MULTI-VIEW METRIC LEARNING FOR MULTI-LABEL IMAGE CLASSIFICATION

Mengying Zhang[1] *Changsheng Li*[2] *Xiangfeng Wang*[3]

[1]National University of Singapore, Singapore
[2]University of Electronic Science and Technology of China, Chengdu, China
[3]East China Normal University, Shanghai, China

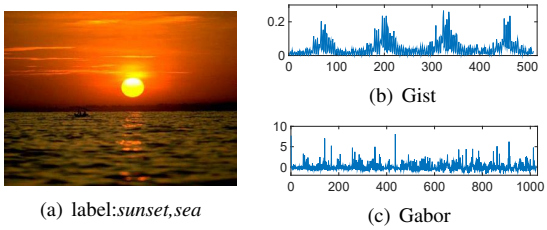
ABSTRACT

Multi-label image classification is a very challenging task, where data are often associated with multiple labels and represented with multiple views. In this paper, we propose a novel multi-view distance metric learning approach to dealing with the multi-label image classification problem. In particular, we attempt to concatenate multiple types of features after learning one optimal distance metric for each view, so as to obtain a better joint representation across multi-view spaces. To preserve the intrinsic geometric structure of the data in the low-dimensional feature space, we introduce a manifold regularization with the adjacency graph being constructed based on all labels. Moreover, we learn another distance metric to capture the dependency of labels, which can further improve the classification performance. Experimental results on publicly available image datasets demonstrate that our method achieves superior performance, compared with the state-of-the-arts.

Index Terms— Multi-label image classification, distance metric learning, multi-view learning

1. INTRODUCTION

In many real-world scenarios, images are often associated with multiple labels instead of one single label and can be represented from multiple views, as shown in Fig. 1. Because of complicated relationships among labels and intrinsic characteristics of data, it is very challenging to predict all labels simultaneously. To deal with this problem, multi-label image classification has been proposed [1, 2, 3], whose goal is to learn a series of classifiers to map an image into a label vector with a fixed size. Multi-label image classification can be simply achieved by casting this task into several binary-class subproblems, in which each subproblem is to predict whether the image is relevant to the corresponding label. This method assumes that different labels are independent. However, in practice, there are often correlations among labels, e.g., plane and sky often appears in an image simultaneously. Empirically and theoretically speaking, such correlations can help predict labels of testing images more accurately [4, 5, 6, 7, 8, 9].



(a) label:*sunset,sea* (b) Gist (c) Gabor

Fig. 1. Image (a) is annotated with two labels: *subset* and *sea*, and can be represented by two views: (b) Gist [10] features (c) Gabor features.

By far, many multi-label learning approaches for image classification have been proposed in the past decade [11, 12, 13, 14, 15, 16]. Most of them aim to leverage relationship of labels into the learning process for better model performance. Because of the space limitation, we briefly survey some distance metric learning based multi-label prediction algorithms (For a complete review, please refer to [17]). The authors in [18] presented an iterative multi-label learning algorithm by alternating the steps between estimating instance-label association and learning distance metrics from the estimated association. A maximum margin output coding (MMOC) formulation was proposed in [19] to learn a distance metric for capturing the correlations between inputs and outputs. Moreover, the method in [20, 21] incorporated k nearest neighbor (kNN) constraints into a similar formulation to [19]. The authors in [22] introduced linear and nonlinear distance metric learning methods to improve the performance of kNN for multi-label data. All the methods mentioned above learn various distance metrics based on only one single view, but ignore the fact that images are often represented by multi-view features.

Recently, multi-view learning has made great progress and developments in various image applications [23]. Comparing to those who only use one single view, it achieves better performance by combining multiple visual features. In the meantime, there are also a few multi-view learning methods proposed to deal with the multi-label image classification problems [24, 25, 26]. In order to integrate multi-view features, [24] introduced a multi-view vector-valued manifold regularization in the proposed formulation. The method in [25] sought a common low-dimensional representation by matrix factorization and then utilized matrix completion to

conduct classification. [26] simultaneously learned a common representation of multiple views and a corresponding projection model between representation and labels. However, these methods utilize the same distance metric for each view, which might not be optimal in real-world applications.

Different from the above methods, in this paper, we propose a novel Multi-view Distance Metric Learning method for Multi-label image classification, named as MDMLM. Specifically, we first learn an optimal distance metric for each view, and then concatenate features of all views based on the learnt metrics, such that a better joint feature representation can be obtained across multi-view spaces. Meanwhile, a manifold regularization with the adjacency graph being constructed by all labels is introduced in the formulation, so as to preserve the intrinsic geometric structure of the data in the low-dimensional feature space. Moreover, we attempt to learn another distance metric in order to capture the dependency of labels. Finally, we develop a simple but effective algorithm to solve the proposed optimization problem. The proposed method is evaluated with exhaustive experiments on public available multi-label image datasets including scene and mirflickr. Experimental results demonstrate that the proposed method achieves significantly better performance compared to the state-of-the-art multi-label classification methods.

2. METHOD

In this section, we first present the preliminaries, then detail the formulation of proposed method, and finally describe the solution to the optimization problem.

2.1. Notations and Preliminaries

Let $\mathcal{S} = \{\mathbf{x}_i, \mathbf{y}_i\}_{i=1}^n$ denote a set of training data. $\mathbf{x}_i \in \mathbb{R}^d$ is the feature representation of the i-th input image. $\mathbf{y}_i = [y_i^1, y_i^2, ..., y_i^L]^T$ are the ground-truth labels, where y_i^j is a binary indicator. $y_i^j = 1$ indicates image $\mathbf{I}_i$ is tagged with the j-th label, and $y_i^j = 0$ otherwise. n and L denote the number of all training images and all possible labels respectively. Our goal is to learn a series of classifiers to map $\mathbf{x}_i$ to a vector $\widehat{\mathbf{y}}_i$, such that $\widehat{\mathbf{y}}_i$ are close to $\mathbf{y}_i$ as much as possible.

To achieve this goal, we introduce the following formulation [20]:

$$\begin{aligned} &\arg \min_{\mathbf{U}\in\mathbb{R}^{l\times s}, \xi_i \geq 0} \frac{1}{2}\|\mathbf{U}\|_F^2 + \frac{C}{n}\sum_{i=1}^{n}\xi_i^2 \\ s.t.\ \ &\|\mathbf{U}^T\mathbf{P}^T\mathbf{x}_i - \mathbf{U}^T\mathbf{y}_i\|_2^2 + \triangle(\mathbf{y}_i, \mathbf{y}) - \xi_i \\ &\leq \|\mathbf{U}^T\mathbf{P}^T\mathbf{x}_i - \mathbf{U}^T\mathbf{y}\|_2^2, \forall \mathbf{y} \in Nei(i), \forall i \end{aligned} \tag{1}$$

where $\|\cdot\|_F$ is the Frobenius norm and $\|\cdot\|_2$ is the l_2 norm. C is a hyperparameter to control the balance between a square-hinge loss function and a regularizer. $\mathbf{y} \in Nei(i)$ denotes the output set of k nearest neighbors of $\mathbf{x}_i$. $\triangle(\mathbf{y}_i, \mathbf{y})$ is the margin, and is defined as $\triangle(\mathbf{y}_i, \mathbf{y}) = \|\mathbf{y}_i - \mathbf{y}\|_1$ [19]. $\mathbf{P} \in \mathbb{R}^{d\times l}$ is the projection matrix, and is simply estimated via the least squares in [20].

The core idea in (1) is to learn a projection matrix $\mathbf{U}$, making for any sample $\mathbf{x}_i$, the prediction distance to the correct labels $\|\mathbf{U}^T\mathbf{P}^T\mathbf{x}_i - \mathbf{U}^T\mathbf{y}_i\|_2^2$ be smaller than the prediction distance to the label set of k nearest neighbors of $\mathbf{x}_i$ $\|\mathbf{U}^T\mathbf{P}^T\mathbf{x}_i - \mathbf{U}^T\mathbf{y}\|_2^2$ by a margin of at least $\triangle(\mathbf{y}_i, \mathbf{y})$.

In (1), the images are represented by only one single view. However, images can be represented by multiple views, e.g., color, shape and texture. The classification performance can be improved if a good fusion strategy to these features can be designed. In addition, in (1), the projection matrix $\mathbf{P}$ is simply estimated via the least squares, and is fixed in the learning process, leaving the solution of (1) less optimal.

2.2. Proposed Formulation

Let $\mathbf{x}_i^m \in \mathbb{R}^{d_m}$ be the feature representation of the i-th input image in the view m, in which d_m is the dimension of $\mathbf{x}_i^m$. $m = 1, 2, \ldots, M$, where M denotes the total number of views. As we know, different views usually come from different feature spaces. Therefore, concatenating these features directly may not be physically meaningful. To overcome this limitation, we learn a linear transformation for each view, and then concatenate the mapped features as a joint feature representation. The final feature can be represented by

$$\mathbf{q}_i = [(\mathbf{q}_i^1)^T, \ldots, (\mathbf{q}_i^M)^T]^T \tag{2}$$

where $\mathbf{q}_i^m$ is the mapped feature representation of the m-th view, and can be obtained by $\mathbf{q}_i^m = \mathbf{W}^m\mathbf{x}_i^m$. $\mathbf{W}^m \in \mathbb{R}^{r_m\times d_m}$ is the learned projection matrix for the m-th view, where r_m is the size of $\mathbf{q}_i^m$.

Based on the new feature representation, we propose a new objective function to reach our goal. The new formulation can be expressed as:

$$\begin{aligned} &\min \mathcal{J}(\mathbf{U}, \mathbf{V}, \mathbf{W}^m) = \frac{1}{2}\|\mathbf{U}\|_F^2 + \frac{C}{n}\sum_{i=1}^{n}\mathcal{L}_i \\ &+\lambda\sum_{i,j=1}^{n}\|\mathbf{V}^T\mathbf{q}_i - \mathbf{V}^T\mathbf{q}_j\|_2^2 A_{ij} + \eta(\|\mathbf{V}\|_F^2 + \sum_{m=1}^{M}\|\mathbf{W}^m\|_F^2) \end{aligned} \tag{3}$$

where λ and η are two tradeoff parameters. $\mathcal{L}_i$ is a hinge loss function, and is defined as $\mathcal{L}_i = [\min(0, \|\mathbf{U}^T\mathbf{V}^T\mathbf{q}_i - \mathbf{U}^T\mathbf{y}\|_2^2 - \|\mathbf{U}^T\mathbf{V}^T\mathbf{q}_i - \mathbf{U}^T\mathbf{y}_i\|_2^2 - \triangle(\mathbf{y}_i, \mathbf{y}))]^2$, where $\mathbf{y}$ belongs to k nearest neighbors of $\mathbf{y}_i$, denoted as $kNN(\mathbf{y}_i)$. $\mathbf{V}^T\mathbf{q}_i$ is the new feature representation of $\mathbf{q}_i$. A_{ij} is the adjacency graph. Here, we use the label space to construct A_{ij} as:

$$A_{ij} = \begin{cases} e^{-\frac{\|\mathbf{y}_i - \mathbf{y}_j\|_2^2}{\sigma}}, & \text{if } \mathbf{y}_i \in kNN(\mathbf{y}_j) \text{ or } \mathbf{y}_j \in kNN(\mathbf{y}_i) \\ 0, & \text{otherwise} \end{cases}$$

In (3), the second term aims to learn a distance metric based on the joint feature representation with multiple views. The third term attempts to preserve the local manifold structure in the new feature space. The last two terms are two regularization terms to limit their complexities.

Our Eq. (3) is fundamentally different from Eq. (1) in the following three aspects:

- Our method takes advantage of features from multiple views, while Eq. (1) only refers to one single view.
- We attempt to preserve the local structure of data in the new low-dimensional feature space, while Eq. (1) ignores this point.
- Instead of Eq. (1) simply estimating the projection matrix $\mathbf{P}$ by the least squares, we seek the optimal solution of the projection matrix in a joint learning framework.

2.3. Optimization

The objective function (3) is not jointly convex with regard to the variables $\mathbf{U}$, $\mathbf{V}$, $\{\mathbf{W}^m\}_{m=1}^M$simultaneously. It is unrealistic to expect an algorithm to easily find the globally optimal solution. Therefore, we employ an alternating optimization strategy to solve this problem and obtain a local optimal solution instead.

Update $\mathbf{U}$ with fixed $\mathbf{V}$ and $\{\mathbf{W}^m\}_{m=1}^M$: with the fixed variables $\mathbf{V}$ and $\{\mathbf{W}^m\}_{m=1}^M$, we update $\mathbf{U}$ in (3) using the gradient descent method. The partial derivative of the objective function (3) with regard to $\mathbf{U}$ is given as:

$$\frac{\partial \mathcal{J}}{\partial \mathbf{U}} = \Bigg\{ \mathbf{I} + \frac{4C}{n} \sum_{(\mathbf{q}_i, \mathbf{y}_i, \mathbf{y}) \in \Theta} \Big[g_i (\mathbf{V}^T \mathbf{q}_i - \mathbf{y})(\mathbf{V}^T \mathbf{q}_i - \mathbf{y})^T - g_i (\mathbf{V}^T \mathbf{q}_i - \mathbf{y}_i)(\mathbf{V}^T \mathbf{q}_i - \mathbf{y}_i)^T \Big] \Bigg\} \mathbf{U},$$

where $\mathbf{I} \in \mathbb{R}^{l \times l}$ is an identity matrix. The variable $g_i = \|\mathbf{U}^T\mathbf{V}^T\mathbf{q}_i - \mathbf{U}^T\mathbf{y}\|_2^2 - \|\mathbf{U}^T\mathbf{V}^T\mathbf{q}_i - \mathbf{U}^T\mathbf{y}_i\|_2^2 - \triangle(\mathbf{y}_i, \mathbf{y})$. Θ denotes the set of $(\mathbf{q}_i, \mathbf{y}_i, \mathbf{y})$ satisfying the following condition: $g_i < 0$. Then, $\mathbf{U}$ is updated by a gradient descent method as

$$\mathbf{U} = \mathbf{U} - \theta \frac{\partial \mathcal{J}}{\partial \mathbf{U}} \tag{4}$$

where θ is the learning rate. In the experiment, θ is updated by $\theta = 0.98 \times \theta$ after each epoch.

Update $\mathbf{V}$ with fixed $\mathbf{U}$ and $\{\mathbf{W}^m\}_{m=1}^M$: We still use the gradient descent method to update $\mathbf{V}$ when fixing $\mathbf{U}$ and $\{\mathbf{W}^m\}_{m=1}^M$. The partial derivative of $\mathcal{J}$ with regard to $\mathbf{V}$ is given as:

$$\frac{\partial \mathcal{J}}{\partial \mathbf{V}} = \frac{4C}{n} \sum_{(\mathbf{q}_i, \mathbf{y}_i, \mathbf{y}) \in \Theta} \Big[g_i \times \mathbf{t}_i \Big] + 2\eta \mathbf{V} + 2\lambda \sum_{i,j=1}^{n} A_{ij} (\mathbf{q}_i - \mathbf{q}_j)(\mathbf{q}_i - \mathbf{q}_j)^T \mathbf{V}$$

where $\mathbf{t}_i = \mathbf{q}_i (\mathbf{y}_i - \mathbf{y})^T \mathbf{U}\mathbf{U}^T$. Then, $\mathbf{V}$ is updated by a gradient descent method as

$$\mathbf{V} = \mathbf{V} - \theta \frac{\partial \mathcal{J}}{\partial \mathbf{V}} \tag{5}$$

Update $\{\mathbf{W}^m\}_{m=1}^M$ with fixed $\mathbf{U}$ and $\mathbf{V}$: When $\mathbf{U}$ and $\mathbf{V}$ are fixed, the Eq. (3) is convex. Taking the partial derivative of (3) with regard to $\mathbf{W}^m, m = 1, 2, \ldots, M$, we can obtain:

$$\frac{\partial \mathcal{J}}{\partial \mathbf{W}^m} = \frac{4C}{n} \sum_{(\mathbf{q}_i, \mathbf{y}_i, \mathbf{y}) \in \Theta} \Big[g_i \times \mathbf{h}_i \Big] + 2\eta \mathbf{W}^m + 2\lambda \sum_{i,j=1}^{n} A_{ij} \Big[\mathbf{Z}_m^T \mathbf{Z}_m \mathbf{W}^m (\mathbf{x}_i^m - \mathbf{x}_j^m)(\mathbf{x}_i^m - \mathbf{x}_j^m)^T + \mathbf{Z}_m^T \Gamma_m (\mathbf{x}_i^m - \mathbf{x}_j^m)^T \Big]$$

where $\mathbf{h}_i = \mathbf{S}_m^T (\mathbf{U}^T \mathbf{y}_i - \mathbf{U}^T \mathbf{y})(\mathbf{x}_i^m)^T$. $\mathbf{S}_m = Col(\mathbf{U}^T \mathbf{V}^T)$, which denotes $\mathbf{S}_m$ is a column subset of $\mathbf{U}^T\mathbf{V}^T$, $m = 1, 2, \ldots, M$, i.e., $\mathbf{S}_1$ is the first r_1 column of $\mathbf{U}^T\mathbf{V}^T$, and $\mathbf{S}_2$ is the second r_2 column of $\mathbf{U}^T\mathbf{V}^T$, and so on. $\mathbf{Z}_m = Col(\mathbf{V}^T)$, and $\Gamma_m = \sum_{s=1, s \neq m}^{M} \mathbf{Z}_s \mathbf{W}^s (\mathbf{x}_i^s - \mathbf{x}_j^s)$.

Then, the optimization is repeated over all variables $\mathbf{U}$, $\mathbf{V}$, and $\{\mathbf{W}^m\}_{m=1}^M$ by fixing the previously updated parameters, until the algorithm converges.

2.4. Prediction

Following [20], we use the k nearest neighbors for multi-label prediction. For a testing sample $\mathbf{q} = [(\mathbf{q}^1)^T, \ldots, (\mathbf{q}^M)^T]^T$, we select k nearest neighbors from the training set by $\|\mathbf{U}^T\mathbf{V}^T(\mathbf{q} - \mathbf{q}_i)\|_2^2$, and then conduct voting based on weighted nearest neighbors for the prediction.

3. EXPERIMENT

3.1. Experimental Datasets and Setting

In order to verify the effectiveness of our method, MDMLM, we test it on two publicly available image datasets, scene and mirflickr. These datasets are widely used for evaluating multi-label image classification algorithms.The scene image dataset consists of 2,000 natural scene images, where each image has five possible class labels, including desert, mountains, sea, sunset and trees. The mirflickr image dataset contains 25,000 images that are representative of a generic domain and are of high quality. There are totally 24 possible labels for each image. For each image, we extract 512-d Gist features and 1024-d Gabor features as two views. The dimension r_m in $\mathbf{W}^m$ are set to 250 and 500 for Gist and Gabor, respectively.

To further evaluate MDMLM's performance, we compare it with several state-of-the-art algorithms. We first compare with LMMO-kNN [21, 20] which is the most related method

Table 1. Quantitative results by our proposed MDMLM and compared methods on the scene dataset.

Method	hamming loss ↓	ranking loss ↓	coverage ↓	one error ↓	average precision ↑
MLSPL	0.2847±0.0059	0.4792±0.0101	2.0462±0.0038	0.6927± 0.0106	0.5127±0.0095
CCA	0.3312±0.0072	0.5021±0.0048	2.0617±0.0029	0.7059± 0.0128	0.4992±0.0073
MV^{3}MR	0.2713±0.0023	0.4532±0.0204	2.0287±0.0056	0.6892± 0.0203	0.5216±0.0062
LMMO-kNN	0.2790±0.0045	0.4519±0.0195	2.0223±0.0013	0.6963± 0.0168	0.5248±0.0108
MDMLM	**0.2646**±0.0018	**0.4434**±0.0021	**2.0166**±0.0071	**0.6864**±0.0023	**0.5378**±0.0015

Table 2. Quantitative results by our proposed MDMLM and compared methods on the mirflickr dataset.

Method	hamming loss ↓	ranking loss ↓	coverage ↓	one error ↓	average precision ↑
MLSPL	0.1983±0.0007	0.3891±0.0081	14.2021±0.0812	0.6211±0.0028	0.4303±0.0061
CCA	0.2121±0.0012	0.4131±0.0072	14.3791±0.0534	0.6417±0.0052	0.4116±0.0034
MV^{3}MR	0.1761±0.0006	0.3427±0.0036	13.4618±0.0626	0.6234±0.0039	0.4331±0.0062
LMMO-kNN	0.1783±0.0004	0.3775±0.0056	13.8820±0.0716	**0.6088**±0.0014	**0.4553**±0.0017
MDMLM	**0.1722**±0.0008	**0.3088**±0.0011	**12.8112**±0.0541	0.6129±0.0038	0.4412±0.0024

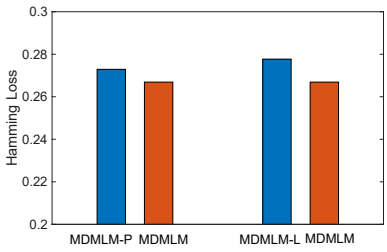


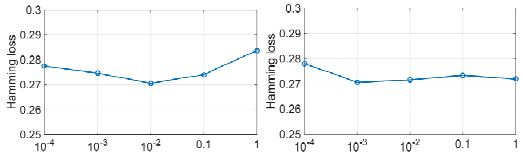
Fig. 2. Verification of the effectiveness of the components. The hamming loss on the scene dataset is used as an example.

to ours. We also compare with MV^{3}MR [24], a typical multi-view algorithm for multi-label data. In addition, we use CCA [27] and MLSPL [8] as another two baselines. For those compared algorithms not referring to multi-view learning, we concatenate all features directly for training and prediction. For a fair comparison, the parameters of all the methods are searched from $\{10^{-4}, 10^{-3}, 10^{-2}, 0.1, 1, 10\}$. The nearest neighbors k is set to 10 throughout the experiments. We use five criteria for evaluation: hamming loss, ranking loss, one error, coverage, and average precision. These criteria are commonly used for evaluating multi-label learning algorithms [28, 29, 8]. For each dataset, we randomly select 30% of data as training data, and use the rest as testing data. We repeat such experiment ten times, and report the average results.

We first test the general performance of our method MDMLM on the two image datasets. Tables 1 and 2 summarize the results of different methods in terms of all the five evaluation criteria. From these tables, our method MDMLM significantly outperforms other methods on the two datasets. MDMLM has better performance than LMMO-kNN, which indicates our idea is effective for multi-label classification.

Moreover, we also verify the effectiveness of our components. Instead of Eq. (1) simply estimating $\mathbf{P}$ via the least squares, we learn such projection matrix in a joint framework, as shown in Eq. (3). To verify its effectiveness, we conduct another experiment: in each iteration of our method, we simply estimate the projection matrix by the least squares, denoted as MDMLM-P. In addition, we also evaluate the effectiveness of preserving local structure of data. In the experiment, we set the parameter λ to zero, denoted as MDMLM-L, which means that the local structure of data is not considered. The results are shown in Fig. 2. MDMLM achieves smaller loss, showing that our component is effective.

We also study the sensitivity of C and λ in our algorithm on the scene dataset. Fig. 3 shows the results, indicating our method is not sensitive to C and λ with wide ranges.



(a) C (b) λ

Fig. 3. Sensitivity study of the parameters in terms of hamming loss on the scene dataset.

4. CONCLUSION

In this paper, we proposed a novel multi-view distance metric learning framework for multi-label image classification, named MDMLM. First, MDMLM learned a projection matrix for each view, and then concatenated these mapped features for better feature fusion. Second, MDMLM took advantage of all labels to construct the adjacency graph in order to preserving the local manifold structure of data in the low-dimensional feature space. Extensive evaluations on the scene and mirflickr datasets showed that our proposed MDMLM significantly outperformed the state-of-the-arts.

5. ACKNOWLEDGMENTS

This work was partially supported by National Natural Science Foundation of China Grant No. 61806044. It also was supported by the Open Project Program of the National Laboratory of Pattern Recognition (NLPR).

6. REFERENCES

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

[2] Z. Zha, X. S. Hua, T. Mei, J. Wang, G. Qi, and Z. Wang, "Joint multi-label multi-instance learning for image classification," in *CVPR*, 2008, pp. 1–8.

[3] S. He, C. Xu, T. Guo, C. Xu, and D. Tao, "Reinforced multi-label image classification by exploring curriculum," in *AAAI*, 2018, pp. 3183–3190.

[4] W. Gao and Z.H. Zhou, "On the consistency of multi-label learning," *Artificial Intelligence*, vol. 199, no. 3, pp. 22–44, 2013.

[5] J. Yan, C. Zhang, H. Zha, M. Gong, C. Sun, J. Huang, S. Chu, and X. Yang, "On machine learning towards predictive sales pipeline analytics," in *AAAI*, 2015, pp. 1945–1951.

[6] C. Li, F. Wei, W. Dong, X. Wang, Q. Liu, and X. Zhang, "Dynamic structure embedded online multiple-output regression for streaming data," *TPAMI*, vol. 41, no. 2, pp. 323–336, 2019.

[7] J. Yan, Y. Li, W. Liu, H. Zha, X. Yang, and S. M Chu, "Graduated consistency-regularized optimization for multi-graph matching," in *ECCV*, 2014, pp. 407–422.

[8] C. Li, F. Wei, J. Yan, X. Zhang, Q. Liu, and H. Zha, "A self-paced regularization framework for multilabel learning," *TNNLS*, vol. 29, no. 6, pp. 2660–2666, 2018.

[9] S. Xiao, J. Yan, C. Li, B. Jin, X. Wang, X. Yang, S. Chu, and H. Zha, "On modeling and predicting individual paper citation count over time," in *IJCAI*, 2016, pp. 2676–2682.

[10] A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," *IJCV*, vol. 42, no. 3, pp. 145–175, 2001.

[11] Z. Li and J. Tang, "Weakly supervised deep metric learning for community-contributed image retrieval," *TMM*, vol. 17, no. 11, pp. 1989–1999, 2015.

[12] X. Li, F. Zhao, and Y. Guo, "Multi-label image classification with a probabilistic label enhancement model," in *UAI*, 2014, pp. 430–439.

[13] J. Wang, Y. Yang, J. Mao, Z. Huang, C. Huang, and W. Xu, "Cnn-rnn: A unified framework for multi-label image classification," in *CVPR*, 2016, pp. 2285–2294.

[14] C. Li, F. Wei, J. Yan, W. Dong, Q. Liu, and H. Zha, "Self-paced multi-task learning," in *AAAI*, 2016, pp. 2175–2181.

[15] Y. Wei, X. Wei, M. Lin, J. Huang, B. Ni, J. Dong, Y. Zhao, and S. Yan, "Hcp: A flexible cnn framework for multi-label image classification," *TPAMI*, vol. 38, no. 9, pp. 1901–1907, 2016.

[16] R. Zhao, J. Li, Y. Chen, J. Liu, Y. Jiang, and X. Xue, "Regional gating neural networks for multi-label image classification," in *BMVC*, 2016, pp. 1–12.

[17] M. Zhang and Z.H. Zhou, "A review on multi-label learning algorithms," *TKDE*, vol. 26, no. 8, pp. 1819–1837, 2014.

[18] R. Jin, S. Wang, and Z.H. Zhou, "Learning a distance metric from multi-instance multi-label data," in *CVPR*, 2009, pp. 896–902.

[19] Y. Zhang and J. Schneider, "Maximum margin output coding," in *ICML*, 2012, pp. 379–386.

[20] W. Liu and I. Tsang, "Large margin metric learning for multi-label prediction," in *AAAI*, 2015, pp. 2800–2806.

[21] W. Liu, D. Xu, I. Tsang, and W. Zhang, "Metric learning for multi-output tasks," *TPAMI*, vol. PP, no. 99, pp. 1–1, 2018.

[22] H. Gouk, B. Pfahringer, and M. Cree, "Learning distance metrics for multi-label classification," in *ACML*, 2016, vol. 63, pp. 318–333.

[23] C. Xu, D. Tao, and C. Xu, "A survey on multi-view learning," *Computer Science*, 2013.

[24] Y. Luo, D. Tao, C. Xu, C. Xu, H. Liu, and Y. Wen, "Multi-view vector-valued manifold regularization for multilabel image classification," *TNNLS*, vol. 24, no. 5, pp. 709–722, 2013.

[25] M. Liu, Y. Luo, D. Tao, C. Xu, and Y. Wen, "Low-rank multi-view learning in matrix completion for multi-label image classification," in *AAAI*, 2015, pp. 2778–2784.

[26] C. Zhang, Z. Yu, Q. Hu, P. Zhu, X. Wang, and X. Liu, "Latent semantic aware multi-view multi-label classification," in *AAAI*, 2018, pp. 4414–4421.

[27] Y. Zhang and J. Schneider, "Multi-label output codes using canonical correlation analysis," in *AISTATS*, 2011, pp. 873–882.

[28] X. Wu and Z.H. Zhou, "A unified view of multi-label performance measures," *arXiv preprint arXiv:1609.00288*, 2016.

[29] C. Li, Q. Liu, J. Liu, and H. Lu, "Learning ordinal discriminative features for age estimation," in *CVPR*, 2012, pp. 2570–2577.