New Graph Structured Sparsity Model for Multi-Label Image Annotations

Xiao Cai[†], Feiping Nie[†], Weidong Cai[‡], Heng Huang[†*]
[†]University of Texas at Arlington, Arlington, Texas 76019, USA
[‡]School of Information Technologies, University of Sydney, NSW 2006, Australia
xiao.cai@mavs.uta.edu, feipingnie@gmail.com, tom.cai@sydney.edu.au, heng@uta.edu

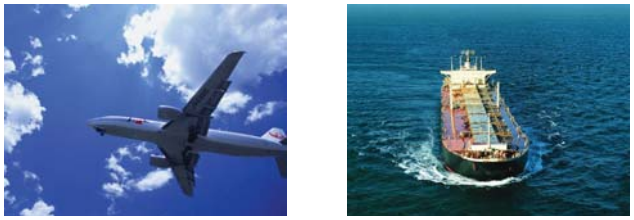
Abstract

In multi-label image annotations, because each image is associated to multiple categories, the semantic terms (label classes) are not mutually exclusive. Previous research showed that such label correlations can largely boost the annotation accuracy. However, all existing methods only directly apply the label correlation matrix to enhance the label inference and assignment without further learning the structural information among classes. In this paper, we model the label correlations using the relational graph, and propose a novel graph structured sparse learning model to incorporate the topological constraints of relation graph in multi-label classifications. As a result, our new method will capture and utilize the hidden class structures in relational graph to improve the annotation results. In proposed objective, a large number of structured sparsity-inducing norms are utilized, thus the optimization becomes difficult. To solve this problem, we derive an efficient optimization algorithm with proved convergence. We perform extensive experiments on six multi-label image annotation benchmark data sets. In all empirical results, our new method shows better annotation results than the state-of-the-art approaches.

1. Introduction

Due to the Internet and visual data sharing websites, the availability of visual data has been dramatically increased in the last decade, which has provided billions of images and videos to computer vision researchers. Annotating these images is crucial for the computer vision system development and validation. However, the task of manually annotating large-scale visual data sets takes a lot of time and effort, and is almost impossible. Thus, how to automatically and accurately annotate the visual data has become one of the central problems in computer vision research.

*Corresponding author. This project was partially supported by U.S. NSF IIS-1117965, IIS-1302675, IIS- 1344152, and ARC grant.



(a) "sky", "plane" (b) "ocean", "ship"

Figure 1. An example of label correlations for class membership inference. Both images have large regions with blue color, and it is hard to decide to annotate them with "sky" or "ocean". However, if we know Fig. 1(a) is annotated to "plane", we have high confidence to annotate it with "sky", rather than "ocean".

Different to traditional single-label multi-class image classifications, in image annotation, each image or video clip is often associated with more than one semantic label, which poses so-called multi-label multi-class classification problem. For example, image in Fig. 1(a) is annotated with semantic words "sky" and "plane", and image in Fig. 1(b) is associated to semantic words "ocean", "ship". The multi-label multi-class classifications have many applications, such as document classification, protein function prediction, and music annotation.

An important difference between single-label classification and multi-label classification is that, the annotation classes in single-label classification are mutually exclusive, but the annotation terms in multi-label classification are correlated to each other. Thus, in multi-label classification, researchers can utilize such annotation label correlations to infer the class memberships from one to another. For example, in Fig. 1, the semantic words "ocean" and "sky" are both strongly related to the blue color, therefore it is difficult to individually decide these two labels based on the color features. However, from the training visual data, we can learn the high correlations between "sky" and "plane", and between "ocean" and "ship". Therefore, if an image is annotated with "plane", as in Fig. 1(a), we are highly confident to annotate the region of blue in the same image as "sky", rather than "ocean". Similarly, for the image anno-

tated with "ship", we will annotate the region of blue as "ocean" as in Fig. 1(b). Many previous multi-label image annotation methods explore such label correlations to improve the classification accuracy [3, 17, 11, 12, 13].

However, all previous methods enhance the multi-label classifications by directly multiplying a label correlation matrix $C \in \Re^{c\times c}$ (which can be calculated by the normalized cosine similarity between classes and c is the number of classes) on the label matrix or coefficient matrix to improve the label propagation or label assignment. None of them explores the structures of classes under the label correlations. Beyond straightforwardly applying the label correlation matrix, in this work, we propose to utilize the class relational graph to model the underlying structures existing in multi-label classes.

The label correlations indeed can be modeled as a class relational graph. For example, using PASCAL 2006 data set, we can model the correlations among annotation term as a relational graph $G = \{V, E\}$ in Fig. 2, where nodes in V are the annotation classes and weights of edges in E are the correlation values between classes (nodes). Some classes, such as "Cat", "Cow", "Sheep", have very small correlations with the rest classes shown in the left panel of Fig. 2 (the values are smaller than 0.02). For demonstration purpose, we threshold them as zero, hence there is no edges connecting these nodes in right panel. If two nodes (classes) have large correlations, the edge weight between them will be large. Such a relational graph model can capture the underlying structural interrelations between classes. How to utilize this relational graph with discovering the hidden classes structures to enhance multi-label classification is computationally challenging.

In this paper, we will propose the novel structured sparsity-inducing norm regularization to incorporate the relational graph information into multi-label classification model. Different to previous methods, which directly use the label correlation values to enhance the classification results, our new method will impose the correlated classes to share the common space, such that the input data relevant to both classes will learn jointly. Our new class relational graph regularization will include a large number of non-smooth structured sparsity-inducing norms, such that the objective function optimization becomes difficult. We will introduce new optimization algorithms to solve the proposed non-smooth convex objective with convergence proof. We perform our new method on six multi-label classification benchmark data sets and compare the results with eight state-of-the-art multi-label classification methods.

2. Multi-Label Classification Using Graph Structured Sparse Learning Model

The existing multi-label learning models cannot incorporate the semantic terms relational graph to enhance the annotation results. To study the feature or class structural relations, many structured sparse learning methods have been proposed in recent research and shown promising results [18, 6, 4, 8, 1, 14, 15, 7]. However, these approaches also cannot incorporate the label relational graph into the classification models.

To address this challenging problem, we propose new graph structured sparsity-inducing norms, which learn the correlated classes in a common space under the relational graph structure.

Given training data of n data points $X = [\mathbf{x}_1, \cdots, \mathbf{x}_n] \in \Re^{d\times n}$, the class indicator matrix of these data points is $Y = [\mathbf{y}_1, \cdots, \mathbf{y}_n]^T \in \Re^{n\times c}$ for c classes, the structured sparsity-inducing norm based classification model is to learn W and the bias $\mathbf{b} \in \Re^{c\times 1}$ by solving:

$$\min_{W,\mathbf{b}} \mathcal{L}(X, W, \mathbf{b}; Y) + \gamma\Omega(W), \tag{1}$$

where $\gamma \geq 0$ is a regularization parameter, and $\mathcal{L}(\cdot)$ is a loss function (*e.g.* least square loss, hinge loss). The regularization term $\Omega(W)$ is the structured sparsity-inducing norm, which usually uses the mixed norms to capture the features and classes structural relations for enhancing the classification tasks.

In multi-label annotations, we have the label (semantic terms) relational graph $G = \{V, E\}$ (*e.g.* the class relational graph constructed in Fig. 2, in which the edge between nodes V_i and V_j is denoted as E_{ij}). If we correctly incorporate such label relational graph into multi-label classification model, the performance can definitely be boosted. Thus, the structured sparsity-inducing norm $\Omega(W)$ is expected to model the label relational graph. Meanwhile, considering the computational efficiency and global optimization, we also hope $\Omega(W)$ to be a convex norm. However, it is challenging to model such graph structured sparsity by the convex norm.

We propose a new graph structured sparsity model to capture the graph structures using the structured sparsity-inducing norms. Our new graph structured sparse multi-label classification model is to solve:

$$\min_{W,\mathbf{b}} \mathcal{L}(X, W, \mathbf{b}; Y) + \gamma \sum_{E_{ij}\in E} \|[\mathbf{w}_i, \mathbf{w}_j]\|_{2,1}, \tag{2}$$

where the $\ell_{2,1}$-norm of the matrix W are defined as $||W||_{2,1} = \sum_{i=1}^{d} ||\mathbf{w}^i||_2$ (also denoted as $\ell_{1,2}$-norm by some researchers). Given a matrix $W = [w_{ij}]$, its ith row and jth column are denoted as $\mathbf{w}^i$ and $\mathbf{w}_j$, respectively. If two classes are correlated, the mixed-norm $\ell_{2,1}$-norm regularization finds inputs relevant to both outputs jointly. Our regularization terms go through all edges in E to include all topological constraints by the structured sparsity-inducing norms. Thus, the learned classification coefficients capture

Figure 2. Relational graph model of multi-label classes. Left: the correlation matrix of all annotation terms. Right: constructed relational graph, where nodes are labels and weights of edges are correlation values between classes.

these graph-based class correlations. Meanwhile, our regularization terms are also convex norms which guarantee the globally optimal results.

Because the weight a_{ij} of the edge connecting nodes V_i and V_j represents the correlation level of these two classes, we also use the weights values to scale the regularization terms. As a result, the highly correlated classes will get large weight in the joint sparsity regularization. Meanwhile, in this work, we use the least square loss (which is faster than hinge and logistic loss functions, and is suitable for large-scale multi-label classifications) and solve the following objective:

$$\min_{W,\mathbf{b}} \left\|X^T W + \mathbf{1}\mathbf{b}^T - Y\right\|_F^2 + \gamma \sum_{i=1}^{c}\sum_{j=1}^{c} a_{ij} \left\|[\mathbf{w}_i, \mathbf{w}_j]\right\|_{2,1} \tag{3}$$

where $\mathbf{1} \in \Re^{n\times 1}$ is the vector with all entries being 1. If there is no edge between nodes V_i and V_j, $a_{ij} = 0$. Taking the derivative w.r.t. $\mathbf{b}$ and setting to zero, we have $\mathbf{b} = \frac{1}{n}Y^T\mathbf{1} - \frac{1}{n}W^T X\mathbf{1}$. We substitute $\mathbf{b}$ into Eq. (3), the problem (3) becomes

$$\min_{W} \left\|HX^T W - HY\right\|_F^2 + \gamma \sum_{i=1}^{c}\sum_{j=1}^{c} a_{ij} \left\|[w_i, w_j]\right\|_{2,1}, \tag{4}$$

where $H = I - \frac{1}{n}\mathbf{1}\mathbf{1}^T$ is the centering matrix.

The objective function in (4) involves many $\ell_{2,1}$-norms as regularization terms, thus the general methods are difficult to apply here. In this paper, we propose an algorithm to solve a general $\ell_{2,1}$-norm minimization problem, based on which we will further derive the algorithm to solve the main objective in (4).

3. General Optimization Framework

Consider a general $\ell_{2,1}$-norm minimization problem as follows:

$$\min_{X\in\mathcal{C}} f(X) + \sum_i \gamma_i \left\|G_i(X)\right\|_{2,1}, \tag{5}$$

where $f(X)$ is an arbitrary function, $G_i(X)$ for each i is an arbitrary matrix-input matrix-output functions, $X \in \mathcal{C}$ is an arbitrary constraint, and assume that the objective has a lower bound. We can see the objective function in (4) is a special case of the problem in (5).

3.1. Iterative Reweighted Algorithm

Because any regularization term $\|G_i(X)\|_{2,1}$ can be written as the reweighted formulation $Tr(G_i^T(X)D^i G_i(X))$, where D^i is a diagonal matrix with the k-th diagonal element as $\frac{1}{2\|[G_i(X)]^k\|_2}$. Thus, we propose an iterative reweighted algorithm to solve problem in (5). The algorithm is described in Algorithm 1. In the following, we will prove that this algorithm will converge and converge to a local or global solution to the problem in (5), when the problem in (5) is non-convex or convex.

Initialize $X \in \mathcal{C}$;
while *not converge* **do**
1. For each i, calculate the diagonal matrix D^i, where the k-th diagonal element is: $\frac{1}{2\|[G_i(X)]^k\|_2}$;
2. Update X by solving: $\min_{X\in\mathcal{C}} f(X) + \sum_i \gamma_i Tr(G_i^T(X)D^i G_i(X))$;

end
Output: X.

Algorithm 1: Algorithm to solve a general $\ell_{2,1}$-norm minimization problem (5).

3.2. Algorithm Convergence Analysis

To prove the convergence of our Algorithm 1, first we introduce the following lemma:

Lemma 1 *Suppose D is a diagonal matrix, where the k-th diagonal element is $\frac{1}{2\|a^k\|_2}$. We have*

$$\|\tilde{A}\|_{2,1} - Tr(\tilde{A}^T D\tilde{A}) \le \|A\|_{2,1} - Tr(A^T DA)\,. \tag{6}$$

Proof: We start the proof from a simple inequality as $-x^2 \leq 0$.

$$\begin{aligned}
& -(\left\|\tilde{\mathbf{a}}^k\right\|_2 - \left\|\mathbf{a}^k\right\|_2)^2 \leq 0 \\
\Rightarrow \quad & 2\left\|\tilde{\mathbf{a}}^k\right\|_2 \left\|\mathbf{a}^k\right\|_2 - \left\|\tilde{\mathbf{a}}^k\right\|_2^2 \leq \left\|\mathbf{a}^k\right\|_2^2 \\
\Rightarrow \quad & \left\|\tilde{\mathbf{a}}^k\right\|_2 - \frac{\left\|\tilde{\mathbf{a}}^k\right\|_2^2}{2\left\|\mathbf{a}^k\right\|_2} \leq \left\|\mathbf{a}^k\right\|_2 - \frac{\left\|\mathbf{a}^k\right\|_2^2}{2\left\|\mathbf{a}^k\right\|_2} \\
\Rightarrow \quad & \sum_k \left\|\tilde{\mathbf{a}}^k\right\|_2 - \sum_k \frac{\left\|\tilde{\mathbf{a}}^k\right\|_2^2}{2\left\|\mathbf{a}^k\right\|_2} \leq \sum_k \left\|\mathbf{a}^k\right\|_2 - \sum_k \frac{\left\|\mathbf{a}^k\right\|_2^2}{2\left\|\mathbf{a}^k\right\|_2} \\
\Rightarrow \quad & \|\tilde{A}\|_{2,1} - Tr(\tilde{A}^T D \tilde{A}) \leq \|A\|_{2,1} - Tr(A^T D A)
\end{aligned}$$

Thus the inequality in the lemma holds.

□

The convergence of the Algorithm 1 is demonstrated in the following theorem:

Theorem 1 *The Algorithm 1 monotonically decreases the value of the objective function (5) in each iteration till the algorithm converges.*

Proof: In the Step 2 of Algorithm 1, we denote the updated X as $\tilde{X}$. We have

$$\begin{aligned}
& f(\tilde{X}) + \sum_i \gamma_i Tr(G_i^T(\tilde{X}) D^i G_i(\tilde{X})) \\
\leq \quad & f(X) + \sum_i \gamma_i Tr(G_i^T(X) D^i G_i(X)) \,. \qquad (7)
\end{aligned}$$

According to Lemma 1 we have

$$\begin{aligned}
& \sum_i \gamma_i \|G_i(\tilde{X})\|_{2,1} - \sum_i \gamma_i Tr(G_i^T(\tilde{X}) D^i G_i(\tilde{X})) \\
\leq \quad & \sum_i \gamma_i \|G_i(X)\|_{2,1} - \sum_i \gamma_i Tr(G_i^T(X) D^i G_i(X)) \,. \qquad (8)
\end{aligned}$$

Summing the inequalities (7) and (8) on both sides, we arrive at

$$\begin{aligned}
& f(\tilde{X}) + \sum_i \gamma_i \left\|G_i(\tilde{X})\right\|_{2,1} \\
\leq \quad & f(X) + \sum_i \gamma_i \|G_i(X)\|_{2,1} \,. \qquad (9)
\end{aligned}$$

Thus the Algorithm 1 monotonically decreases the value of objective function in (5) or remains the objective function value unchanged in each iteration t. Because the objective function in (5) has a lower bound, the Algorithm 1 will converge. When the algorithm has not converged, the Algorithm 1 will monotonically decrease the value of objective function in (5).

□

The following theorem guarantees that the Algorithm 1 will converge to a local or global solution to the problem (5).

Theorem 2 *The Algorithm 1 will converge to a local optimal solution of the objective in (5), and will converge to a global solution if the objective in (5) is a convex function.*

Proof: The Lagrangian function of the problem (5) is:

$$\mathcal{L}(X, \lambda) = f(X) + \sum_i \gamma_i \|G_i(X)\|_{2,1} - h(X, \Lambda) \,, \quad (10)$$

where $h(X, \Lambda)$ is the Lagrangian term to encode the constraint $X \in \mathcal{C}$ in problem (5).

Taking the derivative of $\mathcal{L}(X, \lambda)$ w.r.t X, and setting the derivative to zero, we have:

$$\begin{aligned}
\frac{\partial \mathcal{L}(X, \lambda)}{\partial X} &= f'(X) + \sum_i 2\gamma_i D^i G_i(X) - \frac{\partial h(X, \Lambda)}{\partial X} \\
&= 0 \,, \qquad (11)
\end{aligned}$$

where D is a diagonal matrix, and the k-th diagonal element is $\frac{1}{2\|[G_i(X)]^k\|_2}$.

Suppose the Algorithm 1 converges to a solution X^*, from Step 2 in Algorithm 1, we have:

$$X^* = \arg\min_{X \in \mathcal{C}} f(X) + \sum_i \gamma_i Tr(G_i^T(X)(D^*)^i G_i(X)), \qquad (12)$$

where D is a diagonal matrix with the k-th diagonal element as $\frac{1}{2\|[G_i(X^*)]^k\|_2}$. According to the KKT condition of the problem in Eq. (12), we know that

$$f'(X^*) + \sum_i 2\gamma_i (D^*)^i G_i(X^*) - \frac{\partial h(X^*, \Lambda)}{\partial X^*} = 0 \,. \quad (13)$$

Thus, the solution X^* satisfies Eq. (11), which is the KKT condition of the objective in (5). Therefore, the converged solution X^* is a local solution of the objective in (5). Moreover, if the objective in (5) is a convex function, then the converged solution X^* is a global solution of the objective in (5)

□

In the next section, we will derive the algorithm to solve the objective in (4) based on Algorithm 1.

4. Algorithm to Solve Objective in (4)

According to Algorithm 1, the key step to solve the objective in (4) is to solve the following problem:

$$\begin{aligned}
& \min_W \left\|HX^T W - HY\right\|_F^2 + \\
& \gamma \sum_{i=1}^c \sum_{j=1}^c a_{ij} Tr([\mathbf{w}_i, \mathbf{w}_j]^T D^{ij} [\mathbf{w}_i, \mathbf{w}_j]),
\end{aligned} \qquad (14)$$

where D^{ij} is a diagonal matrix with the k-th diagonal element as $\frac{1}{2\|[\mathbf{w}_i,\mathbf{w}_j]^k\|_2}$.

We simplify the second term in Eq. (14) as following

$$\begin{aligned}
&\sum_{i=1}^{c}\sum_{j=1}^{c} a_{ij} Tr([\mathbf{w}_i,\mathbf{w}_j]^T D^{ij}[\mathbf{w}_i,\mathbf{w}_j]) \\
= &\sum_{i=1}^{c}\sum_{j=1}^{c} (a_{ij}\mathbf{w}_i^T D^{ij}\mathbf{w}_i + a_{ij}\mathbf{w}_j^T D^{ij}\mathbf{w}_j) \\
= &\sum_{i=1}^{c}\mathbf{w}_i^T(\sum_{j=1}^{c} a_{ij}D^{ij})\mathbf{w}_i + \sum_{j=1}^{c}\mathbf{w}_j^T(\sum_{i=1}^{c} a_{ij}D^{ij})\mathbf{w}_j \\
= &\sum_{i=1}^{c}\mathbf{w}_i^T(\sum_{j=1}^{c} a_{ij}D^{ij})\mathbf{w}_i + \sum_{i=1}^{c}\mathbf{w}_i^T(\sum_{j=1}^{c} a_{ji}D^{ji})\mathbf{w}_i
\end{aligned}$$

Because $a_{ij} = a_{ji}$ and $D^{ij} = D^{ji}$, the above equation can be written as:

$$\sum_{i=1}^{c}\mathbf{w}_i^T(2\sum_{j=1}^{c} a_{ij}D^{ij})\mathbf{w}_i\,.$$

Let's denote $M^i = 2\sum\limits_j a_{ij}D^{ij}$, then the problem (14) can be simplified as:

$$\min_W \sum_{i=1}^{c} \left\|HX^T\mathbf{w}_i - H\mathbf{y}_i\right\|_2^2 + \gamma\sum_{i=1}^{c}\mathbf{w}_i^T M^i\mathbf{w}_i\,. \quad (15)$$

We can see that the problem (15) is unrelated between different $\mathbf{w}_i$, and thus can be decoupled to solve the following problem for each $\mathbf{w}_i$:

$$\min_{\mathbf{w}_i} \left\|HX^T\mathbf{w}_i - H\mathbf{y}_i\right\|_2^2 + \gamma\mathbf{w}_i^T M^i\mathbf{w}_i\,. \quad (16)$$

Taking the derivative of Eq. (16) w.r.t. $\mathbf{w}_i$ and setting to zero, we have

$$(XHX^T + \gamma M^i)\mathbf{w}_i - XH\mathbf{y}_i = 0\,. \quad (17)$$

Therefore, we get the optimal solution of the problem (16) as:

$$\mathbf{w}_i = (XHX^T + \gamma M^i)^{-1}XH\mathbf{y}_i\,. \quad (18)$$

Based on the above derivation, the detailed algorithm to solve the objective in (4) is summarized in Algorithm 2.

Because the objective in (4) is a convex problem, according to Theorem 2, we can obtain the global solution with Algorithm 2.

Input: X, A.
Output: $W \in \Re^{d\times c}$.
Initialize $W \in \Re^{d\times c}$;
while *not converge* **do**
1. For each i and j, calculate the diagonal matrix D^{ij}, where the k-th diagonal element is $\frac{1}{2\|[\mathbf{w}_i,\mathbf{w}_j]^k\|_2}$;
2. For each i, calculate the diagonal matrix M^i by: $M^i = 2\sum\limits_j a_{ij}D^{ij}$;
3. For each i, update $\mathbf{w}_i$ by: $\mathbf{w}_i = (XHX^T + \gamma M^i)^{-1}XH\mathbf{y}_i$;
end

Algorithm 2: Algorithm to solve the problem (4).

5. Experimental Results

5.1. Experiment Data

In this section, we will briefly introduce the multi-label image data sets that we used to evaluation the proposed graph structured sparse multi-label learning model.

Barcelona image dataset is composed of urban scenes from Barcelona. And consists of 139 urban scene images in "jpeg" format with minimum resolution of 1600 x 1200. It has 4 overlapping labels: "Buildings", "Flora", "People" and "Sky". Each image is represented by a feature vector of 816 dimensions using the concatenation of LBP [9], GIST [10] and CMT [16].

Natural scene data set [2][1]contains 2407 images represented by a 294-dimensional vector, which are labeled with 6 semantic concepts (labels).

TRECVID 2005 data set[2] contains 61901 sub-shots labeled with 39 concepts. We randomly sample the data such that each concept (label) has at least 100 video key frames. Therefore, we have 3721 images in total and we extract LBP [9], GIST [10] and CMT [16] features from each image. After concatenating the above three visual features, each image is represented by a 816 dimension vector.

PASCAL VOC 2006[3] is a data set for visual object recognition challenge held in 2006. It has 5304 images with 10 classes, i.e. bicycle, bus, car, motorbike, cat, cow, dog, horse, sheep and person. We download the 960 dimension GIST feature image descriptor[4] extracted from all the images. Note that multiple objects from multiple classes may be present in the same image. Therefore, it is a multi-label classification data set.

PASCAL VOC 2007[5] is an extension visual object recognition challenge data set based on PASCAL VOC 2006. It has 9963 images with 4 group annotations and each group can be further divided into the following classes. *Person*: person; *Animal*: bird, cat, cow, dog, horse, sheep; *Vehicle*: aeroplane, bicycle, boat, bus, car, motorbike, train; *Indoor*: bottle, chair, dining table, potted plant, sofa,

[1] http://mulan.sourceforge.net/datasets.html
[2] http://www-nlpir.nist.gov/projects/trecvid/
[3] http://pascallin.ecs.soton.ac.uk/challenges/VOC/voc2006/
[4] https://sites.google.com/site/christophlampert/data
[5] http://pascallin.ecs.soton.ac.uk/challenges/VOC/voc2007/

Table 1. Data Sets Summary

Data Sets	Samples (n)	Features (d)	Labels (c)
BARCELONA	139	816	4
SCENE	2407	294	6
TRECVID05	3721	816	39
PASCAL06	5304	960	10
PASCAL07	9963	512	20
MIRFLICKR08	25000	512	38

tv/monitor. We download the 512 dimension GIST feature as the image descriptor[6] extracted from all the images.

MIR FLICKR 2008[7] is a public image data set used for ACM sponsored image retrieval evaluation. It has 25000 images with 38 classes downloaded from the social photography site Flickr through its public API. After removing the most common annotations, i.e. colors, seasons and place names, the average number of annotation per image is 8.94. In the collection there are 1386 annotations which occur in at least 20 images. We download the 512 dimension GIST image descriptor[8] extracted from all the images.

We summarize the data as the following Table. 1

5.2. Experiment Settings

In our experment, we used the following way to build the graph structure for the annotations. Without losing of generality, we assume the first $l < n$ data are already labeled and each training data $\mathbf{x}_i$ is labeled with a number of annotations $\mathbf{y}_i = \{y_1, \cdots, y_C\}$ represented by $\mathbf{y}_i \in \{0,1\}^C$, such that $\mathbf{y}_i(k) = 1$ if $\mathbf{x}_i$ is annotated with the k-th annotation term and 0 otherwise, $\forall i = 1, 2, ...l$.

Different from conventional single-label classification learning in which classes are mutual exclusive, the annotations are interrelated with one another in multi-label problem. We utilize the following cosine similarity to calculate the annotation affinity matrix

$$A(i,j) = \cos(\mathbf{y}_i, \mathbf{y}_j) = \frac{< \mathbf{y}_i, \mathbf{y}_j >}{(||\mathbf{y}_i|| \times ||\mathbf{y}_j||)} \qquad (19)$$

where $\mathbf{y}_i$ and $\mathbf{y}_j$ are the i-th and j-th column of the indicator matrix of the labeled data $Y \in \mathbb{R}^{l \times c}$ respectively. Thus, a graph $G = (V, E)$ is induced, where $V = A$ and $E \subseteq V \times V$. What is more, in order to remove the "noisy" correlation induced by the outlier data or inevitable inaccurate annotation information of the training data, we set up a filter to set those entries of A in Eq. (19) with values less than 10% of its maximum value to 0. We will used the above calculated annotation graph structure as input for both our method and the comparison approaches.

[6]http://lear.inrialpes.fr/people/guillaumin/data.php
[7]http://press.liacs.nl/mirflickr/
[8]http://lear.inrialpes.fr/people/guillaumin/data.php

In all experiments, we use 5-fold cross validation. Specifically, we split the data evenly into 5 folds and take turns choosing 4 folds for training and using the remaining 1 fold for testing. In each training step, we further divide the training data into 5 parts and pick up 4 parts for testing and chose the remaining 1 part as the validation to tune the best regularization parameter γ from $\log_2\gamma \in \{-20, -19, -18, ..., 19, 20\}$. We repeat the above procedure 5 times and report the average classification results.

Moreover, we compare our proposed method with the following state-of-art multi-label classification methods:

K Nearest Neighbor (KNN), where we set K as 1 (1NN) for its simple and intuitive interpretation, that is, we predict the annotations of the testing data as the ones of its nearest neighbor.

Support Vector Machine (SVM). We consider the annotations as independent ones and use one V.S. the rest strategy to predict the annotations one by one, where we chose the linear kernel and set C as 1.

Besides the above two fundamental methods, in our experiment we compare the following mutli-label dimension reduction methods as well:

Multi-Label LDA (MLDA) [11] is extension of classical Linear Discriminate Analysis for solving the multi-label classification problem. With the new defined with-class and between class scatter matrix, it can take advantage of the prior knowledge, that is, the class-wise correlation.

Multi-Label Informed Latent Semantic Indexing (MLSI) [17] is an approach to extend unsupervised latent semantic indexing (LSI) to utilize the provided supervision information.

Multi-Label Dimension Reduction via Dependent Maximization (MDDM) [19] is a method to identify a lower-dimensional subspace by maximizing the dependence between the original data and the associated annotations.

Multi-Labe Least Square (MLLS) [3] is a method to extract the common subspace shared by multiple annotations. However, the way that MLLS takes advantage of the annotation information is different with our proposed method. In [3], if we denote the data matrix as $X \in \mathbb{R}^{d \times n}$ and annotation indicator matrix as $Y \in \mathbb{R}^{n \times c}$, then MLLS explores the linear annotation information by calculating XYY^TX only without the graph information.

What is more, with the development of feature selection methods, more and more filter methods or their variations can be used to reduce the dimension of feature and further boost the multi-label classification performance. For sake of completeness, we compare them in our experiment as well.

In [5], Kong et al proposed **Multi-Label ReliefF (MRF)** and **Multi-Label F-Statistic (MF)** to extend the traditional reliefF and F-statistic tackling feature selection problem for multi-label data. In addition, they used 1NN as the classi-

Table 2. Classification performance comparison by 5-fold cross validations on the six multi-label image data sets.

Data	Metrics	1NN	MLLS	MDMI	MLSI	SVM	MLDA	MRF	MF	LSG21
BARCELONA	Macro Pre	0.7553	0.7220	0.7147	0.7518	0.7246	0.7091	0.7546	0.7669	**0.8109**
	Macro F1	**0.7596**	0.7235	0.7243	0.7589	0.7567	0.7198	0.6994	0.7222	0.7104
	Micro Pre	0.7296	0.6729	0.6607	0.7101	0.6366	0.6667	0.7498	0.7712	**0.7741**
	Micro F1	**0.7267**	0.6702	0.6672	0.7123	0.6246	0.6721	0.6920	0.7216	0.6690
SCENE	Macro Pre	0.6874	0.4722	0.5898	0.5402	0.6721	0.5936	0.6874	0.6920	**0.6931**
	Micro Pre	0.7053	0.4731	0.6058	0.5559	0.6854	0.6083	0.7054	0.6874	**0.7055**
	Macro F1	0.6825	0.4722	0.5908	0.5409	0.6726	0.5944	0.6825	**0.6844**	0.6748
	Micro F1	0.6864	0.4719	0.6039	0.5523	0.6485	0.6069	0.6863	**0.6896**	0.6821
TRECVID05	Macro Pre	0.5271	0.5152	0.5150	0.5107	0.4582	0.5162	0.4687	0.4688	**0.5617**
	Macro F1	0.5436	0.5311	0.5274	0.5254	0.5501	0.5300	0.4745	0.4745	**0.5664**
	Micro Pre	0.4376	0.4108	0.4110	0.4082	0.3872	0.4115	0.3414	0.3414	**0.4425**
	Micro F1	**0.4356**	0.4147	0.4098	0.4102	0.3933	0.4125	0.4145	0.3422	0.4213
PASCAL06	Macro Pre	0.4613	0.3907	0.4099	0.3987	0.3874	0.4122	0.4468	0.4616	**0.5239**
	Macro F1	0.4681	0.3920	0.4122	0.4016	0.4052	0.4133	0.4541	0.4681	**0.5081**
	Micro Pre	0.4485	0.3727	0.3875	0.3715	0.4131	0.3901	0.4306	0.4447	**0.5226**
	Micro F1	0.4460	0.3724	0.3906	0.3741	0.4055	0.3914	0.4313	0.4657	**0.4957**
PASCAL07	Macro Pre	0.3261	0.3131	0.3096	0.3028	0.3477	0.3135	0.3137	0.3195	**0.3799**
	Macro F1	0.3411	0.3221	0.3187	0.3128	0.3029	0.3229	0.3258	0.3324	**0.4065**
	Micro Pre	0.2346	0.2149	0.2117	0.2082	0.2078	0.2189	0.2202	0.2218	**0.3092**
	Micro F1	0.2320	0.2149	0.2135	0.2110	0.2111	0.2202	0.2139	0.2209	**0.3069**
MIRFLICKR08	Macro Pre	0.3489	0.3417	0.3484	0.3443	0.3574	0.3522	0.3388	0.3452	**0.3814**
	Macro F1	0.3500	0.3536	0.3579	0.3520	0.3574	0.3619	0.4281	0.4368	**0.4672**
	Micro Pre	0.2260	0.2291	0.2308	0.2296	0.2371	0.2243	0.2381	0.2393	**0.2646**
	Micro F1	0.2234	0.2200	0.2247	0.2238	0.2281	0.2287	0.2760	0.2762	**0.3172**

fier to evaluate the multi-label classification performance on 10% to 70% selected features and reported the best multi-label classification result based on a certain number of selected features.

5.3. Multi-Label Classification Results

Two standard multi-label classification performance metrics precision and F1 score are used to evaluate image annotation performances. In our experiment, we report both macro and micro results in Table. 2. As can be observed from the table, first of all, correlations between annotations can indeed boost the classification performance compared with the methods that consider annotation classification independently, like SVM. Moreover, given the same graph structure, our proposed graph structured sparse multi-label learning method can consistently beat those dimension reduction methods as well as feature selection methods invented for multi-label classification on most data sets. For Barcelona data set, because there are only 4 class annotations, the recall of our method is lower than that of KNN method. Therefore, although the precision of our method is higher, we get a less macro and micro F1 score.

5.4. Enhanced Coefficient Matrix W^* by Graph Structured Sparse Learning Model

In Fig. 3, we plot the flowchart of the proposed method for demonstration purpose. Given an image having "Chairs", "Dinning table", "Person" inside, the annotation affinity matrix shows the correlation values between these semantic terms. Because semantic terms "Chairs" and "Dinning table" often appear together and have large correlations, the weight of edge connecting them in the label relational graph G is large. Thus, their regularization term has large contribution in training process (in right-bottom panel), such that the learned coefficient matrix W^* showing these correlations.

From the middle-bottom panel in Fig. 3, we can see that the learned W^* is sparse, shrunk by the $\ell_{2,1}$-norm with the help of pairwise annotation correlation information, which is shown in the top panel. We mark the coefficient visualization results of "Chairs", "Dinning table" by the blue circle. Obviously two semantic terms show similar weight coefficient structures, *i.e.* these two classes share similar visual features. In testing phase, the input feature vector multiplies W^* to predict labels. The existing methods didn't consider the shared structure between correlated semantic terms, hence they predict "Dinning table", but miss "Chairs" in the prediction. Our graph structured sparse multi-label learning model can correctly predict both labels due to the shared similar weight structures in W^*.

6. Conclusion

In this paper, we model the label correlations using the relational graph, and propose a novel graph structured sparse learning model to incorporate the topological con-

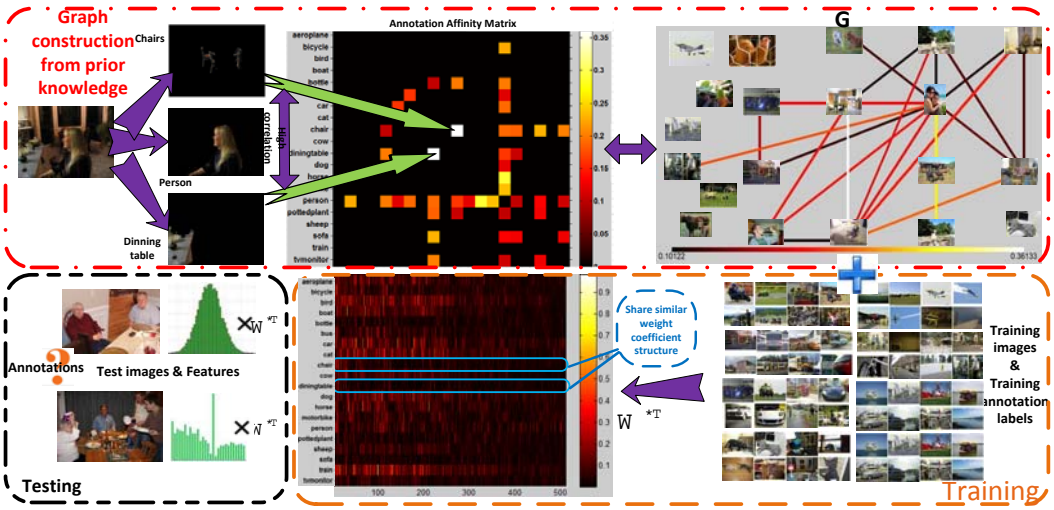


Figure 3. The flowchart of the proposed method. With the higher correlation weight, like the one between "Chairs" and "Dinning table", $\ell_{2,1}$-norm will shrink the coefficient matrix based on different weight values. And the higher weight will boost the multi-label classification via graph structured sparse learning.

straints of relation graph to tackle multi-label classifications problem. Moreover, it is a general method to incorporate graph structure information to the supervised learning. Depending on the constructed graph, we proved that our proposed algorithm can guarantee to converge to global or local solution. Extensive experiments have been conducted on six multi-label data sets. Compared with multiple state-of-art multi-label classification methods, our method consistently achieves superior classification result with respect to both precision and F1 score in macro as well as micro cases.

References

[1] A. Argyriou, T. Evgeniou, and M. Pontil. Multi-task feature learning. *NIPS*, pages 41–48, 2007.

[2] M. Boutell, J. Luo, X. Shen, and C. Brown. Learning multi-label scene classification. *Pattern Recognition*, 37(9):1757–1771, 2004.

[3] S. Ji, L. Tang, S. Yu, and J. Ye. A shared-subspace learning framework for multi-label classification. *TKDD*, 4(2), 2010.

[4] S. Kim and E. Xing. Tree-Guided Group Lasso for Multi-Task Regression with Structured Sparsity. *ICML*, 2010.

[5] D. Kong, C. Ding, H. Huang, and H. Zhao. Multi-label relieff and f-statistic feature selections for image annotation. In *CVPR*, pages 2352–2359, 2012.

[6] C. Micchelli, J. Morales, and M. Pontil. A Family of Penalty Functions for Structured Sparsity. *NIPS*, 2010.

[7] F. Nie, H. Huang, X. Cai, and C. Ding. Efficient and robust feature selection via joint l2,1-norms minimization. *NIPS*, 2010.

[8] G. Obozinski, B. Taskar, and M. Jordan. Multi-task feature selection. *Technical report, Department of Statistics, University of California, Berkeley*, 2006.

[9] T. Ojala, M. Pietikäinen, and T. Mäenpää. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Trans. Pattern Anal. Mach. Intell.*, 24(7):971–987, 2002.

[10] A. Oliva and A. Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision*, 42(3):145–175, 2001.

[11] H. Wang, C. H. Q. Ding, and H. Huang. Multi-label linear discriminant analysis. In *ECCV (6)*, pages 126–139, 2010.

[12] H. Wang, H. Huang, and C. Ding. Image annotation using multi-label correlated green's function. *IEEE Conference on Computer Vision*, pages 1–6, 2009.

[13] H. Wang, H. Huang, and C. Ding. Image annotation using bi-relational graph of images and semantic labels. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2011)*, pages 793–800, 2011.

[14] H. Wang, F. Nie, and H. Huang. Multi-view clustering and feature learning via structured sparsity. *The 30th International Conference on Machine Learning (ICML 2013)*, 2013.

[15] H. Wang, F. Nie, H. Huang, J. Yan, S. Kim, S. Risacher, A. Saykin, and L. Shen. High-Order Multi-Task Feature Learning to Identify Longitudinal Phenotypic Markers for Alzheimer Disease Progression Prediction. *Advances in Neural Information Processing Systems (NIPS)*, pages 1286–1294, 2012.

[16] H. Yu, M. Li, H. Zhang, and J. Feng. Color texture moments for content-based image retrieval. In *ICIP (3)*, pages 929–932, 2002.

[17] K. Yu, S. Yu, and V. Tresp. Multi-label informed latent semantic indexing. In *SIGIR*, pages 258–265, 2005.

[18] M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *Journal of The Royal Statistical Society Series B*, 68(1):49–67, 2006.

[19] Y. Zhang and Z.-H. Zhou. Multilabel dimensionality reduction via dependence maximization. *TKDD*, 4(3), 2010.