Optimal Graph Learning with Partial Tags and Multiple Features for Image and Video Annotation

Lianli Gao[1], Jingkuan Song[2], Feiping Nie[3], Yan Yan[2], Nicu Sebe[2], Heng Tao Shen[4]
[1]University of Electronic Science and Technology of China. [2]University of Trento. [3]University of Texas, Arlington. [4]The University of Queensland.

In multimedia annotation, due to the time constraints and the tediousness of manual tagging, it is quite common to utilize both tagged and untagged data to improve the performance of supervised learning when only limited tagged training data are available [1, 2]. This is often done by adding a geometrically based regularization term in the objective function of a supervised learning model. In this case, a similarity graph is indispensable to exploit the geometrical relationships among the training data points, and the graph construction scheme essentially determines the performance of these graph-based learning algorithms. However, most of the existing works construct the graph empirically and are usually based on a single feature without using the label information.

In this paper, we propose a semi-supervised annotation approach by learning an optimal graph (OGL) from multi-cues (i.e., partial tags and multiple features) which can more accurately embed the relationships among the data points. We further extend our model to address out-of-sample and noisy label issues.

Suppose that for each image, we have for v features. Let $X^t = \{x_i^t\}_{i=1}^n$ denote the feature matrix of the t-th feature of training images, where $t \in \{1,...,v\}$. The traditional graph based semi-supervised learning [3, 4, 5] usually solves the following problem:

$$\min_{F,F_i=Y_l} \sum_{ij} \|f_i - f_j\|_2^2 s_{ij} \tag{1}$$

where f_i and f_j are the labels for the i-th and j-th images, and S is the affinity graph with each entry s_{ij} representing the similarity between two images. The affinity graph $S \in \mathbb{R}^{n\times n}$ is usually defined as follows:

$$s_{ij} = \begin{cases} e^{-\|x_i - x_j\|_2^2/2\sigma^2}, \text{ if } x_i \in \mathcal{N}_K(x_j) \text{ or } x_j \in \mathcal{N}_K(x_i) \\ 0, \text{else} \end{cases} \tag{2}$$

where $\mathcal{N}_K(\cdot)$ is the K-nearest neighbor set and $1 \le (i,j) \le n$. The variance σ will affect the performance significantly, and it is usually empirically tuned. Also, the similarity graph is often derived from single information cue. To address these issues, we propose to learn an optimal graph S from multiple cues.

The multiple cues include the given label information F and the multiple features $X^t = \{x_i^t\}_{i=1}^n$. An optimal graph S should be smooth on all these information cues, which can be formulated as:

$$\min_{S,\alpha} g(F,S) + \mu \sum_{t=1}^{v} \alpha^t h(X^t,S) + \beta r(S,\alpha) \tag{3}$$

where $g(F,S)$ is the penalty function to measure the smoothness of S on the label information F and $h(X^t,S)$ is the loss function to measure the smoothness of S on the feature X^t. $r(S,\alpha)$ are regularizers defined on the target S and α. μ and β are balancing parameters, and α^t determines the importance of each feature.

The penalty function $g(F,S)$ and $h(X^t,S)$ should be defined in the way such that close labels (data points) have high similarity and vice versa. In this paper, we define them as follows:

$$\begin{cases} g(F,S) = \sum_{ij} \|f_i - f_j\|_2^2 s_{ij} \\ h(X^t,S) = \sum_{ij} \left\|x_i^t - x_j^t\right\|_2^2 s_{ij} \\ r(S,\alpha) = \frac{\mu\gamma}{\beta}\|S\|_F^2 + \|\alpha\|_2^2 \end{cases} \tag{4}$$

where f_i and f_j are the labels of data point x_i and x_j. We further constrain that $S \ge 0, S1 = 1, \alpha \ge 0$ and $\alpha^T 1 = 1$. Then we can obtain the objective function for learning the optimal graph by replacing $g(F,S)$, $h(X^t,S)$ and $r(S,\alpha)$ in Eq.3 using Eq.4. By combining Eq.1 with Eq.3, we can obtain the objective function for optimal-graph based SSL, as follows:

$$\begin{aligned} &\min_{S,F,\alpha} \sum_{ij} \|f_i - f_j\|_2^2 s_{ij} + \mu \sum_{tij} \left(\alpha_t \left\|x_i^t - x_j^t\right\|_2^2 s_{ij} \right) \\ &+\mu\gamma\|S\|_F^2 + \beta\|\alpha\|_2^2 \\ &s.t. \left\{ S \ge 0, S1 = 1,\ F_l = Y_l,\ \alpha \ge 0,\ \alpha^T 1 = 1 \right. \end{aligned} \tag{5}$$

We propose an iterative method to minimize the above objective function in Eq.5. Firstly, we initialize $S = \sum_t S^t/v$ with each S^t being calculated using Eq.2, and we initialize $\alpha^t = 1/v$. We further normalize S as $S = (D^{1/2})^T S D^{1/2}$. Once these initial values are given, in each iteration, we first update F given S and α, and then update S and α by fixing the other parameters. Our conclusion is that by learning an optimal graph (OGL) from multi-cues (i.e., partial tags and multiple features), the relationships among the data points can be more accurately embedded. Consequently, the performance of graph-based algorithms using OGL can potentially be improved.

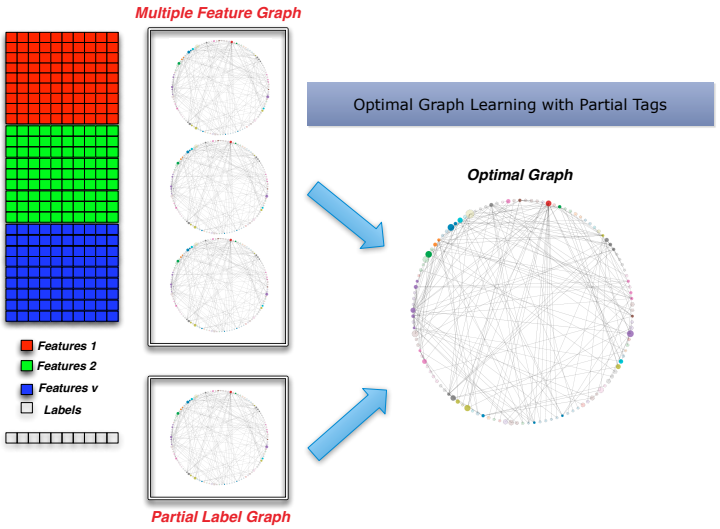

Figure 1: The overview of OGL. Firstly, a similarity graph is constructed on each feature (multiple feature graph) and also on the partial tags (partial label graph) to exploit the relationship among the data points. Partial tags means that tags are provided only for a part of the training data. Then, the optimal graph learning is applied to these graphs to construct an optimal graph, which is integrated with SSL for the task of image and video annotation.

[1] Minmin Chen, Alice Zheng, and Kilian Q. Weinberger. Fast image tagging. In *ICML*, 2013.

[2] Matthieu Guillaumin, Thomas Mensink, Jakob J. Verbeek, and Cordelia Schmid. Tagprop: Discriminative metric learning in nearest neighbor models for image auto-annotation. In *ICCV*, 2009.

[3] Jingkuan Song, Yi Yang, Zi Huang, Heng Tao Shen, and Richang Hong. Multiple feature hashing for real-time large scale near-duplicate video retrieval. In *ACM Multimedia*, 2011.

[4] Jingkuan Song, Yi Yang, Xuelong Li, Zi Huang, and Yang Yang. Robust hashing with local models for approximate similarity search. *IEEE Trans. Cybernetics*, 44(7):1225–1236, 2014.

[5] Dengyong Zhou, Olivier Bousquet, Thomas Navin Lal, Jason Weston, and Bernhard Schölkopf. Learning with local and global consistency. In *NIPS*, 2003.

This is an extended abstract. The full paper is available at the Computer Vision Foundation webpage.