Matrix Completion for Cross-view Pairwise Constraint Propagation

Zheng Yang, Yao Hu, Haifeng Liu*, Huajun Chen, Zhaohui Wu
College of Computer Science, Zhejiang University
Hangzhou, Zhejiang, China
{yzzju,yaoohu,haifengliu,huajunsir,wzh}@zju.edu.cn

ABSTRACT

As pairwise constraints are usually easier to access than label information, pairwise constraint propagation attracts more and more attention in semi-supervised learning. Most existing pairwise constraint propagation methods are based on canonical graph propagation model, which heavily depends on the edge weights in the graph and cannot preserve local and global consistency simultaneously. In order to address this drawback, we cast cross-view pairwise constraint propagation into a problem of low rank matrix completion and propose a Matrix Completion method for cross-view Pairwise Constraint Propagation(MCPCP). With low rank requirement and graph regularization, our MCPCP can preserve local and global consistency simultaneously. We develop an algorithm based on alternating direction method of multipliers(ADMM) to solve the optimization problem. Finally, the effectiveness of MCPCP is demonstrated in cross-view multimedia retrieval.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

General Terms

Algorithms

Keywords

pairwise constraint propagation; matrix completion; multimedia retrieval

1. INTRODUCTION

Pairwise constraints, which specify whether two objects belong to the same class or not, are easier to access than label information and have been widely used for many machine learning tasks, like constrained clustering, multimedia retrieval[8], semi-supervised classification and so on. Pairwise constraint propagation is an effective algorithm to solve these tasks. It aims to fully utilize the available constraints through constraint propagation. LGC[10] and GFH[11] are two canonical label propagation methods, which propagate labels through a graph. Most pairwise constraint propagation methods are based on this kind of graph model. Given a dataset $\mathcal{X}$ with n points, the first p of them are labeled with c labels and the remaining points are unlabeled. LGC regularization framework has the following form:

$$\min_F \frac{1}{2}\left(\sum_{i,j=1}^{n} W_{ij}\left\|\frac{1}{\sqrt{D_{ii}}}F_i-\frac{1}{\sqrt{D_{jj}}}F_j\right\|^2+\mu\sum_{i=1}^{n}\|F_i-Y_i\|^2\right), \tag{1}$$

where $W \in \mathcal{R}^{n\times n}$ is a weight matrix of the dataset, D is a diagonal matrix with $D_{ii} = \sum W_{ij}$ and $\mu > 0$ is a regularization parameter. F and Y are $n \times c$ indicator response matrices. Y is the initial one with $Y_{ij} = 1$ if x_i is labeled as $y_i = j$ and $Y_{ij} = 0$ otherwise. And F is to be learned which labels each point x_i as a label $y_i = \arg\max_{1\le j\le c} F_{ij}$. F_i, F_j and Y_i are all row vectors which represent the label assignments.

As we know, the key to the success of semi-supervised learning is the cluster assumption[11], which indicates nearby objects are likely to have the same label and objects on the same structure are likely to have the same label. The first assumption is local consistency and the second one is global consistency. The two terms in problem (1) aim at these assumptions, respectively. However, the second term cannot preserve global consistency well when the number of labeled samples is not big enough to reveal global structure. Hence, the graph becomes more important as both local and global consistency depend on it. In the past years, many researchers have devoted to constructing better graph[12, 4]. Besides, the second term requires that the final classifying functions should not change too much from the initial label assignments. It is reasonable for the labeled samples but may be ill for unlabeled samples.

In this paper, we focus our attention on cross-view pairwise constraint propagation. Different from canonical methods, we cast it into a matrix completion problem: we use a **relation matrix** to represent the pairwise constraint between objects from different media views. Each entry in the relation matrix is a real number that represents the relevance of two corresponding objects. Some of the entries are available in advance. To complete the relation matrix, we propose a Matrix Complete algorithm for cross-view Pairwise Constraint Propagation(MCPCP). MCPCP

*The corresponding author.

MM'14, November 03 - 07 2014, Orlando, Florida, USA.

http://dx.doi.org/10.1145/2647868.2655057 .

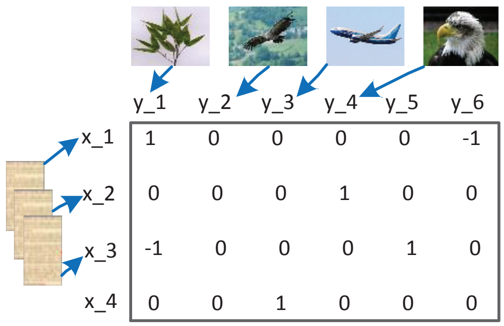

Figure 1: Example of relation matrix Y.

requires that the relation matrix should be low rank and fixes the values at observed entries. It has been proved that the low-rank representation can capture the global structure(such as multiple clusters and subspaces)[6]. These two are fundamental requirements for matrix completion. Besides, MCPCP introduces graph regularization to preserve local consistency. So our MCPCP can preserve local and global consistency simultaneously.

2. CROSS-VIEW PAIRWISE CONSTRAINT PROPAGATION ALGORITHM

In this section, we define cross-view Pairwise Constraint Propagation problem and introduce our Matrix Complete algorithm for cross-view Pairwise Constraint Propagation.

Since cross-view problem can be readily decomposed into a series of two-view subproblems. For the convenience of interpretation, we just focus on two-view problem in this paper. Let $\{\mathcal{X}, \mathcal{Y}\}$ be a two-view dataset, where $\mathcal{X} = \{x_1, x_2, \ldots, x_m\}$ and $\mathcal{Y} = \{y_1, y_2, \ldots, y_n\}$. Pairwise constraints on the two-view dataset consist of must-link constrain set $\mathcal{M} = \{(x_i, y_j) : l(x_i) = l(y_j)\}$ and cannot-link constrain set $\mathcal{C} = \{(x_i, y_j) : l(x_i) \neq l(y_j)\}$. We denote relation matrix $F \in \mathcal{R}^{m\times n}$ to represent the real relevance of objects from different views. Note that, positive f_{ij} correspond to a must-link constraint and negative f_{ij} correspond to a cannot-link. The absolute value of f_{ij} denotes confidence score. With this definition, we denote initial(or observed) relation matrix over the dataset with matrix $Y \in \mathcal{R}^{m\times n}$, where $Y_{ij} = 1$ if $(x_i, y_j) \in \mathcal{M}$, $Y_{ij} = -1$ if $(x_i, y_j) \in \mathcal{C}$, $Y_{ij} = 0$ otherwise. Initial relation matrix Y is illustrated in Figure 1. We achieve pairwise constraint propagation by completing the entries in Y whose values are 0 in the beginning.

Canonical methods achieve this goal by letting every point iteratively spread its constraint information to its neighbors until a global stable state is achieved. Actually, initial relation matrix Y can be viewed as an incomplete matrix if we consider the entries corresponding with $\mathcal{M}$ and $\mathcal{C}$ as observed entries and the others as unknown entries, respectively. Cross-view Pairwise Constraint Propagation aims to learn the complete matrix which can reveal all the relations. Note that, the complete relation matrix F^* is a low-rank matrix. Considering row/column relation vectors as feature vectors of row/column objects, we know that both row objects and column objects can be always clustered into some clusters. It is reasonable to expect that each cluster corresponds to a linear intrinsic subspace, which means there exist a small number of underlying intrinsic relation vectors in F^* such that all the relation vectors are derived from a linear combination of these intrinsic relation vectors. This is called low rank assumption. Therefore, Pairwise Constraint Propagation is a low-rank matrix completion problem.

Now, we consider low-rank Matrix Completion for cross-view Pairwise Constraint Propagation(MCPCP). Liu et al.[7] have demonstrated low-rank representation can capture the global structure(such as multiple clusters and subspaces) of the whole data. A good semi-supervised algorithm should preserve global and local consistency simultaneously. To preserve local consistency, we introduce graph regularization. Given graph $\mathcal{G}_x = (\mathcal{X}, W_x)$ and $\mathcal{G}_y = (\mathcal{Y}, W_y)$ constructed over the dataset $\mathcal{X}$ and $\mathcal{Y}$, we achieve cross-view Pairwise Constraint Propagation by solving the following optimization problem:

$$
\begin{aligned}
&\min_F \mathcal{L}(F) \\
=&rank\|F\| + \beta_x \sum_{i,j=1}^{m} W_{x_{ij}} \left\| \frac{1}{\sqrt{D_{x_{ii}}}} F_{i:} - \frac{1}{\sqrt{D_{x_{jj}}}} F_{j:} \right\|^2 \\
&+ \beta_y \sum_{i,j=1}^{n} W_{y_{ij}} \left\| \frac{1}{\sqrt{D_{y_{ii}}}} F_{:i} - \frac{1}{\sqrt{D_{y_{jj}}}} F_{:j} \right\|^2 \\
&s.t. \quad P_\Omega(F) = Y
\end{aligned}
\tag{2}
$$

where

$$
[P_\Omega(F)]_{ij} = \begin{cases} F_{ij} & \text{if } (i,j) \in \Omega \\ \mathbf{0} & \text{if } (i,j) \notin \Omega \end{cases}, \tag{3}
$$

D_x and D_y are diagonal matrix with $D_{x_{ii}} = \sum_j W_{x_{ij}}$ and $D_{y_{ii}} = \sum_j W_{y_{ij}}$. $\beta_x, \beta_y > 0$ are trade-off parameters between local smoothness and low rankness. And Ω represents the observed(nonzero) entries in Y.

3. OPTIMIZATION

Because of rank minimization term, solving problem (2) is NP-hard. Fortunately, it has been proved that minimization of the rank function can be achieved using the minimizer obtained with the Nuclear Norm under broad conditions[3]. So we relax problem (2) by solving the following function:

$$
\begin{aligned}
\min_F \mathcal{L}(F) =& \|F\|_* + \beta_x Tr(F^T L_x F) + \beta_y Tr(F L_y F^T) \\
& s.t. \quad P_\Omega(F) = Y
\end{aligned}
\tag{4}
$$

where $L_x = I - S_x$ and $L_y = I - S_y$ with $S_x = D_x^{-1/2} W_x D_x^{-1/2}$ and $S_y = D_y^{-1/2} W_y D_y^{-1/2}$. $\|F\|_* = \sum_{i=1}^{min(m,n)} \sigma_i(F)$ is the Nuclear Norm and $\sigma_i(F)$ is the ith largest singular value of F.

Next, we use ADMM method[1] to solve problem (4). In order to make the objective function separable, we first introduce two auxiliary variables H and Q:

$$
\begin{aligned}
\min_F \mathcal{L}(F) =& \|F\|_* + \beta_x Tr(H^T L_x H) + \beta_y Tr(Q L_y Q^T) \\
s.t. \quad & H = F, P_\Omega(H) = Y \\
& Q = F, P_\Omega(Q) = Y
\end{aligned}
\tag{5}
$$

The augmented lagrange function of (5) is

$$\begin{aligned}&\mathcal{L}(F,H,Q,Z1,Z2,\mu)\\ =&\|F\|_* + \beta_x Tr(H^T L_x H) + \beta_y Tr(Q L_y Q^T)\\ &+ Tr(Z1^T(H-F)) + Tr(Z2^T(Q-F))\\ &+ \frac{\mu}{2}[\|H-F\|^2 + \|Q-F\|^2]\end{aligned} \tag{6}$$

where $\mu > 0$ is the penalty parameter. Given the initial setting $F_1 = Y$, $H_1 = Y$, $Q_1 = Y$, $Z1_1 = Y$ and $Z2_1 = Y$, the optimization problem (5) can be solved via the following subproblems:

Subproblem F:

$$\begin{aligned}F_{k+1} =& \arg\min_F \mathcal{L}(F,H_k,Q_k,Z1_k,Z2_k,\mu)\\ =& \arg\min_F \|F\|_* + Tr(Z1_k^T(H_k-F))\\ &+ Tr(Z2_k^T(Q_k-F)) + \frac{\mu}{2}[\|H_k-F\|^2 + \|Q_k-F\|^2]\end{aligned} \tag{7}$$

Ignoring constant terms and with some algebra, this can be rewritten as

$$F_{k+1} = \arg\min_F \|F\|_* + \mu\|F - \frac{1}{2}(H_k + Q_k + \frac{Z1_k + Z2_k}{\mu})\|^2 \tag{8}$$

So the closed form solution of problem (7) is

$$F_{k+1} = \mathcal{D}_{\frac{1}{2\mu}}\big(\frac{1}{2}(H_k + Q_k + \frac{1}{\mu}(Z1_k + Z2_k)\big) \tag{9}$$

where $\mathcal{D}$ is singular value thresholding operator[3].

Subproblem H:

$$\begin{aligned}H_{k+1} =& \arg\min_H \beta_x Tr(H^T L_x H) + Tr(Z1_k^T(H - F_{k+1}))\\ &+ \frac{\mu}{2}\|H - F_{k+1}\|^2.\\ s.t.\quad & P_\Omega(H) = Y\end{aligned} \tag{10}$$

Let the partial derivatives with respect to H vanish. With some algebra, we can get the following equation:

$$(2\beta_x L_x + \mu I)H = \mu F_{k+1} - Z1_k. \tag{11}$$

Since $2\beta_x L_x + \mu I$ is positive definite, we have:

$$H_{k+1} = (2\beta_x L_x + \mu I)^{-1}(\mu F_{k+1} - Z1_k). \tag{12}$$

Actually, the matrix inverse $(2\beta_x L_x + \mu I)^{-1}$ can be calculate in advance and only need to be calculated once. Then, we fix the observed values and obtain

$$H_{k+1} = Y + P_{\Omega^c}(H_{k+1}). \tag{13}$$

Algorithm 1: MCPCP

Input: Y, W_x, W_y, β_x, β_y, ϵ and μ

Initialize: $F_1 = Y$, $H_1 = Y$, $Q_1 = Y$, $Z1_1 = Y$ and $Z2_1 = Y$

Output: F.

1) Calculate L_x by $L_x = I - S_x$, where $S_x = D^{-1/2} W_x D^{-1/2}$ and D is a diagonal matrix with its $D_{ii} = \sum_j W_{xij}$.
2) Calculate L_y by $L_y = I - S_y$, where $S_y = D^{-1/2} W_y D^{-1/2}$ and D is a diagonal matrix with its $D_{ii} = \sum_j W_{yij}$.
3) $A = (2\beta_x L_x + \mu I)^{-1}$.
4) $B = (2\beta_y L_y + \mu I)^{-1}$.
5) **repeat:**
6) Update F_{k+1} by Eq. (9).
7) Update H_{k+1} by $H_{k+1} = A(\mu F_{k+1} - Z1_k)$. Fix values at observed entries $H_{k+1} = P_\Omega(Y) + P_{\Omega^c}(H_{k+1})$.
8) Update Q_{k+1} by $Q_{k+1} = (\mu F_{k+1} - Z2_k)B$. Fix values at observed entries $Q_{k+1} = P_\Omega(Y) + P_{\Omega^c}(Q_{k+1})$.
9) $Z1_{k+1} = Z1_k + \mu(H_{k+1} - F_{k+1})$.
10) $Z2_{k+1} = Z2_k + \mu(Q_{k+1} - H_{k+1})$.
11) **until** $\|F_{k+1} - F_k\|^2 \le \epsilon$.

Subproblem Q:

$$\begin{aligned}Q_{k+1} =& \arg\min_Q \beta_y Tr(Q L_y Q^T) + Tr(Z2_k^T(Q - F_{k+1}))\\ &+ \frac{\mu}{2}\|Q - F_{k+1}\|^2\\ s.t.\quad & P_\Omega(Q) = Y\end{aligned} \tag{14}$$

Similar with *Subproblem H*, we have

$$Q_{k+1} = (\mu F_{k+1} - Z2_k)(2\beta_y L_y + \mu I)^{-1} \tag{15}$$

and

$$Q_{k+1} = Y + P_{\Omega^c}(Q_{k+1}). \tag{16}$$

Subproblem Z1 and Z2:

$Z1$ and $Z2$ can be calculated as follows:

$$Z1_{k+1} = Z1_k + \mu(H_{k+1} - F_{k+1}); \tag{17}$$

$$Z2_{k+1} = Z2_k + \mu(Q_{k+1} - H_{k+1}). \tag{18}$$

We summarize the whole procedure of our solution for problem (5), which is called MCPCP, in Algorithm 1. It is apparent that the main computational cost in each iteration is the computation of SVD in singular value thresholding operator. However, computing the full SVD for MCPCP is unnecessary. So PROPACK[5] can be used to accelerate the computation and make our algorithm more efficient for large-size problems. What is more, we use k-NN graph, which means L_x and L_y are sparse matrixes and the inversion calculation can be solved very efficiently as Cai *.et al* did in GNMF[2]. The convergence of Algorithm 1 is guaranteed by the ADMM[1].

Table 1: The retrieval results on Wiki data set by the MAP scores. The bold numbers are the best results.

Algorithm	Image Query	Text Query	Average
CM	0.2706	0.2294	0.250
Inter-CP	0.552	0.546	0.549
MCPCP	**0.576**	**0.556**	**0.566**

4. EXPERIMENTAL RESULTS

In this section, we evaluate our MCPCP on real world application. As the obtained relation matrix F^* represents the relevance between objects from different views. Naturally, our MCPCP can be used in cross-view retrieval by sorting the relevant relation vectors. The most straightforward way for cross-view retrieval is similarity comparison. We compare our method with two representative methods, Correlation matching(CM) and Inter-CP[8]. The first is a similarity comparison method and the second is a pairwise constraint propagation method.

Dataset: We use Wiki text-image dataset[9] in our experiment. The Wiki dataset consists of 2866 text-image pairs from 10 different classes. Among them, texts are represented by 10 dimensional latent Dirichlet allocation model and each image has a 128 dimensional SIFT histogram feature.

Settings: Different from [8] which divided the data points into training and testing part and obtained all the pairwise constraints of training data from the label information, we adopted another way. We do not divide the data and use less pairwise constraints. The initial pairwise constraints contain must-link constraints obtained from paired points(text-image) and the same number of cannot-link constraints. That is to say, we use 5732 pairwise constraints in our experiment and only half of them should be given manually. These cannot-link constraints are chosen randomly. For graph regularization, we use the normalized correlation as the similarity measure and construct two k-NN graphs on image and text sets, respectively. In our algorithm, there are four parameters. We set $k = 40$ for Inter-CP and our MCPCP, experimentally. All the other parameters are obtained by cross-validation. Specifically, we set $\beta_x = 1$, $\beta_y = 1$, $\mu = 100$ and $k = 40$ for MCPCP. For Inter-CP and our MCPCP, we achieve retrieval by sorting the relevant learned relation vector.

Results: Two tasks of cross-view retrieval are considered: text retrieval using an Image Query and image retrieval using a Text Query. The retrieval results are measured with mean average precision(MAP), a widely used method in the retrieval literature. The results are listed in Table 1. From the results, we can find the performance of MCPCP and Inter-CP is better than that of CM. This is because these two methods not only utilize pairwise constraints but local structure of each view. Considering global and local structure, our MCPCP achieves the best result which has 1.7% improvement in MAP averagely compared with that of Inter-CP.

5. CONCLUSION

In this paper, we proposed a novel method for pairwise constraint propagation by casting it into a problem of low-rank matrix completion. We analyzed this method and pointed out this model can capture global structure of the whole data. In order to preserve local consistency at the same time, we introduced graph regularization. The experimental results in cross-view retrieval demonstrated the effectiveness of our pairwise constraint propagation method. This method can be used in tag-completion, multi-label learning and other semi-supervised learning problems.

6. ACKNOWLEDGEMENTS

This work was supported in part by the National Basic Research Program of China (973 Program) under Grant 2013CB336500, in part by the National Natural Science Foundation of China under Grant 61379071 and Grant 91120302, and in part by the National High-tech R&D Program of China (863 Program) under Grant 2012AA02A601.

7. REFERENCES

[1] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine Learning*, 3(1):1–122, 2011.

[2] D. Cai, X. He, J. Han, and T. S. Huang. Graph regularized nonnegative matrix factorization for data representation. *IEEE Transactions on PAMI*, 33(8):1548–1560, 2011.

[3] J.-F. Cai, E. J. Candès, and Z. Shen. A singular value thresholding algorithm for matrix completion. *SIAM Journal on Optimization*, 20(4):1956–1982, 2010.

[4] M. Karasuyama and H. Mamitsuka. Manifold-based similarity adaptation for label propagation. In *NIPS*, pages 1547–1555, 2013.

[5] R. M. Larsen. Lanczos bidiagonalization with partial reorthogonalization. *DAIMI Report Series*, 27(537), 1998.

[6] G. Liu, Z. Lin, S. Yan, J. Sun, Y. Yu, and Y. Ma. Robust recovery of subspace structures by low-rank representation. *IEEE Transactions on PAMI*, 35(1):171–184, 2013.

[7] G. Liu, Z. Lin, and Y. Yu. Robust subspace segmentation by low-rank representation. In *ICML*, pages 663–670, 2010.

[8] Z. Lu and Y. Peng. Unified constraint propagation on multi-view data. In *AAAI*, 2013.

[9] N. Rasiwasia, J. Costa Pereira, E. Coviello, G. Doyle, G. R. Lanckriet, R. Levy, and N. Vasconcelos. A new approach to cross-modal multimedia retrieval. In *ACM MM*, pages 251–260. ACM, 2010.

[10] D. Zhou, O. Bousquet, T. N. Lal, J. Weston, and B. Schölkopf. Learning with local and global consistency. In *NIPS*, volume 16, pages 321–328, 2003.

[11] X. Zhu, Z. Ghahramani, J. Lafferty, et al. Semi-supervised learning using gaussian fields and harmonic functions. In *ICML*, volume 3, pages 912–919, 2003.

[12] L. Zhuang, H. Gao, Z. Lin, Y. Ma, X. Zhang, and N. Yu. Non-negative low rank and sparse graph for semi-supervised learning. In *CVPR*, pages 2328–2335. IEEE, 2012.