PARTIALLY TAGGED IMAGE CLUSTERING

Qiyue Yin, Shu Wu, Liang Wang

Center for Research on Intelligent Perception and Computing (CRIPAC)
National Laboratory of Pattern Recognition (NLPR)
Institute of Automation, Chinese Academy of Sciences, Beijing, China
{qyyin, shu.wu, wangliang}@nlpr.ia.ac.cn

ABSTRACT

With the exponential growth of tagged images, researchers are resorting to this high semantic tag information to assist the clustering process and promising clustering results have been obtained. However, users may not tag all of their images or some of the images are partially annotated, and this will lead to big performance degradation, which is rarely considered by pervious works. To alleviate this problem, we propose a new framework for image clustering assisted by partially observed tags. Our model enforces the sparse representation obtained through sparse coding and the latent tag representation learned via matrix factorization to be consistent with the partial image-tag observations. Finally, the partitioning of the database is performed using clustering algorithms (e.g., k-means) on the sparse representation. Extensive experiments on three real world datasets demonstrate that the proposed model performs better than the state-of-the-art methods.

Index Terms— Image clustering, partially observed tag information, multi-view clustering, sparse coding

1. INTRODUCTION

Image clustering, which assigns images into different groups, plays an important role for image organization, summarization and visualization [1]. Traditional image clustering algorithms usually resort to the visual features, such as SIFT descriptor. However, using this low level visual features is always ineffective because of the problem of semantic gap [2, 3]. To mitigate this problem, researchers are now exploring the textual information surrounding the images, such as the tags, as complementary high level semantic information to boost the clustering performance.

Several works have been proposed fusing the visual and textual features to improve the clustering performance [4, 5, 6, 7, 8]. Cai et al. [4] proposed a hierarchical clustering model to fuse the visual, textual and link information for the clustering of Web image search results. Similarly, Peng et al [9] utilized tags to obtain the topics as the first clustering layer and then used the visual features for more sophisticated clusters. Furthermore, Rege [8] proposed a co-clustering based framework for simultaneously integrating visual and textual features and then graph theory is applied for the final clustering. Recently, multi-view clustering, which fuses multiple sources of information for clustering tasks, provides a natural way for combining the visual and textual features. It achieves nice results and draws significant attention nowadays [10, 11, 12, 13, 14]. Generally, a wide variety of multi-view clustering works aim at finding a low dimensional embedding among the multiple features, and the complementary information is expected to be maximized in this learned latent space. Typical examples such as [15, 16, 11, 17, 18] obtain promising clustering results.

Though plenty of works have been proposed for image clustering utilizing both visual and textual features, few of them consider the scenario that the textual information is incomplete, which commonly exists in real applications. Compared with the visual features that can be extracted by representative descriptors, the images sometimes may not be annotated, or only given a few tags that are not abundant for the image description. In this circumstances, conventional methods may face a performance degradation for the largely dependency on the complete textual information. It should be noted that several works [17, 18] have been proposed solving the incomplete view problem for multi-view clustering. However, these methods mainly focus on the text data (e.g. Web pages clustering) and may not suitable for image clustering.

In this paper, we proposed a novel framework that focus on image clustering assisted by partially observed tag information. Our model utilizes tags to assist the learning of the visual representation, which consists of two parts. The first part is sparse coding and we learn sparse representation based on the visual feature, which can capture the salient structures of the images. In the second part, we learn latent representation for each tag, and keep the sparse representation and tag representation being consistent with the partial image-tag observations. Furthermore, an importance matrix is employed to deal with the situation that a tag is related to an image but not be observed. Finally, image clustering is achieved by performing clustering algorithms (e.g., k-means) on the learned sparse representation.

Our contribution in this paper is summarized as follows:

A novel framework for image clustering assisted by partially observed tag information is proposed, which is designed for the scenario that the tags for some images are totally missing or partially observed. To the best of our knowledge, this scenario is rarely considered for image clustering.

An effective optimizing algorithm for the proposed model is developed. And extensive experiments on three real world datasets show that the proposed model obtains better clustering results compared with several state-of-the-art methods.

2. PROPOSED MODEL

2.1. Problem Overview

In this paper, we use X^T to represent the transpose of matrix X. X^i and X_i indicate the i-th row and the i-th column of X respectively. X_{ij} is the entity of i-th row and j-th column of X. For two matrices X and I with the same size, we use $\odot$ to denote the element-wise product.

Suppose we have n data points and its visual feature is denoted as $X \in \Re^{p\times n}$ with p as the dimensionality. As for the textual feature T, assume we have q tags and if the i-th tag is annotated to the j-th image, T_{ij} is assigned to be 1, otherwise 0. It should be noted that T can be incomplete, which means some of the images have no tags or some tags of an image may be missing. Our task is to cluster this partially tagged images.

2.2. Formulation

Inspired by sparse coding, we assume that an image can be represented as a spare linear combination of the learned dictionary. Furthermore, we learn latent representation for each tag and use this latent feature to assist the learning of the sparse coding. Thus, the objective we are going to optimize is listed as follows:

$$\begin{aligned}&\min_{B,S,C} \|X - BS\|_F^2 + \alpha\|S\|_1 \\ &\quad + \lambda\left(\sum_{ij\in O}\left(T_{ij} - C^i S_j\right)^2 + \beta\|C\|_F^2\right) \\ &s.t.\ \ \|B_t\|^2 \le 1,\ \ \forall t\end{aligned} \tag{1}$$

where B and S are the learned dictionary and the sparse representation respectively. C is the latent representation for all the tags and O is the observed tag-image set. The parameters α, λ and β are scalars balancing different terms. After the optimization, we can use clustering algorithms, such as k-means, on S for the final data partitioning.

The purpose of the third term in Equation (1) is to enforce the learned two representations to be consistent with the partial image-tag observations. More specifically, we model the consistency using the latent factor model via matrix factorization, namely, we constrain the dot product of the learned tag representation and the sparse representation to approach matrix T. By doing so, the sparse representations of two images will be close if they have similar tag information. The term $\|C\|_F^2$ is a regularizer to avoid over-fitting.

In our hypothesis, $T_{ij} = 0$ can be interpreted into two ways that the i-th tag is not related to image j or it is missing. So we employ an importance matrix $I \in \Re^{q\times n}$ with the same size of T to alleviate the missing situation. And the objective is reformulated as:

$$\begin{aligned}&\min_{B,S,C} \|X - BS\|_F^2 + \alpha\|S\|_1 \\ &\quad + \lambda\left(\sum_{ij} I_{ij}\left(T_{ij} - C^i S_j\right)^2 + \beta\|C\|_F^2\right) \\ &s.t.\ \ \|B_t\|^2 \le 1,\ \ \forall t\end{aligned} \tag{2}$$

For simplicity, we define I as follows:

$$I_{ij} = \begin{cases} a & if\ \ T_{ij} = 1 \\ b & if\ \ T_{ij} = 0 \end{cases} \tag{3}$$

where a and b are two scalars satisfying $a > b > 0$.

Finally, We write the above equation as a compact matrix form:

$$\begin{aligned}&\min_{B,S,C} \|X - BS\|_F^2 + \alpha\|S\|_1 \\ &\quad + \lambda\left(\|L\odot(T - CS)\|_F^2 + \beta\|C\|_F^2\right) \\ &s.t.\ \ \|B_t\|^2 \le 1,\ \ \forall t\end{aligned} \tag{4}$$

where $L = I^{1/2}$ is the element-wise square of matrix I.

3. SOLUTION TO THE PROPOSED MODEL

Since variables B, S and C are coupled together and it may be difficult to solve them jointly, we propose to optimize the three variables alternatively until converge. Thus we will get three very simple sub-problems.

Solve S with B and C fixed. The problem becomes:

$$\min_S \|X - BS\|_F^2 + \lambda\|L\odot(T - CS)\|_F^2 + \alpha\|S\|_1 \tag{5}$$

For each column S_i, we have:

$$\min_{S_i}\left\|\begin{bmatrix} X_i \\ \sqrt{\lambda}L_i\odot T_i\end{bmatrix} - \begin{bmatrix} B \\ \sqrt{\lambda}diag\left(L_i\right)C\end{bmatrix}S_i\right\|^2 + \alpha\|S_i\|_1 \tag{6}$$

where $diag(v)$ denotes diagonal matrix with its diagonal elements being the vector v. This is a standard sparse representation problem, which can be solved using SLEP packages[1].

Solve C with B and S fixed. The problem is written as:

$$\min_C \|L\cdot(T - CS)\|_F^2 + \beta\|C\|_F^2 \tag{7}$$

For each row C^i, the problem is simplified as:

$$\min_{C^i} \left\|T^i diag(L^i) - C^i S diag(L^i)\right\|^2 + \beta\left\|C^i\right\|^2 \tag{8}$$

[1] http://parnec.nuaa.edu.cn/jliu/largeScaleSparseLearning.htm

where we can easily obtain the analytic solutions for C^i.

Solve B with S and C fixed. We have the following problem:

$$\min_{B} \|X - BS\|_F^2 \qquad s.t. \|B_t\|^2 \leq 1 \tag{9}$$

which can be optimized through Lagrangian method. Suppose the size of the dictionary is k, then it becomes:

$$L(B, \phi) = \|X - BS\|_F^2 + \sum_{i=1}^{k} \phi_i (\|B_i\| - 1) \tag{10}$$

where ϕ_i is a positive scalar indicating the Lagrange multiplie. Based on the derivation to B, we can obtain the close form solution as:

$$B = XS^T \left(SS^T + \varphi\right)^{-1} \tag{11}$$

where φ is a diagonal matrix with its i-th entity being $\varphi_{ii} = \phi_i$. And it can be optimized through the Lagrange dual problem $\min_{\varphi_{ii}>0} Tr\left(XS^T(SS^T + \varphi)^{-1} SX^T\right) + Tr(\varphi)$, which is easily solved using conjugate gradient. The whole procedure of the proposed image clustering method is summarized in Algorithm 1.

Algorithm 1 Partially **T**agged **I**mage **C**lustering (**PTIC**)

Input:

Visual feature X, partially observed tag matrix T, the latent dimensionality of S and the number of clusters;

1: Initialize B and C by random matrices.
2: **while** not converge **do**
3: Fix B and C, update S using Equation 6;
4: Fix S and B, update C using Equation 8;
5: Fix S and C, update B using Equation 11;
6: **end while**
7: Perform k-means clustering algorithm on S.

Output:

Image groups based on the presetting number of clusters

4. EXPERIMENTS

4.1. Datasets

Pascal VOC 2007 dataset[2] It consists of 20 categories with a total of 9,963 image-tag pairs. We use Color feature as the visual representation and the dimension of tag feature is 399. Furthermore, those image-tag pairs with multiple categories are removed. Then we have 5,649 image-tag pairs with 30 images have no tags.

NUS WIDE dataset[3] The database is crawled from Flickr and it consists of 269,648 images in 81 categories. Six types of low level features are extracted and the 500D bag of words description is utilized as the visual feature here. As for the tag, the 1,000D tag collection is used. We select the first ten categories with each class consisting of 200 images as a subset to evaluate the proposed method. Note that in this database, 723 images have no tag information.

MIR Flickr dataset[4] It has 15,000 image-tag pairs distributed in 38 categories. The authors provide seven types of low-level features and 2,000D most frequently used tags. The 960D GIST feature is employed as the visual feature here. Furthermore, we select 5 categories with the largest numbers of images as a subset for the experiments. In total, we have 7,933 image-tag pairs with 1,355 images have no tags.

4.2. Experimental settings

We compare our model with the baselines "k-means" and "Sparse Coding" that use no tag information and representative works "PairwiseSC", "CentroidSC" [15] and "PVC" [18] utilizing both visual and tag features. For methods 'PairwiseSC" and "CentroidSC", we follow [15] and choose the mean value of the Euclidean distance between all data points as the standard deviation for constructing the Gaussian kernel. As for "PVC", which is designed for incomplete feature representations, is implemented using the code released by the authors.

For our method "PTIC", the dimension of the sparse coding is chosen to be 300 in the three datasets and we will test its influence in the parameter selection part. Besides, we assign the values of the importance matrix to be 1 and 0.01 in all the experiments and will obtain good results. As k-means is used in all the experiments, it is run 20 times with random initialization. Two widely used metrics, i.e., the accuracy (ACC) and the normalized mutual information (NMI), are utilized to measure the clustering performance. We recommend readers referring to [19] for more details about their definitions.

4.3. Experimental results on the three databases

Table 1 shows the clustering accuracy and normalized mutual information of different methods on the three databases. Overall, it can be seen that our method outperforms all the compared methods. Since tag information has much higher semantic representation than that of visual feature, "k-means" and "Sparse Coding" algorithms obtain much worse results than the other methods using tag information.

"PairwiseSC" and "CentroidSC" aim to find a latent space that makes the visual and tag representations being similar, and this will harm the learning process if some of the images have no tag feature. Compared with them, our model utilizes tags to assist the learning process of the sparse representation, which may be less effected confronting the scenario that the tag information is incomplete.

As for "PVC", it learns a unified latent representation for data points having complete visual and tag features based on

[2] http://pascallin.ecs.soton.ac.uk/challenges/VOC/voc2007/

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

[4] http://www.cs.toronto.edu/ nitish/multimodal/index.html

Methods	VOC		NUS		MIR	
	ACC (%)	NMI (%)	ACC (%)	NMI (%)	ACC (%)	NMI (%)
k-means	12.13	6.11	19.95	6.51	31.66	5.69
Sparse Coding	15.22	6.08	20.13	6.58	32.31	6.04
PairwiseSC	53.20	52.23	38.62	26.00	41.17	9.26
CentroidSC	50.76	49.86	38.51	31.64	41.49	8.48
PVC	52.97	51.51	31.08	23.05	35.71	6.65
PTIC	**56.56**	**53.37**	**42.06**	**34.57**	**43.13**	**9.89**

Table 1. Clustering results in terms of ACC and NMI on VOC, NUS and MIR databases.

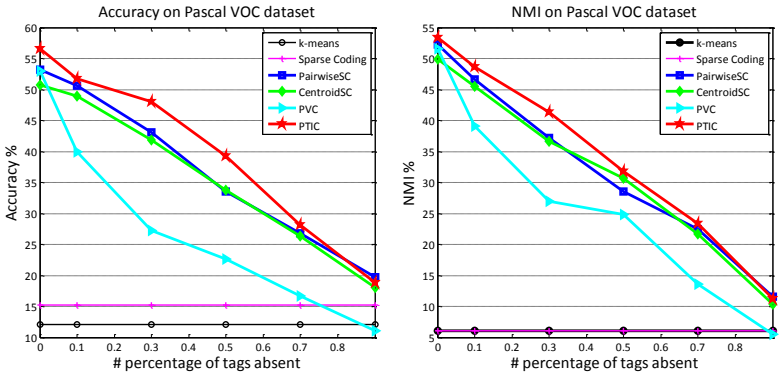

Fig. 1. Clustering accuracy and normalized mutual information when some images have partially observed tags on VOC dataset.

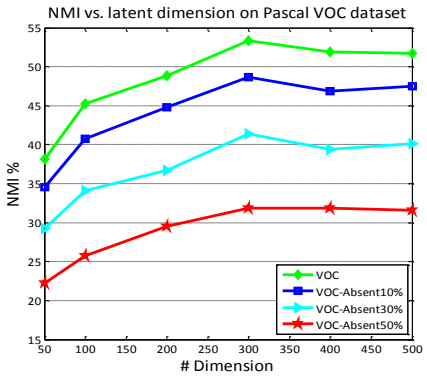

Fig. 2. Clustering normalized mutual information vs. the dimension on VOC dataset.

non-negative matrix factorization, which is effective dealing with the text data. Compared with "PVC", we can obtain the salient structures through sparse coding on the visual feature and assist the learning process through matrix factorization on the tag feature, which is more suitable for partially tagged image clustering.

4.4. Further results of partially observed tag information

In this section, We evaluate the performance of our method facing the situation that some images have partially observed tags, which is different from Section 4.3 that some images have no tags. This is also a practical scenario because users may omit some tags when annotating images. To mimic this scenario, we randomly remove a certain percentage of tags and test the performance of all the methods. We report results on VOC database for space limitation and the other two datasets show similar results.

From Figure 1, it can be seen that our model performs better with the increasing percentages of tags removed. This may because the importance matrix we use can alleviate the tag missing situation. As most of the tags being removed, our model has no prominent improvements over the other methods, and this is reasonable since the tag information is badly polluted to be nice complementary information.

4.5. Parameter Selection

In our model, λ balances the sparse coding of the visual feature and the matrix factorization for the partially observed tag feature. It is selected through searching from the interval [1, 20]. As for the regularizer parameter α, it is chosen following the rules of SLEP package. In this section, we test the clustering performance vs. the latent dimension of the sparse representation. To save space, only the results on VOC dataset are reported, and the other databases show similar results.

In Figure 2, VOC-Absent10%, VOC-Absent30% and VOC-Absent50% mean 10%, 30% and 50% percentages of tags are removed on the VOC dataset. As the dimensions of sparse representation increase, more information can be embedded and better clustering performance will be obtained. However, as the dimension is large enough, the clustering results stop increasing because of the saturated representation ability of the features.

5. CONCLUSION

In this paper, we have proposed a novel image clustering framework that utilize partially observed tags as the complementary information. By enforcing the sparse representation and the learned latent tag representation to be consistent with the partial image-tag observation, we learn better image representation for the final clustering. We have also developed an effective iterative optimization algorithm to solve the above problem. Extensive experiments have demonstrated the effectiveness of our proposed method compared with several state-of-the-art methods.

6. REFERENCES

[1] Alex Rodriguez and Alessandro Laio, "Clustering by fast search and find of density peaks," *Science*, vol. 344, pp. 1492–1496, 2014.

[2] Jile Zhou, Guiguang Ding, and Yuchen Guo, "Latent semantic sparse hashing for cross-modal similarity search," *In SIGIR*, pp. 415–424, 2014.

[3] Hao Ma, Jianke Zhu, Michael R. Lyu, and Irwin King, "Bridging the semantic gap between image contents and tags," *TMM*, vol. 12, pp. 462–473, 2010.

[4] Deng Cai, Xiaofei He, Zhiwei Li, Wei-Ying Ma, and Ji-Rong Wen, "Hierarchical clustering of www image search results using visual, textual and link information," *In WWW*, pp. 952–959, 2004.

[5] Symeon Papadopoulos, Christos Zigkolis, Giorgos Tolias, Yannis Kalantidis, Phivos Mylonas, Yiannis Kompatsiaris, and Athena Vakali, "Image clustering through community detection on hybrid image similarity graphs," *In ICIP*, pp. 2353–2356, 2010.

[6] Symeon Papadopoulos, Christos Zigkolis, Giorgos Tolias, Yannis Kalantidis, Phivos Mylonas, Yiannis Kompatsiaris, and Athena Vakali, "Web image co-clustering based on tag and image content fusion," *In International Conference on Network Infrastructure and Digital Content*, pp. 378–382, 2010.

[7] Pierre-Aiain Moellic, Jean-Emmanuel Hauqeard, and Guillaume Pitel, "Image clustering based on a shared nearest neighbors approach for tagged collections," *In CIVR*, pp. 269–278, 2008.

[8] Manjeet Rege, Ming Dong, and Jing Hua, "Graph theoretical framework for simultaneously integrating visual and textual features for efficient web image clustering," *In WWW*, pp. 317–326, 2008.

[9] Jinye Peng, Yi Shen, and Jianping Fan, "Cross-modal social image clustering and tag cleansing," *Journal of Visual Communication and Image Representation*, vol. 24, pp. 895–910, 2013.

[10] Shiliang Sun, "A survey of multi-view machine learning," *Neural Computing and Applications*, vol. 23, pp. 2031–2038, 2013.

[11] Yuhong Guo, "Convex subspace representation learning from multi-view data," *In AAAI*, pp. 387–393, 2013.

[12] Xiaowen Dong, Pascal Frossard, Pierre Vandergheynst, and Nikolai Nefedov, "Clustering on multi-layer graphs via subspace analysis on grassmann manifolds," *TIP*, vol. 62, pp. 905–918, 2014.

[13] Rongkai Xia, Yan Pan, Lei Du, and Jian Yin, "Robust multi-view spectral clustering via low-rank and sparse decomposition," *In AAAI*, pp. 2149–2155, 2014.

[14] Mingjie Qian and Chengxiang Zhai, "Unsupervised feature selection for multi-view clustering on text-image web news data," *In CIKM*, pp. 1963–1966, 2014.

[15] Abhishek Kumar, Piyush Rai, and Hal Daum Iii, "Co-regularized multi-view spectral clustering," *In ICML*, pp. 1413–1421, 2011.

[16] Jialu Liu, Chi Wang, Jing Gao, and Jiawei Han, "Multi-view clustering via joint nonnegative matrix factorization," *In SDM*, pp. 252–260, 2013.

[17] Anusua Trivedi, Hal Daum III, and Scott L. DuVall, "Multiview clustering with incomplete views," *In NIPS Workshop on Machine Learning for Social Computing*, 2010.

[18] Shao-Yuan Li, Yuan Jiang, and Zhi-Hua Zhou, "Partial multi-view clustering," *In AAAI*, pp. 1968–1974, 2014.

[19] Wen-Yen Chen, Yangqiu Song, Hongjie Bai, Chih-Jen Lin, and E.Y. Chang, "Parallel spectral clustering in distributed systems," *TPAMI*, vol. 33, pp. 568–586, 2010.