The Thirty-Fourth AAAI Conference on Artificial Intelligence (AAAI-20)

Label Enhancement with Sample Correlations via Low-Rank Representation

Haoyu Tang,[1] **Jihua Zhu,**[1*] **Qinghai Zheng,**[1] **Jun Wang,**[2] **Shanmin Pang,**[1] **Zhongyu Li**[1]
[1]School of Software Engineering, Xi'an Jiaotong University, Xian 710049, China
[2]Shanghai Institute for Advanced Communication and Data Science,
School of Communication and Information Engineering,Shanghai University, Shanghai 200444, China
tanghao258@stu.xjtu.edu.cn, zhujh@xjtu.edu.cn

Abstract

Compared with single-label and multi-label annotations, label distribution describes the instance by multiple labels with different intensities and accommodates to more-general conditions. Nevertheless, label distribution learning is unavailable in many real-world applications because most existing datasets merely provide logical labels. To handle this problem, a novel label enhancement method, Label Enhancement with Sample Correlations via low-rank representation, is proposed in this paper. Unlike most existing methods, a low-rank representation method is employed so as to capture the global relationships of samples and predict implicit label correlation to achieve label enhancement. Extensive experiments on 14 datasets demonstrate that the algorithm accomplishes state-of-the-art results as compared to previous label enhancement baselines.

Introduction

Recently, a growing number of studies have focused on the challenging label ambiguity problem. Since single-label learning paradigm where one instance is mapped to one single label has been well studied, multi-label learning (MLL) is highlighted to address this issue. During past years, a collection of scenarios have applied this learning process (Chen et al. 2019; Tsoumakas and Katakis 2007; Huang and Zhou 2012; Zhang and Zhou 2013), which simultaneously assigns multiple labels with identical degrees to each instance. In particular, in the supervised-learning process, each instance is described by a label vector where each value, i.e., the logic label, is either 1 or 0, which represents whether the instance belongs to the relevant label or whether it does not, respectively. Since all labels with the same values contribute equally in the label vector, the relative importance among multiple associated labels, which is supposed to be different under most circumstances, cannot be reflected well.

Therefore, despite MLL's success, in some sophisticated semantics such as facial age estimation and facial expression recognition, the performance of primitive MLL is hindered because a model precisely mapping the instance to a real-valued label vector with quantitative description degrees (i.e., label distribution) is required in these tasks. To meet this demand, the learning process for the above-mentioned model called "label distribution learning" (LDL) (Geng 2016) has attracted significant attention. In LDL, an instance is annotated by a label vector, i.e., the label distribution where each value ranging from 0 to 1 is the description degree of the relevant label and all values add up to 1. As many pieces of literature have demonstrated(Gao et al. 2017; Geng, Yin, and Zhou 2013; Zheng, Jia, and Li 2018), label distributions generally describe attributes of samples more precisely because differences between the relative importance of each label exist in most cases, and implicit cues within the label distributions can be effectively leveraged through LDL for reinforcing the supervised training.

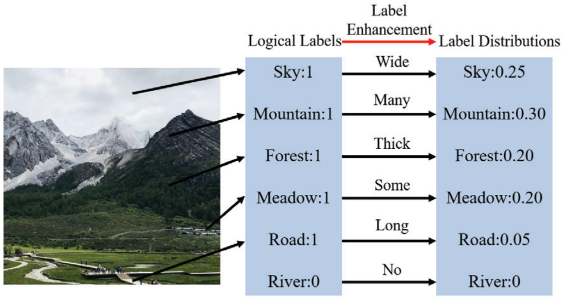


Figure 1: An Example of Label Enhancement

Nevertheless, since manually annotating each instance with label distribution is time-consuming, it is unavailable in most training sets practically(Xu, Lv, and Geng 2019). The requirement of label distribution among different datasets arises some progress in label enhancement (LE), which was proposed by (Xu, Tao, and Geng 2018). It is a pre-processing of the training set, where label distributions are recovered from the off-the-shelf logical labels and the implicit information of given features, as shown in Fig.1.

This definition indicates that the essence of LE is to excavate the information from two folds: the topological structure of feature space and the relationships among logical labels. Several approaches have been proposed according

[*]Corresponding Author

to this principle. To leverage the knowledge in the feature space, some prior efforts (El Gayar, Schwenker, and Palm 2006) assigned the membership degree of each instance to different labels via Fuzzy Clustering Method (FCM) (Melin and Castillo 2005), whereas others constructed graph structures and similarity matrices transferred into the label space later. However, arbitrary elements of the edges in the graph or within the similarity matrix are calculated by the pairwise method (Li, Zhang, and Geng 2015) or the K-nearest neighbors' (KNN) correlations to a certain instance (Xu, Tao, and Geng 2018; Hou, Geng, and Zhang 2016). The downside of these partial-based processes for the graph construction of each instance is that only local topological features have been taken advantage of, and the holistic information of the feature space has been largely untapped. In addition, these approaches always require prior knowledge for hyperparameters. Specifically, if one tunes the parameters such as parameter K in the KNN part slightly, these algorithms' recovery performance varies on a large scale, which tremendously affects the widespread use of these algorithms.

Toward this end, an approach globally unearthing the global structure of the whole feature space and robust to parameters is expected. Thus far, to meet the aforementioned requirements, a novel Label Enhancement with Sample Correlations via low-rank representation (LESC) algorithm is proposed. More specifically, low-rank representation (LRR), which imposes a low-rank constraint on the data subspace representation to capture the global relationship of all instances, is employed to benefit the LE by exploiting the structure of the feature space from a global perspective (Liu and Yan 2011; Yin, Gao, and Lin 2015; Zhai et al. 2019). Intuitively, the constructed low-rank structure could be generally transferred to the label space smoothly, and the lowest-rank representation of feature space is utilized to represent the LRR of the label distributions. As a result, we incorporate the attained LRR into the objective function to explore the hidden cues in the label distribution space. Consequently, we could obtain the optimal recovered label distribution. Extensive experiments have shown that the proposed LESC algorithm is stable to obtain remarkable performance as we expect.

Our contributions can be summarized as follows:

- A novel LESC algorithm is proposed in this paper.It leverages the global-instances relationship to improve the performance of LE.
- By introducing the LRR in the feature space, the intrinsic structure of the feature space is fully exploited for LE.
- Comprehensive experiments conducted on 14 real-world datasets show excellent power and generation compared with several state-of-the-art methods.

Related Work

Label Enhancement

For the convenience of the description of related works, we declare the fundamental notations in advance. The set of labels is $Y = \{y_1, y_2, \cdots, y_o\}$, where o is the size of the label set. For an instance $x_i \in \mathbb{R}^q$, the logical label is denoted as $L_i = \left(l_{x_i}^{y_1}, l_{x_i}^{y_2}, \cdots, l_{x_i}^{y_o}\right)^T$ and $l_{x_i}^{y} \in \{0,1\}$, while the corresponding label distribution is denoted as:

$$D_i = \left(d_{x_i}^{y_1}, d_{x_i}^{y_2}, \cdots, d_{x_i}^{y_o}\right)^T, s.t., \sum_{m=1}^{o} d_{x_i}^{y_m}{=}1 \qquad (1)$$

where $d_{x_i}^{y}$ depicts the degree to which x_i belongs to label y. The goal of the LE process is to recover the associated label distribution of every instance from logical labels in a given training set.

This issue is raised by (Xu, Tao, and Geng 2018), in which the GLLE algorithm was also proposed for the LE process, but some studies concentrated on the same issue before that-for instance, fuzzy clustering method (Melin and Castillo 2005) is applied in (El Gayar, Schwenker, and Palm 2006), which intends to allocate the description values to each instance over diverse clusters. Specifically, features are clustered into t clusters via fuzzy M-means clustering where c_k denotes the $k-th$ cluster center. The cluster membership $\omega_i = \{\omega_{i1}, \omega_{i2}, \cdots, \omega_{it}\}$ for each instance x_i is obtained by calculating the description value over the center c_k as follows:

$$\omega_{ik} = \frac{1}{\sum_{j=1}^{t}\left(\frac{\|x_i - c_k\|_2}{\|x_i - c_j\|_2}\right)^{\frac{1}{\beta-1}}} \qquad (2)$$

where β is larger than 1. Afterward, a zero matrix $Q \in \mathbb{R}^{o\times t}$ is initialized and it is continuously updated by:

$$Q_j = Q_j + \omega_i, s.t., \; l_{x_i}^{y_j} = 1 \qquad (3)$$

where Q_j denotes the $j-th$ row of Q. They constructed prototype label matrix through which classes and clusters are softly associated. After normalizing the columns and rows of Q to sum to 1, the label distribution is computed for each instance x_i using fuzzy composition: $D_i = Q \circ \omega_i$

In addition, other recent studies have focused on the graph-based approaches to tackle the LE problem. They constructed the similarity matrix Q over the features space via various strategies. Hou, Geng, and Zhang recovered the label distribution according to manifold learning (ML), which ensures them to gradually convert the local structure of the feature space into the label space. In particular, to represent this structure, the similarity matrix Q is established based on the assumption that each feature can be represented by the linear combination of its KNN, which means to minimize:

$$\Phi(Q) = \sum_{i=1}^{n}\left\|x_i - \sum_{j\neq i} q_{ij}x_j\right\|^2 \qquad (4)$$

where $q_{ij} = 1$ if x_j is one of x_i's KNNs; otherwise, $q_{ij} = 0$. They further constrained that $\sum_{j=1}^{n} q_{ij} = 1$ for translation invariance. The constructed graph is transferred into the label space to minimize the distance between the target label distribution and the identical linear combination of its KNN label distributions (Roweis and Saul 2000), which infers the optimization of:

$$\phi(D) = \sum_{i=1}^{n}\left\|D_i - \sum_{j\neq i} q_{ij}D_j\right\|^2 \qquad (5)$$

by adding the constraint of $\forall 1 \leq i \leq n, 1 \leq j \leq o, d_{xi}^{j} l_{i}^{j} \geq \lambda$ where $\lambda > 0$. This formula is minimized with respect to the target label distribution D through a constrained quadratic programming process.

Li, Zhang, and Geng regarded the LE as the label propagation (LP) process (Zhu and Goldberg 2009). The pairwise similarity was calculated over the complete feature space and a fully-connected graph was established as:

$$q_{ij} = \begin{cases} \exp\left(-\frac{\|x_i - x_j\|^2}{2\sigma^2}\right), if\ i \neq j \\ 0, if\ i = j \end{cases} \tag{6}$$

where $\forall\ i, j\ \in [1, n]$ and σ is fixed to be 1. The required LP matrix is built from the formula: $P = \tilde{Q}^{-\frac{1}{2}} Q \tilde{Q}^{-\frac{1}{2}}$ with $\tilde{Q} = \textbf{diag}\left[\tilde{q}_1, \tilde{q}_2, \cdots, \tilde{q}_n\right]$ denoting a diagonal matrix where $\tilde{q}_i$ equals to the sum of $i - th$ row element in Q. Thus far, The LP is iteratively implemented, and it is proved that the recovered label distribution matrix $\mathfrak{D} = [D_1; D_2; \cdots; D_n]$ converges to:

$$\mathfrak{D}^* = (1 - \alpha)\left(I - \alpha P\right)^{-1} \Gamma \tag{7}$$

with α denoting the trade-off parameter that controls the contribution between the label propagation P and the initial logical label matrix Γ.

For the GLLE algorithm, the similarity matrix is also constructed in the feature space by partial topological structure. Different from LP, which calculates the pair-wise distance within the whole feature space, the GLLE algorithm computes the distance between a specific instance and its KNNs to define the relevant element in the similarity matrix as follows:

$$q_{ij} = \begin{cases} \exp\left(-\frac{\|x_i - x_j\|^2}{2\sigma^2}\right), if\ x_j \in K\left(i\right) \\ 0, otherwise \end{cases} \tag{8}$$

where $K\left(i\right)$ is the set of x_i's KNNs. Because of the same intuition that these relationships could be converted into the label distribution space, this constructed graph is incorporated into the label space to attain a matrix linearly transforming the logical labels to the label distributions, obtaining the previous state-of-the-art results. Since we normalize each D_i by the softmax normalization for the above-mentioned algorithms, the condition $\sum_{m=1}^{o} d_{x_i}^{y_m}$=1 can be satisfied.

Because it was fully recognized that establishing the similarity matrix based on pair-wise or local feature structure can hinder these approaches' performances, here, the LRR is introduced to excavate the global information and leverage the attained subspace representation to overcome these drawbacks in the proposed algorithm.

LESC Algorithm

In this section, the details of the LESC algorithm are provided. In a training set $S = \{(x_1, L_1), (x_2, L_2), \cdots, (x_n, L_n)\}$, all instances are vertically concatenated along the column to attain the feature matrix $X = [x_1; x_2; \cdots x_n]$, where $x_i \in \mathbb{R}^q$ and $X \in \mathbb{R}^{q \times n}$. After the LE process, a new LDL training set $\varepsilon = \{(x_1, D_1), (x_2, D_2), \cdots, (x_n, D_n)\}$ can be rehabilitated to implement the LDL process. The logical label matrix $\Gamma = [L_1; L_2; \cdots; L_n]$ and the objective label distribution matrix $\mathfrak{D} = [D_1; D_2; \cdots; D_n]$ are created in the same way. For a given instance x_i, it is necessary to find the optimal parameter to recover the best label distribution. This mapping model is represented as follows:

$$D_i = \phi\left(\hat{\theta}, \xi\left(x_i\right)\right) \tag{9}$$

while $\phi(\hat{\theta}, \cdot)$ denotes a linear transformation parameterized by $\hat{\theta}$, $\xi\left(x\right)$ embeds x in a high-dimensional space where the Gaussian kernel function is determined to be employed. Apparently, we need to induce the minimization of the formula to get an optimal $\hat{\theta}$:

$$\min_{\hat{\theta}} \mathcal{L}\left(\hat{\theta}\right) + \lambda_1 \Psi\left(\hat{\theta}\right) \tag{10}$$

where $\mathcal{L}(\hat{\theta})$ denotes a loss function and $\Psi(\hat{\theta})$ is the function that excavates information in the raw feature space and the correlations among labels with λ_1 denoting the trade-off between them.

Since the prior knowledge of the ground-truth label distribution is unavailable, we establish the loss function between the recovered label distributions and the logical labels. The least-squares (LS) loss function is adopted as the first term in (10):

$$\mathcal{L}\left(\hat{\theta}\right) = \sum_{i=1}^{n} \left\| \phi\left(\hat{\theta}, \xi\left(x_i\right)\right) - L_i \right\|^2 \tag{11}$$

In the LRR, all samples and their global relationships are expressed by the linear combination of a small amount of data, which are the bases in the feature space. Accordingly, this property can be transferred to the label space under general conditions. Therefore, it is expected that the low-rank recovery to the label distribution $\mathfrak{D}$ can be expressed, which means to discover a proper $\mathfrak{D}$ for minimizing the distance between $\mathfrak{D}$ and $\mathfrak{D}\hat{C}$, where $\hat{C}$ is the minimized LRR of the feature space. This leads the second term of the optimization formula (10) to be as follows:

$$\Psi\left(\hat{\theta}\right) = \left\| \mathfrak{D} - \mathfrak{D}\hat{C} \right\|_F^2 = \left\| (I - \hat{C}^T)\mathfrak{D}^T \right\|_F^2 \tag{12}$$

To attain the minimizer $\hat{C}$, we aim at seeking the LRR among the feature matrix to excavate the global structure of feature space, i.e., assuming that $X = XC + E$, it is necessary to solve the following regularized rank minimization problem:

$$\min_{C,E} rank\left(C\right) + \lambda_2 \|E\|_l, s.t., X = XC + E \tag{13}$$

where E denotes the sample-specific corruptions, and the minimizer C^* is the so-called "LRR" of feature X with respect to the variable C. λ_2 is the low-rank coefficients which balances the effects between two parts. The rank function could be replaced by the nuclear norm for the convenience

Table 1: Some Information about 14 Datasets.

Dataset	Instances	Features	Labels
Artificial	2601	3	3
Movie	7755	1869	5
SBU_3DFE	2500	243	6
SJAFFE	213	243	6
Yeast-alpha	2465	24	18
Yeast-cdc	2465	24	15
Yeast-cold	2465	24	4
Yeast-diau	2465	24	7
Yeast-dtt	2465	24	4
Yeast-elu	2465	24	14
Yeast-heat	2465	24	6
Yeast-spo	2465	24	6
Yeast-spo5	2465	24	3
Yeast-spoem	2465	24	2

of computing in rank minimization problems, reformulating the above problem as follows:

$$\min_{C,E} \|C\|_* + \lambda_2 \|E\|_{2,1}, s.t., X = XC + E \tag{14}$$

There are diverse approaches to tackle the above convex optimization problem. We adopt the augmented Lagrange multiplier (ALM) approach (Liu et al. 2012) in this paper. Specifically, by introducing an auxiliary variable J, this problem can be equally transformed into the following formula:

$$\begin{aligned} &\min_{J,C,E} \|J\|_* + \lambda_2 \|E\|_{2,1} \\ &s.t.\ X = XC + E, C = J \end{aligned} \tag{15}$$

Since it is regarded as an ALM optimizing problem (Liu et al. 2012), this problem can be settled by minimizing the following augmented Lagrangian function:

$$\begin{aligned} &\mathcal{M} = \|J\|_* + \lambda_2 \|E\|_{2,1} \\ &+\mathrm{tr}\left(Y_1^T(X - XC - E)\right) + \mathrm{tr}\left(Y_2^T(C - J)\right) \\ &+\frac{\mu}{2}\left(\|X - XC - E\|_F^2 + \|C - J\|_F^2\right) \end{aligned} \tag{16}$$

To gain the optimized J, C and E, we update each of them while fixing the other two variables respectively and subsequently updating the corresponding Lagrange multipliers Y_1, Y_2 and μ, during iterations. The detailed solution process can be found in (Liu et al. 2012).

After $\hat{C}$ is optimized, we could represent every term in (10) and consequently obtain the objective function of $\hat{\theta}$:

$$\begin{aligned} P\left(\hat{\theta}\right) &= \sum_{i=1}^{n} \left\|\phi\left(\hat{\theta}, \xi_i\right) - L_i\right\|^2 + \lambda_1 \left\|\left(I - \hat{C}^T\right)\mathfrak{D}^T\right\|_F^2 \\ &= tr\left[\left(\phi\left(\hat{\theta}, \Xi\right) - \Gamma\right)^T \left(\phi\left(\hat{\theta}, \Xi\right) - \Gamma\right)\right] \\ &\quad + \lambda_1 tr\left(\mathfrak{D}\left(I - \hat{C}\right)\left(I - \hat{C}^T\right)\mathfrak{D}^T\right) \end{aligned} \tag{17}$$

where $\Xi = [\xi(x_1), \cdots, \xi(x_n)]$.

Table 2: Introduction to Evaluation Measures.

Measure	Formula
Cheb ↓	$Dis_1(D, \hat{D}) = \max_j \left\| d^{y_j} - \hat{d}^{y_j} \right\|$
Canber ↓	$Dis_2(D, \hat{D}) = \sum_{j=1}^{o} \frac{\left\|d^{y_j} - \hat{d}^{y_j}\right\|}{d^{y_j} + \hat{d}^{y_j}}$
Clark ↓	$Dis_3(D, \hat{D}) = \sqrt{\sum_{j=1}^{o} \frac{\left(d^{y_j} - \hat{d}^{y_j}\right)^2}{\left(d^{y_j} + \hat{d}^{y_j}\right)^2}}$
Cosine ↑	$Sim_1(D, \hat{D}) = \frac{\sum_{j=1}^{o} d^{y_j}\hat{d}^{y_j}}{\sqrt{\sum_{j=1}^{o}\left(d^{y_j}\right)^2}\sqrt{\sum_{j=1}^{o}\left(\hat{d}^{y_j}\right)^2}}$
Intersec ↑	$Sim_2(D, \hat{D}) = \sum_{j=1}^{o} \min\left(d^{y_j}, \hat{d}^{y_j}\right)$

To acquire the optimized $\hat{\theta}^*$, the minimization of this objective function will be solved by an effective quasi-Newton method called the limited memory BFGS (L-BFGS) (Nocedal and Wright 2006), of which the optimizing process is associated with the first-order gradient. Once the formula converges, we feed the optimal $\hat{\theta}^*$ into (9) to form the label distribution D_i. Furthermore, since the defined label distribution needs to meet the requirement $\sum_{m=1}^{o} d_{x_i}^{y_m}$=1, D_i is normalized by the softmax normalization form.

Experiment

Datasets

The fundamental statistics about 13 real-world datasets and a toy dataset employed to evaluate the algorithm are shown in Table 1. Whereas the first three real-world datasets are created from movies and facial expression images, the last datasets from Yeast-alpha to Yeast-spoem are collected from the records of 10 biological experiments on the budding yeast genes(Eisen et al. 1998). The artifical dataset was also adopted in (Xu, Tao, and Geng 2018), which intuitively exhibits the model's ability to recover the label distributions. Each instance $x_i \in \mathbb{R}^3$ is chosen following the rule that the first two dimensions $x_i^{(1)}$ and $x_i^{(2)}$ are formed as a grid with an interval of 0.04 in the range [-1,1], while the third dimension $x_i^{(3)}$ is computed by:

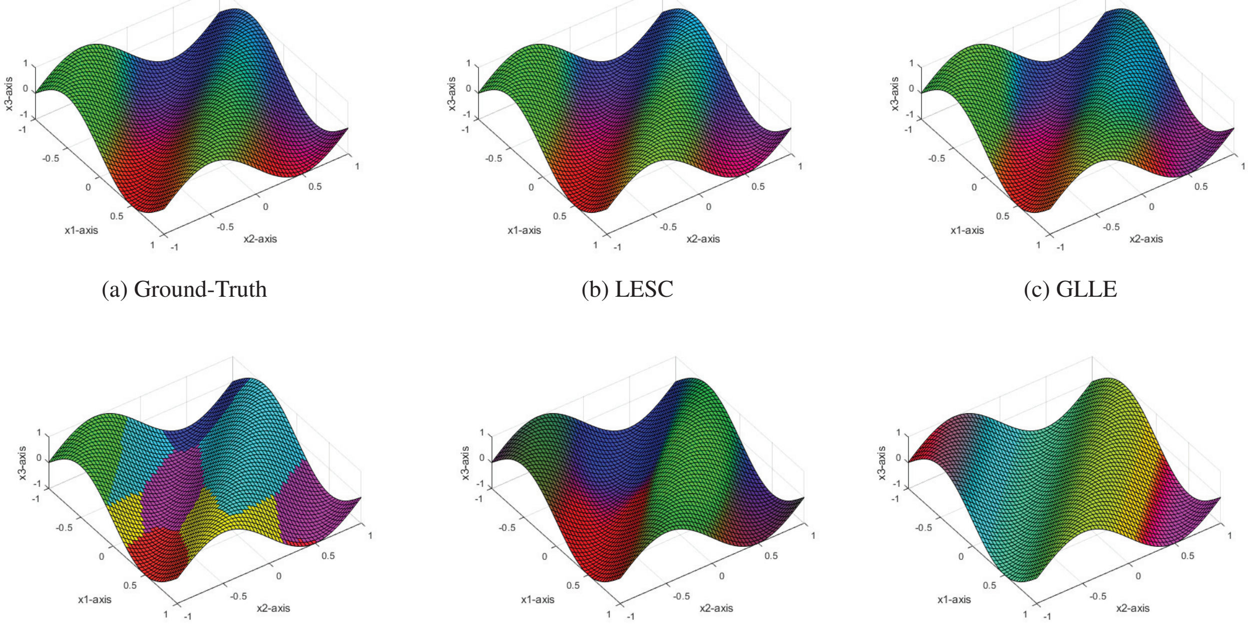
$$x_i^{(3)} = \sin\left(\left(x_i^{(1)} + x_i^{(2)}\right) \times \pi\right) \tag{18}$$

The corresponding label distribution $D_i = \left(d_{x_i}^{y_1}, d_{x_i}^{y_2}, d_{x_i}^{y_3}\right)^T$ is collected through the following equations:

$$w_j = mx_i^{(j)} + n\left(x_i^{(j)}\right)^2 + p\left(x_i^{(j)}\right)^3 + q, j = 1, 2, 3 \tag{19}$$

$$\begin{cases} \varphi_1 = \left(\mathbf{r}_1^\top \mathbf{w}\right)^2 \\ \varphi_2 = \left(\mathbf{r}_2^\top \mathbf{w} + \eta_1\varphi_1\right)^2 \\ \varphi_3 = \left(\mathbf{r}_3^\top \mathbf{w} + \eta_2\varphi_2\right)^2 \end{cases} \tag{20}$$

$$d_{x_i}^{y_j} = \frac{\varphi_j}{\varphi_1 + \varphi_2 + \varphi_3}, j = 1, 2, 3 \tag{21}$$



(a) Ground-Truth

(b) LESC

(c) GLLE

(d) LP

(e) ML

(f) FCM

Figure 2: Visualization of the ground-truth and recovered label distributions on the artificial dataset (regarded as RGB colors, best viewed in color).

where $\mathbf{w} = (w_1, w_2, w_3)$, $m = 1$, $n = 0.5$, $p = 0.2$, $q = 1$, $\mathbf{r}_1 = (4, 2, 1)^T$, $\mathbf{r}_2 = (1, 2, 4)^T$, $\mathbf{r}_3 = (1, 4, 2)^T$, and $\eta_1 = \eta_2 = 0.01$.

Experimental Settings

Evaluation Metric. Since both the recovered and ground-truth label distributions are label vectors, the average distance or similarity between them is calculated to evaluate the LE algorithms thoroughly. For a fair comparison, five measures were selected, where the first three were distance-based measures and the last two were similarity-based measures, reflecting an LE algorithm's performance from different aspects in semantics. As shown in Table 2 where $\hat{D}$ denotes the real label distribution, for these metrics, i.e., Chebyshev distance (Cheb), Canberra metric (Canber), Clark distance (Clark), cosine coefficient (Cosine) and intersection similarity (Intersec), $\downarrow$ states "the smaller the greater" while $\uparrow$ states "the larger the greater".

Implementation Details. To fully investigate the performance of the algorithms, the proposed LESC algorithm and four state-of-the-art algorithms, i.e., FCM (El Gayar, Schwenker, and Palm 2006), LP (Li, Zhang, and Geng 2015), ML (Hou, Geng, and Zhang 2016), and GLLE (Xu, Tao, and Geng 2018) were employed. We list the parameter settings here. The parameters λ_1 and λ_2 are selected among $\{0.0001, 0.001, ..., 10\}$ in our LESC algorithm. As for GLLE, the number of neighbors K is set to $c + 1$ and the parameters λ are set among $\{0.01, 0.1, ..., 100\}$. We also choose the parameter α in LP to be 0.5, the number of neighbors K for ML to be $c + 1$, and the parameter β in FCM to be 2.

We recovered the label distributions from logical labels and the above-mentioned metrics were computed first. Because of the lack of datasets with both logical labels and label distributions, the logical labels had to be binarized from the ground-truth label distributions in the LDL training set so that it was possible to implement LE algorithms and measure the similarity between the recovered label distributions and the ground-truths. To ensure the consistency of evaluation, we binarized the logical labels through the way in GLLE, which conforms to the annotating convention for MLL.

Moreover, it is known that LE aims at recovering the label distributions to strengthen supervised learning, which provides further motivation to evaluate the effectiveness of LE algorithms by conducting LDL. In particular, for a given dataset, the ten-fold cross validation was executed. After each algorithm recovered label distributions under its optimal parameters, the LDL model proposed in (Jia et al. 2019) was trained with the recovered distributions and the ground-truth label distributions. Both trained models were employed to predict the label distributions of the new test set simultaneously, and the same evaluation metrics were calculated between these generated label distributions. Note that we have implemented the LDL experiments with other LDL algorithms and similar results are obtained.

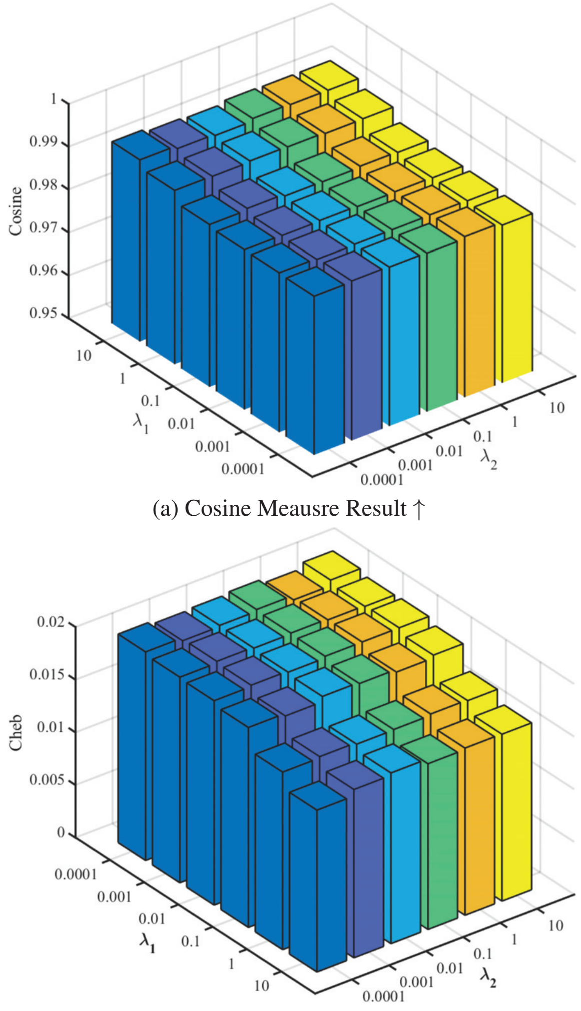
Table 3: Recovery Results (value(rank)) Measured by Cheb and Cosine.

Dataset	Measure Results by Cosine ↑					Measure Results by Cheb ↓				
	FCM	LP	ML	GLLE	LESC	FCM	LP	ML	GLLE	LESC
Artificial	0.933(4)	0.974(3)	0.925(5)	0.980(2)	**0.992(1)**	0.230(5)	0.130(3)	0.227(4)	0.108(2)	**0.057(1)**
Movie	0.773(5)	0.929(2)	0.919(3)	0.900(4)	**0.937(1)**	0.188(5)	0.161(3)	0.164(4)	0.160(2)	**0.121(1)**
SBU_3DFE	0.912(3)	0.922(2)	0.815(5)	0.900(4)	**0.932(1)**	0.135(3)	0.123(2)	0.233(5)	0.141(4)	**0.121(1)**
SJAFFE	0.906(4)	0.941(3)	0.857(5)	0.946(2)	**0.973(1)**	0.132(4)	0.107(3)	0.190(5)	0.100(2)	**0.069(1)**
Yeast-alpha	0.922(3)	0.911(4)	0.756(5)	0.973(2)	**0.992(1)**	0.044(4)	0.040(3)	0.057(5)	0.033(2)	**0.015(1)**
Yeast-cdc	0.929(3)	0.916(4)	0.759(5)	0.959(2)	**0.991(1)**	0.051(4)	0.042(3)	0.071(5)	0.038(2)	**0.019(1)**
Yeast-cold	0.922(4)	0.925(3)	0.784(5)	0.969(2)	**0.986(1)**	0.141(4)	0.137(3)	0.242(5)	0.093(2)	**0.056(1)**
Yeast-diau	0.882(4)	0.915(3)	0.803(5)	0.939(2)	**0.985(1)**	0.124(4)	0.099(3)	0.148(5)	0.084(2)	**0.042(1)**
Yeast-dtt	0.959(3)	0.921(4)	0.763(5)	0.983(2)	**0.991(1)**	0.097(3)	0.128(4)	0.244(5)	0.065(2)	**0.043(1)**
Yeast-elu	0.950(3)	0.918(4)	0.763(5)	0.978(2)	**0.991(1)**	0.052(4)	0.044(3)	0.072(5)	0.030(2)	**0.019(1)**
Yeast-heat	0.883(4)	0.932(3)	0.783(5)	0.980(2)	**0.986(1)**	0.169(5)	0.086(3)	0.165(4)	0.056(2)	**0.046(1)**
Yeast-spo	0.909(4)	0.939(3)	0.803(5)	0.968(2)	**0.975(1)**	0.130(4)	0.090(3)	0.171(5)	0.067(2)	**0.060(1)**
Yeast-spo5	0.922(4)	0.969(3)	0.884(5)	**0.974(1)**	**0.974(1)**	0.162(4)	0.114(3)	0.273(5)	**0.092(1)**	**0.092(1)**
Yeast-spoem	0.878(4)	0.950(3)	0.815(5)	0.968(2)	**0.978(1)**	0.233(4)	0.163(3)	0.400(5)	0.108(2)	**0.087(1)**
Avg.Rank	3.71	3.14	4.86	2.21	**1.00**	4.07	3.00	4.79	2.14	**1.00**

Algorithm Analysis

Recovery Performance. First, to illustrate the recovery performance on the artificial dataset visually, the three-dimensional label distributions were separately converted into the RGB color channels, which were reinforced by the decorrelation stretch process for easier observation. In other words, the label distribution of each point in the feature space could be represented by its color. Thus far, the color patterns can be directly observed to compare both the ground truth and the recovered label distributions. As shown in Fig.2, in contrast to the ground-truth color patterns, our algorithm nearly recovers these patterns identically, while GLLE obtains almost the same results. In addition, the color patterns in other three algorithms are barely satisfactory, which proves the limits of excavating the space structure of features locally.

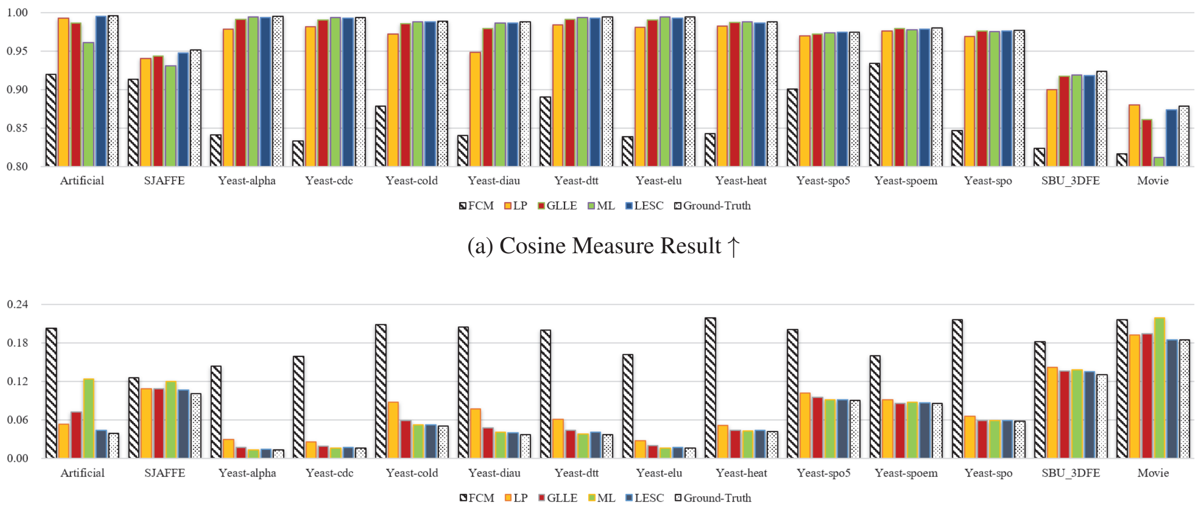
Because of space limitations, we only present the quantitative results of the above five LE algorithms by Cheb and Cosine in Table 3 where the optimal results for each dataset are highlighted with boldface. Note that we could directly borrow the recovery results in (Xu, Tao, and Geng 2018) under the same settings. To exhibit the mean accuracy of the recovered label distribution, the average rank of every algorithm among all datasets is also listed. Some observations stand out in this table. Apparently, the quantitative performances of the artificial dataset are consistent with the recovered color patterns in Fig.2 where the proposed algorithm gets the $1st$ rank. Besides, the changes of rankings under distinct metrics have evidenced the significance of employing a collection of measures. Despite gaps between these metrics, the performance of the proposed algorithm still outperforms those of others for the 13 datasets by a large margin and achieves identical results with the state-of-the-art baselines in the Yeast-spo5 dataset under both shown measures, which substantially surpasses the performance of other algorithms varying in different metrics (e.g., ML ranks the $3rd$ order by Cosine measure while the $4th$ ranking was obtained by the Cheb measure in Movie dataset). These results have strongly demonstrated the proposed algorithm's effectiveness and generalization, and the necessity of employing



(a) Cosine Meausre Result ↑

(b) Cheb Meausre Result ↓

Figure 3: Measure Results of Yeast_alpha Dataset by Cosine ↑ and Cheb ↓.



(a) Cosine Measure Result ↑

(b) Cheb Measure Result ↓

Figure 4: LDL Prediction Results of 14 Datasets by Cosine ↑ and Cheb ↓

Table 4: The Average Ranks on Five Measures.

Measure	FCM	LP	ML	GLLE	LESC
Cheb	4.93	3.57	2.29	2.64	**1.57**
Clark	4.93	3.71	2.29	2.50	**1.57**
Canber	4.93	3.71	2.36	2.50	**1.50**
Cosine	4.93	3.57	2.21	2.64	**1.64**
Intersec	5.00	3.57	2.29	2.57	**1.57**

the LRR.

Parameter Analysis. The effects of two trade-off hyperparameters were analyzed separately on the real-world dataset through fixing one parameter among the aforementioned parameter scope and tuning the other one as well. We only illustrate the case of the dataset Yeast_alpha by Cheb and Cosine measures in Fig.3, while the same rule is obtained in other datasets. When the low-rank coefficient λ_2 varies with the trade-off parameter λ_1 fixed, two shown measure results of the recovery performance fluctuates in a very tiny range that could not even be distinguished. As we increase the parameter λ_1 from 0.0001 to 0.1, the recovery performance also turns out to change within a small scope. When λ_1 is geared to 1 or 10, the results even zooms up to a higher level. Particularly, taking this dataset for reference, it is found that our worst measure result still far exceeds that of the previous state-of-the-art baseline, i.e, 0.987 versus 0.973 (GLLE) by cosine. As discussed before, these phenomena indicate that our algorithm is robust when the low-rank coefficients and the trade-off parameter in the objective function vary by a large margin. This ensures us to generalize our algorithm to different datasets without much effort in terms of adjusting the values of hyperparameters.

LDL Prediction Performance. We train the LDL model and then predict the label distributions of the new test set as discussed before, and the average orders of these results on five measures are listed in Table.4. While LP ranks $3st$ in recovery experiments but ranks $4st$ in LDL prediction, our algorithm still has the highest average ranks across all measures despite huge gaps between these two experiments. The LDL prediction results measured by Cheb and Cosine are illustrated in Fig.4. Our algorithm obtains almost the same prediction results as that of ground-truth on other datasets, which proves the accuracy of the algorithm's recovered label distributions. Moreover, the LESC model ranks $1st$ in nine datasets, while ranking $2nd$ in other datasets, which is superior over other algorithms. Note that we haven't attained the best results in some cases such as Yeast-alpha and Yeast-heat. We believe that this is mainly because of the inaccurate binary of logical labels. However, our prediction results are still comparable to the best ones and to the ground-truth distributions in these datasets, even until the last four decimal places of the Cheb measure.

Conclusion

To excavate the underlying information contained in the feature space through a global approach, a novel algorithm, LESC, was proposed to address the LE issue. The LRR was generated to capture the global intrinsic relationships of instances, and it was subsequently employed to excavate the hidden label information globally. Extensive experimental results among 14 datasets demonstrated the remarkable superiority of the proposed algorithm over several state-of-the-art algorithms in recovering the label distributions and LDL prediction after LE preprocessing on logical labels. Further analysis of the influence of hyperparameters verified the robustness of our algorithm to the variation of parameters.

Acknowledgments

We acknowledge the National Key R&D Program of China (No. 2018AAA0102504) and the National Natural Science

Foundation of China under Grant (nos. 61573273) for funding this research.

References

Chen, Z.-M.; Wei, X.-S.; Wang, P.; and Guo, Y. 2019. Multi-label image recognition with graph convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 5177–5186.

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

El Gayar, N.; Schwenker, F.; and Palm, G. 2006. A study of the robustness of knn classifiers trained using soft labels. In *IAPR Workshop on Artificial Neural Networks in Pattern Recognition*, 67–80. Springer.

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

Geng, X.; Yin, C.; and Zhou, Z.-H. 2013. Facial age estimation by learning from label distributions. *IEEE transactions on pattern analysis and machine intelligence* 35(10):2401–2412.

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

Hou, P.; Geng, X.; and Zhang, M.-L. 2016. Multi-label manifold learning. In *Thirtieth AAAI Conference on Artificial Intelligence*.

Huang, S.-J., and Zhou, Z.-H. 2012. Multi-label learning by exploiting label correlations locally. In *Twenty-sixth AAAI conference on artificial intelligence*.

Jia, X.; Zheng, X.; Li, W.; Zhang, C.; and Li, Z. 2019. Facial emotion distribution learning by exploiting low-rank label correlations locally. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 9841–9850.

Li, Y.-K.; Zhang, M.-L.; and Geng, X. 2015. Leveraging implicit relative labeling-importance information for effective multi-label learning. In *2015 IEEE International Conference on Data Mining*, 251–260. IEEE.

Liu, G., and Yan, S. 2011. Latent low-rank representation for subspace segmentation and feature extraction. In *2011 International Conference on Computer Vision*, 1615–1622. IEEE.

Liu, G.; Lin, Z.; Yan, S.; Sun, J.; Yu, Y.; and Ma, Y. 2012. Robust recovery of subspace structures by low-rank representation. *IEEE transactions on pattern analysis and machine intelligence* 35(1):171–184.

Melin, P., and Castillo, O. 2005. *Hybrid intelligent systems for pattern recognition using soft computing: an evolutionary approach for neural networks and fuzzy systems*, volume 172. Springer Science & Business Media.

Nocedal, J., and Wright, S. 2006. *Numerical optimization*. Springer Science & Business Media.

Roweis, S. T., and Saul, L. K. 2000. Nonlinear dimensionality reduction by locally linear embedding. *science* 290(5500):2323–2326.

Tsoumakas, G., and Katakis, I. 2007. Multi-label classification: An overview. *International Journal of Data Warehousing and Mining (IJDWM)* 3(3):1–13.

Xu, N.; Lv, J.; and Geng, X. 2019. Partial label learning via label enhancement. In *AAAI Conference on Artificial Intelligence*.

Xu, N.; Tao, A.; and Geng, X. 2018. Label enhancement for label distribution learning. In *IJCAI*, 2926–2932.

Yin, M.; Gao, J.; and Lin, Z. 2015. Laplacian regularized low-rank representation and its applications. *IEEE transactions on pattern analysis and machine intelligence* 38(3):504–517.

Zhai, L.; Zhu, J.; Zheng, Q.; Pang, S.; Li, Z.; and Wang, J. 2019. Multi-view spectral clustering via partial sum minimisation of singular values. *Electronics Letters* 55(6):314–316.

Zhang, M.-L., and Zhou, Z.-H. 2013. A review on multi-label learning algorithms. *IEEE transactions on knowledge and data engineering* 26(8):1819–1837.

Zheng, X.; Jia, X.; and Li, W. 2018. Label distribution learning by exploiting sample correlations locally. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Zhu, X., and Goldberg, A. B. 2009. Introduction to semi-supervised learning. *Synthesis lectures on artificial intelligence and machine learning* 3(1):1–130.