Multi-View Matrix Completion for Clustering with Side Information

Peng Zhao, Yuan Jiang, Zhi-Hua Zhou

National Key Laboratory for Novel Software Technology, Nanjing University
Collaborative Innovation Center of Novel Software Technology and Industrialization
Nanjing 210023, China
{zhaop,jiangy,zhouzh}@lamda.nju.edu.cn

Abstract. In many clustering applications, real world data are often collected from multiple sources or with features from multiple channels. Thus, multi-view clustering has attracted much attention during the past few years. It is noteworthy that in many situations, in addition to the data samples, there is some side information descri bing the relation between instances, such as must-links and cannot-links. Though side information has been well exploited in single-view clustering, they have rarely been studied in multi-view scenario. Considering that matrix completion has sound theoretical properties and demonstrates an excellent performance in single-view clustering, in this paper, we propose the first matrix completion based approach for multi-view clustering with side information. Instead of concatenating multiple views into a single one, we enforce the consistency of clustering results on different views as constraints for alternative optimization, and the global optimal solution is obtained since the objective function is jointly convex. The proposed Multi-View Matrix Completion (MVMC) approach exhibits impressive performance in experiments.

Keywords: Multi-View, Clustering, Matrix Completion

1 Introduction

Data clustering is one of the most important tasks in machine learning and data mining. Aiming at grouping data instances into different clusters based on the similarity, clustering has plenty of real applications, such as data summarization [9], text mining [24], bioinformatics [8], etc.

In many applications, data are collected from multiple sources or with feature from different channels. For example, the content and hyperlink information can be thought of two views for webpage dataset [3]. Another example is that the representations in various languages can be regarded as different views for multilingual information retrieval [12]. Since feature information from different views are complementary to each other, multi-view clustering dedicates to leverage information from multiple views to improve the performance of clustering.

It's noteworthy that while plenty of unsupervised clustering algorithms have been widely used, clustering with limited side (supervised) information has gradually obtained more attentions. In general, side information can be divided into two groups, instance-level and label-level. Usually, label-level one is difficult to gather. In contrast, it is often more convenient to collect instance-level information among which the pairwise constraint is one of the most common representations. Pairwise constraints are consisted of two parts: must-link($\mathcal{M}$) and cannot-link($\mathcal{C}$). A must-link (cannot-link) specifies that the pair of instances should (not) be assigned into the same cluster. Pairwise relationship occurs in a variety of applications and domains. For example, when clustering various movies, we may only know two of them should (not) be assigned into the same style which can be viewed as a must-link (cannot-link). Another example is our knowledge that two proteins always co-occur in the Database of Interacting Proteins (DIP) dataset, which can be regarded as a must-link when performing gene clustering [13]. Generally speaking, it is convenient to gather pairwise constraints along with collecting the unlabelled data. Thus, in this paper, we only consider pairwise constraints prototype side information.

Similarly, clustering with side information is also useful for data collected from multiple sources. Existing multi-view clustering approaches cannot directly handle side information properly. Admittedly, by concatenating all the features from multiple views into a single one, one can handle it with a semi-supervised clustering algorithm. However, a simple concatenation has several drawbacks. First, the dimension of concatenation feature matrices is usually high which may trigger the curse of dimensionality and result in a high computational cost. Secondly, the approach of concatenation, in fact, treats different views equally which is not appropriate since the difference between views is ignored. Thus, it's still difficult to efficiently utilize side information in multi-view clustering, due to the trade-off between diversity of feature in multiple views and consistency of side information constraints.

To address this issue, in this paper, we propose a novel clustering approach to utilize side information called **M**ulti-**V**iew **M**atrix **C**ompletion (**MVMC**). Firstly, MVMC constructs a pairwise similarity matrix S_v for the v-th view independently and cast clustering task into a matrix completion problem based on given pairwise constraints and feature information from multiple views. Then, the final pairwise similarity matrix S is learned by controlling S and S_v in different views to approach each other. The global optimal solution is obtained by projective alternative optimization since the objective function is jointly convex. Experimental results on benchmark datasets demonstrate that the proposed MVMC can efficiently utilize side information and outperform other state-of-the-art approaches. Our major contribution is the development of the first approach to tackle constrained multi-view clustering based on matrix completion.

In the following, we start with a brief review of some related work. Then, we propose our MVMC approach and examine the empirical performance of proposed method on several benchmark datasets. Finally, we conclude the paper.

2 Related Work

Multi-view learning has attracted much attention since many real world data are collected from multiple sources or intrinsically have multi-faceted feature representations. In general, various multi-view learning algorithms in different areas can be classified into three groups: 1) co-training, 2) multiple kernel learning, and 3) subspace learning [33]. Multi-view co-training constructs two learners each from one view, and then lets them to provide pseudo-labels for the other learner [3]. And some studies [28,29] show that the diversity of multiple views is the essence of co-training. Multiple kernel learning (MKL) is suitable for multi-view learning because kernels in MKL naturally correspond to different views to improve learning performance [1,15]. Subspace learning algorithms aim at obtaining a common subspace shared by multiple views and then learning models in that shared subspace [17,19].

Multi-view clustering aims at leveraging information from multiple views to improve clustering performance, various multi-view clustering algorithms have been proposed. Roughly, they can be categorized into spectral approaches, subspace approaches and late-fusion approaches. Spectral approaches extend spectral clustering [27] into multi-view data by constructing a measure of similarity between instances [18,23]. The subspace approaches assume that multiple views are generated from a common low-dimensional subspace where the representations of similar instances are close [6,30]. The late-fusion approaches learn a clustering solution from each single view, and then fuse all these intermediate outputs based on consensus [4,35]. The proposed approach in this paper belongs to the first stream.

Clustering with side information in single view scenario has been well developed. Inspired by the work proposed in [32], plenty of algorithms are proposed based on distance metric learning. For example, ITML proposed in [7] learns a metric matrix with side information based on information theory. MCCC proposed in [37] converts clustering to a matrix completion problem.

Matrix Completion (MC) problem was originally proposed for collabative filtering [10]. Assuming that the matrix to be recovered is low-rank, MC finds a matrix X that minimizes the difference with the given observation. However, it is still challenging because rank minimization problem is NP-hard. A major breakthrough in [5] states that minimizing $\mathsf{rank}(X)$, under broad conditions, can be achieved using the minimizer obtained with its convex envelope, the nuclear norm, $\|X\|_*$. In addition, [34] proposed an approach to speed up the process of MC by utilizing side information.

Due to a solid mathematical foundation of MC, it was recently exploited into clustering. For example, a graph-based clustering proposed by [11] identifies clusters from partially observed unweighted graphs via MC. In [36], a crowdsourced clustering is proposed to use the crowd information to recover a similarity metric, which can then be applied on large, growing collections. Besides, a related clustering approach proposed in [37] convert clustering into a MC problem based on side information, which performs well in single view scenario.

All these previous studies on clustering cannot efficiently handle the scene where some side information is provided for multiple views. To the best of our knowledge, this is the first study on multi-view clustering by matrix completion with side information.

3 Our Proposed MVMC Approach

In this section, the matrix completion multi-view clustering assisted with side information model is introduced. Let $\mathcal{D} = \{O_1, O_2, \cdots, O_n\}$ be n instances, and the feature of each instance is collected from m views (channels). Feature in the v-th view is denoted as $X_v = (\mathbf{x}_1^v; \mathbf{x}_2^v; \cdots; \mathbf{x}_n^v)$, where $\mathbf{x}_i^v \in \mathbb{R}^{1\times d_v}$ is the feature of O_i in the v-th view, and d_v is dimension of the v-th view. Let $\mathcal{M}$ ($\mathcal{C}$) denote the set of must-link (cannot-link) constraints, $(i,j) \in \mathcal{M}$ ($(i,j) \in \mathcal{C}$) implies O_i and O_j should (not) be assigned into the same cluster. We define $\Omega = \mathcal{M} \cup \mathcal{C}$ to represent all the pairwise constraints. Meanwhile, let r be the number of clusters.

3.1 Similarity Matrix Construction

For each view, let $\mathbf{u}_i^v \in \{0,1\}^n$ be the membership vector of the i-th cluster in the v-th view, where $\mathbf{u}_{i,j}^v = 1$ if O_j is assigned to the i-th cluster and zero, otherwise. Then the pairwise similarity matrix $S_v \in \{0,+1\}^{n\times n}$ is defined as

$$S_v = \sum\nolimits_{i=1}^{r} \mathbf{u}_i^v (\mathbf{u}_i^v)^{\mathrm{T}} \tag{1}$$

Evidently, $[S_v]_{i,j} = 1$ if O_i and O_j are assigned to the same cluster from the perspective of feature information provided in the v-th view, and zero, otherwise. Furthermore, it is easy to verify that $\mathsf{rank}(S_v) \leq r$, which implicates a low-rank property of similarity matrix.

3.2 Single-View Clustering by Matrix Completion

For a specific view (the subscribe v is omitted in this part for simplicity), finding the best data partition is equivalent to recovering the binary matrix S. Apparently, pairwise constraints are tightly associated with the similarity matrix. More specifically, $[S_v]_{i,j} = 1$ if $(i,j) \in \mathcal{M}$ and $[S_v]_{i,j} = 0$ if $(i,j) \in \mathcal{C}$ for $v = 1, \cdots, m$. Thus, clustering problem with pairwise constraints can be cast into a matrix completion problem, i.e., filling out the missing entries in binary similarity matrix S based on $\mathcal{M}$ and $\mathcal{C}$ (i.e., the partial observations, called S_{ob}) and the feature information from multiple views.

Formally, for a specific view, the binary similarity matrix S can be recovered from the following matrix completion problem,

$$\begin{aligned} \min_{S} \quad & \|S\|_* \\ \text{s.t.} \quad & \mathcal{R}_\Omega(S) = \mathcal{R}_\Omega(S_{\mathrm{ob}}) \end{aligned} \tag{2}$$

where $\|\cdot\|_*$ is nuclear norm, and $\mathcal{R}_\Omega(\cdot) : \mathbb{R}^{n\times n} \mapsto \mathbb{R}^{n\times n}$ is a linear operator which preserve the entry of S in Ω and 0 outside.

However, feature information is not utilized. To efficiently exploit feature information, let $Z = [\mathbf{z}_1, \cdots, \mathbf{z}_k]$ be the first k left singular vectors of X corresponding to the k largest singular values, where $k \geq r$. And we make an assumption to reveal the relationship between X and S:

Assumption: the cluster membership vectors $\{\mathbf{u}_i\}_{i=1}^r$ lie in the subspace of the first k left singular vectors of feature matrix $\{\mathbf{z}_i\}_{i=1}^k$.

A similar assumption is used by the spectral clustering algorithm [22], matrix completion [34] and some others. When assumption holds, i.e., $\mathbf{Span}(\mathbf{u}_1, \cdots, \mathbf{u}_r) \subseteq \mathbf{Span}(\mathbf{z}_1, \cdots, \mathbf{z}_k)$, we know that $\forall i = 1, \cdots, r, \mathbf{u}_i = Z\theta_i$, where $\theta_i \in \mathbb{R}^k$. Then the similarity matrix S can be derived as

$$S = \sum\nolimits_{i=1}^{r} \mathbf{u}_i\mathbf{u}_i^{\mathrm{T}} = \sum\nolimits_{i=1}^{r} Z\theta_i(Z\theta_i)^{\mathrm{T}} = ZMZ^{\mathrm{T}},$$

where $M = \sum_{i=1}^r \theta_i\theta_i^{\mathrm{T}} \in \mathbb{R}^{k\times k}$. Obviously, M is a symmetric positive semidefinite matrix, i.e., $M \in \mathcal{S}_+^k$, where $\mathcal{S}_+^k = \{X \in \mathbb{R}^{k\times k} | X = X^{\mathrm{T}} \text{ and } X \succeq 0\}$.

It's proved in [34] that $\|AXB\|_* = \|X\|_*$ holds when A and B are orthonormal matrices, i.e.,$\mathbf{a_i}^{\mathrm{T}}\mathbf{a_j} = \delta_{i,j}$ and $\mathbf{b_i}^{\mathrm{T}}\mathbf{b_j} = \delta_{i,j}$ for any i and j, where $\delta_{i,j}$ is the Kronecker delta function that outputs 1 if $i = j$ and 0, otherwise. Hence, $\|S\|_* = \|ZMZ^{\mathrm{T}}\|_* = \|M\|_*$.

Besides, since pairwise constraints usually express a belief rather than certainty in many cases, soft constraints are introduced. Incorporating with feature information, Eq. 2 can be reformulated as follows:

$$\min_M \|M\|_* + C\|\mathcal{R}_\Omega(ZMZ^{\mathrm{T}}) - \mathcal{R}_\Omega(S_{\mathrm{ob}})\|_F^2 \tag{3}$$

where $C > 0$ is the regularization parameter introduced to trade off between low-rank property and the consistency of recovery and given side information.

In [37], the fast stochastic subgradient descent method is adopted to solve this optimization problem. And when S has been recovered, spectral clustering algorithm is applied to find the best data partition. This single-view clustering approach is referred as Matrix Completion Constrained Clustering (MCCC).

3.3 From Single-View to Multi-View

When managing to solve the multi-view clustering problem, a simple idea to come up with is to convert multi-view features to a single one. There are two types: the first one is concatenating all the features of multiple views, and then performing semi-supervised single-view algorithms directly on the concatenation; the second one is clustering on each view independently, and selecting the best one w.r.t. the preferred performance measurement index.

Besides, for MCCC, another approach based on the late-fusion arises naturally which performs clustering with pairwise constraints in each view independently, and then concatenates results above all views to obtain final clustering results. Concretely speaking, the pairwise similarity matrix S_v in each view can

be recovered with side information and feature information independently. Then, $S_1, \cdots, S_m$ are fused into a final similarity matrix S as $S = \frac{1}{m}\sum_{v=1}^{m} S_v$. We refer to this approach as MCCC_fusion.

As the information from multiple views is usually complementary to each other, all above fail to combine feature information from multiple views efficiently nevertheless. To address this problem, we directly restrict pairwise similarity matrix S_v and learn the final S. Because the final clustering result should be consistent over all multiple views, the consistency of multiple similarity matrices S_v is enforced. To utilize multi-view feature information, we incorporate them via the assumption claimed previously. Then S_v is expanded as $Z_v M_v Z_v^{\mathrm{T}}$, where $M_v \in \mathbb{R}^{k\times k}$ and $Z_v = [\mathbf{z}_1^v, \cdots, \mathbf{z}_k^v]$, the first k left singular vectors of feature in the v-th view X_v. In fact, k is able to vary over different views. However, it does not make difference to the essence of the problem. Thus, we set k in various views the same in the following.

It's noteworthy to mention that the original nuclear norm term $\|S_v\|_*$ or $\|M_v\|_*$ is non-smooth, which implies that it is inevitable to adopt sub-gradient or proximal approach. Fortunately, since M_v is constrained as a positive semi-definite matrix, then $\|M_v\|_* = \sum_{i=1}^{k} |\sigma_i| = \sum_{i=1}^{k} \mathrm{eig}_i = \mathrm{tr}(M_v)$, where σ_i and eig_i are the i-th singular value and eigenvalue of M, respectively. Thus, the optimization problem can be formulated as follows:

$$\begin{aligned} \min_{S,\{M_v\}_{v=1}^m} & \sum_{v=1}^{m} \left(\mathrm{tr}(M_v) + C_1 \|\mathcal{R}_\Omega(Z_v M_v Z_v^{\mathrm{T}} - S_{\mathrm{ob}})\|_F^2 + C_2 \|Z_v M_v Z_v^{\mathrm{T}} - S\|_F^2 \right) \\ \text{s.t.} \quad & 0 \le S_{i,j} \le 1,\ \forall i,j \in \{1,\cdots,N\}, \\ & M_v \in \mathcal{S}_+^k, \quad v = 1,\cdots,m. \end{aligned} \tag{4}$$

where $C_1, C_2 > 0$ are two regularization parameters. The optimization object function is consisted of three terms, the first two terms are generated from single-view matrix completion, and the last term measures the difference among S_v from multiple views. If we split the Frobenius norm into the square sum of entries, in fact, it is the entry-variance of multiple similarity matrix.

After converting the non-smooth term $\|M_v\|_*$ to a smooth term $\mathrm{tr}(M_v)$, projected gradient descend is adopted which is pretty easy to implement.

3.4 Optimization

In Eq. 4, the constraint regions are convex sets and the objective function is jointly convex w.r.t S and $\{M_v\}_{v=1}^m$. Thus, we developed an iterative algorithm to find the global optimal solution. Firstly, $\{M_v\}_{v=1}^m$ and S are initialized by the given observation S_{ob}. Then the following two steps are repeated until convergence: minimizing $\{M_v\}_{v=1}^m$ over S; and then minimizing S over $\{M_v\}_{v=1}^m$.

1) **Initialization** $\{M_v\}_{v=1}^m$ and S:
Since the observation S_{ob} is given, then $\{M_v\}_{v=1}^m$ and S can be initialized as follows:

$$M_v = Z_v^{\mathrm{T}} S_{\mathrm{ob}} Z_v, \quad S = S_{\mathrm{ob}}. \tag{5}$$

Because each pairwise constraint corresponds to a pair entries in S_{ob} and the value of each entry in S_{ob} is 0/1, this initialization meets the constraint condition in Eq. 4.

2) **Minimizing** object over S with fixed $\{M_v\}_{v=1}^m$:

$$\hat{S} = \underset{0 \le S_{i,j} \le 1}{\arg\min} C_2 \sum\nolimits_{v=1}^{m} \|S - Z_v M_v Z_v^{\mathrm{T}}\|_F^2$$

Obviously, this sub-problem has a closed-form solution,

$$\hat{S} = \mathbf{Proj}_1 \left(\frac{1}{m} \sum_{v=1}^{m} Z_v M_v Z_v^{\mathrm{T}} \right) \tag{6}$$

where $\mathbf{Proj}_1(\cdot)$ is defined as

$$[\mathbf{Proj}_1(X)]_{i,j} = \begin{cases} 0 & \text{if } X_{i,j} < 0; \\ 1 & \text{if } X_{i,j} > 1; \\ X_{i,j} & \text{otherwise.} \end{cases} \tag{7}$$

3) **Minimizing** object over $M_v (v = 1, \cdots, m)$ with fixed S:
Obviously, when fixing S, each M_v can be solved independently. The objective function of sub-problem is

$$\mathcal{L}(M_v) = \mathrm{tr}(M_v) + C_1 \|\mathcal{R}_\Omega(Z_v M_v Z_v^{\mathrm{T}} - S_{\text{ob}})\|_F^2 + C_2 \|Z_v M_v Z_v^{\mathrm{T}} - S\|_F^2 \tag{8}$$

And the optimal solution of sub-problem is

$$\hat{M_v} = \underset{M_v \in \mathcal{S}_+^k}{\arg\min} \mathcal{L}(M_v) \tag{9}$$

$\mathcal{L}(M_v)$ is differential and its gradient $\nabla\mathcal{L}(M_v)$ is

$$\nabla\mathcal{L}(M_v) = I + 2C_1 Z_v^{\mathrm{T}}(\mathcal{R}_\Omega(Z_v M_v Z_v^{\mathrm{T}} - S_{\text{ob}}))Z_v + 2C_2 Z_v^{\mathrm{T}}(Z_v M_v Z_v^{\mathrm{T}} - S)Z_v$$

Besides, it's easy to verify that $\nabla\mathcal{L}(M_v)$ is Lipschitz continuous with constant $L = 2(C_1\|Z_v\|_F^4 + C_2)$. The projective gradient descend method is adopted, the update sequence is defined as:

$$M_v^{(\ell+1)} = \mathbf{Proj}_2 \left(M_v^{(\ell)} - \eta \nabla\mathcal{L}(M_v^{(\ell)}) \right). \tag{10}$$

where η is chosen as $1/L$ for a linear convergence referring to [21]. $\mathbf{Proj}_2$ is a operator projecting M_v back to semi-definite positive cone $\mathcal{S}_+^k$ defined as:

$$\mathbf{Proj}_2(X) = U \max(\sigma, 0) U^{\mathrm{T}} \tag{11}$$

where U and σ correspond to the eigenvectors and eigenvalues of X.

When obtaining final pairwise similarity matrix S, we apply spectral clustering algorithms [27] on S to find the best data partition. The proposed clustering approach above is referred as MVMC (Multi-View Matrix Completion), which is summarized in Algorithm 1.

Convergence Analysis: Because objective function in Eq. 4 is jointly convex with a convex constraints region, Algorithm 1 converges to a global optima.

Algorithm 1 MVMC (Multi-View Matrix Completion)

Input:

1) Multi-view feature: $\mathcal{X} = \{X_v\}_{v=1}^m$, where $X_v = (\boldsymbol{x}_1; \boldsymbol{x}_2; \cdots; \boldsymbol{x}_n)_v \in \mathbb{R}^{N \times d_v}$;
2) The set of pairwise constraints: $\Omega = \mathcal{M} \cup \mathcal{C}$;
3) Regularization parameters: C_1 and C_2;
4) The number of clusters: r.

Output:

Pairwise similarity matrix S and clustering results.

Initialize S and $\{M_v\}_{v=1}^m$ by Eq. 5;
repeat
  Fixing $\{M_v\}_{v=1}^m$ to optimize the objective, update S by Eq. 6 ;
  Fixing S to optimize the objective, update $\{M_v\}_{v=1}^m$ by Eq. 10;
until objective function in Eq. 4 converges.
Performing spectral clustering on S to obtain final clustering results.

4 Experiment

In this section, we compare the performance of proposed approach **MVMC** with several baseline methods over different real world datasets. The baseline methods are representations from two paradigms: multi-view clustering and semi-supervised clustering. The muti-view clustering algorithms are (a) **Co-Reg**, the co-regularized spectral clustering [14], (b) **MKKM**, the multi-view kernel k-means algorithm [26], (c) **RMSC**, robust multi-view spectral clustering based on Markov chain. [31]; The semi-supervised clustering algorithms are (d) **ITML**, the information theoretic metric learning algorithm [7], (e) **MCCC**, matrix completion based constraint clustering [37]. Since there are two ways, i.e., concatenation and best view selection, for semi-supervised algorithms to handle multiple views, ITML and MCCC are separated as ITML_best , ITML_concat and MCCC_best , MCCC_concat. Besides, as we mentioned before, MCCC_fusion is also added into comparison.

4.1 General Experiment Settings

Datasets: The *WebKB* dataset [3] has been widely used in multi-view learning, which contains webpages collected from four universities: Cornell, Texas, Washington and Wisconsin. The webpages are distributed over five clusters and described by two views: the content and citation view. BBCSport consists of 2 views from news articles [14]. The Reuters dataset [2] is built from the Reuters Multilingual test collection, multi-view information is created from different languages, i.e., English, French, German, Italian and Spanish [2]. Statistics of these datasets are summarized in Table 1.

Parameter Settings: There are two regularization parameters C_1 and C_2, cross-validation is applied because of the existence of side information [36,37]. To choose an appropriate k, a trade-off need to be balanced between computational efficiency and violation of assumption. It's noteworthy to mention that k in

Table 1. Statistics of six datasets, the first four datasets are subsets of *WebKB* and d_v denotes the dimension of the v-th view of datasets.

Data Set	#size	#view	#cluster	dimension of each view $d_v(v = 1, \cdots, m)$
Cornell	195	2	5	1703, 195
Texas	187	2	5	1703, 187
Washington	230	2	5	1703, 230
Wisconsin	265	2	5	1703, 265
BBCSport	737	2	5	3183, 3208
Reuters	1600	5	6	2000 for each

each view, in fact, can be different. However, in our experiments, k is chosen as $\min(100, d_v)$ for convenience, where d_v is the dimension of the v-th view.

Side Information: In our experiments, we follow the typical routine of experiments with side information [38,39], where each pairwise constraint is generated by randomly selecting a pair of samples. A must-link constraint is formed if they belong to the same cluster, and cannot-link, otherwise. RATIO is used to measure quantity of side information, i.e. $|\Omega| = \text{RATIO} \cdot n^2$. We vary RATIO from $[0.01, 0.02, \cdots, 0.1]$.

Evaluation: In all the experiments, to evaluate the effectiveness of the proposed approach, we use six different and widely-used criteria to measure clustering performances: F-score, precision, recall, the normalized mutual information (NMI) [25], adjusted rand index(Adj-RI) [20] and average entropy. Note that all the other criteria except for average entropy lie in interval $[0, 1]$, and a higher value indicates a better performance. Meanwhile, a lower average entropy means a more competitive performance.

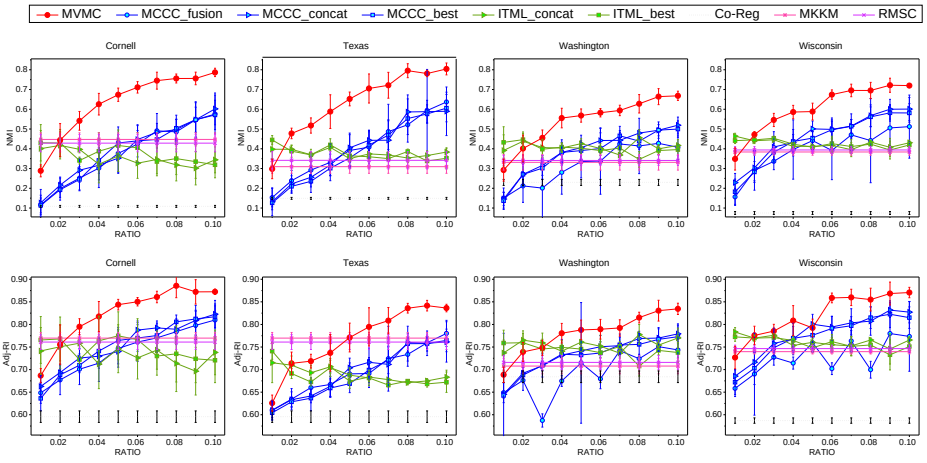

Fig. 1. Comparations of clustering performance with other approaches on *WebKB* dataset (with 4 subsets) w.r.t. NMI and Adjust Rand-Index (the higher, the better). RATIO is used to measure amount of side information which varies from 0.01 to 0.1. On each dataset, 10 test runs were conducted and the average performance as well as standard deviation are presented.

4.2 Results

Due to the space limitation, we only present Figure 1 and Table 2 in experiments part to demonstrate MVMC approach. Figure 1 summarizes the results w.r.t NMI and Adj-RI on *WebKB* dataset. 10 test runs were conducted and the average performance as well as standard deviation are presented.

From Figure 1 we can see that, for all four datasets, firstly, the performance of proposed approach MVMC is gradually better as RATIO increases, which means MVMC can handle side information efficiently. Secondly, comparing with the multi-view clustering, when the side information is extremely scarce, the behavior of MVMC is relatively poor. However, MVMC is able to demonstrate a much better performance with plenty of side information. The reason is that matrix completion cannot give a satisfying recovery with an exceedingly small amount of side information, and when a relatively large amount is given, MVMC can take advantage of side information while multi-view clustering approaches cannot. Thirdly, comparing with the semi-supervised clustering, MVMC almost outperforms all the time especially along with the growth of RATIO. This phenomenon implicates that simple concatenation or late-fusion does not leverage information from multiple views. By exploiting different views via minimizing variance of similarity matrix, MVMC is validated to be effective.

Table 2. Comparations of clustering performance on BBC (abbrv. for BBCSport), REU (abbrv. for Reuters) w.r.t six criteria (except that a lower entropy indicates a better performance, the others lie in [0,1] and the higher, the better). The number of pairwise constraints is chosen as 5,000. On each dataset, 10 test runs were conducted and the average performance as well as standard deviation are presented. Besides, • (○) indicates that MVMC is significantly better (worse) than the compared method (paired t-tests at 95% significance level).

dataset	method	Fscore↑	Precision↑	Recall↑	NMI↑	Adj-RI↑	Avg Entropy↓
BBC	CoReg	.385±.002•	.285±.003•	.606±.011•	.173±.005•	.090±.005•	1.881±0.010•
	MKKM	.745±.013•	.774±.020•	.719±.023•	.661±.016•	.669±.016•	0.724±0.046•
	RMSC	.452±.017•	.472±.017•	.434±.021•	.297±.020•	.290±.020•	1.527±0.043•
	ITML_concat	.681±.072•	.633±.097•	.742±.048•	.624±.056•	.568±.104•	0.882±0.153•
	ITML_best	.560±.065•	.452±.072•	.740±.041•	.518±.054•	.373±.100•	1.198±0.127•
	MCCC_concat	.823±.070•	.783±.085•	.869±.052•	.805±.053•	.772±.088•	0.476±0.135•
	MCCC_best	.768±.057•	.721±.066•	.823±.047•	.750±.047•	.702±.072•	0.609±0.112•
	MCCC_fusion	.861±.088•	.822±.109•	.906±.063•	.867±.053•	.822±.112•	0.336±0.142•
	MVMC	**.990±.003**	**.989±.003**	**.991±.003**	**.982±.005**	**.987±.004**	**0.040±0.011**
REU	CoReg	.346±.001•	.316±.004•	.384±.006•	.274±.002•	.200±.003•	1.902±0.008•
	MKKM	.345±.002•	.319±.015•	.377±.020•	.274±.006•	.201±.009•	1.897±0.028•
	RMSC	.369±.008	.342±.018•	.402±.020•	.303±.017•	.231±.014•	1.825±0.054•
	ITML_concat	.360±.010•	.294±.017•	.466±.022•	.294±.021•	.197±.018•	1.895±0.064•
	ITML_best	.362±.015•	.298±.015•	.464±.033	.305±.020•	.201±.020•	1.866±0.051•
	MCCC_concat	.351±.033•	.359±.034•	.343±.033•	.246±.038•	.218±.041•	1.918±0.097•
	MCCC_best	.334±.029•	.338±.029•	.331±.030•	.231±.033•	.200±.035•	1.976±0.083•
	MCCC_fusion	.459±.051•	.489±.028•	.437±.071•	.377±.071•	.193±.036•	1.496±0.081•
	MVMC	**.528±.030**	**.559±.024**	**.499±.037**	**.472±.027**	**.427±.038**	**1.294±0.061**

Table 2 summarizes the results w.r.t all the six criteria on BBCSport and Reuters. The number of pairwise constraints $|\Omega|$ is both chosen as 5,000. We can see that, MVMC demonstrates a surprisingly better performance than all the other approaches on almost all criteria. It's noteworthy to mention that the randomly sampled pairwise constraints, in fact, only accounts for about 0.9% and 0.2% for BBCSport and Reuters, respectively. It is encouraging that, with such a limited side information, MVMC can still yield a satisfying performance.

5 Conclusions

In this paper, we present MVMC, which is possibly the first attempt to efficiently handle multi-view clustering with side information based on matrix completion. By constructing similarity matrix for each view, we cast clustering into a matrix completion problem. Instead of concatenating multi-views into a single view, we enforce the consistency of clustering results on different views as constraints for alternative optimization, and the global optimal solution is obtained. The proposed MVMC approach exhibits impressive performance in experiments. Studying partial multi-view clustering [16] where each view suffers from some missing features assisted by side information will be an interesting future issue.

Acknowledgement. This research was supported by the National Science Foundation of China (61673201, 61333014)

References

1. Bach, F.R., Lanckriet, G.R., Jordan, M.I.: Multiple kernel learning, conic duality, and the smo algorithm. In: ICML (2004)
2. Bisson, G., Grimal, C.: Co-clustering of multi-view datasets: A parallelizable approach. In: ICDM. pp. 828–833 (2012)
3. Blum, A., Mitchell, T.: Combining labeled and unlabeled data with co-training. In: COLT. pp. 92–100 (1998)
4. Bruno, E., Marchand-Maillet, S.: Multiview clustering: A late fusion approach using latent models. In: SIGIR. pp. 736–737 (2009)
5. Candès, E.J., Recht, B.: Exact matrix completion via convex optimization. Commun. ACM 55(6), 111–119 (2012)
6. Chaudhuri, K., Kakade, S.M., Livescu, K., Sridharan, K.: Multi-view clustering via canonical correlation analysis. In: ICML. pp. 129–136 (2009)
7. Davis, J.V., Kulis, B., Jain, P., Sra, S., Dhillon, I.S.: Information-theoretic metric learning. In: ICML. pp. 209–216 (2007)
8. Dougherty, E.R., Barrera, J., Brun, M., Kim, S., Cesar, R.M., Chen, Y., Bittner, M., Trent, J.M.: Inference from clustering with application to gene-expression microarrays. Computational Biology 9(1), 105–126 (2002)
9. Ganti, V., Gehrke, J., Ramakrishnan, R.: Cactus—clustering categorical data using summaries. In: KDD. pp. 73–83 (1999)
10. Goldberg, D., Nichols, D., Oki, B.M., Terry, D.: Using collaborative filtering to weave an information tapestry. Commun. ACM 35(12), 61–70 (1992)
11. Jalali, A., Chen, Y., Sanghavi, S., Xu, H.: Clustering partially observed graphs via convex optimization. In: ICML. pp. 1001–1008 (2011)
12. Kim, Y.M., Amini, M.R., Goutte, C., Gallinari, P.: Multi-view clustering of multilingual documents. In: SIGIR. pp. 821–822 (2010)
13. Kulis, B., Basu, S., Dhillon, I.S., Mooney, R.J.: Semi-supervised graph clustering: a kernel approach. In: ICML. pp. 457–464 (2005)
14. Kumar, A., Rai, P., Daume, H.: Co-regularized multi-view spectral clustering. In: NIPS 24. pp. 1413–1421 (2011)
15. Lanckriet, G.R., Cristianini, N., Bartlett, P., Ghaoui, L.E., Jordan, M.I.: Learning the kernel matrix with semidefinite programming. JMLR 5, 27–72 (2004)

16. Li, S.Y., Jiang, Y., Zhou, Z.H.: Partial multi-view clustering. In: AAAI (2014)
17. Li, S., Shao, M., Fu, Y.: Multi-view low-rank analysis for outlier detection. In: SDM. pp. 748–756 (2015)
18. Li, Y., Nie, F., Huang, H., Huang, J.: Large-scale multi-view spectral clustering via bipartite graph. In: AAAI. pp. 2750–2756 (2015)
19. Liu, M., Luo, Y., Tao, D., Xu, C., Wen, Y.: Low-rank multi-view learning in matrix completion for multi-label image classification. In: AAAI. pp. 2778–2784 (2015)
20. Manning, C.D., Raghavan, P., Schütze, H.: Introduction to information retrieval (2008)
21. Nesterov, Y.: Introductory Lectures on Convex Optimization: A Basic Course. Springer (2013)
22. Ng, A.Y., Jordan, M.I., Weiss, Y., et al.: On spectral clustering: Analysis and an algorithm. NIPS 15 pp. 849–856 (2002)
23. de Sa, V.R.: Spectral clustering with two views. In: ICML workshop on Learning with Multiple Views. pp. 20–27 (2005)
24. Steinbach, M., Karypis, G., Kumar, V., et al.: A comparison of document clustering techniques. In: KDD workshop on text mining. vol. 400, pp. 525–526 (2000)
25. Strehl, A., Ghosh, J.: Cluster ensembles - a knowledge reuse framework for combining multiple partitions. JMLR 3, 583–617 (2002)
26. Tzortzis, G., Likas, A.: Kernel-based weighted multi-view clustering. In: ICDM. pp. 675–684 (2012)
27. Von Luxburg, U.: A tutorial on spectral clustering. Statistics and computing 17(4), 395–416 (2007)
28. Wang, W., Zhou, Z.H.: Analyzing co-training style algorithms. In: ECML. pp. 454–465 (2007)
29. Wang, W., Zhou, Z.H.: Multi-view active learning in the non-realizable case. In: NIPS 23. pp. 2388–2396 (2010)
30. Wang, Y., Zhang, W., Wu, L., Lin, X., Fang, M., Pan, S.: Iterative views agreement: An iterative low-rank based structured optimization method to multi-view spectral clustering. In: IJCAI. pp. 2153–2159 (2016)
31. Xia, R., Pan, Y., Du, L., Yin, J.: Robust multi-view spectral clustering via low-rank and sparse decomposition. In: AAAI. pp. 2149–2155 (2014)
32. Xing, E.P., Ng, A.Y., Jordan, M.I., Russell, S.: Distance metric learning with application to clustering with side information. NIPS 15 pp. 505–512 (2003)
33. Xu, C., Tao, D., Xu, C.: A survey on multi-view learning. arXiv preprint arXiv:1304.5634 (2013)
34. Xu, M., Jin, R., Zhou, Z.H.: Speedup matrix completion with side information: Application to multi-label learning. In: NIPS 27. pp. 2301–2309 (2013)
35. Ye, H., Zhan, D., Miao, Y., Jiang, Y., Zhou, Z.H.: Rank consistency based multi-view learning: A privacy-preserving approach. In: CIKM. pp. 991–1000 (2015)
36. Yi, J., Jin, R., Jain, A.K., Jain, S., Yang, T.: Semi-crowdsourced clustering: Generalizing crowd labeling by robust distance metric learning. In: NIPS 25. pp. 1772–1780 (2012)
37. Yi, J., Zhang, L., Jin, R., Qian, Q., Jain, A.K.: Semi-supervised clustering by input pattern assisted pairwise similarity matrix completion. In: ICML. pp. 1400–1408 (2013)
38. Zeng, H., Cheung, Y.: Semi-supervised maximum margin clustering with pairwise constraints. TKDE 24(5), 926–939 (2012)
39. Zhang, X., Zong, L., Liu, X., Yu, H.: Constrained nmf-based multi-view clustering on unmapped data. In: AAAI. pp. 3174–3180 (2015)