. S. Dhillon, "Clustering with multiple graphs," in *Proc. IEEE 9th Int. Conf. Data Mining*, 2009, pp. 1016–1021.

[27] Y. Kim, M.-R. Amini, C. Goutte, and P. Gallinari, "Multi-view clustering of multilingual documents," in *Proc. 33rd Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2010, pp. 821–822.

[28] D. Zhou and C. J. Burges, "Spectral clustering and transductive learning with multiple views," in *Proc. 24th Int. Conf. Mach. Learn.*, 2007, pp. 1159–1166.

[29] X. Wang, B. Qian, and I. Davidson, "Improving document clustering using automated machine translation," in *Proc. 21st ACM Int. Conf. Inf. Knowl. Manage.*, 2012, pp. 645–653.

[30] X. Wang, B. Qian, J. Ye, and I. Davidson, "Multi-objective multi-view spectral clustering via Pareto optimization," in *Proc. 13th SIAM Int. Conf. Data Mining*, 2013, pp. 234–242.

[31] K. Chaudhuri, S. M. Kakade, K. Livescu, and K. Sridharan, "Multi-view clustering via canonical correlation analysis," in *Proc. Int. Conf. Mach. Learn.*, 2009, pp. 1–8.

[32] T. M. Selee, T. G. Kolda, W. P. Kegelmeyer, and J. D. Griffin, "Extracting clusters from large datasets with multiple similarity measures using IMSCAND," Sandia Nat. Laboratories, Albuquerque, NM, Tech. Rep. SAND2007-7977, 2007.

[33] L. Grasedyck, D. Kressner, and C. Tobler, "A literature survey of low-rank tensor approximation techniques," *GAMM-Mitteilungen*, vol. 36, no. 1, pp. 53–78, 2013.

[34] X. Liu, S. Ji, W. Glänzel, and B. D. Moor, "Multiview partitioning via tensor methods," *IEEE Trans. Knowl. Data Eng.*, vol. 25, no. 5, pp. 1056–1069, May 2013.

[35] X. Cao, X. Wei, Y. Han, and D. Lin, "Robust face clustering via tensor decomposition," *IEEE Trans. Cybern.*, vol. 45, no. 11, pp. 2546–2557, Nov. 1, 2015.

[36] D. Dueck, "Affinity propagation: Clustering data by passing messages," Ph.D. dissertation, Univ. Toronto, Toronto, ON, 2009.

[37] A. Strehl and J. Ghosh, "Cluster ensembles—a knowledge reuse framework for combining multiple partitions," *J. Mach. Learn. Res.*, vol. 3, pp. 583–617, 2002.

[38] Y. Weiss and W. T. Freeman, "On the optimality of solutions of the max-product belief-propagation algorithm in arbitrary graphs," *IEEE Trans. Inf. Theory*, vol. 47, no. 2, pp. 736–744, Feb. 2001.

[39] D. G. Lowe, "Distinctive image features from Scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, 2004.

[40] D. Dueck and B. J. Frey, "Non-metric affinity propagation for unsupervised image categorization," in *Proc. IEEE 11th Int. Conf. Comput. Vis.*, 2007, pp. 1–8.

[41] L. Fei-Fei and P. Perona. (2005). A Bayesian hierarchical model for learning natural scene categories. *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 524–531 [Online]. Available: http://vision.stanford.edu/Datasets/SceneClass13.rar

[42] K. Grauman and T. Darrell, "Unsupervised learning of categories from sets of partially matching image features," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recog.*, 2006, pp. 2596–2603.

[43] K. Mikolajczyk, B. Leibe, and B. Schiele, "Multiple object class detection with a generative model," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recog.*, 2006, pp. 26–36.

[44] N. Ueffing, M. Simard, S. Larkin, and H. Johnson, "NRC's PORTAGE system for WMT 2007," in *Proc. 2nd Workshop Statistical Mach. Translation*, 2007, pp. 185–188.

[45] K. Bache and M. Lichman. (2013). UCI machine learning repository [Online]. Available: http://archive.ics.uci.edu/ml

[46] A. Strehl, J. Ghosh, and R. J. Mooney, "Impact of similarity measures on Web-page clustering," in *Proc. AAAI Workshop AI Web Search*, 2000, pp. 58–64.

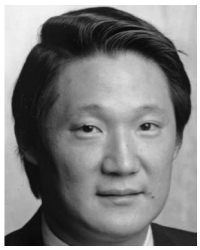
[47] L. Hubert and P. Arabie, "Comparing partitions," *J. Classification*, vol. 2, pp. 193–218, 1985.

[48] M. Meilă, "Comparing clusterings-an axiomatic view," in *Proc. Int. Conf. Mach. Learn.*, 2005, pp. 577–584.

[49] J. Shi and J. Malik. (2000, Aug.). Normalized cuts and image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* [Online]. *22(8)*, pp. 888–905, Available: http://www.cis.upenn.edu/~jshi/software/

Chang-Dong Wang received the PhD degree in computer science in 2013 from Sun Yat-sen University, Guangzhou, China. He is a visiting student at the University of Illinois at Chicago from January 2012 to November 2012. He joined Sun Yat-sen University in 2013 as an assistant professor with the School of Mobile Information Engineering. His current research interests include machine learning and data mining. He has published more than 30 scientific papers in international journals and conferences such as the *IEEE TPAMI*, *IEEE TKDE*, *IEEE TSMC-C*, *Pattern Recognition*, *KAIS*, *Neurocomputing*, ICDM, and SDM. His ICDM 2010 paper received the Honorable Mention for Best Research Paper Awards. He received 2012 Microsoft Research Fellowship Nomination Award. He received 2015 Chinese Association for Artificial Intelligence (CAAI) Outstanding Dissertation. He is a member of the IEEE.

Jian-Huang Lai received the MSc degree in applied mathematics in 1989 and the PhD degree in mathematics in 1999 from Sun Yat-sen University, China. He joined Sun Yat-sen University in 1989 as an assistant professor, where he is currently a professor with the Department of Automation, School of Information Science and Technology and the dean in the School of Information Science and Technology. His current research interests are in the areas of digital image processing, pattern recognition, multimedia communication, and wavelet and its applications. He has published over 100 scientific papers in the international journals and conferences on image processing and pattern recognition, e.g., *IEEE TPAMI*, *IEEE TKDE*, *IEEE TNN*, *IEEE TIP*, *IEEE TSMC (Part B), Pattern Recognition*, ICCV, CVPR, and ICDM. He serves as a standing member of the Image and Graphics Association of China and also serves as a standing director in the Image and Graphics Association of Guangdong. He is a senior member of the IEEE.



Philip S. Yu received the BS degree in EE from National Taiwan University, the MS and PhD degrees in EE from Stanford University, and the MBA degree from New York University. He is a distinguished professor of computer science at the University of Illinois at Chicago and also holds the Wexler chair in Information Technology. Before joining UIC, he was at IBM TJ Watson Research Center, where he was a manager of the Software Tools and Techniques group. His research interest is on big data, including data mining, data stream, database, and privacy. He has published more than 810 papers in refereed journals and conferences. He holds or has applied for more than 300 US patents. He is the editor-in-chief of the *ACM Transactions on Knowledge Discovery from Data*. He is on the steering committee of the IEEE Conference on Data Mining and ACM Conference on Information and Knowledge Management and was a member of the IEEE Data Engineering steering committee. He was the editor-in-chief of the I*EEE Transactions on Knowledge and Data Engineering* (2001-2004). He received the IEEE Computer Society 2013 Technical Achievement Award for "pioneering and fundamentally innovative contributions to the scalable indexing, querying, searching, mining and anonymization of big data," the ICDM 2013 10-year Highest-Impact Paper Award, the EDBT Test of Time Award (2014), and the Research Contributions Award from the IEEE International Conference on Data Mining (2003). He had received several IBM honors including two IBM Outstanding Innovation Awards, an Outstanding Technical Achievement Award, two Research Division Awards, and the 94th Plateau of Invention Achievement Awards. He was an IBM master inventor. He is a fellow of the ACM and IEEE.