Latent Semantics Encoding for Label Distribution Learning *

Suping Xu , **Lin Shang** and **Furao Shen**
State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing 210023, China
Department of Computer Science and Technology, Nanjing University, Nanjing 210023, China
supingxu@smail.nju.edu.cn, {shanglin,frshen}@nju.edu.cn

Abstract

Label distribution learning (LDL) is a newly arisen learning paradigm to deal with label ambiguity problems, which can explore the relative importance of different labels in the description of a particular instance. Although some existing LDL algorithms have achieved better effectiveness in real applications, most of them typically emphasize on improving the learning ability by manipulating the label space, while ignoring the fact that irrelevant and redundant features exist in most practical classification learning tasks, which increase not only storage requirements but also computational overheads. Furthermore, noises in data acquisition will bring negative effects on the generalization performance of LDL algorithms. In this paper, we propose a novel algorithm, i.e., **L**atent **S**emantics **E**ncoding for **L**abel **D**istribution **L**earning (LSE-LDL), which learns the label distribution and implements feature selection simultaneously under the guidance of latent semantics. Specifically, to alleviate noise disturbances, we seek and encode discriminative original physical/chemical features into advanced latent semantic features, and then construct a mapping from the encoded semantic space to the label space via empirical risk minimization. Empirical studies on 15 real-world data sets validate the effectiveness of the proposed algorithm.

1 Introduction

Learning with label ambiguity has increasingly attracted attention in recent machine learning and data mining areas. At present, single-label learning (SLL) and multi-label learning (MLL) [Gibaja Galindo and Ventura, 2014; Zhang and Zhou, 2014; Zhang *et al.*, 2018] are two widely-used paradigms to deal with the label ambiguity problems, where each instance only belongs to a single label in SLL, whereas each instance may be associated to multiple labels simultaneously in MLL. Although SLL and MLL have achieved a lot of success [Boutell *et al.*, 2004; Huang and Zhou, 2012] in classification learning tasks, both of them focus on a relatively broad label ambiguity problem, i.e., "*which labels can describe a particular instance?*", thus the output of an SLL/MLL algorithm often is a set of labels with the implicit assumption that these labels are considered to be equally important. However, sometimes ones need to deal with the further label ambiguity problem of "*to what extent will different labels describe a particular instance?*", in other words, the relative importance of different labels involved in the description of an instance, i.e., a distribution over the set of labels, is expected to be obtained. For example, a facial expression usually is constituted by a variety of different emotional components, such as dejection, pleasure, enthusiasm and so on. Different emotional components make different contributions to building a particular facial expression, and they form an emotion distribution [Zhou *et al.*, 2015] for facial expression. To solve such learning problems with label ambiguity, [Geng and Ji, 2013] first proposed the concept of label distribution learning (LDL), which can be viewed as a more generalized learning paradigm when compared with the traditional SLL and MLL.

Although LDL has been successfully applied to some practical scenarios [Zhou *et al.*, 2015; Geng and Hou, 2015], most of the previous LDL algorithms aim to boost the learning performance by manipulating the label space, such as exploiting the correlations existing among different labels. Nevertheless, similar to SLL/MLL tasks, there may also exist irrelevant and redundant features in LDL tasks, which will result in the increasing of storage requirements and computational overheads. Moreover, in data acquisition, noise disturbances are common due to the limits of the precision and reliability of data collector (temperature/light/heat/pressure sensors, etc.), it could have negative effects on the generalization performance of LDL algorithms. Existing studies on feature selection [Liu and Motoda, 1998] for SLL/MLL tasks have shown that only a subset of relevant features contains the most discriminative information in general, and the ability of classification learning will be improved [Yu and Liu, 2004; Xu *et al.*, 2016] via removing some irrelevant, redundant, and noisy features. It is worth noting that even if both MLL and LDL are placed in the setting of multiple labels, most feature selection algorithms for MLL cannot be well adapted to LDL tasks, since they usually destroy the geometry structure [Guo *et al.*, 2019] of feature space, which leads to the inconsistency between the feature space and the label space, i.e., two instances, whose label distributions are close, may not be close

*Lin Shang is the corresponding author.

to each other in the selected feature space.

To solve the above problem, inspired by [Jian *et al.*, 2016], in this paper, we propose a novel LDL algorithm termed **L**atent **S**emantics **E**ncoding for **L**abel **D**istribution **L**earning, shortly LSE-LDL. Specifically, the latent semantic features are encoded by a regression model from the original physical/chemical features, and the most discriminative original features can be selected by the $\ell_{2,1}$-*norm* regularization term. The encoded latent semantic features can alleviate the negative effects of noise disturbances in data acquisition by the assumption that neighbor instances in the label space keep close to each other in the latent semantic feature space. Under the guidance of latent semantics, the optimization objective of LSE-LDL will focus on learning the label distribution and implementing feature selection simultaneously. Besides, we develop an Alternating-Updating algorithm based on gradient descent for the optimization.

The main contributions of this paper can be summarized as follows: 1) Constructing the latent semantic features for alleviating the negative effects of noise disturbances; 2) Implementing feature selection for eliminating the irrelevant and redundant features; 3) Developing an efficient algorithm to address the optimization problem.

2 LSE-LDL Approach

2.1 Preliminaries

Let $\mathcal{X} \in \mathbb{R}^m$ be the m-dimensional input space, $\mathcal{Y} \in [0,1]^p$ be the p-dimensional label space, and $\{y_1, y_2, \ldots, y_p\}$ be the finite set of p possible labels. Suppose that we have n training instances $\mathbf{x}_i \in \mathcal{X}$ $(i = 1, 2, \ldots, n)$, and each instance $\mathbf{x}_i$ is associated with a label distribution $\mathbf{d}_i = [d_{\mathbf{x}_i}^{y_1}, d_{\mathbf{x}_i}^{y_2}, \ldots, d_{\mathbf{x}_i}^{y_p}]^{\mathrm{T}}$, where $d_{\mathbf{x}_i}^{y_l}$ denotes the description degree of label y_l $(l = 1, 2, \ldots, p)$ to $\mathbf{x}_i$. For each $\mathbf{x}_i$, $\forall d_{\mathbf{x}_i}^{y_l} \in [0,1]$, and $\sum_{l=1}^{p} d_{\mathbf{x}_i}^{y_l} = 1$, which means all the labels can completely describe $\mathbf{x}_i$. For convenience, we denote the training data matrix as $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n]^{\mathrm{T}} \in \mathbb{R}^{n \times m}$ and the label distribution matrix as $\mathbf{D} = [\mathbf{d}_1, \mathbf{d}_2, \ldots, \mathbf{d}_n]^{\mathrm{T}} \in \mathbb{R}^{n \times p}$.

In addition, without loss of generality, for any vector $\mathbf{g} \in \mathbb{R}^m$, we use $\mathbf{g}_j$ to denote the j-th element of $\mathbf{g}$, for any matrix $\mathbf{G} \in \mathbb{R}^{n \times m}$, we use $\mathbf{G}_{ij}$ to denote the (i,j)-th element of $\mathbf{G}$, and use $\mathbf{G}_{i:}$ and $\mathbf{G}_{:j}$ to denote all elements in the i-th row and the j-th column of $\mathbf{G}$, respectively. Tr($\mathbf{G}$) is adopted to denote the trace of square matrix $\mathbf{G}$, and the $\ell_{2,1}$-*norm* of $\mathbf{G}$ is defined as $||\mathbf{G}||_{2,1} = \sum_{i=1}^{n} \sqrt{\sum_{j=1}^{m} \mathbf{G}_{ij}^2} = \sum_{i=1}^{n} ||\mathbf{G}_{i:}||_2$.

2.2 Formulation

The goal of LDL is to generate a learner $\mathbf{f}$ which can predict the label distributions of unseen instances. In general, $\mathbf{f} = [f_1, f_2, \ldots, f_p]^{\mathrm{T}}$ and each f_l $(l = 1, 2, \ldots, p)$ is a sub-learner for label y_l.

We formulate the LDL problems via empirical risk minimization in the following learning error:

$$\min_{\mathbf{f}} \sum_{i=1}^{n} Loss(\mathbf{x}_i, \mathbf{d}_i, \mathbf{f}) + \delta\Omega(\mathbf{f}), \tag{1}$$

where $Loss(\cdot)$ denotes the loss function defined on the training data, $\Omega(\mathbf{f})$ is a regularization term to control the complexity of LDL learner $\mathbf{f}$, and δ is a regularization parameter trading off the two terms.

Some functions [Cha, 2007], which can measure the similarity between the true distribution and the predicted distribution, are suitable to construct the LDL loss function, such as *Canberra*, *Kullback-Leibler divergence*, and *Squared* χ^2, etc. Moreover, ℓ_1-*norm*, ℓ_2-*norm*, and F-*norm* can be the candidates of the regularization term. In this paper, we focus on the commonly adopted *Kullback-Leibler divergence* defined as follows:

$$Loss(\mathbf{x}_i, \mathbf{d}_i, \mathbf{f}) = \sum_{l=1}^{p} \Big(d_{\mathbf{x}_i}^{y_l} \ln \big(\frac{d_{\mathbf{x}_i}^{y_l}}{f_l(\mathbf{x}_i)} \big) \Big), \tag{2}$$

where $d_{\mathbf{x}_i}^{y_l}$ and $f_l(\mathbf{x}_i)$ denote the true and predicted description degrees of label y_l to instance $\mathbf{x}_i$, respectively.

Similar to previous works, we assume each sub-learner f_l follows a maximum entropy model [Berger *et al.*, 1996], i.e.,

$$f_l(\mathbf{x}_i) := p(y_l|\mathbf{x}_i; \boldsymbol{\theta}) = \frac{\exp(\sum_{k=1}^{m} \boldsymbol{\theta}_{lk} x_i^k)}{\sum_{l=1}^{p} \exp(\sum_{k=1}^{m} \boldsymbol{\theta}_{lk} x_i^k)}, \tag{3}$$

where x_i^k is the k-th $(k = 1, 2, \ldots, m)$ feature of instance $\mathbf{x}_i$, and $\boldsymbol{\theta}_{lk}$ is a coefficient with respect to the k-th feature and the l-th label in feature coefficient matrix $\boldsymbol{\theta} \in \mathbb{R}^{p \times m}$. For the second term of Eq.(1), we implement the regularization term by F-*norm* as follows:

$$\Omega(\mathbf{f}) = ||\boldsymbol{\theta}||_F^2 = \sum_{l=1}^{p} \sum_{k=1}^{m} \boldsymbol{\theta}_{lk}^2. \tag{4}$$

By substituting Eqs. (2), (3) and (4) into Eq. (1), we can obtain the following learning error:

$$\min_{\boldsymbol{\theta}} \sum_{i=1}^{n} \sum_{l=1}^{p} \Big(d_{\mathbf{x}_i}^{y_l} \ln \big(\frac{d_{\mathbf{x}_i}^{y_l}}{p(y_l|\mathbf{x}_i; \boldsymbol{\theta})} \big) \Big) + \delta ||\boldsymbol{\theta}||_F^2. \tag{5}$$

To alleviate the possible noisy corruption in the original input space $\mathcal{X}$, we create a mapping $\phi : \mathcal{X} \to \mathcal{Z}$ from the m-dimensional $\mathcal{X}$ to the c-dimensional latent semantic feature space $\mathcal{Z}$ as $\phi(\mathbf{x}_i) = \mathbf{W}^{\mathrm{T}} \mathbf{x}_i$, where $\mathbf{W} \in \mathbb{R}^{m \times c}$ is a feature coefficient matrix. Meanwhile, to eliminate the irrelevant and redundant features, some original features most related to $\mathcal{Z}$ are selected by $||\mathbf{W}||_{2,1}$. Note that the local geometry structures should be consistent between the latent semantic feature space $\mathcal{Z}$ and the label space $\mathcal{Y}$. That is to say, if two instances $\mathbf{x}_i$ and $\mathbf{x}_{i'}$ are close to each other in $\mathcal{Y}$, then they should also be close to each other in $\mathcal{Z}$. Thus, we try to minimize the following:

$$\begin{aligned} \min_{\mathbf{Z}, \mathbf{W}} \Big\{ & \alpha ||\mathbf{X}\mathbf{W} - \mathbf{Z}||_F^2 + \beta ||\mathbf{W}||_{2,1} \\ & + \gamma \sum_{i=1}^{n} \sum_{i'=1}^{n} \mathbf{S}_{ii'} ||\phi(\mathbf{x}_i) - \phi(\mathbf{x}_{i'})||_2^2 \Big\}, \end{aligned} \tag{6}$$

where $\mathbf{Z} = [\phi(\mathbf{x}_1), \phi(\mathbf{x}_2), \ldots, \phi(\mathbf{x}_n)]^{\mathrm{T}} \in \mathbb{R}^{n \times c}$ is a latent semantic feature matrix, and the ℓ_2-*norm* of each row of $\mathbf{W}$, i.e., $||\mathbf{W}_{k:}||_2$, denotes the significance of the k-th feature in approximating the latent semantic feature space $\mathcal{Z}$. With the $\ell_{2,1}$-*norm* regularization term $||\mathbf{W}||_{2,1} = \sum_{k=1}^{m} ||\mathbf{W}_{k:}||_2$, $\mathbf{W}$

becomes the row-sparsity and can eliminate the insignificant features in $\mathcal{X}$ when constructing $\mathcal{Z}$. The parameters α, β and γ make a balance among the reconstruction error of latent semantic feature space, the sparsity of feature mapping model and the degree of preserving local geometry structures. Moreover, $\mathbf{S}_{ii'}$ denotes the similarity between label distributions of $\mathbf{x}_i$ and $\mathbf{x}_{i'}$, and it is defined as follows:

$$\mathbf{S}_{ii'} = \begin{cases} \exp(-\frac{||\mathbf{d}_i - \mathbf{d}_{i'}||_2^2}{\sigma^2}) & (\mathbf{d}_i \in \mathcal{N}_\rho(\mathbf{d}_{i'}) \;||\; \mathbf{d}_{i'} \in \mathcal{N}_\rho(\mathbf{d}_i)); \\ 0 & (\text{otherwise}), \end{cases} \tag{7}$$

where $\mathcal{N}_\rho(\mathbf{d}_i)$ denotes the ρ-nearest neighbors of instance $\mathbf{x}_i$ in the label space $\mathcal{Y}$.

And then, we propose to minimize learning error and implement feature selection simultaneously under the guidance of latent semantics via the following optimization objective:

$$\min_{\boldsymbol{\theta},\mathbf{Z},\mathbf{W}} \Big\{ \sum_{i=1}^{n}\sum_{l=1}^{p} \Big(d_{\mathbf{x}_i}^{y_l} \ln\big(\frac{d_{\mathbf{x}_i}^{y_l}}{p(y_l|\phi(\mathbf{x}_i);\boldsymbol{\theta})}\big)\Big) + \alpha||\mathbf{XW}-\mathbf{Z}||_F^2 + \beta||\mathbf{W}||_{2,1} + \gamma \sum_{i=1}^{n}\sum_{i'=1}^{n} \mathbf{S}_{ii'}||\phi(\mathbf{x}_i)-\phi(\mathbf{x}_{i'})||_2^2 + \delta||\boldsymbol{\theta}||_F^2 \Big\} \tag{8}$$

with

$$p(y_l|\phi(\mathbf{x}_i);\boldsymbol{\theta}) = \frac{\exp(\sum_{j=1}^{c}\boldsymbol{\theta}_{lj}\phi(\mathbf{x}_i)_j)}{\sum_{l=1}^{p}\exp(\sum_{j=1}^{c}\boldsymbol{\theta}_{lj}\phi(\mathbf{x}_i)_j)}, \tag{9}$$

where $\boldsymbol{\theta}_{lj}$ is a coefficient with respect to the j-th ($j = 1, 2, \ldots, c$) latent semantic feature $\phi(\mathbf{x}_i)_j$ and the l-th label y_l in the latent semantic feature coefficient matrix $\boldsymbol{\theta} \in \mathbb{R}^{p\times c}$.

2.3 Learning Algorithm for LSE-LDL

In Eq. (8), following [Nie *et al.*, 2010], the regularization term $||\mathbf{W}||_{2,1}$ can be relaxed to be $2\mathrm{Tr}(\mathbf{W}^{\mathrm{T}}\mathbf{G}\mathbf{W})$, where $\mathbf{G}$ is a diagonal matrix with its diagonal element $\mathbf{G}_{kk} = \frac{1}{2\sqrt{\mathbf{W}_{k:}\mathbf{W}_{k:}^{\mathrm{T}}+\epsilon}}$ ($k = 1, 2, \ldots, m$) and ϵ is a small positive constant. Meanwhile, the term of preserving local geometry structures in Eq. (8) can be induced as:

$$\sum_{i=1}^{n}\sum_{i'=1}^{n} \mathbf{S}_{ii'}||\phi(\mathbf{x}_i)-\phi(\mathbf{x}_{i'})||_2^2 = 2\mathrm{Tr}\big(\mathbf{Z}^{\mathrm{T}}(\mathbf{Diag}-\mathbf{S})\mathbf{Z}\big), \tag{10}$$

where $\mathbf{Diag}$ denotes a diagonal matrix with its diagonal element $\mathbf{Diag}_{ii} = \sum_{i'=1}^{n}\mathbf{S}_{ii'}$, and $(\mathbf{Diag}-\mathbf{S})$ is a graph Laplacian matrix.

Thus, the optimization objective can be formulated as follows:

$$\begin{aligned}\Upsilon(\boldsymbol{\theta},\mathbf{Z},\mathbf{W}) = &\sum_{i=1}^{n}\sum_{l=1}^{p}\Big(d_{\mathbf{x}_i}^{y_l}\ln\big(\frac{d_{\mathbf{x}_i}^{y_l}}{p(y_l|\phi(\mathbf{x}_i);\boldsymbol{\theta})}\big)\Big) \\ &+ \alpha||\mathbf{XW}-\mathbf{Z}||_F^2 + 2\beta\mathrm{Tr}(\mathbf{W}^{\mathrm{T}}\mathbf{G}\mathbf{W}) \\ &+ 2\gamma\mathrm{Tr}\big(\mathbf{Z}^{\mathrm{T}}(\mathbf{Diag}-\mathbf{S})\mathbf{Z}\big) + \delta||\boldsymbol{\theta}||_F^2. \end{aligned} \tag{11}$$

To minimize the optimization objective $\Upsilon(\boldsymbol{\theta},\mathbf{Z},\mathbf{W})$, we adopt an Alternating-Updating framework. Specifically, in each iteration, we update one of the variables $\{\boldsymbol{\theta},\mathbf{Z},\mathbf{W}\}$ while fixing the other two variables.

When we fix $\mathbf{Z}$ and $\mathbf{W}$ to solve $\boldsymbol{\theta}$, the second, third and fourth terms of Eq. (11) are constants and thus can be ignored, then Eq. (11) can be rewritten as:

$$\Upsilon(\boldsymbol{\theta},\mathbf{Z},\mathbf{W}) = \sum_{i=1}^{n}\sum_{l=1}^{p}\Big(d_{\mathbf{x}_i}^{y_l}\ln\big(\frac{d_{\mathbf{x}_i}^{y_l}}{p(y_l|\phi(\mathbf{x}_i);\boldsymbol{\theta})}\big)\Big) + \delta||\boldsymbol{\theta}||_F^2. \tag{12}$$

We optimize Eq. (12) with the gradient descent method and the gradient of $\Upsilon(\boldsymbol{\theta},\mathbf{Z},\mathbf{W})$ w.r.t. $\boldsymbol{\theta}$ is

$$\frac{\partial\Upsilon}{\partial\boldsymbol{\theta}_{lj}} = \sum_{i=1}^{n}\Big[\frac{\exp(\boldsymbol{\theta}_{l:}\phi(\mathbf{x}_i))\phi(\mathbf{x}_i)_j}{\sum_{l=1}^{p}\exp(\boldsymbol{\theta}_{l:}\phi(\mathbf{x}_i))} - \phi(\mathbf{x}_i)_j d_{\mathbf{x}_i}^{y_l}\Big] + 2\delta\boldsymbol{\theta}_{lj}. \tag{13}$$

When we fix $\boldsymbol{\theta}$ and $\mathbf{W}$ to solve $\mathbf{Z}$, the third and fifth terms of Eq. (11) are constants and thus can be ignored, then Eq. (11) can be rewritten as:

$$\begin{aligned}\Upsilon(\boldsymbol{\theta},\mathbf{Z},\mathbf{W}) = &\sum_{i=1}^{n}\sum_{l=1}^{p}\Big(d_{\mathbf{x}_i}^{y_l}\ln\big(\frac{d_{\mathbf{x}_i}^{y_l}}{p(y_l|\phi(\mathbf{x}_i);\boldsymbol{\theta})}\big)\Big) + \\ &\alpha||\mathbf{XW}-\mathbf{Z}||_F^2 + 2\gamma\mathrm{Tr}\big(\mathbf{Z}^{\mathrm{T}}(\mathbf{Diag}-\mathbf{S})\mathbf{Z}\big). \end{aligned} \tag{14}$$

Similarly, we optimize Eq. (14) with the gradient descent method and the gradient of $\Upsilon(\boldsymbol{\theta},\mathbf{Z},\mathbf{W})$ w.r.t. $\mathbf{Z}$ is

$$\begin{aligned}\frac{\partial\Upsilon}{\partial\mathbf{Z}_{ij}} = &\Big[-\sum_{l=1}^{p} d_{\mathbf{x}_i}^{y_l}\boldsymbol{\theta}_{lj} + \frac{\sum_{l=1}^{p}\big(\exp(\boldsymbol{\theta}_{l:}\phi(\mathbf{x}_i))\boldsymbol{\theta}_{lj}\big)}{\sum_{l=1}^{p}\exp(\boldsymbol{\theta}_{l:}\phi(\mathbf{x}_i))}\Big] \\ &+ 2\alpha(\mathbf{Z}_{ij} - \mathbf{X}_{i:}\mathbf{W}_{:j}) + 2\gamma\big[(\mathbf{Diag}_{i:} - \mathbf{S}_{i:}) \\ &+ (\mathbf{Diag}_{:i} - \mathbf{S}_{:i})^{\mathrm{T}}\big]\mathbf{Z}_{:j}. \end{aligned} \tag{15}$$

When we fix $\boldsymbol{\theta}$ and $\mathbf{Z}$ to solve $\mathbf{W}$, the first, fourth and fifth terms of Eq. (11) are constants and thus can be ignored, then Eq. (11) can be rewritten as:

$$\Upsilon(\boldsymbol{\theta},\mathbf{Z},\mathbf{W}) = \alpha||\mathbf{XW}-\mathbf{Z}||_F^2 + 2\beta\mathrm{Tr}(\mathbf{W}^{\mathrm{T}}\mathbf{G}\mathbf{W}). \tag{16}$$

Again, we optimize Eq. (16) with the gradient descent method and the gradient of $\Upsilon(\boldsymbol{\theta},\mathbf{Z},\mathbf{W})$ w.r.t. $\mathbf{W}$ is

$$\frac{\partial\Upsilon}{\partial\mathbf{W}_{kj}} = 2\alpha(\mathbf{X}_{:k})^{\mathrm{T}}(\mathbf{XW}_{:j} - \mathbf{Z}_{:j}) + 4\beta\mathbf{G}_{k:}\mathbf{W}_{:j}. \tag{17}$$

The pseudo codes of LSE-LDL are presented in Algorithm 1 and Algorithm 2, which correspond to the training phase and the testing phase, respectively. In Algorithm 1, the coefficient matrices $\boldsymbol{\theta}$ and $\mathbf{W}$ are initialized with all elements being the random values in the interval $[0, 1]$. The latent semantic feature matrix $\mathbf{Z}$ is initialized with the clustering centers of feature clustering in the original feature space $\mathcal{X}$, and *k-means clustering* is used in this paper. The means of initializing $\mathbf{Z}$ is reasonable since the combinations of some highly correlated features can be considered as one type of latent semantics. Then, an Alternating-Updating framework is used to search the optimum $\boldsymbol{\theta}$, $\mathbf{Z}$, and $\mathbf{W}$. The optimization objective in Eq. (11) is a convex function, thus it will converge to a global minimum. Finally, we can obtain the ranking of the most discriminative original physical/chemical features. In Algorithm 2, we will encode the latent semantics for a given unseen instance $\mathbf{x}_i'$ only via its Q most discriminative original features $[\mathbf{x}_i']_Q$ and their corresponding components $[\mathbf{W}]_Q$ in the optimum $\mathbf{W}$, and some other irrelevant and redundant features are ignored. Based on this, the label distribution of unseen instance is predicted by the maximum entropy model.

Algorithm 1 LSE-LDL (Training Phase)

Input: Training data matrix $\mathbf{X}$, label distribution matrix $\mathbf{D}$;
Parameter: α, β, γ, δ, $\lambda_{\boldsymbol{\theta}}$, $\lambda_{\mathbf{Z}}$, $\lambda_{\mathbf{W}}$ and ϵ;
Output: Q top-ranked features;

1: Initialize $\boldsymbol{\theta}$, $\mathbf{Z}$, $\mathbf{W}$, and Calculate graph laplacian matrix ($\mathbf{Diag} - \mathbf{S}$) from $\mathbf{D}$ by Eq. (7);
2: **Repeat:**
3: Calculate $\frac{\partial \Upsilon}{\partial \boldsymbol{\theta}_{lj}}$ by Eq. (13), and Update $\boldsymbol{\theta}$ by: $\boldsymbol{\theta}_{lj} = \boldsymbol{\theta}_{lj} - \lambda_{\boldsymbol{\theta}} \frac{\partial \Upsilon}{\partial \boldsymbol{\theta}_{lj}}$;
4: Calculate $\frac{\partial \Upsilon}{\partial \mathbf{Z}_{ij}}$ by Eq. (15), and Update $\mathbf{Z}$ by: $\mathbf{Z}_{ij} = \mathbf{Z}_{ij} - \lambda_{\mathbf{Z}} \frac{\partial \Upsilon}{\partial \mathbf{Z}_{ij}}$;
5: Calculate diagonal matrix $\mathbf{G}$ by: $\mathbf{G}_{kk} = \frac{1}{2\sqrt{\mathbf{W}_{k:}\mathbf{W}_{k:}^{\mathrm{T}}+\epsilon}}$;
6: Calculate $\frac{\partial \Upsilon}{\partial \mathbf{W}_{kj}}$ by Eq. (17), and Update $\mathbf{W}$ by: $\mathbf{W}_{kj} = \mathbf{W}_{kj} - \lambda_{\mathbf{W}} \frac{\partial \Upsilon}{\partial \mathbf{W}_{kj}}$;
7: **Until** *convergence or maximum number of iterations*
8: Rank features by $||\mathbf{W}_{k:}||_2$ in a descending order;
9: **return** Q top-ranked features, $\boldsymbol{\theta}$ and $\mathbf{W}$.

Algorithm 2 LSE-LDL (Testing Phase)

Input: Unseen instance $\mathbf{x}'_i$, $\boldsymbol{\theta}$, $\mathbf{W}$ and Q selected features;
Output: Predicted $\mathbf{d}'_i$ of $\mathbf{x}'_i$;

1: Calculate Q-latent semantic feature vector $\phi_Q(\mathbf{x}'_i)$ of $\mathbf{x}'_i$ as: $\phi_Q(\mathbf{x}'_i) = [\mathbf{W}]_Q{}^{\mathrm{T}}[\mathbf{x}'_i]_Q$;
2: $p(y_l|\phi_Q(\mathbf{x}'_i);\boldsymbol{\theta}) = \frac{\exp(\boldsymbol{\theta}_{l:}\phi_Q(\mathbf{x}'_i))}{\sum_{l=1}^{p}\exp(\boldsymbol{\theta}_{l:}\phi_Q(\mathbf{x}'_i))}$;
3: **return** $\mathbf{d}'_i = [p(y_1|\phi_Q(\mathbf{x}'_i);\boldsymbol{\theta}), \ldots, p(y_p|\phi_Q(\mathbf{x}'_i);\boldsymbol{\theta})]^{\mathrm{T}}$.

3 Experiments

In this section, we evaluate the proposed LSE-LDL algorithm on 15 publicly available data sets with six state-of-the-art LDL algorithms over six different evaluation metrics. We implement all LDL algorithms in Matlab R2017b. All the experiments were carried out on a workstation equipped with an Intel Core i7 − 6850K CPU (3.60 GHz) and 32.00 GB memory.

Data sets The 15 data sets are coming from LDL website (http://ldl.herokuapp.com/download). Table 1 summarizes some brief statistics of these data sets, and detailed descriptions of them can be found in [Geng, 2016].

Evaluations Following [Geng, 2016], six widely-used metrics are employed to evaluate the performance of LDL algorithms, including four distance metrics between two distributions (*the lower the value of metric, the better the performance*), i.e., `Chebyshev`, `Clark`, `Canberra`, and `Kullback-Leibler divergence`; and two similarity metrics between two distributions (*the higher the value of metric, the better the performance*), i.e., `Cosine` and `Intersection`.

Baselines and Settings We compare the proposed LSE-LDL algorithm to six state-of-the-art LDL algorithms, including AA-BP [Geng *et al.*, 2013], SA-IIS [Geng *et al.*, 2010], SA-BFGS [Geng *et al.*, 2013], LDLLC [Jia *et al.*, 2018], LALOT [Zhao and Zhou, 2018], and PT-SVM [Geng and Ji, 2013]. All the codes of above compared algorithms are shared by original authors, and we set all parameters to be default values as recommended in original papers. In LSE-LDL, to model the local geometry structures in the latent semantic feature space, σ and ρ are set to be 0.05 and 1% of training instances, respectively. The number of selected features $Q = m$. The regularization parameters in LSE-LDL are tuned with a grid-search strategy by varying their values in the range of $\{0.001, 0.01, 0.1, 1.0, 10\}$. The maximum number of iterations is 5000, and the small positive constant $\epsilon = 0.0001$.

ID	Data sets	n	m	p
1	Yeast-alpha	2,465	24	18
2	Yeast-cdc	2,465	24	15
3	Yeast-cold	2,465	24	4
4	Yeast-diau	2,465	24	7
5	Yeast-dtt	2,465	24	4
6	Yeast-elu	2,465	24	14
7	Yeast-heat	2,465	24	6
8	Yeast-spo	2,465	24	6
9	Yeast-spo5	2,465	24	3
10	Yeast-spoem	2,465	24	2
11	Human Gene	17,892	36	68
12	Natural Scene	2,000	294	9
13	S-JAFFE	213	243	6
14	S-BU_3DFE	2,500	243	6
15	Movie	7,755	1,869	5

Table 1: Statistics of the 15 LDL data sets, where n is number of instances, m is number of features and p is number of labels.

Results On each data set, *ten-fold cross-validation* is employed for the performance evaluation, and the mean value and the standard deviation of ten experimental results are respectively recorded. Among all comparing LDL algorithms, the best predictive performance is highlighted in boldface. In addition, if an LDL algorithm cannot deal with a given data set, its predictive performance is expressed by the symbol of $\times$.

There are originally six different evaluation metrics, due to space limitation, we only present predictive performances w.r.t. `Clark distance` and `Canberra distance` in Table 2 and Table 3, respectively. Predictive performances on other evaluation metrics are similar and therefore omitted.

As shown in Table 2 and Table 3, in all the different evaluation metrics, the proposed LSE-LDL algorithm achieves superior predictive performances against all six compared algorithms. As a whole, across 30 predictive performance results (15 data sets $\times$ 2 evaluation metrics), LSE-LDL ranks in first place among seven comparing algorithms at 70.00% cases, in second place at 16.67% cases, in third place at only 6.67% cases. The results are expected since our algorithm can encode advanced latent semantics for reducing the negative effects of irrelevant, redundant and noisy features, and these semantic features are more relevant to learning the label distribution.

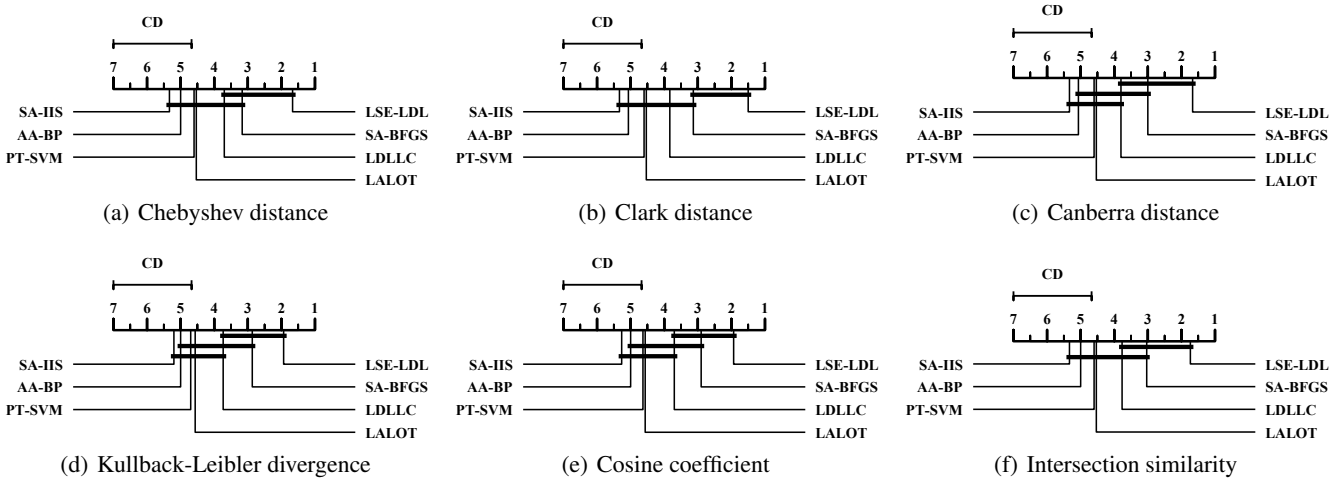
To perform comparisons of predictive performances in

ID	AA-BP	SA-IIS	SA-BFGS	LDLLC	LALOT	PT-SVM	LSE-LDL
1	.8284 ± .5093	.3034 ± .0942	.2099 ± .0824	.2105 ± .0821	.2241 ± .0806	.2204 ± .0807	**.2094 ± .0823**
2	.5887 ± .3459	.2926 ± .1057	.2159 ± .0989	×	.2285 ± .1012	.2262 ± .1013	**.2155 ± .0988**
3	.1543 ± .0855	.1651 ± .0870	.1396 ± .0795	**.1393 ± .0795**	.1491 ± .0838	.1529 ± .0838	**.1393 ± .0793**
4	.2745 ± .1429	.2419 ± .1057	.2008 ± .1034	.2009 ± .1034	.2241 ± .1091	.2401 ± .1052	**.2007 ± .1027**
5	.1201 ± .0729	.1313 ± .0754	.0981 ± .0622	.0981 ± .0623	.1037 ± .0632	.1024 ± .0630	**.0980 ± .0622**
6	.5253 ± .3132	.2756 ± .0915	.1991 ± .0780	×	.2109 ± .0773	.2101 ± .0776	**.1988 ± .0776**
7	.2201 ± .1005	.2240 ± .0906	.1828 ± .0833	.1825 ± .0836	.1889 ± .0867	.1907 ± .0857	**.1823 ± .0834**
8	.2892 ± .1356	.2786 ± .1264	.2499 ± .1270	.2495 ± .1271	.2581 ± .1347	.2725 ± .1310	**.2489 ± .1277**
9	.1888 ± .1243	.1942 ± .1256	.1842 ± .1237	.1840 ± .1239	.1859 ± .1246	.1891 ± .1249	**.1839 ± .1241**
10	×	.1373 ± .1101	.1291 ± .1054	**.1290 ± .1054**	.1363 ± .1134	.1338 ± .1144	.1294 ± .1048
11	3.358 ± .8518	2.128 ± 1.240	**2.108 ± 1.246**	×	×	×	2.116 ± 1.244
12	2.463 ± .3001	2.468 ± .3155	×	×	2.493 ± .3428	2.592 ± .2603	**2.398 ± .2663**
13	.4153 ± .1312	.4587 ± .1301	×	×	.4295 ± .1224	.4446 ± .1336	**.3387 ± .1151**
14	.3993 ± .1553	.4191 ± .1535	×	×	.4535 ± .1578	.4427 ± .1591	**.3924 ± .1450**
15	.5618 ± .2712	.5801 ± .2558	.5541 ± .2837	**.5138 ± .2640**	×	.8530 ± .2453	.7792 ± .3970

Table 2: Predictive performances (mean ± std.) on the 15 LDL data sets evaluated by `Clark distance` ↓.

ID	AA-BP	SA-IIS	SA-BFGS	LDLLC	LALOT	PT-SVM	LSE-LDL
1	2.749 ± 1.719	1.014 ± .3208	.6817 ± .2604	.6835 ± .2588	.7337 ± .2540	.7198 ± .2555	**.6799 ± .2604**
2	1.785 ± 1.021	.8977 ± .3128	.6475 ± .2725	×	.6860 ± .2820	.6800 ± .2842	**.6458 ± .2717**
3	.2662 ± .1471	.2861 ± .1521	.2402 ± .1356	.2399 ± .1357	.2571 ± .1440	.2645 ± .1446	**.2397 ± .1354**
4	.5926 ± .2999	.5284 ± .2262	**.4311 ± .2150**	.4314 ± .2147	.4831 ± .2265	.5136 ± .2173	.4312 ± .2145
5	.2069 ± .1224	.2282 ± .1297	.1688 ± .1018	.1688 ± .1019	.1782 ± .1036	.1762 ± .1034	**.1685 ± .1018**
6	1.544 ± .8953	.8241 ± .2710	**.5831 ± .2127**	×	.6231 ± .2125	.6198 ± .2138	.5834 ± .2113
7	.4425 ± .1956	.4551 ± .1810	.3645 ± .1584	.3639 ± .1587	.3771 ± .1654	.3823 ± .1647	**.3638 ± .1586**
8	.5943 ± .2778	.5733 ± .2588	.5137 ± .2596	.5130 ± .2598	.5330 ± .2808	.5624 ± .2750	**.5127 ± .2620**
9	.2900 ± .1857	.2992 ± .1897	.2829 ± .1841	**.2826 ± .1845**	.2857 ± .1864	.2904 ± .1870	.2828 ± .1852
10	×	.1913 ± .1495	.1796 ± .1423	**.1795 ± .1423**	.1900 ± .1543	.1865 ± .1555	.1801 ± .1414
11	22.89 ± 7.037	14.58 ± 9.779	**14.43 ± 9.821**	×	×	×	14.48 ± 9.811
12	6.784 ± 1.263	6.800 ± 1.352	×	×	7.033 ± 1.396	7.375 ± 1.088	**6.604 ± 1.053**
13	.8644 ± .2950	.9481 ± .2865	×	×	.9020 ± .2859	.9216 ± .3018	**.7014 ± .2541**
14	.8543 ± .3544	.9056 ± .3610	×	×	.9593 ± .3602	.9362 ± .3531	**.8342 ± .3257**
15	1.069 ± .5520	1.115 ± .5256	1.068 ± .5880	**.9827 ± .5383**	×	1.651 ± .5192	1.528 ± .8391

Table 3: Predictive performances (mean ± std.) on the 15 LDL data sets evaluated by `Canberra distance` ↓.



(a) Chebyshev distance (b) Clark distance (c) Canberra distance

(d) Kullback-Leibler divergence (e) Cosine coefficient (f) Intersection similarity

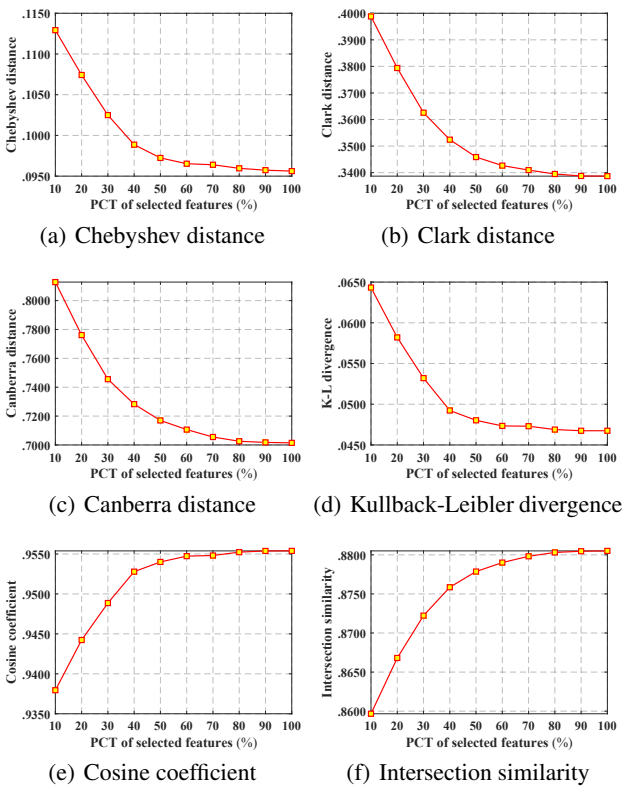
Figure 1: CD diagrams of the comparing LDL algorithms on each evaluation metric.

more well-founded ways, *Friedman test* [Friedman, 1940] is further examined which is a favorable statistical test for comparisons of more than two algorithms over multiple data sets. For each evaluation metric, Friedman test at 0.05 significance level rejects the null hypothesis of "equal" performances among all seven comparing algorithms, and then we adopt *Nemenyi test* to further analyze which algorithms actually differ. The performances of the two algorithms are significantly different if the corresponding average ranks over all the data sets differ by at least one *critical difference* (CD = 2.3254).

To visually show the actual differences of predictive performances among seven comparing LDL algorithms, the CD diagrams [Demšar, 2006] on each evaluation metric are illustrated in Figure 1, where the average rank of each comparing LDL algorithm is marked along the axis (higher ranks to the left). In each metric, if a group of algorithms is not significantly different under Nemenyi test, we will connect them with a thick line.

In summary, out of all 36 comparisons of predictive performances (6 baseline algorithms $\times$ 6 evaluation metrics), our LSE-LDL achieves the statistically comparable predictive performances in only 30.56% cases, and in all the other 69.44% cases, LSE-LDL achieves the statistically superior predictive performances. Besides, no algorithms have outperformed LSE-LDL. The above statistics suggest the statistical superior predictive performances of LSE-LDL as compared to all the other state-of-the-art LDL algorithms.

In order to examine the effectiveness of LSE-LDL in removing some irrelevant and redundant features, we also analyze the influence of the number of selected features for encoding the latent semantics. We run the testing phase of LSE-LDL with different Q varying from 10% to 100% of the original number of features (Stepsize: 10%). Due to space limitation, we only present the predictive performances on *S-JAFFE* with six evaluation metrics. As shown in Figure 2, it is obvious that the performances w.r.t all six evaluation metrics tend to be stable when the percentage (PCT) of selected features is more than 60%. It means that there are 40% irrelevant and redundant features in *S-JAFFE* tasks, and LSE-LDL is effective in reducing storage requirements and computational overheads with superior predictive performances.



(a) Chebyshev distance (b) Clark distance

(c) Canberra distance (d) Kullback-Leibler divergence

(e) Cosine coefficient (f) Intersection similarity

Figure 2: Influence of Q with 6 metrics on *S-JAFFE*

4 Related Work

Over the past several years, several algorithms designed for LDL have been witnessed. Generally, the existing algorithms can be grouped into the three main categories, including algorithm adaptation approaches, problem transformation approaches, and specialized approaches. Algorithm adaptation approaches directly modify some constraint conditions in traditional SLL/MLL algorithms to handle LDL tasks, such as AA-kNN [Geng *et al.*, 2010] and AA-BP [Geng *et al.*, 2013]. Problem transformation approaches transform the LDL task into some corresponding SLL tasks by weighted sampling on training instances, and then handle such SLL tasks through binary classifiers SVM or Naive Bayes, developing two representative problem transformation approaches, i.e., PT-SVM and PT-Bayes [Geng, 2016]. Specialized approaches focus on learning a non-linear conditional probability mass function by directly maximizing/minimizing the entropy/distance between the true and predicted label distributions. As two representative specialized approaches, both of SA-IIS and SA-BFGS [Geng *et al.*, 2013] construct a maximum entropy model with employing Kullback-Leibler divergence as the objective function, and then implement optimization by improved iterative scaling (IIS) [Della Pietra *et al.*, 1997] and quasi-Newton method BFGS [Nocedal and Wright, 2006], respectively. The related survey [Geng, 2016] has shown that specialized approaches are more effective than others in real-world LDL tasks, thus, more recently, researchers are all dedicated to designing highly competitive specialized approaches with considering label correlations, such as LALOT [Zhao and Zhou, 2018] and LDLLC [Jia *et al.*, 2018].

5 Conclusion

In this paper, a novel algorithm called LSE-LDL is proposed, and it can encode the latent semantics for LDL tasks in order to alleviate the negative effects of irrelevant, redundant, and noisy features. The experimental results on 15 real-world LDL data sets validate the effectiveness of the proposed LSE-LDL algorithm.

Acknowledgments

This work is supported by the Natural Science Foundation of China (Nos. 61672276, 61572242), Natural Science Foundation of Jiangsu Province of China (No. BK20161406).

References

[Berger *et al.*, 1996] Adam L. Berger, Vincent J. Della Pietra, and Stephen A. Della Pietra. A maximum entropy approach to natural language processing. *Computational Linguistics*, 22(1):39–71, 1996.

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

[Cha, 2007] Sung-Hyuk Cha. Comprehensive survey on distance/similarity measures between probability density functions. *International Journal of Mathematical Models and Methods in Applied Sciences*, 1(4):300–307, 2007.

[Della Pietra *et al.*, 1997] Stephen A. Della Pietra, Vincent J. Della Pietra, and John D. Lafferty. Inducing features of random fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(4):380–393, 1997.

[Demšar, 2006] Janez Demšar. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, 7:1–30, 2006.

[Friedman, 1940] Milton Friedman. A comparison of alternative tests of significance for the problem of m rankings. *The Annals of Mathematical Statistics*, 11(1):86–92, 1940.

[Geng and Hou, 2015] Xin Geng and Peng Hou. Pre-release prediction of crowd opinion on movies by label distribution learning. In *Proceedings of the 24th International Joint Conference on Artificial Intelligence (IJCAI'15)*, pages 3511–3517, Buenos Aires, Argentina, July 2015.

[Geng and Ji, 2013] Xin Geng and Rongzi Ji. Label distribution learning. In *Proceedings of the 13th IEEE International Conference on Data Mining Workshops (ICDM'13)*, pages 377–383, Dallas, Texas, USA, December 2013.

[Geng *et al.*, 2010] Xin Geng, Kate Smith-Miles, and Zhi-Hua Zhou. Facial age estimation by learning from label distributions. In *Proceedings of the 24nd AAAI Conference on Artificial Intelligence (AAAI'10)*, pages 451–456, Atlanta, Georgia, USA, July 2010.

[Geng *et al.*, 2013] Xin Geng, Chao Yin, and Zhi-Hua Zhou. Facial age estimation by learning from label distributions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(10):2401–2412, 2013.

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

[Gibaja Galindo and Ventura, 2014] Eva L. Gibaja Galindo and Sebastián Ventura. Multi-label learning: a review of the state of the art and ongoing research. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 4(6):411–444, 2014.

[Guo *et al.*, 2019] Yumeng Guo, Fulai Chung, Guozheng Li, Jiancong Wang, and James C. Gee. Leveraging label-specific discriminant mapping features for multi-label learning. *ACM Transactions on Knowledge Discovery from Data*, 13(2):24:1–24:23, 2019.

[Huang and Zhou, 2012] Sheng-Jun Huang and Zhi-Hua Zhou. Multi-label learning by exploiting label correlations Locally. In *Proceedings of the 26th AAAI Conference on Artificial Intelligence (AAAI'12)*, pages 949–955, Toronto, Ontario, Canada, July 2012.

[Jia *et al.*, 2018] Xiuyi Jia, Weiwei Li, Junyu Liu, and Yu Zhang. Label distribution learning by exploiting label correlations. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence (AAAI'18)*, pages 3310–3317, New Orleans, Louisiana, USA, February 2018.

[Jian *et al.*, 2016] Ling Jian, Jundong Li, Kai Shu, and Huan Liu. Multi-label informed feature selection. In *Proceedings of the 25th International Joint Conference on Artificial Intelligence (IJCAI'16)*, pages 1627–1633, New York, NY, USA, July 2016.

[Liu and Motoda, 1998] Huan Liu and Hiroshi Motoda. *Feature Selection for Knowledge Discovery and Data Mining*. Kluwer Academic Publishers, Norwell, MA, USA, 1998.

[Nie *et al.*, 2010] Feiping Nie, Heng Huang, Xiao Cai, and Chris H. Ding. Efficient and robust feature selection via joint $\ell_{2,1}$-norms minimization. In *Proceedings of the 23rd International Conference on Neural Information Processing Systems (NIPS'10)*, pages 1813–1821, Vancouver, British Columbia, Canada, December 2010.

[Nocedal and Wright, 2006] Jorge Nocedal and Stephen J. Wright. *Numerical Optimization, 2nd edition*. Springer, New York, NY, USA, 2006.

[Xu *et al.*, 2016] Suping Xu, Xibei Yang, Hualong Yu, Dong-Jun Yu, Jingyu Yang, and Eric C. C. Tsang. Multi-label learning with label-specific feature reduction. *Knowledge-Based Systems*, 104:52–61, 2016.

[Yu and Liu, 2004] Lei Yu and Huan Liu. Efficient feature selection via analysis of relevance and redundancy. *Journal of Machine Learning Research*, 5:1205–1224, 2004.

[Zhang and Zhou, 2014] Min-Ling Zhang and Zhi-Hua Zhou. A review on multi-label learning algorithms. *IEEE Transactions on Knowledge and Data Engineering*, 26(8):1819–1837, 2014.

[Zhang *et al.*, 2018] Min-Ling Zhang, Yu-Kun Li, Xu-Ying Liu, and Xin Geng. Binary relevance for multi-label learning: an overview. *Frontiers of Computer Science*, 12(2):191–202, 2018.

[Zhao and Zhou, 2018] Peng Zhao and Zhi-Hua Zhou. Label distribution learning by optimal transport. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence (AAAI'18)*, pages 4506–4513, New Orleans, Louisiana, USA, February 2018.

[Zhou *et al.*, 2015] Ying Zhou, Hui Xue, and Xin Geng. Emotion distribution recognition from facial expressions. In *Proceedings of the 23rd ACM international conference on Multimedia (MM'15)*, pages 1247–1250, Brisbane, Australia, October 2015.