2014 IEEE International Conference on Data Mining

Learning Low-Rank Label Correlations for Multi-label Classification with Missing Labels

Linli Xu[†] , Zhen Wang[‡] , Zefan Shen[‡] , Yubo Wang[‡] , and Enhong Chen[†]
School of Computer Science and Technology
University of Science and Technology of China
Hefei, Anhui
[†]{*linlixu,cheneh*}*@ustc.edu.cn,* [‡]{*zwang25,szfan,wybang*}*@mail.ustc.edu.cn*

***Abstract*—Multi-label learning deals with the problem where each training example is associated with a set of labels simultaneously, with the set of labels corresponding to multiple concepts or semantic meanings. Intuitively, the multiple labels are usually correlated in some semantic space while sharing the same input space. As a consequence, the multi-label learning process can be augmented significantly by exploiting the label correlations effectively. Most of the existing approaches share the limitations in that the label correlations are typically taken as prior knowledge, which may not depict the true dependencies among labels correctly; or they do not adequately address the issue of missing labels. In this paper, we propose an integrated framework that learns the correlations among labels while training the multi-label model simultaneously. Specifically, a low rank structure is adopted to capture the complex correlations among labels. In addition, we incorporate a supplementary label matrix which augments the possibly incomplete label matrix by exploiting the label correlations. An alternating algorithm is then developed to solve the optimization problem. Extensive experiments are conducted on a number of image and text data sets to demonstrate the effectiveness of the proposed approach.**

***Keywords*-low rank; label correlation; multi-label learning; missing labels;**

I. Introduction

Multi-label learning deals with the problem where each data example exhibits multiple concepts or semantic meanings and is associated with a set of labels simultaneously. For example, in gene and protein function prediction, multiple functional labels are associated with each gene and protein [1]; in text categorization, a document can be assigned to multiple topics [2]; similarly in image annotation, an image can be tagged with several related keywords [3].

In multi-label learning, each example in the training set is represented by a feature vector and associated with a set of labels. Straightforwardly, one can tackle the problem by dividing the multi-label learning task into a set of independent binary classification problems [3], [4]. This strategy enjoys the advantages of conceptual simplicity and high efficiency. However, it may also lead to degraded performance due to the ignorance of correlations among labels.

Intuitively, the multiple labels are usually correlated in some semantic space while sharing the same input space, and it is essential to exploit the correlations among different labels to facilitate the multi-label learning process. For instance, consider the task of automatically annotating images with textual tags, where each annotation can be treated as a separate class label. As shown in Fig. 1, initially it may be difficult to decide the labels "ocean" and "sky" independently based on the color features, since they are very similar in colors. However, if we are confident that an image should be annotated with "fish", as in Fig. 1(a), then it is more likely that a region of blue in the same image should be annotated with "ocean" rather than "sky". The same principle applies for Fig. 1(b) that has been tagged with "grass", and one can naturally annotate the blue region in the image with "sky".

(a) "ocean", "fish" (b) "sky", "grass"

Figure 1. Label correlations in image annotation: Both images have large regions colored with blue. To decide the labels "sky" and "ocean", one can exploit the relations between "ocean" and "fish", "sky" and "grass" respectively.

Many approaches have been proposed to explore various types of label correlations in multi-label learning [5], [6]. Among them, the label ranking methodology considers correlations between pairs of labels and works by transforming the task into a ranking problem to order the proper labels before the improper labels for each instance [7], [8]. On the other hand, a number of approaches tackle the problem by exploiting high-order correlations among labels, where each label is influenced by the rest of the labels. Representatives include the methods of transformed label space, which work by projecting the original label vectors to low-dimensional label spaces [9], [10]. The major issue of these approaches is the separation of the label projection and model training steps, which implies the possibility that the reduced output

1550-4786/14 $31.00 © 2014 Crown Copyright
DOI 10.1109/ICDM.2014.125

representation may not augment the trained model.

Alternatively, there have been emerging interests in recent multi-label methods that take the correlation information as prior knowledge. For example, a label correlation matrix S can be calculated as the cosine similarity between label vectors and incorporated to enhance the multi-label classification performance [11]–[13]. Specifically, in the work proposed in [11], the matrix S is integrated in the objective function to enhance the prediction of label assignments; while a sparsity-inducing term based on the matrix S is proposed to regularize the multi-label model in [12].

However, there still exist some potential issues that need further concerns. First, in many real applications, it is difficult to get the complete label information for each instance, and only a "partial" set of labels are available. As a result, the methods based on modeling the original label matrix may not accurately capture the relations between labels and features due to the missing labels. Second, the commonly used measures of label correlation such as normalized cosine are usually calculated in a one-to-one way; whereas the correlations among labels can be complex and global with direct and indirect label dependencies, which keeps the one-to-one correlation measure from accurately capturing the real label relations. Third, most of the existing approaches decouple the label correlations from model training in the sense that they either first transform the label space or explicitly calculate the label correlations, followed by training the model. This may result in suboptimal models due to the lack of mutual adaptation of the two steps.

To address the above issues, in this paper, we model the global correlations among labels based on *one-to-all* reconstruction and propose an integrated framework which learns the correlations among labels while training the multi-label model simultaneously. A low rank structure is adopted to capture the local label correlations based on the intuition that a subset of labels can be closely related to each other with similar semantic contexts, while being independent of the rest. Furthermore, to address the issue of incomplete label matrices, we define a supplementary label matrix which augments the original label matrix by exploiting the label correlations. The new supplementary label matrix generally captures richer information regarding label dependence than the original label matrix. With these components, we are able to expand the forms of label correlations and achieve a novel multi-label classification method that captures more complex and flexible dependencies among labels. An alternating algorithm is developed for the optimization. Extensive experiments are conducted on diverse multi-label data sets to demonstrate the effectiveness of the proposed framework.

II. The Unified Framework

In multi-label learning, we are given a data matrix of n training examples $X = [\mathbf{x}_1, ..., \mathbf{x}_n]^\top \in \mathbb{R}^{n\times d}$ and a label matrix $Y = [\mathbf{y}_1, ..., \mathbf{y}_n]^\top \in \mathbb{R}^{n\times c}$ where c is the number of labels and $\mathbf{y}_i \in \{0, 1\}^c$ is a binary label vector indicating the label assignments of the i-th instance. Following traditional supervised learning discipline, a general classification model can be trained by solving the following problem:

$$\min_W \; L(X, W, Y) + \lambda\Omega(W) \tag{1}$$

where $L(\cdot)$ is a loss function, the regularization term $\Omega(W)$ is usually used to capture the specific structures of features and labels with various norms.

In multi-label learning, key challenges exist in the aspects of label correlations and the possibility of incomplete labels. On one hand, it is essential to exploit potential correlations among labels to cope with the exponential-sized output space in the multi-label setting; on the other hand, in real problems with large numbers of labels, it can be difficult to collect the complete label information for each instance, which makes it more complex to capture the label correlations. Specifically, with an incomplete label assignment, the absence of a label does not necessarily mean the lack of association of an instance with that label. As a consequence, approaches that directly model the original label matrix Y as in formulation (1) may not accurately capture the relations among labels and features.

To tackle this problem, we propose to learn the label correlations which can be exploited to augment the incomplete label matrix and obtain a new supplementary label matrix $\hat{Y} \in \mathbb{R}^{n\times c}$ with essentially richer information regarding label correlations. By exploiting the co-occurrence and dependency of related labels, we assume that the predictive confidence $\hat{Y}$ is determined by the available original label information Y and the correlations among different labels, which is modeled by the correlation matrix $S \in \mathbb{R}^{c\times c}$. Formally, motivated by the idea of label dependency propagation [14], the original label matrix can be supplemented by directly multiplying with the label correlation matrix S:

$$\begin{aligned} \hat{Y}_{i,j} &= Y_{i,1} \times S_{1,j} + Y_{i,2} \times S_{2,j} + ... + Y_{i,c} \times S_{c,j} \\ &= \sum_{t=1}^{c} Y_{i,t} \times S_{t,j} \end{aligned} \tag{2}$$

where the element $\hat{Y}_{ij}$ can be regarded as the predictive confidence of the instance $\mathbf{x}_i$ being associated with the j-th label, which is influenced by the prior information of all the other labels in the original label matrix.

A simple example is shown in Fig. 2, where we assume that the only available label information of Image 1 in Fig. 1(a) is "fish". By exploiting the label correlations, a supplementary matrix $\hat{Y}$ is obtained where the predictive confidence of label "ocean" for Image 1 is high due to a strong correlation between the labels "fish" and "ocean", and the predictive confidence of label "sky" is low because of a weak dependence from label "fish" to label "sky". As a consequence, given the supplementary label matrix $\hat{Y}$, one

can augment the label assignment of Image 1 to "fish" and "ocean" from the original label "fish".

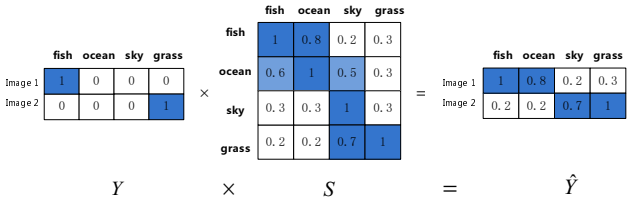

Figure 2. A supplementary label matrix $\hat{Y}$ obtained by multiplying the label correlation matrix S with the original label matrix Y

The example above is based on the assumption that the correlation matrix S can accurately capture the real relations shared among different labels, which will lead to the supplementary matrix $\hat{Y}$ with richer label information. To model the complementary influence of $\hat{Y}$ and S, we propose a unified framework that learns the label correlations while training the classification model simultaneously. Without loss of generality, we adopt the least square loss

$$\begin{aligned} \min_{W,S,E,\hat{Y}} \ & \|XW - \hat{Y}\|_F^2 + \lambda_1 \|W\|_F^2 + \lambda_3 \|E\|_{2,1} \\ \text{s.t. } & \hat{Y} = YS;\ Y = \hat{Y} + E \end{aligned} \tag{3}$$

where in the objective function the original label matrix Y is replaced by the supplementary label matrix $\hat{Y} = YS$. In the meantime, the constraint $Y = \hat{Y} + E$ and the regularization term on E work together and control the difference between Y and $\hat{Y}$. Specifically, to regularize the difference between Y and $\hat{Y}$ while facilitating a label-wise (i.e., column-wise) sparsity on E, we adopt the convex $\ell_{2,1}$ norm as regularization, which can be defined as $\|E\|_{2,1} = \sum_{j=1}^{c} \sqrt{\sum_{i=1}^{n} (E_{ij})^2}$.

The formulation above can then be further rewritten as

$$\begin{aligned} \min_{W,S,E} \ & \|XW - YS\|_F^2 + \lambda_1 \|W\|_F^2 + \lambda_3 \|E\|_{2,1} \\ \text{s.t. } & Y = YS + E \end{aligned} \tag{4}$$

based on which we can observe that the correlation S_{ij} of label pair (i, j) is influenced by all the other labels in (4), which implies a high-order one-to-all dependency rather than a one-to-one correlation. This helps to capture the complex and global correlations that arise from direct and indirect label dependencies.

Besides the global characteristics, the label correlations encoded in S should also capture some local patterns as well. For example, there usually exists grouping of labels such that the labels within a group are strongly correlated with each other, while being independent of the rest. These local patterns essentially imply a low-rank or even a block-diagonal structure of the S matrix, which can be incorporated into the model as follows:

$$\begin{aligned} \min_{W,S,E} \ & \|XW - YS\|_F^2 + \lambda_1 \|W\|_F^2 + \lambda_2 \operatorname{rank}(S) + \lambda_3 \|E\|_{2,1} \\ \text{s.t. } & Y = YS + E \end{aligned}$$

Unfortunately, the rank function is hard to optimize, the nuclear norm $\|\cdot\|_*$ is therefore employed here as a convex approximation of the rank function. The framework of learning low-rank label correlations for multi-label classification with missing labels can then be formulated as

$$\begin{aligned} \min_{W,S,E} \ & \|XW - YS\|_F^2 + \lambda_1 \|W\|_F^2 + \lambda_2 \|S\|_* + \lambda_3 \|E\|_{2,1} \\ \text{s.t. } & Y = YS + E \end{aligned} \tag{5}$$

By solving (5), one can not only learn the correlation matrix S, but also train a multi-label classification model by exploiting the label correlations.

III. Optimization

The optimization problem in (5) is convex and therefore can be optimized globally. In this section, we propose an alternating iterative algorithm to solve for the label correlation matrix S and the multi-label parameters W.

A. Computing W Given S, E

With S and E given, W is the only variable and the problem turns into:

$$\min_{W} \ \|XW - YS\|_F^2 + \lambda_1 \|W\|_F^2 \tag{6}$$

which is a classic ridge regression problem with a closed-form solution:

$$W = (X^\top X + \lambda_1 I)^{-1} X^\top YS \tag{7}$$

B. Computing S Given W

Compared to the closed-form solution of W given S, it is not trivial to optimize S given W due to the two non-smooth regularization terms in (5). Here we first introduce an auxiliary variable Z to make the objective function separable. Problem (5) can then be rewritten as:

$$\begin{aligned} \min_{Z,S,E} \ & \|XW - YS\|_F^2 + \lambda_2 \|Z\|_* + \lambda_3 \|E\|_{2,1} \\ \text{s.t. } & Y = YS + E,\ S = Z. \end{aligned} \tag{8}$$

By introducing augmented Lagrangian multipliers and incorporating the equality constraints into the cost function, the problem is transformed into:

$$\begin{aligned} \min_{S,Z,E,\Lambda_1,\Lambda_2} \ & \|XW - YS\|_F^2 + \lambda_2 \|Z\|_* + \lambda_3 \|E\|_{2,1} \\ & + \frac{\rho}{2} \|Y - YS - E + \frac{\Lambda_1}{\rho}\|_F^2 - \frac{1}{2\rho} \|\Lambda_1\|_F^2 \\ & + \frac{\rho}{2} \|S - Z + \frac{\Lambda_2}{\rho}\|_F^2 - \frac{1}{2\rho} \|\Lambda_2\|_F^2. \end{aligned} \tag{9}$$

Then the inexact ALM (IALM) method is applied to solve for each variable in (9) iteratively with blockwise coordinate descent procedures. Each iteration of IALM involves updating one variable, with the other variables fixed to their most recent values [15]. The updating rules are as follows:

[Update Z^{k+1}] According to singular value thresholding [16], the solution is given by

$$Z^{k+1} = J_{\frac{\lambda_2}{\rho}}(S^k + \frac{\Lambda_2^k}{\rho}) \tag{10}$$

where $J_\lambda(A) = U_A S_\lambda(\Sigma_A) V_A^\top$ is the singular value operator with $A = U_A \Sigma_A V_A^\top$ being the singular value decomposition of A, and $S_\lambda(A_{ij}) = \text{sign}(A_{ij}) \max(0, |A_{ij}| - \lambda)$ is the soft-thresholding operator.

[Update S^{k+1}] Taking the derivative of the objective and setting it to zero, we have:

$$S^{k+1} \quad = \quad [(2+\rho)Y^\top Y + \rho I]^{-1} T \tag{11}$$

where

$$T = 2Y^\top XW + \rho(Y^\top Y - Y^\top E^k + Z^{k+1}) + Y^\top \Lambda_1^k - \Lambda_2^k.$$

[Update E^{k+1}] According to the $\ell_{2,1}$ minimization operator [17], the solution can be computed as follows:

$$E^{k+1}(:,i) = \begin{cases} \frac{\|\mathbf{q}_i^k\| - \frac{\lambda_3}{\rho}}{\|\mathbf{q}_i^k\|} \mathbf{q}_i^k & \text{if } \|\mathbf{q}_i^k\| > \frac{\lambda_3}{\rho} \\ 0 & \text{otherwise} \end{cases} \tag{12}$$

where $Q^k = Y - YS^{k+1} + \frac{\Lambda_1^k}{\rho}$, $\mathbf{q}_i^k$ is the i-th column of Q^k and $E^{k+1}(:,i)$ is the i-th column of the optimal solution E^{k+1}.

[Update Multipliers $\Lambda_1^{k+1}, \Lambda_2^{k+1}$] $\Lambda_1^{k+1}, \Lambda_2^{k+1}$ can be updated directly by

$$\begin{aligned} \Lambda_1^{k+1} &= \Lambda_1^k + \rho(Y - YS^{k+1} - E^{k+1}) \\ \Lambda_2^{k+1} &= \Lambda_2^k + \rho(S^{k+1} - Z^{k+1}) \end{aligned} \tag{13}$$

Note that all the updates above are in closed forms and the iterative updates for all the variables in the inexact ALM algorithm are outlined in Algorithm 1.

Algorithm 1 Solve Problem (8) via Inexact ALM

Input: data matrix X, label matrix Y, weight matrix W, parameter $\lambda_1, \lambda_2, \lambda_3$
Output: S, Z, E
Initialize $Z = S = 0$, $E = 0$, $\Lambda_1 = 0$, $\Lambda_2 = 0$, $\rho = 10^{-6}$, $\max_\rho = 10^{10}$, $\mu = 1.1$, $\epsilon = 10^{-6}$
while not converged **do**
  fix S, E and update variable Z according to (10)
  fix Z, E and update variable S according to (11)
  fix Z, S and update variable E according to (12)
  fix S, Z, E and update the multipliers Λ_1 and Λ_2 according to (13)
  update the parameter ρ by$\rho = \min(\rho\mu, \max_\rho)$
  check the convergence conditions:
  $\|Y - YS - E\|_\infty < \epsilon$,$\|Z - S\|_\infty < \epsilon$
end while

The overall procedure of learning low-rank label correlations for multi-label classification (ML-LRC) can be summarized in Algorithm 2.

Algorithm 2 The ML-LRC Framework

Input: training data set $\{X, Y\}$, parameter $\lambda_1, \lambda_2, \lambda_3$
Output: W, S
Initialize $S = I$
repeat
  fix S and update W according to (7)
  fix W and solve for S using Algorithm 1
until convergence

Table I
CHARACTERISTICS OF THE DATA SETS

Data set	$\|D\|$	$\dim(D)$	$L(D)$	$F(D)$	$LC(D)$	Domain
Emotions	593	72	6	numeric	1.869	music
Birds	645	260	19	numeric	1.014	audio
Enron	1702	1001	53	nominal	3.378	text
Image	2000	294	5	numeric	1.236	image
Scene	2407	294	6	numeric	1.074	image
Pascal06	5304	960	10	numeric	1.263	image
Bibtex	7395	1836	159	nominal	2.402	text

IV. EXPERIMENTS

A. Experimental Setup

1) Data Sets: Experiments are run on 7 real-world multi-label data sets from diverse domains. The characteristics of the data sets are summarized in Table I. For each data set D, we use $|D|$, $\dim(D)$, $L(D)$, $F(D)$, $LC(D)$ to denote the number of instances, number of features, number of possible labels, feature types and label cardinality which corresponds to the average number of labels per instance.

2) Evaluation Metrics: As discussed in [6], the generalization performance of a multi-label classification method is not only measured from a classification perspective, but also measured from a label ranking perspective. In this paper, *Ranking loss*, *One-error*, *Coverage*, and *Average AUC* are used to measure the performance of multi-label algorithms from different aspects. For Average AUC, the larger the values the better the performance, while for the other three metrics, smaller values indicate better performance.

3) Baselines: To examine the effectiveness of our framework, ML-LRC is firstly compared with Ridge Regression, which can be considered as a degenerated version of ML-LRC without exploiting label correlations. ML-LRC is also compared with 3 state-of-the-art multi-label methods: ML-KNN [4] which adapts the k-nearest neighbor principle to generate a set of independent classifiers; and MLLS [18] which models the correlations among labels by a common subspace shared by all the classifiers. Finally, in order to verify the mutual reinforcement of calculating the label correlations and training the model simultaneously in our framework, we compare with LSG21 [12] which decouples the two steps.

4) Parameter Settings: For the competing algorithms, we use parameter configurations as suggested in the corresponding papers. Furthermore, the regularization parameters of all

the methods are tuned using 5-fold cross validation.

B. Classification Results

All the algorithms are run on the data sets with 5-fold cross validation, where Tables II to V report the results of various algorithms in terms of different evaluation metrics. For the larger data set "Bibtex", LSG21 [12] has not trained a predictive model in a reasonable time, and is marked as DNF (Did Not Finish) in the tables. On each data set, the mean of the evaluation metrics is recorded. A bold number indicates the best performance on the corresponding data set.

From the results shown in Table II to Table V, we can observe that the proposed framework achieves better or comparable performance in terms of all the 4 measures on all the data sets. Specifically, when comparing to Ridge Regression, which is a degenerated version of ML-LRC without exploiting label correlations, the performance of our approach is significantly superior by 15.51% and 12.94% averaged on all data sets in terms of RankLoss and Coverage respectively, with only slight advantages of 3.85%, 2.46% regarding the other two measures, verifying the effectiveness of exploiting label correlations for boosting the classification performance.

In the meantime, the proposed ML-LRC outperforms ML-KNN and MLLS by 15.90% and 7.45% respectively, averaged over all the data sets and all the measures. The lower performance of ML-KNN may be due to the limitation that ML-KNN handles the classification tasks independently. On the other hand, the comparison between our method and MLLS indicates more accurate label dependencies are encoded in the correlation matrix S in our framework than the shared subspace in [18].

Last, the performance of ML-LRC is comparable with LSG21 regarding Average AUC, with an improvement of 15.17%, 7.17% and 12.57% in terms of RankLoss, One-error and Coverage respectively. This verifies that simultaneously calculating the label correlations and training the model in our framework can facilitate the classification performance.

C. Label Correlations

Besides evaluating the multi-label classification performance, we also examine the label correlations that are learned simultaneously. We first take the correlation matrix S learned from "Pascal06" which is a data set for visual object recognition with 10 class labels including bicycle, bus, car, cat, cow, dog, horse, motorbike, person and sheep; and then obtain a relational matrix A after some post-processing steps such as normalization, thresholding etc. The relational matrix A can be further represented in a semantic relational graph $G = \{V, E\}$ shown in Fig. 3, where nodes in V are the annotation labels and weights of the edges in E correspond to the correlation values between labels.

Table II
SUMMARY OF PERFORMANCE IN TERMS OF *ranking loss*.

Data	Algorithm				
	ML-LRC	ML-KNN	Ridge-Rg	LSG21	MLLS
Emotions	**0.1531**	0.2701	0.1830	0.1696	0.1697
Birds	**0.2308**	0.2820	0.2592	0.2481	0.2879
Enron	**0.0789**	0.0865	0.1439	0.1671	0.1311
Image	**0.1626**	0.1822	0.1718	0.1753	0.1676
Scene	**0.0745**	0.0827	0.0802	0.0826	0.0811
Pascal06	**0.1392**	0.1737	0.1481	0.1457	0.1436
Bibtex	**0.0732**	0.1094	0.0889	DNF	0.0856

Table III
SUMMARY OF PERFORMANCE IN TERMS OF *one-error*.

Data	Algorithm				
	ML-LRC	ML-KNN	Ridge-Rg	LSG21	MLLS
Emotions	**0.2582**	0.4518	0.3022	0.2838	0.2721
Birds	**0.4814**	0.7104	0.4958	0.4845	0.5724
Enron	0.2491	0.2438	0.2485	0.3247	**0.2166**
Image	**0.3084**	0.3378	0.3240	0.3281	0.3134
Scene	**0.2304**	0.2351	0.2361	0.2375	0.2333
Pascal06	**0.4082**	0.4702	0.4142	0.4126	0.4107
Bibtex	**0.3459**	0.4307	0.3519	DNF	0.3492

Table IV
SUMMARY OF PERFORMANCE IN TERMS OF *coverage*.

Data	Algorithm				
	ML-LRC	ML-KNN	Ridge-Rg	LSG21	MLLS
Emotions	**1.7290**	2.2903	1.8855	1.8241	1.8225
Birds	**6.4379**	7.3257	7.1480	6.9790	7.7991
Enron	**11.8582**	12.6572	19.7554	21.5806	18.5457
Image	**0.9249**	1.0029	0.9708	0.9712	0.9446
Scene	**0.4556**	0.4994	0.4858	0.4991	0.4908
Pascal06	**1.6699**	1.9898	1.7639	1.7375	1.7274
Bibtex	**22.5382**	31.3515	26.8587	DNF	25.7612

Table V
SUMMARY OF PERFORMANCE OF IN TERMS OF *average AUC*.

Data	Algorithm				
	ML-LRC	ML-KNN	Ridge-Rg	LSG21	MLLS
Emotions	**0.8370**	0.7145	0.8195	0.8243	0.8254
Birds	**0.7323**	0.5661	0.7042	0.7023	0.6888
Enron	**0.7635**	0.6658	0.7151	0.7348	0.7081
Image	**0.8319**	0.8189	0.8221	0.8273	0.8296
Scene	**0.9295**	0.9286	0.9204	0.9234	0.9275
Pascal06	**0.8606**	0.8198	0.8535	0.8551	0.8578
Bibtex	**0.9228**	0.8134	0.9108	DNF	0.9218

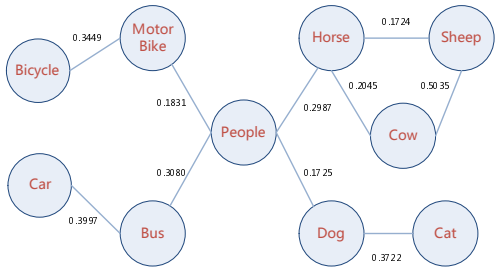

Figure 3. The semantic relational graph obtained by ML-LRC for the Pascal06 data set

D. Incomplete Labels

In order to evaluate the robustness of the proposed method to missing labels, we take the "Pascal06" data set and mask a ratio of the labels, and then compare with Ridge Regression, MLLS and LSG21 which all directly model the label matrices. We vary the ratios of observed labels in "Pascal06" data set from 30% to 80% with 10% as the interval, and Fig. 4 to 5 present the curves of various metrics with different ratios of observed labels. It can be demonstrated that ML-LRC is superior to all the other methods with different performance measures and different levels of incomplete label information, which justifies the capability of ML-LRC handling data with missing labels.

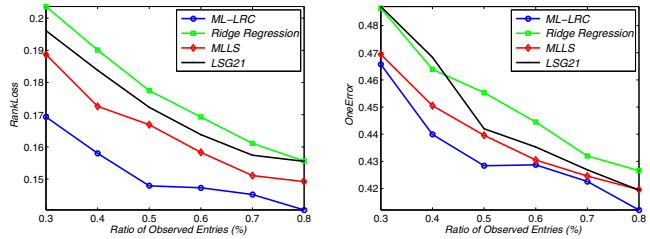

Figure 4. RankLoss and OneError with different ratios of observed labels

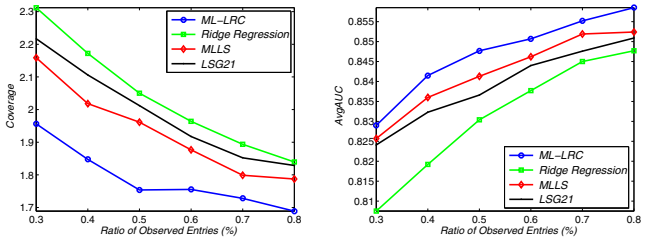

Figure 5. Coverage and Average AUC with different ratios of observed labels

V. Conclusion

In this paper we propose an integrated framework ML-LRC which learns the correlations among labels while training the multi-label model simultaneously. A low rank structure is adopted to capture the complex correlations among labels. In the meantime, to address the issue of incomplete labels, we incorporate a supplementary label matrix which augments the original label matrix by exploiting the label correlations. With the complementary interactions of model training and correlation learning, the proposed method not only exhibits a superiority in label prediction, but also captures more complex and flexible dependencies among labels. Moreover, scenarios of missing labels can be handled effectively by exploiting the label correlations.

Acknowledgment

Research supported by the National Natural Science Foundation of China (No. 61375060) and the Fundamental Research Funds for the Central Universities (WK0110000036).

References

[1] Z. Barutcuoglu, R. Schapire, and O. Troyanskaya, "Hierarchical multi-label prediction of gene function," *Bioinformatics*, vol. 22, no. 7, pp. 830–836, 2006.

[2] N. Ueda and K. Saito, "Parametric mixture models for multi-labeled text," in *NIPS 15*. MIT Press, 2003.

[3] M. Boutell, J. Luo, X. Shen, and C. Brown, "Learning multi-label scene classification," *Pattern Recognition*, vol. 37, no. 9, pp. 1757–1771, 2004.

[4] M. Zhang and Z. Zhou, "Ml-knn: A lazy learning approach to multi-label learning," *Pattern Recognition*, vol. 40, no. 7, pp. 2038–2048, 2007.

[5] J. Read, B. Pfahringer, G. Holmes, and E. Frank, "Classifier chains for multi-label classification," *Machine Learning*, vol. 85, no. 3, pp. 333–359, 2011.

[6] M. Zhang and Z. Zhou, "A review on multi-label learning algorithms," *IEEE Transactions on Knowledge and Data Engineering*, vol. 99, 2013.

[7] A. Elisseeff and J. Weston, "A kernel method for multi-labelled classification," in *NIPS 14*, 2002, pp. 681–687.

[8] J. Frnkranz, E. Hllermeier, and E. Mencła, "Multilabel classification via calibrated label ranking," *Machine Learning*, vol. 73, no. 2, pp. 133–153, 2008.

[9] D. Hsu, S. Kakade, and J. Langford, "Multi-label prediction via compressed sensing," in *NIPS*, 2009, pp. 772–780.

[10] T. Zhou, D. Tao, and X. Wu, "Compressed labeling on distilled labelsets for multi-label learning," *Machine Learning*, vol. 88, no. 1-2, pp. 69–126, 2012.

[11] H. Wang, H. Huang, and C. Ding, "Image annotation using multi-label correlated green's function," in *ICCV*, 2009, pp. 2029–2034.

[12] X. Cai, F. Nie, W. Cai, and H. Huang, "New graph structured sparsity model for multi-label image annotation," in *ICCV*, 2013.

[13] H. Wang, C. Ding, and H. Huang, "Multi-label linear discriminant analysis," in *ECCV*, 2010, pp. 126–139.

[14] B. Fu, G. Xu, and Z. W. amd L. Cao, "Leveraging supervised label dependency propagation for multi-label learning," in *ICDM*, 2013, pp. 1061–1066.

[15] Z. Lin, A. Ganeshm, J. Wright, and Y. Ma, "Fast convex optimization algorithm for exact recovery of a corrupted low-rank matrix," Tech. Rep. UILU-ENG-09-2214, 2009.

[16] J. Cai, E. Cands, and Z. Shen, "A singular value thresholding algorithm for matrix completion," *SIAM Journal on Optimization*, vol. 20, no. 4, pp. 1956–1982, 2010.

[17] G. Liu, Z. Lin, and Y. Yu, "Robust subspace segmentation by low-rank representation," in *ICML*, 2010, pp. 663–670.

[18] S. Ji, L. Tang, S. Yu, and J. Ye, "Extracting shared subspace for multi-label classification," in *KDD*, 2008, pp. 381–389.