# Multi-Label Classification with Label-Specific Feature Generation: A Wrapped Approach

Ze-Bang Yu, and Min-Ling Zhang

**Abstract**—Label-specific features serve as an effective strategy to learn from multi-label data, where a set of features encoding specific characteristics of each label are generated to help induce multi-label classification model. Existing approaches work by taking the two-stage strategy, where the procedure of label-specific feature generation is independent of the follow-up procedure of classification model induction. Intuitively, the performance of resulting classification model may be suboptimal due to the decoupling nature of the two-stage strategy. In this paper, a wrapped learning approach is proposed which aims to jointly perform label-specific feature generation and classification model induction. Specifically, one (kernelized) linear model is learned for each label where label-specific features are simultaneously generated within an embedded feature space via empirical loss minimization and pairwise label correlation regularization. Comparative studies over a total of sixteen benchmark data sets clearly validate the effectiveness of the wrapped strategy in exploiting label-specific features for multi-label classification.

**Index Terms**—Multi-label classification, label-specific features, label correlation, wrapped procedure

## 1 Introduction

In recent years, multi-label classification techniques have been widely used in solving real-world tasks involving objects with rich semantics [18], [20], [24], [25], [28], [30], [35], [39], [43]. Formally speaking, let $\mathcal{X} = \mathbb{R}^m$ denote the $m$-dimensional feature space and $\mathcal{Y} = \{\omega_1, \omega_2, \dots, \omega_l\}$ denote the label space consisting of $l$ class labels. The task of multi-label classification is learn a predictive function $h : \mathcal{X} \to 2^{\mathcal{Y}}$ from the training set $\mathcal{D} = \{(\boldsymbol{x}_i, Y_i) \mid 1 \le i \le n\}$, where $\boldsymbol{x}_i \in \mathcal{X}$ is a feature vector and $Y_i \subseteq \mathcal{Y}$ is the set of relevant labels associated with $\boldsymbol{x}_i$.

Generally, the desired multi-label classification model is instantiated by learning a set of discriminative functions $\boldsymbol{g} = \{g_1, g_2, \dots, g_l\}$, i.e. $h(\boldsymbol{x}) = \{\omega_j \mid g_j(\boldsymbol{x}) > 0,\ 1 \le j \le l\}$, where $g_j : \mathcal{X} \to \mathbb{R}$ determines the relevancy of $\omega_j$ w.r.t. $\boldsymbol{x}$. The most straightforward strategy is to employ the very single feature representation $\boldsymbol{x}$ in learning all the discriminative functions. However, the distinct characteristics of each class label may not be fully considered by employing identical feature representation for model induction. For instance, in multi-label text categorization, features corresponding to word terms *GDP*, *income* and *currency* are informative in discriminating economic and non-economic documents, while features corresponding to word terms *Olympics*, *movie* and *celebrity* are informative in discriminating entertainment and non-entertainment documents.

Recently, the strategy of *label-specific features* has been proposed to learning from multi-label data where the relevancy of each label is determined with tailored features of its own. Existing approaches following the label-specific features strategy usually work in a two-stage manner [13], [15], [32], [38], [44], [48], [50], where the procedure of generating label-specific features is independent of the subsequent procedure of the training classification model. Nonetheless, the performance of resulting classification model may be affected by the decoupling nature of two-stage mechanism. As an inherent building block of the multi-label learning system, it is beneficial to jointly consider the generation of label-specific features along with the induction of classification model. In this way, informative feedbacks from the classification model can be incorporated in the generation of label-specific features, and vice versa.

Recently, the strategy of label-specific features has been jointly considered with several tasks in multi-label classification such as label correlation exploitation [19] and missing label completion [8], [36]. In [19], similarity constraints over predictive outputs of a pair of class labels are exploited to help induce the label-specific features. In [36], label completion techniques are utilized to enrich the missing labeling information which guide the generation of label-specific features to be used by stand-alone classification models such as BSVM [14]. In [8], positive as well as negative label correlations are utilized to help complete missing labeling information as well as label-specific features generation. Although the task of label correlation exploitation or missing label completion has been jointly considered in previous studies on label-specific features [8], [19], [36], the task of classification model induction has been rarely jointly considered with label-specific features generation.

In light of these observations, a novel approach named Wrap, i.e. *WRAPping multi-label classification with label-specific features generation*, is proposed. Specifically, for each class label, one (kernelized) linear model is learned by simultaneously generating its label-specific features in an embedded feature space. The resulting problem is solved by alternating optimization based on empirical loss minimization and pairwise label correlation regularization. Extensive experiments over a total of sixteen benchmark data sets

- *Ze-Bang Yu and Min-Ling Zhang are with the School of Computer Science and Engineering, Southeast University, Nanjing 210096, China, and the Key Laboratory of Computer Network and Information Integration (Southeast University), Ministry of Education, China. Email: {yuzb, zhangml}@seu.edu.cn (Corresponding author: Min-Ling Zhang)*

clearly show the benefits of wrapping label-specific features generation with predictive model induction for multi-label classification.

The rest of this paper is organized as follows. Firstly, related works are briefly discussed. Secondly, technical details of the proposed WRAP approach are presented. Thirdly, comparative experimental results are reported. Finally, we conclude this paper.

## 2 RELATED WORK

Multi-label classification aims to learn a mapping function from the feature space to the *powerset* of label space. Due to the combinatorial nature of the predicted label set, most approaches focus on exploiting correlations among class labels to facilitate the induction of multi-label classification model [11], [51]. Roughly speaking, the order of label correlations can be considered in a *first-order* manner by treating each class label independently [1], [49], a *second-order* manner by exploiting pairwise interactions between class labels [2], [10], [21], and a *high-order* manner by exploiting interactions among a subset of or all class labels [17], [27], [34].

In addition to label correlations exploitation, another effective way to facilitate multi-label classification is to manipulate the feature space. Dimensionality reduction [31] or feature selection [26] over the original feature space serve as the most common strategy for feature manipulation. Furthermore, there have been other feature manipulation strategies for muti-label classification such as generating discriminative meta-level features from the original features [3], [45], aligning latent spaces for features and labels via DNNs [5], [6], [42], [46], exploiting distance metric [12], [22], [33] or multi-view representation [40], [41], [47] for multi-label data. It is worth noting that, as a common practice adopted by these feature manipulation strategies, identical feature representation is utilized in the discrimination processes of all class labels.

Different from existing multi-label feature manipulation strategies, label-specific features have been proposed as an alternative feature manipulation strategy to multi-label classification and has attracted significant research attentions in recent years. The key idea is to derive tailored feature representation for each class label rather than relying on the identical feature representation for follow-up discriminative modeling process. Specifically, the process of generating label-specific features can be conducted in two different manners, i.e. in the *transformed* feature space or in the *original* feature space.

Initial attempt towards label-specific features works in the first manner by deriving tailored feature representation in transformed feature space [50]. For each class, $k$-means clustering is employed to analyze the distributional property of positive and negative instances of each class label, where the identified clustering centers are used as the embedding bases for feature transformation. To enhance the label-specific features generation process, several customized strategies can be employed such as redundant information removal with attribute reduction [44], structured sparsity regularization over label-specific features generation process [7], [16], label-specific features expansion with nearest neighbor rules [38], linear discriminant analysis for informative label-specific features excavation [13], joint missing labeling information completion and label-specific features generation [8], [36], and multi-granularity label-specific features generation [23], etc.

On the other hand, label-specific features can be derived in another manner by retaining specific subset of features for different class labels within the original feature space. Correspondingly, feature selection for each class label can be conducted to enable retaining specific subset of original features, such as imposing sparse [15] or non-sparse [37] assumption over the selected subset of features, invoking LASSO and spectral clustering techniques [32] for feature subset selection over meta-labels, and learning to weight original features via linear regression [19] or regularized optimization [48], etc.

Generally, existing multi-label classification approaches based on label-specific features work in two independent generation-then-induction stages. In the next section, a first attempt towards wrapped label-specific features generation and multi-label classification model induction will be introduced.

## 3 THE WRAP APPROACH

Given the multi-label training set $\mathcal{D} = \{(\boldsymbol{x}_i, Y_i) \mid 1 \leq i \leq n\}$, let $\mathbf{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n]^\top \in \mathbb{R}^{n\times m}$ denote the instance matrix in the $m$-dimensional feature space and $\mathbf{Y} = [\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_n]^\top \in \{0,1\}^{n\times l}$ denote the binary label matrix in the label space with $l$ class labels. Here, we have $\boldsymbol{y}_i = [y_{i1}, y_{i2}, \ldots, y_{il}]^\top$ where $y_{ij} = +1$ if $\omega_j \in Y_i$ and $y_{ij} = 0$ otherwise.

Furthermore, WRAP employs an embedding matrix $\mathbf{V} \in \mathbb{R}^{m\times d}$ which maps the original $m$-dimensional feature space into a $d$-dimensional feature space ($d < m$) for label-specific features generation. Correspondingly, one linear model w.r.t each class label is assumed for wrapped label-specific features generation and predictive model induction:

$$g_j(\boldsymbol{x}) = \boldsymbol{u}_j^\top \cdot \mathbf{V}^\top \cdot \boldsymbol{x} + b_j \quad (1 \leq j \leq l), \tag{1}$$

Here, $\boldsymbol{u}_j \in \mathbb{R}^d$ and $b_j$ correspond to weight vector and bias of the linear model assumed for class label $\omega_j$ in the embedded feature space.

### 3.1 Linear Wrapping for Label-Specific Features

To enable label-specific features generation in conjunction with the assumed classification model, WRAP aims to optimize the following objective function:

$$\min_{\mathbf{U},\mathbf{V},\boldsymbol{b}} \ \frac{1}{2}||\mathbf{XVU} + \mathbf{1}_n\boldsymbol{b}^\top - \mathbf{Y}||_2^2 \tag{2}$$
$$+\frac{\lambda_1}{2}||\mathbf{V}||_2^2 + \lambda_3||\mathbf{U}||_1,$$

Here, we have $\mathbf{U} = [\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_l] \in \mathbb{R}^{d\times l}$, $\boldsymbol{b} = [b_1, b_2, \ldots, b_l]^\top \in \mathbb{R}^{l\times 1}$, and $\mathbf{1}_n = [1, 1, \ldots, 1]^\top \in \mathbb{R}^{n\times 1}$.

Conceptually, the first term in Eq.(2) measures the predictive loss of the learned models while the second term in Eq.(2) controls model complexity of the embedding matrix. The third term in Eq.(2) corresponds to the $\ell_1$ norm of the weight matrix $\mathbf{U}$, which aims to introduce sparsity into the weight vector of linear models. In this way, WRAP naturally fulfills the task of label-specific features generation in the embedded feature space by retaining features with non-zero weight for each class label.

To enable the exploitation of label correlations, WRAP adapts the objective function in Eq.(2) as follows:

$$\begin{aligned}\min_{\mathbf{U},\mathbf{V},\boldsymbol{b}}\ & \frac{1}{2}||\mathbf{X}\mathbf{V}\mathbf{U}+\mathbf{1}_n\boldsymbol{b}^\top-\mathbf{Y}||_2^2+\frac{\lambda_1}{2}||\mathbf{V}||_2^2 \\ & +\frac{\lambda_2}{2}\mathrm{tr}\left(\mathbf{U}\mathbf{C}\mathbf{U}^\top\right)+\lambda_3||\mathbf{U}||_1 \\ \text{s.t.}:\ & \boldsymbol{u}_j^\top\boldsymbol{u}_j=1,\quad 1\leq j\leq l,\end{aligned}\tag{3}$$

Here, $\mathrm{tr}\left(\mathbf{U}\mathbf{C}\mathbf{U}^\top\right)=\sum_{j=1}^{l}\sum_{k=1}^{l}c_{jk}\boldsymbol{u}_j^\top\boldsymbol{u}_k$ considers the correlations between any pair of linear models. Specifically, $\mathbf{C}=[c_{jk}]_{l\times l}\in\mathbb{R}^{l\times l}$ with $c_{jk}=-\sum_{i=1}^{n}y_{ij}y_{ik}$ being the negation of the number of training examples on which $\omega_j$ and $\omega_k$ co-occur as relevant labels. Therefore, the third term in Eq.(3) enforces that the weight vectors of two linear models should be similar to each other if the corresponding class labels have strong correlations. The resulting problem is solved by following the alternating optimization procedure.

**Fix $\mathbf{V}$ and $\boldsymbol{b}$, Optimize $\mathbf{U}$** The optimization problem in this step corresponds to:

$$\begin{aligned}\min_{\mathbf{U}}\ & f(\mathbf{U})+\lambda_3||\mathbf{U}||_1 \\ \text{s.t.}:\ & \boldsymbol{u}_j^\top\boldsymbol{u}_j=1,\quad 1\leq j\leq l,\end{aligned}\tag{4}$$

where $f(\mathbf{U})=\frac{1}{2}||\mathbf{X}\mathbf{V}\mathbf{U}+\mathbf{1}_n\boldsymbol{b}^\top-\mathbf{Y}||_2^2+\frac{\lambda_2}{2}\mathrm{tr}\left(\mathbf{U}\mathbf{C}\mathbf{U}^\top\right)$ is the smooth part of the convex objective function in Eq.(4). Correspondingly, the iterative procedure based on proximal gradient descent is invoked to minimize Eq.(4). It is not difficult to show that:

$$\nabla_{\mathbf{U}}f(\mathbf{U})=(\mathbf{X}\mathbf{V})^\top(\mathbf{X}\mathbf{V}\mathbf{U}+\mathbf{1}_n\boldsymbol{b}^\top-\mathbf{Y})+\lambda_2\mathbf{U}\mathbf{C},\tag{5}$$

Furthermore, $f(\mathbf{U})$ satisfies the $L$-Lipschitz condition:

$$\begin{aligned}&||\nabla_{\mathbf{U}}(f(\mathbf{U}_1)-\nabla_{\mathbf{U}}(f(\mathbf{U}_2)||_2 \\ &\leq(||\mathbf{X}\mathbf{V}||_2^2+\lambda_2||\mathbf{Y}||_2^2)\cdot||\mathbf{U}_1-\mathbf{U}_2||,\end{aligned}\tag{6}$$

with Lipschitz constant $L=||\mathbf{X}\mathbf{V}||_2^2+\lambda_2||\mathbf{Y}||_2^2$. Given the solution $\mathbf{U}^{(t)}$ at $t$-th iteration, the weight matrix $\mathbf{U}$ is updated in the next iteration as follows:

$$\mathbf{Z}^{(t)}=\mathbf{U}^{(t)}-\frac{1}{L}\nabla_{\mathbf{U}}f(\mathbf{U}^{(t)}),\tag{7}$$

$$\mathbf{S}_{ij}^{(t)}=\mathrm{sign}\left(\mathbf{Z}_{ij}^{(t)}\right)\max\left(\left|\mathbf{Z}_{ij}^{(t)}\right|-\frac{\lambda_3}{L},0\right),\tag{8}$$

$$\mathbf{U}_{ij}^{(t+1)}=\mathbf{S}_{ij}^{(t)}/||\boldsymbol{s}_j||_2,\tag{9}$$

where $\boldsymbol{s}_j=\left[\mathbf{S}_{1j}^{(t)},\ldots,\mathbf{S}_{dj}^{(t)}\right]^\top$ corresponds to the $j$-th column of $\mathbf{S}^{(t)}$.

**Fix $\mathbf{U}$, Optimize $\mathbf{V}$ and $\boldsymbol{b}$** The optimization problem in this step corresponds to:

$$\min_{\mathbf{V},\boldsymbol{b}}\ \frac{1}{2}||\mathbf{X}\mathbf{V}\mathbf{U}+\mathbf{1}_n\boldsymbol{b}^\top-\mathbf{Y}||_2^2+\frac{\lambda_1}{2}||\mathbf{V}||_2^2,\tag{10}$$

It is not difficult to show that the minimizer of Eq.(10) w.r.t. $\mathbf{V}$ should satisfy the following condition:

$$\mathbf{X}^\top\mathbf{X}\mathbf{V}\mathbf{U}\mathbf{U}^\top+\mathbf{X}^\top(\mathbf{1}_n\boldsymbol{b}^\top-\mathbf{Y})\mathbf{U}^\top+\lambda_1\mathbf{V}=\mathbf{0},\tag{11}$$

Let $\mathbf{A}=\mathbf{X}^\top\mathbf{X}$, $\mathbf{B}=\mathbf{U}\mathbf{U}^\top$and $\mathbf{C}=\mathbf{X}^\top(\mathbf{1}_n\boldsymbol{b}^\top-\mathbf{Y})\mathbf{U}^\top$. The symmetric matrix $\mathbf{A}$ can be factorized into $\mathbf{P}\mathbf{\Lambda}\mathbf{P}^\top$, where $\mathbf{P}$ is an orthonormal matrix whose columns store the eigenvectors of $\mathbf{A}$ and $\mathbf{\Lambda}$ is a diagonal matrix whose diagonal elements store the eigenvalues of $\mathbf{A}$. Similarly, $\mathbf{B}$ can be factorized into $\mathbf{Q}\mathbf{\Gamma}\mathbf{Q}^\top$ with orthonormal eigenvector matrix $\mathbf{Q}$ and diagonal eigenvalue matrix $\mathbf{\Gamma}$. Accordingly, Eq.(11) is equivalent to:

$$\mathbf{P}\mathbf{\Lambda}\mathbf{P}^\top\mathbf{V}\mathbf{Q}\mathbf{\Gamma}\mathbf{Q}^\top+\mathbf{C}+\lambda_1\mathbf{V}=\mathbf{0},\tag{12}$$

By multiplying $\mathbf{P}^\top$ ($\mathbf{Q}$) to the left (right) of both sides of Eq.(12), we can have:

$$\mathbf{\Lambda}\mathbf{P}^\top\mathbf{V}\mathbf{Q}\mathbf{\Gamma}+\mathbf{P}^\top\mathbf{C}\mathbf{Q}+\lambda_1\mathbf{P}^\top\mathbf{V}\mathbf{Q}=\mathbf{0},\tag{13}$$

Therefore, the closed-form solution for $\mathbf{V}$ corresponds to:

$$\mathbf{V}=\mathbf{P}\left((-\mathbf{P}^\top\mathbf{C}\mathbf{Q})\oslash(\mathbf{\Lambda}\mathbf{1}_m\mathbf{1}_d^\top\mathbf{\Gamma}+\lambda_1\mathbf{1}_m\mathbf{1}_d^\top)\right)\mathbf{Q}^\top,\tag{14}$$

Here, $\oslash$ represents the Hadamard division operator. In addition, the closed-form solution for $\boldsymbol{b}$ corresponds to:

$$\boldsymbol{b}=-\frac{1}{n}\cdot(\mathbf{X}\mathbf{V}\mathbf{U}-\mathbf{Y})^\top\mathbf{1}_n,\tag{15}$$

The complete procedure of WRAP in *linear* mode is summarized in Table 1. Firstly, the model parameters $\mathbf{U}$, $\mathbf{V}$ and $\boldsymbol{b}$ are randomly initialized with $\mathbf{U}^\top\mathbf{U}=\mathbf{I}$ (Steps 2 and 4). After that, an iterative procedure is invoked to optimize $\{\mathbf{U}\}$ and $\{\mathbf{V},\boldsymbol{b}\}$ in an iterative manner (Steps 5-12). Finally, the label set for unseen instance is predicted by querying the induced model (Steps 13 and 26).

### 3.2 Kernelized Wrapping for Label-Specific Features

Furthermore, a nonlinear version of the WRAP approach can be derived by introducing the kernel trick [29]. Given the kernel function $\kappa:\mathcal{X}\times\mathcal{X}\mapsto\mathbb{R}$, let $\psi:\mathbb{R}^m\mapsto\mathbb{R}^{\mathcal{H}_\kappa}$ be the induced (implicit) nonlinear mapping from the original feature space $\mathcal{X}$ to the higher-dimensional Reproducing Kernel Hilbert Space (RKHS). Accordingly, the embedding matrix is set as $\mathbf{V}=\mathbf{\Psi}\mathbf{H}$ where $\mathbf{\Psi}=[\psi(\boldsymbol{x}_1),\psi(\boldsymbol{x}_2),\ldots,\psi(\boldsymbol{x}_n)]\in\mathbb{R}^{\mathcal{H}_\kappa\times n}$ and $\mathbf{H}\in\mathbb{R}^{n\times d}$. Then, the prediction of resulting model on unseen instance $\boldsymbol{x}$ corresponds to:

$$g_j(\boldsymbol{x})=\boldsymbol{u}_j^\top\cdot\mathbf{H}^\top\cdot\varphi(\boldsymbol{x})+b_j\quad(1\leq j\leq l),\tag{16}$$

where $\varphi(\boldsymbol{x})=[\kappa(\boldsymbol{x}_1,\boldsymbol{x}),\kappa(\boldsymbol{x}_2,\boldsymbol{x}),\ldots,\kappa(\boldsymbol{x}_n,\boldsymbol{x})]^\top$.

Furthermore, the objective function of Eq.(3) can be rewritten as:

$$\begin{aligned}\min_{\mathbf{U},\mathbf{H},\boldsymbol{b}}\ & \frac{1}{2}||\mathbf{K}\mathbf{H}\mathbf{U}+\mathbf{1}_n\boldsymbol{b}^\top-\mathbf{Y}||_2^2+\frac{\lambda_1}{2}||\mathbf{H}^\top\mathbf{K}\mathbf{H}||_2^2 \\ & +\frac{\lambda_2}{2}\mathrm{tr}\left(\mathbf{U}\mathbf{C}\mathbf{U}^\top\right)+\lambda_3||\mathbf{U}||_1 \\ \text{s.t.}:\ & \boldsymbol{u}_j^\top\boldsymbol{u}_j=1,\quad 1\leq j\leq l,\end{aligned}\tag{17}$$

where $\mathbf{K}=[k_{ij}]_{n\times n}$ corresponds to the kernel matrix with $k_{ij}=\kappa(\boldsymbol{x}_i,\boldsymbol{x}_j)=\langle\psi(\boldsymbol{x}_i),\psi(\boldsymbol{x}_j)\rangle_{\mathcal{H}_\kappa}$. Similarly, the resulting problem can be solved by optimizing $\{\mathbf{U}\}$ and $\{\mathbf{H},\boldsymbol{b}\}$ in an alternative manner.

**Fix $\mathbf{H}$ and $\boldsymbol{b}$, Optimize $\mathbf{U}$** By re-writing $f(\mathbf{U})$ in Eq.(4) as $f^\kappa(\mathbf{U})=\frac{1}{2}||\mathbf{K}\mathbf{H}\mathbf{U}+\mathbf{1}_n\boldsymbol{b}^\top-\mathbf{Y}||_2^2+\frac{\lambda_2}{2}\mathrm{tr}\left(\mathbf{U}\mathbf{C}\mathbf{U}^\top\right)$ with the following gradient:

$$\begin{aligned}&\nabla_{\mathbf{U}}f^\kappa(\mathbf{U}) \\ &=(\mathbf{K}\mathbf{H})^\top(\mathbf{K}\mathbf{H}\mathbf{U}+\mathbf{1}_n\boldsymbol{b}^\top-\mathbf{Y})+\lambda_2\mathbf{U}\mathbf{C},\end{aligned}\tag{18}$$

TABLE 1
Pseudo-code of WRAP.

```
Inputs:
$\mathcal{D}$:   the multi-label training set $\{(\boldsymbol{x}_i, Y_i)\,|1 \leq i \leq n\}$ ($\mathcal{X} = \mathbb{R}^m$, $\mathcal{Y} = \{\omega_1, \omega_2, \ldots, \omega_l\}$, $\boldsymbol{x}_i \in \mathcal{X}$, $Y_i \subseteq \mathcal{Y}$)
$d$:   dimensionality of embedded feature space
$\lambda_1, \lambda_2, \lambda_3$:   regularization parameters w.r.t. the complexity, correlation and sparsity terms of the objective function
Outputs:
$Y$:   predicted label set for unseen instance $\boldsymbol{x}$
Process:
Set $\mathbf{Y} = [\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_n]^\top$ with $\boldsymbol{y}_i = [y_{i1}, y_{i2}, \ldots, y_{il}]^\top$, where $y_{ij} = +1$ if $\omega_j \in Y_i$ and $y_{ij} = 0$ otherwise;
Randomly initialize $\mathbf{U} \in \mathbb{R}^{d \times l}$ ($\mathbf{U}^\top\mathbf{U} = \mathbf{I}$), $\boldsymbol{b} \in \mathbb{R}^{l \times 1}$;
if linear mode then
  Randomly initialize $\mathbf{V} \in \mathbb{R}^{m \times d}$;
  Factorize the symmetric matrix $\mathbf{X}^\top\mathbf{X}$ into $\mathbf{P\Lambda P}^\top$ with $\mathbf{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n]^\top$;
  repeat
    $\mathbf{U}^{(0)} = \mathbf{U}$;
    $t = 0$;
    repeat
      Calculate $\nabla_{\mathbf{U}} f(\mathbf{U}^{(t)})$ w.r.t. Eq.(5); Set $L = ||\mathbf{XV}||_2^2 + \lambda_2||\mathbf{Y}||_2^2$; Update $\mathbf{U}^{(t+1)}$ w.r.t. Eqs.(7)-(9);
      $t = t + 1$;
    until convergence
    $\mathbf{U} = \mathbf{U}^{(t)}$;
    Factorize the symmetric matrix $\mathbf{UU}^\top$ into $\mathbf{Q\Gamma Q}^\top$;
    Set $\mathbf{C} = \mathbf{X}^\top(\mathbf{1}_n\boldsymbol{b}^\top - \mathbf{Y})\mathbf{U}^\top$;
    Update $\mathbf{V}$ and $\boldsymbol{b}$ according to Eq.(14) and Eq.(15) respectively;
  until convergence
  Calculate $g_j(\boldsymbol{x})$ $(1 \leq j \leq l)$ according to Eq.(1);
else
  Randomly initialize $\mathbf{H} \in \mathbb{R}^{n \times d}$;
  Set $\mathbf{K} = [k_{ij}]_{n \times n}$ with $k_{ij} = \kappa(\boldsymbol{x}_i, \boldsymbol{x}_j)$ and then factorize it into $\mathbf{R\Phi R}^\top$;
  repeat
    $\mathbf{U}^{(0)} = \mathbf{U}$;
    $t = 0$;
    repeat
      Calculate $\nabla_{\mathbf{U}} f^\kappa(\mathbf{U}^{(t)})$ w.r.t. Eq.(18); Set $L = ||\mathbf{KH}||_2^2 + \lambda_2||\mathbf{Y}||_2^2$; Update $\mathbf{U}^{(t+1)}$ w.r.t. Eqs.(7)-(9) by replacing $f(\cdot)$ with $f^\kappa(\cdot)$;
      $t = t + 1$;
    until convergence
    $\mathbf{U} = \mathbf{U}^{(t)}$;
    Factorize the symmetric matrix $\mathbf{UU}^\top$ into $\mathbf{Q\Gamma Q}^\top$;
    Set $\mathbf{C}^\kappa = \mathbf{K}^\top(\mathbf{1}_n\mathbf{b}^\top - \mathbf{Y})\mathbf{U}^\top$;
    Update $\mathbf{H}$ and $\boldsymbol{b}$ according to Eq.(23) and Eq.(24) respectively;
  until convergence
  Calculate $g_j(\boldsymbol{x})$ $(1 \leq j \leq l)$ according to Eq.(16);
end if
Return $Y = \{\omega_j \mid g_j(\boldsymbol{x}) > 0.5,\ 1 \leq j \leq l\}$.
```

Accordingly, $f^\kappa(\mathbf{U})$ also satisfies the $L$-Lipschitz condition with $L = ||\mathbf{KH}||_2^2 + \lambda_2||\mathbf{Y}||_2^2$. Given the solution $\mathbf{U}^{(t)}$ at $t$-th iteration, the weight matrix $\mathbf{U}$ is updated according to Eqs.(7)-(9) by replacing $f(\cdot)$ with $f^\kappa(\cdot)$.

**Fix U, Optimize H and $\boldsymbol{b}$** The optimization problem in this step corresponds to:

$$\min_{\mathbf{H},\boldsymbol{b}} \ \frac{1}{2}||\mathbf{KHU} + \mathbf{1}_n\boldsymbol{b}^\top - \mathbf{Y}||_2^2 + \frac{\lambda_1}{2}||\mathbf{H}^\top\mathbf{KH}||_2^2, \tag{19}$$

It is not difficult to show that the minimizer of Eq.(19) w.r.t. $\mathbf{H}$ should satisfy the following condition:

$$\mathbf{K}^\top\mathbf{KHUU}^\top + \mathbf{K}^\top(\mathbf{1}_n\boldsymbol{b}^\top - \mathbf{Y})\mathbf{U}^\top + \lambda_1\mathbf{KH} = \mathbf{0}, \tag{20}$$

Let $\mathbf{C}^\kappa = \mathbf{K}^\top(\mathbf{1}_n\mathbf{b}^\top - \mathbf{Y})\mathbf{U}^\top$. Furthermore, the kernel matrix $\mathbf{K}$ can be factorized into $\mathbf{R\Phi R}^\top$ with orthonormal eigenvector matrix $\mathbf{R}$ and diagonal eigenvalue matrix $\mathbf{\Phi}$. Accordingly, Eq.(20) is equivalent to:

$$\mathbf{R\Phi R}^\top(\mathbf{KH})\mathbf{Q\Gamma Q}^\top + \mathbf{C}^\kappa + \lambda_1(\mathbf{KH}) = \mathbf{0}, \tag{21}$$

By multiplying $\mathbf{R}^\top$ ($\mathbf{Q}$) to the left (right) of both sides of Eq.(21), we can have:

$$\mathbf{\Phi R}^\top(\mathbf{KH})\mathbf{Q\Gamma} + \mathbf{R}^\top\mathbf{C}^\kappa\mathbf{Q} + \lambda_1\mathbf{R}^\top(\mathbf{KH})\mathbf{Q} = \mathbf{0}, \tag{22}$$

Therefore, the closed-form solution for $\mathbf{H}$ corresponds to:

$$\mathbf{H} = \tag{23}$$

TABLE 2
Characteristics of the experimental data sets.

| Data set | $\|\mathcal{S}\|$ | $dim(\mathcal{S})$ | $L(\mathcal{S})$ | $LCard(\mathcal{S})$ | $LDen(\mathcal{S})$ | $DL(\mathcal{S})$ | Domain |
|---|---|---|---|---|---|---|---|
| CAL500 | 502 | 68 | 174 | 26.044 | 0.150 | 502 | music |
| emotions | 593 | 72 | 6 | 1.869 | 0.311 | 27 | music |
| birds | 645 | 260 | 19 | 1.014 | 0.053 | 133 | audio |
| medical | 978 | 1,449 | 45 | 1.245 | 0.028 | 94 | text |
| language log | 1,460 | 1,004 | 75 | 1.180 | 0.016 | 286 | text |
| enron | 1,702 | 1,001 | 53 | 3.378 | 0.064 | 753 | text |
| scene | 2,407 | 294 | 6 | 1.074 | 0.179 | 15 | images |
| yeast | 2,417 | 103 | 14 | 4.237 | 0.303 | 198 | biology |
| slashdot | 3,782 | 1,079 | 22 | 1.181 | 0.054 | 156 | text |
| corel5k | 5,000 | 499 | 374 | 3.522 | 0.009 | 3,175 | images |
| bibtex | 7,395 | 1,836 | 159 | 2.402 | 0.015 | 2,856 | text |
| corel16k | 13,766 | 500 | 153 | 2.859 | 0.019 | 4,803 | images |
| delicious | 16,105 | 500 | 983 | 19.020 | 0.019 | 15,806 | text |
| eurlex-sm | 19,348 | 5,000 | 201 | 2.213 | 0.011 | 2,504 | text |
| tmc2007 | 28,596 | 500 | 22 | 2.158 | 0.098 | 1,341 | text |
| mediamill | 43,907 | 120 | 101 | 4.376 | 0.043 | 6,555 | video |

$$\begin{aligned}&\mathbf{K}^{-1}\left(\mathbf{R}\left((-\mathbf{R}^\top\mathbf{C}^\kappa\mathbf{Q})\oslash(\mathbf{\Phi}\mathbf{1}_n\mathbf{1}_d^\top\mathbf{\Gamma}+\lambda_1\mathbf{1}_n\mathbf{1}_d^\top)\right)\mathbf{Q}^\top\right)\\&=\mathbf{R}\mathbf{\Phi}^{-1}\left((-\mathbf{R}^\top\mathbf{C}^\kappa\mathbf{Q})\oslash(\mathbf{\Phi}\mathbf{1}_n\mathbf{1}_d^\top\mathbf{\Gamma}+\lambda_1\mathbf{1}_n\mathbf{1}_d^\top)\right)\mathbf{Q}^\top,\end{aligned}$$

In addition, the closed-form solution for $\boldsymbol{b}$ corresponds to:

$$\boldsymbol{b}=-\frac{1}{n}\cdot(\mathbf{KHU}-\mathbf{Y})^\top\mathbf{1}_n, \tag{24}$$

The complete procedure of WRAP in *kernelized* mode is also summarized in Table 1. Similar to the linear mode, the model parameters $\mathbf{U}$, $\mathbf{H}$ and $\boldsymbol{b}$ are randomly initialized (Steps 2 and 15). After that, an iterative procedure is invoked to optimize $\{\mathbf{U}\}$ and $\{\mathbf{H}, \boldsymbol{b}\}$ in an iterative manner (Steps 16-23). Finally, the label set for unseen instance is predicted by querying the induced model (Steps 24 and 26).

## 4 EXPERIMENTS

### 4.1 Experimental Setup

In this paper, a total of sixteen benchmark multi-label data sets have been employed for extensive comparative studies.[1] Table 2 summarizes characteristics of each experimental data set $\mathcal{S}$, including the *number of examples* ($|\mathcal{S}|$), *number of features* ($dim(\mathcal{S})$), *number of class labels* ($L(\mathcal{S})$), *label cardinality* ($LCard(\mathcal{S})$, i.e. average number of relevant labels per example), *label density* ($LDen(\mathcal{S})$), i.e. label cardinality over $L(\mathcal{S})$), and *number of distinct label sets* ($DL(\mathcal{S})$) appearing in $\mathcal{S}$ [27], [51].

The performance of WRAP is compared against the following seven well-established multi-label classification approaches with parameter configurations suggested in respective standard implementations, where three of them work by employing the strategy of label-specific features:

- BR [1]: A classical multi-label classification approach which decomposes the original multi-label classification task into a set of binary classification tasks, one per class label. [parameter configuration: $C = 1$];
- ECC [27]: An ensemble-based learning approach for multi-label classification, which builds an ensemble of $N$ classifier chains over the class labels in random order. [parameter configuration: $N = 5$];
- RAKEL [34]: A transformation-based learning approach for multi-label classification, which transforms the original multi-label classification task into $N$ multi-class classification tasks over $k$ randomly chosen class labels. [parameter configuration: $N = l$, $k = 3$];
- ML-KNN [52]: A popular $k$NN-based learning approach for multi-label classification, which can be regarded as a degenerated counterpart of the label-specific feature strategy by utilizing only the original feature representation. [parameter configuration: $k = 10$];
- LIFT [50]: The seminal multi-label classification approach based on label-specific features which generates tailored features by conducting clustering analysis over the set of positive examples and negative examples w.r.t. each class label. [parameter configuration: $r = 0.1$];
- MLSF [32]: Another comparing multi-label classification approach based on label-specific features which generates tailored features by retaining different subset of original features for a group of class labels. [parameter configuration: $K = \lceil l/10 \rceil$, $\epsilon = 0.01$, $\alpha = 0.8$, $\gamma = 0.01$];
- LFLC [23]: The third comparing multi-label classification approach based on label-specific features which generates tailored features by analyzing local and global feature-to-label correlations. [parameter configuration: grid search for $\lambda \in \{1, 3, \ldots, 19\}$ with step-size 2, $\eta \in \{1e{-}10, \ldots, 1e{-}5\}$ with a multiple of $e$ at each step, $\beta = 10^4$].

As shown in Table 1, for the proposed WRAP approach (*linear* mode), we have the following parameter configuration in this paper: grid search for $\lambda_1, \lambda_2 \in \{0, 1, \ldots, 10\}$ with step-size 1, $\lambda_3 = 0.1$ and $d = \lfloor \alpha \min(m, l) \rfloor$ with

[1] Data sets publicly available at http://mulan.sourceforge.net/datasets-mlc.html

TABLE 3
Experimental results of the comparing approaches on the first eight data sets (↓: the smaller the better; ↑: the larger the better).

| Approach | Hamming loss ↓ | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | CAL500 | emotions | birds | medical | language log | enron | scene | yeast |
| Wrap | 0.136 ± 0.002 | 0.200 ± 0.007 | 0.045 ± 0.048 | 0.010 ± 0.002 | 0.015 ± 0.001 | 0.047 ± 0.001 | 0.110 ± 0.006 | 0.199 ± 0.005 |
| Wrap$^{\kappa}$ | 0.136 ± 0.005 | 0.184 ± 0.023 | 0.045 ± 0.006 | 0.009 ± 0.001 | 0.015 ± 0.001 | 0.045 ± 0.002 | 0.078 ± 0.003 | 0.185 ± 0.005 |
| Br | 0.165 ± 0.003 | 0.207 ± 0.021 | 0.070 ± 0.010 | 0.010 ± 0.001 | 0.019 ± 0.001 | 0.061 ± 0.002 | 0.131 ± 0.005 | 0.202 ± 0.006 |
| Br$^{\kappa}$ | 0.137 ± 0.001 | 0.176 ± 0.009 | 0.047 ± 0.005 | 0.028 ± 0.001 | 0.016 ± 0.000 | 0.051 ± 0.001 | 0.075 ± 0.003 | 0.187 ± 0.001 |
| Ecc | 0.147 ± 0.002 | 0.226 ± 0.011 | 0.050 ± 0.004 | 0.010 ± 0.001 | 0.015 ± 0.001 | 0.050 ± 0.001 | 0.122 ± 0.003 | 0.211 ± 0.008 |
| Rakel | 0.279 ± 0.009 | 0.254 ± 0.034 | 0.250 ± 0.013 | 0.034 ± 0.001 | 0.051 ± 0.003 | 0.162 ± 0.009 | 0.164 ± 0.017 | 0.291 ± 0.009 |
| Ml-knn | 0.147 ± 0.003 | 0.200 ± 0.007 | 0.051 ± 0.006 | 0.015 ± 0.002 | 0.016 ± 0.000 | 0.055 ± 0.002 | 0.086 ± 0.005 | 0.199 ± 0.008 |
| Lift | 0.139 ± 0.003 | 0.187 ± 0.010 | 0.045 ± 0.009 | 0.012 ± 0.001 | 0.015 ± 0.000 | 0.046 ± 0.002 | 0.081 ± 0.007 | 0.194 ± 0.006 |
| Lift$^{\kappa}$ | 0.227 ± 0.019 | 0.418 ± 0.034 | 0.131 ± 0.041 | 0.038 ± 0.002 | 0.029 ± 0.001 | 0.112 ± 0.008 | 0.380 ± 0.007 | 0.324 ± 0.010 |
| Mlsf | 0.138 ± 0.004 | 0.224 ± 0.021 | 0.050 ± 0.003 | 0.010 ± 0.001 | 0.015 ± 0.001 | 0.051 ± 0.001 | 0.128 ± 0.007 | 0.210 ± 0.005 |
| Lflc | 0.136 ± 0.004 | 0.197 ± 0.004 | 0.045 ± 0.003 | 0.010 ± 0.001 | 0.015 ± 0.001 | 0.046 ± 0.001 | 0.107 ± 0.005 | 0.197 ± 0.004 |
| Approach | One error ↓ | | | | | | | |
| | CAL500 | emotions | birds | medical | language log | enron | scene | yeast |
| Wrap | 0.113 ± 0.033 | 0.265 ± 0.025 | 0.660 ± 0.031 | 0.125 ± 0.025 | 0.730 ± 0.032 | 0.220 ± 0.021 | 0.272 ± 0.008 | 0.223 ± 0.008 |
| Wrap$^{\kappa}$ | 0.115 ± 0.038 | 0.224 ± 0.033 | 0.656 ± 0.040 | 0.130 ± 0.016 | 0.724 ± 0.020 | 0.203 ± 0.030 | 0.181 ± 0.008 | 0.214 ± 0.025 |
| Br | 0.119 ± 0.045 | 0.268 ± 0.024 | 0.712 ± 0.045 | 0.134 ± 0.020 | 0.750 ± 0.008 | 0.271 ± 0.015 | 0.317 ± 0.030 | 0.233 ± 0.033 |
| Br$^{\kappa}$ | 0.115 ± 0.039 | 0.228 ± 0.042 | 0.668 ± 0.034 | 0.147 ± 0.029 | 0.698 ± 0.017 | 0.247 ± 0.018 | 0.181 ± 0.018 | 0.215 ± 0.012 |
| Ecc | 0.299 ± 0.042 | 0.312 ± 0.033 | 0.712 ± 0.035 | 0.148 ± 0.027 | 0.752 ± 0.016 | 0.267 ± 0.026 | 0.338 ± 0.018 | 0.262 ± 0.023 |
| Rakel | 0.596 ± 0.019 | 0.310 ± 0.057 | 0.819 ± 0.034 | 0.582 ± 0.029 | 0.910 ± 0.032 | 0.755 ± 0.066 | 0.351 ± 0.023 | 0.348 ± 0.022 |
| Ml-knn | 0.158 ± 0.014 | 0.372 ± 0.004 | 0.837 ± 0.029 | 0.405 ± 0.023 | 0.905 ± 0.016 | 0.468 ± 0.018 | 0.317 ± 0.013 | 0.254 ± 0.023 |
| Lift | 0.137 ± 0.014 | 0.241 ± 0.030 | 0.691 ± 0.036 | 0.158 ± 0.016 | 0.729 ± 0.022 | 0.232 ± 0.023 | 0.195 ± 0.024 | 0.225 ± 0.024 |
| Lift$^{\kappa}$ | 0.476 ± 0.257 | 0.734 ± 0.050 | 0.966 ± 0.015 | 0.728 ± 0.040 | 0.986 ± 0.018 | 0.584 ± 0.198 | 0.823 ± 0.016 | 0.776 ± 0.105 |
| Mlsf | 0.128 ± 0.040 | 0.319 ± 0.021 | 0.701 ± 0.055 | 0.161 ± 0.017 | 0.755 ± 0.012 | 0.283 ± 0.028 | 0.345 ± 0.010 | 0.261 ± 0.022 |
| Lflc | 0.120 ± 0.022 | 0.277 ± 0.031 | 0.656 ± 0.037 | 0.136 ± 0.016 | 0.721 ± 0.025 | 0.231 ± 0.009 | 0.250 ± 0.023 | 0.225 ± 0.025 |
| Approach | Ranking loss ↓ | | | | | | | |
| | CAL500 | emotions | birds | medical | language log | enron | scene | yeast |
| Wrap | 0.175 ± 0.006 | 0.157 ± 0.028 | 0.088 ± 0.011 | 0.020 ± 0.006 | 0.157 ± 0.015 | 0.080 ± 0.006 | 0.103 ± 0.006 | 0.170 ± 0.004 |
| Wrap$^{\kappa}$ | 0.175 ± 0.004 | 0.140 ± 0.018 | 0.088 ± 0.011 | 0.020 ± 0.009 | 0.155 ± 0.010 | 0.074 ± 0.003 | 0.062 ± 0.004 | 0.158 ± 0.006 |
| Br | 0.180 ± 0.006 | 0.162 ± 0.014 | 0.104 ± 0.007 | 0.024 ± 0.010 | 0.113 ± 0.004 | 0.085 ± 0.004 | 0.115 ± 0.013 | 0.174 ± 0.011 |
| Br$^{\kappa}$ | 0.180 ± 0.002 | 0.139 ± 0.015 | 0.085 ± 0.008 | 0.022 ± 0.008 | 0.102 ± 0.012 | 0.080 ± 0.006 | 0.058 ± 0.007 | 0.157 ± 0.009 |
| Ecc | 0.214 ± 0.007 | 0.183 ± 0.017 | 0.097 ± 0.008 | 0.026 ± 0.004 | 0.112 ± 0.011 | 0.088 ± 0.005 | 0.138 ± 0.004 | 0.202 ± 0.009 |
| Rakel | 0.286 ± 0.006 | 0.185 ± 0.018 | 0.290 ± 0.040 | 0.544 ± 0.005 | 0.601 ± 0.035 | 0.335 ± 0.030 | 0.137 ± 0.017 | 0.246 ± 0.008 |
| Ml-knn | 0.220 ± 0.005 | 0.214 ± 0.009 | 0.125 ± 0.023 | 0.059 ± 0.012 | 0.156 ± 0.006 | 0.111 ± 0.003 | 0.105 ± 0.007 | 0.221 ± 0.007 |
| Lift | 0.185 ± 0.005 | 0.141 ± 0.015 | 0.106 ± 0.014 | 0.028 ± 0.006 | 0.133 ± 0.013 | 0.078 ± 0.005 | 0.063 ± 0.009 | 0.168 ± 0.010 |
| Lift$^{\kappa}$ | 0.672 ± 0.017 | 0.682 ± 0.035 | 0.512 ± 0.030 | 0.786 ± 0.039 | 0.846 ± 0.014 | 0.796 ± 0.009 | 0.710 ± 0.012 | 0.687 ± 0.075 |
| Mlsf | 0.204 ± 0.004 | 0.181 ± 0.014 | 0.107 ± 0.027 | 0.048 ± 0.020 | 0.118 ± 0.019 | 0.084 ± 0.005 | 0.122 ± 0.010 | 0.201 ± 0.008 |
| Lflc | 0.177 ± 0.004 | 0.164 ± 0.008 | 0.086 ± 0.010 | 0.020 ± 0.002 | 0.143 ± 0.013 | 0.080 ± 0.007 | 0.084 ± 0.006 | 0.171 ± 0.009 |
| Approach | Average precision ↑ | | | | | | | |
| | CAL500 | emotions | birds | medical | language log | enron | scene | yeast |
| Wrap | 0.520 ± 0.006 | 0.807 ± 0.019 | 0.352 ± 0.026 | 0.907 ± 0.017 | 0.341 ± 0.024 | 0.710 ± 0.012 | 0.832 ± 0.007 | 0.761 ± 0.006 |
| Wrap$^{\kappa}$ | 0.518 ± 0.010 | 0.830 ± 0.021 | 0.352 ± 0.040 | 0.903 ± 0.013 | 0.350 ± 0.011 | 0.720 ± 0.022 | 0.892 ± 0.006 | 0.781 ± 0.016 |
| Br | 0.502 ± 0.007 | 0.800 ± 0.019 | 0.314 ± 0.035 | 0.892 ± 0.020 | 0.330 ± 0.010 | 0.661 ± 0.004 | 0.807 ± 0.020 | 0.754 ± 0.020 |
| Br$^{\kappa}$ | 0.504 ± 0.008 | 0.830 ± 0.018 | 0.353 ± 0.029 | 0.889 ± 0.015 | 0.376 ± 0.010 | 0.690 ± 0.013 | 0.895 ± 0.011 | 0.779 ± 0.008 |
| Ecc | 0.451 ± 0.013 | 0.779 ± 0.019 | 0.317 ± 0.023 | 0.887 ± 0.017 | 0.336 ± 0.015 | 0.664 ± 0.012 | 0.787 ± 0.009 | 0.716 ± 0.015 |
| Rakel | 0.314 ± 0.008 | 0.778 ± 0.025 | 0.275 ± 0.055 | 0.426 ± 0.033 | 0.168 ± 0.024 | 0.558 ± 0.030 | 0.791 ± 0.022 | 0.670 ± 0.012 |
| Ml-knn | 0.445 ± 0.012 | 0.739 ± 0.019 | 0.235 ± 0.034 | 0.706 ± 0.015 | 0.200 ± 0.012 | 0.527 ± 0.009 | 0.812 ± 0.009 | 0.703 ± 0.008 |
| Lift | 0.497 ± 0.008 | 0.823 ± 0.005 | 0.325 ± 0.040 | 0.872 ± 0.014 | 0.353 ± 0.011 | 0.699 ± 0.011 | 0.885 ± 0.014 | 0.764 ± 0.014 |
| Lift$^{\kappa}$ | 0.635 ± 0.069 | 0.684 ± 0.018 | 0.165 ± 0.039 | 0.278 ± 0.037 | 0.057 ± 0.014 | 0.276 ± 0.015 | 0.452 ± 0.011 | 0.853 ± 0.060 |
| Mlsf | 0.479 ± 0.010 | 0.779 ± 0.014 | 0.326 ± 0.025 | 0.868 ± 0.020 | 0.324 ± 0.016 | 0.663 ± 0.010 | 0.793 ± 0.007 | 0.725 ± 0.015 |
| Lflc | 0.518 ± 0.011 | 0.799 ± 0.012 | 0.361 ± 0.023 | 0.899 ± 0.010 | 0.355 ± 0.022 | 0.709 ± 0.009 | 0.851 ± 0.021 | 0.762 ± 0.014 |
| Approach | Macro-averaging AUC ↑ | | | | | | | |
| | CAL500 | emotions | birds | medical | language log | enron | scene | yeast |
| Wrap | 0.555 ± 0.011 | 0.835 ± 0.029 | 0.815 ± 0.041 | 0.619 ± 0.030 | 0.540 ± 0.040 | 0.635 ± 0.014 | 0.906 ± 0.005 | 0.688 ± 0.008 |
| Wrap$^{\kappa}$ | 0.549 ± 0.016 | 0.855 ± 0.018 | 0.816 ± 0.029 | 0.621 ± 0.029 | 0.552 ± 0.014 | 0.693 ± 0.022 | 0.950 ± 0.006 | 0.727 ± 0.010 |
| Br | 0.502 ± 0.007 | 0.826 ± 0.023 | 0.718 ± 0.051 | 0.644 ± 0.052 | 0.553 ± 0.020 | 0.619 ± 0.028 | 0.886 ± 0.011 | 0.629 ± 0.013 |
| Br$^{\kappa}$ | 0.508 ± 0.012 | 0.862 ± 0.012 | 0.836 ± 0.032 | 0.655 ± 0.069 | 0.571 ± 0.010 | 0.675 ± 0.028 | 0.951 ± 0.003 | 0.713 ± 0.012 |
| Ecc | 0.497 ± 0.013 | 0.815 ± 0.013 | 0.732 ± 0.050 | 0.646 ± 0.057 | 0.544 ± 0.052 | 0.650 ± 0.019 | 0.878 ± 0.004 | 0.650 ± 0.017 |
| Rakel | 0.522 ± 0.026 | 0.800 ± 0.014 | 0.701 ± 0.050 | 0.655 ± 0.075 | 0.746 ± 0.027 | 0.813 ± 0.036 | 0.905 ± 0.010 | 0.686 ± 0.016 |
| Ml-knn | 0.691 ± 0.020 | 0.802 ± 0.010 | 0.847 ± 0.022 | 0.662 ± 0.068 | 0.703 ± 0.027 | 0.836 ± 0.032 | 0.938 ± 0.004 | 0.709 ± 0.008 |
| Lift | 0.518 ± 0.006 | 0.854 ± 0.011 | 0.775 ± 0.057 | 0.626 ± 0.028 | 0.582 ± 0.039 | 0.681 ± 0.036 | 0.946 ± 0.005 | 0.688 ± 0.012 |
| Lift$^{\kappa}$ | 0.956 ± 0.010 | 1.000 ± 0.000 | 0.947 ± 0.053 | 0.667 ± 0.044 | 0.768 ± 0.031 | 0.928 ± 0.021 | 1.000 ± 0.000 | 1.000 ± 0.000 |
| Mlsf | 0.528 ± 0.012 | 0.816 ± 0.020 | 0.660 ± 0.038 | 0.654 ± 0.012 | 0.558 ± 0.036 | 0.503 ± 0.024 | 0.892 ± 0.010 | 0.631 ± 0.008 |
| Lflc | 0.553 ± 0.010 | 0.834 ± 0.013 | 0.835 ± 0.040 | 0.652 ± 0.026 | 0.557 ± 0.004 | 0.686 ± 0.028 | 0.924 ± 0.005 | 0.684 ± 0.009 |

$\alpha = 0.9$.[2] Accordingly, the kernel version is denoted as Wrap$^{\kappa}$ with the following parameter configuration: $\lambda_1 = \lambda_2 = \lambda_3 = 0.1, \alpha = 0.9$, Gaussian kernel with grid search for $\sigma \in \{2^0, 2^1, \ldots, 2^{15}\}$ with a multiple of 2 at each step. For Br and Lift, linear kernel SVM [4] is utilized to instantiate the base learner for both approaches. Accordingly, the kernelized versions are denoted as Br$^{\kappa}$ and Lift$^{\kappa}$ respectively by utilizing Gaussian kernel SVM as their base learner.

For performance evaluation, five widely-used multi-label metrics are utilized including *hamming loss*, *one-error*, *ranking loss*, *average precision* and *macro-averaging AUC* [11], [51]. For the first three evaluation metrics, the *smaller* the metric value the better the performance. For the other two evaluation metrics, the *larger* the metric value the better the performance.

[2] As shown in Eq.(2), for the $m \times d$ embedding matrix $\mathbf{V}$ and the $d \times l$ weight vector matrix $\mathbf{U}$, the rank of $\mathbf{VU}$ is upper-bounded by $\min(m, l)$, i.e. $r(\mathbf{VU}) \leq \min(r(\mathbf{V}), r(\mathbf{U})) \leq \min(m, l)$. Correspondingly, we set $d$ to be some fraction of $\min(m, l)$ with $\alpha = 0.9$.

TABLE 4
Experimental results of the comparing approaches on the other six data sets (↓: the smaller the better; ↑: the larger the better).

| Approach | Hamming loss ↓ | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | slashdot | corel5k | bibtex | corel16k1 | delicious | eurlex-sm | tmc2007 | mediamill |
| Wrap | 0.036 ± 0.002 | 0.009 ± 0.000 | 0.012 ± 0.000 | 0.019 ± 0.001 | 0.018 ± 0.000 | 0.006 ± 0.000 | 0.060 ± 0.000 | 0.030 ± 0.000 |
| Wrap$^{\kappa}$ | 0.036 ± 0.002 | 0.009 ± 0.000 | 0.013 ± 0.000 | 0.019 ± 0.001 | 0.019 ± 0.000 | 0.010 ± 0.001 | 0.047 ± 0.002 | 0.028 ± 0.000 |
| Br | 0.048 ± 0.001 | 0.011 ± 0.000 | 0.016 ± 0.000 | 0.019 ± 0.000 | 0.018 ± 0.000 | 0.002 ± 0.000 | 0.056 ± 0.000 | 0.029 ± 0.000 |
| Br$^{\kappa}$ | 0.054 ± 0.001 | 0.009 ± 0.000 | 0.014 ± 0.000 | 0.019 ± 0.000 | 0.018 ± 0.000 | 0.007 ± 0.000 | 0.047 ± 0.001 | 0.029 ± 0.000 |
| Ecc | 0.042 ± 0.001 | 0.010 ± 0.000 | 0.013 ± 0.000 | 0.020 ± 0.000 | 0.019 ± 0.000 | 0.005 ± 0.000 | 0.057 ± 0.001 | 0.031 ± 0.000 |
| Rakel | 0.075 ± 0.002 | 0.038 ± 0.001 | 0.041 ± 0.001 | 0.169 ± 0.003 | 0.250 ± 0.002 | 0.044 ± 0.001 | 0.375 ± 0.016 | 0.271 ± 0.007 |
| Ml-knn | 0.060 ± 0.001 | 0.010 ± 0.000 | 0.015 ± 0.000 | 0.020 ± 0.000 | 0.019 ± 0.000 | 0.005 ± 0.000 | 0.067 ± 0.001 | 0.029 ± 0.000 |
| Lift | 0.038 ± 0.001 | 0.009 ± 0.000 | 0.012 ± 0.000 | 0.019 ± 0.000 | 0.018 ± 0.000 | 0.017 ± 0.003 | 0.054 ± 0.001 | 0.030 ± 0.000 |
| Lift$^{\kappa}$ | 0.127 ± 0.019 | 0.019 ± 0.000 | 0.029 ± 0.005 | 0.038 ± 0.000 | 0.031 ± 0.003 | 0.021 ± 0.001 | 0.140 ± 0.051 | 0.053 ± 0.006 |
| Mlsf | 0.039 ± 0.002 | 0.009 ± 0.000 | 0.012 ± 0.000 | 0.019 ± 0.000 | 0.018 ± 0.000 | 0.002 ± 0.000 | 0.057 ± 0.001 | 0.030 ± 0.000 |
| Lflc | 0.037 ± 0.001 | 0.009 ± 0.000 | 0.012 ± 0.000 | 0.019 ± 0.000 | 0.024 ± 0.000 | 0.352 ± 0.016 | 0.093 ± 0.001 | 0.030 ± 0.000 |
| Approach | One error ↓ | | | | | | | |
| | slashdot | corel5k | bibtex | corel16k1 | delicious | eurlex-sm | tmc2007 | mediamill |
| Wrap | 0.369 ± 0.030 | 0.621 ± 0.032 | 0.363 ± 0.012 | 0.635 ± 0.005 | 0.340 ± 0.010 | 0.098 ± 0.005 | 0.196 ± 0.005 | 0.156 ± 0.003 |
| Wrap$^{\kappa}$ | 0.373 ± 0.007 | 0.617 ± 0.017 | 0.385 ± 0.011 | 0.673 ± 0.008 | 0.406 ± 0.014 | 0.610 ± 0.078 | 0.185 ± 0.007 | 0.182 ± 0.003 |
| Br | 0.423 ± 0.011 | 0.667 ± 0.017 | 0.399 ± 0.011 | 0.738 ± 0.016 | 0.343 ± 0.008 | 0.046 ± 0.002 | 0.176 ± 0.005 | 0.162 ± 0.006 |
| Br$^{\kappa}$ | 0.371 ± 0.020 | 0.653 ± 0.025 | 0.354 ± 0.007 | 0.712 ± 0.011 | 0.326 ± 0.009 | 0.110 ± 0.007 | 0.128 ± 0.005 | 0.162 ± 0.005 |
| Ecc | 0.436 ± 0.018 | 0.757 ± 0.014 | 0.402 ± 0.015 | 0.693 ± 0.003 | 0.467 ± 0.010 | 0.044 ± 0.003 | 0.178 ± 0.005 | 0.176 ± 0.005 |
| Rakel | 0.706 ± 0.029 | 0.866 ± 0.013 | 0.785 ± 0.008 | 0.961 ± 0.007 | 0.941 ± 0.024 | 0.582 ± 0.007 | 0.635 ± 0.078 | 0.700 ± 0.032 |
| Ml-knn | 0.696 ± 0.018 | 0.791 ± 0.009 | 0.837 ± 0.006 | 0.805 ± 0.007 | 0.573 ± 0.012 | 0.334 ± 0.006 | 0.334 ± 0.007 | 0.182 ± 0.004 |
| Lift | 0.383 ± 0.012 | 0.675 ± 0.017 | 0.381 ± 0.013 | 0.668 ± 0.007 | 0.324 ± 0.008 | 0.903 ± 0.068 | 0.160 ± 0.002 | 0.164 ± 0.003 |
| Lift$^{\kappa}$ | 0.956 ± 0.009 | 0.984 ± 0.007 | 0.971 ± 0.002 | 0.930 ± 0.011 | 0.936 ± 0.021 | 0.932 ± 0.086 | 0.629 ± 0.126 | 0.899 ± 0.043 |
| Mlsf | 0.414 ± 0.021 | 0.667 ± 0.010 | 0.402 ± 0.013 | 0.743 ± 0.019 | 0.358 ± 0.009 | 0.049 ± 0.002 | 0.179 ± 0.002 | 0.193 ± 0.008 |
| Lflc | 0.366 ± 0.011 | 0.631 ± 0.018 | 0.351 ± 0.007 | 0.645 ± 0.010 | 0.343 ± 0.008 | 0.866 ± 0.028 | 0.207 ± 0.004 | 0.158 ± 0.004 |
| Approach | Ranking loss ↓ | | | | | | | |
| | slashdot | corel5k | bibtex | corel16k1 | delicious | eurlex-sm | tmc2007 | mediamill |
| Wrap | 0.097 ± 0.009 | 0.142 ± 0.004 | 0.102 ± 0.003 | 0.145 ± 0.001 | 0.118 ± 0.002 | 0.013 ± 0.001 | 0.043 ± 0.001 | 0.048 ± 0.001 |
| Wrap$^{\kappa}$ | 0.105 ± 0.004 | 0.143 ± 0.002 | 0.073 ± 0.004 | 0.187 ± 0.009 | 0.135 ± 0.001 | 0.143 ± 0.020 | 0.066 ± 0.004 | 0.048 ± 0.001 |
| Br | 0.098 ± 0.004 | 0.123 ± 0.004 | 0.086 ± 0.003 | 0.162 ± 0.002 | 0.121 ± 0.001 | 0.009 ± 0.001 | 0.043 ± 0.001 | 0.036 ± 0.000 |
| Br$^{\kappa}$ | 0.087 ± 0.007 | 0.122 ± 0.004 | 0.074 ± 0.003 | 0.154 ± 0.002 | 0.109 ± 0.001 | 0.014 ± 0.001 | 0.029 ± 0.001 | 0.036 ± 0.001 |
| Ecc | 0.107 ± 0.005 | 0.148 ± 0.003 | 0.088 ± 0.002 | 0.180 ± 0.002 | 0.141 ± 0.002 | 0.007 ± 0.001 | 0.043 ± 0.000 | 0.092 ± 0.002 |
| Rakel | 0.681 ± 0.028 | 0.537 ± 0.012 | 0.743 ± 0.010 | 0.424 ± 0.015 | 0.488 ± 0.008 | 0.125 ± 0.007 | 0.438 ± 0.018 | 0.222 ± 0.014 |
| Ml-knn | 0.187 ± 0.011 | 0.149 ± 0.004 | 0.252 ± 0.006 | 0.196 ± 0.001 | 0.160 ± 0.002 | 0.013 ± 0.000 | 0.082 ± 0.002 | 0.046 ± 0.000 |
| Lift | 0.093 ± 0.006 | 0.122 ± 0.002 | 0.075 ± 0.004 | 0.158 ± 0.004 | 0.129 ± 0.001 | 0.050 ± 0.001 | 0.039 ± 0.001 | 0.043 ± 0.000 |
| Lift$^{\kappa}$ | 0.933 ± 0.021 | 0.924 ± 0.011 | 0.976 ± 0.024 | 0.840 ± 0.009 | 0.966 ± 0.022 | 0.952 ± 0.030 | 0.772 ± 0.116 | 0.621 ± 0.031 |
| Mlsf | 0.094 ± 0.006 | 0.143 ± 0.004 | 0.092 ± 0.015 | 0.183 ± 0.006 | 0.127 ± 0.002 | 0.008 ± 0.001 | 0.038 ± 0.001 | 0.054 ± 0.005 |
| Lflc | 0.106 ± 0.008 | 0.162 ± 0.008 | 0.098 ± 0.003 | 0.171 ± 0.004 | 0.169 ± 0.002 | 0.308 ± 0.017 | 0.073 ± 0.001 | 0.047 ± 0.001 |
| Approach | Average precision ↑ | | | | | | | |
| | slashdot | corel5k | bibtex | corel16k1 | delicious | eurlex-sm | tmc2007 | mediamill |
| Wrap | 0.715 ± 0.025 | 0.324 ± 0.008 | 0.578 ± 0.011 | 0.354 ± 0.002 | 0.367 ± 0.006 | 0.894 ± 0.003 | 0.831 ± 0.002 | 0.704 ± 0.003 |
| Wrap$^{\kappa}$ | 0.710 ± 0.011 | 0.325 ± 0.004 | 0.571 ± 0.011 | 0.314 ± 0.006 | 0.306 ± 0.005 | 0.383 ± 0.059 | 0.827 ± 0.003 | 0.708 ± 0.002 |
| Br | 0.673 ± 0.004 | 0.280 ± 0.009 | 0.539 ± 0.008 | 0.277 ± 0.008 | 0.329 ± 0.004 | 0.946 ± 0.006 | 0.839 ± 0.003 | 0.725 ± 0.003 |
| Br$^{\kappa}$ | 0.714 ± 0.016 | 0.290 ± 0.009 | 0.585 ± 0.005 | 0.299 ± 0.005 | 0.375 ± 0.002 | 0.874 ± 0.005 | 0.885 ± 0.003 | 0.725 ± 0.004 |
| Ecc | 0.663 ± 0.009 | 0.252 ± 0.006 | 0.538 ± 0.011 | 0.296 ± 0.005 | 0.299 ± 0.005 | 0.951 ± 0.004 | 0.839 ± 0.003 | 0.640 ± 0.003 |
| Rakel | 0.415 ± 0.028 | 0.257 ± 0.005 | 0.232 ± 0.007 | 0.201 ± 0.008 | 0.228 ± 0.006 | 0.792 ± 0.009 | 0.493 ± 0.042 | 0.264 ± 0.023 |
| Ml-knn | 0.456 ± 0.020 | 0.187 ± 0.006 | 0.193 ± 0.002 | 0.220 ± 0.006 | 0.199 ± 0.001 | 0.719 ± 0.004 | 0.706 ± 0.003 | 0.667 ± 0.002 |
| Lift | 0.692 ± 0.011 | 0.290 ± 0.006 | 0.566 ± 0.010 | 0.324 ± 0.004 | 0.348 ± 0.003 | 0.365 ± 0.039 | 0.852 ± 0.002 | 0.728 ± 0.003 |
| Lift$^{\kappa}$ | 0.183 ± 0.010 | 0.087 ± 0.006 | 0.079 ± 0.024 | 0.183 ± 0.001 | 0.088 ± 0.011 | 0.092 ± 0.030 | 0.503 ± 0.065 | 0.415 ± 0.046 |
| Mlsf | 0.683 ± 0.017 | 0.267 ± 0.009 | 0.538 ± 0.014 | 0.273 ± 0.004 | 0.331 ± 0.006 | 0.944 ± 0.002 | 0.844 ± 0.003 | 0.714 ± 0.005 |
| Lflc | 0.713 ± 0.006 | 0.312 ± 0.010 | 0.588 ± 0.004 | 0.340 ± 0.006 | 0.362 ± 0.002 | 0.146 ± 0.023 | 0.796 ± 0.002 | 0.698 ± 0.003 |
| Approach | Macro-averaging AUC ↑ | | | | | | | |
| | slashdot | corel5k | bibtex | corel16k1 | delicious | eurlex-sm | tmc2007 | mediamill |
| Wrap | 0.703 ± 0.033 | 0.581 ± 0.010 | 0.870 ± 0.004 | 0.744 ± 0.008 | 0.755 ± 0.004 | 0.888 ± 0.016 | 0.927 ± 0.003 | 0.851 ± 0.003 |
| Wrap$^{\kappa}$ | 0.732 ± 0.014 | 0.683 ± 0.008 | 0.916 ± 0.002 | 0.639 ± 0.022 | 0.612 ± 0.013 | 0.697 ± 0.035 | 0.876 ± 0.044 | 0.758 ± 0.022 |
| Br | 0.702 ± 0.049 | 0.524 ± 0.015 | 0.874 ± 0.003 | 0.673 ± 0.003 | 0.738 ± 0.003 | 0.901 ± 0.010 | 0.913 ± 0.002 | 0.839 ± 0.004 |
| Br$^{\kappa}$ | 0.697 ± 0.020 | 0.527 ± 0.009 | 0.895 ± 0.004 | 0.683 ± 0.004 | 0.734 ± 0.007 | 0.897 ± 0.016 | 0.947 ± 0.001 | 0.839 ± 0.007 |
| Ecc | 0.674 ± 0.036 | 0.531 ± 0.018 | 0.870 ± 0.004 | 0.658 ± 0.005 | 0.702 ± 0.002 | 0.906 ± 0.019 | 0.911 ± 0.003 | 0.776 ± 0.004 |
| Rakel | 0.806 ± 0.040 | 0.804 ± 0.007 | 0.976 ± 0.001 | 0.915 ± 0.001 | 0.845 ± 0.003 | 0.916 ± 0.016 | 0.895 ± 0.013 | 0.712 ± 0.012 |
| Ml-knn | 0.809 ± 0.020 | 0.790 ± 0.023 | 0.938 ± 0.002 | 0.922 ± 0.001 | 0.918 ± 0.002 | 0.919 ± 0.012 | 0.919 ± 0.002 | 0.933 ± 0.000 |
| Lift | 0.732 ± 0.006 | 0.595 ± 0.009 | 0.909 ± 0.004 | 0.689 ± 0.006 | 0.700 ± 0.004 | 0.824 ± 0.027 | 0.921 ± 0.002 | 0.810 ± 0.007 |
| Lift$^{\kappa}$ | 0.845 ± 0.025 | 0.822 ± 0.006 | 1.000 ± 0.000 | 1.000 ± 0.000 | 1.000 ± 0.000 | 0.916 ± 0.019 | 1.000 ± 0.000 | 1.000 ± 0.000 |
| Mlsf | 0.685 ± 0.010 | 0.520 ± 0.004 | 0.873 ± 0.006 | 0.669 ± 0.007 | 0.730 ± 0.004 | 0.978 ± 0.004 | 0.928 ± 0.002 | 0.842 ± 0.003 |
| Lflc | 0.715 ± 0.053 | 0.568 ± 0.012 | 0.875 ± 0.005 | 0.726 ± 0.004 | 0.757 ± 0.003 | 0.549 ± 0.025 | 0.929 ± 0.004 | 0.832 ± 0.007 |

## 4.2 Experimental Results

On each data set, five-fold cross-validation is performed where the mean metric value as well as standard deviation on each evaluation metric is recorded for the comparing approaches. Correspondingly, Tables 3 and 4 report the detailed experimental results over all benchmark data sets.

Furthermore, we employ the widely-used *Friedman test* [9] to statistically analyze the relative performance among the comparing approaches. Let $k$, $T$ and $r_i^j$ denote the number of comparing approaches, the number of data sets and the rank of the $j$-th algorithm on the $i$-th data set respectively. In case of ties, the mean ranks are shared among the corresponding algorithms. Accordingly, let $R_j = \frac{1}{T}\sum_{i=1}^{T} r_i^j$ denote the average rank for the $j$-th algorithm over all data sets. Then, the following Friedman statistic $F_F$ will be distributed according to $F$-distribution under the null hypothesis of all algorithms having "equal" performance, with $k-1$ numerator degrees of freedom and $(k-1)(T-1)$ denominator degrees of freedom:

TABLE 5
Summary of the Friedman statistics $F_F$ in terms of each evaluation metric and the critical value at 0.05 significance level for Wrap and Wrap$^\kappa$ (# comparing approaches $k = 8$, # data sets $T = 16$).

| Evaluation metric | $F_F$ for Wrap | $F_F$ for Wrap$^\kappa$ | Critical value |
|---|---|---|---|
| *Hamming loss* | 11.8375 | 28.4366 | |
| *One-error* | 27.0767 | 45.0447 | |
| *Ranking loss* | 16.7699 | 31.9451 | 2.0980 |
| *Average precision* | 25.2445 | 21.7841 | |
| *Macro-averaging AUC* | 7.8070 | 16.4233 | |

$$F_F = \frac{(T-1)\chi_F^2}{T(k-1)-\chi_F^2}, \text{ where}$$
$$\chi_F^2 = \frac{12T}{k(k+1)}\left[\sum_{j=1}^{n} R_j^2 - \frac{k(k+1)^2}{4}\right], \tag{25}$$

Table 5 summarizes the Friedman statistics $F_F$ in terms of each evaluation metric and the corresponding critical value at 0.05 significance level for Wrap and Wrap$^\kappa$ (# learning approaches $k = 8$, # data sets $T = 16$).[3] As shown in Table 5, the null hypothesis of equal performance among the comparing approaches is clearly rejected for all evaluation metrics.

To show whether Wrap achieves significantly different performance against the other comparing approaches, we employ *Holm's procedure* [9] as the post-hoc test by treating Wrap as the control approach. The same procedure also applies to Wrap$^\kappa$ by treating it as the control approach. Without loss of generality, Wrap is denoted as the first comparing approach $\mathcal{A}_1$. For the other $k-1$ comparing approaches $\mathcal{A}_j$ ($2 \leq j \leq k$), the one having the ($j$-1)-th largest average rank over all data sets is denoted as $\mathcal{A}_j$. Then, we have the test statistic for comparing $\mathcal{A}_1$ (i.e. Wrap) and $\mathcal{A}_j$ as follows:

$$z_j = (R_1 - R_j) \bigg/ \sqrt{\frac{k(k+1)}{6T}} \quad (2 \leq j \leq k), \tag{26}$$

Accordingly, we use $p_j$ to denote the $p$-value of $z_j$ under normal distribution. At significance level $\alpha$, the Holm's procedure works by sequentially checking whether $p_j < \alpha/(k-j+1)$ holds in ascending order of $j$. Specifically, the Holm's procedure terminates at $j^*$ with $j^*$ being the first $j$ such that $p_j < \alpha/(k-j+1)$ does not hold.[4] Accordingly, Wrap is deemed to have significantly different performance against $\mathcal{A}_j$ with $j \in \{2, \ldots, j^*-1\}$.

Tables 6 and 7 report the statistics by taking Holm's procedure as the post-hoc test at 0.05 significance level, where Wrap and Wrap$^\kappa$ are treated as the control approach respectively. Accordingly, we can have the following observations based on the reported experimental results:

TABLE 6
Comparison of Wrap (control approach) against other comparing approaches with *Holm's procedure* as the post-hoc test in terms of each evaluation metric (significance level $\alpha = 0.05$, # comparing approaches $k = 8$).

| *Hamming loss* | | | | |
|---|---|---|---|---|
| $j$ | approach | $z_j$ | $p_j$ | $\alpha/(k-j+1)$ |
| 2 | Rakel | -5.449 | 5.073e-8 | 0.007 |
| 3 | Ml-knn | -2.346 | 1.900e-2 | 0.008 |
| 4 | Ecc | -2.309 | 2.092e-2 | 0.010 |
| 5 | Br | -2.093 | 3.636e-2 | 0.013 |
| 6 | Mlsf | -0.794 | 4.273e-1 | 0.017 |
| 7 | Lflc | -0.361 | 7.182e-1 | 0.025 |
| 8 | Lift | 0.361 | 1.000e0 | 0.050 |

| *One-error* | | | | |
|---|---|---|---|---|
| $j$ | approach | $z_j$ | $p_j$ | $\alpha/(k-j+1)$ |
| 2 | Rakel | -6.423 | 1.336e-10 | 0.007 |
| 3 | Ml-knn | -5.377 | 7.592e-8 | 0.008 |
| 4 | Mlsf | -3.789 | 1.513e-4 | 0.010 |
| 5 | Ecc | -3.500 | 4.649e-4 | 0.013 |
| 6 | Br | -1.876 | 6.060e-2 | 0.017 |
| 7 | Lift | -1.335 | 1.818e-1 | 0.025 |
| 8 | Lflc | -0.794 | 4.273e-1 | 0.050 |

| *Ranking loss* | | | | |
|---|---|---|---|---|
| $j$ | approach | $z_j$ | $p_j$ | $\alpha/(k-j+1)$ |
| 2 | Rakel | -5.521 | 3.372e-8 | 0.007 |
| 3 | Ml-knn | -3.753 | 1.749e-4 | 0.008 |
| 4 | Ecc | -1.985 | 4.718e-2 | 0.010 |
| 5 | Mlsf | -1.407 | 1.593e-1 | 0.013 |
| 6 | Lflc | -1.335 | 1.818e-1 | 0.017 |
| 7 | Br | -0.108 | 9.138e-1 | 0.025 |
| 8 | Lift | 0.541 | 1.000e0 | 0.050 |

| *Average precision* | | | | |
|---|---|---|---|---|
| $j$ | approach | $z_j$ | $p_j$ | $\alpha/(k-j+1)$ |
| 2 | Rakel | -5.918 | 3.262e-9 | 0.007 |
| 3 | Ml-knn | -5.629 | 1.811e-8 | 0.008 |
| 4 | Ecc | -3.428 | 6.080e-4 | 0.010 |
| 5 | Mlsf | -2.815 | 4.884e-3 | 0.013 |
| 6 | Br | -2.201 | 2.772e-2 | 0.017 |
| 7 | Lift | -0.722 | 4.705e-1 | 0.025 |
| 8 | Lflc | -0.650 | 5.160e-1 | 0.050 |

| *Macro-averaging AUC* | | | | |
|---|---|---|---|---|
| $j$ | approach | $z_j$ | $p_j$ | $\alpha/(k-j+1)$ |
| 2 | Ecc | -2.490 | 1.278e-2 | 0.007 |
| 3 | Br | -1.804 | 7.120e-2 | 0.008 |
| 4 | Mlsf | -1.083 | 2.790e-1 | 0.010 |
| 5 | Lift | 0.180 | 1.000e0 | 0.013 |
| 6 | Lflc | 0.577 | 1.000e0 | 0.017 |
| 7 | Rakel | 0.650 | 1.000e0 | 0.025 |
| 8 | Ml-knn | 2.815 | 1.000e0 | 0.050 |

- Across all data sets, Wrap and Wrap$^\kappa$ achieve the best or second best performance in 93.8%, 68.7%, 56.3% and 62.5% cases in terms of *hamming loss*, *one-error*, *ranking loss* and *average precision* respectively. Although Lift$^\kappa$ achieves the best performance in terms of *macro-averaging AUC*, its performance on the other evaluation metrics are less competitive.
- Across all data sets, Wrap achieves a lower average rank than the comparing approaches in 29 out of 35 cases (7 comparing approaches × 5 evaluation metrics). Furthermore, Wrap$^\kappa$ achieves a lower average

[3] For fairer comparison, Br and Lift are compared against Wrap while Br$^\kappa$ and Lift$^\kappa$ are compared against Wrap$^\kappa$. The other approaches Ecc, Rakel, Ml-knn, Mlsf and Lflc are compared against both Wrap and Wrap$^\kappa$.

[4] If $p_j < \alpha/(k-j+1)$ holds for all $j$, $j^*$ is set to be $k+1$.

TABLE 7
Comparison of Wrap$^\kappa$ (control approach) against other comparing approaches with *Holm's procedure* as the post-hoc test in terms of each evaluation metric (significance level $\alpha = 0.05$, # comparing approaches $k = 8$).

| | | *Hamming loss* | | |
|---|---|---|---|---|
| $j$ | approach | $z_j$ | $p_j$ | $\alpha/(k-j+1)$ |
| 2 | Rakel | -6.315 | 2.706e-10 | 0.007 |
| 3 | Lift$^\kappa$ | -6.026 | 1.680e-9 | 0.008 |
| 4 | Ml-knn | -3.067 | 2.161e-3 | 0.010 |
| 5 | Ecc | -2.634 | 8.435e-3 | 0.013 |
| 6 | Mlsf | -1.516 | 1.296e-1 | 0.017 |
| 7 | Lflc | -1.443 | 1.489e-1 | 0.025 |
| 8 | Br$^\kappa$ | -1.227 | 2.199e-1 | 0.050 |

| | | *One-error* | | |
|---|---|---|---|---|
| $j$ | approach | $z_j$ | $p_j$ | $\alpha/(k-j+1)$ |
| 2 | Lift$^\kappa$ | -6.062 | 1.343e-9 | 0.007 |
| 3 | Rakel | -5.124 | 2.992e-7 | 0.008 |
| 4 | Ml-knn | -3.717 | 2.019e-4 | 0.010 |
| 5 | Mlsf | -2.418 | 1.562e-2 | 0.013 |
| 6 | Ecc | -2.057 | 3.970e-2 | 0.017 |
| 7 | Lflc | -0.036 | 9.712e-1 | 0.025 |
| 8 | Br$^\kappa$ | 0.289 | 1.000e0 | 0.050 |

| | | *Ranking loss* | | |
|---|---|---|---|---|
| $j$ | approach | $z_j$ | $p_j$ | $\alpha/(k-j+1)$ |
| 2 | Lift$^\kappa$ | -5.918 | 3.262e-9 | 0.007 |
| 3 | Rakel | -4.475 | 7.660e-6 | 0.008 |
| 4 | Ml-knn | -2.887 | 3.892e-3 | 0.010 |
| 5 | Ecc | -1.443 | 1.489e-1 | 0.013 |
| 6 | Mlsf | -0.830 | 4.066e-1 | 0.017 |
| 7 | Lflc | -0.577 | 5.637e-1 | 0.025 |
| 8 | Br$^\kappa$ | 1.480 | 1.000e0 | 0.050 |

| | | *Average precision* | | |
|---|---|---|---|---|
| $j$ | approach | $z_j$ | $p_j$ | $\alpha/(k-j+1)$ |
| 2 | Lift$^\kappa$ | -5.052 | 4.376e-7 | 0.007 |
| 3 | Rakel | -4.475 | 7.660e-6 | 0.008 |
| 4 | Ml-knn | -4.186 | 2.842e-5 | 0.010 |
| 5 | Ecc | -2.309 | 2.092e-2 | 0.013 |
| 6 | Mlsf | -1.660 | 9.694e-2 | 0.017 |
| 7 | Lflc | -0.180 | 9.138e-1 | 0.025 |
| 8 | Br$^\kappa$ | 0.613 | 1.000e0 | 0.050 |

| | | *Macro-averaging AUC* | | |
|---|---|---|---|---|
| $j$ | approach | $z_j$ | $p_j$ | $\alpha/(k-j+1)$ |
| 2 | Ecc | -1.732 | 8.326e-2 | 0.007 |
| 3 | Mlsf | -0.505 | 6.134e-1 | 0.008 |
| 4 | Lflc | 0.505 | 1.000e0 | 0.010 |
| 5 | Rakel | 1.010 | 1.000e0 | 0.013 |
| 6 | Br$^\kappa$ | 1.119 | 1.000e0 | 0.017 |
| 7 | Ml-knn | 2.815 | 1.000e0 | 0.025 |
| 8 | Lift$^\kappa$ | 4.871 | 1.000e0 | 0.050 |

rank than Ecc and Mlsf in terms of all evaluation metrics, and achieves a lower average rank than Ml-knn, Lift$^\kappa$ and Lflc in terms of all evaluation metrics except *macro-averaging AUC*.

- Wrap significantly outperforms Rakel in terms of all evaluation metrics except *macro-averaging AUC*, significantly outperforms Ml-knn in terms of *one-error*, *ranking loss* and *average precision*, and significantly outperforms Mlsf in terms of *one-error* and *average precision*. Furthermore, Wrap$^\kappa$ significantly outperforms Rakel, Ml-knn and Lift$^\kappa$ in terms of all evaluation metrics except *macro-averaging AUC*, and significantly outperforms Ecc in terms of *hamming loss*.
- Similar to Lift$^\kappa$, Ml-knn achieves better performance than the comparing approaches in terms of *macro-averaging AUC*, which measures the predictive performance by taking each class label independently. Nonetheless, its performance becomes worse on the other evaluation metrics which measure the predictive performance by considering the ranking relations among class labels.

## 4.3 Further Analysis

### 4.3.1 Usefulness of Embedding

For the proposed Wrap approach, an embedding matrix $\mathbf{V}$ is employed to facilitate the procedure of joint label-specific features generation and classification model induction. In this subsection, the usefulness of taking the embedding operation is further investigated. Specifically, a degenerated version of the proposed approach named Wrap$_{deg}$ is considered which works by fixing the embedding matrix $\mathbf{V}$ as an identity matrix.

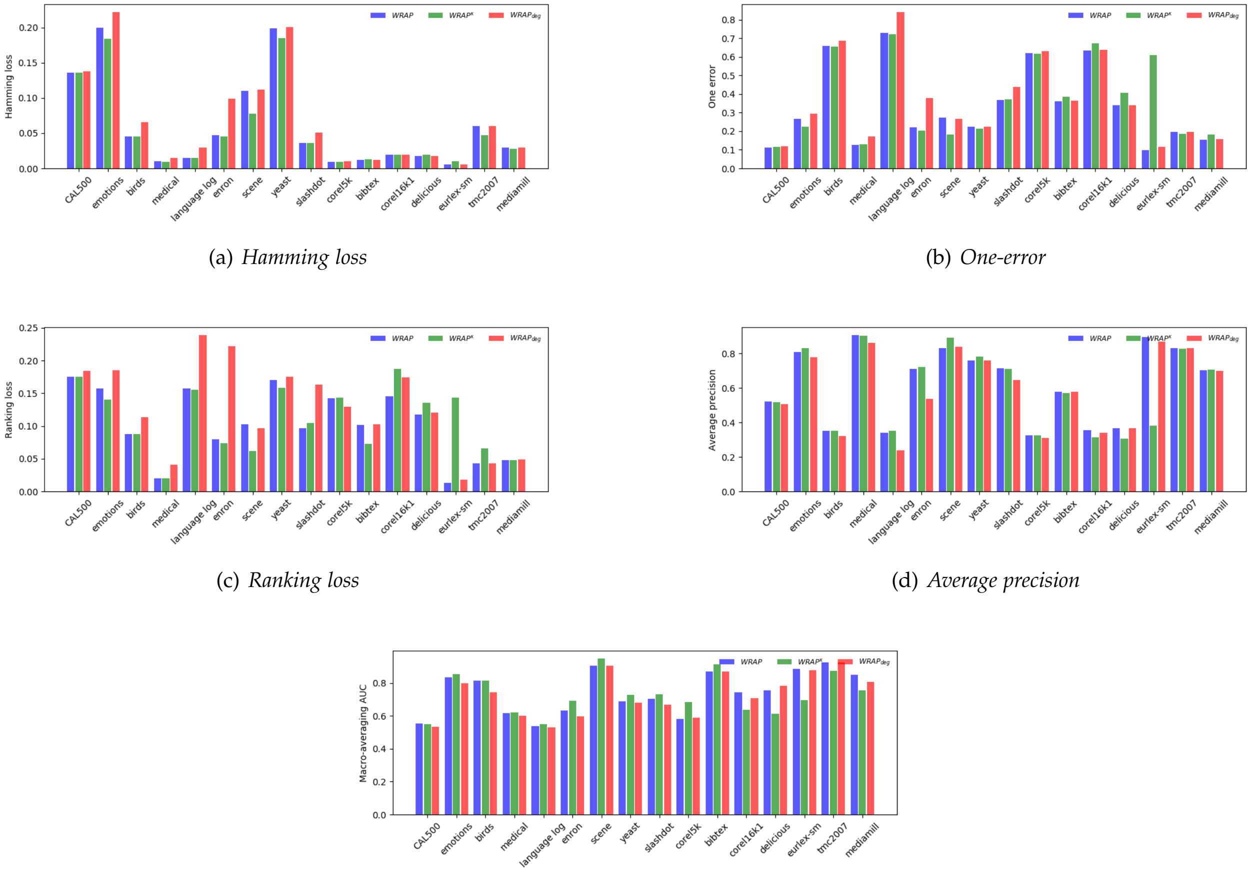
Fig. 1 illustrates the predictive performance of Wrap, Wrap$^\kappa$ and Wrap$_{deg}$ in terms of each evaluation metric. Pairwise $t$-test at 0.05 significance level show that, out of 80 cases (16 data sets × 5 evaluation metrics), the proposed approach achieves superior or at least comparable performance to its degenerated version in 97.5% (linear mode) and 72.5% (kernel mode) cases respectively. These results clearly show the usefulness of exploiting the embedding matrix adaptively in the wrapped optimization procedure to improve the generalization performance of the resulting model.

### 4.3.2 Sensitivity Analysis

To show the performance sensitivity of the proposed approach, Fig. 2 gives an illustrative example on how the performance of Wrap (linear mode) changes with varying configurations of parameters $\alpha$ $(d = \alpha \min(m, l))$, $\lambda_1$, $\lambda_2$ and $\lambda_3$ (data set: `slashdot`; evaluation metric: *hamming loss*). As shown in Fig. 2, the performance of Wrap is relatively stable as the parameter values change within a reasonable range. This serves as a desirable property in using the proposed approach, which can be observed on other data sets and evaluation metrics as well.

### 4.3.3 Convergence Analysis

To show the convergence property of the proposed approach, Fig. 3 gives an illustrative example on how the objective function value of Wrap (linear mode; Eq.(3)) and Wrap$^\kappa$ (kernel mode; Eq.(17)) changes as the number of iterations increases on four data sets `birds`, `CAL500`, `emotions` and `medical`. As shown in Fig. 3, the objective function value decreases significantly in initial iterations and gradually converges as the optimization procedure proceeds. Therefore, for the experiments conducted in this paper, the optimization procedure for the proposed approach is terminated if the decrease in objective function value is less than $10^{-4}$ after one alternating iteration.



(a) *Hamming loss*

(b) *One-error*

(c) *Ranking loss*

(d) *Average precision*

(e) *Macro-averaging AUC*

Fig. 1. Performance of WRAP, WRAP$^{\kappa}$ and the degenerated version WRAP$_{deg}$ in terms of each evaluation metric.

## 5 Conclusion

In this paper, the strategy of label-specific features for multi-label classification is investigated. Different to existing approaches which work in the generation-then-induction two-stage manner, a novel approach is proposed which performs label-specific feature generation and classification model induction in a joint manner. The wrapping procedure is instantiated based on (kernelized) linear implementation with empirical loss minimization and pairwise label correlation regularization. Extensive experiments show that the proposed wrapping approach serves as a promising solution for multi-label classification based on label-specific features.

The proposed approach serves as one feasible solution to jointly consider generating label-specific features and inducing classification model, it is interesting to investigate alternative instantiations for wrapped multi-label classification with label-specific features generation in the future.

## References

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

[2] C. Brinker, E. Loza Mencía, and J. Fürnkranz, "Graded multilabel classification by pairwise comparisons," in *Proceedings of the 14th IEEE International Conference on Data Mining*, Shenzhen, China, 2014, pp. 731–736.

[3] S. Canuto, M. A. Gonçalves, and F. Benevenuto, "Exploiting new sentiment-based meta-level features for effective sentiment analysis," in *Proceedings of the 9th ACM International Conference on Web Search and Data Mining*, San Francisco, CA, 2016, pp. 53–62.

[4] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, no. 3, p. Article 27, 2011.

[5] C. Chen, H.-B. Wang, W.-W. Liu, X.-Y. Zhao, T.-L. Hu, and G. Chen, "Two-stage label embedding via neural factorization machine for multi-label classification," in *Proceedings of the 33rd AAAI Conference on Artificial Intelligence*, Honolulu, HI, 2019, pp. 3304–3311.

[6] Z.-M. Chen, X.-S. Wei, P. Wang, and Y.-W. Guo, "Multi-label image recognition with graph convolutional networks," in *Proceedings of the 32nd IEEE Conference on Computer Vision and Pattern Recognition*, Long Beach, CA, 2019, pp. 5177–5186.

[7] Z.-S. Chen and M.-L. Zhang, "Multi-label learning with regularization enriched label-specific features," in *Proceedings of the 11th Asian Conference on Machine Learning*, Nagoya, Japan, 2019, pp. 411–424.

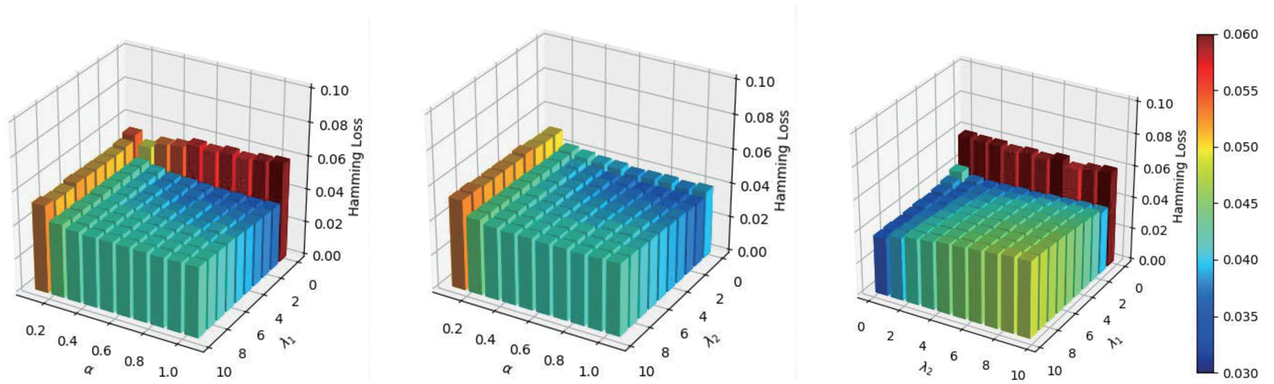
[8] Z. Cheng and Z. Zeng, "Joint label-specific features and label correlation for multi-label learning with missing label," *Applied Intelligence*, vol. 50, pp. 4029–4049, 2020.

[9] J. Demšar, "Statistical comparisons of classifiers over multiple data sets," *Journal of Machine Learning Research*, vol. 7, no. Jan, pp. 1–30, 2006.

[10] J. Fürnkranz, E. Hüllermeier, E. Loza Mencía, and K. Brinker, "Multilabel classification via calibrated label ranking," *Machine Learning*, vol. 73, no. 2, pp. 133–153, 2008.

[11] E. Gibaja and S. Ventura, "A tutorial on multilabel learning," *ACM Computing Surveys*, vol. 47, no. 3, p. Article 52, 2015.

[12] H. Gouk, B. Pfahringer, and M. Cree, "Learning distance metrics for multi-label classification," in *Proceedings of the 8th Asian Conference on Machine Learning*, Hamilton, New Zealand, 2016, pp. 318–

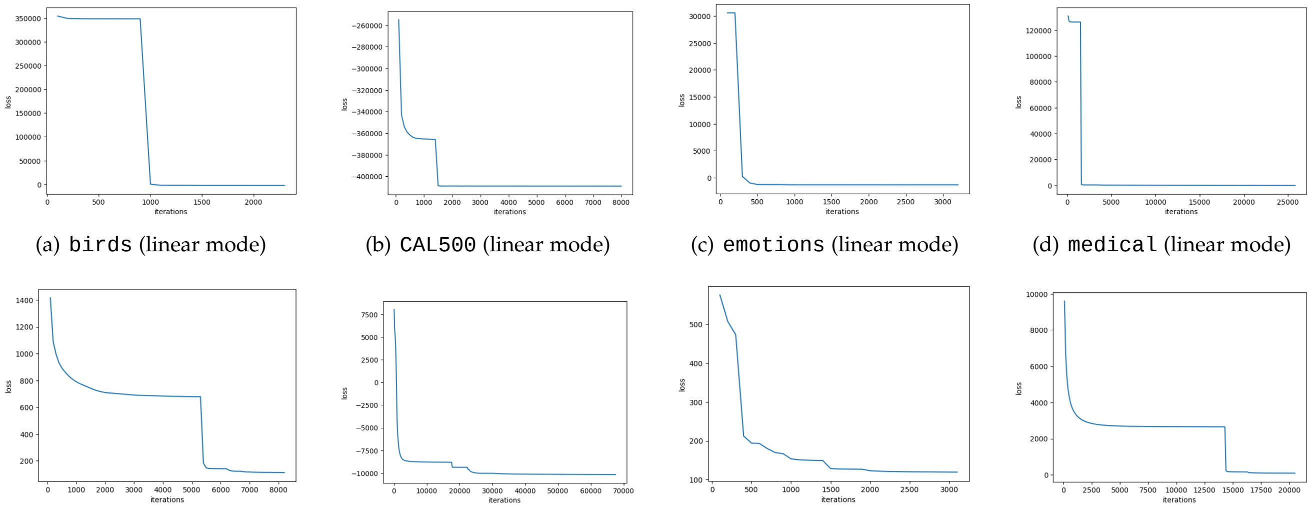


(a) $\alpha \in \{0.2, 0.4, \ldots, 1.0\}$ and $\lambda_1 \in \{0, 2, \ldots, 10\}$ ($\lambda_2 = 2, \lambda_3 = 0.1$)

(b) $\alpha \in \{0.2, 0.4, \ldots, 1.0\}$ and $\lambda_2 \in \{0, 2, \ldots, 10\}$ ($\lambda_1 = 2, \lambda_3 = 0.1$)

(c) $\lambda_1, \lambda_2 \in \{0, 2, \ldots, 10\}$ ($\alpha = 0.9, \lambda_3 = 0.1$)

Fig. 2. Predictive performance of Wrap (linear mode) with varying parameter configurations (data set: `slashdot`; evaluation metric: *hamming loss*).



(a) `birds` (linear mode)

(b) `CAL500` (linear mode)

(c) `emotions` (linear mode)

(d) `medical` (linear mode)

(e) `birds` (kernel mode)

(f) `CAL500` (kernel mode)

(g) `emotions` (kernel mode)

(h) `medical` (kernel mode)

Fig. 3. The objective function value (i.e. predictive loss) of Wrap (first row) and Wrap$^{\kappa}$ (second row) changes as the number of iterations increases on four data sets `birds`, `CAL500`, `emotions` and `medical`.

333.

[13] Y. Guo, F. Chung, G. Li, J. Wang, and J. C. Gee, “Leveraging label-specific discriminant mapping features for multi-label learning,” *ACM Transactions on Knowledge Discovery from Data*, vol. 13, no. 2, p. Article 24, 2019.

[14] H. Hartono, O. S. Sitompul, T. Tulus, and E. B. Nababan, “Biased support vector machine and weighted-smote in handling class imbalance problem,” *International Journal of Advances in Intelligence Informatics*, vol. 4, no. 1, pp. 21–27, 2018.

[15] J. Huang, G. Li, Q. Huang, and X. Wu, “Learning label-specific features and class-dependent labels for multi-label classification,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 28, no. 12, pp. 3309–3323, 2016.

[16] ——, “Joint feature selection and classification for multilabel learning,” *IEEE Transactions on Cybernetics*, vol. 48, no. 3, pp. 876–889, 2018.

[17] M. Huang, F. Zhuang, X. Zhang, X. Ao, Z. Niu, M.-L. Zhang, and Q. He, “Supervised representation learning for multi-label classification,” *Machine Learning*, vol. 108, no. 5, pp. 747–763, 2019.

[18] B.-B. Jia and M.-L. Zhang, “Multi-dimensional classification via stacked dependency exploitation,” *Science China Information Sciences*, vol. 63, no. 12, p. Article 222102, 2020.

[19] X.-Y. Jia, S.-S. Zhu, and W.-W. Li, “Joint label-specific features and correlation information for multi-label learning,” *Journal of Computer Science and Technology*, vol. 35, no. 2, pp. 247–258, 2020.

[20] T. Li, S. Gao, and Y. Xu, “Deep multi-similarity hashing for multi-label image retrieval,” in *Proceedings of the 26th ACM International Conference on Information and Knowledge Management*, Singapore, 2017, pp. 2159–2162.

[21] Y. Li, Y. Song, and J. Luo, “Improving pairwise ranking for multi-label image classification,” in *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, HI, 2017, pp. 3617–3625.

[22] W. Liu and I. W. Tsang, “Large margin metric learning for multi-label prediction,” in *Proceedings of the 29th AAAI Conference on Artificial Intelligence*, Austin, TX, 2015, pp. 2800–2806.

[23] J. Ma, H. Zhang, and T. W. S. Chow, “Multilabel classification with label-specific features and classifiers: A coarse- and fine-tuned framework,” *IEEE Transactions on Cybernetics*, in press.

[24] Y. Ma, C. Cui, J. Yu, J. Guo, G. Yang, and Y. Yin, “Multi-task MIML learning for pre-course student performance prediction,” *Frontiers of Computer Science*, vol. 14, no. 5, p. Article 145313, 2020.

[25] J. Nam, Y.-B. Kim, E. Loza-Mencía, S. Park, R. Sarikaya, and J. Fürnkranz, “Learning context-dependent label permutations for multi-label classification,” in *Proceedings of the 36th International Conference on Machine Learning*, Long Beach, CA, 2019, pp. 4733–4742.

[26] R. B. Pereira, A. Plastino, B. Zadrozny, and L. H. C. Merschmann, “Categorizing feature selection methods for multi-label classification,” *Artificial Intelligence Review*, vol. 49, no. 1, pp. 57–78, 2018.

[27] 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.

[28] T. N. Rubin, A. Chambers, P. Smyth, and M. Steyvers, "Statistical topic models for multi-label document classification," *Machine Learning*, vol. 88, no. 1-2, pp. 157–208, 2012.

[29] B. Schölkopf and A. J. Smola, *Learning with Kernels: Support Vector Machines, Regularization, Optimization and Beyond*. Camridge, MA: MIT Press, 2001.

[30] F. Sun, J. Tang, H. Li, G.-J. Qi, and T. S. Huang, "Multi-label image categorization with sparse factor representation," *IEEE Transactions on Image Processing*, vol. 23, no. 3, pp. 1028–1037, 2014.

[31] L. Sun, S. Ji, and J. Ye, *Multi-label Dimensionality Reduction*. Boca Ration, FL: Chapman and Hall/CRC, 2013.

[32] L. Sun, M. Kudo, and K. Kimura, "Multi-label classification with meta-label-specific features," in *Proceedings of the 23rd International Conference on Pattern Recognition*, Cancun, Mexico, 2016, pp. 1612–1617.

[33] Y.-P. Sun and M.-L. Zhang, "Compositional metric learning for multi-label classification," *Frontiers of Computer Science*, vol. 15, no. 5, p. Article 155320, 2021.

[34] G. Tsoumakas, I. Katakis, and I. Vlahavas, "Random k-labelsets for multi-label classification," *IEEE Transactions on Knowledge and Data Engineering*, vol. 23, no. 7, pp. 1079–1089, 2011.

[35] X. Wang and G. Sukthankar, "Multi-label relational neighbor classification using social context features," in *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Chicago, IL, 2013, pp. 464–472.

[36] Y. Wang, W. Zheng, Y. Cheng, and D. Zhao, "Joint label completion and label-specific features for multi-label learning algorithm," *Soft Computing*, vol. 24, pp. 6553–6569, 2020.

[37] W. Weng, Y.-N. Chen, C.-L. Chen, S.-X. Wu, and J.-H. Liu, "Non-sparse label specific features selection for multi-label classification," *Neurocomputing*, vol. 377, pp. 85–94, 2020.

[38] W. Weng, Y. Lin, S. Wu, Y. Li, and Y. Kang, "Multi-label learning based on label-specific features and local pairwise label correlation," *Neurocomputing*, vol. 273, pp. 385–394, 2018.

[39] B. Wu, E. Zhong, A. Horner, and Q. Yang, "Music emotion recognition by multi-label multi-layer multi-instance multi-view learning," in *Proceedings of the 22nd ACM International Conference on Multimedia*, Orlando, FL, 2014, pp. 117–126.

[40] X. Wu, Q.-G. Chen, Y. Hu, D. Wang, X. Chang, X. Wang, and M.-L. Zhang, "Multi-view multi-label learning with view-specific information extraction," in *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, Macau, China, 2019, pp. 3884–3890.

[41] Y. Xing, G. Yu, D. Carlotta, J. Wang, and Z. Zhang, "Multi-label co-training," in *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, Stockholm, Sweden, 2018, pp. 2882–2888.

[42] J.-H. Xu, H.-D. Tian, Z.-Y. Wang, Y. Wang, F. Chen, and W.-X. Kang, "Joint input and output space learning for multi-label image classification," *IEEE Transactions on Multimedia*, in press.

[43] M. Xu and L.-Z. Guo, "Learning from group supervision: The impact of supervision deficiency on multi-label learning," *Science China Information Sciences*, vol. 64, no. 3, p. Article 130101, 2021.

[44] S. Xu, X. Yang, H. Yu, D.-J. Yu, J. Yang, and E. C. C. Tsang, "Multi-label learning with label-specific feature reduction," *Knowledge-Based Systems*, vol. 104, pp. 52–61, 2016.

[45] Y. Yang and S. Gopal, "Multilabel classification with meta-level features in a learning-to-rank framework," *Machine Learning*, vol. 88, no. 1-2, pp. 47–68, 2012.

[46] C. K. Yeh, W. C. Wu, W. J. Ko, and Y. C. F. Wang, "Learning deep latent spaces for multi-label classification," in *Proceedings of the 31st AAAI Conference on Artificial Intelligence*, San Francisco, CA, 2017, pp. 2838–2844.

[47] C. Zhang, Z. Yu, Q. Hu, P. Zhu, X. Liu, and X. Wang, "Latent semantic aware multi-view multi-label classification," in *Proceedings of the 32nd AAAI Conference on Artificial Intelligence*, New Orleans, LA, 2018, pp. 4414–4421.

[48] J. Zhang, C. Li, D. Cao, Y. Lin, S. Su, L. Dai, and S. Li, "Multi-label learning with label-specific features by resolving label correlations," *Knowledge-Based Systems*, vol. 159, pp. 148–157, 2018.

[49] M.-L. Zhang, Y.-K. Li, Y.-Y. Liu, and X. Geng, "Binary relevance for multi-label learning: An overview," *Frontiers of Computer Science*, vol. 12, no. 2, pp. 191–202, 2018.

[50] M.-L. Zhang and L. Wu, "LIFT: Multi-label learning with label-specific features," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 1, pp. 107–120, 2015.

[51] M.-L. Zhang and Z.-H. Zhou, "A review on multi-label learning algorithms," *IEEE Transactions on Knowledge and Data Engineering*, vol. 26, no. 8, pp. 1819–1837, 2014.

[52] ——, "Ml-knn: A lazy learning approach to multi-label learning," *Pattern recognition*, vol. 40, no. 7, pp. 2038–2048, 2007.