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, China.
lianli.gao@uestc.edu.cn

[2]University of Trento, Italy
{jingkuan.song,yan.yan,niculae.sebe}@unitn.it

[3]University of Texas, Arlington
feipingnie@gmail.com

[4]The University of Queensland, Australia
shenht@itee.uq.edu.au

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. 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. Extensive experiments on four public datasets show the consistent superiority of OGL over state-of-the-art methods by up to 12% in terms of mean average precision.

1. Introduction

Recently, we have witnessed an exponential growth of user generated videos and images, due to the booming of social networks, such as Facebook and Flickr. Consequently, there are increasing demands to effectively organize and access these multimedia data via tagging. One promising direction is to combine labeled data (often of limited amount) and a huge pool of unlabeled data in forming abundant training resources for optimizing annotation models, which is referred to semi-supervised learning (SSL). This is often done by adding a geometrically based regularization term in the objective function of a supervised learning model. To exploit the geometrical relationships among the training data points, the algorithms treat both labeled and unlabeled samples as vertices (nodes) in a graph and build pairwise edges between these vertices which are weighed by the affinities (similarities) between the corresponding sample pairs.

SSL has been widely studied [3, 28, 13, 32, 4] and applied to many challenging tasks [4, 2, 3] such as image annotation and image retrieval. By exploiting a large number of unlabeled data with reasonable assumptions, SSL can reduce the need of expensive labeled data and thus achieve promising results especially for noisy labels [25]. The harmonic function approach [35] and local and global consistency (LGC) [34] are two representative graph-based SSL methods. The harmonic function approach [35] emphasizes the harmonic nature of the energy function and LGC considers the spread of label information in an iterative way. While these two methods are transductive, manifold regularization (MR) [1, 21] is inductive. In practice, MR extends regression and SVM to semi-supervised learning methods such as Laplacian Regularized Least Squares (LapRLS) and Laplacian Support Vector Machines (LapSVM) respectively by adding a geometrically based regularization term [17].

Recently, many applications [33, 25, 24] were proposed. Zhang *et al.* [33] extended LDA to semi-supervised discriminant analysis, Tang *et al.* [25] addressed the noisy label issue for the task of semi-supervised image labeling, and Song *et al.* [24] utilized weak-label information for cross-media retrieval.

Since an informative graph is critical for the graph-based algorithms, its construction has been extensively studied [27, 3, 18, 22, 30, 23]. The most popular way to construct a graph is the K-nearest neighbor (or ϵ-range neighbor) method, where, for each data point, the samples are connected with its K-nearest neighbors (or ϵ-range-neighbor). Then the Gaussian-kernel can be used to quantify the graphs. However, the tuning of σ in the Gaussian-kernel approach is empirical [27]. Recently, it has been proposed to learn the graph by considering the pairwise distance-based and the reconstruction coefficients-based methods. The former is based on the assumption that close data points should have a high similarity and vice v rsa. The latter ass mes that each data point can be reconstructed as a linear combination of the other data points. These two methods show different strengths and weaknesses in various applications. However, most of these graphs are constructed on a single information cue (e.g., visual feature, labels), and an optimal graph utilizing multiple cues has rarely been addressed.

To address the above issues, in this work, we propose learning an optimal graph (OGL) from multi- ues (i.e., partial tags and multiple features), which can more accurately encode the relationships between data points. Then, we incorporate the learned optimal graph with the SSL model, and we further extend this model to address out-of-sample extension and noisy label issues. It is worthwhile to highlight the following aspects of the proposed approach here:

- As far as we know, this is the first work to explicitly learn an optimal graph both from labels and multiple features and we propose an efficient way to solve the optimization problem. The learned optimal graph can automatically determine the confidence of the partial tags and different visual features to more precisely reflect the relationships among data points.
- Our optimal graph learning method is a general framework, and theoretically, it can be incorporated with other graph-based learning methods. Specifically, we integrate OGL with the SSL method, and evaluate our OGL for image and video annotation.
- We further discuss and propose solutions for out-of-sample extension, noisy label issues and different graph construction methods for our models, to deal with the real applications.
- Experiments for image and video annotation on real world datasets demonstrate the superiority of OGL over existing graph construction methods.

2. Related Work

There are generally two ways to build a similarity graph. One is based on pairwise distances (e.g., Euclidean distance), and the other is based on reconstruction coefficients. The first method is based on a reasonable assumption that close data points should have a high similarity and vice versa. The second method assumes that each data point could be represented as a linear combination of the other points. When the data are clean, i.e., the data points are strictly sampled from the subspaces, several approaches are able to recover the subspaces [16]. However, in real applications, the data set may lie at the union of multiple subspaces or contain noise and outliers [18]. As a result, inter-class data points may be connected with very high weights. Hence, eliminating the effects of errors becomes a major challenge. To address these pro lems, several algorithms have been proposed, e.g., Lo ally Linear Manifold Clust ring (LLMC) [7], Agglomerative Lossy Compression (ALC) [19], Sparse Subspace Clustering (SSC) [5], L1-graph [3, 29], Low Rank Representation (LRR) [10, 9], Latent Low Rank Representation (LatLRR) [11], Fixed Rank Representation (FRR) [12], L2Graph [18] (please refer to [26] for a comprehensive survey on these algorithms).

Of the above methods, SSC [5] and L1-graph [3] obtain a sparse similarity graph from the sparsest coefficients. One of the main differences between these two techniques is that [5] formulates the noise and outliers in the objective function and provides more theoretical analysis, whereas [3] derives a series of algorithms upon the L1-graph for various tasks. The popular LRR model [10, 9] and its extensions [11, 12] are very similar to SSC, except that they aim to obtain a similarity graph from the lowest-rank representation rather than the sparsest one. Both ℓ_1 and rank-minimization-based methods can automatically select neighbors for each data point by adopting the sparse solution, and have achieved impressive results in numerous applications. H wever, their computational complexity is proportional to the cube of the problem size Moreover, SSC requires that t e corruption over each data point has a sparse structure, and LRR assumes that only a small portion of the data is contaminated, otherwise the performance will degrade. In fact, these wo problems are mainly caused by the adopted error-handling strategy, i.e., remo ing the errors from the data set to obtain a clean dictionary over which each sample is encoded [18].

Differently, our OGL (described in the next section) eliminates the effects of errors and builds a more precise graph by considering both partial tags and multiple features, and automati ally assigning a higher confidence to good performed graphs constructed on each information cue.

3. Our Approach

In this section, we introduce our method OGL which consists of two phases (see Fig. 1). 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

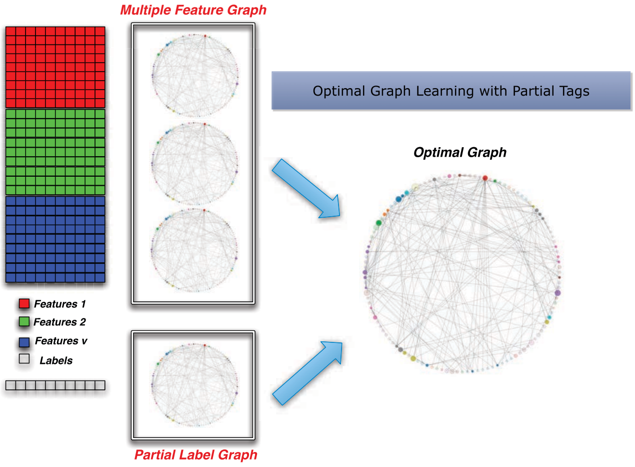

Figure 1. The overview of OGL.

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. Note that in theory our approach can be integrated with all kinds of graph-based algorithms.

3.1. Terms and Notations

We first introduce the notations which will be used in the rest of the paper. $X = \{x_1, x_2, ..., x_n\}$ represents a set of n images, and $y_i = \{0, 1\}^c$ is the label for the i-th image ($1 \leq i \leq n$), and c is the number of annotations/classes. The first l points x_i ($i \leq l$) are labeled as Y_l, and the remaining u points x_i ($l + 1 \leq i \leq n$) are unlabeled. The goal of transductive graph-based SSL is to predict the label F_u of the unlabeled points. Define a $n \times c$ matrix $F = \begin{bmatrix} F_l \\ F_u \end{bmatrix}$ with $F_l = Y_l$ and $F_u = \{0\}^{u*c}$.

Suppose that for each image, we have v features. Let $X^t = \{x_i^t\}_{i=1}^n$ denote the feature matrix of the t-th view of training images, where $t \in \{1, ..., v\}$.

3.2. Optimal graph-based SSL

The traditional graph based semi-supervised learning usually solves the following problem:

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

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 $1 \leq (i, j) \leq n$. The variance σ will affect the performance significantly, and it i usually empirically tu ed. Also, the simil rity graph is often derived from si gle nformation 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:

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

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 $g(F, S)$ should be defined in such a way t at close labels ave h gh similarity and vic versa. In this paper, we define it as follows:

$$g(F, S) = \sum_{ij} \|f_i - f_j\|_2^2 s_{ij} \tag{4}$$

where f_i and f_j are the labels of data points x_i and x_j. Similarly, $h(X^t, S)$ can be defined as:

$$h\left(X^t, S\right) = \sum_{ij} \left\|x_i^t - x_j^t\right\|_2^2 s_{ij} \tag{5}$$

Note that for simplicity, we use distance based method to learn the similarity graph here. Other options based on the reconstruction coefficients methods can be utilized to achieve better performance, which is discussed in the next section. Instead of preserving all the pairwise distances, we consider preserving the pair distances of the K-nearest neighbors here, i.e., if x_i^t and x_j^t (or f_i *and* f_j) are not K-nearest neighbors of each other, their distance will be set to a large constant. The regularizer term $r(S, \alpha)$ is defined as:

$$r(S, \alpha) = \frac{\mu\gamma}{\beta} \|S\|_F^2 + \|\alpha\|_2^2 \tag{6}$$

We further constrain that $S \geq 0$, $S\mathbf{1} = \mathbf{1}$, $\alpha \geq 0$ and $\alpha^T\mathbf{1} = \mathbf{1}$, where $\mathbf{1} \in \mathbb{R}^{N \times 1}$ is a column vector with all ones. 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, Eq.5 and Eq.6. By combining Eq.1 with Eq.3, we can obtain the objective function for optimal-graph based SSL, as follows:

$$\begin{aligned}&\min_{S,F,\alpha} \sum_{ij} \|f_i - f_j\|_2^2 s_{ij} + \beta \|\alpha\|_2^2 \\ &+\mu \left(\sum_{tij} \left(\alpha_t \left\|x_i^t - x_j^t\right\|_2^2 s_{ij} \right) + \gamma \|S\|_F^2 \right) \\ &s.t. \begin{cases} S \geq 0, S\mathbf{1} = \mathbf{1} \\ F_l = Y_l \\ \alpha \geq 0, \alpha^T \mathbf{1} = \mathbf{1} \end{cases}\end{aligned} \tag{7}$$

3.3. Iterative optimization

We propose an iterative method to minimize the above objective function in Eq.7. 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. These steps are described below.

3.3.1 Update F

By fixing S and α, we can obtain F by optimizing Eq.7. This is equivalent to optimize the following obj ctive function:

$$\min_{F,F_l=Y_l} \sum_{ij} \|f_i - f_j\|_2^2 s_{ij} = \min_{F,F_l=Y_l} \left\|F\left(I - S\right) F^T\right\|_F^2 \tag{8}$$

where I is an identity matrix. Let $L = I - S$, and differentiate the objective function in Eq.8 with respect to F, we obtain:

$$\begin{aligned}&LF = 0 \Rightarrow \begin{bmatrix} L_{ll} & L_{lu} \\ L_{ul} & L_{uu} \end{bmatrix} \begin{bmatrix} F_l \\ F_u \end{bmatrix} = 0 \\ &\Rightarrow \begin{cases} L_{ll} F_l + L_{lu} F_u = 0 \\ L_{ul} F_l + L_{uu} F_u = 0 \end{cases}\end{aligned} \tag{9}$$

Then we can obtain:

$$F_u^* = -{L_{uu}}^{-1} L_{ul} F_l \tag{10}$$

3.3.2 Update S

By fixing F and α, we can obtain S by optimizing Eq.7. It is equivalent to optimize the following objective function:

$$\sum_{ij} \|f_i - f_j\|_2^2 s_{ij} + \mu \sum_{tij} \left(\alpha_t \left\|x_i^t - x_j^t\right\|_2^2 s_{ij} \right) \tag{11}$$
$$+\mu\gamma \|S\|_F^2$$

It can be reformulated as:

$$\begin{aligned}&\min_{S,S\geq 0,S1=1} \sum_i tr\left(\mu\gamma s_i {s_i}^T + \left(a_i + \mu b_i\right) {s_i}^T\right) \\ &\Rightarrow \min_{S,S\geq 0,S1=1} \sum_i tr\left(s_i {s_i}^T + \tfrac{a_i+\mu b_i}{\mu\gamma} {s_i}^T\right)\end{aligned} \tag{12}$$

and it is equivalent to:

$$\min_{S,S\geq 0,S1=1} \sum_i \left\| s_i + \frac{a_i + \mu b_i}{2\mu\gamma} \right\|_2^2 \tag{13}$$

where $b_i = \{b_{ij}, 1 \leq j \leq n\}$ with $b_{ij} = \sum\limits_t \alpha_t \left\|x_i^t - x_j^t\right\|_2^2$ and $a_i = \{a_{ij}, 1 \leq j \leq n\} \in R^{1\times n}$ with $a_{ij} = \|y_i - y_j\|_2^2$.

The problem in Eq.13 is simplex and we use the accelerated projected gradie t method to linearly solve this problem. The critical step of the projected gradient method is to solve the following proximal problem:

$$\min_{x\geq 0, x^T 1 = 1} \frac{1}{2} \|x - c\|_2^2 \tag{14}$$

This proximal problem can be solved using KKT approach. Then each s_i can be efficiently solved, and we can get the updated graph S.

3.3.3 Update α

By fixing F and S, we can obtain α by optimizing Eq.7. It is equivalent to optimize the following objective function:

$$\begin{aligned}&\min_{\alpha\geq 0,\alpha^T 1=1} \mu \sum_t \alpha_t \left(\sum_{ij} \left\|x_i^t - x_j^t\right\|_2^2 s_{ij} \right) + \beta \|\alpha\|_2^2 \\ &\Rightarrow \min_{\alpha\geq 0,\alpha^T 1=1} \mu d\alpha + \beta \|\alpha\|_2^2\end{aligned} \tag{15}$$

where $d = \{d_t, 1 \leq t \leq v\}$ with $d_t = \sum\limits_{ij} \left\|x_i^t - x_j^t\right\|_2^2 s_{ij}$. It can be reformulated in the form of problem in Eq. 4 and can be solved similarly to obtain α.

We update F, S and α iteratively until the objective function Eq.7 converges.

4. Extensions of OGL

In this section, we discuss several issues for graph based learning methods in real applications and provide the solutions.

4.1. Out-of-sample extension

Out-of-sample refers to learning an annotation function that is able to label new data points. This can be achieved by adding a fitting model and a regularizer to the objective function in Eq.7, e.g., $\|XW + \mathbf{1}b - F\|_F^2 + \eta \|W\|_F^2$, where $W \in \mathbb{R}^{m\times c}$, $b \in \mathbb{R}^{1\times c}$ and $\mathbf{1}$ is a vector of all ones. To obtain the optimal solution W^* and b^*, we set the derivatives of the objective function with respect to W and b equals to zero. We have:

$$b^* = \frac{1}{n}\left(\mathbf{1}^T F - \mathbf{1}^T XW\right) \tag{16}$$

$$W^{*} = \left(X^{T} L_{c} X + \eta I\right)^{-1} X^{T} L_{c} F \qquad (17)$$

where X is the concatenation of different features X^{t}, and $L_{c} = I - \mathbf{1}\mathbf{1}^{T}$. Then $\|XW + \mathbf{1}b - F\|_{F}^{2} + \eta \|W\|_{F}^{2}$ can be reformulated as:

$$tr(F^{T} B F) \qquad (18)$$

where $B = L_{c} - L_{c} X \left(X^{T} L_{c} X + \eta I\right)^{-1} X^{T} L_{c}$. Then, by adding this fitting model, F can be obtained by solving:

$$\begin{array}{l} \min\limits_{F} trF^{T}\left(I - S + \omega B\right) F + tr(F - Y)^{T} U \left(F - Y\right) \\ \Rightarrow F^{*} = (I + U - S + \omega B)^{-1} UY \end{array} \qquad (19)$$

where ω is the parameter for the fitting model. Then we can obtain the annotation function W and b. Note that other fitting models can lso be applied here, e.g., SVM, fast image tagging [2]. In [2], Chen et al. address the incomplete tagging problem by introducing a term $\tilde{B}$ to enrich the existing jointly learn the annotation function $\tilde{W}$ and tag enrich function $\tilde{B}$, as follows: $\sum_{i} \left\|\tilde{B} f_{i} - x_{i} \tilde{W}\right\|_{2}^{2}$. For simplicity, we utilize the least square regression model to tackle the out-of-sample problem. Moreover, the performance can be further improved by incorporating better fitting models.

4.2. Noisy labels

The user-provided tags may be noisy. To address this issue, instead of limiting that the predicted labels F_{l} are strictly equal to the given hard labels Y_{l}, we introduce a soft error term $\|F_{l} - Y_{l}\|_{F}^{2}$ to release this constraint. Then, by fixing S and α, F is obtained by solving:

$$\begin{array}{l} \min\limits_{F} \sum\limits_{ij} \|f_{i} - f_{j}\|_{2}^{2} s_{ij} + \mu \|F_{l} - Y_{l}\|_{F}^{2} \\ = \min\limits_{F} trF^{T}\left(I - S\right) F + tr(F - Y)^{T} U \left(F - Y\right) \end{array} \qquad (20)$$

where $U \in \mathbb{R}^{n \times n}$ is a diagonal matrix. By setting the derivative of the Eq.20 *w.r.t* F to zero, we have:

$$\begin{array}{l} F^{*} - SF^{*} + U\left(F^{*} - Y\right) = 0 \\ \Rightarrow F^{*} = (I + U - S)^{-1} UY \end{array} \qquad (21)$$

where $Y = \begin{bmatrix} Y_{l} \\ Y_{u} \end{bmatrix}$ with $Y_{u} = \{0\}^{u*c}$. Experimental results show that Eq.21 has superior performance over Eq.10.

4.3. Different graph construction models

In our work, we utilize a distance-based method to construct the similarity graph in Eq.4 and Eq.5 for simplicity. OGL can be further extended by using different graph construction models. One possible way is to adopt the reconstruction coefficients methods, which can be calculated by solving:

$$\min_{S} \sum_{i} \|x_{i} - D_{i} s_{i}\| \;\; s.t \;\; \mathbf{1}^{T} s_{i} = 1 \qquad (22)$$

where $s_{i} \in \mathbb{R}^{n \times 1}$ is the coefficient of x_{i} over D_{i} and D_{i} consists of the K-nearest neighbors of x_{i} in Euclidean space.

Recently, some studies have exploite the inherent sparsity of sparse representation to obtain a block-diagonal affinity matrix, e.g., SSC [5] and the L1-graph [3]. In [3], the L1-graph is proposed for image analysis to solve the following problem:

$$\min_{S} \|s_{i}\|_{1} \;\; s.t. \;\; \|x_{i} - X_{i} s_{i}\|_{2} < \delta \qquad (23)$$

where $s_{i} \in \mathbb{R}^{n \times 1}$ is the sparse representation of x_{i} over the dictionary X_{i} and δ is the error tolerance.

An ther recently proposed method, LRR [10, 9], aims to find the lowest-rank represen ation, r ther than the sparsest, by solving:

$$\min_{S,E} \;\; rank\left(S\right) + \lambda \|E\|_{2,1}, \;\; s.t. \;\; X = XS + E \qquad (24)$$

where $S \in \mathbb{R}^{n \times n}$ is the coefficient matrix of X over the data set itself and E is the reconstruction error. These graph construction methods have reported superiority over distance-based graph construction method. Our OGL can be further improved if these graph construction methods were adopted.

5. Experiments

We evaluate our algorithm on the task of image and video annotation. Firstly, we study the influence of the parameters in our algorithm. The , we compare our results with state-of-the-art algorithms on four standard datasets.

5.1. Experimental settings

5.1.1 Datasets

We consider four publicly available datasets that have been widely used in previous work.

IXMAS. This video dataset consists of 12 action classes (e.g., check watch, cross arms and scratch head). Each action is executed three times by 12 actors and is recorded with five cameras observing the subjects from very different perspectives.

NUS-WIDE. This image dataset contains 269,648 images downloaded from Flickr. Tagging ground-truth for 81 semantic concepts is provided for evaluation. Similar to [14], we only use the images associated with the 10 most frequent concept annotations obtaining 167,577 mages in total.

ESP Game. It contains a wide variety of images including logos, drawings, and personal photos. Following [2], we use the subset of 20,000, out of the 60,000 images publicly available.