Knowledge-Based Systems 121 (2017) 153–162

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys

# Multi-view learning via multiple graph regularized generative model

Shaokai Wang[a], Eric Ke Wang[a], Xutao Li[a], Yunming Ye[a,*], Raymond Y.K. Lau[b], Xiaolin Du[c]

[a] Department of Computer Science, Shenzhen Graduate School, Harbin Institute of Technology, Shenzhen, China
[b] Department of Information Systems, City University of Hong Kong, Hong Kong
[c] College of Computer Science, Beijing University of Technology, Beijing, China

ARTICLE INFO

*Article history:*
Received 21 June 2016
Revised 9 December 2016
Accepted 17 January 2017
Available online 3 February 2017

*Keywords:*
Multi-view learning
Generative model
Manifold learning

ABSTRACT

Topic models, such as probabilistic latent semantic analysis (PLSA) and latent Dirichlet allocation (LDA), have shown impressive success in many fields. Recently, multi-view learning via probabilistic latent semantic analysis (MVPLSA), is also designed for multi-view topic modeling. These approaches are instances of generative model, whereas they all ignore the manifold structure of data distribution, which is generally useful for preserving the nonlinear information. In this paper, we propose a novel multiple graph regularized generative model to exploit the manifold structure in multiple views. Specifically, we construct a nearest neighbor graph for each view to encode its corresponding manifold information. A multiple graph ensemble regularization framework is proposed to learn the optimal intrinsic manifold. Then, the manifold regularization term is incorporated into a multi-view topic model, resulting in a unified objective function. The solutions are derived based on the Expectation Maximization optimization framework. Experimental results on real-world multi-view data sets demonstrate the effectiveness of our approach.

## 1. Introduction

Many data sets in real world are collected from multiple sources or represented by multiple views, where different views describe distinct perspectives of the data. For example, in computer vision, each image or video can be represented by multiple types of features to describe different aspects of visual characteristics, such as SIFT [1], HOG [2], LBP [3], CENTRIST [4]. In text corpus, one document may be translated into multiple different languages, and the translation in each language can be considered as a view. We usually cannot concatenate all multiple views into one single view for finding patterns, since each view has its own specific statistical property. Thus, many multi-view learning algorithms have been proposed to explore the rich information among different views [5–9].

Topic models, such as probabilistic Latent Semantic Analysis (pLSA) [10] and Latent Dirichlet Allocation (LDA) [11], are useful approaches for exploiting co-occurrence information on text or visual data, which can be used to find high-level latent topics. In text analysis, these methods model each document as a mixture over a fixed set of underlying topics, where each topic is characterized as a distribution over words. These approaches have shown impressive success in discovering low-rank hidden structures for textual and visual data [12–14]. Recently, Zhuang et al. [15] proposed MV-PLSA method, which is a multi-view topic modeling algorithm via Probabilistic Latent Semantic Analysis. The algorithm jointly models the co-occurrences of features and documents from different views. However, the limitation of these topic models is that they all fail to consider the intrinsic geometrical structure of the data distribution.

The geometrical structure of data distribution can be modeled by using a manifold regularization term. Many previous studies [16,17] have shown that high-dimensional data, such as texts and images, are often found to be embedded on a low-rank nonlinear manifold, and learning the manifold structure can provide better dimensionality reduction mapping.

Recently, Cai et al. proposed two topic models, Laplacian Probabilistic Latent Semantic Indexing (LapPLSI) [18] and Locally-consistent Topic Modeling (LTM) [19], which use manifold information based on PLSA. Specifically, the manifold structure is modeled by a nearest neighbor graph. The graph is then utilized to smooth the probability density functions. By doing so, these two models obtain better probabilistic distributions and show higher discriminative power than PLSA and LDA on document clustering. However, both models are designed for single-view data.

In this paper, we propose a new multi-view clustering method via generative model, which leverages an ensemble regularization term to consider the manifold information in multiple views. In particular, we construct a nearest neighbor graph to model the underlying manifold structure for each view. A multiple graph ensem-

[*] Corresponding author.
*E-mail address:* yeyunming@hit.edu.cn (Y. Ye).

http://dx.doi.org/10.1016/j.knosys.2017.01.022
0950-7051/© 2017 Elsevier B.V. All rights reserved.

ble regularization framework is proposed to combine the intrinsic manifold information with these graphs. Then, the manifold regularization term is incorporated into Probabilistic Latent Semantic Analysis to form our objective function. More specifically, the main contributions of this paper include:

- We propose a novel Probabilistic Latent Semantic Analysis based Multiple Graph regularized Generative Model for multi-view clustering (MGGM). The MGGM can effectively utilize the multi-view manifold information and achieve an improvement by 6% against the counterpart without using the information.
- Different views have different underlying manifold structure. In order to obtain the intrinsic manifold approximation based on multiple views, we propose a multiple graph ensemble regularization framework, which can automatically combine the nearest neighbor graphs of different views to approximate the intrinsic manifold.
- We have conducted comprehensive experiments on text and image data sets. Experimental results show that our algorithm is substantially better than state-of-the-art multi-view clustering methods in terms of clustering accuracy.

The rest of this paper is organized as follows: Section 2 reviews related work. Section 3 presents our proposed MGGM method. Experimental results are given in Section 4. Finally, we conclude the paper in Section 5.

## 2. Related work

Suppose that we have $N$ data points $X = \{x_1, \dots, x_N\}$, which are from $K$ classes $C = \{c_1, \dots, c_K\}$. There are $T$ views $\{f_1, \dots, f_T\}$ for the data, where $f_t$ denotes the $t$th view label. Each data point $x_i \in X$ has a feature vector $< w^t_{i,1}, \dots, w^t_{i,M_t} >$ for view $f_t$, where the number of features is $M_t$. Our goal in multi-view clustering is to partition $X = \{x_1, \dots, x_N\}$ into $K$ clusters by exploiting the information contained in all $T$ different views $\{f_1, \dots, f_T\}$.

In the past decades, exploring the rich information among multiple views arouses vast amount of interest [5,7,15,20–26]. One of the earliest schemes for multi-view learning is co-training, which is introduced by Blum and Mitchell [20]. It learns the model by maximizing the mutual agreement on predictions of the unlabeled data. Specifically, co-training built a classifier separately on each view, and then each classifier's predictions on new unlabeled instances are used to enlarge the training set of the other. Many works have been developed following the idea of co-training [21–24]. For example, Nigam and Ghani [21] developed a Co-EM algorithm. Yu et al. [23] proposed a Bayesian undirected graphical model for co-training through Gaussian process. Kumar and Daum [24] developed a multi-view spectral clustering algorithm. Recently, Song et al. [5] proposed a model for data missing completion by seamlessly exploring the knowledge from multiple sources, and applied it to the application of volunteerism tendency prediction. Song et al. [6] explored users' social content from multiple social networks to predict their interests. Nie et al. [7] developed an approach to model the progression of chronic diseases based on multimedia and multimodal observational data. However, all the co-training based approaches make the conditional independent assumption to work well, but the assumption in practice is usually too strong to be satisfied. Hence, these methods may not work effectively [27].

Probabilistic Latent Semantic Analysis (PLSA) was originally developed for latent topic analysis for text data [10]. Recently some multi-view clustering methods based on PLSA are proposed. For example, Yu et al. [28,29] proposed a collaborative PLSA for multi-view clustering, whose central idea is to maximize the mutual agreements between different views. Zhuang et al. [15] recently extended PLSA to multi-view scenario and proposed MVPLSA. Fig. 1 shows the graphical model. In the MVPLSA model, based on the observation that different features may be grouped together to indicate a high-level concept, let $\{z^t_1, \dots, z^t_{Q^t}\}$ in each view $f_t$ be the high-level latent topic. The high-level latent topic $z^t_q$ and the data class $c_k$ are two latent variables. The data point $x_i$, feature $w^t_{i,j}$ and view label $f_t$ are three visible variables. Given this graphical model the joint probability over the variables is

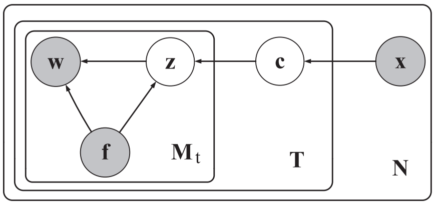

**Fig. 1.** The graphical model of MVPLSA.

$$P(x_i, w^t_{i,j}, f_t) = P(x_i)P(f_t)P(w^t_{i,j}|x_i, f_t)$$

$$P(w^t_{i,j}|x_i, f_t) = \sum_{k=1}^{K}\sum_{q=1}^{Q^t} P(w^t_{i,j}|z^t_q, f_t)P(z^t_q|c_k, f_t)P(c_k|x_i). \qquad (1)$$

Let $n(x_i, w^t_{i,j}, f_t)$ be the frequency of feature $w^t_{i,j}$ occurring in $x_i$, and $M_t$ be the number of features in the $t$th view $f_t$. The model parameters $P(c_k|x_i)$, $P(z^t_q|c_k, f_t)$ and $P(w^t_{i,j}|z^t_q, f_t)$ can be learned by maximizing the log-likelihood

$$\begin{aligned}\mathcal{L} &= \sum_{i=1}^{N}\sum_{t=1}^{T}\sum_{j=1}^{M_t} n(x_i, w^t_{i,j}, f_t)\log P(x_i, w^t_{i,j}, f_t)\\ &\propto \sum_{i=1}^{N}\sum_{t=1}^{T}\sum_{j=1}^{M_t} n(x_i, w^t_{i,j}, f_t)\log \sum_{k=1}^{K}\\ &\sum_{q=1}^{Q^t} P(w^t_{i,j}|z^t_q, f_t)P(z^t_q|c_k, f_t)P(c_k|x_i). \qquad (2)\end{aligned}$$

The MVPLSA model can jointly model the co-occurrences from multiple views and obtain additional gains. However, this model fails to consider the intrinsic geometrical structure of the data distribution, which is generally useful for preserving the nonlinear information.

## 3. Method

In this paper, we derive a PLSA-based Multiple Graph Regularized Generative Model which considers both multi-view and manifold information. We construct a nearest neighbor graph to model the underlying manifold structure for each view. The Laplacian matrix is then computed relying on the corresponding graph of each view. Multiple manifold regularization terms are separately constructed for each view. A multiple graph ensemble regularization framework is proposed to combine multiple manifolds. Then, the regularization term is incorporated into MVPLSA, resulting in our objective function.

### *3.1. Manifold regularization term*

The smoothness assumption considers that if two data samples $x_i$ and $x_j$ are close on the manifold, then the corresponding conditional probability distributions $P(C|x_i)$ and $P(C|x_j)$ are also close. By setting up a multiple graph ensemble regularization term in the generative model, the manifold structure of each view can be taken into account.

Here, we consider the Kullback–Leibler divergence to measure the distance between two distributions based on multi-view data. Let $P_i(c_k) = P(c_k|x_i)$ be the probability that instance $x_i$ is labeled as $c_k$. Let $P_i(C) = \{P(c_k|x_i)\}_k$ and $P_s(C) = \{P(c_k|x_s)\}_k$, and the Kullback–Leibler divergence between $P_i(C)$ and $P_s(C)$ is defined as

$$D\big(P_i(C)||P_s(C)\big) = \sum_k P_i(c_k)\log\frac{P_i(c_k)}{P_s(c_k)}.$$

Kullback–Leibler divergence is not symmetric, we use the following symmetric Kullback–Leibler divergence

$$\mathcal{D}\big(P_i(C), P_s(C)\big) = \frac{1}{2}\left(D(P_i(C)||P_s(C)) + D(P_s(C)||P_i(C))\right).$$

The local manifold structure can be effectively modeled through a nearest neighbor graph on a scatter of data points [16]. Let matrix $E \in \mathbb{R}^N \times \mathbb{R}^N$ be the adjacency graph, wherein each element $E_{ij}$ is the weight of the edge between instance $x_i$ and instance $x_j$. $E$ is the intrinsic manifold approximation of the multi-view data which is calculated by a linear combination of manifolds of each view. Here we first assume that $E$ is known. We'll introduce how to obtain $E$ in Section 3.2.

By using the symmetric Kullback–Leibler divergence, we can measure the smoothness of conditional probability distribution $P(C|x_i)$ on the intrinsic manifold of the multi-view data. The regularization term can be formulated as

$$\mathcal{R} = \sum_{i=1}^{N}\sum_{s=1}^{N}\mathcal{D}\big(P_i(C), P_s(C)\big)E_{is}. \tag{3}$$

Minimizing the regularization term $\mathcal{R}$ can guarantee the smoothness of conditional probability distribution $P(C|x_i)$ on the intrinsic manifold.

To solve multi-view learning problem, we aim to estimate conditional probability distribution $P(c_k|x_i)$ by maximizing log-likelihood in the generative model and the conditional probability distributions smoothness on the intrinsic manifold of the multi-view data. Formally, assuming the intrinsic manifold approximation $E$ is given, $P(c_k|x_i)$ can be estimated by maximizing the regularized log-likelihood as follows:

$$\begin{aligned}\max_P \quad \mathcal{O} &= \mathcal{L} - \lambda_1\mathcal{R}\\ &= \sum_{i=1}^{N}\sum_{t=1}^{T}\sum_{j=1}^{M_t} n(x_i, w^t_{i,j}, f_t)\log\sum_{k=1}^{K}\sum_{q=1}^{Q^t} P(w^t_{i,j}|z^t_q, f_t)\\ &\quad\times P(z^t_q|c_k, f_t)P(c_k|x_i)\\ &\quad -\lambda_1\sum_{i=1}^{N}\sum_{s=1}^{N}\mathcal{D}\big(P_i(C), P_s(C)\big)E_{is},\end{aligned} \tag{4}$$

where $\lambda_1$ is a regularization parameter, which can be tuned in a cross validation manner.

Next, we'll introduce how to obtain the intrinsic manifold approximation $E$.

### 3.2. Multiple graph ensemble regularization

In the literature, the local manifold structure is often modeled through a nearest neighbor graph [16]. Hence, we construct a nearest neighbor graph to model the underlying manifold structure for each view. Let us define the edge weight matrix $U^t$ of view $f_t$ as follows:

$$U^t_{is} = \begin{cases} 1, & if\ x_i \in N_p(x_s)\ or\ x_s \in N_p(x_i)\ in\ view\ f_t\\ 0, & otherwise.\end{cases} \tag{5}$$

where $N_p(x_i)$ denotes the set of $p$ nearest neighbors of $x_i$ (with respect to the Euclidean distance). As a result, $T$ graphs $\{U^t\}_{t=1}^T$can be obtained. we would like to combine them to approximate the intrinsic manifold. However, there are no explicit rules to combine them.

Inspired by the ensemble manifold regularizer [30,31], which assumes that the intrinsic manifold approximately lies in the convex hull of the previously given manifold candidates, we adopt a linear combination way to estimate the intrinsic manifold, i.e.,

$$E = \sum_{t=1}^{T}\mu_t U^t,\ s.t.\ \sum_{t=1}^{T}\mu_t = 1,\ \mu_t \geq 0,\ for\ t = 1, \ldots, T. \tag{6}$$

The key issue of such approximation is determining the combination parameters $\{\mu_t\}_{t=1}^T$. Next, we'll introduce how to learn them effectively.

Let $\{L^t\}_{t=1}^T$ be a set of graph Laplacian matrices in multiple views [32], where $L^t = D^t - U^t$, $D$ and $D^t$ is a diagonal matrix with the $i$th diagonal entry $D_{ii} = \sum_{j=1}^N E_{ij}$, $D^t_{ii} = \sum_{j=1}^N U^t_{ij}$. We have $L = D - E = \sum_{t=1}^T \mu_t L^t$. Based on the smoothness assumption, similar to Eq. (3), we have:

$$\begin{aligned}\overline{\mathcal{R}} &= \frac{1}{2}\sum_{k=1}^{K}\sum_{i,s=1}^{N}\big(P(c_k|x_i) - P(c_k|x_s)\big)^2 E_{is}\\ &= \sum_{k=1}^{K}\left(\sum_{i=1}^{N}P(c_k|x_i)^2 D_{ii} - \sum_{i,s=1}^{N}P(c_k|x_i)E_{is}P(c_k|x_s)\right)\\ &= \sum_{t=1}^{T}\mu_t\big(Tr(P^{\mathrm{T}}L^tP)\big)\end{aligned} \tag{7}$$

$Tr(P^{\mathrm{T}}L^tP)$ can be used to measure the smoothness of conditional probability distribution $P$ on the manifold structure of view $f_t$. The smaller $Tr(P^{\mathrm{T}}L^tP)$ is, the more smooth conditional probability distribution $P$ on the manifold structure of view $f_t$ is. In order to explore the complementary property of different views, we prefer to assign a larger weight $\mu_t$ to smaller $Tr(P^{\mathrm{T}}L^tP)$. In other words, the smaller $Tr(P^{\mathrm{T}}L^tP)$ is, the view $f_t$ plays the more important role in learning the conditional probability distribution $P$.

Following this idea, we propose a multiple graph ensemble regularization framework to learn the hyperparameters $\{\mu_t\}_{t=1}^T$. When $P$ is fixed, the optimal $\{\mu_t\}_{t=1}^T$ corresponds to the solution of the following minimization problem:

$$\begin{aligned}&\min_{\boldsymbol{\alpha}} \quad \sum_{t=1}^{T}\alpha_t^{\frac{1}{\lambda_2}}Tr(P^{\mathrm{T}}L^tP)\\ &s.t.\quad \textstyle\sum_{t=1}^T\alpha_t = 1,\ \alpha_t \geq 0,\ \mu_t = \alpha_t^{\frac{1}{\lambda_2}}\ for\ t = 1, \ldots, T.\end{aligned} \tag{8}$$

where $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_T)^{\mathrm{T}}$, $\lambda_2 \in (0, 1)$ is a parameter to control the distribution of weights $\alpha_t^{\frac{1}{\lambda_2}}$.

Applying the Lagrange multiplier method, we have the Lagrange function

$$L(\boldsymbol{\alpha}, \gamma) = \sum_{t=1}^{T}\alpha_t^{\frac{1}{\lambda_2}}Tr(P^{\mathrm{T}}L^tP) + \gamma\left(1 - \sum_{t=1}^{T}\alpha_t\right) \tag{9}$$

where $\gamma$ is the Lagrange multiplier. By setting the derivative of $L(\boldsymbol{\alpha}, \gamma)$ with respect to $\alpha_t$ and $\gamma$ to zero, we can obtain the closed-form of $\mu_t$ as

$$\mu_t = \alpha_t^{\frac{1}{\lambda_2}} = \frac{Tr(P^{\mathrm{T}}L^tP)^{\frac{1}{\lambda_2-1}}}{\left(\sum_{t=1}^T Tr(P^{\mathrm{T}}L^tP)^{\frac{\lambda_2}{\lambda_2-1}}\right)^{\frac{1}{\lambda_2}}} \tag{10}$$

Because $\lambda_2 \in (0, 1)$, Eq. (10) shows that smaller $Tr(P^{\mathrm{T}}L^tP)$ will be assigned with larger weights. As a result, the important views are selected and irrelevant or noisy views are eliminated.

After learning the optimal graph hyperparameter $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_T)^{\mathrm{T}}$, we obtain the intrinsic manifold approximation

$$E = \sum_{t=1}^{T}\mu_t U^t. \tag{11}$$

### 3.3. Parameter estimation

In our model, $\{P(c_k|x_i), P(z_q^t|c_k, f_t), P(w_{i,j}^t|z_q^t, f_t)\}$ is the set of parameters to be estimated. For clarity, we define $\Theta = \{P(c_k|x_i), P(z_q^t|c_k, f_t), P(w_{i,j}^t|z_q^t, f_t)\}$. The standard procedure for maximum likelihood estimation with latent variables is the Expectation Maximization (EM) algorithm. For our problem, we develop an EM-like algorithm to solve the optimization problem in Eq. (4).

E-step: By applying the Bayes'formula on Eq. (1), we compute the posterior probabilities for the latent variables $P(z_q^t, c_k|x_i, w_{i,j}^t, f_t)$:

$$P(z_q^t, c_k|x_i, w_{i,j}^t, f_t) = \frac{P(w_{i,j}^t|z_q^t, f_t)P(z_q^t|c_k, f_t)P(c_k|x_i)}{\sum_{k=1}^K \sum_{q=1}^{Q^t} P(w_{i,j}^t|z_q^t, f_t)P(z_q^t|c_k, f_t)P(c_k|x_i)}. \tag{12}$$

M-step: In M-step, we maximize the expected complete data log-likelihood:

$$\begin{aligned}\mathcal{Q}(\Theta) = &\sum_{i=1}^{N}\sum_{t=1}^{T}\sum_{j=1}^{M_t} n(x_i, w_{i,j}^t, f_t) \sum_{k=1}^{K}\sum_{q=1}^{Q^t} P(z_q^t, c_k|x_i, w_{i,j}^t, f_t) \\ &\log P(w_{i,j}^t|z_q^t, f_t)P(z_q^t|c_k, f_t)P(c_k|x_i) \\ &-\lambda_1 \sum_{i=1}^{N}\sum_{s=1}^{N} \mathcal{D}\big(P_i(c), P_s(c)\big)E_{ij}.\end{aligned} \tag{13}$$

There are two parts in $\mathcal{Q}(\Theta)$. The second part does not include $P(z_q^t|c_k, f_t)$ and $P(w_{i,j}^t|z_q^t, f_t)$. Thus, the re-estimation equation for $P(z_q^t|c_k, f_t)$ and $P(w_{i,j}^t|z_q^t, f_t)$ are the same as that of MVPLSA [15]:

$$\begin{aligned} P(z_q^t|c_k, f_t) &= \frac{\sum_{i=1}^N \sum_{j=1}^{M_t} n(x_i, w_{i,j}^t, f_t)P(z_q^t, c_k|x_i, w_{i,j}^t, f_t)}{\sum_{q=1}^{Q^t}\sum_{i=1}^N \sum_{j=1}^{M_t} n(x_i, w_{i,j}^t, f_t)P(z_q^t, c_k|x_i, w_{i,j}^t, f_t)} \\ P(w_{i,j}^t|z_q^t, f_t) &= \frac{\sum_{i=1}^N \sum_{k=1}^{K} n(x_i, w_{i,j}^t, f_t)P(z_q^t, c_k|x_i, w_{i,j}^t, f_t)}{\sum_{j=1}^{M_t}\sum_{i=1}^N \sum_{k=1}^{K} n(x_i, w_{i,j}^t, f_t)P(z_q^t, c_k|x_i, w_{i,j}^t, f_t)}. \end{aligned} \tag{14}$$

For parameters $\{P(c_k|x_i)\}$, we can obtain the re-estimation equation by using Lagrange multiplier method similar to LTM [19]:

$$Y_k = (\Omega + \lambda_1 L)^{-1} V_k, \tag{15}$$

where $Y_k = [P(c_k|x_1), \ldots, P(c_k|x_N)]^T$, $\Omega$ denotes a $N$-by-$N$ diagonal matrix with $\rho_i = \sum_{t=1}^T \sum_{j=1}^{M_t} n(x_i, w_{i,j}^t, f_t)$ as entries. $V_k$ is a $N$-dimensional vector with $\sum_{t=1}^T \sum_{j=1}^{M_t} \sum_{q=1}^{Q^t} n(x_i, w_{i,j}^t, f_t)P(z_q^t, c_k|x_i, w_{i,j}^t, f_t)$ as entries. $L$ is a $N$-by-$N$ graph Laplacian matrix, i.e., $L = D - E$, Here $D$ denotes a $N$-by-$N$ diagonal matrix whose entries are row sums of $E$, i.e., $D_{ii} = \sum_s E_{is}$.

In the EM iterations, the E-step (Eq. (12)) and M-step (Eq. (14) and (15)) are alternated repeatedly until convergence of the objective $\mathcal{O}$ is reached. We summarized the procedure of our MGGM algorithm in Algorithm 1.

## 4. Experiments

### 4.1. Datasets

In this section, we evaluate the proposed multi-view learning framework on five real-world multi-view data sets. A summary of the characteristics of these data sets are shown in Table 1.

**Algorithm 1** Multiple graph regularized generative model.

```
Require: multi-view data $X = \{X^{(1)}, \ldots, X^{(T)}\}$, parameters $\lambda_1$, $\lambda_2$, the number of latent topics for each view $Q^t$, the number of nearest neighbors $p$, termination condition value $\epsilon$
Ensure: $P(c_k|x_i)$, $i = 1, \ldots, N$, $k = 1, \ldots, K$
initialize $P(z_q^t|c_k, f_t)$ and $P(w_{i,j}^t|z_q^t, f_t)$ randomly, the initialization of $P(c_k|x_i)$ is detailed in Section 4.6, initialize $\boldsymbol{\mu}$, $\boldsymbol{\mu} \leftarrow 1/T$.
Compute the graph matrix $U^t$ as in Eq. (5)
Compute the intrinsic manifold approximation $E$ as in Eq. (11)
for $it = 1 \rightarrow Max\_It$ do
    E-step: calculate $P(z_q^t, c_k|x_i, w_{i,j}^t, f_t)$ using Eq. (12)
    M-step:
    update $P(z_q^t|c_k, f_t)$ and $P(w_{i,j}^t|z_q^t, f_t)$ using Eq. (14)
    update $P(c_k|x_i)$ using Eq. (15)
    update $\boldsymbol{\mu}$ according to Eq. (10)
    update $E$ using Eq. (11)
    if $(\mathcal{O}(\Theta^{it+1}) - \mathcal{O}(\Theta^{it})) < \epsilon$ then
        break;
    end if
end for
```

**Table 1**
Details of the multi-view data sets used in our experiments.

| Datasets | DBLP | Reuters | Handwritten | Cora | NUS |
|---|---|---|---|---|---|
| #Instances | 4057 | 1200 | 2000 | 12,004 | 30,000 |
| #View 1 | 6167 | 2000 | 240 | 292 | 226 |
| #View 2 | 3787 | 2000 | 76 | 12,004 | 74 |
| #View 3 | – | 2000 | 216 | – | 129 |
| #View 4 | – | – | 47 | – | – |
| #View 5 | – | – | 6 | – | – |
| #Classes | 4 | 6 | 10 | 10 | 31 |

**DBLP Dataset** This bibliographic network dataset is adapted from the original dataset of Ming Ji.[1] The dataset consists of 4057 authors which are classified into four classes: "Database", "Data Mining", "AI", "Information Retrieval". It consists of two views for each author: the 6167 dimensions paper name view and the 3787 dimensions term view. The paper name view is the normalized vector of the names of an author's papers, and the term view is the normalized vector of the terms of an author's papers.

**Reuters Multilingual Dataset** [2] This collection contains feature characteristics of documents originally written in five different languages , and their translations, over a common set of 6 categories. The translation in each language can be considered as a view. There are three views for each documents: the English translation view, the French translation view and the German translation view.

**Handwritten Dataset** [3] This handwritten dataset is from the UCI repository. The dataset consists of 2000 examples. There are five views for each digit: 240 pixel averages in 2 × 3 windows, 76 Fourier coefficients of the character shapes, 216 profile correlations, 47 Zernike moments, and 6 morphological features.

**Cora Dataset** This bibliographic network dataset is adapted from the original Cora dataset [33]. Scientific articles in the dataset are classified into 10 thematic categories. The content view of each article corresponds to the normalized TF-IDF vector of the abstract of the article, which includes 292 dimensions. The relational view corresponds to the citation relation between two articles, which has 12,004 dimensions.

[1] http://web.engr.illinois.edu/~mingji1/DBLP_four_area.zip.

[2] http://membres-lig.imag.fr/grimal/data.html.

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

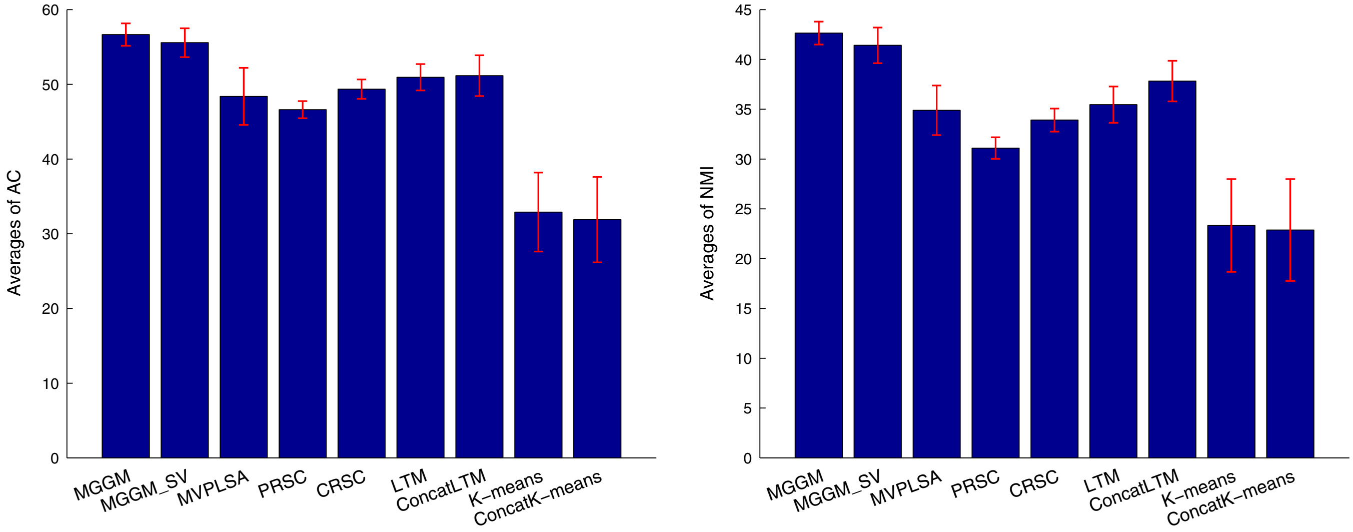

**Fig. 2.** A comparison of average clustering results in terms of AC and NMI.

**NUS-WIDE-Object (NUS) Dataset** [4] This object recognition dataset consists of 30,000 images in 31 classes. We use three views for each image including 226 dimension color moments (CM), 74 dimension edge distribution and 129 dimension wavelet texture.

### 4.2. Comparison methods

We compare the performance of the proposed MGGM algorithm with several baseline methods:

- K-means: K-means is the most common clustering method and we utilize it as a baseline method. The best single view result is reported as final result for evaluation.
- ConcatK-means: We concatenate the features of all the views, and then run K-means directly on this concatenated view representation.
- Locally-consistent Topic Modeling (LTM) [19]: LTM trains the models from each single view. This model provides a principled way to utilize the single view manifold information. The best single view result is thus reported for evaluation.
- ConcatLTM: After concatenating the features of all the views, we then run LTM directly on this concatenated view representation.
- MVPLSA [15]: MVPLSA is a generative model for multi-view learning via Probabilistic Latent Semantic Analysis.
- Pairwise Co-regularized Spectral Clustering (PRSC) [24]: PRSC pushes pairwise spectral embeddings of different feature types close to each other by utilizing object-object similarity matrices, followed by k-means clustering on embedding results. There is a hyperparameter to trade off the spectral clustering objective and the spectral embedding (dis)agreement. We set the parameter to be 0.01 as suggested.
- Centroid Co-regularized Spectral Clustering (CRSC) [24]: CRSC pushes all spectral embeddings of different feature types close to a centroid embedding by utilizing object-object similarity matrices, and then runs K-means to identify clusters. Again, we use the default value 0.01 for the parameter in this method.

### 4.3. Evaluation metrics

The clustering result is evaluated by comparing the obtained label of each data point with the label provided by the dataset. As adopted in [34], two clustering metrics are used for testing the performance: clustering accuracy (AC) and normalized mutual information (NMI). The result is evaluated by comparing the cluster label of each sample with the label provided by the data set. The AC is used to measure the percentage of correct labels predicted. Given a data set containing $N$ instances, for the $i$th instance, let $r_i$ be the cluster label obtained by applying different algorithms and $s_i$ be the label provided by the date set. The AC is defined as follows:

$$AC = \frac{\sum_{i=1}^{N} \delta(s_i, map(r_i))}{N}$$

where $\delta(x, y)$ is the delta function that equals one if $x = y$ and zero otherwise, and $map(r_i)$ is the permutation mapping function that maps each cluster label $r_i$ to the equivalent label from the data set. The best mapping can be found using the Kuhn-Munkres algorithm.

In the clustering applications, mutual information is used to measure how similar two sets of clusters are. Given two clustering results $C$ and $C'$. Their mutual information metric $MI(C, C')$ is defined as follows:

$$MI(C, C') = \sum_{c_i \in C, c'_j \in C'} p(c_i, c'_j) log \frac{p(c_i, c'_j)}{p(c_i)p(c'_j)}$$

where $p(c_i)$ and $p(c'_j)$ denote the probabilities that an instance arbitrarily selected from the data set belongs to the clusters $c_i$ and $c'_j$, respectively, and $p(c_i, c'_j)$ is the joint probability that the arbitrarily selected instance belongs to the clusters $c_i$ and $c'_j$ at the same time. $MI(C, C')$ takes values between 0 and $max(H(C), H(C'))$, where $H(C)$ and $H(C')$ are the entropies of $C$ and $C'$, respectively. It reaches the maximum $max(H(C), H(C'))$ if the two sets of clusters are identical, and reaches 0 if the two sets are independent. In our experiments, we use the $NMI(C, C')$, which takes values between 0 and 1

$$NMI(C, C') = \frac{MI(C, C')}{max(H(C), H(C'))}$$

### 4.4. Results

Results of different models are presented in Table 2. In order to eliminate the evaluation bias in the experiments, 10 test runs were conducted, and the average performance as well as the standard

[4] http://lms.comp.nus.edu.sg/research/NUS-WIDE.htm.

**Table 2**
Performance of different models on real data sets. Bold performance corresponds to the best model.

| Method | AC(%) | | | | | NMI(%) | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | DBLP | Reuters | Handwritten | Cora | NUS | DBLP | Reuters | Handwritten | Cora | NUS |
| MGGM | **80.68** ± **1.74** | **51.78** ± **3.80** | **95.51** ± **0.29** | **41.15** ± **0.18** | 13.98± 0.13 | **51.48** ± **1.98** | **32.99** ± **2.08** | **91.39** ± **0.34** | **24.65** ± **0.20** | 12.66± 0.29 |
| MGGM_SV | 77.99 ± 3.42 | 50.82 ± 3.38 | 95.23 ± 0.81 | 39.76 ± 0.12 | **14.06** ± **0.20** | 48.43 ± 3.91 | 31.67 ± 2.15 | 90.71 ± 1.01 | 23.21 ± 0.08 | **12.95** ± **0.16** |
| MVPLSA | 74.43 ± 8.77 | 47.62 ± 3.44 | 72.08 ± 2.94 | 39.69 ± 0.11 | 13.84 ± 0.53 | 45.55 ± 6.14 | 29.65 ± 1.92 | 68.21 ± 1.83 | 23.18 ± 0.06 | 12.51 ± 0.29 |
| PRSC | 70.79 ± 1.03 | 48.08 ± 1.17 | 80.79 ± 1.16 | 21.48 ± 1.22 | 11.82 ± 0.28 | 37.08 ± 1.01 | 28.06 ± 1.12 | 72.26 ± 1.06 | 7.54 ± 1.11 | 10.56 ± 0.14 |
| CRSC | 70.87 ± 1.03 | 48.11 ± 1.77 | 90.42 ± 1.22 | 23.88 ± 1.17 | 13.45 ± 0.35 | 37.20 ± 1.02 | 27.75 ± 1.18 | 82.00 ± 1.29 | 10.15 ± 1.15 | 12.44 ± 0.16 |
| LTM | 76.85 ± 1.28 | 42.61 ± 3.93 | 93.10 ± 0.17 | 29.34 ± 1.66 | 12.77 ± 0.12 | 46.21 ± 1.70 | 23.81 ± 3.46 | 87.60 ± 0.20 | 15.14 ± 1.90 | 4.57 ± 0.34 |
| ConcatLTM | 73.30 ± 6.70 | 45.53 ± 1.98 | 94.28 ± 0.20 | 29.14 ± 2.03 | 13.55 ± 0.85 | 45.22 ± 4.90 | 28.13 ± 1.64 | 89.27 ± 0.29 | 15.76 ± 1.31 | 10.69 ± 0.63 |
| K-means | 40.32 ± 3.37 | 23.12 ± 7.22 | 66.56 ± 8.08 | 20.77 ± 2.45 | 13.73 ± 0.66 | 15.77 ± 4.74 | 10.95 ± 8.37 | 69.28 ± 3.57 | 8.25 ± 1.90 | 12.43 ± 0.21 |
| ConcatK-means | 38.61 ± 5.87 | 24.13 ± 7.13 | 66.37 ± 6.75 | 17.63 ± 3.09 | 12.64 ± 0.39 | 12.42 ± 6.24 | 13.14 ± 7.67 | 69.56 ± 3.58 | 7.86 ± 2.95 | 11.43 ± 0.16 |

deviation are reported. In the table, MGGM_SV denotes the best results of MGGM when only using single graph $U^t$ (5NN graph of view $f_t$) to construct the regularization term Eq. (3). LTM (K-means) denotes the best single view results when running each view using the LTM (K-means) technique. For a better comparison, we depict the average performances of each model over all data sets in Fig. 2.

From Table 2 and Fig. 2, we have the following observations:

1. One observes that MGGM achieves a better performance than the comparison approaches. Compared with the single view model LTM, our method is able to achieve an accuracy improvement by more than 11%, which validates the usefulness of data integration in clustering.
2. The methods with manifold regularization outperform the ones without manifold regularization, i.e., multi-view algorithm MGGM and MGGM_SV are superior to MVPLSA and CRSC in terms of AC and NMI values for all the data sets. Especially, MGGM is able to achieve an improvement by 6% against the MVPLSA on DBLP data set(the AC value of MGGM is 80.68% while the result of MVPLSA is 74.43%). It indicates that the manifold regularizer preserves the nonlinear information. It also implies that the geometry information in data space is crucial for clustering performance.
3. The performance of MGGM is better than MGGM_SV. This suggests that the composite manifold learned by multiple graph ensemble regularization can select the most effective graphs to approximate the intrinsic manifold. As a result, the nonlinear structure of data distribution is preserved effectively.

### 4.5. *Convergence, parameters $\lambda_1$, $\lambda_2$, the number of latent topic* Q *and nearest neighbors* p

In this subsection, we investigate the convergence of the EM-like algorithm in Algorithm 1, and whether maximizing $\mathcal{O}$ in Eq. (4) leads to a better clustering performance. In addition, the sensitivity of the selection of parameters $\lambda_1$, $\lambda_2$, the number of latent topics and nearest neighbors are also investigated. To simplify our model, let the numbers of latent topics for all attribute views be the same, i.e., $Q^1 = Q^2 = \ldots = Q^T = Q$. The number of nearest neighbors $p$ is empirically set to be 5. For parameter selection, we determine the parameter $Q$ firstly. We set $Q$ the same value as in MVPLSA algorithm. Then we fix $Q$, and search for the optimal parameters $\lambda_1$ and $\lambda_2$ simultaneously.

**Convergence** Fig. 3 shows the convergence curves of the MGGM algorithm on DBLP and Reuters data sets. The x-axis is the number of iterations. The line with square marker indicates the values of objective function in different iterations. From the figure, we can see that the algorithm converges as the iteration number increases. The other two lines shows the AC and NMI performance of MGGM on DBLP and Reuters data sets. We can see the trend that the AC and NMI of MGGM increase when the objective value $\mathcal{O}$ increases. This result indicates that the optimization of $\mathcal{O}$ improves clustering performance.

**The number of latent topic *Q*** We investigate the sensitivity of the number of latent topic $Q$. Fig. 4 shows the AC and NMI of MGGM with various $Q$ values in DBLP, Reuters, Handwritten and Cora data sets, when we fix $\lambda_1 = 300, 300, 15, 000, 300$ and $\lambda_2 = 0.8, 0.95, 0.95, 0.8$ respectively. We find that MGGM model is robust with the tuning of $Q$. We set $Q = 500, 900, 100, 2000, 1000$ for DBLP, Reuters, Handwritten, Cora and NUS data sets, respectively.

**Parameters $\lambda_1$ and $\lambda_2$** We investigate the sensitivity of the parameters $\lambda_1$ and $\lambda_2$. Fig. 5 shows the AC and NMI of MGGM with various values of $\lambda_1$ and $\lambda_2$ on DBLP, Reuters, Handwritten and Cora data sets, when we fix $Q = 500, 900, 100, 2000,$ respectively. $\lambda_1$ adjusts the weight of manifold regularizer to a reasonable value.

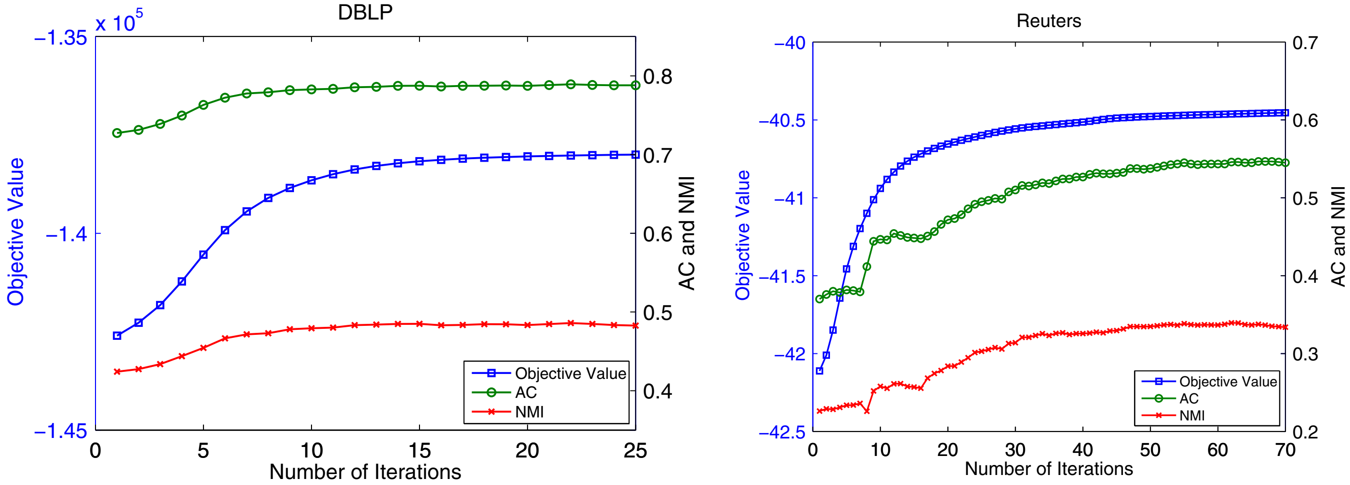

**Fig. 3.** Convergence of optimization on DBLP and Reuters data sets.

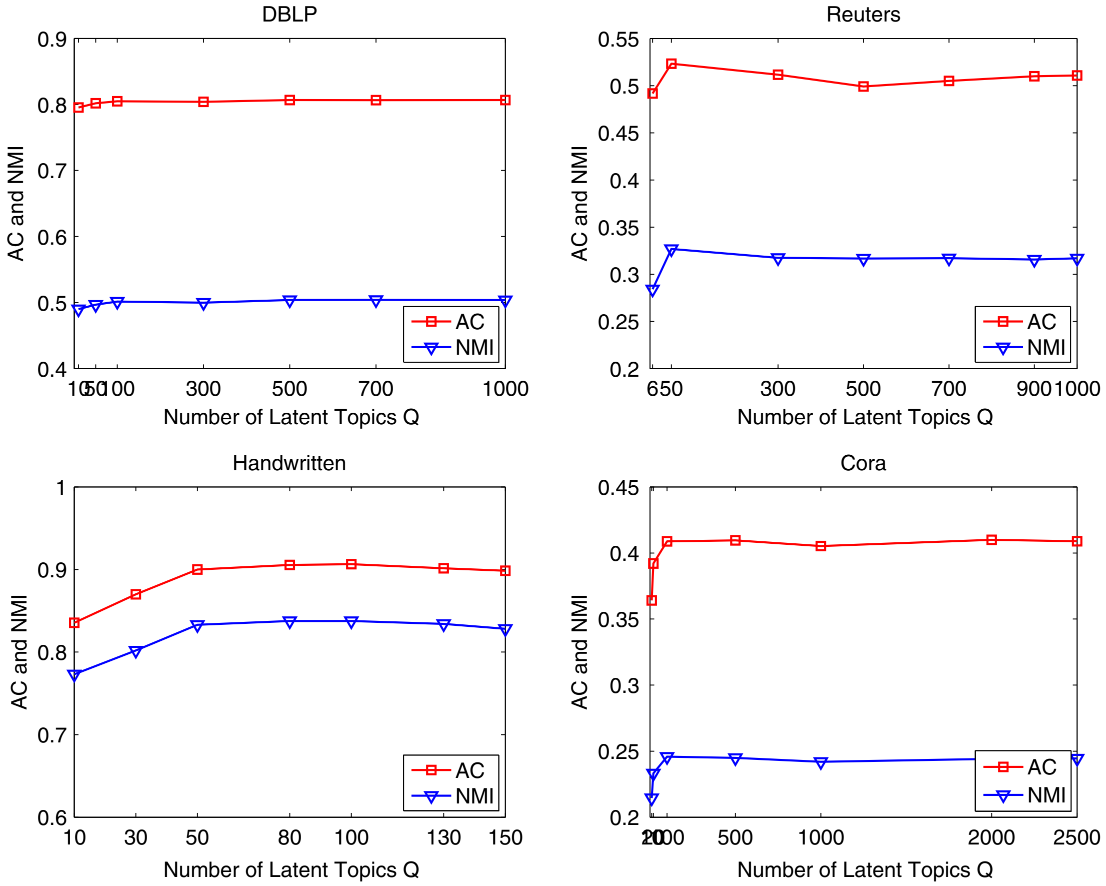

**Fig. 4.** Tuning $Q$ on the four data sets.

From the figure, we observe that the accuracy is poor when $\lambda_1$ is small, and the accuracy increases when $\lambda_1$ becomes large. This demonstrates the advantages of the manifold regularizer in the proposed method. We set $\lambda_1 = 300, 300, 15,000, 300, 0.1$ for DBLP, Reuters, Handwritten, Cora and NUS data sets, respectively. $\lambda_2 \in (0, 1)$ is a parameter to control the distribution of graph hyper-parameters $\{\mu_t\}_{t=1}^T$. According to the tuning test in Fig. 5, we set $\lambda_2 = 0.8, 0.95, 0.95, 0.8, 0.95$ for DBLP, Reuters, Handwritten, Cora and NUS data sets, respectively.

**The number of nearest neighbors *p*** We investigate the impact of the number of nearest neighbors $p$. Fig. 6 shows the AC and NMI of MGGM with various $p$ values on DBLP and Reuters data sets. As we can see, MGGM is very stable with respect to the number of nearest neighbors $p$. It achieves good performance with the p varying from 3 to 5, and the performance decreases very slightly as the p increases. We set $p = 5$ for all the data sets.

### 4.6. Initialization method

As EM-like algorithm often suffers from the local minimum problem, it is very important to study how different initialization methods can help to alleviate the issue for the proposed method.

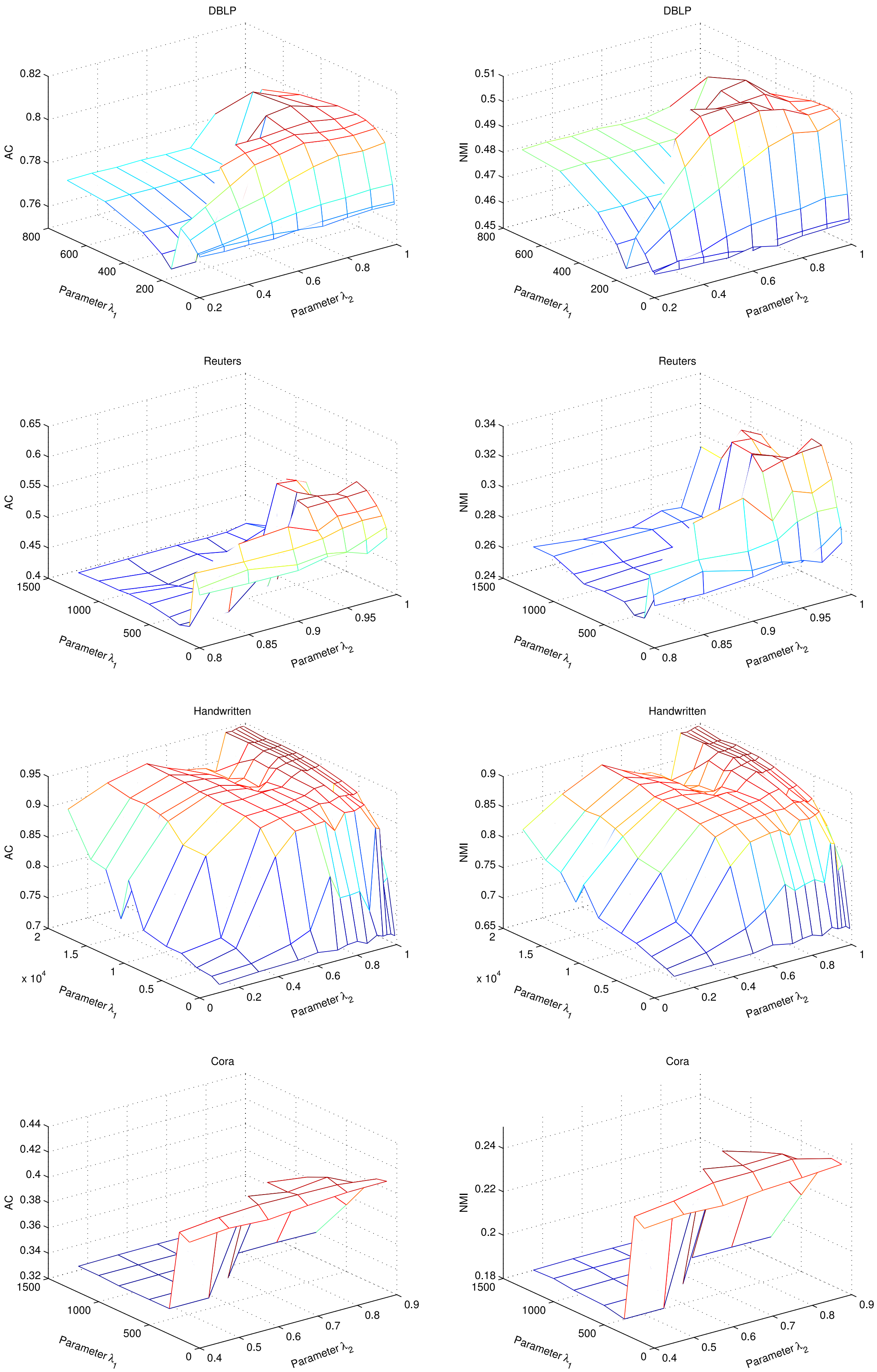

**Fig. 5.** Tuning $\lambda_1$ and $\lambda_2$ on the four data sets.

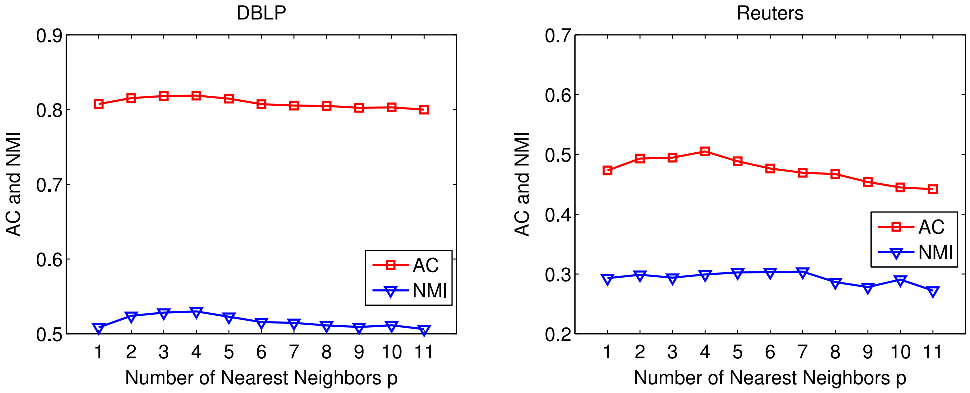

**Fig. 6.** Tuning $p$ on DBLP and Reuters data sets.

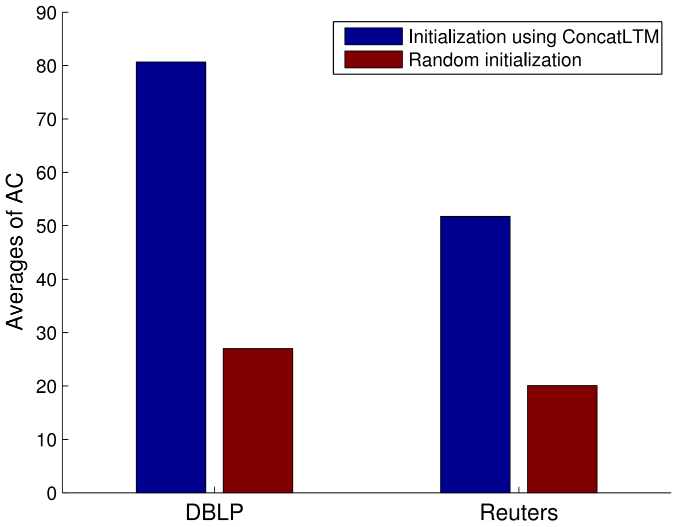

**Fig. 7.** The performance of different initial values $P(c_k|x_i)$ on DBLP and Reuters datasets.

Here we compare two methods for the initialization of $P(c_k|x_i)$. One initialization method utilizes the ConcatLTM, and the other method adopts random initialization.

Fig. 7 shows the results of ConcatLTM initialization and random initialization in terms of AC value on DBLP and Reuters datasets. From the figure, we observe that MGGM with random initialization performs worse than with ConcatLTM initialization. This indicates that it is very important to initialize properly the probability $P(c_k|x_i)$. In the experiments, we utilize ConcatLTM method to initialize $P(c_k|x_i)$.

### 4.7. Computational complexity

In this subsection, we provide a computational cost analysis of MGGM. As defined previously, let $K$, $T$, and $N$ be the number of clusters, the number of the views, and the number of data points, respectively. Let the numbers of latent topics and view features be the same for all attribute views, i.e., $Q^1 = Q^2 = \ldots = Q^T = Q$, $M_1 = M_2 = \ldots = M_T = M$. The cost of the E-step and the update of the log-likelihood $\mathcal{L}$ in MGGM are the same as that of MVPLSA, which is $O(KQTMN)$. The cost of the regularization term $\mathcal{R}$ in Eq. (3) is $O(KN^2)$. The cost of the intrinsic manifold approximation $E$ in Eqs. (10) and (11) is $O(TKN^2)$. Hence, the overall cost of MGGM is $O(KQTMN + TKN^2 + KN^2)$.

Fig. 8 shows the running time of MGGM and MVPLSA on NUS data set with the number of instances varying from 5000 to 30000. The experiment is conducted on a computer which has Intel Xeon (R) 2.10 GHz 24 processors, 128 GB memory. We can see that the time costs of MGGM and MVPLSA are comparable when the number of instances is 5000 or 10000. However, when the number reaches 20,000 or 30,000, the running time of MGGM is quite longer than MVPLSA. The reason is that the time costs of updates on regularization term $\mathcal{R}$ and intrinsic manifold approximation are quadratic to the number of instances. However, as seen in Section 4.4, MGGM delivers substantially better clustering results than MVPLSA.

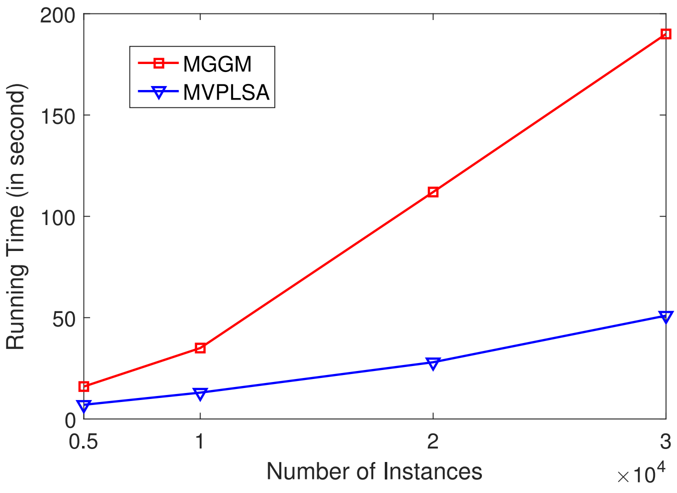

**Fig. 8.** Running time of MGGM and MVPLSA on NUS data set with different number of instances.

## 5. Conclusion

In this paper, we have introduced a novel multi-view learning method via multiple graph regularized generative model to exploit multi-view manifold information in clustering. In particular, we combined multiple previously given nearest neighbor graphs of all views to approximate the intrinsic manifold, where the manifold structure among multiple views is caught. Furthermore, a multiple graph ensemble regularization term was formulated to make the low dimensional representations effectively encoding the intrinsic manifold information. Then the manifold regularization term was incorporated into a generative model based on the Probabilistic Latent Semantic Analysis method, resulting in a unified objective function. We derived an EM-like algorithm to solve the optimization problem. Experiments on the real-world data sets have demonstrated that our algorithm can effectively maintains the underlying manifold structure of multiple views, and therefore enhance the learning performance.

## Acknowledgments

This research was supported in part by NSFC under Grant No. 61572158, 61272538, 61562027, 61602132, Shenzhen Science and Technology Program under Grant No. JCYJ20160330163900579, Shenzhen Strategic Emerging Industries Program under Grant No. JSGG20150512145714247, Social Science Planning Project of Jiangxi Province under Grant No. 15XW12, Education Department of Jiangxi Province under Grant No. GJJ150494, Harbin Institute of Technology Shenzhen Graduate School Startup Foundation No. BE45001008, and Beijing Postdoctoral Research Foundation. Raymond Y.K. Lau's work was supported by a grant from the Research Grants Council of the Hong Kong Special Administrative Region, China (Project: CityU11502115), and the Shenzhen Municipal Science and Technology R&D FundingCBasic Research Program (Project No. JCYJ20140419115614350)

## References

[1] D.G. Lowe, Distinctive image features from scale-invariant keypoints, Int. J. Comput. Vis. 60 (2) (2004) 91–110.

[2] N. Dalal, B. Triggs, Histograms of oriented gradients for human detection, in: IEEE Conference on Computer Vision and Pattern Recognition, 2005, pp. 886–893.

[3] T. Ojala, M. Pietikainen, T. Maenpaa, Multiresolution gray-scale and rotation invariant texture classification with local binary patterns, IEEE Trans. Pattern Anal. Mach. Intell. 24 (7) (2002) 971–987.

[4] J. Wu, J.M. Rehg, Where am I: place instance and category recognition using spatial pact, in: IEEE Conference on Computer Vision and Pattern Recognition, 2008, pp. 1–8.

[5] X. Song, L. Nie, L. Zhang, M. Akbari, T. Chua, Multiple social network learning and its application in volunteerism tendency prediction, in: Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2015a, pp. 213–222.

[6] X. Song, L. Nie, L. Zhang, M. Liu, T. Chua, Interest inference via structure-constrained multi-source multi-task learning, in: Proceedings of the International Joint Conference on Artificial Intelligence, 2015b, pp. 2371–2377.

[7] L. Nie, L. Zhang, Y. Yang, M. Wang, R. Hong, T. Chua, Beyond doctors: future health prediction from multimedia and multimodal observations, in: Proceedings of the 23rd annual ACM conference on multimedia conference, 2015, pp. 591–600.

[8] H. Zhang, G. Liu, T. Chow, W. Liu, Textual and visual content-based anti-phishing: a Bayesian approach, IEEE Trans. Neural Networks 22 (10) (2011) 1532–1546.

[9] Q. Wu, M. Ng, Y. Ye, X. Li, Y. Li, Multi-label collective classification via markov chain based learning method, Knowl. Based Syst. (63) (2014) 1–14.

[10] T. Hofmann, Probabilistic latent semantic indexing, in: Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 1999, pp. 50–57.

[11] D. Blei, A. Ng, M. Jordan, Latent dirichlet allocation, J. Mach. Learn. Res. 3 (2003) 993–1022.

[12] L. Zhuang, H. Gao, J. Luo, Z. Lin, Regularized semi-supervised latent dirichlet allocation for visual concept learning, Neurocomputing 119 (2013) 26–32.

[13] Y. Zhang, W. Wei, A jointly distributed semi-supervised topic model, Neurocomputing 134 (2014) 38–45.

[14] S. Wang, Y. Ye, X. Li, X. Huang, R. Lau, Semi-supervised collective classification in multi-attribute network data, Neural Process. Lett. (2016), doi:10.1007/s11063-016-9517-y.

[15] F. Zhuang, G. Karypis, X. Ning, Q. He, Z. Shi, Multi-view learning via probabilistic latent semantic analysis, Inf. Sci. 199 (2012) 20–30.

[16] M. Belkin, P. Niyogi, Laplacian eigenmaps and spectral techniques for embedding and clustering, in: Neural Information Processing Systems, 2001, pp. 586–691.

[17] T. Jin, J. Yu, J. You, K. Zeng, C. Li, Z. Yu, Low-rank matrix factorization with multiple hypergraph regularizer, Pattern Recognit. 48 (3) (2015) 1011–1022.

[18] D. Cai, Q. Mei, J. Han, C. Zhai, Modeling hidden topics on document manifold, in: Proceedings of the 17th International Conference on Information and Knowledge Management, 2008, pp. 911–920.

[19] D. Cai, X. Wang, X. He, Probabilistic dyadic data analysis with local and global consistency, in: Proceedings of the 26th Annual International Conference on Machine Learning, 2009, pp. 105–112.

[20] A. Blum, M. Tom, Combining labeled and unlabeled data with co-training, in: Proceedings of the 11th Annual Conference on Learning Theory, 1998, pp. 92–100.

[21] K. Nigam, G. Rayid, Analyzing the effectiveness and applicability of co-training, in: Proceedings of the 9th International Conference on Information and Knowledge Management, 2000, pp. 86–93.

[22] W. Wang, Z. Zhou, A new analysis of co-training, in: Proceedings of the 27th International Conference on Machine Learning, 2010, pp. 1135–1142.

[23] S. Yu, B. Krishnapuram, R. Rosales, R. Rao, Bayesian co-training, J. Mach. Learn. Res. 12 (2011) 2649–2680.

[24] A. Kumar, P. Rai, H. Daum, Co-regularized multi-view spectral clustering, in: Advances in Neural Information Processing Systems, 2011, pp. 1413–1421.

[25] Y. Fang, H. Zhang, Y. Ye, X. Li, Detecting hot topics from twitter: a multiview approach, J. Inf. Sci. 40 (5) (2014) 578–593.

[26] Q. Wu, M. Tan, X. Li, H. Min, N. Sun, Nmfe-sscc: non-negative matrix factorization ensemble for semi-supervised collective classification, Knowl. Based Syst. (89) (2015) 160–172.

[27] M. Belkin, P. Niyogi, V. Sindhwani, Manifold regularization: a geometric framework for learning from labeled and unlabeled examples, J. Mach. Learn. Res. 7 (2006) 2399–2434.

[28] Y. Jiang, J. Liu, Z. Li, H. Lu, Collaborative PLSA for multi-view clustering, in: Proceedings of the 21th International Conference on Pattern Recognition, 2012a, pp. 2997–3000.

[29] Y. Jiang, J. Liu, Z. Li, P. Li, H. Lu, Co-regularized PLSA for multi-view clustering, in: Proceedings of the 11th Asian Conference on Computer Vision, 2012b, pp. 202–213.

[30] B. Geng, D. Tao, C. Xu, L. Yang, X. Hua, Ensemble manifold regularization, IEEE Trans. Pattern Anal. Mach. Intell. 34 (6) (2012) 1227–1233.

[31] M. Karasuyama, H. Mamitsuka, Multiple graph label propagation by sparse integration, IEEE Trans. Neural Netw. Learn. Syst. 24 (12) (2013) 1999–2012.

[32] X. He, P. Niyogi, Locality preserving projections, in: Advances in Neural Information Processing Systems, 2004, pp. 145–153.

[33] Y. Jacob, L. Denoyer, P. Gallinari, Classification and annotation in social corpora using multiple relations, in: Proceedings of the 20th ACM International Conference on Information and Knowledge Management, 2011, pp. 1215–1220.

[34] H. Wang, C. Weng, J. Yuan, Multi-feature spectral clustering with minimax optimization, in: IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 4106–4113.