Supplementary Materials for Dual Set Multi-Label Learning

Chong Liu,[1,2] **Peng Zhao,**[1,2] **Sheng-Jun Huang,**[3] **Yuan Jiang,**[1,2] **Zhi-Hua Zhou**[1,2]
[1] National Key Laboratory for Novel Software Technology, Nanjing University, Nanjing 210023, China
[2] Collaborative Innovation Center of Novel Software Technology and Industrialization, Nanjing 210023, China
[3] College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing 211106, China
{liuc, zhaop, jiangy, zhouzh}@lamda.nju.edu.cn, huangsj@nuaa.edu.cn

Details of Benchmark Approaches

In this section, we will introduce several benchmark approaches to deal with dual set multi-label learning problems.

Independent Decomposition

For each label set, the Independent Decomposition method first constructs the corresponding multi-class training sets:

$$\mathcal{D}^j = \{(\boldsymbol{x}_i, y_i^j) | 1 \leq i \leq m, j \in \{a, b\}\},$$

where $y_i^j \in \mathcal{Y}^j$. Afterwards, a multi-class algorithm $\mathcal{A}$ can be used to train a model

$$h^j : \mathcal{X} \to \mathcal{Y}^j,$$

for the set j, i.e.

$$h^j \leftarrow \mathcal{A}(\mathcal{D}^j).$$

For an unseen instance $\boldsymbol{x}$, the algorithm predicts its two labels by querying each multi-class classifier and then combining their predict labels as the final result:

$$[h^a(\boldsymbol{x}), h^b(\boldsymbol{x})].$$

Obviously, the main drawback of Independent Decomposition is that its classifiers are learned on each label set independently, so it neglects the relationship between the two label sets.

Co-Occurrence Based Decomposition

Suppose label co-occurrence set $\mathcal{C}$ is used to record all the label co-occurrence cases in the training set. At first, $\mathcal{C}$ is empty. Then the algorithm will scan the training set, and for i-th instance, if (y_i^a, y_i^b) does not appear in $\mathcal{C}$, it will be added in $\mathcal{C}$. Then, the original training set will be:

$$\mathcal{D} = \{(\boldsymbol{x}_i, t_i) | 1 \leq i \leq m\},$$

where $t_i = 1, \cdots, |\mathcal{C}|$ is the new class label. Next, a multi-class algorithm $\mathcal{A}$ can be used to train a classifier:

$$h \leftarrow \mathcal{A}(\mathcal{D}),$$

predictions are made by querying the multi-class classifier.

Co-Occurrence Based Decomposition suffers from two problems. First, the number of co-occurrence cases could be large. An extreme case is that each label in $\mathcal{Y}^a$ co-occur with each label in $\mathcal{Y}^b$, which will result in all $L_1 \times L_2$ classes for the transformed multi-class problem, where L_1, L_2 refer to sizes of dual label sets, respectively. This will lead to insufficient training data for each class, and low efficiency of the overall task. Second, this method can only predict the label combinations which have been occurred in the training set, and thus may lead to poor performance on the testing set.

Label Stacking

Given the training set $\mathcal{D} = \{(\boldsymbol{x}_i, y_i^a, y_i^b) | 1 \leq i \leq m\}$, without loss of generality, we assume that $L_1 \geq L_2$, and a multi-class algorithm $\mathcal{A}$ is applied to learn two classifiers:

$$h^b \leftarrow \mathcal{A}(\{(\boldsymbol{x}_i, y_i^b) | 1 \leq i \leq m\}),$$
$$h^a \leftarrow \mathcal{A}(\{([\boldsymbol{x}_i, y_i^b], y_i^a) | 1 \leq i \leq m\}),$$

where $[x_1, x_2]$ concatenates x_1 and x_2. For an unseen instance $\boldsymbol{x}$, its second label is predicted by

$$h^b(\boldsymbol{x}),$$

then its first label is predicted by

$$h^a([\boldsymbol{x}, h^b(\boldsymbol{x})]).$$

Here the prediction result $h^b(\boldsymbol{x})$ is a 0-1 matrix, where each 1 indicates that the instance is associated with certain label, 0 otherwise.

Label Stacking differs from Independent Decomposition in its ability to exploit inter-set label relationship. Moreover, different to Co-Occurrence Based Decomposition, Label Stacking only makes one label set help the other one rather than making two sets help each other. The reason of the assumption $L_1 \geq L_2$ lies in the fact that generally multi-class learning with fewer classes performs better, which is able to provide better pseudo-labels for the other label set.

Proofs for Theoretical Results

In this part, we provide detailed proofs for theoretical results in the main paper.

Proof for Theorem 1

For the case with dual sets of labels, we follow (Mohri, Rostamizadeh, and Talwalkar 2012) to define the following margin:

Definition 1. *A hypothesis h is defined based on a scoring function $g : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$. Then, the label associated to point $\boldsymbol{x}$ is the one resulting in the largest score $g(\boldsymbol{x}, y)$, which defines the following mapping from $\mathcal{X}$ to $\mathcal{Y}$,*

$$h : \boldsymbol{x} \to \arg\max_{y \in \mathcal{Y}} g(\boldsymbol{x}, y),$$

then definition of the margin $\rho_h(\boldsymbol{x}, y)$ of the function g at a labeled example $(\boldsymbol{x}, y)$,

$$\bar{\rho}_h(\boldsymbol{x}, y) = g(\boldsymbol{x}, y) - \max_{y' \neq y} g(\boldsymbol{x}, y').$$

Similarly, we define the margin for directly learning from all labels as follows:

Definition 2. *A hypothesis h is defined based on a scoring function $g : \mathcal{X} \times [\mathcal{Y}^a \times \mathcal{Y}^b] \to \mathbb{R}$. Then definition of the margin $\rho_h(\boldsymbol{x}, y)$ of the function g at a labeled example $(\boldsymbol{x}, y)$, $y = [y^a, y^b]$,*

$$\bar{\bar{\rho}}_h(\boldsymbol{x}, y) = \min\{g(\boldsymbol{x}, y^a), g(\boldsymbol{x}, y^b)\} - \max_{y' \neq y^a \wedge y' \neq y^b} g(\boldsymbol{x}, y').$$

Based on Definition 1 and 2 given above, we have the following theorem:

Theorem 1. *For dual-set multi-label learning problems, h^a and h^b are classifiers trained on the instance space $\mathcal{X}$ and label space $\mathcal{Y}^a$, $\mathcal{Y}^b$ respectively. h is a classifier trained directly from $\mathcal{X} \times [\mathcal{Y}^a \times \mathcal{Y}^b]$, namely,*

$$h : \boldsymbol{x} \to \arg\max_{y^a, y^b \in [\mathcal{Y}^a \times \mathcal{Y}^b]} h(\boldsymbol{x}, y),$$

where $y = [y^a, y^b]$, then margin of learning from dual label set is larger than that of directly learning from all labels:

$$\min\{\bar{\rho}_{h^a}(\boldsymbol{x}, y^a), \bar{\rho}_{h^b}(\boldsymbol{x}, y^b)\} \geq \bar{\bar{\rho}}_h(\boldsymbol{x}, y).$$

Proof. From the definition

$$\bar{\bar{\rho}}_h(\boldsymbol{x}, y) = \min\{h^a(\boldsymbol{x}, y^a), h^b(\boldsymbol{x}, y^b)\} - \max_{y' \neq y^a \wedge y' \neq y^b} h(\boldsymbol{x}, y'),$$

without loss of generality, we could assume that $h^a(\boldsymbol{x}, y^a) \leq h^b(\boldsymbol{x}, y^b)$, then

$$\begin{aligned}
\bar{\bar{\rho}}_h(\boldsymbol{x}, y) &= h^a(\boldsymbol{x}, y^a) - \max_{y' \neq y^a \wedge y' \neq y^b} h(\boldsymbol{x}, y') \\
&\leq h^a(\boldsymbol{x}, y^a) - \max_{y' \neq y^a \wedge y' \in [L_1]} h^a(\boldsymbol{x}, y') \\
&= \bar{\rho}_{h^a}(\boldsymbol{x}, y^a).
\end{aligned}$$

Similarly, we could also prove that $\bar{\bar{\rho}}_h(\boldsymbol{x}, y) \leq \bar{\rho}_{h^b}(\boldsymbol{x}, y^b)$, thus we have

$$\min\{\bar{\rho}_{h^a}(\boldsymbol{x}, y^a), \bar{\rho}_{h^b}(\boldsymbol{x}, y^b)\} \geq \bar{\bar{\rho}}_h(\boldsymbol{x}, y).$$

□

Remark. From Theorem 1, we can see that the margin of h is bounded by the minimum of margin of h^a and h^b. The margin is the larger the better. Thus, this bound implies the effectiveness of splitting the whole label set into two disjoint label sets. This exactly accords with our intuition, that we should consider the hierarchical structure in label sets.

Proofs for Theorem 2

Consider the approach that splits label sets into dual sets, we name it as splitting approach:

$$h^{spl}(\boldsymbol{x}) = [h^a(\boldsymbol{x}), h^b(\boldsymbol{x})],$$

then, we give the definitions of empirical margin loss and risks based on *hamming loss* as follows,

Definition 3. *(Empirical Margin Loss (Mohri, Rostamizadeh, and Talwalkar 2012))*

$$\hat{R}_\rho(h) = \frac{1}{m} \sum_{i=1}^{m} \Phi_\rho(\rho_h(\boldsymbol{x}_i, y_i)),$$

where $\Phi_\rho(\cdot)$ is the margin loss function defined as,

$$\Phi_\rho = \begin{cases} 0, & \text{if } \rho \leq x \\ 1 - x/\rho, & \text{if } 0 \leq x \leq \rho \\ 1. & \text{if } x \leq 0 \end{cases}$$

Remark. Since margin loss function is a monotonously non-increasing function, it means that the larger margin is, the less loss will be.

Definition 4. *(Risks Based on Hamming Loss)*

$$R(h) = \mathbb{E}_{(\boldsymbol{x}, y) \sim \mathcal{D}} \left[\frac{1}{L_1 + L_2} \sum_{\ell=1}^{L_1 + L_2} [\![h_\ell(\boldsymbol{x}) \neq y_\ell]\!] \right],$$

$$R(h^a) = \mathbb{E}_{(\boldsymbol{x}, y^a) \sim \mathcal{D}} \left[\frac{1}{L_1} \sum_{\ell=1}^{L_1} [\![h^a_\ell(\boldsymbol{x}) \neq y^a_\ell]\!] \right],$$

$$R(h^b) = \mathbb{E}_{(\boldsymbol{x}, y^b) \sim \mathcal{D}} \left[\frac{1}{L_2} \sum_{\ell=1}^{L_2} [\![h^b_\ell(\boldsymbol{x}) \neq y^b_\ell]\!] \right].$$

And we have the key observation:

Observation. The losses of these approaches satisfy,

$$[\![h_\ell(\boldsymbol{x}) \neq y_\ell]\!] \leq \max\{[\![h^a_\ell(\boldsymbol{x}) \neq y^a_\ell]\!], [\![h^b_\ell(\boldsymbol{x}) \neq y^b_\ell]\!]\}.$$

Proof. Since $[\![\cdot]\!]$ is either 1 or 0, we only need to bound the case when the right hand side is equal to 0.

As we know that $h(\boldsymbol{x}) = [h^a(\boldsymbol{x}), h^b(\boldsymbol{x})]$ and $y = [y^a, y^b]$, when $[\![h^a_\ell(\boldsymbol{x}) \neq y^a_\ell]\!] = 0 \wedge [\![h^b_\ell(\boldsymbol{x}) \neq y^b_\ell]\!] = 0$, we have left hand side as $[\![h_\ell(\boldsymbol{x}) \neq y_\ell]\!] = 0$. □

Based on Definition 3 and 4, we have the following generalization bound of the approach that splits the total label set into dual label sets:

Theorem 2. *Let $H = \{(\boldsymbol{x}, y^a, y^b) \in \mathcal{X} \times [\mathcal{Y}^a \times \mathcal{Y}^b] \to \mathbf{w}^{\mathrm{T}} \phi(\boldsymbol{x}) | \sum_{\ell=1}^{L_1 + L_2} \|\mathbf{w}\|^2_{\mathbb{H}} \leq \Lambda^2\}$ be a hypothesis set with $y^a = 1, \cdots, L_1, y^b = 1, \cdots, L_2$, where $\phi : \mathcal{X} \to \mathbb{H}$ is a feature mapping induced by some positive definite kernel κ. Assume that $S \subset \{\boldsymbol{x} : \kappa(\boldsymbol{x}, \boldsymbol{x}) \leq r^2\}$, and fix $\rho > 0$, then for any $\delta > 0$, with probability at least $1 - \delta$, the following generalization bound holds for all $h^{spl} = [h^a, h^b] \in H$:*

$$R(h^{spl}) \leq \hat{R}_\rho(h^{spl}) + \frac{2r\Lambda}{\rho} \sqrt{\frac{\max\{L_1, L_2\}}{m}} + 3\sqrt{\frac{\log(2/\delta)}{m}}.$$

To prove Theorem 2, we firstly provide Lemma 1, Lemma 2, and Lemma 3, which play importance roles in the following proofs.

Lemma 1. *The risks of the approaches satisfy,*

$$R(h^{spl}) \leq \max\{R(h^a), R(h^b)\}.$$

Proof.

$$\begin{aligned}
R(h^{spl}) &= \mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{D}}\left[\frac{1}{L_1+L_2}\sum_{\ell=1}^{L_1+L_2}[\![h_\ell^{spl}(\boldsymbol{x}) \neq y_\ell]\!]\right] \\
&= \frac{1}{L_1+L_2}\mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{D}}\left[\sum_{\ell=1}^{L_1}[\![h_\ell^{spl}(\boldsymbol{x}_i) \neq y_{i,\ell}]\!]\right] \\
&\quad + \frac{1}{L_1+L_2}\mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{D}}\left[\sum_{\ell=L_1+1}^{L_1+L_2}[\![h_\ell^{spl}(\boldsymbol{x}_i) \neq y_{i,\ell}]\!]\right] \\
&= \frac{L_1}{L_1+L_2}R(h^a) + \frac{L_2}{L_1+L_2}R(h^b) \\
&\leq \frac{1}{L_1+L_2}\max\{R(h^a), R(h^b)\}(L_1+L_2) \\
&= \max\{R(h^a), R(h^b)\}.
\end{aligned}$$

□

Lemma 2. *The empirical risks of these approaches satisfy,*

$$\max\{\hat{R}_\rho(h^a), \hat{R}_\rho(h^b)\} \leq \hat{R}_\rho(h^{spl}).$$

Proof.

$$\begin{aligned}
&\max\{\hat{R}_\rho(h^a), \hat{R}_\rho(h^b)\} \\
=&\frac{1}{m}\max\left\{\sum_{i=1}^m \Phi_\rho(\bar{\rho}_{h^a}(\boldsymbol{x}_i, y_i^a)), \sum_{i=1}^m \Phi_\rho(\bar{\rho}_{h^b}(\boldsymbol{x}_i, y_i^b))\right\} \\
\leq&\frac{1}{m}\sum_{i=1}^m \max\left\{\Phi_\rho(\bar{\rho}_{h^a}(\boldsymbol{x}_i, y_i^a)), \Phi_\rho(\bar{\rho}_{h^b}(\boldsymbol{x}_i, y_i^b))\right\} \\
\leq&\frac{1}{m}\sum_{i=1}^m \Phi_\rho(\bar{\bar{\rho}}_{h^{spl}}(\boldsymbol{x}_i, y_i)) = \hat{R}_\rho(h^{spl}).
\end{aligned}$$

The last inequality holds due to Theorem 1 and the fact margin loss function $\Phi(\cdot)$ is monotonically non-increasing. □

Based on similar proof skills in (Lei et al. 2015), we use Gaussian Complexity (Bartlett and Mendelson 2002) to prove a bound which exhibits a radical dependence on the maximal number of labels.

Lemma 3. *Let* $H = \{(\boldsymbol{x}, y) \in \mathcal{X}\times\mathcal{Y} \to \mathbf{w}^{\mathrm{T}}\phi(\boldsymbol{x}) | \sum_{\ell=1}^L \|\mathbf{w}\|_{\mathbb{H}}^2 \leq \Lambda^2\}$ *be a hypothesis set with* $y = 1, \cdots, L$*, where* $\phi : \mathcal{X} \to \mathbb{H}$ *is a feature mapping induced by some positive definite kernel* κ*. Assume that* $S \subset \{\boldsymbol{x} : \kappa(\boldsymbol{x}, \boldsymbol{x}) \leq r^2\}$*, and fix* $\rho > 0$*, then, for any* $\delta > 0$*, with probability at least* $1-\delta$*, the following generalization bound holds for all* $h \in H$*,*

$$R(h) \leq \hat{R}_\rho(h) + \frac{2r\Lambda}{\rho}\sqrt{\frac{L}{m}} + 3\sqrt{\frac{\log(2/\delta)}{m}}.$$

Now, we proceed to prove Theorem 2.

Proof.

$$\begin{aligned}
R(h^{spl}) &\leq \max\{R(h^a), R(h^b)\} \\
&\leq \max\left\{\hat{R}_\rho(h^a) + \frac{2r\Lambda}{\rho}\sqrt{\frac{L_1}{m}} + 3\sqrt{\frac{\log(2/\delta)}{m}},\right. \\
&\qquad \left.\hat{R}_\rho(h^b) + \frac{2r\Lambda}{\rho}\sqrt{\frac{L_2}{m}} + 3\sqrt{\frac{\log(2/\delta)}{m}}\right\} \\
&\leq \max\{\hat{R}_\rho(h^a), \hat{R}_\rho(h^b)\} \\
&\quad + \frac{2r\Lambda}{\rho}\sqrt{\frac{\max\{L_1, L_2\}}{m}} + 3\sqrt{\frac{\log(2/\delta)}{m}} \\
&\leq \hat{R}_\rho(h^{spl}) + \frac{2r\Lambda}{\rho}\sqrt{\frac{\max\{L_1, L_2\}}{m}} + 3\sqrt{\frac{\log(2/\delta)}{m}}.
\end{aligned}$$

□

Remark. From Theorem 2, we can see that it makes sense to split label sets to deal with dual-set multi-label learning since the convergence rate of generalization error is standard as $O(1/\sqrt{m})$. Besides, the error bound exhibits a radical dependence on the maximal number of labels in dual sets. This also implies a relatively balanced splitting on the label sets may improve the performance.

Details of Datasets

In this section, we introduce detailed information about the collection and process of all three datasets mentioned in main paper.

Calligrapher-Font Joint Classification

The first dataset is *Calligrapher-Font* dataset and this is the largest dataset of this paper, whose task is to predict the calligrapher and font simultaneously given images of Chinese characters. The dataset comes from 31 different calligraphy works written by 14 ancient famous Chinese calligraphers in 5 kinds of fonts. Each calligraphy work contains hundreds of Chinese characters. Overall, there are 23195 Chinese characters in the dataset. Every Chinese character corresponds to a grey image. They are processed to be 100×100 pixels black-and-white images in order to remove the effect of brightness and size. Finally they are extracted Dense SIFT (Lowe 2004) feature, which are set to be 512 dimensions.

Brand-Type Joint Classification

The second dataset is *Brand-Type* dataset, which is aimed at predicting the car brand and type given the car image. The dataset comes from 2247 colorful car images collected on the Internet, which belongs to 7 brands and 3 types. All images are processed to be $224\times224\times3$ pixels images in order to fit the input size of the Vgg-Verydeep-16 net (Simonyan and Zisserman 2014). Then the net is used to extract CNN feature for each image, which are 4096 dimensions. In detail, the output of softmax layer before the last layer is used as the feature vector.

Frequency-Gender Joint Classification

The last dataset is *frequency-gender* dataset, whose task is, given voice feature information, to predict the frequency range of voice and the gender of the speaker simultaneously. It comes from www.primaryobjects.com and its original job is to identify the gender of a voice. There are 3168 voice instances marked with male or female as their labels and 20 voice attributes are used to describe them. Among the attributes, we adapt the mean frequency to be our first label set. Since the human voice frequency ranges from 0Hz to 280Hz, we cut the range into 7 parts, every 40Hz for an interval, and label the mean frequency from 1 to 7. Very few instances are labeled as 1 or 2, i.e., lower than 80Hz, which are removed in order to avoid serious class imbalance problems. In this way, we get the first label set, whose number of class is 5. And we directly take the original male/female label as our second label set.

References

Bartlett, P. L., and Mendelson, S. 2002. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research* 3:463–482.

Lei, Y.; Dogan, Ü.; Binder, A.; and Kloft, M. 2015. Multi-class svms: From tighter data-dependent generalization bounds to novel algorithms. In *Advances in Neural Information Processing Systems 28*, 2035–2043.

Lowe, D. G. 2004. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision* 60(2):91–110.

Mohri, M.; Rostamizadeh, A.; and Talwalkar, A. 2012. *Foundations of machine learning*. MIT press.

Simonyan, K., and Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. *CoRR* abs/1409.1556.