GRAPH-BASED SEMI-SUPERVISED LEARNING WITH MULTI-LABEL

Zheng-Jun Zha [†], *Tao Mei* [‡], *Jingdong Wang*[‡], *Zengfu Wang* [†], *Xian-Sheng Hua* [‡]

[†]University of Science and Technology of China, Hefei, 230027, P. R. China
[‡]Microsoft Research Asia, Beijing, 100080, P. R. China

ABSTRACT

Conventional graph-based semi-supervised learning methods predominantly focus on single label problem. However, it is more popular in real-world applications that an example is associated with multiple labels simultaneously. In this paper, we propose a novel graph-based learning framework in the setting of semi-supervised learning with multi-label. The proposed approach is characterized by simultaneously exploiting the inherent correlations among multiple labels and the label consistency over the graph. We apply the proposed framework to video annotation and report superior performance compared to key existing approaches over the TRECVID 2006 corpus.

Index Terms— graph-based learning, multi-label, semi-supervised learning

1. INTRODUCTION

In real-world applications of machine learning, one often faces a lack of sufficient labeled data since human labeling is a labor-intensive and time-consuming process. However, in many cases, the unlabeled data are usually plentiful. Consequently, semi-supervised learning, which attempts to leverage the unlabeled data in addition to labeled data, has attracted much attention. Many different semi-supervised learning techniques have been proposed. Extensive review can be found in [1].

As a major family of semi-supervised learning, graph-based methods have recently attracted significant interest due to their effectiveness and efficiency [2] [3] [4]. Almost all the graph-based methods essentially estimate a function on the graph such that it has two properties: 1) it should be close to the given labels on the labeled examples, and 2) it should be smooth on the whole graph. This can be expressed in a regularization framework where the first term is a loss function to penalize the deviation from the given labels, and the second term is a regularizer to prefer the label smoothness. The typical graph-based methods are similar to each other, and differ slightly in the loss function and the regularizer. For example, Zhu *et al*. [3] proposed a Gaussian random field and Harmonic function (GRF) method. They defined a quadratic loss function with infinity weight to clamp the labeled examples, and formulated the regularizer based on the graph combinatorial Laplacian. In [4], Belkin *et al*. mentioned that p-Laplacian can be used as a regularizer. Zhou *et al*. [2] presented the Local and global consistency (LGC) method. They defined a quadratic loss function and used the normalized combinatorial Laplacian in the regularizer.

Fig. 1. Sample video clips associated with multiple labels.

The existing graph-based approaches mainly limit in dealing with single label problems. However, many real-world tasks are naturally posed as multi-label problems, where an example can be associated with multiple labels simultaneously. For example, in video annotation, a video clip can be annotated with multiple labels at the same time, such as 'sky," "mountain," and "water" (see Fig. 1). In text categorization, a document can be classified into multiple categories, e.g., a document can belong to "novel," "Jules Verne's writing," and "books on travelling." A direct way to tackle the typical graph-based learning under multi-label setting is to translate it into a set of independent single label problems [5]. The drawback with this approach is that it does not take into account the inherent correlations among multiple labels. However, researchers have proven the value of label correlations in various application fields [6] [7] [8].

To address the above issue, in this paper, we propose a novel graph-based semi-supervised learning framework, which can simultaneously explore the correlations among multiple labels and the label consistency over the graph. Specifically, the framework employs two types of regularizers. One is used to prefer the label smoothness on the graph, the other is adopted to address that the multi-label assignments for each example should be consistent with the inherent label correlations. We applied the proposed framework to video annotation and conducted experiments over TRECVID 2006 development set [9]. We reported the su-

This work was performed when Zheng-Jun Zha was visiting Microsoft Research Asia as a research intern.

perior performance compared to key existing graph-based learning approaches.

The rest of this paper is organized as follows. We give a short introduction to the existing graph-based semi-supervised learning framework in Section 2. Section 3 provides the detailed description of the proposed framework. Experimental results on TRECVID 2006 data set are reported in Section 4 followed by concluding remarks in Section 5.

2. GRAPH-BASED SEMI-SUPERVISE LEARING WITH SINGLE LABEL

The graph-based semi-supervised methods define a weighted graph such that the nodes correspond to labeled and unlabeled data points and the edges reflect the similarities between data points. The existing algorithms are all based on the common assumption that the labels are smooth on the graph. Then, they essentially estimate a function over the graph such that it satisfies two conditions: 1) it should be close to the given labels, and 2) it should be smooth on the whole graph. Generally, these two conditions are presented in a regularization form.

Let $\mathcal{X} = \{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N\}$ be a set of N data points in $\mathbb{R}^d$. The first L points are labeled as $\mathbf{y}_l = [y_1, y_2, \cdots, y_L]^T$ with $y_i \in \{0, 1\}$ $(1 \leqslant i \leqslant L)$, and the task is to label the remaining points $\{\mathbf{x}_{L+1}, \cdots, \mathbf{x}_N\}$. Denote the graph by $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where the node set $\mathcal{V} = \mathcal{L} \cup \mathcal{U}$ with $\mathcal{L}$ corresponding to $\{\mathbf{x}_1, \cdots, \mathbf{x}_L\}$ and $\mathcal{U}$ corresponding to $\{\mathbf{x}_{L+1}, \cdots, \mathbf{x}_N\}$. The edges $\mathcal{E}$ are weighted by the $n \times n$ affinity matrix $\mathbf{W}$ with $\mathbf{W}_{ij}$ indicating the similarity measure between $\mathbf{x}_i$ and $\mathbf{x}_j$, and $\mathbf{W}_{ii}$ is set to 0. Let $\mathbf{f} = [f_1, f_2, \cdots, f_L, f_{L+1}, \cdots, f_N]^T$ denote the predicted labels of $\mathcal{X}$.

Mathematically, the graph-based methods aim to find an optimal $\mathbf{f}^*$ essentially by minimizing the following objective function

$$E(\mathbf{f}) = E_l(\mathbf{f}) + E_s(\mathbf{f}), \tag{1}$$

where $E_l(\mathbf{f})$ is a loss function to penalize the deviation from the given labels, and $E_s(\mathbf{f})$ is a regularizer to prefer the label smoothness. For example, the Gaussian random field method [3] formulates $E_l(\mathbf{f})$ and $E_s(\mathbf{f})$ as

$$E_l(\mathbf{f}) = \infty \sum\nolimits_{i \in L} (f_i - y_i)^2 = (\mathbf{f} - \mathbf{y})^T \mathbf{\Lambda} (\mathbf{f} - \mathbf{y}),$$
$$E_s(\mathbf{f}) = \frac{1}{2} \sum\nolimits_{i,j \in L \cup U} \mathbf{W}_{ij} (f_i - f_j)^2 = \mathbf{f}^T \mathbf{\Delta} \mathbf{f},$$

where $\mathbf{\Lambda}$ is a diagonal matrix with $\mathbf{\Lambda}_{ii} = \infty, i \leqslant L$ and $\mathbf{\Lambda}_{ii} = 0, i > L$. $\mathbf{\Delta} = \mathbf{D} - \mathbf{W}$ is the combinatorial graph Laplacian, $\mathbf{D}$ is a diagonal matrix with its (i, i)-element equal to the sum of the ith row of $\mathbf{W}$. In Local and Global Consistency method [2], $E_l(\mathbf{f})$ is defined as $(\mathbf{f} - \mathbf{y})^T (\mathbf{f} - \mathbf{y})$ and $E_s(\mathbf{f})$ is formulated as $\mathbf{f}^T \mathbf{D}^{-1/2} \mathbf{\Delta} \mathbf{D}^{-1/2} \mathbf{f}$, where $\mathbf{D}^{-1/2} \mathbf{\Delta} \mathbf{D}^{-1/2}$ is the normalized combinatorial Laplacian.

As aforementioned, the existing graph-based methods mainly address the semi-supervised problem for single label scenario, and they are sub-optimal for multi-label scenario, which is more challenging but much closer to real-world applications.

3. GRAPH-BASED SEMI-SUPERVISED LEARNING WITH MULTI-LABEL

In this section, we address the semi-supervised K-label problem. Define an $N \times K$ label matrix $\mathbf{Y}$, where $\mathbf{Y}_{ik}$ is 1 if $\mathbf{x}_i$ is a labeled sample with its label k, and 0 otherwise. Define an $N \times K$ matrix $\mathbf{F} = [\mathbf{f}_1, \mathbf{f}_2, \cdots, \mathbf{f}_N]^T$, where $\mathbf{F}_{ik}$ is the confidence that $\mathbf{x}_i$ is associated with label k.

3.1. Motivation

In multi-label scenario, it is observed that the labels assigned to a sample (e.g., a video clip in video annotation task) are usually consistent with the inherent label correlations. For example, "car," and "road" are usually co-assigned to a certain video clip since they often appear simultaneously, while "explosion fire" and "waterscape" are generally not assigned to a sample at the same time since they usually do not co-occur.

Motivated by this observation, we propose a novel graph-based learning framework, such that the vector-valued function over the graph has three properties: 1) it should be close to the given labels, 2) it should be smooth on the whole graph, and 3) it should be consistent with the label correlations. The former two properties have been addressed in existing graph-based methods. The third one is novel and the key contribution of this paper.

3.2. Formulation

This unified regularization framework consists of three components: a loss function $E_l(\mathbf{F})$, and two types of regularizers $E_s(\mathbf{F})$ and $E_c(\mathbf{F})$. Specifically, $E_l(\mathbf{F})$ corresponds to the first property to penalize the deviation from the given multi-label assignments, $E_s(\mathbf{F})$ is a regularizer to address the multi-label smoothness, and $E_c(\mathbf{F})$ is a regularizer to prefer the third property. Then, the proposed framework can be formulated to minimize

$$E(\mathbf{F}) = E_l(\mathbf{F}) + E_s(\mathbf{F}) + E_c(\mathbf{F}), \tag{2}$$

where $E_l(\mathbf{F})$ and $E_s(\mathbf{F})$ can be specified in a way similar to that adopted in existing graph-based methods. Here we define them as the same as those in GRF [3], i.e.,

$$E_l(\mathbf{F}) = \mathrm{tr}((\mathbf{F} - \mathbf{Y})^T \mathbf{\Lambda} (\mathbf{F} - \mathbf{Y})),$$
$$E_s(\mathbf{F}) = \mathrm{tr}(\mathbf{F}^T \mathbf{\Delta} \mathbf{F}),$$

where $\mathrm{tr}(\mathbf{M})$ is the trace of matrix $\mathbf{M}$.

We make use of the correlations among multiple labels to define $E_c(\mathbf{F})$. To capture the label correlations, we introduce a $K \times K$ symmetric matrix $\mathbf{C}$ with $\mathbf{C}_{ij}$ represents

the correlation between label i and label j. Then, we specify $E_c(\mathbf{F}) = -\text{tr}(\mathbf{F}\mathbf{C}\mathbf{F}^T)$ to make the predicted multiple labels for each sample satisfy the correlations. Eqn. (2) is then specifically formulated as the following

$$\begin{aligned} E(\mathbf{F}) = \quad & \text{tr}((\mathbf{F}-\mathbf{Y})^T\mathbf{\Lambda}(\mathbf{F}-\mathbf{Y})) \\ & +(1-\alpha)\,\text{tr}(\mathbf{F}^T\Delta\mathbf{F}) - \alpha\,\text{tr}(\mathbf{F}\mathbf{C}\mathbf{F}^T), \quad (3) \end{aligned}$$

where α is the trade-off parameter. For the sake of simplicity, we name this algorithm instance as ML-GRF.

3.3. Solution

Differentiating $E(\mathbf{F})$ with respect to $\mathbf{F}$, we have

$$\mathbf{\Lambda}(\mathbf{F}-\mathbf{Y}) + (1-\alpha)\mathbf{\Delta}\mathbf{F} - \alpha\mathbf{F}\mathbf{C} = 0. \quad (4)$$

Let $\mathbf{F} = [\mathbf{F}_l^T\mathbf{F}_u^T]^T$, and represent the matrix $\mathbf{W}$ (and similarly $\mathbf{D}$) in the form of block matrices:

$$\mathbf{W} = \begin{bmatrix} \mathbf{W}_{ll} & \mathbf{W}_{lu} \\ \mathbf{W}_{ul} & \mathbf{W}_{uu} \end{bmatrix}. \quad (5)$$

Then Eqn. (4) can be wrote as

$$\mathbf{P}_{uu}\mathbf{F}_u + \mathbf{F}_u\mathbf{C} = \mathbf{S}, \quad (6)$$

where $\mathbf{P}_{uu} = (1-1/\alpha)(\mathbf{D}_{uu} - \mathbf{W}_{uu})$ and $\mathbf{S} = (1-1/\alpha)\mathbf{W}_{ul}\mathbf{Y}$. Eqn. (6) is essentially a Sylvester equation [10], which is widely used in control theory.

Vectorizing the unknown matrix $\mathbf{F}_u$, Eqn. (6) can be transformed to a linear equation:

$$(\mathbf{I}\otimes\mathbf{P}_{uu} + \mathbf{C}^T\otimes\mathbf{I})\text{vec}(\mathbf{F}_u) = \text{vec}(\mathbf{S}), \quad (7)$$

where $\otimes$ is the Kronecker product, $\mathbf{I}$ is the identity matrix, and $\text{vec}(\mathbf{M})$ is the vectorization of the matrix $\mathbf{M}$. We can then obtain $\mathbf{F}_u$ from $\text{vec}(\mathbf{F}_u)$, which is equal to $(\mathbf{I}\otimes\mathbf{P}_{uu} + \mathbf{C}^T\otimes\mathbf{I})^{+}\text{vec}(\mathbf{S})$.

In addition, we can also define $E_l(\mathbf{F})$ and $E_s(\mathbf{F})$ as those adopted in LGC [2] method, i.e.,

$$\begin{aligned} E_l(\mathbf{F}) &= \text{tr}((\mathbf{F}-\mathbf{Y})^T(\mathbf{F}_l-\mathbf{Y})), \\ E_s(\mathbf{F}) &= \text{tr}((\mathbf{F}-\mathbf{Y})^T\mathbf{D}^{-1/2}\mathbf{\Delta}\mathbf{D}^{-1/2}(\mathbf{F}_l-\mathbf{Y})). \end{aligned}$$

Then Eqn. (2) is specifically formulated as

$$\begin{aligned} E(\mathbf{F}) = \quad & \text{tr}((\mathbf{F}-\mathbf{Y})^T(\mathbf{F}-\mathbf{Y})) - \alpha\,\text{tr}(\mathbf{F}\mathbf{C}\mathbf{F}^T) \\ & +(1-\alpha)\text{tr}((\mathbf{F}-\mathbf{Y})^T\mathbf{D}^{-1/2}\mathbf{\Delta}\mathbf{D}^{-1/2}(\mathbf{F}-\mathbf{Y})). \end{aligned}$$

We name this approach as ML-LGC and the corresponding solution is similar to that of ML-GRF. We omit the discussion due to lack of space.

4. EXPERIMENTS

In this section, we evaluate the proposed framework on a widely used benchmark video data set and compared it against two state-of-the-art graph-based methods.

4.1. Implementation Issues

In many real-world applications, the labeled data is usually insufficient. Thus the correlation matrix $\mathbf{C}$ obtained from that limited data is unreliable. To tackle this difficulty, we propose an iterative solution. In each iteration t, the labels are predicted by solving Eqn. (6) with $\mathbf{C}^{t-1}$ estimated at last iteration $(t-1)$, and the data points labeled with high certainty are incorporated to re-estimated $\mathbf{C}^t$.

Suppose there are M data points with the predicted label matrix $\mathbf{F}_p = [\mathbf{f}_1, \mathbf{f}_2, \cdots, \mathbf{f}_M]^T$. For the label vector $\mathbf{f}_i$ associated with $\mathbf{x}_i$, we calculate the certainty as $g(\mathbf{f}_i) = \frac{1}{K}\sum_{j=1}^{K}\exp\{-f_{ij}(2\mu_j - f_{ij})\}$, where $\mu_j = \frac{1}{M}\sum_{i=1}^{M} f_{ij}$. Then, given a label matrix $\mathbf{F}_c = [\mathbf{f}_1\ \mathbf{f}_2\ \cdots\ \mathbf{f}_n]^T$ on n data points with certain labeling, $\mathbf{C}$ is calculated as $\mathbf{C}_{ij} = \exp(-||\mathbf{f}_i^{'} - \mathbf{f}_j^{'}||^2/2\sigma_c^2)$, where $\mathbf{f}_i^{'}$ is the ith column of $\mathbf{F}_c$, and $\sigma_c = \text{E}(\parallel \mathbf{f}_i^{'} - \mathbf{f}_j^{'} \parallel)$ is the average distance.

4.2. Evaluations

We conduct experiments to compare the proposed approaches (i.e., ML-GRF and ML-LGC) against other two representative graph-based methods: GRF [3] and LGC [2]. The comparison is conducted over TREVID 2006 development set [9], which contains 137 broadcast news videos. These videos are segmented into 61901 sub-shots and 39 concepts are labeled on each sub-shot according to LSCOM-Lite annotations [11]. The development set is separated into four partitions with 90 videos as training, 16 videos as validation, 16 videos as fusion, and 15 videos as test [12].

The low-level feature we use here is 225-Dimensional block-wise color moment, which are extracted over 5×5 fixed grids, each block is described by 9-Dimensional features. We use the Gaussian kernel function to calculate the weighted matrix $\mathbf{W}$. The kernel bandwidth σ and the trade-off parameter α are respectively selected via the validation process.

For performance evaluation, we use the TRECVID performance metric: Average Precision (AP) [9] to evaluate and compare the approaches on each concept. Through averaging the AP over all 39 concepts, we obtain the mean average precision (MAP), which is the overall evaluation result. Table 1 shows the MAP of our two approaches, the LGC and GRF, and Fig. 2 illustrates the AP of these four approaches. The following observation can be obtained:

- By exploiting the label correlations, the proposed methods outperform the approaches which treat the semantic labels separately and neglect their integration. In details, ML-LGC achieves around 7.2% improvements on MAP compared to LGC, while ML-GRF obtains about 6.5% improvements compared to GRF.

- ML-LGC outperforms LGC over 35 of all the 39 concepts. Some of the improvements are significant, such

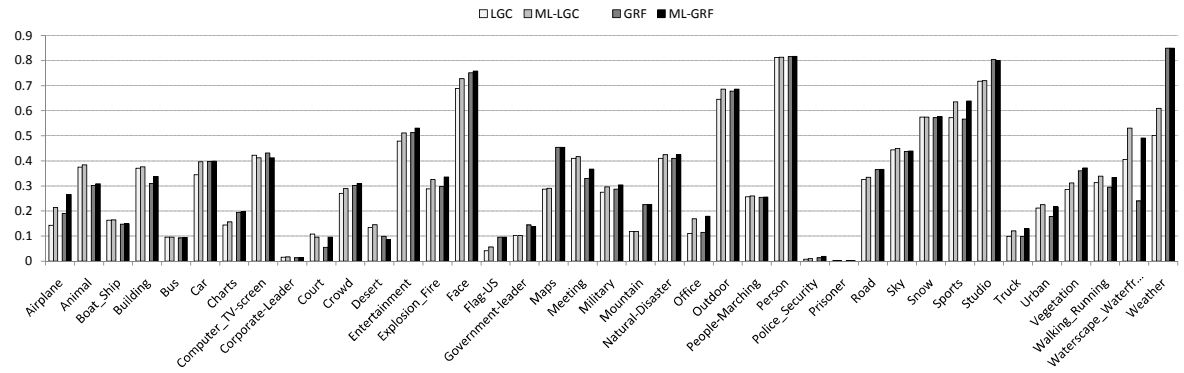

Fig. 2. The AP comparison of LGC [2], GRF [3], and the proposed approaches (i.e., ML-LGC and ML-GRF) .

Table 1. Comparison of MAP for the four approaches.

Approach	MAP	Improvement
LGC [2]	0.307	-
ML-LGC	**0.329**	+7.2%
GRF [3]	0.325	-
ML-GRF	**0.346**	+6.5%

as the 52%, 49%, and 31% improvements on "office," "airplane," and "waterscape," respectively. Compared to GRF, ML-GRF obtains improvements on 32 concepts, such as "waterscape" (105%), "office" (55%), and "airplane" (40%).

- The proposed methods degrade slightly on a few concepts. The main reason is that each of these concepts has weak interactions with other concepts. As a result, the presence/absence of these concepts cannot benefit from those of the others.

In summary, compared to the existing graph-based methods, the proposed approaches consistently outperform the performance on the diverse 39 concepts.

5. CONCLUSIONS

We have proposed a novel graph-based semi-supervised learning framework to address the multi-label problems, which simultaneously takes into account both the correlations among multiple labels and the label consistency over the graph. In addition to the label smoothness, the proposed framework also addresses that the multi-label assignment to each sample should be consistent with the inherent label correlations. Experiments on the benchmark TRECVID data set demonstrated that the novel framework is superior to key existing graph-based methods, in both overall performance and the consistency of performance on diverse concepts.

6. REFERENCES

[1] X. Zhu, "Semi-supervised learning literature survey," Tech. Rep. 1530, Computer Sciences, University of Wisconsin-Madison, 2005.

[2] D. Zhou, O. Bousquet, T.N. Lal, J. Weston, and B. Scholkopf, "Learning with local and global consistency," in *Advances in Neural Information Processing Systems (NIPS)*, Cambridge, MA, 2004, vol. 16, pp. 321–328, MIT Press.

[3] X. Zhu, Z. Ghahramani, and J. D. Lafferty, "Semi-supervised learning using gaussian fields and harmonic functions," in *Proceedings of the Twentieth International Conference on Machine Learning (ICML)*, Washington, DC, 2003, pp. 912–919.

[4] M. Belkin, I. Matveeva, and P. Niyogi, "Regularization and semi-supervised learning on large graphs," in *Annual Conference on Computational Learning Theory (COLT)*. 2004, vol. 3120 of *Lecture Notes in Computer Science*, pp. 624–638, Springer.

[5] T. Mei, X.-S. Hua, W. Lai, L. Yang, Z.-J. Zha, and *et al.*, "Msra-ustc-sjtu at trecvid 2007: High-level feature extraction and search," in *TREC Video Retrieval Evaluation Online Proceedings*, 2007.

[6] N. Ghamrawi and A. McCallum, "Collective multi-label classification," in *Proceedings of the ACM International Conference on Information and Knowledge Management (CIKM)*, New York, NY, 2005, pp. 195–200, ACM.

[7] G.-J. Qi, X.-S. Hua, Y. Rui, J. Tang, T. Mei, and H.-J. Zhang, "Correlative multi-label video annotation.," in *Proceedings of the ACM International Conference on Multimedia (MM)*. 2007, pp. 17–26, ACM.

[8] Y. Liu, R. Jin, and L. Yang, "Semi-supervised multi-label learning by constrained non-negative matrix factorization," in *Proceedings of AAAI Conference on Artificial Intelligence(AAAI)*, Boston, 2006, pp. 666–671.

[9] TRECVID, "http://www-nlpir.nist.gov/projects/trecvid/," .

[10] P. Lancaster and M. Tismenetsky, "The theory of matrices," *Mathematical Social Sciences*, vol. 13, no. 1, February 1987.

[11] M. Naphade, L. Kennedy, J. R. Kender, S. F. Chang, J. R. Smith, P. Over, and A. Hauptmann, "LSCOM-lite: A light scale concept ontology for multimedia understanding for TRECVID 2005," in *Technical Report. RC23612, IBM T.J. Watson Research Center*, 2005.

[12] A. Yanagawa, S.-F Chang, L. Kennedy, and W. Hsu, "Columbia university's baseline detectors for 374 LSCOM semantic visual concepts," in *Columbia University ADVENT Technical Report#222-2006-8*, 2007.