Journal of King Saud University – Computer and Information Sciences 34 (2022) 10094–10108

HOSTED BY

Contents lists available at ScienceDirect

Journal of King Saud University – Computer and Information Sciences

journal homepage: www.sciencedirect.com

A semi-supervised label distribution learning model with label correlations and data manifold exploration

Ruiqi Guo[a], Yong Peng[a,*], Wanzeng Kong[a,b], Fan Li[c]

[a] School of Computer Science and Technology, Hangzhou Dianzi University, Hangzhou 310018, China
[b] Zhejiang Key Laboratory of Brain-Machine Collaborative Intelligence, Hangzhou 310018, China
[c] CAAC Key Laboratory of Flight Techniques and Flight Safety, Civil Aviation Flight University of China, Guanghan 618307, China

ARTICLE INFO

Article history:
Received 19 July 2022
Revised 16 September 2022
Accepted 11 October 2022
Available online 15 October 2022

Keywords:
Label distribution learning
Semi-supervised learning
Joint optimization
Label correlation
Graph regularization

ABSTRACT

Label distribution learning is a novel machine learning paradigm to deal with label ambiguity, which is the generalization of the traditional single-label learning and multi-label learning paradigms. Though label distribution learning has attracted a lot of attentions recently, data sets with label distributions rather than logical labels have always been scarce and most of the existing models put emphasis mainly on the supervised learning, neglecting the utilization of unlabeled samples. In this paper, we propose a semi-supervised Label Distribution Learning model with label Correlations and data Manifold exploration (sLDLCM). On one hand, sLDLCM is a semi-supervised extension of existing label distribution learning models, which can effectively make use of both labeled and unlabeled data for better capturing the underlying data properties; on the other hand, sLDLCM jointly estimates the label distributions of unlabeled samples and the other model variables. Besides, both label correlations and local data manifold are explored in sLDLCM. Extensive experiments are conducted on six real-world data sets including two facial expression, one movie rating, two bioinformatics and one visual sentiment analysis data sets. Comparative studies demonstrate that the proposed sLDLCM model achieves better performance the state-of-the-arts in terms of five evaluation metrics.

1. Introduction

Label distribution learning is a new machine learning paradigm that portrays the importance of different labels to samples by means of label distributions. The two popular learning paradigms in machine learning and data mining, single-label learning and multi-label learning, can be considered as two special cases of label distribution learning (Tarekegn et al., 2021; Tarekegn et al., 2021; Wang et al., 2021; Wang et al., 2021). Unlike the single-label learning and multi-label learning which only tend to answer the question of which label can be used to describe the sample, label distribution learning attempts to answer the question of how should each label describe the sample. By portraying the relative importance of each label to the sample, label distribution learning can better characterize the relationship between labels and samples. For instance, in the field of facial expression recognition, the expression of a human face may be compounded by multiple basic emotions (*e.g.*, sad, anger, and fear); therefore, classifying an expression into a certain emotion by manually setting a threshold may be not accurate enough to characterize the multiple emotions. Instead, a combination of multiple emotions with different strengths is more appropriate to characterize the expression. Since the label distribution learning paradigm was introduced by Geng *et al.* (Geng, 2016; Geng, 2016), it has attracted much attention in academic community due to its powerful ability in dealing with the label ambiguity problem (Chen et al., 2018; Chen et al., 2018; Gao et al., 2017; Gao et al., 2017). As a result, lots of efforts were made in developing more competitive learning models and applying label distribution learning in diverse applications such as remote sensing (Luo et al., 2021; Luo et al., 2021), facial expression recognition (Li et al., 2020; Li et al., 2020; Wen et al., 2020; Wen et al., 2020), head pose estimation (Xu et al., 2019; Xu et al., 2019; Liu et al., 2021; Liu et al., 2021), age estimation(Liao et al., 2020; Liao et al., 2020; Si et al., 2022; Si et al., 2022) and some others (Ren and Geng, 2017; Ren and Geng, 2017).

[*] Corresponding author.
E-mail address: yongpeng@hdu.edu.cn (Y. Peng).

Peer review under responsibility of King Saud University.

Production and hosting by Elsevier

https://doi.org/10.1016/j.jksuci.2022.10.008
1319-1578/© 2022 The Author(s). Published by Elsevier B.V. on behalf of King Saud University.

Though rapid progresses were recently made in label distribution learning, there still some limitations among the existing models. A common problem is that most of them are only applicable to supervised learning paradigm, that is, the training samples should be fully labeled. However, in real-world scenario, sometimes it is hard to acquire labeled samples such as the medical data or it is time-consuming to manually annotate a large number of samples (Pattanaik et al., 2020; Pattanaik et al., 2020). Especially, in label distribution learning, it is difficult to manually determine the descriptive degree of labeling on samples one by one. Though some label enhancement methods were proposed to automatically transform single-label data sets to label distributed ones, their overall performance needs to be further improved to satisfy the real applications, leading to the scarce of labeled samples at present. Therefore, it is of great necessity to incorporate the abundant unlabeled samples into learning process(Aamir and Zaidi, 2021; Aamir and Zaidi, 2021). That is, semi-supervised extension of existing label distribution learning models is meaningful by exploring the underlying data properties of both labeled and unlabeled data.

In this paper, we propose a semi-supervised label distribution learning based on a semi-supervised least square regression formula whose regression target is enforced to satisfy the probability distribution constraints. As pointed by existing studies, exploring label correlations is generally beneficial to improve the learning performance. For example, there is a high probability that butterflies and flowers appear in an image at the same time while it is rare to have desert and river appear simultaneously. In our model, we also consider the label correlations by estimating the labels of unlabeled samples in real time. Moreover, the local invariance idea is considered to enforce that similar samples share similar label distributions. The newly formulated model is termed the semi-supervised Label Distribution Learning with label Correlations and local Manifold (sLDLCM).

Compared with the existing label distribution learning models, sLDLCM has the following contributions.

- sLDLCM is a semi-supervised label distribution learning model, which can better characterize the underlying data properties by involving both labeled and unlabeled data in model learning. To some extent, it alleviates the scarcity of labeled samples in label distribution learning.
- sLDLCM takes the label correlations into consideration. Since the calculation of label correlations requires all the labels of samples, sLDLCM jointly estimates the label distributions of unlabeled samples which in turn facilitates the calculation of label correlations.
- sLDLCM has some secondary contributions including the consideration of local data manifold and its out-of-sample prediction ability. Moreover, an efficient optimization algorithm is proposed to solve its objective function.

The remainder of this paper is structured as follows. Section 2 provides a brief introduction to the recent progresses in label distribution learning and some other techniques. Section 3 describes the model formulation and optimization method to the proposed sLDLCM model. Experiments are conducted and the analysis on experimental results are provided in Section 4. Finally, Section 5 concludes the whole paper.

Notations. In this paper, we use upper case and lower case letters to respectively denote matrices and vectors. For matrix $\mathbf{A} \in \mathbb{R}^{n\times m}$, $\mathbf{a}^i$ and $\mathbf{a}_j$ denote its ith row and its jth column, respectively. a_{ij} denotes the element in the jth column of the ith row. The ℓ_2-norm of $\mathbf{A}$ is defined as $\|\mathbf{A}\|_2 = \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{m} a_{ij}^2}$ and its $\ell_{2,1}$-norm is $\|\mathbf{A}\|_{2,1} = \sum_{i=1}^{n}\sqrt{\sum_{j=1}^{m} a_{ij}^2} = \sum_{i=1}^{n}\|\mathbf{a}^i\|_2$.

2. Revisit of label distribution learning

In this section, some related studies are briefly reviewed to serve as the background knowledge to the present work.

Traditionally, machine learning methods are difficult to characterize the level of significance for different labels because the logical label requires one sample falls into one category or another. This hard classification lacks flexibility to deal with certain issues, such as facial emotion recognition and multi-label ranking. As a new machine learning paradigm, label distribution learning can be a promising solution to the fuzzy labeling problem (Geng et al., 2013; Geng et al., 2013; Geng et al., 2014; Geng et al., 2014). In the past decade, label distribution learning has been extensively studied, whose theoretical foundation has been gradually consolidated (Geng, 2016; Geng, 2016).

Most of the existing label distribution learning algorithms can be classified into three categories. The first one is called problem transformation (PT); that is, the involved models attempt to transform the labeled distribution learning problem into a single-labeled learning problem. For example, PT-SVM transforms the training samples into weighted single-labeled samples, and then resamples the training set to generate a new training set based on the weights of each sample, and finally predicts the probability of each label as the label distribution of the sample (Geng, 2016; Geng, 2016). The second class is called algorithm adaptation (AA), which extends the traditional algorithm to a labeled distribution learning algorithm by modifying the constraints. The representative one is the AA-kNN (Wang and Geng, 2019; Wang and Geng, 2019), which is adapted from k nearest neighbors and uses the mean of the label distributions of the k nearest neighbors of a sample to predict the label distribution of that sample. Building on the decision tree, Shen et al. proposed label distribution learning forests(Shen et al., 2017; Shen et al., 2017), which simulates the label distribution by mixing the predictions of the leaf nodes and defines a strictly descending loss function based on the distribution. For the third class, we can generally term the involved algorithm as specialized algorithms (SA), which were designed to fit the label distribution problem better by exploiting certain properties inherent to label distribution learning such as SA-IIS (Geng et al., 2014; Geng et al., 2014) and SA-BFGS (Geng, 2016; Geng, 2016). SA-IIS assumes the label distribution model to be the maximum entropy model, whose objective function can be optimized by an improved iterative scaling (IIS) strategy. SA-BFGS further improvements the performance of SA-IIS by using the quasi-Newton method BFGS to circumvent the explicit calculation of an inverse Hessian matrix, leading to improved efficiency (Bollapragada et al., 2018; Bollapragada et al., 2018).

These algorithms can better identify the significance of different labels, but they usually require a high volume of supervised information. The performance of the algorithms degrades faster in the absence of labeled data. Therefore, the involvement of unlabeled data in the learning process of the model may result in a large enhancement of the model.

Generally, few label distribution learning models were investigated to involve the unlabeled data into learning or estimate the label information of unlabeled samples in model optimization. Below are three existing ones to our knowledge. In (Hou et al., 2017; Hou et al., 2017), an semi-supervised adaptive LDL (SALDL) was proposed to solve the scarcity of labeled samples in facial age estimation. However, the specially designed Gaussian distribution is only applicable for modeling the facial aging process, which makes SALDL is not a general semi-supervised label distribution model and therefore is not flexible enough in dealing with other tasks. Xu and Zhou proposed an incomplete label distribution learning (IncomLDL) model which employs a trace norm minimization objective function to explore the label correlations and is optimized by a proximal gradient descend (Xu and Zhou, 2017; Xu and Zhou, 2017). Further, the GRME (fragmentary LDL algorithm via graph regularized maximum entropy criteria) was proposed for recovering missing label factors by making better use of inter-label correlations via the graph Laplacian matrix (Xu et al., 2021; Xu and Zhou, 2017). GRME is based on two assumptions; one is that the reconstructed label matrix should be consistent with the missing label matrix, and the other is if two data points are close to each other in the feature space, they should have similar label distributions.

Due to that some labels have a high probability of appearance simultaneously, studying potential correlations between labels can better model the label distributions(Ma et al., 2020; Ma et al., 2020). Recently, researchers have attempted to improve model performance by taking label correlations into account. Jia et al. encoded the label correlation as a distance and uses the distance matrix for promoting model convergence (Jia et al., 2018; Jia et al., 2018; Jia et al., 2021; Jia et al., 2021). In (Ren et al., 2019; Ren et al., 2019), a LDL model by leveraging label-specific features (LDLSF) was proposed, in which the Pearson's coefficient was used to measure label correlation. To be specific, l_1-norm and $l_{2,1}$-norm regularization terms were used to respectively learn label-specific and label-common features.

In the work of this paper, we intend to construct a graph Laplacian matrix reflecting the relationship between samples using the feature information of labeled and unlabeled samples. To utilize the label correlations better, we also constrain the model from the output. Thus we can use the data information more effectively and predict the label distribution of the samples more accurately while solving the data scarcity problem.

3. The proposed sLDLCM model

In this section, we introduce the model formulation and optimization of sLDLCM. Besides, some discussions on its convergence property, computational complexity are provided.

3.1. Model Formulation

In semi-supervised setting, we are often given l labeled samples, $\mathbf{X}_L = [\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_l] \in \mathbb{R}^{d \times l}$, and u unlabeled samples, $\mathbf{X}_U = [\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_u] \in \mathbb{R}^{d \times u}$. Here $n = l + u$ is the number of samples and d is the feature dimensionality. We denote $\mathbf{X} = [\mathbf{X}_L, \mathbf{X}_U] \in \mathbb{R}^{d \times n}$ is the complete data collection matrix in which the i-th column $\mathbf{x}_i \in \mathbb{R}^d$ is the i-th sample. Accordingly, we have the label distribution matrix $\mathbf{Y}_L = [\mathbf{y}_1; \mathbf{y}_2; \cdots; \mathbf{y}_l] \in \mathbb{R}^{l \times c}$ corresponding to $\mathbf{X}_L$ in which the i-th row $\mathbf{y}_i \in \mathbb{R}^{1 \times c}$ corresponds to the label distribution of sample $\mathbf{x}_i$ and c is the number of labels. Similarly, we denote $\mathbf{Y} = [\mathbf{Y}_L; \mathbf{Y}_U] \in \mathbb{R}^{n \times c}$ as the complete label distribution matrix corresponding to $\mathbf{X}$ but $\mathbf{Y}_U \in \mathbb{R}^{u \times c}$ is unknown. Based on the definition of label distribution learning, the elements in each row of $\mathbf{Y}$ satisfy the two constraints of non-negativity and row-normalization, *i.e.*, $y_{ij} \geqslant 0$ and $\sum_{j=1}^{c} y_{ij} = 1$. Here, y_{ij} is no longer the predicted probability, but the proportion of the j-th label in the full description of the i-th sample (Geng, 2016; Geng, 2016). Then, sLDLCM aims to estimate $\mathbf{Y}_U \in \mathbb{R}^{u \times c}$ as accurate as possible given $\mathbf{X}$ and $\mathbf{Y}_L$.

In Fig. 1, we show the overall framework of our proposed sLDLCM model. From this figure, we find that sLDLCM has three important factors. First, both unlabeled and labeled samples are involved in the sLDLCM model learning process for better capturing the underlying data properties. Second, sLDLCM takes both label correlations and local data manifold into consideration. Third, the label distributions corresponding to the unlabeled samples are jointly optimized with the other model variables. Below we translate these factors into mathematical expressions, which finally formulates the objective function of sLDLCM.

Assuming that the feature space and the label space are linearly correlated, we use the least square regression to model the relationship between them for simplicity. Mathematically, the data matrix $\mathbf{X}$ and the label distribution matrix $\mathbf{Y}$ is connected by a projection matrix $\mathbf{W}$ such that the approximation error should be minimized. That is,

$$\begin{aligned} &\min_{\mathbf{W}, \mathbf{Y}_U} \tfrac{1}{2} \|\mathbf{X}^T \mathbf{W} - \mathbf{Y}\|_2^2, \\ &s.t. \quad \mathbf{Y}_U \geqslant \mathbf{0}, \mathbf{Y}_U \mathbf{1}_u = \mathbf{1}_c, \end{aligned} \tag{1}$$

where $\mathbf{1}_u \in \mathbb{R}^u$ and $\mathbf{1}_c \in \mathbb{R}^c$ are two all-one column vectors. $\mathbf{0} \in \mathbb{R}^{u \times c}$ is an all-zero matrix which shares the same size with $\mathbf{Y}_U$. Obviously, objective (1) unifies the estimation of projection matrix and the label distribution matrix of unlabeled samples together in a succinct form, which is an elementary implementation of semi-supervised label distribution learning.

To enhance the learning ability of objective (1), we further consider two additional properties of label-correlations and local data manifold. Inspired by (Ren et al., 2019; Ren et al., 2019), we realize that each label might correlate several specific features, and also some labels might share some common features. Then, we consider $\mathbf{W}$ as the superposition of two matrices $\mathbf{P}$ and $\mathbf{Q}$, based on which the ℓ_1-norm and $\ell_{2,1}$-norm are respectively imposed to identify label-specific and label-shared features. Then, problem (1) has now been transformed as

$$\begin{aligned} &\min_{\mathbf{P}, \mathbf{Q}, \mathbf{Y}_U} \tfrac{1}{2} \|\mathbf{X}^T (\mathbf{P} + \mathbf{Q}) - \mathbf{Y}\|_2^2 + \lambda_1 \|\mathbf{P}\|_1 + \lambda_2 \|\mathbf{Q}\|_{2,1} \\ &s.t. \quad \mathbf{Y}_U \geqslant \mathbf{0}, \mathbf{Y}_U \mathbf{1}_u = \mathbf{1}_c, \end{aligned} \tag{2}$$

where $\mathbf{P}$ is the weight matrix constrained by l_1-regularization and $\mathbf{Q}$ is the weight matrix constrained by $l_{2,1}$-regularization. λ_1 and λ_2 are regularization parameters to balance the impacts of different terms in (2).

Further, it should be taken into account that label correlations strongly affects the algorithm performance. We examine the impact of label correlations from two differing perspectives. On the one hand, we attempt to establish a relationship between each label to constrain the output label, so that higher correlation between two labels indicates the more similar output of these two labels. On the other hand, we injects prior knowledge on some particular applications into the model by building a weighted graph in the feature space.

$$\begin{aligned} \min_{\mathbf{P}, \mathbf{Q}, \mathbf{Y}_U} \frac{1}{2} &\|\mathbf{X}^T (\mathbf{P} + \mathbf{Q}) - \mathbf{Y}\|_2^2 + \lambda_1 \|\mathbf{P}\|_1 + \lambda_2 \|\mathbf{Q}\|_{2,1} \\ &+ \frac{\lambda_3}{2} \sum_{i=1}^{n} \left\{ \sum_{j,k=1}^{c} r_{jk} \left(\mathbf{x}_i^T (\mathbf{p}_j + \mathbf{q}_j) - \mathbf{x}_i^T (\mathbf{p}_k + \mathbf{q}_k) \right)^2 \right\} \\ &+ \frac{\lambda_4}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} s_{ij} \|\mathbf{y}^i - \mathbf{y}^j\|_2^2, \ s.t. \quad \mathbf{Y}_U \geqslant \mathbf{0}, \mathbf{Y}_U \mathbf{1}_u = \mathbf{1}_c. \end{aligned} \tag{3}$$

Here r_{jk} is the Pearson's coefficient to measure the correlation between the j-th and k-th labels, which respectively correspond to the j-th column and k-th column of $\mathbf{Y}$. Obviously, if r_{jk} is large, then the discrepancy between $\mathbf{x}_i^T (\mathbf{p}_j + \mathbf{q}_j)$ and $\mathbf{x}_i^T (\mathbf{p}_k + \mathbf{q}_k)$ should be small. s_{ij} is an graph affinity matrix to measure the similarity between the i-th and j-th samples. s_{ij} is defined as

$$s_{ij} = \begin{cases} \exp\left(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{2\sigma^2} \right), & \text{if} \quad i \neq j, \\ 0, & \text{otherwise}, \end{cases} \tag{4}$$

where the bandwidth parameter σ is set to one in the following experiments. Since $\mathbf{y}^i \in \mathbb{R}^{1 \times c}$ and $\mathbf{y}^j \in \mathbb{R}^{1 \times c}$ are label distributions

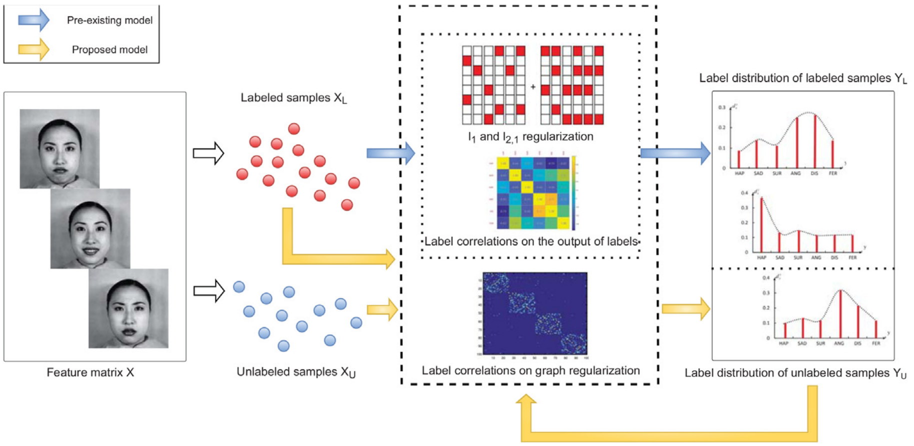

Pre-existing model
Proposed model
Feature matrix X
Labeled samples X$_L$
Unlabeled samples X$_U$
l$_1$ and l$_{2,1}$ regularization
Label correlations on the output of labels
Label correlations on graph regularization
Label distribution of labeled samples Y$_L$
Label distribution of unlabeled samples Y$_U$
HAP SAD SUR ANG DIS FER

Fig. 1. The overall framework of our proposed sLDLCM model.

respectively corresponding to samples $\mathbf{x}_i$ and $\mathbf{x}_j$, they are enforced to be similar if s_{ij} is large. It is obvious that the fourth and fifth items in objective (3) constrains the label distribution matrix from both horizontal and vertical dimensions.

To facilitate the following optimization process, we convert problem (3) into the following compact matrix form

$$\begin{aligned}&\min_{\mathbf{W},\mathbf{P},\mathbf{Q},\mathbf{Y}_U} \tfrac{1}{2}\|\mathbf{X}^T\mathbf{W}-\mathbf{Y}\|_2^2+\lambda_1\|\mathbf{P}\|_1+\lambda_2\|\mathbf{Q}\|_{2,1}\\ &+\tfrac{\lambda_3}{2}\mathrm{Tr}(\mathbf{X}^T\mathbf{W}(\mathbf{T}-\mathbf{R})\mathbf{W}^T\mathbf{X})+\tfrac{\lambda_4}{2}\mathrm{Tr}(\mathbf{W}^T\mathbf{X}(\mathbf{D}-\mathbf{S})\mathbf{X}^T\mathbf{W})\\ &s.t.\quad \mathbf{W}=\mathbf{P}+\mathbf{Q},\mathbf{Y}_U\geqslant\mathbf{0},\mathbf{Y}_U\mathbf{1}_u=\mathbf{1}_c,\end{aligned}\qquad (5)$$

where $\mathbf{T}$ is a diagonal matrix and $t_{ii}=\sum_{j=1}^n r_{ij}$, $\mathbf{D}$ is also a diagonal degree matrix and $d_{ii}=\sum_{j=1}^n \mathbf{s}_{ij}$.

3.2. Model Optimization

There are multiple variables, *i.e.*, $\mathbf{W},\mathbf{P},\mathbf{Q}$ and $\mathbf{Y}_U$, in the objective function (5) of the proposed sLDLCM model. Below we propose to optimize it under the alternating direction method of multipliers framework (ADMM) (He et al., 2020; He et al., 2020).

- Update $\mathbf{W}$. The augmented Lagrangian function $\mathscr{L}(\mathbf{W})$ with respect to variable $\mathbf{W}$ in the t-th iteration is

$$\begin{aligned}&\min_{\mathbf{W}}\frac{1}{2}\|\mathbf{X}^T\mathbf{W}-\mathbf{Y}\|_2^2+\langle\Gamma,\mathbf{W}-\mathbf{P}-\mathbf{Q}\rangle+\frac{\mu}{2}\|\mathbf{W}-\mathbf{P}-\mathbf{Q}\|_2^2\\ &+\frac{\lambda_3}{2}\mathrm{Tr}(\mathbf{X}^T\mathbf{W}(\mathbf{T}-\mathbf{R})\mathbf{W}^T\mathbf{X})+\frac{\lambda_4}{2}\mathrm{Tr}(\mathbf{W}^T\mathbf{X}(\mathbf{V}-\mathbf{A})\mathbf{X}^T\mathbf{W}).\end{aligned}\qquad (6)$$

By taking the derivative of the above equation with respect to $\mathbf{W}$ and setting it to zero, we have

$$\begin{aligned}&\mathbf{X}(\mathbf{X}^T\mathbf{W}-\mathbf{Y})+\mathbf{\Gamma}+\mu_t(\mathbf{W}-\mathbf{P}-\mathbf{Q})+\lambda_3\mathbf{X}\mathbf{X}^T\mathbf{W}(\mathbf{T}-\mathbf{R})\\ &+\lambda_4\mathbf{X}(\mathbf{V}-\mathbf{A})\mathbf{X}^T\mathbf{W}=\mathbf{0},\end{aligned}\qquad (7)$$

which is equivalent to

$$\begin{aligned}&\mathbf{W}(\mathbf{I}+\lambda_3(\mathbf{T}-\mathbf{R}))+(\mathbf{X}\mathbf{X}^T)^{-1}(\mu_t\mathbf{I}+\lambda_4\mathbf{X}(\mathbf{V}-\mathbf{A})\mathbf{X}^T)\mathbf{W}\\ &=(\mathbf{X}\mathbf{X}^T)^{-1}(\mathbf{X}\mathbf{Y}+\mu_t(\mathbf{P}+\mathbf{Q})-\mathbf{\Gamma}).\end{aligned}\qquad (8)$$

We assume that the null space of data $\mathbf{X}$ has been removed and then $\mathbf{S}_t=\mathbf{X}\mathbf{X}^T$ is invertible. This is a standard Sylvester equation and can be solved efficiently (Wang et al., 2016; Wang et al., 2016).

- Update $\mathbf{P}$. The Lagrangian function $\mathscr{L}(\mathbf{P})$ with respect to variable $\mathbf{P}$ in the t-th iteration is

$$\begin{aligned}\mathbf{P}^{t+1}&=\arg\min_{\mathbf{P}}\lambda_1\|\mathbf{P}\|_1+\frac{\mu_k}{2}\|\mathbf{W}^{t+1}-\mathbf{P}-\mathbf{Q}^t\|_2^2\\ &+\left\langle\Gamma^t,\mathbf{W}^{t+1}-\mathbf{P}-\mathbf{Q}^t\right\rangle.\\ &=\arg\min_{\mathbf{P}}\frac{\lambda_1}{\mu_k}\|\mathbf{P}\|_1+\frac{1}{2}\left\|\mathbf{P}-(\mathbf{W}^{t+1}-\mathbf{Q}^t+\frac{\mathbf{\Gamma}^t}{\mu_t})\right\|_2^2.\end{aligned}\qquad (9)$$

The solution to the above problem is shown below

$$\mathbf{P}^{t+1}=\mathscr{S}_{\frac{\lambda_1}{\mu_t}}\left[\mathbf{W}^{t+1}-\mathbf{Q}^t+\frac{\mathbf{\Gamma}^t}{\mu_t}\right],\qquad (10)$$

where $\mathscr{S}(\cdot)$ is the soft shrinkage operator defined in A.

- Update $\mathbf{Q}$. The Lagrangian function $\mathscr{L}(\mathbf{Q})$ with respect to variable $\mathbf{Q}$ in the t-th iteration is

$$\begin{aligned}\mathbf{Q}^{t+1}&=\arg\min_{\mathbf{Q}}\lambda_2\|\mathbf{Q}\|_{2,1}+\frac{\mu}{2}\|\mathbf{W}^{t+1}-\mathbf{P}^{t+1}-\mathbf{Q}^t\|_2^2\\ &+\left\langle\Gamma,\mathbf{W}^{t+1}-\mathbf{P}^{t+1}-\mathbf{Q}^t\right\rangle.\\ &=\arg\min_{\mathbf{Q}}\frac{\lambda_2}{\mu_t}\|\mathbf{Q}\|_{2,1}+\frac{1}{2}\left\|\mathbf{Q}-(\mathbf{W}^{t+1}-\mathbf{P}^{t+1}+\frac{\mathbf{\Gamma}^t}{\mu_t})\right\|_F^2.\end{aligned}\qquad (11)$$

Therefore, the updating rule for variable $\mathbf{Q}$ is

$$\mathbf{Q}^{t+1}=\mathbf{\Omega}_{\frac{\lambda_2}{\mu_k}}\left[\mathbf{W}^{t+1}-\mathbf{P}^{t+1}+\frac{\mathbf{\Gamma}^t}{\mu_t}\right],\qquad (12)$$

which is the $\ell_{2,1}$-norm minimization operator defined in A.

- Update $\mathbf{Y}_U$. We propose to optimize $\mathbf{Y}_U$ row-wisely since problem (1) can be decoupled for each $i|_{i=l+1}^{l+u}$. By define $\mathbf{M}\triangleq\mathbf{X}^T\mathbf{W}$, for each i, we have

$$\min_{\mathbf{y}^i\geqslant\mathbf{0},\mathbf{y}^i\mathbf{1}=1}\frac{1}{2}\|\mathbf{m}^i-\mathbf{y}^i\|_2^2,\qquad (13)$$

which defines an Euclidean projection on a simplex constraint. Here, $\mathbf{m}^i$ and $\mathbf{y}^i$ denote the i-th row of $\mathbf{M}$ and $\mathbf{Y}$, respectively. The optimization procedure to problem (13) is provided in Algorithm 1 whose detailed derivation is provided in B.

```
Algorithm 1: The algorithm to solve objective function (13)
Input: vector $\mathbf{m}_i\in\mathbb{R}^c$;
Output: vector $\mathbf{y}_i\in\mathbb{R}^c$.
1: Compute $\mathbf{q}=\mathbf{m}_i-\frac{\mathbf{1}\mathbf{1}^T}{c}\mathbf{m}_i+\frac{1}{c}\mathbf{1}_c$;
2: Use Newton's method to obtain the root $\bar{\beta}^*$ of (B.13);
3: Obtain the optimal solution $y_{ij}^*=(q_j-\bar{\beta}^*)_+$ for $j=1,\cdots,c$;
```

Based on the above analysis, we summarize the optimization procedure to sLDLCM model objective function in Algorithm 2.

Algorithm 2: The optimization of sLDLCM model objective function (5).

```
Input: Data matrix $\mathbf{X} = [\mathbf{X}_L; \mathbf{X}_U] \in \mathbb{R}^{d\times n}$, the label distribution
    matrix $\mathbf{Y}_L \in \mathbb{R}^{l\times c}$, parameters $\lambda_1$, $\lambda_2$, $\lambda_3$ and $\lambda_4$; $\varepsilon = 10^{-8}$,
    $\mu = 10^{-6}$, $\mu_{\max} = 10^8$ and $\rho = 1.1$;
Output: The estimated label distribution matrix $\mathbf{Y}_U \in \mathbb{R}^{u\times c}$;
Initialize the similarity matrix $\mathbf{S}$ according to (4),
    $\mathbf{Y}_U = \mathbf{1}_u\mathbf{1}_u^T./c$, and $\mathbf{P}$, $\mathbf{Q}$ randomly;
Calculate the label correlation matrix $\mathbf{R}$ according to
    $\mathbf{Y} = [\mathbf{Y}_L; \mathbf{Y}_U]$;
while not converged do
    Update $\mathbf{W}^{t+1}$ by solving the Sylvester Eq. (8);
    Update $\mathbf{P}^{t+1}$ by (10);
    Update $\mathbf{Q}^{t+1}$ by (12);
    Update the $i|_{i=l+1}^{n}$ row of $\mathbf{Y}^{t+1}$ by solving problem (13)
      with Algorithm 1;
    Update the label correlation matrix $\mathbf{R}$ based on $\mathbf{Y}^{t+1}$;
    Update Lagrangian multiplier $\mathbf{\Gamma}$ by
      $\mathbf{\Gamma}^{t+1} = \mathbf{\Gamma}^{t} + \mu_t(\mathbf{W}^{t+1} - \mathbf{P}^{t+1} - \mathbf{Q}^{t+1})$;
    Update $\mu$ by $\mu^{t+1} = \min(\mu_{\max}, \rho\mu^t)$;
    Check the convergence condition $\|\mathbf{W} - \mathbf{P} - \mathbf{Q}\|_\infty < \varepsilon$;
end while
```

3.3. Discussions

Below we provide a brief introduction of the computational complexity and convergence property of sLDLCM.

On the computational complexity analysis, we perform the analysis based on the big $\mathcal{O}$ notation. Obviously, the main complexity comes from the loop in Algorithm 2; specifically, the four blocks respectively updating variables $\mathbf{W}, \mathbf{P}, \mathbf{Q}$ and $\mathbf{Y}_U$. The complexity of updating $\mathbf{W}$ is $\mathcal{O}(nd^2 + d^3)$ as it needs to solve a standard Sylvester equation. Updating $\mathbf{P}$ by soft shrinkage costs $\mathcal{O}(dc)$ complexity while it is $\mathcal{O}(dc^2)$ in updating $\mathbf{Q}$ by $\ell_{2,1}$-norm minimization. Since the complexity of updating each row of $\mathbf{Y}_U$ by Algorithm 1 is $\mathcal{O}(c)$, we need $\mathcal{O}(uc)$ to update $\mathbf{Y}_U$. Besides, the complexity of refreshing $\mathbf{R}$ is $\mathcal{O}(c^2)$. Assuming that the number of iterations of Algorithm 2 is t, the computational complexity is $\mathcal{O}(t(nd^2 + d^3 + dc^2 + dc + uc + c))$. Considering that in usual cases we have $n \approx u > d \gg c$ in semi-supervised learning, the overall complexity of optimizing sLDLCM objective function is $\mathcal{O}(tnd^2)$.

Regarding the convergence property of the ADMM algorithm, it has been well investigated when the number of blocks (*i.e.*, unknown variables) is at most two. However, since in Algorithm 2 there are four blocks corresponding to the four variables, $\mathbf{W}, \mathbf{P}, \mathbf{Q}$ and $\mathbf{Y}$ and the objective function (5) is not smooth, it is theoretically difficult to prove the convergence of our proposed algorithm. In reality, it is expected that sLDLCM has good convergence properties according to the analysis below. First, there is a unique solution in updating $\mathbf{W}$ each time based on (1). Second, we have analytical solutions when updating $\mathbf{P}$ and $\mathbf{Q}$, respectively. Third, when updating each row of $\mathbf{Y}_U$, the two involved Lagrangian multipliers can be uniquely determined, leading to its unique solution. Therefore, we declare that the convergence of the optimization procedure in Algorithm 2 can be guaranteed. In the experiments, we will show the monotonically decreasing of objective function values in terms of the number of iterations.

Proposition 1. The Sylvester Eq. (8) has a unique solution.

Proof. The left coefficient matrix is $(\mathbf{XX}^T)^{-1}(\mu_t\mathbf{I} + \lambda_4\mathbf{X}(\mathbf{V} - \mathbf{A})\mathbf{X}^T)$ positive definite when the null space of $\mathbf{X}$ is removed. Then, its eigenvalues are positive, *i.e*, $\alpha_i > 0$. Similarly, the right coefficient matrix $\mathbf{I} + \lambda_3(\mathbf{T} - \mathbf{R})$ is also positive definite; therefore, its eigenvalues are also positive, *i.e.*, $\beta_j > 0$. Hence, for any eigenvalues of both coefficient matrices, we have $\alpha_i + \beta_j > 0$. According to (Wang et al., 2016; Wang et al., 2016), the Sylvester Eq. (8) has a unique solution.

4. Experiments

In this section, in order to demonstrate the effectiveness of the proposed sLDLCM model, we compare it with several state-of-the-art models on six benchmark label distributed data sets.

4.1. Data Sets

Six data sets are used in the following experiments including s-JAFFE, SBU_3DFE, Movie, Yeast-cold, Yeast-dtt and Twitter_LDL, which are briefly described as follows. The data set s-JAFFE consisted of 213 grayscale facial expression images. The original size of the image is 256*256 pixels, and the eyes of each image are placed in the same position by cropping. The features of the image are extracted by the Local Binary Patterns method (LBP), where we set the radius to 2 and the number of neighbors to 16. The extracted 243-dimensional LBP histogram is used as the final feature. Each image is described by Ekman's 6 basic emotion labels, namely anger, disgust, fear, happiness, sadness, and surprise. Each image was rated by 60 investigators on the 6 basic emotion labels using score on 0–5 scale, and the mean score of each emotion indicates the emotion label. The normalized intensities of 6 basic emotion labels constitute the label distribution of a image. Similar to s-JAFFE, the data set SBU_3DFE consists of 2500 facial expression images and is scored in the same manner as for s-JAFFE by 23 investigators. The Movie data set contains the rating information of 7755 Netflix movies. Features contain categorical attributes (e.g., first actor, second actor, country, language, director and music production, etc.) and numerical attributes (e.g., year, duration and Budget, etc.). For the categorical attributes, we set a threshold

Table 1
Statistics of six real-world data sets

Name	Samples	Features	Labels	Domains
s-JAFFE	213	243	6	Facial expression
SBU_3DFE	2,500	243	6	Facial expression
Movie	7,755	1,869	5	Movie rating
Yeast-cold	2,465	24	4	Bioinformatics
Yeast-dtt	2,465	24	4	Bioinformatics
Twitter_LDL	**10,045**	**168**	**8**	**Visual sentiment analysis**

Table 2
Evaluation metrics of label distribution learning.

Name	Formula
Sϕrensen ↓	$\frac{\sum_{j=1}^{c}\lvert P_j-Q_j\rvert}{\sum_{j=1}^{c}\lvert P_j+Q_j\rvert}$
Kullback–Leibler($K-L$) ↓	$\sum_{j=1}^{c}P_j\ln\frac{P_j}{Q_j}$
Chebyshev ↓	$\max_j\lvert P_j-Q_j\rvert$
Intersection ↑	$\sum_{j=1}^{c}\min(P_j,Q_j)$
Cosine ↑	$\frac{\sum_{j=1}^{c}P_jQ_j}{\sqrt{\sum_{j=1}^{p}P_j^2}\sqrt{\sum_{j=1}^{p}Q_j^2}}$

value θ (typically 5 or 10) and categories with occurrences lower than θ are reassigned a new value of "Other", which allows filtering out categories with limited impact. Subsequently, the categorical attributes are represented by a unique heat code encoding and the numerical attributes are normalized. The final feature will be represented as a vector of 1869 dimensions. A total of 478,656 viewers contributed 54,242,292 ratings on five scales to the Movie data set. Each rating scale scores into five levels. The normalized intensities of 5 rating criteria constitute the label distribution of a movie. The Yeast-cold and Yeast-dtt data sets contain 2465 genes collected from the experiment on the Saccharomyces cerevisiae, which are represented by a 24-dimensional vector. Gene labels were obtained by evaluating gene expression levels at discrete time points in the experiment. The normalized intensities of 4 gene expression levels constitute the label distribution of a genes. The Twitter_LDL data set was collected from the Twitter website and contains 10045 real images. The label of each image consists of 8 emotions (anger, happiness, awe, satisfaction, disgust, excitement, fear and sadness), 8 viewers rate the strength of the above 8 emotions, and finally the ratings of all viewers are combined to generate the label distribution. We also use the LBP method to extract the features of the original image, and then use the Principal Component Analysis to reduce it to 168 dimensions.

Some important characteristics of these five data sets are summarized in Table 1.

4.2. Experimental settings

Five evaluation metrics are used to quantitatively compare the performance of different label distribution learning models, whose definitions are provided in Table 2. These metrics can be divided into two groups. The former three metrics belong to one group,

Table 3
The learning results of the nine compared models under the 10-fold cross-validation paradigm. ↑ (↓) indicates that the higher (lower) the value, the better (worse) the performance.

Measure	Algorithm	s-JAFFE	SBU_3DFE	Movie	Yeast-cold	Yeast-dtt	**Twitter_LDL**
Sϕrensen↓	sLDLCM	**0.1089**	**0.1342**	**0.1659**	**0.0587**	**0.0413**	**0.3757**
	AA-kNN	0.1285	0.1402	0.1752	0.0593	0.0421	**0.3921**
	PT-SVM	0.1578	0.1473	0.2165	0.0657	0.0440	**0.5489**
	SA-BFGS	0.1384	0.1364	0.1772	0.0592	0.0418	**0.3842**
	CPNN	0.1413	0.1609	0.1877	0.0604	0.0423	**0.4177**
	LDL-SVR	0.1386	0.1484	0.1678	0.0592	0.0416	**0.3965**
	LDLSF	0.1128	0.1356	0.1817	0.0591	0.0417	**0.3765**
	IncomLDL	0.1411	0.1442	0.1848	0.0608	0.0420	**0.4006**
	GRME	0.1255	0.1400	0.1686	0.0588	0.0420	**0.3809**
$K-L$↓	sLDLCM	**0.0408**	**0.0591**	**0.1039**	**0.0121**	**0.0061**	**0.7024**
	AA-kNN	0.0540	0.0659	0.1090	0.0122	0.0063	**0.8125**
	PT-SVM	0.0774	0.0921	0.2561	0.0145	0.0068	**1.1019**
	SA-BFGS	0.0670	0.0632	0.1185	0.0122	0.0063	**0.7945**
	CPNN	0.0614	0.0825	0.1326	0.0127	0.0064	**0.8137**
	LDL-SVR	0.0619	0.0723	0.1063	0.0122	0.0062	**0.7636**
	LDLSF	0.0419	0.0618	0.2038	**0.0121**	0.0063	**0.8119**
	IncomLDL	0.0720	0.0706	0.1312	0.0123	0.0063	**0.7768**
	GRME	0.0612	0.0613	0.1064	**0.0121**	0.0062	**0.8023**
Chebyshev↓	sLDLCM	**0.0833**	**0.1075**	**0.1159**	**0.0507**	**0.0360**	**0.2988**
	AA-kNN	0.1024	0.1184	0.1218	0.0512	0.0363	**0.3092**
	PT-SVM	0.1216	0.1425	0.2098	0.0565	0.0380	**0.4341**
	SA-BFGS	0.1025	0.1129	0.1262	0.0511	0.0361	**0.3107**
	CPNN	0.1094	0.1369	0.1337	0.0522	0.0367	**0.3271**
	LDL-SVR	0.1113	0.1259	0.1174	0.0511	0.0362	**0.3225**
	LDLSF	0.0873	0.1080	0.1273	0.0511	**0.0360**	**0.3045**
	IncomLDL	0.1112	0.1171	0.1364	0.0520	0.0365	**0.3143**
	GRME	0.1030	0.1138	0.1215	0.0509	0.0361	**0.3084**
Intersection↑	sLDLCM	**0.8911**	**0.8653**	**0.8347**	**0.9413**	**0.9588**	**0.6223**
	AA-kNN	0.8715	0.8598	0.8248	0.9407	0.9582	**0.6079**
	PT-SVM	0.8422	0.8502	0.7059	0.9343	0.9560	**0.4511**
	SA-BFGS	0.8616	0.8636	0.8228	0.9408	0.9582	**0.6158**
	CPNN	0.8587	0.8391	0.8123	0.9396	0.9577	**0.5823**
	LDL-SVR	0.8614	0.8516	0.8322	0.9408	0.9584	**0.6035**
	LDLSF	0.8871	0.8637	0.8183	0.9409	0.9583	**0.6202**
	IncomLDL	0.8521	0.8440	0.8152	0.9399	0.9582	**0.5967**
	GRME	0.8634	0.8600	0.8314	0.9412	0.9585	**0.6191**
Cosine↑	sLDLCM	**0.9635**	**0.9449**	**0.9333**	**0.9887**	**0.9941**	**0.8238**
	AA-kNN	0.9488	0.9349	0.9283	0.9885	0.9938	**0.8104**
	PT-SVM	0.9272	0.9132	0.8320	0.9862	0.9935	**0.6420**
	SA-BFGS	0.9408	0.9407	0.9239	0.9886	0.9938	**0.8214**
	CPNN	0.9414	0.9197	0.9133	0.9881	0.9939	**0.7871**
	LDL-SVR	0.9414	0.9293	0.9304	0.9886	**0.9941**	**0.7835**
	LDLSF	0.9612	0.9440	0.9224	0.9886	**0.9941**	**0.8212**
	IncomLDL	0.9426	0.9356	0.9213	0.9883	0.9938	**0.8126**
	GRME	0.9481	0.9400	0.9305	0.9886	0.9940	**0.8218**

Table 4
The learning results of the nine compared models under the 4-fold cross-validation paradigm. ↑ (↓) indicates that the higher (lower) the value, the better (worse) the performance.

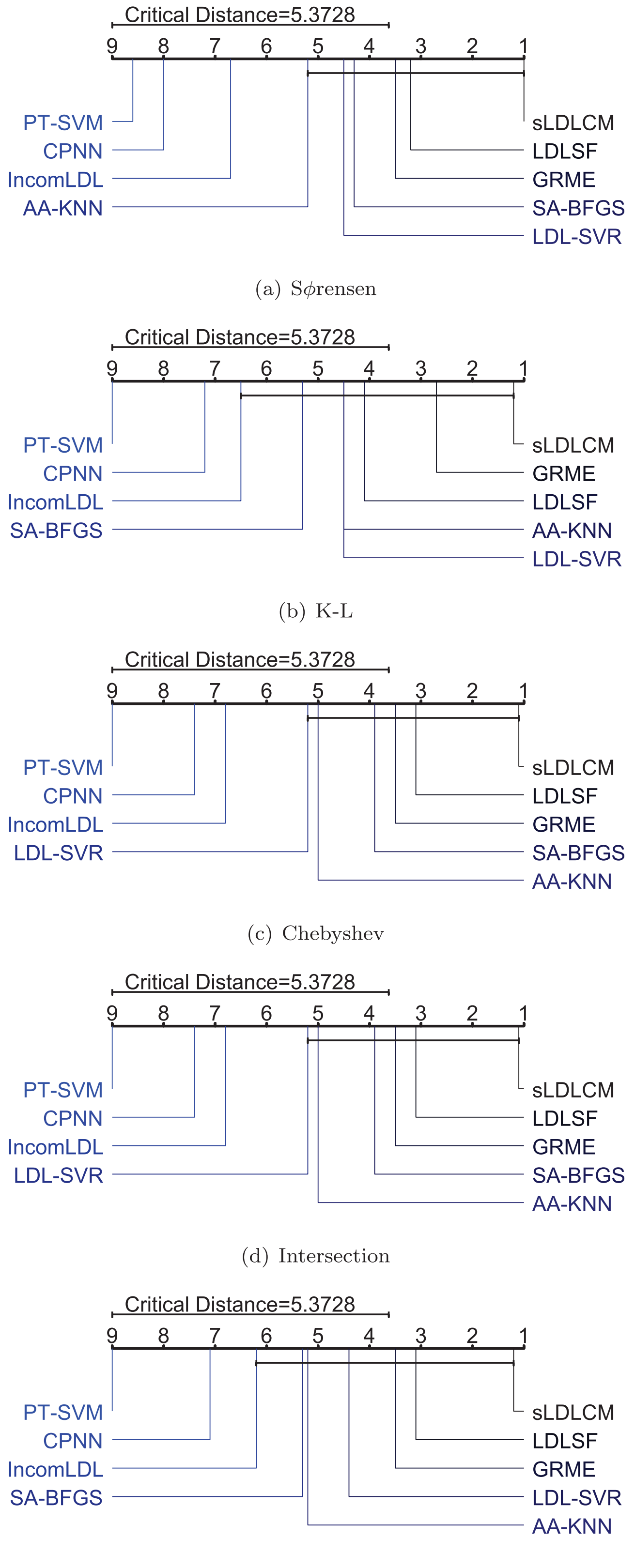
Measure	Algorithm	s-JAFFE	SBU_3DFE	Movie	Yeast-cold	Yeast-dtt	Twitter_LDL
Sϕrensen↓	sLDLCM	**0.1109**	**0.1352**	**0.1751**	**0.0591**	**0.0417**	**0.3763**
	AA-kNN	0.1324	0.1425	0.1785	0.0593	0.0422	**0.3946**
	PT-SVM	0.1600	0.3251	0.2544	0.0691	0.0444	**0.5664**
	SA-BFGS	0.1391	0.1395	0.1852	0.0594	0.0419	**0.3878**
	CPNN	0.1466	0.1611	0.1891	0.0603	0.0427	**0.4201**
	LDL-SVR	0.1413	0.1585	0.1802	0.0592	0.0419	**0.3984**
	LDLSF	0.1181	0.1360	0.1896	0.0592	0.0418	**0.3791**
	IncomLDL	0.1589	0.1579	0.1989	0.0629	0.0424	**0.4196**
	GRME	0.1310	0.1416	0.1794	0.0592	0.0422	**0.3850**
$K-L$↓	sLDLCM	**0.0423**	**0.0612**	**0.1194**	**0.0122**	**0.0062**	**0.7034**
	AA-kNN	0.0560	0.0678	0.1201	0.0124	0.0064	**0.8300**
	PT-SVM	0.0803	0.3181	0.2784	0.0158	0.0070	**1.1295**
	SA-BFGS	0.0655	0.0648	0.1300	0.0124	0.0064	**0.8054**
	CPNN	0.0671	0.0836	0.1347	0.0126	0.0066	**0.8202**
	LDL-SVR	0.0644	0.0809	0.1222	**0.0122**	0.0063	**0.7746**
	LDLSF	0.0564	0.0617	0.2585	**0.0122**	0.0063	**0.8489**
	IncomLDL	0.0801	0.0855	0.1517	0.0133	0.0067	**0.7927**
	GRME	0.0668	0.0644	0.1221	0.0123	0.0068	**0.8061**
Chebyshev↓	sLDLCM	**0.0857**	**0.1081**	**0.1223**	**0.0510**	**0.0360**	**0.2995**
	AA-kNN	0.1046	0.1201	0.1272	0.0513	0.0366	**0.3112**
	PT-SVM	0.1241	0.2302	0.2314	0.0607	0.0386	**0.4503**
	SA-BFGS	0.1052	0.1149	0.1322	0.0512	0.0363	**0.3124**
	CPNN	0.1145	0.1378	0.1338	0.0520	0.0371	**0.3311**
	LDL-SVR	0.1132	0.1345	0.1322	0.0511	0.0362	**0.3239**
	LDLSF	0.0904	0.1090	0.1328	0.0512	0.0361	**0.3075**
	IncomLDL	0.1284	0.1196	0.1494	0.0542	0.0366	**0.3296**
	GRME	0.1129	0.1182	0.1252	0.0512	0.0362	**0.3117**
Intersection↑	sLDLCM	**0.8891**	**0.8648**	**0.8269**	**0.9409**	**0.9583**	**0.6219**
	AA-kNN	0.8676	0.8570	0.8218	0.9406	0.9581	**0.6054**
	PT-SVM	0.8400	0.6749	0.7236	0.9309	0.9556	**0.4336**
	SA-BFGS	0.8609	0.8605	0.8148	0.9406	0.9556	**0.6120**
	CPNN	0.8534	0.8389	0.8109	0.9397	0.9573	**0.5799**
	LDL-SVR	0.8587	0.8415	0.8198	0.9408	0.9581	**0.6016**
	LDLSF	0.8819	0.8640	0.8104	0.9408	0.9582	**0.6179**
	IncomLDL	0.8411	0.8421	0.8011	0.9371	0.9576	**0.5802**
	GRME	0.8553	0.8572	0.8227	0.9406	0.9582	**0.6140**
Cosine↑	sLDLCM	**0.9618**	**0.9442**	**0.9292**	**0.9886**	**0.9941**	**0.8238**
	AA-kNN	0.9468	0.9332	0.9253	0.9883	0.9937	**0.8091**
	PT-SVM	0.9240	0.7689	0.8101	0.9849	0.9934	**0.6181**
	SA-BFGS	0.9386	0.9364	0.9173	0.9882	0.9937	**0.8167**
	CPNN	0.9304	0.9187	0.9122	0.9881	0.9938	**0.7836**
	LDL-SVR	0.9391	0.9213	0.9157	0.9884	0.9940	**0.7826**
	LDLSF	0.9559	0.9435	0.9158	**0.9886**	**0.9941**	**0.8184**
	IncomLDL	0.9243	0.9289	0.9066	0.9875	0.9937	**0.8124**
	GRME	0.9466	0.9365	0.9223	0.9884	0.9939	**0.8184**

Table 5
Friedman statistic τ_F for each evaluation criterion at the significance level of 0.05. In this table, the left side corresponds to the 10-fold cross-validation paradigm and the right part is the 4-fold cross-validation paradigm.

Measure	τ_F	critical value	Measure	τ_F	critical value
Sϕrensen	21.471	2.180	Sϕrensen	23.800	2.180
$K-L$	11.526		$K-L$	9.969	
Chebyshev	20.714		Chebyshev	20.714	
Intersection	27.877		Intersection	23.162	
Cosine	11.928		Cosine	19.629	

which measure the distance between ground truth label vector and estimated label vector. Therefore, ↓ is used to denote that the smaller these metric values, the better learning performance. The latter two metrics are belonging to the other group, which measure the similarity between two vectors. By contrast, ↑ means that the larger these metric values, the better learning performance. Here, we use P_j to designate the j-th element of the true label distribution and Q_j to designate the j-th element of the predicted label distribution. The detailed definitions of these measures are provided in Table 2.

On the control experiments, we compare sLDLCM with some state-of-the-art label distribution learning models including AA-kNN, PT-SVM, SA-BFGS, CPNN (conditional probability neural network), LDL-SVR (Geng and Hou, 2015; Geng and Hou, 2015), LDLSF (label distribution learning with label-specific features) (Ren et al., 2019; Ren et al., 2019), IncomLDL (incomplete label distribution learning) (Xu and Zhou, 2017; Xu and Zhou, 2017) and GRME (fragmentary LDL via graph regularized maximum entropy criteria) (Xu et al., 2021; Xu et al., 2021). The parameters involved in respective models are set as suggested by the original papers. In our experi-



(a) Sϕrensen

(b) K-L

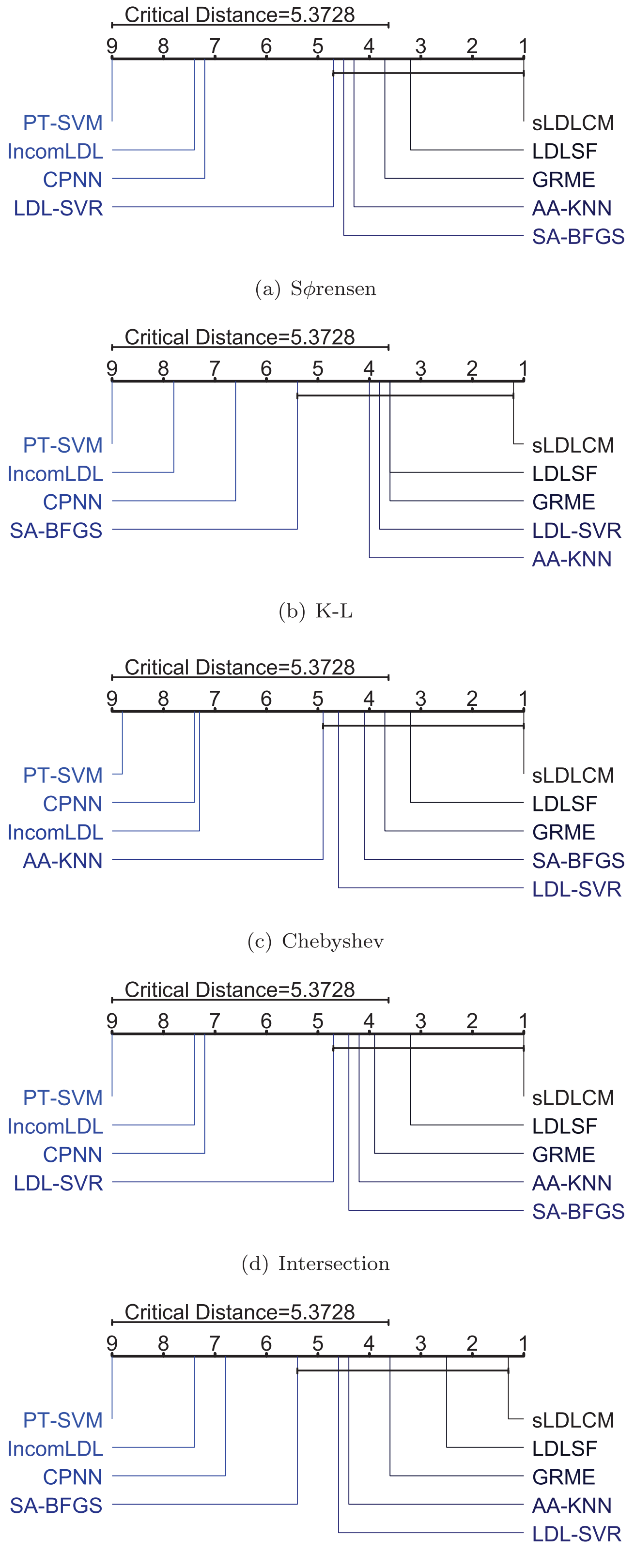
(c) Chebyshev

(d) Intersection

(e) Cosine

Fig. 2. CD diagram of the comparison algorithm using 10-fold cross-validation in each evaluation criterion. (CD = 4.9047 at 5% level of significance).

ments we find that sLDLCM is insensitive to the regularization parameters λ_1 and λ_2, which are respectively set to empirical values 10^{-5} and 10^{-3}. For λ_3 and λ_4, they are searched from candidate values $\{10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}\}$. For all elements in $\mathbf{Y}_U$, they are initialized as $\frac{1}{c}$, where c is the number of labels in the corresponding dataset. Such uniform distribution indicates that we have no prior



(a) Sϕrensen

(b) K-L

(c) Chebyshev

(d) Intersection

(e) Cosine

Fig. 3. CD diagram of the comparison algorithm using 4-fold cross-validation in each evaluation criterion. (CD = 4.9047 at 5% level of significance).

Table 6
Average rank differences among the nine models under ten-fold cross-validation with Sϕrensen evaluation criteria for Nemenyi test.

	AA-kNN	PT-SVM	SA-BFGS	CPNN	LDL-SVR	LDLSF	IncomLDL	GRME
sLDLCM	4.17	**7.67**	3.25	**7**	3.75	2	**5.75**	2.42
AA-kNN		3.5	0.92	2.83	0.42	2.17	1.58	1.75
PT-SVM			4.42	0.67	3.92	**5.67**	1.92	**5.25**
SA-BFGS				3.75	0.5	1.25	2.5	0.83
CPNN					3.25	**5**	1.25	4.58
LDL-SVR						1.75	2	1.33
LDLSF							3.75	0.42
IncomLDL								3.33

Table 7
Average rank differences among the nine models under ten-fold cross-validation with $K-L$ evaluation criteria for Nemenyi test.

	AA-kNN	PT-SVM	SA-BFGS	CPNN	LDL-SVR	LDLSF	IncomLDL	GRME
sLDLCM	3.75	**7.83**	3.91	**6.16**	2.91	3.25	4.75	1.91
AA-kNN		4.08	0.16	2.41	0.84	0.5	1	1.84
PT-SVM			3.92	1.67	4.92	4.58	3.08	**5.92**
SA-BFGS				2.25	1	0.66	0.84	2
CPNN					3.25	2.91	1.41	4.25
LDL-SVR						0.34	1.84	1
LDLSF							1.5	1.34
IncomLDL								2.84

Table 8
Average rank differences among the nine models under ten-fold cross-validation with Chebyshev evaluation criteria for Nemenyi test.

	AA-kNN	PT-SVM	SA-BFGS	CPNN	LDL-SVR	LDLSF	IncomLDL	GRME
sLDLCM	3.75	**7.92**	3	**6.42**	4.42	1.84	**5.59**	2.34
AA-kNN		4.17	0.75	2.67	0.67	1.91	1.84	1.41
PT-SVM			**4.92**	1.5	3.5	**6.08**	2.33	**5.58**
SA-BFGS				3.42	1.42	1.16	2.59	0.66
CPNN					2	4.58	0.83	4.08
LDL-SVR						2.58	1.17	2.08
LDLSF							3.75	0.5
IncomLDL								3.25

Table 9
Average rank differences among the nine models under ten-fold cross-validation with Intersection evaluation criteria for Nemenyi test.

	AA-kNN	PT-SVM	SA-BFGS	CPNN	LDL-SVR	LDLSF	IncomLDL	GRME
sLDLCM	3.83	**7.67**	3.58	**7**	3.58	2.17	**6.17**	2
AA-kNN		3.84	0.25	3.17	0.25	1.66	2.34	1.83
PT-SVM			4.09	0.67	4.09	**5.5**	1.5	**5.67**
SA-BFGS				3.42	0	1.41	2.59	1.58
CPNN					3.42	4.83	0.83	5
LDL-SVR						1.41	2.59	1.58
LDLSF							4	0.17
IncomLDL								4.17

Table 10
Average rank differences among the nine models under ten-fold cross-validation with Cosine evaluation criteria for Nemenyi test.

	AA-kNN	PT-SVM	SA-BFGS	CPNN	LDL-SVR	LDLSF	IncomLDL	GRME
sLDLCM	4.16	**7.83**	3.75	**5.91**	3.83	2.08	4.83	2.08
AA-kNN		3.67	0.41	1.75	0.33	2.08	0.67	2.08
PT-SVM			4.08	1.92	4	**5.75**	3	**5.75**
SA-BFGS				2.16	0.08	1.67	1.08	1.67
CPNN					2.08	3.83	1.08	3.83
LDL-SVR						1.75	1	1.75
LDLSF							2.75	0
IncomLDL								2.75

Table 11
Average rank differences among the nine models under fourfold cross-validation with Sϕrensen evaluation criteria for Nemenyi test.

	AA-kNN	PT-SVM	SA-BFGS	CPNN	LDL-SVR	LDLSF	IncomLDL	GRME
sLDLCM	3.42	**8**	3.42	**6.33**	3.92	2	**6.33**	2.58
AA-kNN		4.58	0	2.91	0.5	1.42	2.91	0.84
PT-SVM			4.58	1.67	4.08	**6**	1.67	**5.42**
SA-BFGS				2.91	0.5	1.42	2.91	0.84
CPNN					2.41	4.33	0	3.75
LDL-SVR						1.92	2.41	1.34
LDLSF							4.33	0.58
IncomLDL								3.75

Table 12
Average rank differences among the nine models under fourfold cross-validation with $K-L$ evaluation criteria for Nemenyi test.

	AA-kNN	PT-SVM	SA-BFGS	CPNN	LDL-SVR	LDLSF	IncomLDL	GRME
sLDLCM	3.16	**7.83**	3.5	**5.33**	2.25	3.08	**5.66**	3.66
AA-kNN		4.67	0.34	2.17	0.91	0.08	2.5	0.5
PT-SVM			4.33	2.5	5.58	4.75	2.17	4.17
SA-BFGS				1.83	1.25	0.42	2.16	0.16
CPNN					3.08	2.25	0.33	1.67
LDL-SVR						0.83	3.41	1.41
LDLSF							2.58	0.58
IncomLDL								2

Table 13
Average rank differences among the nine models under fourfold cross-validation with Chebyshev evaluation criteria for Nemenyi test.

	AA-kNN	PT-SVM	SA-BFGS	CPNN	LDL-SVR	LDLSF	IncomLDL	GRME
sLDLCM	3.58	**7.83**	3.25	**6.5**	3.83	2	**6.25**	2.75
AA-kNN		4.25	0.33	2.92	0.25	1.58	2.67	0.83
PT-SVM			4.58	1.33	4	5.83	1.58	**5.08**
SA-BFGS				3.25	0.58	1.25	3	0.5
CPNN					2.67	4.5	0.25	3.75
LDL-SVR						1.83	2.42	1.08
LDLSF							4.25	0.75
IncomLDL								3.5

Table 14
Average rank differences among the nine models under fourfold cross-validation with Intersection evaluation criteria for Nemenyi test.

	AA-kNN	PT-SVM	SA-BFGS	CPNN	LDL-SVR	LDLSF	IncomLDL	GRME
sLDLCM	3.33	**8**	3.33	**6.33**	3.92	2	**6.33**	2.75
AA-kNN		4.67	0	3	0.59	1.33	3	0.58
PT-SVM			4.67	1.67	4.08	**6**	1.67	**5.25**
SA-BFGS				3	0.59	1.33	3	0.58
CPNN					2.41	4.33	0	3.58
LDL-SVR						1.92	2.41	1.17
LDLSF							4.33	0.75
IncomLDL								3.58

Table 15
Average rank differences among the nine models under fourfold cross-validation with Cosine evaluation criteria for Nemenyi test.

	AA-kNN	PT-SVM	SA-BFGS	CPNN	LDL-SVR	LDLSF	IncomLDL	GRME
sLDLCM	3.5	**7.83**	4	**5.66**	4.25	1.25	**5.83**	2.16
AA-kNN		4.33	0.5	2.16	0.75	2.25	2.33	1.34
PT-SVM			3.83	2.17	3.58	**6.58**	2	**5.67**
SA-BFGS				1.66	0.25	2.75	1.83	1.84
CPNN					1.41	4.41	0.17	3.5
LDL-SVR						3	1.58	2.09
LDLSF							4.58	0.91
IncomLDL								3.67

Table 16
Average rank differences between sLDLCM and the other eight algorithms for Bonferroni-Dunn test.

	AA-kNN	PT-SVM	SA-BFGS	CPNN	LDL-SVR	LDLSF	IncomLDL	GRME
Sϕrensen 10CV	4.17	**7.67**	3.25	**7**	3.75	2	**5.75**	2.42
$K-L$ 10CV	3.75	**7.83**	3.91	**6.16**	2.91	3.25	**4.75**	1.91
Chebshev 10CV	3.75	**7.92**	3	**6.42**	**4.42**	1.84	**5.59**	2.34
InterSection 10CV	3.83	**7.67**	3.58	**7**	3.58	2.17	**6.17**	2
Cosine 10CV	4.16	**7.83**	3.75	**5.91**	3.83	2.08	**4.83**	2.08
Sϕrensen 4CV	3.42	**8**	3.42	**6.33**	3.92	2	**6.33**	2.58
$K-L$ 4CV	3.16	**7.83**	3.5	**5.33**	2.25	3.08	**5.66**	3.66
Chebshev 4CV	3.58	**7.83**	3.25	**6.5**	3.83	2	**6.25**	2.75
InterSection 4CV	3.33	**8**	3.33	**6.33**	3.92	2	**6.33**	2.75
Cosine 4CV	3.5	**7.83**	4	**5.66**	4.25	1.25	**5.83**	2.16

knowledge on the label distribution of each sample. Matrix $\mathbf{W}$ is initialized by solving problem (1). Both matrices $\mathbf{P}$ and $\mathbf{Q}$ are initialized to $\frac{1}{2}\mathbf{W}$.

4.3. Results and analysis

In the following experiments, two paradigms, *i.e.*, 10-fold cross-validation and 4-fold cross-validation, are conducted on each data set for comparative studies among these label distribution learning models. To be specific, in each run of the 10-fold cross-validation paradigm, 90% samples are labeled while 10% samples are unlabeled. Accordingly, 75% samples are labeled and 25% samples are unlabeled in the 4-fold cross-validation paradigm. The average values of different evaluation metrics are reported in Tables 3 and 4, where the best result in each case is highlighted in bold. To better distinguish whether there are significant differences among the obtained results by these nine models, statistical tests were performed by the Friedman test, Nemenyi test and Bonferroni-Dunn test(López-Vázquez and Hochsztain, 2019; López-Vázquez and Hochsztain, 2019; Nandhini et al., 2020; Nandhini et al., 2020). Usually, the Friedman test is used to determine whether the performance of these models is the same, whose underlying hypothesis is that "all the models have the same label distribution learning performance". The Friedman test calculates the Friedman statistics τ_F according to

$$\tau_F = \frac{(N-1)\tau_{\chi^2}}{N(k-1)-\tau_{\chi^2}}, \tag{14}$$

and

$$\tau_{\chi^2} = \frac{k-1}{k}\cdot\frac{12N}{k^2-1}\sum_{i=1}^{k}\left(r_i-\frac{k+1}{2}\right)^2, \tag{15}$$

where k represents the number of models (*i.e.*, 9), N represents the number of data sets (*i.e.*, 6) and r_i denotes the mean rank of i-th algorithm. For example, in terms of the Sϕrensen metric in Table 3, the ranks of AA-kNN on the five data sets are 4, 5, 4, 6, 7, respectively while they are 3, 4, 3, 2, 5 for GRME. Table 5 presents the Friedman statistics τ_F in terms of different evaluation metrics when the significance level is 0.05. The corresponding critical value is 2.180 when $k=9$ and $N=6$.

As depicted in Table 5, it is obvious that the hypothesis that "all the label distribution learning models have the same performance" should be rejected when the significance level is 0.05. Therefore, the Nemenyi test serves as the post hoc test to further distinguish these models, based on which the critical distance (CD) can be calculated by

$$CD = q_\alpha\sqrt{\frac{k(k+1)}{6N}}, \tag{16}$$

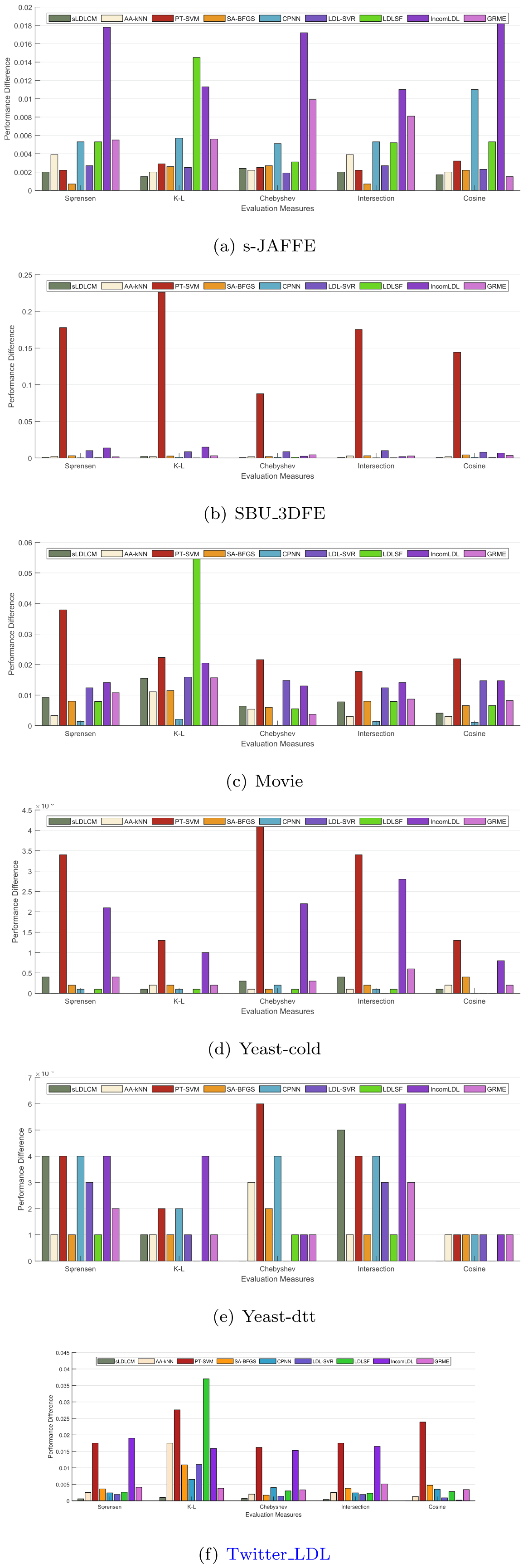
where q_α is the critical value of the Tukey distribution and it is 3.102 when the number of compared models is 9. If the gap between the average ranks of two models exceeds the CD (*i.e.*, 4.9047), we reject the hypothesis that "the two algorithms have no significantly different performance" and believe that there exist significant differences between their results. The CD diagrams of these compared models in terms of different evaluation metrics are respectively shown in Figs. 2 and 3. The differences in average ranks among the nine models that were subjected to the Nemenyi test under the five evaluation metrics are shown in Tables 6–15, where the terms greater than CD are bolded. In addition, Bonferroni-Dunn test is performed to better indicate the differences between our algorithm and a given algorithm. When the number of compared models was 9, the corresponding q_α is 2.724. CD was calculated using Eq. (16) (*i.e.*, 4.3070), and the detailed results are shown in Table 16.

Based on the above experimental results, we have the following insights.

- (1) In most of the cases listed in Tables 3 and 4, sLDLCM obtained the best performance among these compared models, followed by the LDLSF and GRME. LDLSF is a supervised model in which the label-specific (-common) features and label correlations are considered. GRME on one hand considers the local data invariance and on the other hand uses maximal entropy criterion to regularize the label distribution of data. However, as a semi-supervised model, GRME has no out-of-sample extension ability.
- (2) From the statistical tests, we find that sLDLCM has the highest rank in terms of all the evaluation metrics. As a whole, sLDLCM integrates four parts together, *i.e.*, joint label estimation of unlabeled samples, exploitation of label-specific (-common) features, label correlations and local data manifold, contributing to excellent label distribution learning performance.
- (3) Generally, sLDLCM achieved better performance in the 10-fold cross-validation than that in the 4-fold cross-validation across all used data sets. This indicates that the more labeled samples we are given, the better the sLDLCM can make use of the data label information and simultaneously explore the label correlations. In Fig. 4, we plot the performance differences of the compared models in the two learning paradigms. We find that sLDLCM has very small performance degeneration in the 4-fold cross-validation paradigm in comparison with the 10-fold one, except the only two cases, *i.e.*, the Sϕrensen and Intersection metrics on the Yeast-dtt data set. This shows that the involving unlabeled data into learning process is beneficial for sLDLCM to better capture the data properties.

4.4. Convergence and parameter sensitivity analysis

Besides the theoretical analysis in Section 3.3, here we employ two examples to experimentally evaluate the convergence property of sLDLCM. In Fig. 5, we show the convergence curves of sLDLCM on the SBU_3DFE and s-JAFFE data sets. From this figure,



(a) s-JAFFE

(b) SBU_3DFE

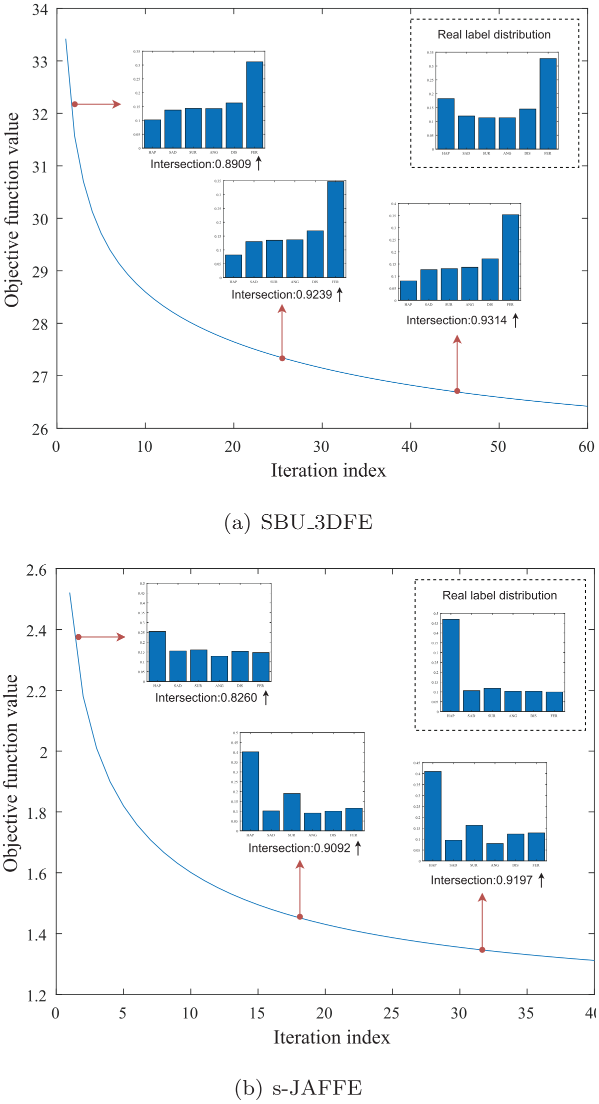
(c) Movie

(d) Yeast-cold

(e) Yeast-dtt

(f) Twitter_LDL

Fig. 4. The performance differences of the compared models in the two learning paradigms.



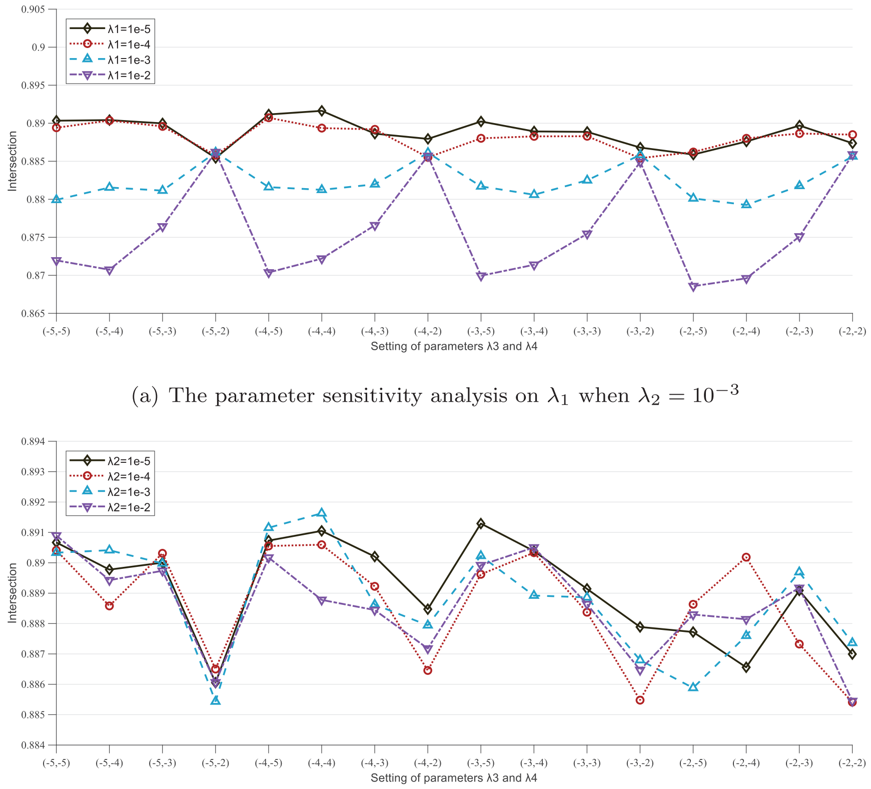
(a) SBU_3DFE

(b) s-JAFFE

Fig. 5. Convergence curves of sLDLCM on SBU_3DFE and s-JAFFE data sets.

we can observe that as the number of iterations increases, the objective function values of sLDLCM monotonically decrease. Similar trends can be found from the other data sets. These indicate the desirable convergence properties of our proposed sLDLCM model. In addition, in each example we tracked the predicted label distributions of one sample in terms of iterations. In initial stages, the predicted label distribution focuses mainly on the most significant label, and therefore could not well capture the relationship among all the labels. As the number of iterations increases, the predicted label distribution gradually approaches the ground truth label distribution. In sLDLCM, the involved three components work together effectively for better modeling the label distributions. One is the joint estimation of label distributions of unlabeled samples (*i.e.*,$\mathbf{Y}_U$), the other is the exploitation of the label correlations, and the third one is using the graph regularization to constrain the label distributions.

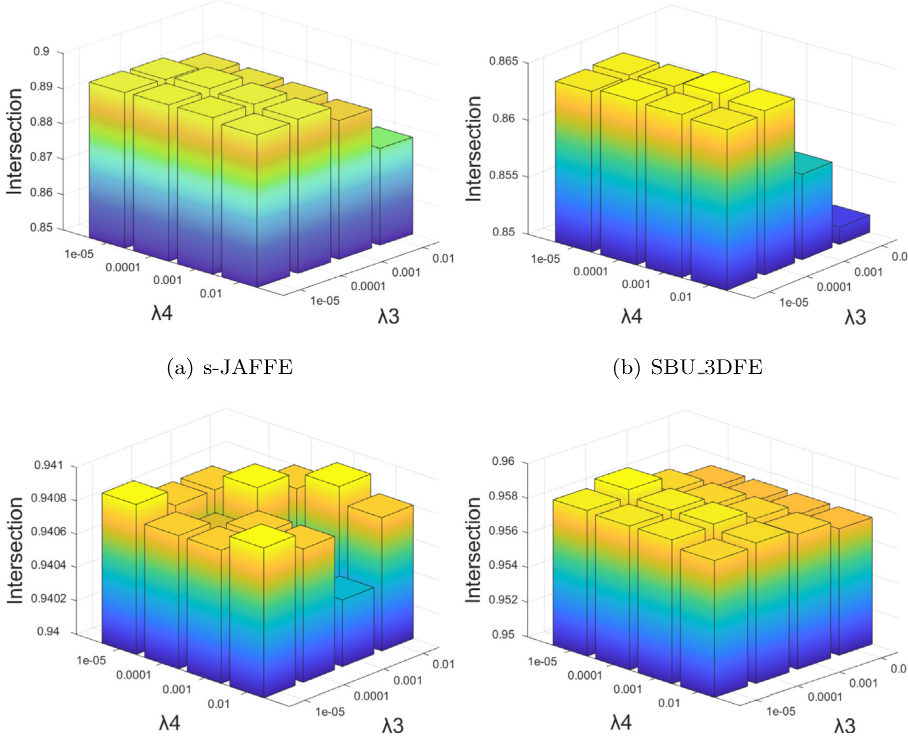
There are four regularization parameters in the sLDLCM model objective function (3). As declared in Section 4.2, sLDLCM is insensitive to λ_1 and λ_2, which were set to fixed values in experiments. The optimal values of the remaining two parameters, λ_3 and λ_4, were determined by grid search. Taking the s-JAFFE data set as



(a) The parameter sensitivity analysis on λ_1 when $\lambda_2 = 10^{-3}$

(b) The parameter sensitivity analysis on λ_2 when $\lambda_1 = 10^{-5}$

Fig. 6. The parameter sensitivity analysis on λ_1 and λ_2.



(a) s-JAFFE

(b) SBU_3DFE

(c) Yeast-cold

(d) Yeast-dtt

Fig. 7. The performance of sLDLCM in terms of different (λ_3, λ_4) s.

an example, the two subfigures in Fig. 6 show the variations of sLDLCM learning performance in terms of λ_1 and λ_2, respectively. The abscissa values (a, b) mean that $\lambda_3 = 10^a$ and $\lambda_4 = 10^b$. The top two lines in Fig. 6(a), respectively corresponding to $\lambda_1 = 10^{-5}$ and $\lambda_1 = 10^{-4}$, are approximately overlapping. Even when $\lambda_1 = 10^{-3}$, sLDLCM also achieved promising performance. In Fig. 6 (b), it can be observed that all the four lines approximately coincide; therefore, we conclude that sLDLCM is not sensitive to the selection of different λ_2s. To reduce the burden of parameter tuning, we simply fixed the values of λ_1 and λ_2 as 10^{-5} and 10^{-3}. In Fig. 7, some example data sets are used to show how the sLDLCM performance is affected by parameters λ_3 and λ_4. From this figure, we find that there are many candidate values of (λ_3, λ_4) to make sLDLCM achieve good performance.

5. Conclusions

In this paper, we proposed a semi-supervised labeled distribution learning algorithm termed sLDLCM, which consists of four ingredients to work together. First, a semi-supervised least square regression is used to estimate the label distributions of unlabeled samples. Second, ℓ_1-norm and $\ell_{2,1}$-norm are used to respectively learn label-specific and label-common features. Third, the correlations among different labels are taken into account. Finally, the local invariance of data is used to learn similar label distributions for similar samples. We conducted extensive experiments on six real-world data sets to demonstrate the effectiveness of sLDLCM by comparing it with the state-of-the-arts. Experimental results demonstrated that the proposed sLDLCM model is competent for label distribution learning. In the future, we will consider the nonlinear extension of the current sLDLCM model for capturing the possible nonlinear structure in data.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

This work was partially supported by the Natural Science Foundation of Zhejiang Province (LY21F030005), National Natural Science Foundation of China (61971173, U20B2074), Fundamental Research Funds for the Provincial Universities of Zhejiang (GK209907299001-008), China Postdoctoral Science Foundation (2017M620470), CAAC Key Laboratory of Flight Techniques and Flight Safety (FZ2021KF16) and Graduate Scientific Research Foundation of Hangzhou Dianzi University (CXJJ2022088).

Appendix A. Two operators

Below we describe two operators, *soft shrinkage* and *$\ell_{2,1}$-norm minimization*, which are involved in the optimization to sLDLCM objective function.

- soft shrinkage operator $\mathscr{S}_\epsilon[x]$ is defined as

$$\mathscr{S}_\epsilon[x] = \begin{cases} x - \epsilon, & \text{if } x > \epsilon, \\ x + \epsilon, & \text{if } x < -\epsilon, \\ 0, & \text{otherwise}, \end{cases} \tag{A.1}$$

where x and ϵ is usually a small positive value. This operator can be extended to vectors and matrices by applying it elementwisely as

$$\mathscr{S}_\epsilon[\mathbf{M}] = \arg\min_{\mathbf{X}} \epsilon\|\mathbf{X}\|_1 + \frac{1}{2}\|\mathbf{X} - \mathbf{M}\|_2^2. \tag{A.2}$$

- $\ell_{2,1}$-norm minimization operator. Given a matrix $\mathbf{M} = [\mathbf{m}_1, \mathbf{m}_2, \cdots, \mathbf{m}_i, \cdots]$. Assuming that the optimal solution of

$$\mathbf{\Omega}_\lambda[\mathbf{W}] \triangleq \|\mathbf{W}\|_{2,1} + \frac{1}{2}\|\mathbf{W} - \mathbf{M}\|_2^2 \tag{A.3}$$

is $\mathbf{W}^*$, the i-th column of $\mathbf{W}^*$ can be obtained by

$$\mathbf{W}_i^* = \begin{cases} \frac{\|\mathbf{m}_i\|_2 - \lambda}{\|\mathbf{m}_i\|_2}\mathbf{q}_i, & \text{if } \lambda < \|\mathbf{m}_i\|_2, \\ 0, & \text{otherwise}. \end{cases} \tag{A.4}$$

Appendix B. Optimization to problem (13)

To simplify the following notations, we use $\mathbf{m}_i$ and $\mathbf{y}_i$ to represent the transpose of $\mathbf{m}^i$ and $\mathbf{y}^i$, respectively. Then, we have the Lagrangian function of problem (13) as

$$\mathscr{L}(\mathbf{y}_i, \eta, \boldsymbol{\beta}) = \|\mathbf{y}_i - \mathbf{m}_i\|_2^2 - \eta(\mathbf{y}_i^T\mathbf{1}_c - 1) - \boldsymbol{\beta}^T\mathbf{y}_i, \tag{B.1}$$

where η is a scalar and $\boldsymbol{\beta} \in \mathbb{R}^c$ is a vector. Below we provide analysis that both η and $\boldsymbol{\beta}$ can be determined. We suppose that the optimal solution to the proximal problem (13) is $\mathbf{y}_i^*$, and the associated Lagrangian multipliers are η^* and $\boldsymbol{\beta}^*$. Then, according to the KKT (Karush–Kuhn–Tucker) condition, we have the following equations and inequalities

$$\begin{cases} \forall j \in \{1, 2\cdots, c\}, & y_{ij}^* - m_{ij} - \eta^* - \beta_j^* = 0, \\ \forall j \in \{1, 2\cdots, c\}, & y_{ij}^* \geqslant 0, \\ \forall j \in \{1, 2\cdots, c\}, & \beta_j^* \geqslant 0, \\ \forall j \in \{1, 2\cdots, c\}, & y_{ij}^*\beta_j^* = 0, \end{cases} \tag{B.2}$$

where y_{ij}^* is the j-th element of vector $\mathbf{y}_i^*$. The vector form of Eq. (B.2) is

$$\mathbf{y}_i^* - \mathbf{m}_i - \eta^*\mathbf{1}_c - \boldsymbol{\beta}^* = \mathbf{0}. \tag{B.6}$$

Considering the normalization constraint $\mathbf{y}_i^T\mathbf{1}_c = 1$, the above equation can be reformulated into

$$\eta^* = \frac{1 - \mathbf{1}_c^T\mathbf{m}_i - \mathbf{1}_c^T\boldsymbol{\beta}^*}{c}. \tag{B.7}$$

By substituting (B.7) into (B.6), we have

$$\mathbf{y}_i^* = \mathbf{m}_i - \frac{\mathbf{1}_c\mathbf{1}_c^T}{c}\mathbf{m}_i + \frac{1}{c}\mathbf{1}_c - \frac{\mathbf{1}_c^T\boldsymbol{\beta}^*}{c}\mathbf{1}_c + \boldsymbol{\beta}^*. \tag{B.8}$$

Denote $\bar{\beta}^* = \frac{\mathbf{1}_c^T\boldsymbol{\beta}^*}{c}$ and $\mathbf{q} = \mathbf{m}_i - \frac{\mathbf{1}_c\mathbf{1}_c^T}{c}\mathbf{m}_i + \frac{1}{c}\mathbf{1}_c$, the above equation can be rewritten as

$$\mathbf{y}_i^* = \mathbf{q} + \boldsymbol{\beta}^* - \bar{\beta}^*\mathbf{1}_c. \tag{B.9}$$

Therefore, for each $j = 1, \cdots, c$, we have

$$y_{ij}^* = q_j + \beta_j^* - \bar{\beta}^*. \tag{B.10}$$

According to equations (B.3)-(B.5) and (B.10), we know $q_j + \beta_j^* - \bar{\beta}^* = (q_j - \bar{\beta}^*)_+$, where $(f(\cdot))_+ = \max(f(\cdot), 0)$. Therefore, we have

$$y_{ij}^* = (q_j - \bar{\beta}^*)_+. \tag{B.11}$$

Now if the optimal $\bar{\beta}^*$ can be determined, the optimal solution $\mathbf{y}_i^*$ can be obtained from (B.11). Eq. (B.10) can be rewritten as $\beta_j^* = y_{ij}^* + \bar{\beta}^* - q_j$ such that $\beta_j^* = (\bar{\beta}^* - q_j)_+$. Therefore, $\bar{\beta}^*$ can be calculated as

$$\bar{\boldsymbol{\beta}}^* = \frac{1}{c}\sum_{j=1}^{c}(\bar{\beta}^* - q_j)_+ . \tag{B.12}$$

According to the constraint $\mathbf{y}_i^T\mathbf{1} = 1$ and (B.11), we define the following function

$$f(\bar{\beta}) = \sum_{j=1}^{c}(q_j - \bar{\beta})_+ - 1, \tag{B.13}$$

and the optimal $\bar{\beta}^*$ should satisfy $f(\bar{\beta}^*) = 0$. When (B.13) equals to zero, the optimal $\bar{\beta}^*$ can be obtained via Newton method, namely

$$\bar{\beta}^{(k+1)} = \bar{\beta}^{(k)} - \frac{f(\bar{\beta}^{(k)})}{f\prime(\bar{\beta}^{(k)})}. \tag{B.14}$$

We know that $f(\bar{\beta})$ is a piecewise linear and monotonically increasing function. When $q_j \geqslant \bar{\beta}, f(\bar{\beta}) = \sum_{j=1}^{c} q_j - \bar{\beta} - 1$ and we have $f\prime(\bar{\beta}) = -1$. When $q_j \leqslant \bar{\beta}, f(\bar{\beta}) = -1$ and its derivative $f\prime(\bar{\beta}) = 0$. Therefore, we can obtain $f\prime(\bar{\beta})$ by counting the number of positive values in $(q_j - \bar{\beta})|_{j=1}^{c}$.

References

Aamir, M., Zaidi, S.M.A., 2021. Clustering based semi-supervised machine learning for ddos attack classification. Journal of King Saud University-Computer and Information Sciences 33, 436–446.

Bollapragada, R., Nocedal, J., Mudigere, D., Shi, H.-J., & Tang, P.T.P. (2018). A progressive batching L-BFGS method for machine learning. In International Conference on Machine Learning (pp. 620–629).

Chen, M., Wang, X., Feng, B., Liu, W., 2018. Structured random forest for label distribution learning. Neurocomputing 320, 171–182.

Gao, B.-B., Xing, C., Xie, C.-W., Wu, J., Geng, X., 2017. Deep label distribution learning with label ambiguity. IEEE Transactions on Image Processing 26, 2825–2838.

Geng, X., 2016. Label distribution learning. IEEE Transactions on Knowledge & Data Engineering 28, 1734–1748.

Geng, X., & Hou, P. (2015). Pre-release prediction of crowd opinion on movies by label distribution learning. In Proceedings of International Joint Conference on Artificial Intelligence (pp. 3511–3517).

Geng, X., Wang, Q., & Xia, Y. (2014). Facial age estimation by adaptive label distribution learning. In Proceedings of International Conference on Pattern Recognition (pp. 4465–4470).

Geng, X., Yin, C., Zhou, Z.-H., 2013. Facial age estimation by learning from label distributions. IEEE Transactions on Pattern Analysis and Machine Intelligence 35, 2401–2412.

He, B., Ma, F., Yuan, X., 2020. Optimally linearizing the alternating direction method of multipliers for convex programming. Computational Optimization and Applications 75, 361–388.

Hou, P., Geng, X., Huo, Z.-W., & Lv, J.-Q. (2017). Semi-supervised adaptive label distribution learning for facial age estimation. In Proceedings of AAAI Conference on Artificial Intelligence (pp. 2015–2021).

Jia, X., Li, W., Liu, J., & Zhang, Y. (2018). Label distribution learning by exploiting label correlations. In Proceedings of AAAI Conference on Artificial Intelligence (pp. 3310–3317).

Jia, X., Li, Z., Zheng, X., Li, W., Huang, S.-J., 2021. Label distribution learning with label correlations on local samples. IEEE Transactions on Knowledge and Data Engineering 33, 1619–1631.

Li, P., Hu, Y., Wu, X., He, R., Sun, Z., 2020. Deep label refinement for age estimation. Pattern Recognition 100 (107178), 1–12.

Liao, L., Zhang, X., Zhao, F., Lou, J., Wang, L., Xu, X., Zhang, H., & Li, G. (2020). Multi-branch deformable convolutional neural network with label distribution learning for fetal brain age prediction. In IEEE International Symposium on Biomedical Imaging (pp. 424–427).

Liu, T., Wang, J., Yang, B., Wang, X., 2021. Ngdnet: Nonuniform gaussian-label distribution learning for infrared head pose estimation and on-task behavior understanding in the classroom. Neurocomputing 436, 210–220.

López-Vázquez, C., Hochsztain, E., 2019. Extended and updated tables for the friedman rank test. Communications in Statistics-Theory and Methods 48, 268–281.

Luo, J., Wang, Y., Ou, Y., He, B., Li, B., 2021. Neighbor-based label distribution learning to model label ambiguity for aerial scene classification. Remote Sensing 13 (755), 1–24.

Ma, A., You, F., Jing, M., Li, J., Lu, K., 2020. Multi-source domain adaptation with graph embedding and adaptive label prediction. Information Processing & Management 57 (102367), 1–19.

Nandhini, M., Rajalakshmi, M., Sivanandam, S.N., 2020. Performance analysis of predictive association rule classifiers using healthcare datasets. IETE Technical Review 39, 143–156.

Pattanaik, P., Mittal, M., Khan, M.Z., Panda, S., 2020. Malaria detection using deep residual networks with mobile microscopy. Journal of King Saud University-Computer and Information Sciences 34, 1700–1705.

Ren, T., Jia, X., Li, W., Chen, L., & Li, Z. (2019). Label distribution learning with label-specific features. In Proceedings of International Joint Conference on Artificial Intelligence (pp. 3318–3324).

Ren, Y., & Geng, X. (2017). Sense beauty by label distribution learning. In Proceedings of International Joint Conference on Artificial Intelligence (pp. 2648–2654).

Shen, W., Zhao, K., Guo, Y., & Yuille, A. (2017). Label distribution learning forests. In Proceedings of International Conference on Neural Information Processing Systems (pp. 834–843).

Si, S., Wang, J., Peng, J., & Xiao, J. (2022). Towards speaker age estimation with label distribution learning. In IEEE International Conference on Acoustics, Speech and Signal Processing (pp. 4618–4622).

Tarekegn, A.N., Giacobini, M., Michalak, K., 2021. A review of methods for imbalanced multi-label classification. Pattern Recognition 118 (107965), 1–12.

Wang, J., & Geng, X. (2019). Theoretical analysis of label distribution learning. Proceedings of the AAAI Conference on Artificial Intelligence, 33, 5256–5263.

Wang, J., Wang, X., Tian, F., Liu, C.H., Yu, H., 2016. Constrained low-rank representation for robust subspace clustering. IEEE transactions on cybernetics 47, 4534–4546.

Wang, R., Ridley, R., Qu, W., Dai, X., et al., 2021. A novel reasoning mechanism for multi-label text classification. Information Processing & Management 58 (102441), 1–15.

Wen, X., Li, B., Guo, H., Liu, Z., Hu, G., Tang, M., & Wang, J. (2020). Adaptive variance based label distribution learning for facial age estimation. In Proceedings of European Conference on Computer Vision (pp. 379–395).

Xu, C., Gu, S., Tao, H., Hou, C., 2021. Fragmentary label distribution learning via graph regularized maximum entropy criteria. Pattern Recognition Letters 145, 147–156.

Xu, L., Chen, J., Gan, Y., 2019. Head pose estimation using improved label distribution learning with fewer annotations. Multimedia Tools and Applications 78, 19141–19162.

Xu, M., & Zhou, Z.-H. (2017). Incomplete label distribution learning. In Proceedings of International Joint Conference on Artificial Intelligence (pp. 3175–3181).