Partial Multi-View Clustering*

Shao-Yuan Li Yuan Jiang Zhi-Hua Zhou
National Key Laboratory for Novel Software Technology
Nanjing University, Nanjing 210023, China
{lisy,jiangy,zhouzh}@lamda.nju.edu.cn

Abstract

Real data are often with multiple modalities or coming from multiple channels, while multi-view clustering provides a natural formulation for generating clusters from such data. Previous studies assumed that each example appears in all views, or at least there is one view containing all examples. In real tasks, however, it is often the case that every view suffers from the missing of some data and therefore results in many *partial examples*, i.e., examples with some views missing. In this paper, we present possibly the first study on partial multi-view clustering. Our proposed approach, PVC, works by establishing a latent subspace where the instances corresponding to the same example in different views are close to each other, and similar instances (belonging to different examples) in the same view should be well grouped. Experiments on two-view data demonstrate the advantages of our proposed approach.

Introduction

In many tasks, data are with multiple modalities or coming from multiple channels. Multi-view clustering provides a natural formulation for clustering with such data, where each view corresponds to one modality or information channel. For example, in web page grouping, the web page content and its linkage information can be regarded as two views; in web image retrieval, the visual information of images and their textual tags can be regarded as two views.

Here, each view is actually a feature set. Formally, given a data set $D = \{(\mathbf{x}_i^1, \mathbf{x}_i^2, \cdots, \mathbf{x}_i^v, \mathbf{y}_i), i = 1, \ldots, N\}$, where $X_i = (\mathbf{x}_i^1, \mathbf{x}_i^2, \cdots, \mathbf{x}_i^v)$ is the ith example, and $\mathbf{y}_i \in \mathcal{Y}$ is its cluster label, $\mathbf{x}_i^j$ is the instance of the ith example in the jth view. Multi-view clustering aims at clustering X_i into it's corresponding cluster $\mathbf{y}_i$, where the number of clusters is usually prefixed. Note that in multi-view clustering, each view is a set of features, and the 'feature grouping' information is exploited. This makes it significantly different from multi-dimensional clustering where each dimension is a single feature.

A number of approaches have been proposed for multi-view clustering. Roughly speaking, they can be categorized into spectral approaches and subspace approaches. The former are extensions from single-view spectral clustering approaches with the help of some similarity measures (de Sa 2005; Zhou and Burges 2007; Kumar and Daumé 2011; Kumar, Rai, and Daumé 2011), whereas the latter generally try to identify a latent subspace across the multiple views (Hardoon and Shawe-taylor 2009; Chaudhuri et al. 2009; Shon et al. 2006; Salzmann et al. 2010; White et al. 2012; Guo 2013; Liu et al. 2013).

It is noteworthy that previous studies on multi-view clustering either assumed that all examples have full information in all views, or that there exists at least one view which contains all the examples, i.e., there exists some $g \in \{1, ..v\}$ such that all examples $\mathbf{x}_1^g, \mathbf{x}_2^g, \cdots, \mathbf{x}_N^g$ are available. In real tasks, however, it's often the case that every view suffers from some missing information, which results in many *partial examples*. For example, in disease diagnosis, the blood test and the neuroimage can be regarded as two views of each individual, and it often occurs that some individuals would only like to take one test; in bi-lingual documents grouping, the two languages can be seen as two views and many documents have only single language part; in speakers grouping according to audio-visual appearance, the audio and visual can be seen as two views of a speaker and some speakers have only audio or visual information.

If we want to apply existing multi-view clustering approaches to partial examples, we can either remove the examples that suffer from missing information, or preprocess the partial examples by first filling in the missing information. The first strategy clearly contradicts with the target of clustering which aims at distributing all examples to their corresponding cluster, whereas our experiments show that the second strategy is not really a good choice either. In this paper, we propose the PVC (Partial multi-View Clustering) approach to handle such data, which works based on NMF by learning a latent subspace where the instances belonging to the same example are close to each other and similar instances from the same view should be well grouped. Following (Piyush et al. 2010; Eaton, Desjardins, and Jacob 2010), we focus on two-view data in this paper and experimental results validate the advantages of our PVC approach.

In the following we start with a brief review of some related work. Then, we propose our PVC approach and report the experimental results. Finally, we conclude the paper .

*This research was supported by NSFC (61273301, 61333014).

Related Work

The exploitation of multiple modalities or information from multiple channels have attracted much attention. For example, Blum and Mitchell (1998) treated the the web page content and its linkage as two information sources in web page classification. Li *et al.*(2009) regarded the visual information and surrounding texts of images as two information sources in image retrieval. In speech recognition, the audio and visual sources are used together as two sources (Ngiam et al. 2011; Potamianos et al. 2004).

Multi-view learning provides a natural formulation for learning with such kind of multi-source data. Many studies focused on the exploitation of the information compatibility of different views for learning with unlabeled data. Co-training (Blum and Mitchell 1998) is possibly the most famous representative of multi-view semi-supervised learning. It constructs two learners each from one view, and then lets them to provide pseudo-labels for the other learner. Some studies (Wang and Zhou 2007; 2010b) disclose that such kind of disagreement-based approaches (Zhou and Li 2010) do not really need the existence of multiple views, and the diversity among the learners are the real essence. However, with suitable multiple views, semi-supervised learning with even a single labeled example has been shown to be possible (Zhou, Zhan, and Yang 2007). Moreover, the existence of multiple views enables exponential sample complexity improvement for active learning in non-realizable case (Wang and Zhou 2010a). Utilizing information from multiple sources in many specific studies like active learning (Zhou, Chen, and Dai 2006), multi-task learning (He and Lawrence 2011), multi-instance learning (Zhang, He, and Lawrence 2013),et al. has also been found very useful.

Expecting a better clustering result by exploiting information from multiple views, various multi-view clustering approaches have been proposed. Roughly, they can be categorized into spectral approaches and subspace approaches. With the help of some similarity measure between examples, spectral clustering (von Luxburg 2007) has been extended to multi-view data. de Sa(2005) constructs a bipartite similarity graph and propose their spectral clustering algorithm based on the minimizing-disagreement idea. Zhou and Burges (2007) define a mixture of Markov chains on similarity graph of each view and generalize spectral clustering to multiple views. Kumar and Daume(2011) propose a co-training approach where the similarity matrix of one view is constrained by the spectral embedding of the other view. Kumar *et al.* (2011) further propose two co-regularization approaches to enforce the clustering hypotheses on different views to agree with each other.

Assuming the multiple views are generated from one common subspace, subspace approaches aim at learning a latent intrinsic subspace where the representations of instances in each view are close for similar examples. By finding two projections of two set variables such that their correlations in the projected space are maximized, CCA (Hotelling 1936) is one of the earliest technique applied on two view data. It was further generalized to kernel variant Kernel CCA (Hardoon and Shawe-taylor 2009) and data with more than two views (Chaudhuri et al. 2009; Hardoon, Szedmak, and Shawe-taylor 2004). Factorizing each view as the linear combination of shared latent representation and view-specific parts, several approaches have been proposed. Salzmann *et al.*(2010) introduced an orthogonality constraint of view-private parts to penalize the redundant between views; White *et al.* (2012) and Guo (2013) formulated the subspace learning as a convex optimization problem with a sparsity norm regularization. Being one of the most effective techniques used in single-view latent subspace based clustering, non-negative matrix factorization(NMF) (Lee and Seung 1999) was recently exploited by multi-view setting (Greene and Cunningham 2009; Akata, Thurau, and Bauckhage 2011; Liu et al. 2013) and show good performance.

All these previous studies on multi-view clustering assumed that all examples present in all views. Piyush *et al.*(2010) proposed an approach which uses one view's kernel matrix as the similarity matrix and complete the missing view's kernel using Laplacian regularization. It is clear that this approach requires that there exists at least one view containing all the examples. Eaton *et al.*(2010) considered semi-supervised multi-view clustering where a set of *must-link* and *cannot-link* constraints are provided, whereas our approach proposed in this paper can be extended to semi-supervised setting with a similar strategy.

To the best of our knowledge, we are presenting the first multi-view clustering approach which is able to handle the case that each view suffers from missing information and there are many partial examples.

Our Proposed PVC Approach

For the convenience of discussion, assume that we are handling two-view data, i.e., given a data set of N instances $D = \{(\mathbf{X}, \mathbf{y})\} = \{(X^1, X^2, \mathbf{y})\} = \{(X_i, \mathbf{y}_i), i = 1, \ldots, N\}$, where $X_i = (\mathbf{x}_i^1, \mathbf{x}_i^2)$ is the ith example of two views, and $\mathbf{y}_i \in \mathcal{Y}$ is its cluster label. $\mathbf{x}_i^1 \in \mathbb{R}^{1\times d_1}(\mathbf{x}_i^2 \in \mathbb{R}^{1\times d_2})$ is the instance of the ith example in the first(second)view of dimension $d_1(d_2)$. In the partial view setting, a partial view example set $\hat{\mathbf{X}} = \{\hat{\mathbf{X}}^{(1,2)}, \hat{\mathbf{X}}^{(1)}, \hat{\mathbf{X}}^{(2)})\}$ instead of $\mathbf{X}$ is given, where $\hat{\mathbf{X}}^{(1,2)}$, $\hat{\mathbf{X}}^{(1)}$, $\hat{\mathbf{X}}^{(2)}$ denotes the examples present and only present in both views, the first view, and the second view, respectively. The goal of partial view clustering, same as normal multi-view clustering, is to cluster the examples into their corresponding clusters. We assume that the number of clusters are prefixed by users.

The Formulation

We assume that the number of examples present and only present in both views, the first view, and the second view is c, m and n, i.e., $\hat{\mathbf{X}}^{(1,2)}$=$[(\mathbf{x}_1^1, \mathbf{x}_1^2); \ldots ; (\mathbf{x}_c^1, \mathbf{x}_c^2)] \in \mathbb{R}^{c\times(d_1+d_2)}$, $\hat{\mathbf{X}}^{(1)}$=$[\mathbf{x}_{c+1}^1; \ldots ; \mathbf{x}_{c+m}^1] \in \mathbb{R}^{m\times d_1}$, and $\hat{\mathbf{X}}^{(2)}$=$[\mathbf{x}_{c+m+1}^2; \ldots ; \mathbf{x}_{c+m+n}^2] \in \mathbb{R}^{n\times d_2}$, $N = c + m + n$.

In the partial view setting, $\hat{\mathbf{X}}^{(1,2)}, \hat{\mathbf{X}}^{(1)}, \hat{\mathbf{X}}^{(2)}$ are represented by heterogeneous features of dimension $(d_1 + d_2), d_1, d_2$, which makes their clustering not so direct. But examining the problem from view perspective, in each individual view, their instances are sharing the same feature space; and the two different views are bridged by the shared common examples. If we can learn a common latent subspace for the two views, like previous subspace-based multi-view learning, where different instances belonging to the same example are close, while at the same time for each view, the cluster property that similar instances are grouped is well captured in the latent subspace, then we can directly exploit any standard single view clustering method, like k-means, in this subspace, and do not need to consider whether to complete the partial view examples.

Let $\hat{\mathbf{X}}^{(1,2)}= [\mathbf{X}_c^{(1)}, \mathbf{X}_c^{(2)}]$ be composed of instances $\mathbf{X}_c^{(1)} \in \mathbb{R}^{c\times d_1}$, $\mathbf{X}_c^{(2)} \in \mathbb{R}^{c\times d_2}$ coming from two views. We now have the instances of each view denoted as $\bar{\mathbf{X}}^{(1)} = [\mathbf{X}_c^{(1)}; \hat{\mathbf{X}}^{(1)}] \in \mathbb{R}^{(c+m)\times d_1}$, $\bar{\mathbf{X}}^{(2)} = [\mathbf{X}_c^{(2)}; \hat{\mathbf{X}}^{(2)}] \in \mathbb{R}^{(c+n)\times d_2}$. As we deal with webpage clustering task in the experiment, for such text data, non-negative matrix factorization(NMF) (Lee and Seung 1999) has been an highly effective technique used in single-view clustering tasks, which actually assumes that the observed instances are generated by additive combination of an underlying set of hidden basis. The best fit between NMF and our problem is that, the latent subspace(hidden basis), what we are trying to learn, is just exactly what NMF aims at finding. Incorporating the NMF technique into our problem, for each view, its latent subspace learning can be formulated as:

$$\min_{U^{(1)}\geq 0, \bar{P}^{(1)}\geq 0} \|\bar{\mathbf{X}}^{(1)} - \bar{P}^{(1)}U^{(1)}\|_F^2 + \lambda\Omega(\bar{P}^{(1)}), \tag{1}$$

$$\min_{U^{(2)}\geq 0, \bar{P}^{(2)}\geq 0} \|\bar{\mathbf{X}}^{(2)} - \bar{P}^{(2)}U^{(2)}\|_F^2 + \lambda\Omega(\bar{P}^{(2)}), \tag{2}$$

where $U^{(1)} \in \mathbb{R}^{t\times d_1}$ and $U^{(2)} \in \mathbb{R}^{t\times d_2}$ are the basis matrix for each view's latent space, and $\bar{P}^{(1)} = [P_c^{(1)}; \hat{P}^{(1)}] \in \mathbb{R}^{(c+m)\times t}$, $\bar{P}^{(2)} = [P_c^{(2)}; \hat{P}^{(2)}] \in \mathbb{R}^{(c+n)\times t}$ are the latent representation of instances in the latent space. The same latent space dimension t is shared between the two views. And λ is the tradeoff parameter for the regularization term $\Omega(P)$. By Eq.1 and Eq.2, the latent space basis U and corresponding instance latent representation P are simultaneously learned to minimize the instance reconstruction error, which enforces all instances from each individual view to be smoothly gathered in the latent space.

So far, the latent space are learned independently for each view. For the partial view setting, for examples available in both views $\mathbf{X}_c^{(1)}, \mathbf{X}_c^{(2)}$, their latent representation $P_c^{(1)}, P_c^{(2)}$ should also be close. Combining this idea and Eq.1, Eq.2, by enforcing $P_c^{(1)} = P_c^{(2)} = P_c$, we have the following minimization problem

$$\begin{aligned}
\min_{\{U^{(v)}, \bar{P}^{(v)}\}_{v=1}^{2}} O \equiv & \left\| \begin{bmatrix} \mathbf{X}_c^{(1)} \\ \hat{\mathbf{X}}^{(1)} \end{bmatrix} - \begin{bmatrix} P_c \\ \hat{P}^{(1)} \end{bmatrix} U^{(1)} \right\|_F^2 + \lambda \|\bar{P}^{(1)}\|_1 \\
+ & \left\| \begin{bmatrix} \mathbf{X}_c^{(2)} \\ \hat{\mathbf{X}}^{(2)} \end{bmatrix} - \begin{bmatrix} P_c \\ \hat{P}^{(2)} \end{bmatrix} U^{(2)} \right\|_F^2 + \lambda \|\bar{P}^{(2)}\|_1 \\
\text{s.t.} \quad & U^{(1)} \geq 0, U^{(2)} \geq 0, \\
& \bar{P}^{(1)} \geq 0, \bar{P}^{(2)} \geq 0, \qquad (3)
\end{aligned}$$

where $\bar{P}^{(1)} = [P_c; \hat{P}^{(1)}]$, $\bar{P}^{(2)} = [P_c; \hat{P}^{(2)}]$ are the latent representation of instances for two views. Now we can have the homogeneous feature representation for all examples as $P = [P_c; \hat{P}^{(1)}; \hat{P}^{(2)}] \in \mathbb{R}^{(c+m+n)\times t}$, whether they are originally partial or not. Any standard clustering approach can be applied on such representation. Note that Eq.3 is different from previous subspace based multi-view clustering approaches, which either requires $\bar{P}^{(1)}$ and $\bar{P}^{(2)}$ are the same, or do not require $\bar{P}^{(1)}$ and $\bar{P}^{(2)}$ to share any common part. In Eq.3, $\bar{P}^{(1)}$ and $\bar{P}^{(2)}$ share one same part P_c and at the same time has their own individual part $\hat{P}^{(1)}$, $\hat{P}^{(2)}$. Moreover, learned by using all available instances of each view, the individual basis matrix $U^{(1)}$ and $U^{(2)}$ are connected by the common P_c. Lasso is used for $\Omega(P)$ in this work as one of the mostly often used regularization for text analysis.

The Algorithm

To address the optimization in Eq.3, which is convex in latent representations $P_c, \hat{P}^{(v)}$ given the basis matrix $U^{(v)}$ and vice versa, but not jointly convex in both, we propose an iterative update procedure and prove its convergence. Firstly, the basis matrices are initialized by the initialization step and then the following two steps are repeated until convergence:1) minimizing O over $P_c, \hat{P}^{(v)}$ with fixed $U^{(v)}$; and 2) minimizing O over $U^{(v)}$ with fixed $P_c, \hat{P}^{(v)}$.

Initialization: Since the iterative AO procedure's efficiency is greatly affected by the initialization step, in this paper, we learn the initial value of $U^{(v)}$ rather than random allocation:

$$\begin{aligned}
\min_{U^{(1)}, U^{(2)}, P_c} \quad & O_{init} \equiv \|\mathbf{X}_c^{(1)} - P_c U^{(1)}\|_F^2 + \|\mathbf{X}_c^{(2)} - P_c U^{(2)}\|_F^2 \\
& + \lambda \|P_c\|_1 \qquad (4) \\
\text{s.t.} \quad & U^{(1)} \geq 0, U^{(2)} \geq 0, P_c \geq 0.
\end{aligned}$$

It can be seen that $U^{(1)}, U^{(2)}$ are essentially initialized by applying traditional NMF multi-view clustering methods on examples without partial views. This initialization is also solved by iterative AO optimization. At each iteration, O_{init} is minimized alternatively over P_c and $U^{(v)}$. Fixing P_c, $U^{(1)}$ and $U^{(2)}$ can be independently optimized by :

$$\min_{U^{(1)}\geq 0} O_{init}(U^{(1)}) \equiv \|\mathbf{X}_c^{(1)} - P_c U^{(1)}\|_F^2, \tag{5}$$

$$\min_{U^{(2)}\geq 0} O_{init}(U^{(2)}) \equiv \|\mathbf{X}_c^{(2)} - P_c U^{(2)}\|_F^2. \tag{6}$$

Fixing $U^{(1)}, U^{(2)}$, P_c is optimized by:

$$\begin{aligned}
\min_{P_c \geq 0} O_{init}(P_c) \equiv & \|\mathbf{X}_c^{(1)} - P_c U^{(1)}\|_F^2 + \|\mathbf{X}_c^{(2)} - P_c U^{(2)}\|_F^2 \\
& + \lambda \|P_c\|_1. \qquad (7)
\end{aligned}$$

1). Minimizing O over $P_c, \hat{P}^{(v)}$ with fixed $U^{(v)}$ Given the basis matrix $U^{(v)}$ for each view, the computation of $P_c, \hat{P}^{(v)}$ do not depend on each other. Therefore, Eq. 3 reduces to:

$$\min_{\hat{P}^{(1)}\geq 0} O(\hat{P}^{(1)}) \equiv \|\hat{\mathbf{X}}^{(1)} - \hat{P}^{(1)}U^{(1)}\|_F^2 + \lambda\|\hat{P}^{(1)}\|_1, \tag{8}$$

$$\min_{\hat{P}^{(2)}\geq 0} O(\hat{P}^{(2)}) \equiv \|\hat{\mathbf{X}}^{(2)} - \hat{P}^{(2)}U^{(2)}\|_F^2 + \lambda\|\hat{P}^{(2)}\|_1, \tag{9}$$

$$\begin{aligned}
\min_{P_c \geq 0} O(P_c) \equiv & \|\mathbf{X}_c^{(1)} - P_c U^{(1)}\|_F^2 + \|\mathbf{X}_c^{(2)} - P_c U^{(2)}\|_F^2 \\
& + \lambda \|P_c\|_1. \qquad (10)
\end{aligned}$$

Noting the same formulation of Eq. 10 as Eq. 7, so at the first iteration, P_c has already been obtained from initialization.

Algorithm 1 The PVC Approach

```
Input: data set $\{\hat{\mathbf{X}}^{(1,2)}, \hat{\mathbf{X}}^{(1)}, \hat{\mathbf{X}}^{(2)}\}$; parameters: $\{t, \lambda\}$
Output: Basis matrices $\{U^{(1)}, U^{(2)}\}$; latent representations $\{P_c, \hat{P}^{(1)}, \hat{P}^{(2)}\}$
Algorithm:
1: Initialize $U^{(1)}, U^{(2)}$ by Eq.4.
2: repeat
3:    Fixing $U^{(1)}, U^{(2)}$, update $\hat{P}^{(1)}, \hat{P}^{(2)}, P_c$ by Eq.8-10 .
4:    Fixing $\hat{P}^{(1)}, \hat{P}^{(2)}, P_c$, update $U^{(1)}, U^{(2)}$ by Eq.11-12.
5: until Eq. 3 converges
```

2). Minimizing O over $U^{(v)}$ with fixed $P_c, \hat{P}^{(v)}$ Given $P_c, \hat{P}^{(v)}$, the latent representations for instances of each view can be obtained as $\bar{P}^{(1)} = [P_c; \hat{P}^{(1)}]$ and $\bar{P}^{(2)} = [P_c; \hat{P}^{(2)}]$, minimizing O over $U^{(v)}$ now independently reduces to:

$$\min_{U^{(1)} \geq 0} O(U^{(1)}) \equiv \|\bar{\mathbf{X}}^{(1)} - \bar{P}^{(1)} U^{(1)}\|_F^2, \tag{11}$$

$$\min_{U^{(2)} \geq 0} O(U^{(2)}) \equiv \|\bar{\mathbf{X}}^{(2)} - \bar{P}^{(2)} U^{(2)}\|_F^2. \tag{12}$$

To solve the optimization problem in Eq. 5-Eq. 12, which are (lasso regularized) NMF with one factor fixed, in this paper, we employ the GCD(greedy coordinate descent) approach proposed by Hsieh and Dhillon (2011), which is about 10 times faster than cyclic coordinate descent scheme and proved to converge. To get a stable sparsity tradeoff parameter for different data set, λ is normalized by the data size during the optimization, i.e., for Eq.8,9,10(7), λ is respectively timed by coefficient $(d_1 + d_2)/t, d_1/t, d_2/t$. Algorithm 1 summarizes the PVC approach.

Convergence Property We prove that our PVC approach in Algorithm 1 converges to a local minima solution.

Theorem 1 *The objective function value of Eq. 3 is nonincreasing under the optimization procedure in Algorithm 1.*

Lemma 1 (Hsieh and Dhillon 2011) *For least squares NMF, if a basis matrix, latent representation pair sequence $\{(U_j, P_j)\}$ is generated by GCD, then every limit point of this sequence is a stationary point.*

Proof of Theorem 1: To prove Theorem 1, we only need to prove that the objective function value of Eq. 3 is nonincreasing after each step in line 3, 4. With fixed $U^{(1)}, U^{(2)}$, the objective function value of Eq. 3 with respect to $\hat{P}^{(1)}, \hat{P}^{(2)}, P_c$ equals the sum of the objective function value of Eq. 8- 10. With fixed $\hat{P}^{(1)}, \hat{P}^{(2)}, P_c$, the objective value of Eq. 3 with respect to $U^{(1)}, U^{(2)}$ equals the sum of Eq. 11- 12. By Lemma 1, the objective function value of Eq. 8-Eq. 12 are guaranteed to converge to some local minima. So the objective function value of Eq. 3 is guaranteed to nonincrease after each step in line 3, 4.■

Experiment

In this section, we compare the PVC approach with six baseline methods over four webpage data sets.

Data Sets: The WebKB data set[1] (Blum and Mitchell 1998) has been widely used in multi-view learning (Guo 2013; Zhang and Huan 2012), which contains webpages collected from four universities: *Cornell, Texas, Washington and Wisconsin*. The webpages are distributed over five classes: *student, project, course, staff* and *faculty* and described by two views: the *content* view and the *citation* view. Each webpage is described by 1703 words in the *content* view, and the number of citation links between other pages in the *citation* view. Statistics of the data sets are summarized in Table 1.

Table 1: Statics of four webpage data sets. # size, # view, # cluster denotes the number of examples, views, and clusters of each data set. # featCon and # featCit denotes the number of features in the content view and citation view.

Data Set	# size	# view	# cluster	# featCon	# featCit
Cornell	195	2	5	1703	195
Texas	187	2	5	1703	187
Washington	230	2	5	1703	230
Winsconsin	265	2	5	1703	265

To simulate the partial view setting, we randomly select a fraction of webpages to be *partial examples*, i.e., they are described by either the *content* or the *citation* view, but not both, and the remaining ones appear in both the *content* and the *citation* view. Following the formulation section, two different settings for the *partial examples* are considered in this paper, the first is: both $m > 0$ and $n > 0$, i.e., both views suffer from missing information about webpages; the second is: either $m = 0$ or $n = 0$, i.e., at least one view, either the *content* or the *citation* view, does not suffer from any missing information, being the *'complete view'*. Note the second setting is what Piyush *et al.*(2010) have considered, which is a special case of our partial view setting. To simplify the assignment of *partial examples* to their corresponding view for the first setting, we evenly distribute them to the two views in the experiment. And for the second setting, two tasks, either with the *content* or the *citation* view being the *'complete view'* are conducted for each data set. Each time we randomly select 10% to 90% examples, with 20% as interval, as *partial examples*. Such process is repeated 10 times and the average and standard deviation results are recorded.

Baseline Algorithms: Two subspace based multi-view clustering methods CCA, ConvexSub and four spectral multi-view clustering methods are included as baselines.[2]

CCA: We use the LSCCA package[3] implementation of CCA and kernel CCA to first extract the latent representation and then perform k-means. The clustering results for whichever gives the best performance is recorded.

ConvexSub: The subspace-based multi-view clustering method developed by (Guo 2013).

MinDisSC: The multi-view spectral clustering method developed by (de Sa 2005).

CentroidSC: The centroid multi-view spectral method developed by (Kumar, Rai, and Daumé 2011).

PairwiseSC: The pairwise multi-view spectral clustering method developed by (Kumar, Rai, and Daumé 2011).

[1] http://membres-liglab.imag.fr/grimal/data.html

[2] Except for CCA(Kernel CCA), we use the implementation codes for other baselines provided by their authors.

[3] http://www.public.asu.edu/ jye02/Software/CCA/index.html

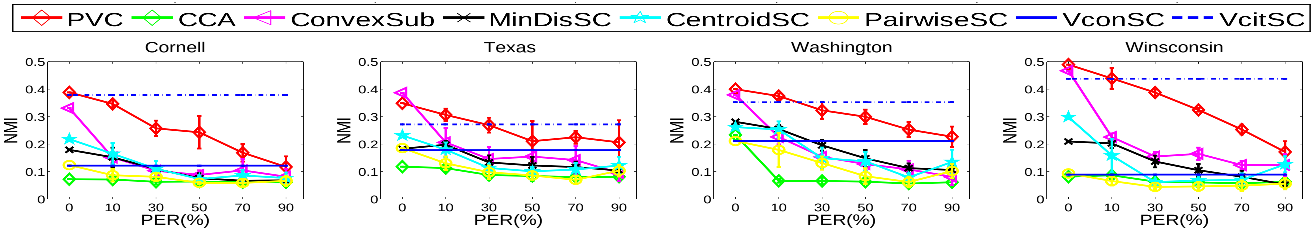

Figure 1: The NMI(the higher, the better) results for the four data sets when both views miss information about examples. Partial examples are evenly distributed to the *content* and *citation* views. *PER*(partial example ratio) is the ratio of partial examples.

SingleView: Regarding the clustering result within each single view when the partial example ratio becomes zeros, i.e., all examples are complete examples, as upper bound for partial view clustering, we run spectral clustering (von Luxburg 2007) respectively on the *content* view and *citation* view, denoted as **VconSC** and **VcitSC**.

Notice that for the first *partial example* setting, the approach of (Piyush et al. 2010) can't be used to help the baselines to handle partial examples, for a fair comparison, in this experiments they are facilitated with the ALM (Augmented Lagrange Multipliers)[4] (Lin, Chen, and Ma 2010) matrix completion method by first filling in the missing information of the partial examples. For the second *partial example* setting, the approach of (Piyush et al. 2010) is used to facilitate kernel CCA and the four spectral methods.

For kernel based methods kernel CCA and the four spectral baselines, Gaussian kernel is used and the width parameter is set as the median pairwise distance between instances. Parameters of baselines are well tuned to achieve the best performance (consistent with the values recommended in their corresponding papers). For PVC, the latent dimension t is set as the number of clusters, the default choice for most subpace approaches. The sparsity tradeoff parameter λ is fixed as 0.01 for all data sets. We run 20 rounds iterations for PVC in the experiment, which is empirically shown enough to almost converge and achieve good enough clustering performance. The parameter study and convergence process are later discussed. The k-means algorithm is performed on the latent representation of examples to get the final clustering result for our PVC approach and CCA. Since k-means might be sensitive to initializations, we run it for 20 times and record the mean result.

The normalized mutual information(NMI) clustering evaluation measure (Kumar, Rai, and Daumé 2011; Guo 2013) is used in this paper. Results for the two different settings are shown in Figure 1 and Figure 2 -3 respectively.

Results

Figure 1 summarize the results (the average and standard deviation) for the fist setting, i.e., both the *content* and *citation* view suffer from missing information, on the four data sets for *PER*(partial example ratio) varying from 10% to 90% with 20% as interval. Taken as the upper bound of each comparison methods, the results of all methods when *PER* is 0%, i.e., all examples are described by both the *content* and *citation* view, are also recorded.

From Figure 1 we can see that, for all four data sets, comparing the two spectral clustering results on each complete single view, VconSC always performs better than VcitSC, which coincides with our expectation that the *content* view is more informative than the *citation*view. When the partial example ratio *PER* is 0%, the data set actually comes to the traditional multi-view data that each example appears in all views. In such case, PVC is also able to perform better than most baselines except for ConvexSub, which specifically designed a convex optimization algorithm and was validated to get better performance than direct AO optimization by their experiment. As the partial example ratio *PER* varies from 10% to 90%, our PVC approach always achieves the best performance among all methods. Although matrix completion is powerful in recovering missing values for low rank matrix, ALM seems less effective in the view missing case; which may be caused by that while the data are missing block-wise for the partial view data setting, matrix completion requires the missing locations to be random.

Possibly the most inspiring thing about PVC we can get from Figure 1 is that, even when *PER* equals 90%, i.e., as large as 90% examples are partial, PVC is still able to get better performance than VcitSC on the single complete *citation* view. Moreover, for *Cornell* with 0% *PER*, out of expectation, the specifically designed multi-view methods even perform worse than the single view approach VconSC. These two points may suggest an interesting problem of view selection for examples in multiple view clustering.

Similar results for PVC and baselines for the second special partial example setting, i.e., either the *citation* or the *content*view is *'complete'*, are also obtained in Figure 2-3. As the approach of (Piyush et al. 2010) greatly depends on the similarity matrix constructed from the complete view, it would only be helpful when the complete view is informative enough, like in Figure 3, the *complete* content view. Otherwise, the performance is not satisfiable. However, even the complete view is less informative, like in Figure 2 the complete *citation* view, our PVC still performs much better than all baselines as long as PER is no more than 70%.

Parameter Study In the above, parameters are fixed for PVC. Here we explore the effect of the sparsity trade off parameter λ to clustering performance. Tuning λ from $\{10^{-6}, 10^{-5}, \ldots, 10^{-1}, 1\}$ for three different partial example ratios 30%, 50% and 70%, we only present the results for

[4] http://perception.csl.illinois.edu/matrix-rank/home.html

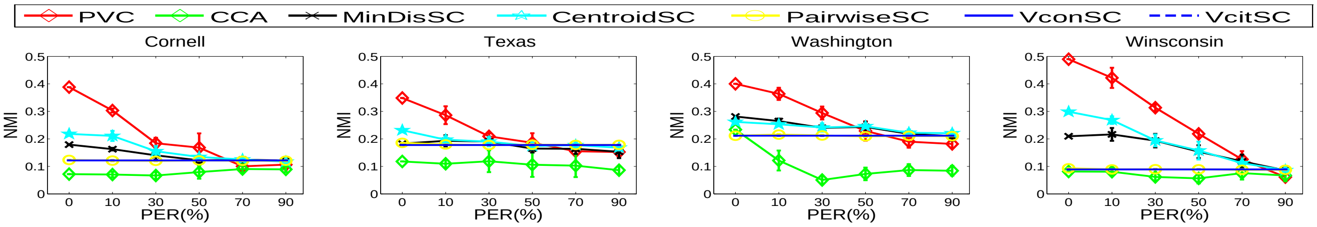

Figure 2: The NMI(the higher, the better) results on the four data sets when the *citation* view is complete and only the *content* view misses information about examples. *PER*(partial example ratio) is the ratio of partial examples.

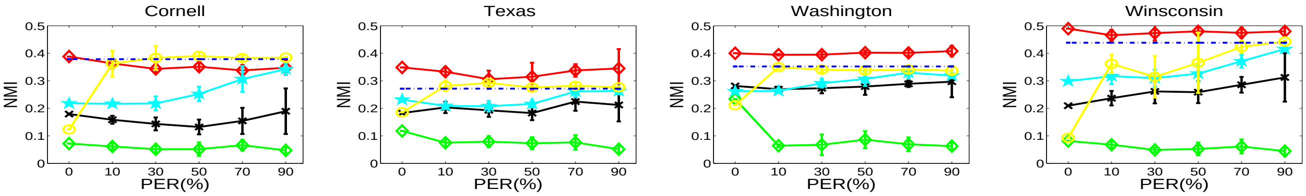

Figure 3: The NMI(the higher, the better) results on the four data sets when the *content* view is complete and only the *citation* view misses information about examples. *PER*(partial example ratio) is the ratio of partial examples.

the first partial example setting, i.e., both views suffer from missing information, in Figure 4 due to space limitation.

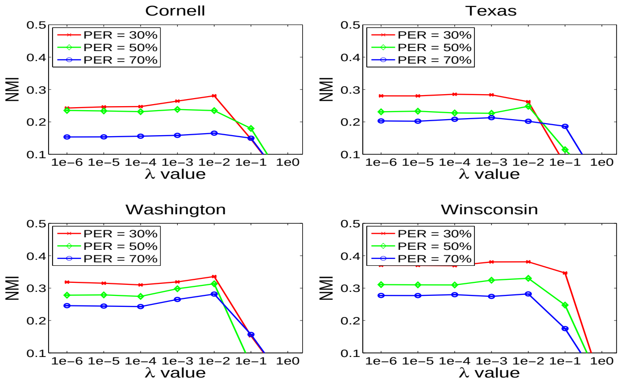

Figure 4: Influence of the sparsity parameter λ on the four data sets with three different partial example ratio PER.

From Figure 4 we can see that PVC achieves stably good performance when λ is around 10^{-2}, which is the value we used in the experiment. It's not strange that all data sets share the same preference for parameter λ since we have normalize out the data set size in each optimization step.

Convergence Study We have proved in previous section that the PVC objective function is convergent. Due to space limitation, we here only show the convergence curve and corresponding NMI performance for the first partial example setting, i.e., both views suffer from missing information, with partial example ratios 70% in Figure 5. It can be seen that the objective function value monotonically decreases as the iteration round increases. Though it takes a lot of rounds to converge, the GCD in each iteration runs very fast and actually 20 round is enough to get good clustering results.

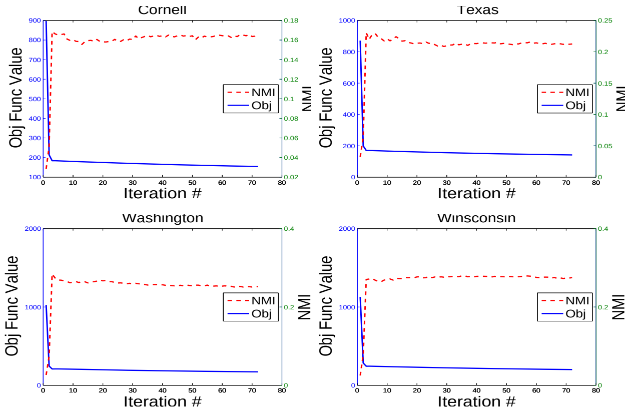

Figure 5: Objective function value convergence curve and corresponding NMI performance curve *vs* number of iterations of PVC with partial example ratio $PER = 70\%$.

Conclusion

In this paper, we present possibly the first attempt to deal with multi-view clustering with partial views where each view may suffer from missing of some data. Based on NMF, our proposed PVC approach establishes a latent subspace where the instances corresponding to the same example in different views are close to each other, and the instances (belonging to different examples) in the same view are well grouped. Experimental results validate the effectiveness of PVC. In this paper we focus on two-view data. The number of views in real tasks is often smaller than five, and it is not difficult to extend and apply our approach. Extending to more views, however, will suffer from computational problem. In the future we will study how to extend this subspace based partial view learning idea to data with more views, and to nonlinear latent subspace cases.

References

Akata, Z.; Thurau, C.; and Bauckhage, C. 2011. Non-negative matrix factorization in multimodality data for segmentation and label prediction. In *Proceedings of the 16th Computer Vision Winter Workshop*.

Blum, A., and Mitchell, T. 1998. Combining labeled and unlabeled data with co-training. In *Proceedings of the 11th Annual Conference on Computational Learning Theory*, 92–100.

Chaudhuri, K.; Kakade, S.; Livescu, K.; and Sridharan, K. 2009. Multi-view clustering via canonical correlation analysis. In *Proceedings of the 26th International Conference on Machine Learning*, 129–136.

de Sa, V. 2005. Spectral clustering with two views. In *Proceedings of the International Conference on Machine Learning Workshop on Learning with Multiple Views*.

Eaton, E.; Desjardins, M.; and Jacob, S. 2010. Multi-view clustering with constraint propagation for learning with an incomplete mapping between views. In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, 389–398.

Greene, D., and Cunningham, P. 2009. A matrix factorization approach for integrating multiple data views. In *Proceedings of the 2009 European Conference on Machine Learning and Knowledge Discovery in Databases I*, 423–438.

Guo, Y. 2013. Convex subspace representation learning from multi-view data. In *Proceedings of 27th AAAI Conference on Artificial Intelligence*, 387–393.

Hardoon, D., and Shawe-taylor, J. 2009. Convergence analysis of kernel canonical correlation analysis: theory and practice. *Machine Learning* 74:23–38.

Hardoon, D.; Szedmak, S.; and Shawe-taylor, J. 2004. Canonical correlation analysis: an overview with application to learning methods. *Neural Computation* 16(12):2639–2664.

He, J., and Lawrence, R. 2011. A graph-based framework for multi-task multi-view learning. In *Proceedings of the 28th International Conference on Machine Learning*, 25–32.

Hotelling, H. 1936. Relations between two sets of variates. *Biometrika* 28(3/4):321–377.

Hsieh, C., and Dhillon, I. 2011. Fast coordinate descent methods with variable selection for non-negative matrix factorization. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1064–1072.

Kumar, A., and Daumé, H. 2011. A co-training approach for multi-view spectral clustering. In *Proceedings of the 28th International Conference on Machine Learning*, 393–400.

Kumar, A.; Rai, P.; and Daumé, H. 2011. Co-regularized multi-view spectral clustering. In *Advances in Neural Information Processing Systems*, 1413–1421.

Lee, D., and Seung, H. 1999. Learning the parts of objects by non-negative matrix factorization. *Nature* 401(6755):788–791.

Li, M.; Xue, X.-B.; and Zhou, Z.-H. 2009. Exploiting multi-modal interactions: a unified framework. In *Proceedings of the 21st International Joint Conference on Artificial Intelligence*, 1120–1125.

Lin, Z.; Chen, M.; and Ma, Y. 2010. The augmented lagrange multiplier method for exact recovery of corrupted low-rank matrices. *http://arxiv.org/abs/1009.5055*.

Liu, J.; Wang, C.; Gao, J.; and Han, J. 2013. Multi-view clustering via joint nonnegative matrix factorization. In *Proceedings of the 13th SIAM International Conference on Data Mining*, 252–260.

Ngiam, J.; Khosla, A.; Kim, M.; Nam, J.; Lee, H.; and Ng, A. 2011. Multimodal deep learning. In *Proceedings of the 28th International Conference on Machine Learning*, 689–696.

Piyush, R.; Anusua, T.; Daumé, H.; and Scott, L. 2010. Multiview clustering with incomplete views.

Potamianos, G.; Neti, C.; Luettin, J.; and Matthews, I. 2004. Audio-visual automatic speech recognition: An overview. *Issues in Visual and Audio-Visual Speech Processing* 22:23.

Salzmann, M.; Ek, C.; Urtasun, R.; and Darrell, T. 2010. Factorized orthogonal latent spaces. In *Proceedings of the 13th International Conference on Artificial Intelligence and Statistics*, 701–708.

Shon, A.; Grochow, K.; Hertzmann, A.; and Rao, R. 2006. Learning shared latent structure for image synthesis and robotic imitation. In *Advances in Neural Information Processing Systems*, 1233–1240.

von Luxburg, U. 2007. A tutorial on spectral clustering. *Statistics and Computing* 17(4):395–416.

Wang, W., and Zhou, Z.-H. 2007. Analyzing co-training style algorithms. In *Proceedings of the 18th European Conference on Machine Learning*, 454–465.

Wang, W., and Zhou, Z.-H. 2010a. Multi-view active learning in the non-realizable case. In *Advances in Neural Information Processing Systems*, 2388–2396.

Wang, W., and Zhou, Z.-H. 2010b. A new analysis of co-training. In *Proceedings of the 27th International Conference on Machine Learning*, 1135–1142.

White, M.; Zhang, X.; Schuurmans, D.; and Yu, Y. 2012. Convex multi-view subspace learning. In *Advances in Neural Information Processing Systems*, 1682–1690.

Zhang, J., and Huan, J. 2012. Inductive multi-task learning with multiple view data. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 543–551.

Zhang, D.; He, J.; and Lawrence, R. 2013. Mi2ls: multi-instance learning from multiple informationsources. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 149–157.

Zhou, D., and Burges, C. 2007. Spectral clustering and transductive learning with multiple views. In *Proceedings of the 24th International Conference on Machine Learning*, 1159–1166.

Zhou, Z.-H., and Li, M. 2010. Semi-supervised learning by disagreement. *Knowledge and Information Systems* 24(3):415–439.

Zhou, Z.-H.; Chen, K.-J.; and Dai, H.-B. 2006. Enhancing relevance feedback in image retrieval using unlabeled data. *ACM Transactions on Information Systems* 24(2):219–244.

Zhou, Z.-H.; Zhan, D.-C.; and Yang, Q. 2007. Semi-supervised learning with very few labeled training examples. In *Proceedings of the 22nd AAAI Conference on Artificial Intelligence*, 675–680.