Label Distribution Learning Machine

Jing Wang [1 2] **Xin Geng** [1 2]

Abstract

Although Label Distribution Learning (LDL) has witnessed extensive classification applications, it faces the challenge of objective mismatch – the objective of LDL mismatches that of classification, which has seldom been noticed in existing studies. Our goal is to solve the objective mismatch and improve the classification performance of LDL. Specifically, we extend the margin theory to LDL and propose a new LDL method called **L**abel **D**istribution **L**earning **M**achine (LDLM). First, we define the label distribution margin and propose the **S**upport **V**ector **R**egression **M**achine (SVRM) to learn the optimal label. Second, we propose the adaptive margin loss to learn label description degrees. In theoretical analysis, we develop a generalization theory for the SVRM and analyze the generalization of LDLM. Experimental results validate the better classification performance of LDLM.

1. Introduction

Label Distribution Learning (LDL) (Geng, 2016) is a novel learning paradigm, in which each instance is annotated with a label distribution. Essentially, a label distribution is a multi-dimensional vector, whose elements are called the label description degrees indicating the relative importance of labels. Fig.1 shows an image from the famous JAFFE (Lyons et al., 1998) dataset with a ground-truth label “ANG”. The mean ratings for the six expressions are re-scaled to a label distribution $\{0.09, 0.14, 0.10, 0.30, 0.25, 0.12\}$, which models the different importance of labels. LDL directly learns a mapping from instances to label distributions. Compared with single-label learning (SLL) and multi-label learning (MLL), LDL straightly considers label ambiguity (Gao et al., 2017) and attracts lots of attention from researchers.

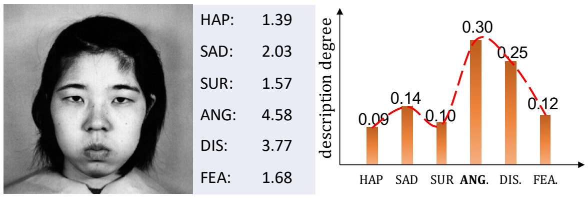


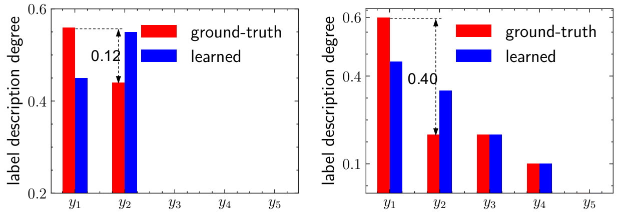
Figure 1. An image from the JAFFE dataset (Lyons et al., 1998) with a ground-truth label “ANG”. The label distribution is obtained by re-scaling the mean ratings, which explicitly modes the relative importance of different emotions.

LDL has been applied to varieties of real-world classification problems, such as emotion recognition (Li & Deng, 2019; Yang et al., 2017), multi-label learning (Zhang et al., 2021), age estimation (Shen et al., 2017), facial beauty perception (Xie et al., 2015), head-pose estimation (Geng et al., 2020), etc. A common practice is as follows. First, in the training phase, an LDL model is learned from the training set (with label distribution) by minimizing the distance between the model’s outputs and the ground-truth label distribution (Geng, 2016). Second, in the test phase, for a test instance, the label having the highest predicted label description degree by the learned model is treated as the predicted label (Wang & Geng, 2019a). For example, in the application of age estimation, Shen et al. (2017) first learned an LDL function from the facial images described by (age) label distribution. Then, for an unknown image, simply the age having the highest predicted label description degree is regarded as the predicted age.

Although LDL has found wide applications, it faces the challenge of **objective mismatch** (Gao et al., 2018; Wang & Geng, 2019a). The objective of LDL is to learn the whole label distribution (e.g., $\{0.09, 0.14, 0.10, 0.30, 0.25, 0.12\}$ in Fig. 1), while the goal of classification is to learn the optimal label (e.g., “ANG” in Fig. 1). One may not expect good classification performance even if the label distribution is well learned. To see that, we present an example in Fig. 2, where the red and the blue bars represent the ground-truth and the learned label distributions, respectively. For Fig. 2(a), the L_1-norm loss of the learned LDL function equals 0.22, and the predicted label y_2 is different from the optimal label y_1. In contrast, for Fig. 2(b), the L_1-norm loss of the learned LDL function equals 0.3 while the prediction

[1] School of Computer Science and Engineering, Southeast University, Nanjing, China [2] Key Laboratory of Computer Network and Information Integration (Southeast University), Ministry of Education. Correspondence to: Xin Geng <xgeng@seu.edu.cn>.

Proceedings of the 38th International Conference on Machine Learning, PMLR 139, 2021. Copyright 2021 by the author(s).



(a) $\ell_1 = 0.22, \ell_{0/1} = 1$

(b) $\ell_1 = 0.3, \ell_{0/1} = 0$

Figure 2. Illustration of objective mismatch. The red and blue bars are the ground-truth the learned label distributions, respectively. Fig. 2(a) has a lower L_1-norm loss than Fig. 2(b), while Fig. 2(b) has a lower 0/1 loss than Fig. 2(a). Note that Fig. 2(a) is superior to Fig. 2(b) in terms of LDL and inferior to Fig. 2(b) in terms of classification, which justifies the objective mismatch.

incurs no loss. To summarize, although Fig. 2(a) has a smaller loss for LDL, Fig.2(b) has a smaller classification loss, which justifies the objective mismatch. The reason lies in that LDL may ignore the optimal label for learning the whole label distribution. The objective mismatch may lead to performance deterioration of LDL (Gao et al., 2018).

To alleviate the objective mismatch and improve the classification performance of LDL, we extend the margin theory (Cortes & Vapnik, 1995) to LDL and propose a novel method named **L**abel **D**istribution **L**earning **M**achine (LDLM). Specifically, we define the label distribution margin that directly connects classification with LDL. Inspired by the theory (Theorem 1 and Corollary 1), we propose the **S**upport **V**ector **R**egression **M**achine (SVRM) to learn the optimal label. Besides, to sufficiently exploit the supervision information of label distribution, we define the adaptive margin loss to learn label description degrees. In the theoretical analysis, we develop a generalization theory for the SVRM and analyze the generalization of LDLM. Finally, experimental results validate the better classification performance of LDLM. Our main contributions are summarized as follows:

1. We define the label distribution margin, which directly connects LDL with classification (Theorems 1 and 2).
2. We extend margin theory to LDL and design a new LDL method called LDLM that uses SVRM and adaptive margin loss to learn label distribution.
3. We develop a generalization theory for SVRM. Besides, we prove the better generalization of LDLM.

The rest of the paper is organized as follows. First, Section 2 briefly reviews some related work. Second, Section 3 presents the LDLM in detail. Third, Section 4 analyzes the generalization. Fourth, Section 5 reports the experimental results. Finally, Section 6 concludes.

2. Related Work

This work is related to two branches of research, including label distribution learning and margin theory, which are briefly discussed as follows.

Geng et al. (2013) first proposed LDL for age estimation. They used label distribution to model the smoothness of aging process and proposed two algorithms IIS-LLD and CPNN to learn the age label distribution. For a test image, the age label having the highest predicted description degree is regarded as the predicted age. Geng & Hou (2015) employed label distribution to cover all rating information from users in pre-release rating on movies and put forward LDL-SVR to learn from such rating distribution. Shen et al. (2017) used differentiable decision trees to learn label distribution and designed LDLFs. Compared with existing parametric LDL methods, LDLFs can model any form of label distributions and can be combined with representation learning (Shen et al., 2017). Considering the ambiguity of acne severity counting and grade, Wu et al. (2019) adopted two label distributions to model the number of lesions and the acne severity of a face image. They designed a multi-task model to learn the label distributions. Although these works apply LDL to classification tasks, none of them have ever noticed the challenge of objective mismatch.

Gao et al. (2018) first observed the objective mismatch (they called it inconsistency) in age estimation – the training objective is to learn the age label distribution (measured by KL-divergence), while the test goal is to predict the ground-truth age (measured by MAE). To solve the objective mismatch, they jointly learned the age label distribution and the ground-truth age (Gao et al., 2018). Nevertheless, the approach is only suitable for real-valued label space. Wang & Geng (2019a) put forward LDL4C that is a specialized LDL algorithm for classification. LDL4C solves the objective mismatch by re-weighing *w.r.t.* information entropy. However, LDL4C is a heuristic method and has no theory guarantees. Compared with Gao et al. (2018) and Wang & Geng (2019a), our work proposes a general LDL method and has a strong theoretical foundation.

Margin theory was first introduced by Vapnik (1995), which maximizes the margin of data and directly leads to Support Vector Machine (SVM) (Cortes & Vapnik, 1995). Later, Vapnik (1995) extended margin to regression problem and proposed Support Vector Regression (SVR) that fits an ϵ-insensitive tube of data. Most importantly, margin theory is a statistical tool that has been applied to analyze the generalization of algorithms, such as boosting (Gao & Zhou, 2013), multi-class classification (Kuznetsov et al., 2014), optimal margin distribution machine (Zhang & Zhou, 2020), etc. Our work extends margin theory to LDL. We define a new margin, i.e., label distribution margin that connects LDL with classification, and design a new method LDLM.

3. The LDLM Approach

This section elaborates on the proposed method. First, we introduce the notations. Second, we propose the SVRM. Third, we introduce the adaptive margin loss. Fourth, we explain the optimization method.

3.1. Notations

Let $\mathcal{X} \in \mathbb{R}^q$ be the input space and $\mathcal{Y} = \{y_1, , y_2, \cdots, y_m\}$ be the label space. Let $\mathcal{D}$ be the (unknown) underlying distribution over $\mathcal{X}$. In LDL, each $\boldsymbol{x}$ is annotated with a label distribution $D = \{d_{\boldsymbol{x}}^{y_1}, d_{\boldsymbol{x}}^{y_2}, \cdots, d_{\boldsymbol{x}}^{y_m}\}$, where $d_{\boldsymbol{x}}^{y_j}$ is called the label description degree and satisfies $\sum_{j=1}^{m} d_{\boldsymbol{x}}^{y_j} = 1$ and $d_{\boldsymbol{x}}^{y_j} \geq 0$ (Geng, 2016). Let $\boldsymbol{x}_i$ and D_i be the ith training instance and label distribution and $S = \{(\boldsymbol{x}_1, D_1), \cdots, (\boldsymbol{x}_n, D_n)\}$ be a training set. Moreover, let $[n]$ take the identity of the set $\{1, 2, \cdots, n\}$, $\text{sign}(\cdot)$ be the sign function, and $\mathbb{I}(\cdot)$ be the indicator function.

For each $\boldsymbol{x} \in \mathcal{X}$, let $y_{\boldsymbol{x}}$ be the optimal label that has the highest label description degree, which is defined by

$$y_{\boldsymbol{x}} = \arg\max_{\bar{y} \in \mathcal{Y}} d_{\boldsymbol{x}}^{\bar{y}}. \tag{1}$$

Let $\hat{d}$ be a learned LDL function and $\hat{D} = \{\hat{d}_{\boldsymbol{x}}^{y_1}, \cdots, \hat{d}_{\boldsymbol{x}}^{y_m}\}$ be the predicted label distribution of $\boldsymbol{x}$. Define the predicted label of $\boldsymbol{x}$ by

$$\hat{y}_{\boldsymbol{x}} = \arg\max_{\bar{y} \in \mathcal{Y}} \hat{d}_{\boldsymbol{x}}^{\bar{y}}, \tag{2}$$

which has the highest predicted label description degree.

3.2. Support Vector Regression Machine (SVRM)

Our goal is to learn the optimal label. To start, we define the **Label Distribution Margin** (LDM).

Definition 1. *For each $\boldsymbol{x}$, the label distribution margin is defined by*

$$\alpha_{\boldsymbol{x}} = \max_{\bar{y} \in \mathcal{Y}} d_{\boldsymbol{x}}^{\bar{y}} - \max_{\bar{y} \in \mathcal{Y} \setminus \{y_{\boldsymbol{x}}\}} d_{\boldsymbol{x}}^{\bar{y}},$$

which is the difference between the highest and the second highest label description degrees

Theorem 1. *For each $\boldsymbol{x} \in \mathcal{X}$, if the predicted label distribution satisfies the following inequality*

$$\sum_j |d_{\boldsymbol{x}}^{y_j} - \hat{d}_{\boldsymbol{x}}^{y_j}| \leq \alpha_{\boldsymbol{x}},$$

the predicted label satisfies $\hat{y}_{\boldsymbol{x}} = y_{\boldsymbol{x}}$.

Theorem 1 says that for each instance, if the L_1-norm loss of a learned LDL function is less than or equal to the LDM, the predicted label is guaranteed to equal the optimal label. Indeed, LDM indicates the **hardness** of label distribution. For an instance with large (small) LDM, the misclassification threshold is high (low) (e.g., Fig. 2(b) has a higher mis-classification threshold than that of Fig. 2(a)). To put it differently, it's less likely to mis-classify an instance with large LDM, and vice versa. Inspired by that, we can first modify the label distribution of each instance to maximize LDM and then learn the modified label distribution. For each $\boldsymbol{x}_i$, we define the **single-label distribution** $L = \{l_{\boldsymbol{x}}^{y_1}, l_{\boldsymbol{x}}^{y_2}, \cdots, l_{\boldsymbol{x}}^{y_m}\}$, where $l_{\boldsymbol{x}}^{y_j}$ equals 1 if $y_j = y_{\boldsymbol{x}}$ and 0 otherwise. Note that L has the largest LDM and the highest mis-classification threshold as well.

Corollary 1. *For each $\boldsymbol{x} \in \mathcal{X}$, and any $\rho \leq 1$, if the predicted label distribution satisfies the following inequality*

$$\sum_j |\hat{d}_{\boldsymbol{x}}^{y_j} - l_{\boldsymbol{x}}^{y_j}| \leq \rho,$$

the predicted label satisfies $\hat{y}_{\boldsymbol{x}} = y_{\boldsymbol{x}}$.

According to the above corollary, similar to Support Vector Regression (SVR) (Smola & Schölkopf, 2004), we define the **ρ-insensitive loss** for any $0 < \rho < 1$ by the following

$$|\xi|_\rho = \begin{cases} 0 & \text{if } |\xi| \leq \rho \\ (|\xi| - \rho)/(1 - \rho) & \text{otherwise,} \end{cases} \tag{3}$$

where $|\xi| = \sum_j |\hat{d}_{\boldsymbol{x}}^{y_j} - l_{\boldsymbol{x}}^{y_j}|$. It is a surrogate loss for 0/1 loss because $|\xi|_\rho$ is larger than or equal to 0 if $|\xi| \leq 1$ and is larger than 1 otherwise.

Since our goal is to learn the optimal label, it is encouraged to have the largest predicted label description degree. To achieve that, we use large margin (Cortes & Vapnik, 1995). Specifically, we encourage the label description degree of the optimal label to be larger than those of other labels by a margin ρ. Adding together the ρ-insensitive loss and large margin, we formulate the learning problem as

$$\begin{aligned} \min_{\boldsymbol{W}, \boldsymbol{\xi}, \boldsymbol{\zeta}} \; & \frac{\lambda_1}{2} \|\boldsymbol{W}\|_{\text{F}}^2 + \sum_{i=1}^{n} \frac{\boldsymbol{\xi}_i}{1 - \rho} + \lambda_2 \sum_{i,j} \frac{\boldsymbol{\zeta}_{i,j}}{\rho} \\ \text{s.t.} \quad & \|\boldsymbol{W}^\top \cdot \boldsymbol{x}_i - L_i\|_1 - \rho \leq \boldsymbol{\xi}_i, \; \boldsymbol{\xi}_i \geq 0, \\ & \boldsymbol{w}_{y_{\boldsymbol{x}_i}} \cdot \boldsymbol{x}_i - \boldsymbol{w}_j \cdot \boldsymbol{x}_i \geq \rho - \boldsymbol{\zeta}_{i,j}, \; \boldsymbol{\zeta}_{i,j} \geq 0, \\ & \text{for } i = 1, \cdots, n \text{ and } j : y_j \neq y_{\boldsymbol{x}_i}, \end{aligned} \tag{4}$$

where $\boldsymbol{W} = [\boldsymbol{w}_1, \cdots, \boldsymbol{w}_m]$ is the model parameter, $\boldsymbol{w}_{y_{\boldsymbol{x}_i}}$ is the column of $\boldsymbol{W}$ corresponding to $y_{\boldsymbol{x}_i}$, λ_1 is a regularization parameter, λ_2 is a balancing parameter, and $\boldsymbol{\xi}_i$ and $\boldsymbol{\zeta}_{i,j}$ are slack variables.

Eq. (4) jointly optimizes the ρ-insensitive loss and margin loss (Mohri et al., 2018), which can be regarded as a combination of SVR and SVM. Thereby, we call it **S**upport **V**ector **R**egression **M**achine (SVRM). Fig. 3 compares SVM, SVR, and SVRM in the case of binary classification. Fig. 3(a) shows that for SVM, the points outside the marginal hyper-planes are correctly classified. Fig. 3(b) shows that for SVR, the points inside the marginal hyper-planes are classified correctly according to Corollary 1. Fig. 3(c) shows that for SVRM, only the points lying between the marginal hyperplanes are support vectors, and other points are correctly classified by SVR or SVM. Thereby, compared with SVM and SVR, SVRM has fewer support vectors, which implies better generalization. Since SVRM learns single-label distribution, it can be applied to any SLL problems.

3.3. LDL with Adaptive Margin Loss

As discussed in Section 1, label description degrees tell the relative importance of labels. However, SVRM only concerns the label description degree of the optimal label and ignores those of other labels, which loses lots of supervised information. Indeed, the label description degrees of other labels are also important to the performance of LDL. To see that, for the example of Fig. 1, the label "DIS" has the sub-optimal label description degree, which is inferior to "ANG" but superior to other labels. The label description degrees of other labels can guide an LDL model to select the sub-optimal label (e.g., "DIS" for Fig. 1) as the predicted label when it fails to predict the optimal one, which improves the generalization.

Formally speaking, for each $\boldsymbol{x}$, we define the **sub-optimal label** by

$$y'_{\boldsymbol{x}} = \arg\max\nolimits_{\bar{y} \in \mathcal{Y} \setminus \{y_{\boldsymbol{x}}\}} d_{\boldsymbol{x}}^{\bar{y}}, \tag{5}$$

which has the second highest label description degree. Next, we define the **second label distribution margin**.

Definition 2. *For each $\boldsymbol{x}$, the second label distribution margin is defined by*

$$\beta_{\boldsymbol{x}} = \max_{\bar{y} \in \mathcal{Y} \setminus \{y_{\boldsymbol{x}}\}} d_{\boldsymbol{x}}^{\bar{y}} - \max_{\bar{y} \in \mathcal{Y} \setminus \{y_{\boldsymbol{x}}, y'_{\boldsymbol{x}}\}} d_{\boldsymbol{x}}^{\bar{y}}, \tag{6}$$

which is the difference between the second highest and the third highest label description degrees.

Theorem 2. *For each $\boldsymbol{x} \in \mathcal{X}$, if the predicted label distribution satisfies the following inequality*

$$\sum_{j : y_j \neq y_{\boldsymbol{x}}} |d_{\boldsymbol{x}}^{y_j} - \hat{d}_{\boldsymbol{x}}^{y_j}| \leq \beta_{\boldsymbol{x}}, \tag{7}$$

the predicted label satisfies $\hat{y}_{\boldsymbol{x}} = y_{\boldsymbol{x}}$ or $\hat{y}_{\boldsymbol{x}} = y'_{\boldsymbol{x}}$.

The theorem says that, for each instance, if the L_1-norm loss (*w.r.t.* all labels except the optimal one) of a learned LDL function is less than or equal to the second LDM, the predicted label equals either the optimal label or the sub-optimal one. Next, we define the $\beta_{\boldsymbol{x}}$-insensitive loss by

$$\ell_{\beta_{\boldsymbol{x}}}(\xi) = \begin{cases} 0 & \text{if } \xi \leq \beta_{\boldsymbol{x}} \\ \xi - \beta_{\boldsymbol{x}} & \text{otherwise,} \end{cases} \tag{8}$$

where $\xi = \sum_{j : y_j \neq y_{\boldsymbol{x}}} |d_{\boldsymbol{x}}^{y_j} - \hat{d}_{\boldsymbol{x}}^{y_j}|$. Since the $\beta_{\boldsymbol{x}}$-insensitive loss adapts to the second LDM of $\boldsymbol{x}$, we call it **Adaptive Margin Loss**. Next, adding the adaptive margin loss to model (4), the problem can be further cast as

$$\begin{aligned} \min_{\boldsymbol{W}, \boldsymbol{\xi}, \boldsymbol{\zeta}, \boldsymbol{\phi}} \; & \frac{\lambda_1}{2} \|\boldsymbol{W}\| + \sum_{i=1}^{n} \frac{\boldsymbol{\xi}_i}{1 - \rho} + \lambda_2 \sum_{i,j} \frac{\boldsymbol{\zeta}_{i,j}}{\rho} + \lambda_3 \sum_{i=1}^{n} \boldsymbol{\phi}_i \\ \text{s.t.} \; & \|\boldsymbol{W}^\top \cdot \boldsymbol{x}_i - L_i\|_1 - \rho \leq \boldsymbol{\xi}_i, \; \boldsymbol{\xi}_i \geq 0, \\ & \boldsymbol{w}_{y_{\boldsymbol{x}_i}} \cdot \boldsymbol{x}_i - \boldsymbol{w}_j \cdot \boldsymbol{x}_i \geq \rho - \boldsymbol{\zeta}_{i,j}, \; \boldsymbol{\zeta}_{i,j} \geq 0, \\ & \sum_{j : y_j \neq y_{\boldsymbol{x}_i}} |\boldsymbol{w}_j \cdot \boldsymbol{x}_i - d_{\boldsymbol{x}_i}^{y_j}| - \beta_{\boldsymbol{x}_i} \leq \boldsymbol{\phi}_i, \; \boldsymbol{\phi}_i \geq 0, \\ & \text{for } i = 1, \cdots, n \end{aligned} \tag{9}$$

where λ_3 is a balancing parameter, and $\boldsymbol{\phi}_i$ is a slack variable.

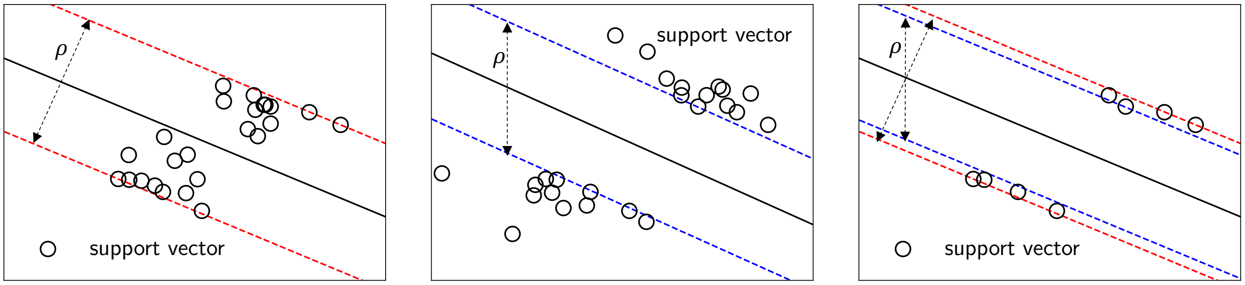
The first and the second constraints of Eq. (9) encourage our model to choose the optimal label as the predicted label. Meanwhile, the third constraint of Eq. (9) encourages our model to choose as the predicted label the sub-optimal label even if it fails to predict the optimal one. As a result, our model can sufficiently exploit the supervision information of label distribution.

3.4. Optimization Method

Eq. (9) is difficult to solve due to a larger number of constraints. Notably, there are $2n + (m-1)n$ constraints, which may overwhelm the memory limit for large datasets. Inspired by the Pegasos method (Shalev-Shwartz et al., 2011), which uses Stochastic Gradient Descent (SGD) to solve SVM efficiently, we apply SGD to optimize problem (9) as well. The details of the algorithm are presented in Algorithm 1, where line 10 calculates the sub-gradient, and line 11 updates the parameters using the sub-gradient.

Algorithm 1 SGD for solving LDLM

```
 1: Input: training set $S$, parameters $\lambda_1$, $\lambda_2$, $\lambda_3$, and $\rho$, batch size $\theta$, learning rate $\eta$, and maximum iteration $T$
 2: Output: model parameter $\boldsymbol{W}$
 3: Initialize $\boldsymbol{W}^0 = \mathbf{0}$;
 4: for $t = 1$ to $T$ do
 5:     Generate a batch $A_t \subseteq [n]$, where $|A_t| = \theta$;
 6:     $\bar{A}_t = \{i \in A_t : \sum_j |\boldsymbol{w}_j \cdot \boldsymbol{x}_i - l_{\boldsymbol{x}_i}^{y_j}| > \rho\}$;
 7:     $\hat{A}_t = \{i \in A_t : \sum_{j : y_j \neq y_{\boldsymbol{x}_i}} |\boldsymbol{w}_j \cdot \boldsymbol{x}_i - d_{\boldsymbol{x}_i}^{y_j}| > \beta_{\boldsymbol{x}_i}\}$;
 8:     for $j = 1$ to $m$ do
 9:         $A'_t = \{i \in A_t : \boldsymbol{w}_{y_{\boldsymbol{x}_i}} \cdot \boldsymbol{x}_i - \boldsymbol{w}_j \cdot \boldsymbol{x}_i < \rho\}$;
10:         $\bigtriangledown_{\boldsymbol{w}_j} = \lambda_1 \boldsymbol{w}_j + {}^1/_{1-\rho} \sum_{i \in \bar{A}_i} \mathrm{sign}(\boldsymbol{w}_j \cdot \boldsymbol{x}_i - l_{\boldsymbol{x}_i}^{y_j}) \boldsymbol{x}_i + {}^{\lambda_2}/_{\rho} \sum_{i \in A'_t} \boldsymbol{x}_i + \lambda_3 \sum_{i \in \hat{A}_t} \mathrm{sign}(\boldsymbol{w}_j \cdot \boldsymbol{x}_i - d_{\boldsymbol{x}_i}^{y_j}) \boldsymbol{x}_i$;
11:         $\boldsymbol{w}_j^t = \boldsymbol{w}_j^{t-1} - \eta \cdot \bigtriangledown_{\boldsymbol{w}_j}$;
12:     end for
13: end for
14: Return: $\boldsymbol{W}^t$
```



(a) SVM

(b) SVR

(c) SVRM

Figure 3. Comparison among SVM, SVR, and SVRM in the case of binary classification, where the solid lines and the dotted lines are the separating hyper-planes and the marginal hyper-planes, respectively. Fig. 3(a) shows SVM, where the points inside the marginal hyper-planes are support vectors. Fig. 3(b) illustrates SVR, which fits a ρ-insensitive zone. According to Corollary 1, the points inside the ρ-insensitive zone are guaranteed to be correctly classified. Fig. 3(c) demonstrates SVRM, which is a combination of SVM and SVR. Only the points between the marginal hyper-planes are not guaranteed to be correctly classified (i.e., support vectors). Compared with SVM and SVR, SVRM has fewer support vectors and better generalization.

4. Theoretical Analysis

4.1. Generalization Analysis of SVRM

For each $\boldsymbol{x}$, $y_{\boldsymbol{x}}$ is regarded as the ground-truth label. For a real-value function f, define the risk $R(f) = \mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}}[\mathbb{I}(\hat{y}_{\boldsymbol{x}} \neq y_{\boldsymbol{x}})]$, the empirical ρ-insensitive loss $\hat{R}_\rho(f) = \frac{1}{n}\sum_{i=1}^{n} |\|f(\boldsymbol{x}_i) - L_i\|_1|_\rho$, and the empirical margin loss $\tilde{R}_\rho(f) = \frac{1}{n}\sum_{i,j}\max\{\rho - f(\boldsymbol{x}_i, y_{\boldsymbol{x}_i}) + f(\boldsymbol{x}, y_j), 0\}/\rho$. For simplicity, assume $\sup_{\boldsymbol{x}} \|\boldsymbol{x}\|_2 \le r$.

Theorem 3. *Let $\mathcal{F} = \{\boldsymbol{x} \mapsto \boldsymbol{W}^\top \cdot \boldsymbol{x} : \|\boldsymbol{w}_j\|_2 \le \Lambda\}$ be the hypothesis space. Fix $1 > \rho > 0$. For any $\delta > 0$, with probability at least $1 - \delta$, the bounds hold for all $f \in \mathcal{F}$,*

$$R(f) \le \hat{R}_\rho(f) + \frac{2\sqrt{2}r\Lambda m}{(1-\rho)\sqrt{n}} + \sqrt{\frac{\log {}^1\!/\delta}{2n}},$$

$$R(f) \le \min\left\{\hat{R}_\rho(f) + \frac{2\sqrt{2}r\Lambda m}{(1-\rho)\sqrt{n}}, \tilde{R}_\rho(f) + \frac{4r\Lambda m}{\rho\sqrt{n}}\right\} + \sqrt{\frac{\log {}^2\!/\delta}{2n}}.$$

The first bound upper bounds the risk of SVR (with the ρ-insensitive loss) by the sum of three terms, where the first one is the empirical loss, the second one is a complexity term (Bartlett & Mendelson, 2002), and the last one is a by-product, which can be ignored. We extend the margin theory and support the use of SVR for classification. The second bound upper bounds the risk of SVRM by the sum of two terms, where the first one is credited to the combination of SVM and SVR, and the second one can be ignored. Theorem 3 establishes $\mathcal{O}(m/\sqrt{n})$ bounds, which admits a linear dependence on the number of classes.

For SVM, the complexity term is determined by ${}^1\!/\rho$ (Mohri et al., 2018). For SVR (with the ρ-insensitive loss), the complexity term depends on ${}^1\!/{1-\rho}$. SVRM seeks a trade-off between SVM and SVR – a larger (smaller) value of ρ increases (decreases) the complexity of SVR but decreases (increases) that of SVM. For example, as shown in Fig 5, a larger value of ρ increases the number of support vectors of SVM and decreases that of SVR.

4.2. Generalization Analysis of LDLM

Suppose that the label distribution function is the conditional probability distribution function, i.e., $d_{\boldsymbol{x}}^{y_j} = \mathbb{P}(y = y_j \mid \boldsymbol{x})$, where y is the random label variable (*w.r.t.* the conditional probability distribution). Let L_1^* be the Bayes error (Devroye et al., 2013), i.e., $L_1^* = \mathbb{P}(y_{\boldsymbol{x}} \neq y)$.

Theorem 4 (Wang & Geng (2019b))**.** *Let $\hat{d}$ be a learned LDL function. Then, the following bound holds*

$$\mathbb{P}(\hat{y}_{\boldsymbol{x}} \neq y) - L_1^* \le \mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}}\left[\sum_{\bar{y}} |\hat{d}_{\boldsymbol{x}}^{\bar{y}} - d_{\boldsymbol{x}}^{\bar{y}}|\right].$$

Theorem 4 says that the error of a learned LDL function would approach the Bayes error if it is close to the ground-truth LDL function in L_1-norm loss.

Define $\mathcal{N} = \{\boldsymbol{x} \in \mathcal{X} \mid \hat{d}_{\boldsymbol{x}}^{y_{\boldsymbol{x}}} - \hat{d}_{\boldsymbol{x}}^{y_l} < \rho, \exists l : y_l \neq y_{\boldsymbol{x}}\}$ (e.g., the zone inside the marginal hyper-planes in Fig. 3(a)), and $\mathcal{M} = \{\boldsymbol{x} \in \mathcal{X} \mid \sum_j |\hat{d}_{\boldsymbol{x}}^{y_j} - l_{\boldsymbol{x}_i}^{y_j}| > \rho\}$ (e.g., the zone outside the marginal hyper-planes in Fig. 3(b)). Note that $\hat{y}_{\boldsymbol{x}} = y_{\boldsymbol{x}}$ if $\boldsymbol{x} \notin \mathcal{N}$ or $\boldsymbol{x} \notin \mathcal{M}$. Let $\mathcal{D}_{\mathcal{N}\cap\mathcal{M}}$ be the distribution over $\mathcal{N} \cap \mathcal{M}$ (e.g., the zone between the marginal hyperplanes in Fig. 3(c)). A tighter bound can be proved.

Theorem 5. *Let $\hat{d}$ be a learned LDL function. Let $\mathcal{N}$ and $\mathcal{M}$ be defined above. Then, the following bound holds*

$$\mathbb{P}(\hat{y}_{\boldsymbol{x}} \neq y) - L_1^* \le \mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}_{\mathcal{N}\cap\mathcal{M}}}\left[\sum_{\bar{y}} |\hat{d}_{\boldsymbol{x}}^{\bar{y}} - d_{\boldsymbol{x}}^{\bar{y}}|\right].$$

Theorem 5 says that the error of our model would approach the Bayes error if the L_1-norm loss of the instances in $\mathcal{N} \cap \mathcal{M}$ approaches 0. That is, the error is only determined by the set $\mathcal{N} \cap \mathcal{M}$. Our bound is tighter than Theorem 4.

For a learned LDL function $\hat{d}$, define the empirical loss by $\hat{R}_\beta(\hat{d}) = \frac{1}{n}\sum_{i=1}^{n} \ell_\beta(\sum_{\bar{y} \neq y_{\boldsymbol{x}_i}} |\hat{d}^{\bar{y}}_{\boldsymbol{x}_i} - d^{\bar{y}}_{\boldsymbol{x}_i}|) + \beta$ for $\beta > 0$ ($\beta_{\boldsymbol{x}}$ is fixed to β for the convenience of analysis). Define $L_2^* = \mathbb{P}(y'_{\boldsymbol{x}} \neq y)$. We can prove the next theorem.

Theorem 6. *Let $\mathcal{F}$ be the hypothesis space defined in Theorem 3. Fix $1 > \rho > 0$ and $\beta \geq 0$ such that $\beta \leq \beta_{\boldsymbol{x}}$ for all $\boldsymbol{x} \in \mathcal{X}$. Then, for any $\delta > 0$, with probability at least $1 - \delta$, the following bound holds for all $f \in \mathcal{F}$*

$$\mathbb{P}(\hat{y}_{\boldsymbol{x}} \neq y) \leq \min\left\{L_1^* + \hat{R}_\rho(f) + \frac{2\sqrt{2}r\Lambda m}{(1-\rho)\sqrt{n}},\right.$$
$$\left. L_2^* + \hat{R}_\beta(f) + \frac{2\sqrt{2}m\Lambda r}{\sqrt{n}}\right\} + \sqrt{\frac{\log {}^{2}/_{\delta}}{2n}}.$$

Theorem 6 bounds the error of LDLM by the sum of two terms. The first term is due to that LDLM learns both the optimal label and the label description degrees of other labels. The second term is a by-product, which can be ignored. The bound shows the importance of both the optimal label and the label description degrees of other labels for the generalization of LDLM.

5. Experiments

5.1. Methodology

Experimental Datasets The experiments are conducted on 17 real-world datasets, characteristics of which are summarized in Table 1. In detail, the first 15 datasets are collected by Geng (2016), where the first ten (from Alpha to Spoem) are from the clustering analysis of genome-wide expression in Yeast *Saccharomyces cerevisiae* (Eisen et al., 1998), the Scene is a multi-label image dataset whose label distributions are transformed from rankings (Geng & Luo, 2014), the Gene is obtained from the research on the relation between gene and diseases (Yu et al., 2012), the Movie is collected from user ratings on movies (Geng & Hou, 2015), and the SJAFFE and SBU_3DFE are collected from JAFFE (Lyons et al., 1998) and BU_3DFE (Yin et al., 2006), respectively. The M2B (Nguyen et al., 2012) and SCUT-FBP (Xie et al., 2015) are about facial beauty perception, which are pre-processed as (Ren & Geng, 2017).

Table 1. Characteristics of the experimental datasets.

ID	Dataset	*#Examples*	*#Features*	*#Labels*
1	Alpha	2,465	24	18
2	Cdc	2,465	24	15
3	Cold	2,465	24	4
4	Diau	2,465	24	7
5	Dtt	2,465	24	4
6	Elu	2,465	24	14
7	Heat	2,465	24	6
8	Spo	2,465	24	6
9	Spo5	2,465	24	3
10	Spoem	2,465	24	2
11	Scene	2,000	294	9
12	Gene	17,892	36	68
13	Movie	7,755	1,869	5
14	SJAFFE	213	243	6
15	SBU_3DFE	2,500	243	6
16	M2B	1,240	250	5
17	SCUT-FBP	1,500	300	5

Evaluation Metrics Since we aim at improving the classification performance of LDL, the suggested LDL metrics by Geng (2016) are not used. Two metrics are adopted. The first one is 0/1 loss, i.e., $\ell_{0/1}(y_{\boldsymbol{x}}, \hat{y}) = \mathbb{I}(\hat{y} \neq y_{\boldsymbol{x}})$ ($y_{\boldsymbol{x}}$ is regarded as the ground-truth label), which indicates the classification performance of the comparing approaches. The second one is the error probability (Wang & Geng, 2019a)

$$\ell_{\mathrm{ep}}(y, \hat{y}) = \mathbb{P}(y \neq \hat{y} \mid \boldsymbol{x}) = 1 - \mathbb{P}(y = \hat{y} \mid \boldsymbol{x}) = 1 - d^{\hat{y}}_{\boldsymbol{x}},$$

where the third equation is by the assumption that label distribution function is the conditional probability function, i.e., $\mathbb{P}(y = \hat{y} \mid \boldsymbol{x}) = d^{\hat{y}}_{\boldsymbol{x}}$. Error probability indicates the generalization ability of the comparing methods.

Baselines We compare LDLM with two SLL methods (SVR and SVM) and five LDL methods (SA-BFGS, LDL-SVR, EDL-LRL, LDLFs, and LDL4C), which are as follows

- SVR (Sanchez-Fernandez et al., 2004) and SVM (Chang & Lin, 2011): SVR learns the single-label distribution, and SVM learns the optimal label.
- SA-BFGS (Geng, 2016): It applies the maximum entropy model to learn label distribution, where KL-divergence is used as the loss function.
- LDL-SVR (Geng & Hou, 2015): It adds a sigmoid transformation to the output of an SVR model to fit label distribution.
- EDL-LRL[1] (Jia et al., 2019): It exploits local label correlation by capturing low-rank structure locally when learning label distribution.
- LDLFs[2] (Shen et al., 2017): It uses the differentiable decision trees to learn label distribution, which is an ensemble method.

[1] Code: https://github.com/NJUST-IDAM/EDL-LRL.

[2] Code: https://github.com/shenwei1231/caffe-LDLForests

- LDL4C (Wang & Geng, 2019a): It's a specialized LDL algorithm for classification, where the objective mismatch is alleviated using the weighting method.

The parameters of the methods are set as follows. For SVR, SVM, and LDL-SVR, the linear kernel is applied and $C = 1$. For SVR and LDL-SVR, $\epsilon = 0.1$. For SVM, the one-vs-one strategy is used. For EDL-LRL and LDLFs, the default parameters are used. For LDL4C, we tune the parameters as suggested by Wang & Geng (2019a). For LDLM, $\lambda_1 = 0.001$, λ_2 and λ_3 are tuned from the candidate set $\{10^{-3}, \cdots, 1\}$, and $\rho = 0.01$. We tune the parameters of each method by ten-fold cross-validation. Moreover, we implement LDLM in Python and carry out the experiments on a Linux server with a 2.70GHz CPU and 62GB memory.

5.2. Results and Discussion

Results of LDLM We run each method with the best parameters for 10 times random partitions (90% training and 10% testing). Tables 2 and 3 tabulate the experimental results (mean±std.%) on the 17 datasets[3] in terms of 0/1 loss and error probability, where the best performance is highlighted in boldface for each dataset. Since LDLFs overfits on SJAFFE, the results of LDLFs on SJAFFE is not available. Furthermore, we conduct pairwise t-test at a significance of 0.05 and use •/◦ to indicate whether LDLM is statistically superior/inferior to the comparing methods. From Tables 2 and 3, we can make four observations:

1. LDLM ranks the first in 76.5% cases in terms of 0/1 loss and 58.9% cases in terms of error probability. Besides, LDLM is significantly superior to the comparing methods in 65.5% and 54% cases in terms of 0/1 loss and error probability, respectively.
2. LDLM outperforms SVM and SVR by a large margin in terms of both 0/1 loss and error probability because LDLM combines SVM and SVR. Besides, it considers label description degrees of all labels while only the optimal label is learned in SVR and SVM.
3. Compared with SA-BFGS, LDL-SVR, EDL-LRL, and LDLFs, LDLM has statistically better performance in terms of 0/1 loss and comparable performance in terms of error probability. For one thing, LDLM alleviates the objective mismatch by ρ-insensitive loss and large margin. For another, LDLM uses the adaptive margin loss to preserve generalization.
4. LDLM achieves comparable performance with LDL4C with the win/tie/lose counts of 4/13/0 and 1/16/0 in terms of 0/1 loss and error probability, respectively. LDLM has better mean performance than LDL4C with the top-1 times of 13 vs. 2 and 10 vs. 5 in terms of 0/1 loss and error probability, respectively. Besides, LDLM has theory guarantees but LDL4C is heuristic.

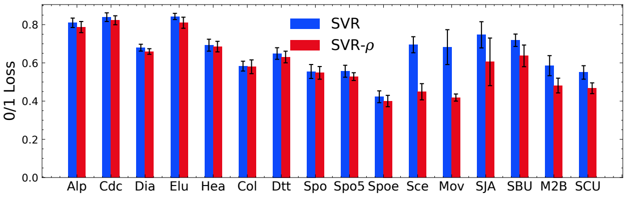


Figure 4. Performance comparison between SVR and SVR-ρ in terms of 0/1 loss. Note that SVR-ρ has better performance than SVR, which validates the effectiveness of the ρ-insensitive loss.

Our method achieves better classification performance and competitive generalization ability at the same time.

Classification Results of SVR with ρ-Insensitive Loss We denote SVR with ρ-insensitive loss by SVR-ρ (by setting $\lambda_2 = 0$ and $\lambda_3 = 0$ of LDLM). Corollary 1 and Theorem 3 well support SVR-ρ for classification. To further show that, we run SVR and SVR-ρ on the experimental datasets (the optimal label is regarded as the ground-truth label) for 10 times random data partitions (90% training and 10% testing). Fig. 4 shows the comparison results in terms of 0/1 loss. From Fig. 4, we see that SVR-ρ has much better classification performance than SVR, which shows the advantage of ρ-insensitive loss for classification.

Classification Results of SVRM To show the effectiveness of SVRM, we compare it with SVR and SVM since it can be viewed as a combination of SVM and SVR. We run the three methods with ω of the training data (ω changing from 10% to 90%) and repeat for 10 times random partitions (90% training and 10% testing). Due to limited space, we only present the comparison results on Movie, SBU_3DFE, SCUT-FBP, and M2B in Fig. 5. As shown in Fig. 5, SVRM converges fast and has better classification performance than SVM and SVR, which suggests that SVRM is a competitive method. The reason lies in that SVRM is a combination of SVM and SVR, which has fewer support vectors (as shown in Fig. 3) and better performance.

Ablation Study LDLM learns the label description degrees of all labels to preserve generalization. Here, we conduct an ablation study to show the usefulness of that. Notice that SVR-ρ is different from LDLM ($\lambda_2 = 0$) in that the latter considers the label description degrees of all labels except the optimal one. So, we compare LDLM against SVR-ρ in terms of error probability.

[3] Each dataset is denoted by its first three letters. Besides, Spo5 and Spoem are denote by Spo5 and Spoe to distinguish from Spo.

Table 2. Experimental results (mean±std.%) in terms of 0/1 loss.

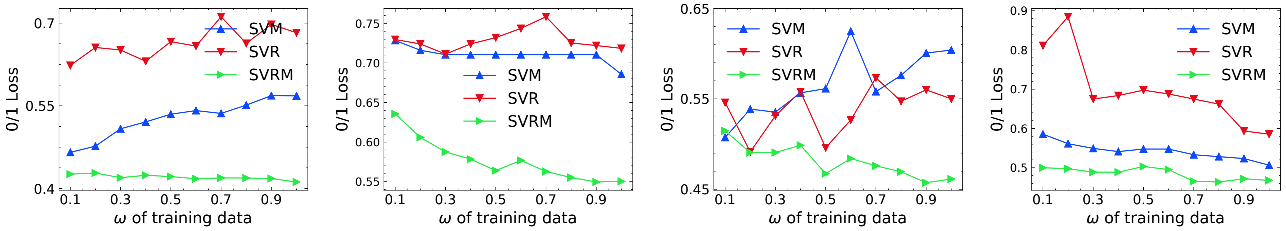
	SVR	SVM	SA-BFGS	LDL-SVR	EDL-LRL	LDLFs	LDL4C	LDLM
Alp	80.97±2.46●	78.74±2.91	89.74±2.47●	90.83±2.02●	89.70±2.37●	88.03±2.77●	78.70±2.34	**78.34±3.66**
Cdc	83.94±2.27●	82.47±2.25	82.56±2.20	82.43±1.99	82.60±2.14	82.31±1.92	81.78±2.20	**81.62±2.87**
Dia	67.95±1.79●	68.07±1.89●	69.66±3.88●	70.83±3.75●	69.90±3.96●	69.90±3.17●	66.45±1.73●	**65.27±1.09**
Elu	84.30±1.61●	81.01±3.19	90.39±1.86●	90.87±1.81●	90.43±1.82●	89.29±2.22●	**80.28±1.35**	80.32±2.55
Hea	69.25±3.13●	67.88±4.04	70.14±2.88●	70.55±2.13●	70.02±2.88●	68.03±2.07	67.54±3.21	**66.66±2.76**
Col	58.30±2.57	57.93±3.70	58.05±3.60	58.01±3.57	58.09±3.55	**56.63±3.43**	57.53±3.00	56.71±2.61
Dtt	64.91±3.03●	65.48±3.86●	63.24±2.37	63.25±2.05	63.45±2.26	**62.31±3.10**	62.68±2.72	62.43±2.75
Spo	55.50±3.65●	54.77±3.32	55.66±3.53●	56.23±3.38●	55.70±3.57●	57.77±3.76●	54.73±1.89	**54.69±3.29**
Spo5	55.62±3.13●	54.85±2.82●	57.08±2.90●	60.77±3.77●	56.84±2.81●	53.59±2.23	53.43±3.05●	**52.82±2.04**
Spoe	42.27±3.10●	49.86±4.57●	43.57±2.64●	46.33±3.11●	43.49±2.62●	42.84±2.68●	40.08±2.23	**39.51±2.18**
Sce	69.50±4.20●	41.90±3.50	61.80±3.59●	71.90±2.79●	62.10±3.27●	73.50±6.66●	41.95±2.37	**41.35±3.53**
Gen	93.23±0.38●	95.71±0.43●	95.67±0.53●	98.31±0.22●	96.03±0.48●	96.16±0.74●	92.75±0.80	**92.52±0.41**
Mov	68.26±9.12●	57.52±2.78●	45.97±1.47●	41.88±1.44	47.72±2.07●	44.33±1.76●	**40.86±1.56**	41.10±1.94
SJA	74.70±6.86●	74.70±6.86●	51.23±10.5●	80.65±8.24●	80.65±8.24●	N/A	39.39±9.80	**38.96±10.5**
SBU	71.84±3.24●	68.72±3.50●	55.88±2.56	65.68±3.55●	66.12±2.79●	63.48±3.51●	56.92±2.77●	**54.92±3.29**
M2B	58.55±5.25●	52.10±4.02●	53.87±5.55●	50.40±4.29	50.81±3.71●	48.71±2.11●	48.06±3.02●	**46.61±3.74**
SCU	55.00±3.53●	62.87±4.76●	69.80±3.32●	46.80±3.30	61.33±4.49●	46.40±2.82	46.53±2.27	**45.80±2.97**
top-1	0	0	0	0	0	2	2	**13**
w./t./l.	16/1/0	10/7/0	13/4/0	11/6/0	14/3/0	10/6/0	4/13/0	

Table 3. Experimental results (mean±std.%) in terms of error probability.

	SVR	SVM	SA-BFGS	LDL-SVR	EDL-LRL	LDLFs	LDL4C	LDLM
Alp	94.46±0.07●	94.52±0.07●	94.28±0.04●	94.28±0.03●	94.28±0.04●	94.26±0.05	94.26±0.02	**94.25±0.04**
Cdc	93.03±0.05●	92.96±0.05●	92.89±0.05●	92.88±0.06	92.89±0.05	92.88±0.06	92.88±0.05	**92.87±0.05**
Dia	84.62±0.16●	85.01±0.26●	84.30±0.17	84.31±0.14	84.30±0.16	84.28±0.13	84.28±0.10	**84.27±0.13**
Elu	92.88±0.10●	92.92±0.13●	92.62±0.06	92.61±0.05	92.62±0.06	92.62±0.05	92.60±0.05	**92.60±0.04**
Hea	82.63±0.20●	82.56±0.29●	82.43±0.20●	82.43±0.19●	82.43±0.20●	82.30±0.14	82.33±0.18	**82.29±0.19**
Col	73.10±0.26	72.97±0.35	73.01±0.32●	72.98±0.33	73.01±0.31●	**72.93±0.33**	72.96±0.31	72.96±0.33
Dtt	74.28±0.23●	74.40±0.30●	74.19±0.19●	74.20±0.20●	74.19±0.19●	74.12±0.13	74.12±0.21	**74.09±0.21**
Spo	81.06±0.43	81.01±0.43	81.07±0.42	81.08±0.41	81.08±0.41	81.23±0.43●	**81.00±0.41**	**81.00±0.43**
Spo5	65.53±0.62●	65.50±0.48●	65.43±0.49●	66.31±0.71●	65.40±0.48●	**64.68±0.48**	65.26±0.58	64.97±0.67
Spoe	47.47±0.69●	48.69±0.77●	47.06±0.54	48.32±0.86●	47.04±0.55	47.19±0.53●	47.00±0.62	**46.90±0.67**
Sce	80.06±2.48●	65.48±2.83	66.80±2.56●	66.43±2.44●	65.85±2.31	76.44±3.78●	**64.50±1.58**	65.05±2.42
Gen	98.26±0.06●	98.39±0.06	98.20±0.04	98.24±0.02	98.21±0.04	98.20±0.04	**98.16±0.06**	98.21±0.06
Mov	75.58±4.42●	71.59±0.66●	68.47±0.20●	67.65±0.27	68.86±0.46●	68.10±0.29●	**67.43±0.30**	67.58±0.33
SJA	83.64±2.18●	83.64±2.18●	76.89±1.25●	81.88±1.13●	81.88±1.13●	N/A	**75.73±2.04**	75.74±1.28
SBU	82.11±0.43●	81.17±0.55●	**76.77±0.57**	80.06±0.53●	79.91±0.50●	79.33±0.67●	77.34±0.54	76.92±0.63
M2B	61.30±3.48●	56.86±2.25●	57.68±4.05●	55.15±2.77	55.28±2.71●	54.36±2.14	53.58±2.40	**53.54±1.86**
SCU	60.20±1.81●	64.80±3.99●	71.63±2.06●	54.35±1.13●	65.05±2.87●	54.13±1.10	54.14±1.41●	**54.02±1.15**
top-1	0	0	1	0	0	2	5	**10**
w./t./l.	15/2/0	13/4/0	11/6/0	9/8/0	10/7/0	5/11/0	1/16/0	

Table 4. Performance comparison between LDLM ($\lambda_2 = 0$) and SVR-ρ in terms of error probability. We summarize the results of the Wilcoxon signed-rank test for LDLM against SVR-ρ in the last column.

	Diau	Elu	SJAFFE	Scene	SBU_3DFE	M2B	LDLM **vs.** SVR-ρ
SVR-ρ	84.68±0.14%	92.85±0.12%	79.60±2.50%	66.96±2.59%	79.53±1.42%	54.36±2.14%	**win**[9.8e-4]
LDLM	**84.27±0.13%**	**92.60±0.04%**	**76.61±2.03%**	**65.83±1.92%**	**77.64±0.42%**	**54.00±2.15%**	



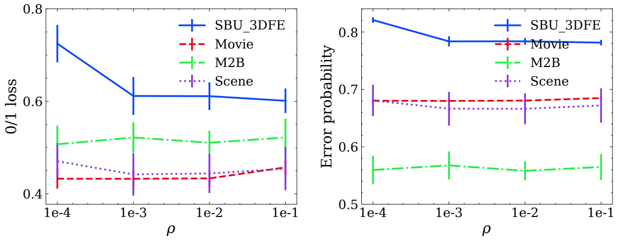
(a) Movie

(b) SBU_3DFE

(c) SCUT-FBP

(d) M2B

Figure 5. Performance comparison for SVRM against SVM and SVR in terms of 0/1 loss on Movie, SBU_3DFE, SCUT-FBP, and M2B with ω training data, where ω changes from 0.1 to 0.9.



(a) 0/1 loss

(b) Error probability

Figure 6. Sensitivity of ρ on SBU_3DFE, Movie, M2B, and Scene in terms of 0/1 loss and error probability.

Due to limited space, we report part of the results in Table 4. To show whether learning the label description degrees can significantly improve the generalization, we conduct the Wilcoxon signed-rank test (Demšar, 2006) for LDLM against SVR-ρ, which is summarized in the last column of Table 4. At a significance of 0.05, LDLM is statistically superior to SVR-ρ, which validates the effectiveness of learning the label description degrees of all labels. Besides, it also validates the advantage of label distribution compared with single-label.

Parameter Sensitivity Analysis LDLM has a key parameter, i.e., the margin ρ. To show the sensitivity of ρ, we run LDLM with ρ selecting from the candidate set $\{10^{-4}, \cdots, 10^{-1}\}$. Fig. 6 show the sensitivity of ρ in terms of 0/1 loss and error probability on SBU_3DFE, Movie, M2B, and Scene. According to Fig. 6, LDLM with $\rho = 0.01$ has better performance.

6. Conclusion

This paper proposes a new LDL method named LDLM to address the objective mismatch of LDL in classification. We first define the label distribution margin and propose SVRM to learn the optimal label. Moreover, we propose the adaptive margin loss to learn the label description degrees of other labels. Theoretical analysis shows the generalization of SVRM and the better generalization of LDLM. Experimental results show that SVRM is a competitive method, and LDLM has better classification performance than the comparing methods.

However, there are still some limitations of our method, such as SVRM is a linear model and LDLM only applies to SLL problems. In the future, we will explore 1) how to apply kernel trick to SVRM, and 2) how to extend LDLM to MLL problems.

Acknowledgements

This research was supported by the National Key Research and Development Plan of China (No. 2017YFB1002801), the National Natural Science Foundation of China (62076063), the Collaborative Innovation Center of Novel Software Technology and Industrialization, and the Collaborative Innovation Center of Wireless Communications Technology.

References

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

Chang, C.-C. and Lin, C.-J. Libsvm: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2(3):1–27, May 2011.

Cortes, C. and Vapnik, V. Support-vector networks. *Machine Learning*, 20(3):273–297, Sep. 1995.

Demšar, J. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, 7 (1):1–30, 2006.

Devroye, L., Györfi, L., and Lugosi, G. *A probabilistic theory of pattern recognition*, volume 31. Springer Science & Business Media, 2013.

Eisen, M. B., Spellman, P. T., Brown, P. O., and Botstein, D. Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences*, 95(25):14863–14868, Dec. 1998.

Gao, B., Xing, C., Xie, C., Wu, J., and Geng, X. Deep label distribution learning with label ambiguity. *IEEE Transactions on Image Processing*, 26(6):2825–2838, June 2017.

Gao, B., Zhou, H., Wu, J., and Geng, X. Age estimation using expectation of label distribution learning. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, pp. 712–718, July 2018.

Gao, W. and Zhou, Z. On the doubt about margin explanation of boosting. *Artificial Intelligence*, 203:1–18, Oct. 2013.

Geng, X. Label distribution learning. *IEEE Transactions on Knowledge and Data Engineering*, 28(7):1734–1748, July 2016.

Geng, X. and Hou, P. Pre-release prediction of crowd opinion on movies by label distribution learning. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, pp. 3511–3517, July 2015.

Geng, X. and Luo, L. Multilabel ranking with inconsistent rankers. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3742–3747, June 2014.

Geng, X., Yin, C., and Zhou, Z. Facial age estimation by learning from label distributions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(10): 2401–2412, Oct. 2013.

Geng, X., Qian, X., Huo, Z., and Zhang, Y. Head pose estimation based on multivariate label distribution. *IEEE Transactions on Pattern Analysis and Machine Intelligence, early access*, Oct. 2020. doi: 10.1109/TPAMI.2020.3029585.

Jia, X., Zheng, X., Li, W., Zhang, C., and Li, Z. Facial emotion distribution learning by exploiting low-rank label correlations locally. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9833–9842, June 2019.

Kuznetsov, V., Mohri, M., and Syed, U. Multi-class deep boosting. In *Advances in Neural Information Processing Systems*, volume 27, pp. 2501–2509, Dec. 2014.

Li, S. and Deng, W. Blended emotion in-the-wild: Multi-label facial expression recognition using crowdsourced annotations and deep locality feature learning. *International Journal of Computer Vision*, 127(6-7):884–906, Nov. 2019.

Lyons, M., Akamatsu, S., Kamachi, M., and Gyoba, J. Coding facial expressions with gabor wavelets. In *Proceedings of the IEEE Conference on Automatic Face and Gesture Recognition*, pp. 200–205, Apr. 1998.

Mohri, M., Rostamizadeh, A., and Talwalkar, A. *Foundations of Machine Learning*. Adaptive Computation and Machine Learning. MIT Press, Cambridge, MA, 2 edition, 2018. ISBN 978-0-262-03940-6.

Nguyen, T. V., Liu, S., Ni, B., Tan, J., Rui, Y., and Yan, S. Sense beauty via face, dressing, and/or voice. In *Proceedings of the 20th ACM International Conference on Multimedia*, pp. 239–248, Oct. 2012.

Ren, Y. and Geng, X. Sense beauty by label distribution learning. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, pp. 2648–2654, Aug. 2017.

Sanchez-Fernandez, M., de-Prado-Cumplido, M., Arenas-Garcia, J., and Perez-Cruz, F. SVM multiregression for nonlinear channel estimation in multiple-input multiple-output systems. *IEEE Transactions on Signal Processing*, 52(8):2298–2307, Aug. 2004.

Shalev-Shwartz, S., Singer, Y., Srebro, N., and Cotter, A. Pegasos: primal estimated sub-gradient solver for SVM. *Mathematical Programming*, 127(1):3–30, Oct. 2011.

Shen, W., Zhao, K., Guo, Y., and Yuille, A. L. Label distribution learning forests. In *Advances in Neural Information Processing Systems 30*, pp. 834–843, Dec. 2017.

Smola, A. J. and Schölkopf, B. A tutorial on support vector regression. *Statistics and Computing*, 14(3):199–222, Aug. 2004.

Vapnik, V. N. *The nature of statistical learning theory*. Springer-Verlag New York, Inc., 1995.

Wang, J. and Geng, X. Classification with label distribution learning. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pp. 3712–3718, Aug. 2019a.

Wang, J. and Geng, X. Theoretical analysis of label distribution learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 5256–5263, July 2019b.

Wu, X., Wen, N., Liang, J., Lai, Y., She, D., Cheng, M., and Yang, J. Joint acne image grading and counting via label distribution learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10641–10650, Oct. 2019.

Xie, D., Liang, L., Jin, L., Xu, J., and Li, M. Scut-fbp: A benchmark dataset for facial beauty perception. In *Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics*, pp. 1821–1826, Oct. 2015.

Yang, J., She, D., and Sun, M. Joint image emotion classification and distribution learning via deep convolutional neural network. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, pp. 3266–3272, Aug. 2017.

Yin, L., Wei, X., Sun, Y., Wang, J., and Rosato, M. J. A 3D facial expression database for facial behavior research. In *Proceedings of the 7th International Conference on Automatic Face and Gesture Recognition*, pp. 211–216, Apr. 2006.

Yu, J., Jiang, D., Xiao, K., Jin, Y., Wang, J., and Sun, X. Discriminate the falsely predicted protein-coding genes in aeropyrum pernix k1 genome based on graphical representation. *Communications in Mathematical and in Computer Chemistry*, 67:845–866, 2012.

Zhang, M., Zhang, Q., Fang, J., Li, Y., and Geng, X. Leveraging implicit relative labeling-importance information for effective multi-label learning. *IEEE Transactions on Knowledge and Data Engineering*, 33(5):2057–2070, May 2021.

Zhang, T. and Zhou, Z. Optimal margin distribution machine. *IEEE Transactions on Knowledge and Data Engineering*, 32(6):1143–1156, June 2020.