Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence

Multi-Kernel Multi-Label Learning with Max-Margin Concept Network

Wei Zhang[1], Xiangyang Xue[1], Jianping Fan[2], Xiaojing Huang[1], Bin Wu[1], Mingjie Liu[1]
[1]School of Computer Science, Fudan University, Shanghai, China
[2]Department of Computer Science, UNC-Charlotte, NC28223, USA
{weizh, xyxue}@fudan.edu.cn jfan@uncc.edu {hxj, wubin, mjliu}@fudan.edu.cn

Abstract

In this paper, a novel method is developed for enabling Multi-Kernel Multi-Label Learning. Inter-label dependency and similarity diversity are simultaneously leveraged in the proposed method. A concept network is constructed to capture the inter-label correlations for classifier training. Maximal margin approach is used to effectively formulate the feature-label associations and the label-label correlations. Specific kernels are learned not only for each label but also for each pair of the inter-related labels. By learning the eigenfunctions of the kernels, the similarity between a new data point and the training samples can be computed in the online mode. Our experimental results on real datasets (web pages, images, music, and bioinformatics) have demonstrated the effectiveness of our method.

1 Introduction

For many real-world applications, semantics richness requires multiple labels to sufficiently describe the data, thus one object (image, video, text, etc.) might be related with more than one semantic concepts simultaneously. For example, in the image annotation task, an image that shows a bird flying in the sky is associated with two labels (concepts) $bird$ and sky at the same time. Multi-label learning deals with the data associated with more than one concepts simultaneously and has already been applied to web page classification, text categorization, image annotation, bioinformatics etc. One strategy for multi-label learning is to deem multi-label problem with c labels as a classification problem with 2^c classes, and standard multi-class algorithms can be applied straightforward [Tsoumakas and Katakis, 2007]. The main drawbacks of this strategy include: 1) high cost of computation; 2) most classes might have no positive training data [Hariharan *et al.*, 2010]. An alternative strategy for multi-label learning is to independently decompose the task into c binary classification problems, one per label [Boutell *et al.*, 2004; Li *et al.*, 2009]; however, it would lose the correlations between labels, which is significant to the performance of multi-label classification. For example, the concepts $bird$ and sky often co-occur in the same image, while $bird$ and $office$ may seldom co-occur. To exploit the correlations between labels, many algorithms have been introduced recently, such as CMLF (Collective Multi-Label with Features) [Ghamrawi and McCallum, 2005], M^{3}N(Max-Margin Markov Network) [Tasker *et al.*, 2003], SSVM(Structural SVM) [Tsochantaridis *et al.*, 2004], SMML (Structured Max-Margin Learning) [Xue *et al.*, 2010] and CML(Correlative Multi-Label framework) [Qi *et al.*, 2007]. Another strategy is to transform multi-label learning into a ranking problem(ranking the proper labels before others for each data) [Elisseeff and Weston, 2002]. The above existing algorithms all employ the same feature extractor for different concepts and ignore the similarity diversity, which might be unsuitable for the real applications. For example, suppose that there are two images: one contains the concepts sky and $bird$, the other contains the concepts sky and $building$. These two images are similar when the concept sky is concerned; however, they are dissimilar to each other when the concept $bird$ or $building$ is concerned.

It is well-accepted that extracting more suitable features and designing more accurate similarity functions play an essential role in achieving more precise classification [Sonnenburg *et al.*, 2007]. With the proliferation of kernel-based methods such as SVM, kernel function or kernel matrix has been widely used to implement feature transformation or determine the data similarity. Many existing algorithms employ the same kernel for all the labels (concepts) and show that Gaussian kernel is powerful [Jebara, 2004]. However, the diverse data similarity cannot be characterized effectively by using one single kernel and multiple kernels are necessary [Tang *et al.*, 2009; Bach *et al.*, 2004]. To overcome the disadvantage of traditional one-kernel-fit-all setting, some algorithms learn multiple kernels for each label (concept) [Xiang *et al.*, 2010; Rakotomamonjy *et al.*, 2007]; however, the inter-label correlations are not leveraged sufficiently for achieving more effective multi-kernel learning.

In this paper, a novel method is developed for achieving Multi-Kernel Multi-Label Learning with Max-Margin Concept Network such that inter-label dependency and similarity diversity are sufficiently leveraged at the same time. The concept network is constructed for characterizing the inter-label correlations more effectively, so that we can leverage such inter-label correlations for classifier training and enhancing the discrimination power of the classifiers significantly. The site potentials encode the feature-label associations while the edge potentials actually capture the label-label correlations that are dependent on the features. A maximal margin approach is used to formulate the above site and the edge potentials. Based on the design of the potential functions in our model, we decouple our objective function label by label; nevertheless, the inter-label interactions remain to be captured, which differs in a crucial way from the state-of-art algorithms. In order to embed the label information and the inter-label (inter-concept) correlations, we learn specific kernels not only for each label but also for each pair of inter-related labels. On the other hand, those multiple kernels share the common basis which can be learned by spectral decomposition of a Gram kernel matrix. Furthermore, by learning the eigenfunctions of the kernels, the similarity between a new data point and the training samples can be computed in the online mode.

The rest of this paper is organized as follows: In Section 2 we formulate the proposed model for multi-kernel multi-label learning. We focus on the multi-kernel learning technique and model inference with eigenfunction in Section 3 and 4, respectively. Our experimental results on real datasets (web pages, images, music, and bioinformatics.) are given in Section 5. Finally, we conclude this paper in Section 6.

2 The Proposed Model

In a multi-label learning framework, multiple labels for each sample are represented as a $c-$dimensional binary vector $\mathbf{y} = [\mathbf{y}_1, \ldots, \mathbf{y}_c]$, where $\mathbf{y}_l = 1 (l = 1, \ldots, c)$ indicates that the sample belongs to the class l, and $\mathbf{y}_l = 0$ otherwise. We build a discriminative model $\theta^\top \Phi(\mathbf{x}, \mathbf{y})$ which scores the feature-label pair $(\mathbf{x}, \mathbf{y})$, and the parameter vector θ can be learned from the labeled samples $\{(\mathbf{x}^1, \mathbf{y}^1), \ldots, (\mathbf{x}^n, \mathbf{y}^n)\}$. For any test sample $\mathbf{x}$, the associated labels can be inferred by $\hat{\mathbf{y}} = arg \max_{\mathbf{y}} \theta^\top \Phi(\mathbf{x}, \mathbf{y})$.

In real world, semantic concepts usually do not appear independently but occur correlatively. A concept network is constructed to characterize the inter-label correlations more precisely and to learn the inter-related classifiers in the feature space. Each concept (label) l corresponds to one certain node (site) in the concept network. If concepts l and t are inter-related, there is an edge between the corresponding two nodes, denoted by $l \sim t$. Given n labeled training samples $\{(\mathbf{x}^1, \mathbf{y}^1), \ldots, (\mathbf{x}^n, \mathbf{y}^n)\}$, we can define empirical conditional probabilities $p(t|l) = \frac{\sum_{i=1}^n \mathbf{y}_l^i \mathbf{y}_t^i}{\sum_{i=1}^n \mathbf{y}_l^i}$ and $p(l|t) = \frac{\sum_{i=1}^n \mathbf{y}_l^i \mathbf{y}_t^i}{\sum_{i=1}^n \mathbf{y}_t^i}$, and then connect an edge between l and t if $\frac{p(t|l)p(l|t)}{p(t|l)+p(l|t)} > \mathbb{P}_0$, where $\mathbb{P}_0$ is a predefined threshold.

The concept network consists of two components: concepts (labels) and the inter-concept correlations. To capture the feature-concept associations and the inter-concept correlations in a unified framework, our model is formulated as:

$$\theta^\top \Phi(\mathbf{x}, \mathbf{y}) = \sum_{l=1}^{c} \pi_l v_l^\top \varphi_l(\mathbf{x}) + \sigma \sum_{l \sim t} \pi'_{lt} w_{lt}^\top \varphi_{lt}(\mathbf{x}) \tag{1}$$

where σ is a trade-off parameter, $\pi_l = \mathbf{1}_{(\mathbf{y}_l=1)} - \epsilon \mathbf{1}_{(\mathbf{y}_l=0)}$ and $\pi'_{lt} = \mathbf{1}_{(\mathbf{y}_l=\mathbf{y}_t=1)} - \epsilon' \mathbf{1}_{(\mathbf{y}_l \neq \mathbf{y}_t)}$. $\big(\mathbf{1}_{(\cdot)}$ is an indicator taking on value 1 if the predication is true and 0 otherwise. $0 < \epsilon, \epsilon' < 1$ are used to deal with class-imbalance by biasing toward the positive samples.$\big)$ v_l and w_{lt} are the sub-vectors of θ associated with the node (label) l and the edge (label-label pair) $l \sim t$ on the concept network, respectively. $\varphi_l(\mathbf{x})$ and $\varphi_{lt}(\mathbf{x})$ are (nonlinear) functions mapping sample features $\mathbf{x}$ to kernel spaces with respective to the node l and the edge $l \sim t$, respectively. Since there exists a gap between the similarity for observations and the similarity for semantic concepts in many applications, different concepts (labels) concern different features and it is better to learn different mapping functions for different concepts. We would employ multi-kernel technique to implement both the concept specific and the pairwise concept specific feature mappings (for detail see Section 3) such that similarity diversity can be effectively characterized.

The first part of Eq. (1) is the site potential of the concept network, which captures the association between the labels and the features, and maximizing the site potential is equivalent to maximizing the margin between sample $\mathbf{x}$ and the hyperplane for each concept in the kernel space. Meanwhile, the second part of Eq. (1) is the edge potential taking into account label-label correlations, where $\mathbf{y}_l = \mathbf{y}_t = 1$ means that semantic concepts l and t co-occur while $\mathbf{y}_l \neq \mathbf{y}_t$ indicates that one of these two concepts is present and the other is absent, so maximizing such edge potential is equivalent to maximizing the margin between the samples and the hyperplane which cuts the kernel feature space into two halves (one corresponding to $\mathbf{y}_l = \mathbf{y}_t = 1$ while the other $\mathbf{y}_l \neq \mathbf{y}_t$). By considering both site and edge potentials in a unified framework, we sufficiently leverage the associations between features and labels, and the correlations among labels and their dependence on the features. To learn the proposed model, the objective function is defined as:

$$\begin{aligned} &\min_{\theta, \xi^i} f(\theta) = \frac{1}{2} \|\theta\|^2 + \lambda \sum_{i=1}^{n} \xi^i \\ &s.t. \quad \theta^\top [\Phi(\mathbf{x}^i, \mathbf{y}^i) - \Phi(\mathbf{x}^i, \mathbf{y})] \geq \Delta(\mathbf{y}^i, \mathbf{y}) - \xi^i \\ &\forall i \in \{1, \ldots, n\}, \forall \mathbf{y} \in \{1, 0\}^c \end{aligned} \tag{2}$$

where ξ^i is a slack variable, and $\theta^\top [\Phi(\mathbf{x}^i, \mathbf{y}^i) - \Phi(\mathbf{x}^i, \mathbf{y})]$ can be viewed as the margin between the prediction and the true label. $\Delta(\mathbf{y}^i, \mathbf{y}) = \sum_{l=1}^c \mathbf{1}_{(\mathbf{y}_l^i \neq \mathbf{y}_l)}$ represents the multi-label loss scaling with the number of wrong labels in $\mathbf{y}$.There are $n \times 2^c$ constraints and the optimization problem is too complex to be solved directly. Based on the design of our model, we factor the proposed global model formulation as the sum of local models:

$$\theta^\top \Phi(\mathbf{x}, \mathbf{y}) = \sum_{l=1}^{c} \vartheta_l^\top \Psi_l(\mathbf{x}, \mathbf{y}_l, \mathbf{y}_{\mathcal{N}_l}) \tag{3}$$

and each local model with respect to concept l is as follows:

$$\vartheta_l^\top \Psi_l(\mathbf{x}, \mathbf{y}_l, \mathbf{y}_{\mathcal{N}_l}) = \pi_l v_l^\top \varphi_l(\mathbf{x}) + \sigma \sum_{t \in \mathcal{N}_l} \pi'_{lt} w_{lt}^\top \varphi_{lt}(\mathbf{x}) \tag{4}$$

where ϑ_l is the parameter sub-vector of θ corresponding to concept l, and $\mathcal{N}_l = \{t | t \sim l\}$ denotes the set of related concepts for l. $\mathbf{y}_l$ is the lth component of multi-label vector $\mathbf{y}$, and $\mathbf{y}_{\mathcal{N}_l}$ is the subvector of $\mathbf{y}$ corresponding the related concepts for l. Like [Xiang *et al.*, 2010; Sontag *et al.*, 2010], our optimization can be approximately decoupled into c interdependent subproblems. For each $l \in \{1, \ldots, c\}$,

$$\min_{\vartheta_l, \xi_l^i} f_l(\vartheta_l) = \frac{1}{2}\|\vartheta_l\|^2 + \lambda_l \sum_{i=1}^{n} \xi_l^i \tag{5}$$

$$s.t. \quad \vartheta_l^\top [\Psi_l(\mathbf{x}^i, \mathbf{y}_l^i, \mathbf{y}_{\mathcal{N}_l}^i) - \Psi_l(\mathbf{x}^i, \mathbf{y}_l, \mathbf{y}_{\mathcal{N}_l}^i)] \geq \mathbf{1}_{(\mathbf{y}_l^i \neq \mathbf{y}_l)} - \xi_l^i \quad \forall i \in \{1, \ldots, n\}, \forall \mathbf{y}_l \in \{1, 0\} \tag{6}$$

Since $\mathbf{y}_l, \mathbf{y}_l^i \in \{1, 0\}$, there are only two cases: either $\mathbf{y}_l = \mathbf{y}_l^i$ or $\mathbf{y}_l = 1 - \mathbf{y}_l^i$. If $\mathbf{y}_l = \mathbf{y}_l^i$, the constraints in (6) always hold; so, we can only focus on the case $\mathbf{y}_l = 1 - \mathbf{y}_l^i$ and the constraints in (6) can be further written as:

$$\vartheta_l^\top [\Psi_l(\mathbf{x}^i, \mathbf{y}_l^i, \mathbf{y}_{\mathcal{N}_l}^i) - \Psi_l(\mathbf{x}^i, 1 - \mathbf{y}_l^i, \mathbf{y}_{\mathcal{N}_l}^i)] \geq 1 - \xi_l^i \quad \forall i \in \{1, \ldots, n\} \tag{7}$$

$\vartheta_l^\top \Psi_l(\mathbf{x}^i, \mathbf{y}_l^i, \mathbf{y}_{\mathcal{N}_l}^i)$ is the local model score based on the observation $\mathbf{x}^i$ and the *completely true* labels, while $\vartheta_l^\top \Psi_l(\mathbf{x}^i, 1 - \mathbf{y}_l^i, \mathbf{y}_{\mathcal{N}_l}^i)$ is the local model score based on the observation $\mathbf{x}^i$ and the *almost true* labels. In the decoupled formulation, the model parameter sub-vector ϑ_l can be learned with ease. Although the model parameter sub-vectors are learned label by label, the correlations between labels are still be taken into account due to the second item in the right-side of Eq. (4). Now, there are only n constraints in the optimization problem (5) s.t. (7) for each $l \in \{1, \ldots, c\}$, which is similar to 2-class SVM. The dual of the optimization problem is as follows:

$$\max_{\alpha_l^i} \sum_{i=1}^{n} \alpha_l^i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_l^i \alpha_l^j (\Delta\Psi_l^i)^\top \Delta\Psi_l^j \quad s.t. \quad \lambda \geq \alpha_l^i \geq 0, \forall i \in \{1, \ldots, n\}; \tag{8}$$

where α_l^i denotes the dual variable, and

$$\Delta\Psi_l^i = \Psi_l(\mathbf{x}^i, \mathbf{y}_l^i, \mathbf{y}_{\mathcal{N}_l}^i) - \Psi_l(\mathbf{x}^i, 1 - \mathbf{y}_l^i, \mathbf{y}_{\mathcal{N}_l}^i) \tag{9}$$

The primal variable ϑ_l can be computed from the dual variables: $\vartheta_l = \sum_{i=1}^{n} \alpha_l^i \Delta\Psi_l^i$. According to (4) and (9), we have:

$$(\Delta\Psi_l^i)^\top \Delta\Psi_l^j = \beta_l \mathcal{K}_l(\mathbf{x}^i, \mathbf{x}^j) + \sum_{t \in \mathcal{N}_l} \beta'_{lt} \mathcal{K}_{lt}(\mathbf{x}^i, \mathbf{x}^j) \tag{10}$$

where $\mathcal{K}_l(\mathbf{x}^i, \mathbf{x}^j) = \varphi_l^\top(\mathbf{x}^i)\varphi_l(\mathbf{x}^j)$, $\mathcal{K}_{lt}(\mathbf{x}^i, \mathbf{x}^j) = \varphi_{lt}^\top(\mathbf{x}^i)\varphi_{lt}(\mathbf{x}^j)$. The coefficients β_l and β'_{lt} can easily be derived from (4) and (9). (For saving space, we do not present them here.)

3 Multi-Kernel Learning

Similarity is important to the classification performance and Gaussian (RBF) kernel actually characterizes the similarity between samples. We first define an original kernel regardless of label information as: $\mathcal{K}(\mathbf{x}^i, \mathbf{x}^j) = \varphi^\top(\mathbf{x}^i)\varphi(\mathbf{x}^j) = exp\{-\rho \mathbf{d}(\mathbf{x}^i, \mathbf{x}^j)\}$, where $\mathbf{d}(\mathbf{x}^i, \mathbf{x}^j)$ denotes the distance between $\mathbf{x}^i$ and $\mathbf{x}^j$, and ρ is a scaling parameter. $\mathcal{K}(\mathbf{x}^i, \mathbf{x}^j)$ measures the similarity between samples $\mathbf{x}^i$ and $\mathbf{x}^j$. For all samples in the training set $\{\mathbf{x}^1, ..., \mathbf{x}^n\}$, the pairwise similarities consist a Gram kernel matrix K with $K(i,j) = \mathcal{K}(\mathbf{x}^i, \mathbf{x}^j)$, which is symmetric and can be decomposed as: $K = \sum_{k=1}^{n} \eta_k \mathbf{u}_k \mathbf{u}_k^\top$, where $\mathbf{u}_k$ is the eigenvector with respective to eigenvalue η_k. Let $\mathbf{u}_k(i)$ and $\mathbf{u}_k(j)$ denote the ith and jth components of the eigenvector $\mathbf{u}_k$ respectively, we get $K(i,j) = \sum_{k=1}^{n} \eta_k \mathbf{u}_k(i)\mathbf{u}_k(j)$. It has been shown that the eigenvalue spectrum of the Gram matrix decays rapidly when the RBF kernel is employed [Williams and Seeger, 2000]. Thus, to reduce the complexity, we can also just select m dominant eigenvectors with large eigenvalues: $K \approx \sum_{k=1}^{m} \eta_k \mathbf{u}_k \mathbf{u}_k^\top$ and $K(i,j) \approx \sum_{k=1}^{m} \eta_k \mathbf{u}_k(i)\mathbf{u}_k(j), (m < n)$.

In order to incorporate the label information, we learn the concept-specific kernel matrix for the l-th label, denoted as K_l, by maximizing the similarities between data with the same label. Meanwhile, inspired by [Liu *et al.*, 2009; Sun *et al.*, 2010; Yan *et al.*, 2007], we require K_l to be in the neighborhood of the original Gram matrix K. The cost function is as follows:

$$\max_{K_l} Y_{\cdot l}^\top K_l Y_{\cdot l} - \gamma_l \|K_l - K\|_F^2 \tag{11}$$

where $Y_{\cdot l}$ is the $l-$th column of the matrix $Y \in \{0,1\}^{n \times c}$. $Y_{\cdot l}$ corresponds to the labels with respect to the $l-$th concept for all samples and the quadratic form $Y_{\cdot l}^\top K_l Y_{\cdot l}$ measures the sum of the similarities between data with the label l, the Frobenius matrix norm $\|K_l - K\|_F^2$ measures the divergence between $K_l$and K, and γ_l is the controlling parameter. Assume that the concept-specific kernel matrix K_l shares the common basis as K: $K_l = \sum_{k=1}^{m} \omega_{lk} \mathbf{u}_k \mathbf{u}_k^\top$, then the cost function (11) can be expressed as:

$$\max_{\omega_{lk}} \sum_{k=1}^{m} \omega_{lk} Y_{\cdot l}^\top \mathbf{u}_k \mathbf{u}_k^\top Y_{\cdot l} - \gamma_l \sum_{k=1}^{m} (\omega_{lk} - \eta_k)^2 \tag{12}$$

Similarly, to sufficiently leverage the correlations among the semantic concepts and their dependence on the input features, the pairwise label specific kernel matrix K_{lt} can be learned by maximizing the following cost function:

$$\max_{K_{lt}} Y_{\cdot l}^\top K_{lt} Y_{\cdot t} - \Upsilon_{lt} \|K_{lt} - K\|_F^2 \tag{13}$$

where Υ_{lt} is also the controlling parameter and the quadratic form $Y_{\cdot l}^\top K_l Y_{\cdot t}$ measures the sum of the similarities between data with the label l and t. Maximizing $Y_{\cdot l}^\top K_l Y_{\cdot t}$ means that two images should be similar in the kernel space if they are associated with two inter-related labels l and t respectively. It will enhance the discrimination power of the classifiers by learning from the samples associated with other related labels on the concept network. Again, we let K_{lt} share the common basis as K: $K_{lt} = \sum_{k=1}^{m} \zeta_{ltk} \mathbf{u}_k \mathbf{u}_k^\top$, and the cost function (13) is rewritten as:

$$\max_{\zeta_{ltk}} \sum_{k=1}^{m} \zeta_{ltk} Y_{\cdot l}^\top \mathbf{u}_k \mathbf{u}_k^\top Y_{\cdot t} - \Upsilon_{lt} \sum_{k=1}^{m} (\zeta_{ltk} - \eta_k)^2 \qquad (14)$$

Both (12) and (14) are optimization problems of quadratic functions and can be solved directly.

4 Model Inference by Eigenfunction: Online Mode

For any new image $\mathbf{x}^{new}$, the inference problem is to find the optimal label configuration $\hat{\mathbf{y}}^{new} = arg\max_{\mathbf{y}} \theta^\top \Phi(\mathbf{x}^{new}, \mathbf{y})$. The size of multi-label space is exponential to the number of classes, and it is intractable to enumerate all possible label configurations to find the best one. Therefore we employ an approximate inference technique called $Iterated\ Conditional\ Modes$ (ICM) [Winkler, 1995] due to its effectiveness. First, we initialize a multi-label configuration (e.g., determine each label by $max_{y_l} \pi_l v_l^\top \varphi_l(\mathbf{x})$ without allowing for inter-label dependency initially). Then, in each iteration, given $\mathbf{y}_{\mathcal{N}_l}$, we sequentially update $\mathbf{y}_l$ using the local model: if $\vartheta_l^\top \Psi_l(\mathbf{x}^{new}, \mathbf{y}_l = 1, \mathbf{y}_{\mathcal{N}_l})$ is larger than $\vartheta_l^\top \Psi_l(\mathbf{x}^{new}, \mathbf{y}_l = 0, \mathbf{y}_{\mathcal{N}_l})$ then $\mathbf{y}_l = 1$; otherwise $\mathbf{y}_l = 0$. Since $\vartheta_l = \sum_{i=1}^n \alpha_l^i \Delta \Psi_l^i$, the prediction rule actually uses kernels and dual variables as well. To get $\mathcal{K}_l(\mathbf{x}^{new}, \mathbf{x}^i)$ and $\mathcal{K}_{lt}(\mathbf{x}^{new}, \mathbf{x}^i)$, we first calculate: $\mathcal{K}(\mathbf{x}^{new}, \mathbf{x}^i) = exp\{-\rho \mathbf{d}(\mathbf{x}^{new}, \mathbf{x}^i)\}$. According to [Williams and Seeger, 2000], the eigenfunctions of kernel $\mathcal{K}$ satisfy:

$$\int \mathcal{K}(\mathbf{x}', \mathbf{x}) p(\mathbf{x}) \phi_k(\mathbf{x}) d\mathbf{x} = \daleth_k \phi_k(\mathbf{x}') \qquad (15)$$

where $\phi_k(.)$ is a eigenfunction and $p(\mathbf{x})$ is the probability density in the input space. $p(\mathbf{x})$ can be estimated by empirical distribution, and Eq. (15) can be approximated as:

$$\frac{1}{n} \sum_{i=1}^{n} \mathcal{K}(\mathbf{x}', \mathbf{x}^i) \phi_k(\mathbf{x}^i) = \daleth_k \phi_k(\mathbf{x}') \qquad (16)$$

Then $\frac{1}{n} K[\phi_k(\mathbf{x}^1), ..., \phi_k(\mathbf{x}^n)]^\top = \daleth_k [\phi_k(\mathbf{x}^1), ..., \phi_k(\mathbf{x}^n)]^\top$ because $K(i,j) = \mathcal{K}(\mathbf{x}^i, \mathbf{x}^j)$. We can find $\daleth_k = \frac{1}{n}\eta_k$ and $\phi_k(\mathbf{x}^i) = \mathbf{u}_k(i)$, where η_k is the eigenvalue of the Gram matrix K and $\mathbf{u}_k(i)$ is the ith component of the eigenvector $\mathbf{u}_k$. Using Eq. (16), for any new data

$$\phi_k(\mathbf{x}^{new}) = \frac{1}{\eta_k} \sum_{i=1}^{n} \mathcal{K}(\mathbf{x}^{new}, \mathbf{x}^i) \mathbf{u}_k(i) \qquad (17)$$

Thus $\mathcal{K}_l(\mathbf{x}^{new}, \mathbf{x}^i)$ and $\mathcal{K}_{lt}(\mathbf{x}^{new}, \mathbf{x}^i)$ can be computed as follows:

$$\begin{aligned} \mathcal{K}_l(\mathbf{x}^{new}, \mathbf{x}^i) &= \sum_{k=1}^{m} \omega_{lk} \phi_k(\mathbf{x}^{new}) \mathbf{u}_k(i) \\ \mathcal{K}_{lt}(\mathbf{x}^{new}, \mathbf{x}^i) &= \sum_{k=1}^{m} \zeta_{ltk} \phi_k(\mathbf{x}^{new}) \mathbf{u}_k(i) \end{aligned} \qquad (18)$$

5 Experiments

In the experiments, we compare our method $'Ours'$ with the state-of-the-art methods: 1) RML [Petterson and Caetano, 2010]; 2) ML-KNN [Zhang and Zhou, 2007]; 3) Tang's method [Tang *et al.*, 2009]; and 4) RankSVM [Elisseeff and Weston, 2002]. We consider four real applications: web page classification, image annotation, music emotion tagging, and gene categorization.

$Web\ Page\ Classification$. We first conduct the experiment on a collection of eleven datasets[1] for real Web pages linked from the domain $yahoo.com$. Each dataset contains 5000 documents (2,000 for training and 3,000 for testing), and about $15\% \sim 45\%$ of them belong to multiple categories simultaneously. Each Web page uses the "Bag-of-Words" representation [Dumais *et al.*, 1998]. The detailed description of these datasets is given in Table 1.

Datasets	dim	c	**Datasets**	dim	c
Arts	462	26	Business	438	30
Computers	681	33	Education	550	33
Entertainment	640	21	Health	612	32
Recreation	606	22	Reference	793	33
Science	743	40	Social	1047	39
Society	636	27			

Table 1: Eleven datasets of real Web pages linked from the "$yahoo.com$" domain. Each dataset contains 5,000 documents. dim denotes the dimensionality of data feature vector, and c denotes the number of classes.

$Image\ Annotation$. The experiments are conducted on two image datasets: MSRC (MicroSoft Research Cambridge) and Scene. 1) MSRC dataset contains 591 images (300 for training and 291 for testing) with 23 concepts in total. There are about 3 tags on average per image. We ignore the concepts $horse$ and $mountain$ since they have few positive samples. Thus there are totally 21 concepts. For each MSRC image, we first extract 44-dim features including RGB histogram, HSV histogram, HUE histogram, SAT histogram, mean texture response, and texture response histogram. 2) The Scene dataset [Boutell *et al.*, 2004] contains 2407 images (1211 for training and 1196 for testing) with totally 6 labels. In the LUV space, each image is divided into the 7×7 grids and the mean and variance of each band are computed. Thus, each Scene image is described by 294-dim feature vector.

$Music\ Emotion\ Tagging$. The dataset Emotion [2] used for this task consists of 593 songs (391 for training and 202 for test). Each sample is represented by a 72-dimensional feature vector. There are 6 types of emotions in total: amazed-surprised, happy-pleased, relaxing-calm, quiet-still, sad-lonely, and angry-fearful.

$Gene\ Categorization$. The final experiment is to predict the gene functional classes, which is conducted on the microarray expression dataset called Yeast[3] with 2,417 samples (

[1]http://www.kecl.ntt.co.jp/as/members/ueda/yahoo.tar.gz
[2]http://mulan.sourceforge.net/datasets.html
[3]http://mips.gsf.de/proj/yeast

Criteria	Methods	*yahoo.com* Web Pages Datasets										
		Arts	Busi.	Comp.	Educ.	Ente.	Health	Recre.	Refe.	Scie.	Social	Society
Micro-	Ours	0.364	**0.729**	**0.501**	**0.416**	**0.481**	**0.604**	0.369	**0.527**	0.363	**0.602**	0.421
F1	RML	0.365	0.703	0.452	0.252	0.253	0.563	0.326	0.457	0.241	0.147	0.241
↑	ML-KNN	0.132	0.704	0.413	0.280	0.233	0.295	0.130	0.230	0.226	0.562	0.299
	Tang's	0.231	0.706	0.393	0.259	0.368	0.533	0.226	0.435	0.192	0.544	0.312
	RankSVM	**0.389**	0.709	0.475	0.412	0.468	0.563	**0.396**	0.517	**0.369**	0.559	**0.433**
Macro-	Ours	**0.189**	**0.178**	**0.192**	**0.141**	**0.243**	**0.268**	**0.245**	**0.149**	**0.169**	**0.157**	**0.153**
F1	RML	0.165	0.149	0.083	0.097	0.154	0.187	0.176	0.080	0.115	0.096	0.113
↑	ML-KNN	0.077	0.117	0.102	0.089	0.118	0.146	0.074	0.051	0.090	0.137	0.080
	Tang's	0.095	0.106	0.104	0.080	0.158	0.211	0.124	0.087	0.068	0.085	0.089
	RankSVM	0.144	0.083	0.054	0.092	0.187	0.152	0.218	0.119	0.102	0.062	0.094
Ham-	Ours	**0.057**	**0.026**	**0.036**	**0.038**	0.055	**0.037**	**0.057**	**0.025**	**0.031**	**0.021**	**0.052**
ming	RML	0.058	0.032	0.037	0.050	0.059	0.041	**0.057**	0.027	0.051	0.101	0.096
Loss	ML-KNN	0.061	0.027	0.041	0.039	0.063	0.047	0.062	0.032	0.033	0.022	0.054
↓	Tang's	0.094	0.092	0.097	**0.038**	**0.053**	0.222	**0.057**	0.087	0.057	0.072	0.056
	RankSVM	0.063	0.027	0.042	0.048	0.062	0.042	0.064	0.034	0.038	0.027	0.060

Table 2: Experimental results on *yahoo.com* Web Pages Datasets. ↑ indicates 'the larger, the better'; ↓ indicates 'the smaller, the better'. The best performances are bolded for each evaluation criterion.

Criteria	Methods	Datasets	
		MSRC	Scene
Micro-F1	Ours	**0.556**	**0.744**
↑	RML	0.394	0.656
	ML-KNN	0.429	0.699
	Tang's	0.553	0.707
	RankSVM	0.479	0.631
Macro-F1	Ours	**0.442**	**0.751**
↑	RML	0.256	0.660
	ML-KNN	0.164	0.692
	Tang's	0.303	0.735
	RankSVM	0.200	0.638
Hamming	Ours	0.099	**0.089**
Loss	RML	0.231	0.109
↓	ML-KNN	0.094	0.099
	Tang's	**0.090**	0.130
	RankSVM	0.099	0.127

Table 3: Experimental results on MSRC and Scene datasets. ↑ indicates 'the larger, the better'; ↓ indicates 'the smaller, the better'. The best performances are bolded for each evaluation criterion.

Criteria	Methods	Datasets	
		Emotion	Yeast
Micro-F1	Ours	**0.705**	**0.665**
↑	RML	0.683	0.504
	ML-KNN	0.670	0.644
	Tang's	0.651	0.658
	RankSVM	0.619	0.651
Macro-F1	Ours	**0.695**	**0.443**
↑	RML	0.683	0.423
	ML-KNN	0.645	0.370
	Tang's	0.581	0.385
	RankSVM	0.609	0.359
Hamming	Ours	**0.195**	0.196
Loss	RML	0.241	0.204
↓	ML-KNN	0.202	0.195
	Tang's	0.240	**0.190**
	RankSVM	0.234	0.201

Table 4: Experimental results on Music-Emotion and Yeast datasets. ↑ indicates 'the larger, the better'; ↓ indicates 'the smaller, the better'. The best performances are bolded for each evaluation criterion.

1,500 for training and 917 for testing). Each sample is represented by a 103-dimensional vector. The average number of labels per gene in the training set is about 4, and the total number of labels is 14.

Table 2 shows the experimental results of our method in comparison with other related methods on *yahoo.com* Web Pages Datasets. Table 3 and 4 give the results on MSRC, Scene, Music-Emotion, and Yeast datasets. We use multi-label classification criteria Micro-F1, Macro-F1, and Hamming Loss to evaluate the performance: Micro-F1 computes the F1 measure on the predictions of different labels as a whole; Macro-F1 averages the F1 measure on the predictions of different labels; Hamming Loss calculates how many times an instance-label pair is misclassified [Zhang and Zhou, 2010; Sun *et al.*, 2010]. On each evaluation criterion, the best result is highlighted in boldface. Best parameters are chosen by tuning in experiments. In our method, the threshold $\mathbb{P}_0$ is concerned with the concept network construction and the parameter σ in Eq. (1) controls the influence of label-label correlation on multi-label learning. γ_l in Eq. (11) and Υ_{lt} in Eq. (13) are the trade-off between the common kernel and the multiple specific kernels. From the results, we find that our method performs better than other methods in most cases. Our method sufficiently leverages feature-label association, inter-label dependency, and similarity diversity at the same time, which inherits all merits of the state-of-the-art methods.

6 Conclusions

Inter-label dependency and similarity diversity are simultaneously leveraged in the proposed multi-kernel multi-label learning method. A concept network is first constructed for characterizing the inter-label correlations effectively, and the maximal margin technique effectively captures the feature-label associations and the label-label correlations. By decoupling the multi-label learning task into inter-dependant subproblems label by label, the proposed method learns multiple interrelated classifiers jointly. Specific kernels not only for each label but also for each pair of inter-related labels are learned to embed the label information and the inter-label (inter-concept) correlations. Similarity between a new data point and the training samples can be computed easily via the eigenfunctions of the kernels.

Acknowledgments

We would like to thank the anonymous reviewers for their helpful comments. This work was supported in part by the 973 Program (No.2010CB327906), the NSF of China (No.60903077 and No.60873178), the STCSM's Innovation Program(No. 10511500703), and the Shanghai Leading Academic Discipline Project (No.B114).

References

[Bach *et al.*, 2004] Francis R. Bach, Gert R. G. Lanckriet, and Michael I. Jordan. Multiple kernel learning, conic duality, and the smo algorithm. In *ICML*, 2004.

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

[Dumais *et al.*, 1998] S. T. Dumais, J. Platt, D. Heckerman, and M. Sahami. Inductive learning algorithms and representation for text categorization. In *7th ACM IKM*, 1998.

[Elisseeff and Weston, 2002] Andre Elisseeff and Jason Weston. A kernel method for multi-labelled classification. In *NIPS*, 2002.

[Ghamrawi and McCallum, 2005] Nadia Ghamrawi and Andrew McCallum. Collective multi-label classification. In *CIKM*, 2005.

[Hariharan *et al.*, 2010] Bharath Hariharan, Lihi Zelnik-Manor, S.V.N. Vishwanathan, and Manik Varma. Large scale max-margin multi-label classification with priors. In *ICML*, 2010.

[Jebara, 2004] Tony Jebara. Multi-task feature and kernel selection for svms. In *ICML*, 2004.

[Li *et al.*, 2009] Ying-Xin Li, Shuiwang Ji, Sudhir Kumar, Jieping Ye, and Zhi-Hua Zhou. Drosophila gene expression pattern annotation through multi-instance multi-label learning. In *IJCAI*, 2009.

[Liu *et al.*, 2009] Jun Liu, Jianhui Chen, Songcan Chen, and Jieping Ye. Learning the optimal neighborhood kernel for classification. In *IJCAI*, pages 1144–1149, 2009.

[Petterson and Caetano, 2010] James Petterson and Tiberio Caetano. Reverse multi-label learning. In *NIPS*, 2010.

[Qi *et al.*, 2007] Guo-Jun Qi, Xian-Sheng Hua, Yong Rui, Jinhui Tang, Tao Mei, and Hong-Jiang Zhang. Correlative multi-label video annotation. In *ACM MM*, 2007.

[Rakotomamonjy *et al.*, 2007] Alain Rakotomamonjy, Francis R. Bach, Stephane Canu, and Yves Grandvalet. More efficiency in kernel learning. In *ICML*, 2007.

[Sonnenburg *et al.*, 2007] S. Sonnenburg, G. Rtsch, C. Schfer, and B. Schlkopf. Large scale multiple kernel learning. *JMLR*, (7), 2007.

[Sontag *et al.*, 2010] David Sontag, Ofer Meshi, Tommi Jaakkola, and Amir Globerson. More data means less inference: A pseudo-max approach to structured learning. In *NIPS*, 2010.

[Sun *et al.*, 2010] Yu-Yin Sun, Yin Zhang, and Zhi-Hua Zhou. Multi-label learning with weak label. In *AAAI*, 2010.

[Tang *et al.*, 2009] Lei Tang, Jianhui Chen, and Jieping Ye. On multiple kernel learning with multiple labels. In *IJCAI*, 2009.

[Tasker *et al.*, 2003] Ben Tasker, Carlos Guestrin, and Daphne Koller. Max-margin markov network. In *NIPS*, 2003.

[Tsochantaridis *et al.*, 2004] Ioannis Tsochantaridis, Thomas Hofmann, Thorsten Joachims, and Yasemin Altun. Support vector machine learning for interdependent and structured output spaces. In *ICML*, 2004.

[Tsoumakas and Katakis, 2007] G. Tsoumakas and I. Katakis. Multi-label classification: An overview. *Int. Journal of Data Warehousing and Mining*, 3, 2007.

[Williams and Seeger, 2000] Christopher Williams and Matthias Seeger. The effect of the input density distribution on kernel-based classifiers. In *ICML*, 2000.

[Winkler, 1995] G. Winkler. Image analysis, random fields and dynamic monte carlo methods: A mathematical introduction. *Springer-Verlag, Berlin, Heidelberg*, 1995.

[Xiang *et al.*, 2010] Yu Xiang, Xiangdong Zhou, Zuotao Liu, Tat-Seng Chua, and Chong-Wah Ngo. Semantic context modeling with maximal margin conditional random fields for automatic image annotation. In *CVPR*, 2010.

[Xue *et al.*, 2010] Xiangyang Xue, Hangzai Luo, and Jianping Fan. Structured max-margin learning for multi-label image annotation. In *CIVR*, 2010.

[Yan *et al.*, 2007] Rong Yan, Jelena Tesic, and John Smith. Model-shared subspace boosting for multi-label classification. In *KDD*, pages 834–843, 2007.

[Zhang and Zhou, 2007] Min-Ling Zhang and Zhi-Hua Zhou. Ml-knn: A lazy learning approach to multi-label learning. *Pattern Recognition*, 40(7):2038–2048, 2007.

[Zhang and Zhou, 2010] Yin Zhang and Zhi-Hua Zhou. Multi-label dimensionality reduction via dependence maximization. *ACM Trans Knowledge Discovery from Data*, 2010.