Constrained Instance Clustering in Multi-Instance Multi-Label Learning

Citation	Pei, Y., & Fern, X. Z. (2014). Constrained instance clustering in multi-instance multi-label learning. Pattern Recognition Letters, 37, 107-114. doi:10.1016/j.patrec.2013.07.002
DOI	10.1016/j.patrec.2013.07.002
Publisher	Elsevier
Version	Accepted Manuscript
Terms of Use	http://cdss.library.oregonstate.edu/sa-termsofuse

Constrained Instance Clustering in Multi-Instance Multi-Label Learning

Yuanli Pei*, Xiaoli Z. Fern

School of Electrical Engineering and Computer Science, Oregon State University, Corvallis, OR 97331, USA

Abstract

In multi-instance multi-label (MIML) learning, datasets are given in the form of bags, each of which contains multiple instances and is associated with multiple labels. This paper considers a novel instance clustering problem in MIML learning, where the bag labels are used as background knowledge to help group instances into clusters. The goal is to recover the class labels or to find the subclasses within each class. Prior work on constraint-based clustering focuses on pairwise constraints and can not fully utilize the bag-level label information. We propose to encode the bag-label knowledge into soft bag constraints that can be easily incorporated into any optimization based clustering algorithm. As a specific example, we demonstrate how the bag constraints can be incorporated into a popular spectral clustering algorithm. Empirical results on both synthetic and real-world datasets show that the proposed method achieves promising performance compared to state-of-the-art methods that use pairwise constraints.

Keywords:
MIML, Instance Clustering, Constrained Clustering, Bag Constraints, Spectral Clustering

1. Introduction

The Multi-Instance Multi-Label (MIML) learning framework [23] has been successfully applied in a variety of applications including computer vision [5, 15, 21] and audio analysis [14]. In MIML, datasets are given in the form of bags and each bag contains multiple instances. It is assumed that there exists a class structure such that each instance in the bag belongs to one of the classes. However, the instance class labels are not directly observed. Instead, the class labels are only provided at the bag level, which is the union of all instance labels within the bags. The goal of MIML learning is then to build a classifier to predict the labels for an unseen bag [22, 23] or to annotate the label of each instance within the bag [1].

In this paper, we consider a novel instance clustering problem within the MIML framework, where the goal is to group instances from all bags into clusters. In particular, we seek to find a cluster structure that corresponds to or refines the existing class structure. That is, we assume that each class contains one or more subclasses and our goal is to find such subclasses via clustering. In our motivating application, we want to understand the structure of bird song within each species. Here a bag corresponds to the spectrogram of a 10-second field recording of multiple birds, and each instance corresponds to a segment in the spectrogram capturing a single bird utterance (a syllable). The labels of a bag are the set of species (one or more) present in the recording. Birds from a single species may vocalize in different modes. For instance, the sound made by a woodpecker has at least two distinct modes: pecking and calling. We are interested in finding such distinct modes within each species by applying clustering techniques to instances. Ideally we would perform clustering on instances of the same species to learn such modes. However, this is impractical because the labels are only provided at the bag level and we do not have accurate instance-level species labels. Therefore, we cast this problem as an instance clustering problem with bag-level class labels as side information.

Existing literature on clustering with side information primarily focuses on pairwise Must-Link (ML) and Cannot-Link (CL) constraints [6–8, 12, 13, 17, 19]. Note that one could potentially generate ML and CL constraints based on the bag-level labels, but they incorporate only limited information for MIML datasets (as will be discussed in Sec.4.3) and are not effective for our problem. Another closely related topic is MIML instance annotation [1, 16, 21], where an instance classifier is learned from MIML data that predicts the class label of each instance. The key difference between MIML instance annotation and our work is that we are interested in finding the refinement of class structure for the instances, whereas instance annotation only focuses on recovering the class labels of instances based-on the bag-level labels.

In this paper, we propose to incorporate the bag-level side information in the form of *bag constraints*. Our approach defines two similarity measures between bags based on *class* labels and *cluster* labels respectively. By requiring the two similarities to order pairs of bags consistently, we encode bag-level label knowledge into soft constraints, which can be easily incorporated into traditional clustering objectives as a penalty term. In particularly, we incorporate such constraints into a popular spectral clustering algorithm and validate the effectiveness of the resulting method on both synthetic and real-world datasets. Experiments show that our method produces good clustering

*Corresponding Author. Tel: +1 (541) 737-3617. Fax: +1 (541) 737-1300.
Email addresses: peiy@eecs.oregonstate.edu (Yuanli Pei), xfern@eecs.oregonstate.edu (Xiaoli Z. Fern)
URL: http://web.engr.oregonstate.edu/~xfern (Xiaoli Z. Fern)

results compared to spectral clustering methods with pairwise constraints.

2. Problem Statement

In our problem, the data consists of M bags $\{\mathbb{B}_1, \cdots, \mathbb{B}_M\}$, where each bag $\mathbb{B}_i$ contains n_i instances, i.e., $\mathbb{B}_i = \{x_{i1}, \cdots, x_{in_i}\}$ with $x_{iq} \in \mathcal{R}^d$. As prior knowledge, each $\mathbb{B}_i$ is associated with a set of class labels, denoted by $\mathbb{Y}_i \subseteq \{1, \cdots, C\}$, where C is the total number of distinct classes. Denote $\mathcal{X} = \bigcup_{m=1}^{M} \mathbb{B}_m$ and let $N = \sum_{m=1}^{M} n_m$ be the total number of instances[1] in $\mathcal{X}$, our goal is to partition the N instances in $\mathcal{X}$ into K disjoint clusters that respect the class boundaries. That is, if x_p and x_q belong to the same cluster, they must belong to the same class, while the converse is true only if $K = C$, in which case we wish to recover the classes perfectly by clustering. In the case of $K > C$, some classes may contain multiple clusters that correspond to subclasses of the existing classes.

3. Bag Constraints for MIML Instance Clustering

In our setup, the desired cluster labels are closely related to the class labels. To capture this relationship, we introduce two different representations for each pair of bags using their class-label set and cluster-label set respectively, and require these two representations to induce similarities that behave similarly in terms of their ranking orders. That is, if a pair of bags $\mathbb{B}_i$ and $\mathbb{B}_j$ is more similar to each other than another pair $\mathbb{B}_r$ and $\mathbb{B}_s$ according to their class labels, the similarity should maintain the same order when measured using cluster labels. This will allow us to find a clustering solution that implicitly respects the class labels.

More formally, we use (i, j) to represent a pair of bags $\mathbb{B}_i$ and $\mathbb{B}_j$. Let $\Omega_L(i, j)$ be the *class-label similarity* between $\mathbb{B}_i$ and $\mathbb{B}_j$, and let $\Omega_A(i, j)$ be their *cluster-label similarity*.[2] Conceivably, a good clustering result is such that a large value of $\Omega_L(i, j)$ corresponds to a large value of $\Omega_A(i, j)$. For example, for a pair of bags $\mathbb{B}_i$ and $\mathbb{B}_j$ with a certain number of class labels, the more class labels they share, the larger the value $\Omega_L(i, j)$ will be and correspondingly we expect the value $\Omega_A(i, j)$ to be larger.

Using the above defined notation, we introduce the bag constraints as follows:

$$[\Omega_L(i, j) - \Omega_L(r, s)][\Omega_A(i, j) - \Omega_A(r, s)] \geq 0, \ \forall i, j, r, s \in \{1, \ldots, M\} \tag{1}$$

The first term on the left hand side of the above inequality compares the difference of class-label similarities between (i, j) and (r, s). The second term computes the corresponding difference of the cluster-label similarities. By requiring the nonnegativity of the product, the inequality requires the two similarities to consistently order any pairs of bags. In this way, the bag constraints indirectly enforces the consistency between class labels and cluster labels for all bags.

The above bag constraints can be easily incorporated into any optimization based clustering algorithm. Let f_A be the objective to be maximized by a clustering algorithm, the bag constraints can be incorporated as

$$\max_A \quad f_A + \frac{\alpha}{2M^2} \sum_{(i,j)} \sum_{(r,s)} [\Omega_L(i, j) - \Omega_L(r, s)][\Omega_A(i, j) - \Omega_A(r, s)] \tag{2}$$

where M is the total number of bags, $2M^2$ is introduced as a normalizer to make α invariant to different number of bags, and the parameter α controls the trade-off between the bag constraints and the original clustering objective.

4. Incorporate Bag Constraints to Spectral Clustering

In this section, we incorporate the bag constraints into spectral clustering by modifying the *Normalized LinkRatio* objective. We show that this leads to a standard spectral clustering problem with a modified similarity matrix.

4.1. Preliminaries on Spectral Clustering

We first briefly review the spectral clustering. Let $A = [a_1, \cdots, a_K]$ be a *partition matrix*, where each column a_k is a binary assignment vector for cluster $\mathbb{X}_k$, with $a_{qk} = 1$ if instance x_q is assigned to cluster $\mathbb{X}_k$ and 0 otherwise. Let W be the symmetric *similarity matrix* of instances. Define the *degree matrix* $D = \text{Diag}(W\mathbf{1}_N)$, where $\text{Diag}(\cdot)$ forms a diagonal matrix with elements of the input vector as the diagonal elements, $\mathbf{1}_N$ denotes a N-dimensional vector of all 1's, and N is the total number of vertices. The K-way spectral clustering with *Normalized LinkRatio* objective is defined as [18]

$$\max_A \quad \frac{1}{K} \sum_{k=1}^{K} \frac{a_k^T W a_k}{a_k^T D a_k} \tag{3}$$

$$\text{s.t.} \quad A \in \{0, 1\}^{N \times K}, \quad A\mathbf{1}_K = \mathbf{1}_N. \tag{4}$$

Rewrite the objective as

$$\frac{1}{K} \sum_{k=1}^{K} \frac{a_i^T W a_i}{a_i^T D a_i} = \sum_{k=1}^{K} \frac{a_k^T D^{1/2} D^{-1/2} W D^{-1/2} D^{1/2} a_k}{a_k^T D a_k}.$$

Define $z_k = \frac{D^{1/2} a_k}{\|D^{1/2} a_k^T\|}$, and $Z = [z_1, \cdots, z_K]$. Ignoring the discrete constraint for Z at this stage, one can formulate a new clustering problem with respect to variable Z as

$$\max_Z \quad \text{tr}(Z^T D^{-1/2} W D^{-1/2} Z) \tag{5}$$

$$\text{s.t.} \quad Z^T Z = I \tag{6}$$

where the constraint (6) comes from the definition of Z. The solution of Z for this new problem is the eigenvectors associated with the K largest eigenvalues of $D^{-1/2} W D^{-1/2}$ [2]. Correspondingly, a discrete solution A of the original problem can be obtained by taking a rounding procedure from Z (e.g., using Kmeans or the approach proposed in [18]).

[1] In this paper, we assume that all instances are distinct.

[2] At this point, we do not specify the function forms of $\Omega_L(\cdot, \cdot)$ and $\Omega_A(\cdot, \cdot)$, since they can be problem-specified. However, this does not prevent us from viewing them as geometrical similarities.

4.2. *Spectral Clustering with Bag Constraints*

To incorporate the bag constraints, we need to define the two similarity functions in Eq. (1), the class-label similarity function $\Omega_L(\cdot)$ and the cluster-label similarity $\Omega_A(\cdot)$. Ideally, $\Omega_L(\cdot)$ should satisfy the following conditions: (1) In the case where class label information between two bags $\mathbb{B}_i$ and $\mathbb{B}_j$ is unambiguous, (i.e., they do not share class label or they both belong to the same single class), $\Omega_L(i, j)$ should achieve minimum or maximum values; (2) In the ambiguous case where bags $\mathbb{B}_i$ and $\mathbb{B}_j$ have multiple labels and $\mathbb{Y}_i \cap \mathbb{Y}_j \neq \phi$, the smaller the quantity $\frac{|\mathbb{Y}_i \cap \mathbb{Y}_j|}{|\mathbb{Y}_i \cup \mathbb{Y}_j}|$ ($|\mathbb{Y}_i|$ is the number of distinct classes in $\mathbb{Y}_i$) is, i.e., the smaller the relative "common-label" set is, the smaller $\Omega_L(i, j)$ should be.

Based on the above considerations, we define the following class-label similarity function. Let y_i be the $C \times 1$ binary class indicator vector for bag $\mathbb{B}_i$, with elements $y_{ic} = 1/|\mathbb{Y}_i|$ if $c \in \mathbb{Y}_i$, and $y_{ic} = 0$ otherwise. Denote $Y = [y_1, \cdots, y_M]$, where $y_m = \mathbf{0}$ for any bag $\mathbb{B}_m$ that is not labeled. The *class-label similarity* between (i, j) is defined as

$$\Omega_L(i, j) = y_i^T y_j \tag{7}$$

To define $\Omega_A(\cdot)$, denote the *bag indicator matrix* $B = [b_1, \cdots, b_M]$, with column vector $b_i \in \{0, 1\}^{N \times 1}$ and the element $b_{qi} = 1$ if instance $x_q \in \mathbb{B}_i$, and $b_{qi} = 0$ otherwise. The *cluster structure* of bag $\mathbb{B}_i$ can be captured by the $K \times 1$ column vector $Z^T D^{-1/2} b_i$. The k-th element in the cluster structure vector is $\frac{|\mathbb{X}_k \cap \mathbb{B}_i|}{||D^{1/2} a_k^T||}$, where $|\mathbb{X}_k \cap \mathbb{B}_i|$ counts the number of instances in bag $\mathbb{B}_i$ that belong to cluster $\mathbb{X}_k$. Essentially, $Z^T D^{-1/2} b_i$ forms a histogram of the cluster labels in bag $\mathbb{B}_i$ and normalizes each count by a quantity that can be roughly interpreted as the volume of the cluster.[3] This normalization allows the similarity measure to balance the contributions of clusters of different sizes. We now define the *cluster-label similarity* between (i, j) as

$$\Omega_A(i, j) = (Z^T D^{-1/2} b_i)^T (Z^T D^{-1/2} b_j) = b_i^T D^{-1/2} Z Z^T D^{-1/2} b_j \tag{8}$$

Substituting $\Omega_L(i, j)$ and $\Omega_A(i, j)$ into the inequality of bag constraints Eq. (1) , we have

$$\begin{aligned}(y_i^T y_j - y_r^T y_s)(b_i^T D^{-1/2} Z Z^T D^{-1/2} b_j - b_r^T D^{-1/2} Z Z^T D^{-1/2} b_s) \geq 0 \\ \Leftrightarrow \operatorname{tr}(Z^T D^{-1/2} (y_i^T y_j - y_r^T y_s)(b_j b_i^T - b_s b_r^T) D^{-1/2} Z) \geq 0\end{aligned}$$

where $y_i^T y_j - y_r^T y_s$ is a scalar. This inequality constraint is imposed for two pairs of bags. To incorporate the bag constraints for all pairs of bags, we follow the method introduced in Eq. (2) and add the following penalty term to the *Normalized LinkRatio* objective

Algorithm 1 *Spectral Clustering with Bag Constraints*

Input: A set of bags $\{\mathbb{B}_i\}_{i=1}^M$, $\mathbb{B}_i = \{x_{i1}, \cdots, x_{in_i}\}$; a set of known label sets associated with bags $\{(\mathbb{Y}_i, \mathbb{B}_i)\}$; parameter α; the number of instance clusters K.
Output: Instance clustering result.

```
1: Create instance similarity matrix $W \in \mathcal{R}^{N \times N}$; form the diagonal
   degree matrix $D = \mathrm{Diag}(W\mathbf{1}_N)$.
2: Form the label indicator matrix $Y$ and the bag indicator matrix $B$,
   as described in Sec. 4. Construct the bag-constraint matrix $Q =
   B(Y^T Y - \mu I)B^T$.
3: Compute the normalized similarity matrix with bag constraints
   $W' = D^{-1/2}(W + \alpha Q)D^{-1/2}$.
4: Find the $K$ largest eigenvectors of $W', v_1, \cdots, v_K$; form the matrix
   $V = [v_1, \cdots, v_K] \in \mathcal{R}^{N \times K}$.
5: Re-normalize the rows of $V$ to have unit length yielding $V' \in \mathcal{R}^{N \times K}$,
   i.e., $V'_{ij} = V_{ij} / (\sum_j V_{ij}^2)^{1/2}$ .
6: Treat each row of $V'$ as a point in $\mathcal{R}^K$ and cluster $V'$ via Kmeans.
   Assign the original instance $x_q$ to cluster $\mathbb{X}_k$ if and only if the $q$-th
   row of $V'$ is assigned to $\mathbb{X}_k$.
```

$$\frac{\alpha}{2M^2} \sum_{(i,j)} \sum_{(r,s)} [\Omega_L(i, j) - \Omega_L(r, s)][\Omega_A(i, j) - \Omega_A(r, s)] \tag{9}$$

$$= \frac{\alpha}{2M^2} \operatorname{tr}(Z^T D^{-1/2} \sum_{(i,j)} \sum_{(r,s)} (y_i^T y_j - y_r^T y_s)(b_j b_i^T - b_s b_r^T) D^{-1/2} Z) \tag{10}$$

$$= \alpha \cdot \operatorname{tr}\left(Z^T D^{-1/2} B (Y^T Y - \mu I) B^T D^{-1/2} Z\right) \tag{11}$$

$$= \alpha \cdot \operatorname{tr}\left(Z^T D^{-1/2} Q D^{-1/2} Z\right),$$

$$\text{with } \mu = \frac{\operatorname{sum}(Y^T Y)}{M^2} \text{ and } Q = B(Y^T Y - \mu I)B^T. \tag{12}$$

The two summations in Eq. (9) sum over all possible configurations of (i, j) and (r, s), and the function sum$(\cdot)$ in Eq. (12) sums over all elements of input matrix. The detailed derivation from Eq. (10) to Eq. (11) can be found in Appendix A.

Adding the above bag constraints as a penalty term to the *Normalized LinkRatio* objective (5) and taking into account the original constraint (6), we can rewrite the spectral clustering with bag constraints as

$$\max_Z \quad tr(Z^T D^{-1/2} (W + \alpha Q) D^{-1/2} Z) \tag{13}$$

$$\text{s.t.} \quad Z^T Z = I \tag{14}$$

It is easy to see that this formulation is equivalent to the standard spectral clustering Eq. (5) and Eq. (6) with a modified similarity matrix. We can then apply the general approach of spectral clustering to solve this optimization problem.

The spectral clustering algorithm with bag constraints is summarized in Algorithm 1. Note that in step 1, one can choose any method to compute the similarity matrix W so that the data similarities are properly captured (Existing methods include the ones in [9, 10, 20]). We applied the Kmeans rounding procedure in Step 6. One can, of course, apply any other appropriate rounding procedure. Step 2 involves several matrix multiplications. Since the dimension of Y is $C \times M$ and B is $N \times M$, the

[3] The normalization factor for cluster $\mathbb{X}_k$ is $||D^{1/2} a_k^T||$, where a_k is the binary indicator vector for cluster $\mathbb{X}_k$ and D is the degree matrix.

Table 1: **Relation of Bag Constraints and Pairwise Constraints for** $K = C$.

Cases	$Q_{p,q}$	ML/CL
$\lvert\mathbb{Y}_i\rvert = 1,\ x_p, x_q \in \mathbb{B}_i$	$1 - \mu$	ML
$\mathbb{Y}_i = \mathbb{Y}_j,\ \lvert\mathbb{Y}_i\rvert = \lvert\mathbb{Y}_j\rvert = 1,\ x_p \in \mathbb{B}_i,\ x_q \in \mathbb{B}_j$	1	ML
$\mathbb{Y}_i \cap \mathbb{Y}_j = \phi,\ x_p \in \mathbb{B}_i,\ x_q \in \mathbb{B}_j$	0	CL
$\lvert\mathbb{Y}_i\rvert > 1,\ x_p, x_q \in \mathbb{B}_i$	$\frac{1}{\lvert\mathbb{Y}_i\rvert} - \mu$	N/A
$\mathbb{Y}_i \cap \mathbb{Y}_j \neq \phi, \lvert\mathbb{Y}_i\rvert > 1$ or $\lvert\mathbb{Y}_j\rvert > 1, x_p \in \mathbb{B}_i, x_q \in \mathbb{B}_j$	$(0, 1)$	N/A

complexity of Step 2 is dominated by $\mathcal{O}(N^2M)$. Step 4 computes the top K eigenvectors of W', which is the most computationally expensive part. Using the Lanczos method, the complexity of Step 4 is $\mathcal{O}(KTnnz(W'))$, where T is the number of Lanczos iteration steps and $nnz(W')$ is the number of nonzero elements in matrix W' [6]. Hence, the overall complexity is not increased by introducing bag constraints.

4.3. Relation to ML/CL Pairwise Constraints

As analyzed previously, in some cases pairwise constraints can be induced from the bag-level labels. When $K = C$, partitioning instances into K clusters is similar to predicting the class labels for instances. In this case, if a bag only has a single label then all instances within the bag belong to the same cluster and thus ML constraints can be imposed. Similarly, if two single-label bags have the same label, ML constraints should also be imposed on all pairs of instances formed across the two bags. For two bags that do not share any label, since they can not belong to the same cluster, CL constraints can be imposed on any instance pairs formed across the two bags. When $K > C$, some classes may correspond to more than one clusters. Thus, we can not impose ML constraints even for instances pairs that come from a single class label. However, CL constraints are still possible when two bags do not share any label.

The bag-constraint matrix Q introduced in Sec. 4 has some important properties that are closely related to pairwise constraints. We summarized these properties in the following proposition.

Proposition 1 (Properties of Q)**.** *Let* $\mathbb{Y}_i$ *and* $\mathbb{Y}_j$ *be the sets of class labels for bag* $\mathbb{B}_i$ *and bag* $\mathbb{B}_j$ *respectively. Let* $|\mathbb{Y}_i|$ *and* $|\mathbb{Y}_j|$ *be the sizes of the label set* $\mathbb{Y}_i$ *and* $\mathbb{Y}_j$*, respectively. Denote* $Q_{p,q}$ *as the value of the entry in Q that corresponds to the pair of instances* x_p *and* x_q*. Then the value of* $Q_{p,q}$ *can be determined according to Table 1.*

Proof. By the definition of Q in Eq. (12), we know that if $x_p, x_q \in \mathbb{B}_i$, $Q_{p,q} = \frac{1}{|\mathbb{Y}_i|} - \mu$, and that if $x_p \in \mathbb{B}_i, x_q \in \mathbb{B}_j$, $Q_{p,q} = \frac{|\mathbb{Y}_i \cap \mathbb{Y}_j|}{|\mathbb{Y}_i| \cdot |\mathbb{Y}_j|}$. It is thus easy to verify the first four cases. For the last case, since $|\mathbb{Y}_i \cap \mathbb{Y}_j| \neq \phi$, it follows that $|\mathbb{Y}_i \cap \mathbb{Y}_j| > 0$. Because the denominator $|\mathbb{Y}_i| \cdot |\mathbb{Y}_j|$ is also positive, we know $Q_{p,q} > 0$. Also given that $|\mathbb{Y}_i| > 1$ or $|\mathbb{Y}_j| > 1$, we know $|\mathbb{Y}_i \cap \mathbb{Y}_j| < |\mathbb{Y}_i| \cdot |\mathbb{Y}_j|$. Hence, $Q_{p,q} < 1$. □

It can be seen that, when ML constraints can be inferred for x_p and x_q, the value of $Q_{p,q}$ is 1 or $1 - \mu$ (approximately equal to 1 since μ is usually very small), which is the maximum of the constraint matrix. The value of $Q_{p,q}$ reaches 0 when CL constraints can be inferred. In other cases where some overlap exists between the class labels of two bags and no ML or CL can be imposed, the value of $Q_{p,q}$ lies in range $(0, 1)$ and the magnitude depends on the extend of the overlap. The more overlap their label sets have, the larger the value of $Q_{p,q}$ is. As such, we can view pairwise ML and CL constraints as only able to accommodate the cases where $Q_{p,q}$ takes extreme values. In contrast, our proposed method can capture different levels of ambiguity by allowing $Q_{p,q}$ to take a continuous value between zero and one, which potentially leads to more effective usage of the bag-level label information.

Table 2: **MIML Datasets Information**. "Single-Label bags" is the number of bags that contain only a single class; "Multi-Label bags" is the number of bags that have multiple labels; "Avg. Inst." is the average number of instances in each bag; "Avg. Bag Label" is the average number of class labels in each bag.

Dataset	Birdsong	MSRC v2	Carroll	Frost
Classes	13	23	24	24
Dimension	38	48	16	16
Single-Label Bags	199	130	1	12
Multi-Label Bags	349	461	165	132
Total Bags	548	591	166	144
Total Inst.	4998	1758	717	565
Avg. Inst.	9.12	2.97	4.32	3.92
Avg. Bag Label	2.02	2.51	3.93	3.60

5. Empirical Evaluation

We conduct experiments on synthetic and real-world MIML datasets to evaluate the proposed bag-constrained spectral clustering method. The baseline methods include both unconstrained spectral clustering and existing spectral clustering algorithms with pairwise constraints.

5.1. Datasets Description

We use two real-world datasets and two synthetic datasets to evaluate our method. These datasets are previously used by a recent study on instance annotation for MIML [1] and are available online.[4] The summary of the datasets is provided in Table 2.

HJA Birdsong is a real-word MIML dataset with 548 bags, each representing the spectrogram of a 10-second birdsong recording. Each instance corresponds to a bird song syllable in the spectrogram described by a 38-dimensional feature vector. There are 10232 instances, 4998 of which are provided with ground-truth class labels. For evaluation purpose, we use the filtered dataset, which only contains the labeled instances in each bag. Note that the ground-truth instance labels are only used in the evaluation.

MSRC v2 is the second version (v2) of Microsoft Research Cambridge (MSRC) image dataset,[5] containing 591 images and 23 classes. Each image is considered as a bag and regions in the images are viewed as instances. Each instance is described by a 16-dimensional histogram of gradients and a 32-dimensional histogram of colors.

[4] `http://web.engr.oregonstate.edu/~briggsf/kdd2012datasets`

[5] `http://research.microsoft.com/en-us/projects/objectclassrecognition/`

Letter-Carroll and **Letter-Frost** are two synthetic datasets generated using the Letter Recognition dataset from the UCI Machine Learning repository[6] and two poems.[7] To generate these datasets, words in the poems are viewed as bags and letters in each word are the instances and are randomly sampled (without replacement) from the Letter Recognition dataset. Bag-level labels are formed as the union of all letter labels in the word.

All datasets are standardized such that the mean of each feature is 0 and the standard deviation is 1. Instance similarities are computed using the *local scaling* factor proposed in [20]. Specifically, the similarity between instances x_p and x_q is computed by $W_{pq} = \exp\left(-\frac{\|x_p - x_q\|^2}{2\sigma_p\sigma_q}\right)$, where σ_p and σ_q are *local scaling* factors. The local scaling factor σ_q is defined as $\sigma_q = \|x_q - x_q^{(t)}\|$, where $x_q^{(t)}$ is the t-th nearest neighbor of x_q. We adopt $t = 7$ as recommended in [20], which is also shown to be effective in [11].

5.2. Baseline Methods

Our baseline methods include unconstrained spectral clustering (SP)and two constrained spectral clustering algorithms, Spectral Learning (*SpLearn*) algorithm proposed in [7] and constrained spectral clustering by regularization (*SpReg*) method proposed in [6]. SpLearn incorporates ML and CL constraints by directly modifying the entries of similarity matrix to 1 for ML constraints and to 0 for CL constraints. SpReg encodes ML constraints by adding a penalty term into the Normalized Cut objective. Note that SpReg only incorporates ML constraint but not CL constraints. To apply unconstrained spectral clustering (SP) to MIML instance clustering, we ignore the bag structure as well as the bag-level labels. For constrained spectral clustering, we create ML and CL constraints according to Table 1. The parameter β in SpReg that controls the enforcement of constraints is set to 20, the same value as that used in [6].

5.3. Parameter Selection

In our algorithm, the parameter α is introduced to balance the trade-off between instance-feature similarity and the bag constraints. A large value of α imposes stronger restriction on the clustering solution to conform to the bag constraints and a small value of α produces clustering results without being heavily influenced by such constraints. We tested the performance of our method over a range of α values (from 0 to 1, by a 0.005 increment) on all our datasets and the results have shown that a value in the range $[0.5, 1]$ typically leads to significantly improved clustering performance. Figure 1 shows the performance of our method on the two real-world datasets as a function of α. In all the following experiments, the parameter α is set to 0.7.

5.4. Experiments and Discussions

We conducted experiments in two different scenarios. In the first scenario K is set to C and the goal is to group the instances based on their classes. In the second scenario, we have $K > C$ and some classes are represented by more than one cluster. In both scenarios, we test our algorithm with two implementations in order to thoroughly evaluate its performance. The first implementation (CSP) is a direct implementation of algorithm 1. The second implementation (CSP w.o.clml) is designed to test how well our algorithm could perform if we ignore the information that can be captured by ML and CL constraints. For this implementation, we set the entries that correspond to ML or CL constraints in the bag-constraint matrix Q to 0 and leave the rest unchanged. The two implementations are identical otherwise.

5.4.1. Scenario 1: $K = C$

In this scenario, we evaluate the performance of our method by changing the percentage of labeled bags. In particular, we vary the percentage of labeled bags from 20% to 100% of the whole dataset, with a 20% increment. For a fixed percentage, we randomly subsample bags (without replacement) to create the bag-constraint matrix and pairwise ML/CL constraints. The experiment is repeated for 20 random runs and the results are averaged.

We use two criteria to evaluate the clustering performance, *Normalized Mutual Information* (NMI) and *Class Purity*. The NMI is defined as

$$NMI = \frac{2I(\mathbf{X};\mathbf{C})}{H(\mathbf{X}) + H(\mathbf{C})} \tag{15}$$

where **X** and **C** are the numerical cluster and class label vectors, $I(\cdot;\cdot)$ computes the mutual information, and $H(\cdot)$ calculates the entropy. To compute *Class Purity*, each cluster is assigned to the most frequent class in the cluster, and then the accuracy of this assignment is measured by comparing the assigned labels with the ground-truth class labels. Formally,

$$\text{purity}(\mathbb{X}, \mathbb{C}) = \frac{1}{N}\sum_k \max_j |\mathbb{X}_k \cap \mathbb{C}_j| \tag{16}$$

where $\mathbb{X} = \{\mathbb{X}_1, \ldots, \mathbb{X}_K\}$ is the set of clusters and $\mathbb{C} = \{\mathbb{C}_1, \ldots, \mathbb{C}_C\}$ is the set of classes.

The NMI and Class Purity results are reported in Fig. 2 and Fig. 3, respectively. From these results, we have the following observations and conclusions:

- Both CSP and CSP w.o.clml *outperforms* SP significantly as more bags are labeled.

- Our method is *comparable* with SpLearn and *outperforms* SpReg when the average number of bag-level labels is *small* (HJA Birdsong). In this case, the ambiguity of instance labels induced from the bag-level labels is low, and many ML and CL constraints can be inferred. Such constraints can be properly incorporated by SpLearn to improve clustering. However, the number of ML constraints is relatively smaller compared to that of CL constraint, and thus the constraints do not help SpReg as much. In the

[6] `http://archive.ics.uci.edu/ml/machine-learning-databases/letter-recognition`

[7] The poem that generates the Letter-Carroll dataset is "*Jabberwocky*" written by Lewis Carroll in his 1872 novel *Through the Looking-Glass, and What Alice Found there*. The other poem that is used to create the Letter-Frost dataset is "*The Road Not Taken*" by *Robert Frost*, published in 1916 in the collection *Mountain Interval*.

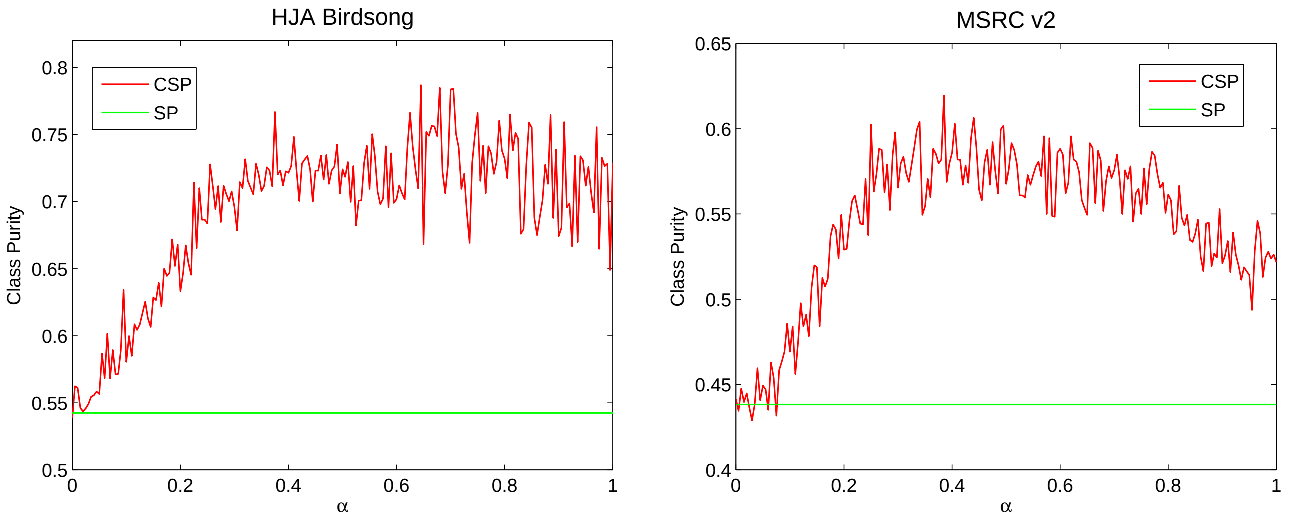

Figure 1: Class Purity results as a function of α (see Sec. 5.4.1 for the introduction of Class Purity). Higher value suggests better clustering performance. SP is unconstrained spectral clustering. CSP is the proposed spectral clustering with bag constraints.

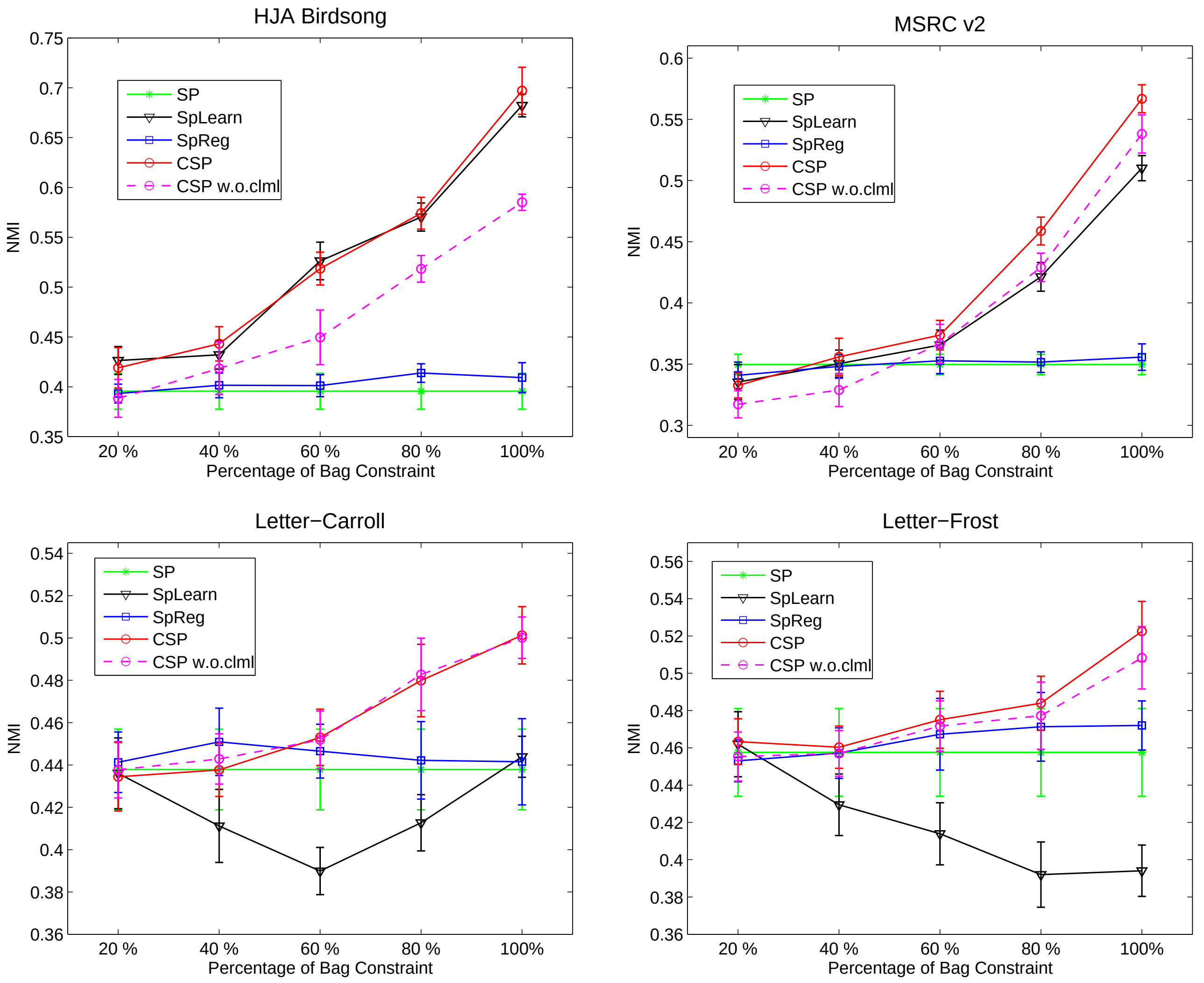

Figure 2: Scenario of $K = C$: NMI results as a function of constraints creating from different percentage of labeled bags. Results are averaged over 20 random runs of independently sampled constraint sets; error bars are reported with mean and standard deviation.

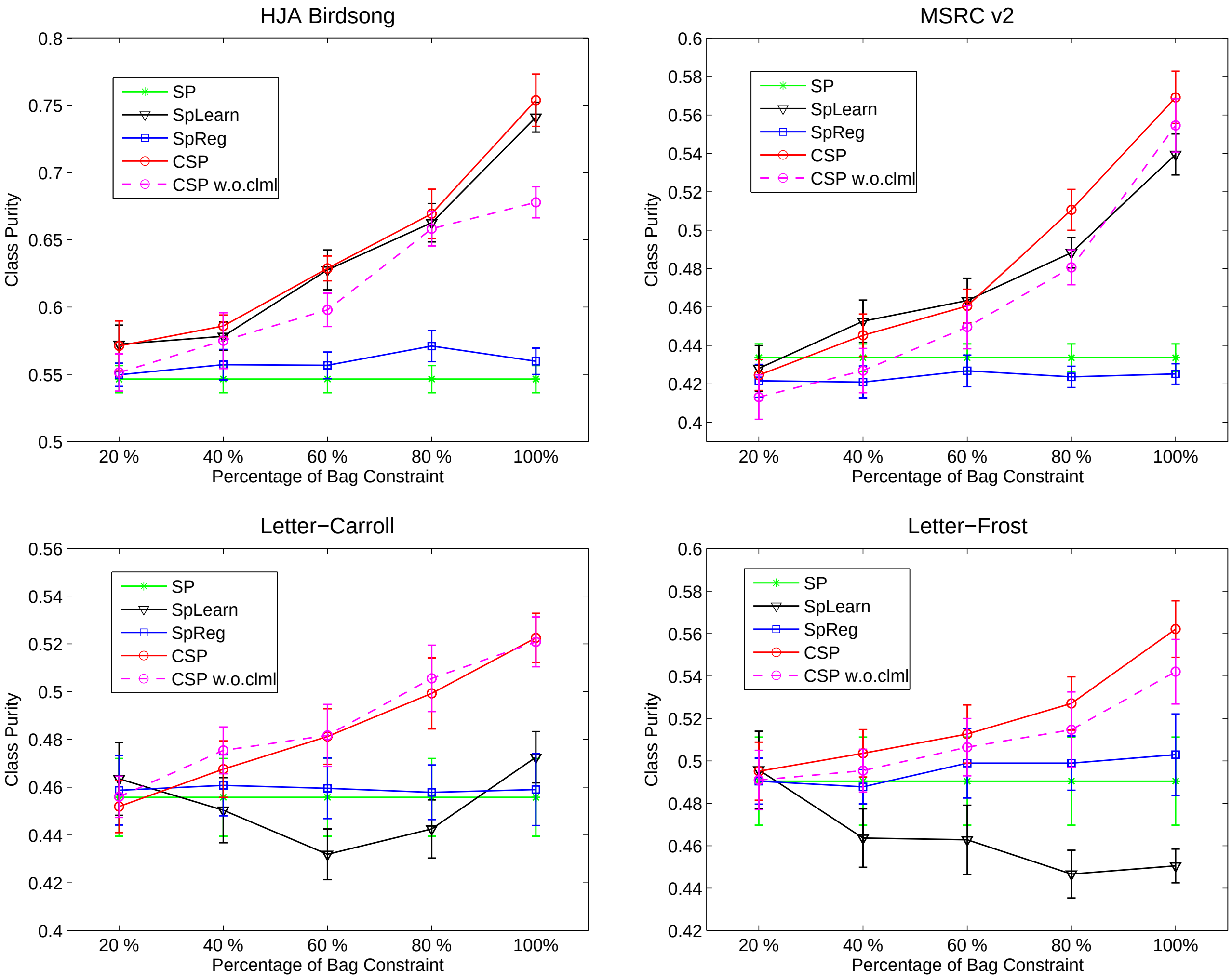

Figure 3: Scenario of $K = C$: Class Purity results as a function of different number of constraints. See Figure 2 caption.

meantime, the proposed method enables the clustering algorithm to incorporate information beyond what can be captured by the basic ML and CL constraints, which allows it to achieve competitive performance compared to SpLearn and SpReg.

- Our method *outperforms* SpLearn and SpReg when the number of average bag-level labels is relatively *larger* (MSRC v2, Letter-Carroll and Letter-Frost). In this case, very limited ML and CL constraints can be inferred, and our method with bag constraints better captures the side information in the bag-level labels. The fact that CSP and CSP w.o.clml performs almost the same in Letter-Carroll datasets indirectly demonstrates that ML and CL constraints can hardly be inferred. This gives more explanation why our method outperforms SpLearn and SpReg.

- On Letter-Carroll and Letter-Frost, while our method still *outperforms* SP, SpReg shows *no gain* and SpLearn actually leads to *degraded* clustering performance. Similar negative results have been reported in [4], which showed that constraint sets generated based on the ground truth labels can sometimes lead to degraded clustering performance. Further examination on these two datasets indicates that their bag-labels mostly induce CL constraints, which cannot be used by SpReg (thus explaining its flat performance). Moreover, the degraded performance by SpLearn suggests that CL constraints alone might not provide good guidance for MIML instance clustering. It is interesting to note that prior research [3] has demonstrated that *CL* constraints can sometimes make the solution space overly constrained, leading to more difficult clustering problem. This provides a possible explanation for the degraded performance of SpLearn.

5.4.2. Scenario 2: $K > C$

In the scenario of $K > C$, we evaluate the performance of our method by changing the number of clusters K. For each dataset, we assume that all bags are labeled at the bag-level and vary the number of clusters from $K = C$ to roughly $2C$ with 7 steps. In the case of $K > C$, ML constraints can not be extracted (see discussion in Sec. 4.3). Hence, no constraints can be incorporated into SpReg and only CL constraints could be incorporated into SpLearn. We therefore do not consider the SpReg baseline in this scenario and remove ML constraints in SpLearn. The

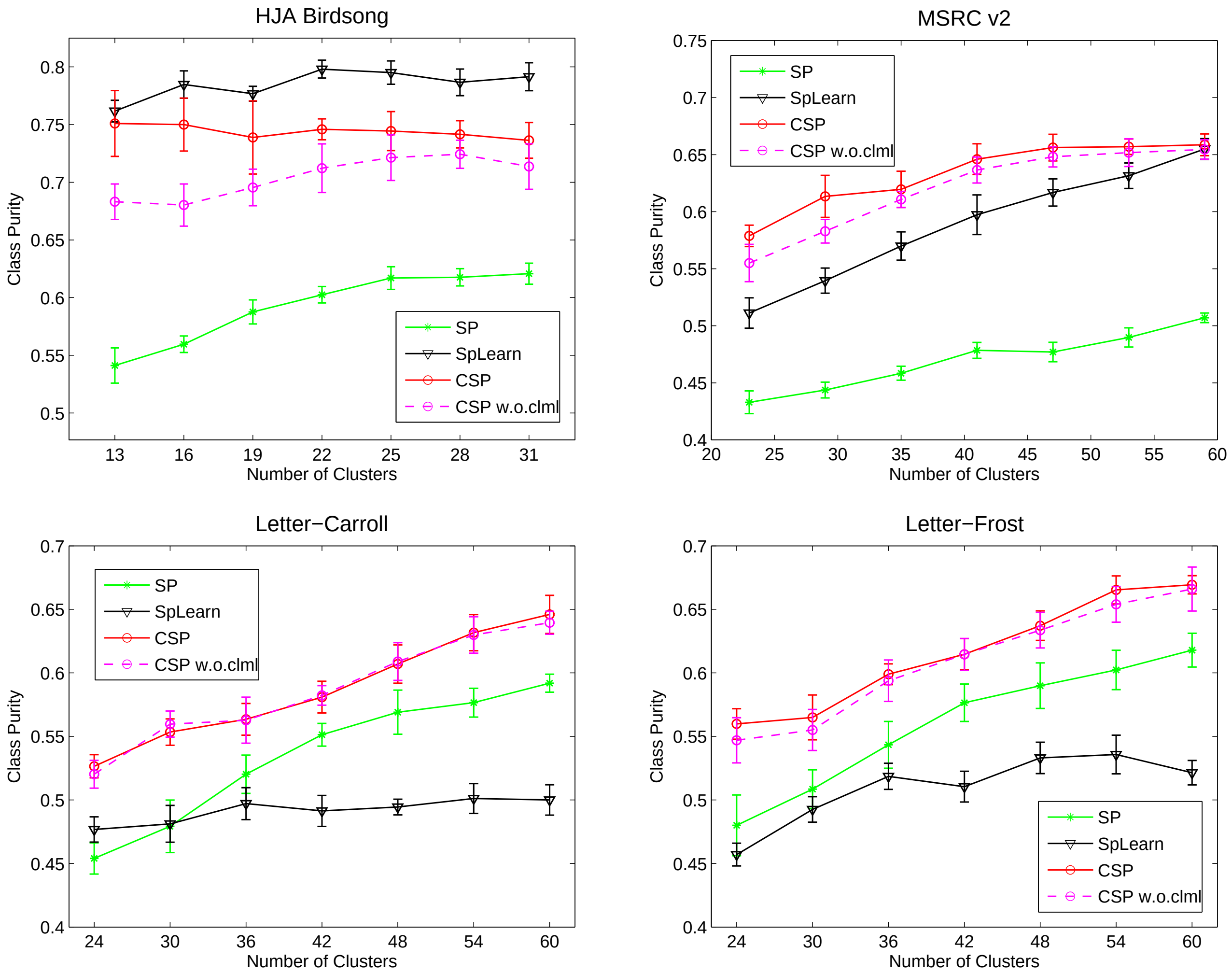

Figure 4: Scenario of $K > C$: Class Purity results as a function of the number of clusters. Results are averaged over 20 random runs. Error bars are reported with mean and standard deviation.

parameters setting is the same with the previous experiment.

We report the averaged Class Purity results over 20 runs in Fig. 4. NMI results are highly similar, and thus omitted to avoid redundancy. We can see that our method still outperforms SP consistently and significantly.

In the datasets with large average number of class labels (MSRC v2, Letter-Carroll, and Letter-Frost), we again observe that CSP and CSP w.o.clml performs similarly, which shows that few CL or ML constraints could be extracted. This is one of the possible reasons that SpLearn can not compete with our method for these datasets. Nonetheless, When the average number of class labels is small (HJA Birdsong), SpLearn excels. One possible explanation is that the bag label information in this dataset is relatively unambiguous and many pairwise constraints can be extracted. While our method can handle ambiguous information much better, SpLearn deals with unambiguous information more directly. Note that when we remove the single-label bags in HJA Birdsong and conduct the same experiment, we observed that our method is comparable with SpLearn (the result is not reported due to space limit). These results suggest that our method is more suitable for MIML datasets containing large numbers of multi-label bags.

6. Conclusion

In this paper, we introduce a novel instance clustering problem in the MIML framework, where the bag-level labels are used as side information to inform the clustering of instances. The goal is to recover the classes or to discover subclasses within each class. Traditional constraint-based clustering methods can not fully leverage the knowledge provided by the bag-level class labels. In contrast, we present a simple yet effective principle that incorporates the bag-level label information as bag constraints. The proposed constraints can be readily integrated into any optimization-based clustering algorithm by adding a penalty term to the objective. In this paper, we demonstrate how the bag constraints can be incorporated into spectral clustering and empirically validate its effectiveness on both synthetic and real-world MIML datasets. The results show that the proposed bag-constrained method for spectral clustering generally outperforms state-of-the-art spectral clustering algorithms that use pairwise ML and CL constraints and is most suitable for MIML datasets that contain relatively large number of multi-label bags.

7. Acknowledgements

This material is based upon work supported by the National Science Foundation under Grant No. IIS-1055113.

Appendix A.

In this appendix, we show the derivation from Eq. (10) to Eq. (11). First we focus on the summation term inside the trace operation. Since the two summation sum over all possible configurations of (i, j) and (r, s), we have the following rearrangement

$$\begin{aligned}
& \sum_{(i,j)} \sum_{(r,s)} (y_i^T y_j - y_r^T y_s)(b_j b_i^T - b_s b_r^T) \\
= \quad & \sum_{(i,j)} \sum_{(r,s)} \left[(y_i^T y_j) b_j b_i^T + (y_r^T y_s) b_s b_r^T - (y_r^T y_s) b_j b_i^T - (y_i^T y_j) b_s b_r^T \right] \\
= \quad & 2[\sum_{(i,j)} \sum_{(r,s)} (y_i^T y_j) b_j b_i^T - \sum_{(i,j)} \sum_{(r,s)} (y_i^T y_j) b_s b_r^T] \\
= \quad & 2[\sum_{(i,j)} b_j (y_i^T y_j) b_i^T \sum_{(r,s)} 1 - \sum_{(i,j)} (y_i^T y_j) \sum_{(r,s)} b_s b_r^T] \\
= \quad & 2[M^2 \sum_{(i,j)} b_j (y_i^T y_j) b_i^T - \text{sum}(Y^T Y) B B^T] \\
= \quad & 2M^2 \cdot [\sum_{(i,j)} b_j (y_j^T y_i) b_i^T - \frac{\text{sum}(Y^T Y)}{M^2} B B^T] \\
= \quad & 2M^2 \cdot (B Y^T Y B^T - \frac{\text{sum}(Y^T Y)}{M^2} B B^T) \\
= \quad & 2M^2 \cdot B(Y^T Y - \mu I) B^T
\end{aligned}$$

where sum$(\cdot)$ denotes summing over all elements of matrix, $\mu = \frac{\text{sum}(Y^T Y)}{M^2}$, and I is the identity matrix. In the derivation, we have used the fact that $y_i^T y_j$ is a scalar and $y_i^T y_j = y_j^T y_i$. Correspondingly, Eq. (10) is derived to Eq. (11) as

$$\begin{aligned}
& \frac{\alpha}{2M^2} \cdot tr\left(Z^T D^{-1/2} [2M^2 B(Y^T Y - \mu I) B^T] D^{-1/2} Z \right) \\
= \quad & \alpha \cdot tr\left(Z^T D^{-1/2} B(Y^T Y - \mu I) B^T D^{-1/2} Z \right).
\end{aligned}$$

References

[1] Briggs, F., Fern, X. Z., Raich, R., 2012. Rank-loss Support Instance Machines for MIML Instance Annotation. In: Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining. KDD '12. ACM, New York, NY, USA, pp. 534–542.

[2] Chung, F. R. K., 1997. Spectral Graph Theory. Vol. 92. American Mathematical Society.

[3] Davidson, I., Ravi, S., 2006. Identifying and Generating Easy Sets of Constraints For Clustering. In: Proceedings of the 21st AAAI conference on Artificial Intelligence. AAAI Press, pp. 336–341.

[4] Davidson, I., Wagstaff, K., Basu, S., 2006. Measuring Constraint-Set Utility for Partitional Clustering Algorithms. In: PKDD. pp. 115–126.

[5] Feng, S., Xu, D., Jan. 2010. Transductive Multi-Instance Multi-Label learning algorithm with application to automatic image annotation. Expert Syst. Appl. 37 (1), 661–670.

[6] Ji, X., Xu, W., 2006. Document Clustering with Prior Knowledge. In: SIGIR '06: Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval. ACM, New York, NY, USA, pp. 405–412.

[7] Kamvar, S. D., Klein, D., Manning, C. D., 2003. Spectral Learning. In: International Joint Conference on Artificial Intelligence, IJCAI. pp. 561–566.

[8] Kulis, B., Basu, S., Dhillon, I., Mooney, R., Jan. 2009. Semi-supervised Graph Clustering: a Kernel Approach. Machine Learning Journal 74 (1), 1–22.

[9] Ng, A. Y., Jordan, M. I., Weiss, Y., 2001. On Spectral Clustering: Analysis and an algorithm. In: Advances in Neural Information Processing Systems, NIPS. MIT Press, pp. 849–856.

[10] Shi, J., Malik, J., Aug. 2000. Normalized Cuts and Image Segmentation. IEEE Trans. Pattern Anal. Mach. Intell. 22 (8), 888–905.

[11] Sugiyama, M., May 2007. Dimensionality Reduction of Multimodal Labeled Data by Local Fisher Discriminant Analysis. Journal of Machine Learning Research, JMLR 8, 1027–1061.

[12] Wang, F., Ding, C. H. Q., Li, T., 2009. Integrated KL (K-means - Laplacian) Clustering: A New Clustering Approach by Combining Attribute Data and Pairwise Relations. In: SIAM SDM. pp. 38–48.

[13] Wang, X., Davidson, I., 2010. Flexible Constrained Spectral Clustering. In: Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data mining. KDD '10. ACM, New York, NY, USA, pp. 563–572.

[14] Xu, X.-S., Xue, X., Zhou, Z.-H., 2011. Ensemble Multi-Mnstance Multi-Label Learning Approach for Video Annotation Task. In: Proceedings of the 19th ACM international conference on Multimedia. MM '11. ACM, New York, NY, USA, pp. 1153–1156.

[15] Xue, X., Zhang, W., Zhang, J., Wu, B., Fan, J., Lu, Y., 2011. Correlative multi-label multi-instance image annotation. In: IEEE International Conference on Computer Vision. IEEE, pp. 651–658.

[16] Yang, S.-H., Bian, J., Zha, H., 2010. Hybrid Generative/Discriminative Learning for Automatic Image Annotation. In: Conference on Uncertainty in Artificial Intelligence, UAI. AUAI Press, pp. 683–690.

[17] Yu, S. X., Shi, J., 2001. Grouping with Bias. In: Advances in Neural Information Processing Systems, NIPS. MIT Press, Cambridge, MA, pp. 1327–1334.

[18] Yu, S. X., Shi, J., 2003. Multiclass Spectral Clustering. In: Proceedings of the Ninth IEEE International Conference on Computer Vision. ICCV '03. IEEE Computer Society, Washington, DC, USA, pp. 313–319.

[19] Yu, S. X., Shi, J., Jan. 2004. Segmentation Given Partial Grouping Constraints. IEEE Trans. Pattern Anal. Mach. Intell. 26 (2), 173–183.

[20] Zelnik-Manor, L., Perona, P., 2004. Self-Tuning Spectral Clustering. In: Advances in Neural Information Processing Systems, NIPS. MIT Press, Cambridge, MA, USA, pp. 1601–1608.

[21] Zha, Z.-J., Hua, X.-S., Mei, T., Wang, J., Qi, G.-J., Wang, Z., June 2008. Joint Multi-Label Multi-Instance Learning for Image Classification. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR. IEEE Computer Society, pp. 1–8.

[22] Zhang, M.-L., Zhou, Z.-H., Dec. 2008. M^{3}MIML: A Maximum Margin Method for Multi-Instance Multi-Label Learning. In: Proceedings of the 2008 Eighth IEEE International Conference on Data Mining. ICDM '08. IEEE Computer Society, Washington, DC, USA, pp. 688–697.

[23] Zhou, Z.-H., Zhang, M.-L., 2007. Multi-Instance Multi-Label Learning with Application to Scene Classification. In: Advances in Neural Information Processing Systems, NIPS. MIT Press, Cambridge, MA, USA, pp. 1609–1616.