模式分析与机器智能
工业和信息化部重点实验室
MIIT Key Laboratory of
Pattern Analysis & Machine Intelligence

ParNeC 模式识别与神经计算研究组
PAttern Recognition and NEural Computing

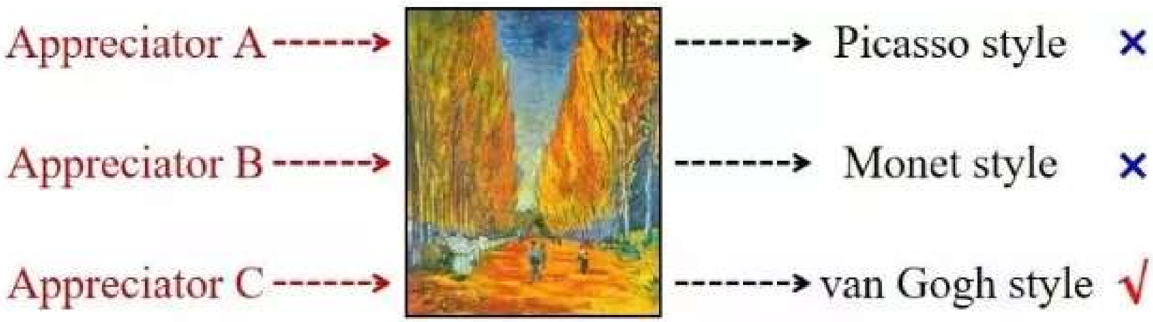
Partial Label Learning with Dissimilarity Propagation guided Candidate Label Shrinkage

Yuheng Jia[1,3]*, **Fuchao Yang**[2], **Yongqiang Dong**[1]*
[1] School of Computer Science and Engineering, Southeast University, Nanjing 210096, China
[2] College of Software Engineering, Southeast University Nanjing 210096, China
[3] Key Laboratory of New Generation Artificial Intelligence Technology and Its Interdisciplinary Applications (Southeast University), Ministry of Education, China
yhjia@seu.edu.cn, yangfc@seu.edu.cn, dongyq@seu.edu.cn

NeurIPS 2023

Background

Partial Label Learning (PLL)



Partial label learning (PLL), also known as superset-label learning or ambiguous label learning, is a representative weakly supervised learning framework which learns from inaccurate supervision information.
In partial label learning, each instance is associated with a set of candidate labels with only one being ground-truth and others being false positive.
As the ground-truth label of a sample conceals in the corresponding candidate label set, which can not be directly acquired during the training process, partial label learning task is a quite challenging problem.

Related work

To tackle the mentioned challenge, existing works mainly focus on disambiguation

Averaging-based approaches

Such as PL-KNN(2005) averages the candidate labels of neighboring samples to make the prediction.

Identification-based approaches

The ground-truth label is treated as a latent variable and can be identified through an iterative optimization procedure such as EM. Labeling confidence based strategy is proposed in many state-of-the-art identification based approaches for better disambiguation.

Deep-learning based models

PICO(2022) is a contrastive learning-based approach devised to tackle label ambiguity in partial label learning.
PRODEN(2020) is a model where the simultaneous updating of the model and identification of true labels are seamlessly integrated.

Motivation

SDIM first built a pairwise dissimilarity matrix through the candidate label sets, and then maximized the difference of the label confidence between two samples if their pairwise dissimilarity between them is large according to the constructed dissimilarity matrix. However, the dissimilarity matrix constructed by SDIM is predefined and relatively sparse, which depresses its effectiveness.

$$r_{ij} = \begin{cases} 1, & \text{if } \mathbf{y}_i^\top \mathbf{y}_j = 0 \\ 0, & \text{if } \mathbf{y}_i^\top \mathbf{y}_j \neq 0 \end{cases}$$

Basic Model

Let $X = [x_1, x_2, \dots, x_m]^T \in \mathbb{R}^{m\times d}$ denote the feature matrix
$Y = [y_1, y_2, \dots y_m]^T \in \{0,1\}^{m\times q}$ represents the partial label matrix

To fulfill PLL, we first build the following constrained regression model

$$\min_{\mathbf{W},\mathbf{F}} \|\mathbf{XW}-\mathbf{F}\|_F^2 + \lambda \|\mathbf{W}\|_F^2$$
$$\text{s.t. } \mathbf{F1}_q = \mathbf{1}_m, \mathbf{0}_{m\times q} \leq \mathbf{F} \leq \mathbf{Y},$$

We initialize the label confidence matrix as $F_{ij} = \frac{1}{\sum_j y_{ij}}\ if\ y_{ij} = 1$, otherwise $F_{ij} = 0$.

We assume that the mapping from the features to the ground-truth label may be easier, while that to the false-positive label residing in the candidate label set is relatively harder. Accordingly, optimizing Eq. (1) will help disambiguate the candidate labels and produce a preliminary label confidence matrix by exploring the useful information in the feature space.

Dissimilarity Propagation guided Candidate Label Shrinkage

To further exploit the valuable information in the label space, we first use candidate labels to construct a dissimilarity matrix $D_0 \in R_{m\times m}$, i.e.,

$$\mathbf{D}_{0ij} = \begin{cases} 1, & \text{if } y_i y_j^{\top} = 0 \\ 0, & \text{otherwise.} \end{cases}$$

Therefore, D_0 indicates the semantic dissimilarity of samples.

We then multiply the label confidence matrix with its transpose to create a similarity matrix termed as FF^T, whose (i, j)-th element indicates the similarity between x_i and x_j.

As the semantic dissimilarity matrix D_0 and similarity matrix FF^T form an adversarial relationship, we use this adversarial prior to shrink the solution space of F by:

$$\min_{\mathbf{F}} \left\| \mathbf{D}_0 \odot \mathbf{F}\mathbf{F}^{\top} \right\|_1 ,$$

ParN$_e$C 模式识别与神经计算研究组 PAttern Recognition and NEural Computing

Dissimilarity Propagation guided Candidate Label Shrinkage

$$\min_{\mathbf{F}} \left\| \mathbf{D}_0 \odot \mathbf{F}\mathbf{F}^{\mathsf{T}} \right\|_1 ,$$

Unfortunately, directly minimizing Eq. (3) cannot help produce a better label confidence matrix F, because D_0 is inferred from the candidate label set, and the (i, j)-th element of D_0 is positive only when $y_i y_j^T = 0$, while F is upper bounded by Y, the (i, j)-th element of FF^T is positive only when $y_i y_j^T \neq 0$. That is the locations of the positive elements of D_0 and FF^T are complementary.

Dissimilarity Propagation guided Candidate Label Shrinkage

Specifically, we leverage the local geometric structure of samples to enhance D_0. Note that each column of D (e.g., $D_{.i}$) can represent the dissimilarity relationships between a sample (e.g., x_i) and the other samples. If two samples x_i and x_j are close to each other in the feature space, their dissimilarity relationships ($D_{.i}$ and $D_{.j}$) should also be similar. To capture the feature similarity, we build a local geometric matrix $S \in \mathbb{R}_{m\times m}$ using a radial basis function (RBF) kernel:

$$\mathbf{S}_{ij} = \begin{cases} \exp(-||x_i - x_j||_2^2/\sigma^2), & \text{if } j \in \mathcal{N}_i \\ 0, & \text{otherwise}, \end{cases}$$

Based on S, the dissimilarity propagation guided candidate label shrinkage module becomes

$$\min_{\mathbf{F}} \left\| \mathbf{D}_0 \odot \mathbf{F}\mathbf{F}^\top \right\|_1, \quad \Longrightarrow \quad \min_{\mathbf{F},\mathbf{D}} \left\| \mathbf{D} \odot \mathbf{F}\mathbf{F}^\top \right\|_1 + \sum_{i,j=1}^{m} \mathbf{S}_{ij} \left\| \mathbf{D}_{.i} - \mathbf{D}_{.j} \right\|_2^2$$

$$\text{s.t. } \mathbf{0}_{m\times m} \leq \mathbf{D} \leq \mathbf{1}_{m\times m}, \mathbf{D}_{ij} = \mathbf{D}_{0ij}, \text{if } \mathbf{D}_{0ij} = 1.$$

Optimization and Setting

Taking all the above considerations into account, the proposed model finally becomes:

$$\min_{\mathbf{W},\mathbf{F},\mathbf{D}} \|\mathbf{XW}-\mathbf{F}\|_F^2 + \lambda\|\mathbf{W}\|_F^2 + \alpha\left\|\mathbf{D}\odot\mathbf{FF}^{\mathsf{T}}\right\|_1 + \beta\mathrm{Tr}(\mathbf{DLD}^{\mathsf{T}})$$
$$\text{s.t. } \mathbf{F1}_q = \mathbf{1}_m, \mathbf{0}_{m\times q} \le \mathbf{F} \le \mathbf{Y}, \mathbf{0}_{m\times m} \le \mathbf{D} \le \mathbf{1}_{m\times m}, \mathbf{D}_{ij} = \mathbf{D}_{0ij}, \text{if } \mathbf{D}_{0ij} = 1,$$

where $L \in \mathbb{R}^{m\times m} = D_S - S$ is a graph Laplacian matrix, and $D_S \in \mathbb{R}^{m\times m}$ is a diagonal matrix with the i-th diagonal element $D_{Sii} = \sum_{j=1}^{m} S_{ij}$. $Tr(\cdot)$ returns the trace of a matrix. $\alpha, \beta \ge 0$ are two hyper-parameters to balance different terms.

We adopt IALM to solve the problem in Eq. (6). To simplify Eq. (6), we introduce an auxiliary matrix $A = D \in \mathbb{R}^{m\times m}$ and the solution can be obtained by solving the following augmented Lagrange equation:

$$\min_{\mathbf{W},\mathbf{F},\mathbf{D},\mathbf{A}} \|\mathbf{XW}-\mathbf{F}\|_F^2 + \lambda\|\mathbf{W}\|_F^2 + \alpha\left\|\mathbf{A}\odot\mathbf{FF}^{\mathsf{T}}\right\|_1 + \beta\mathrm{Tr}(\mathbf{DLD}^{\mathsf{T}}) + \langle\mathbf{\Phi}, \mathbf{D}-\mathbf{A}\rangle + \frac{\mu}{2}\|\mathbf{D}-\mathbf{A}\|_F^2$$
$$\text{s.t. } \mathbf{F1}_q = \mathbf{1}_m, \mathbf{0}_{m\times q} \le \mathbf{F} \le \mathbf{Y}, \mathbf{0}_{m\times m} \le \mathbf{A} \le \mathbf{1}_{m\times m}, \mathbf{A}_{ij} = \mathbf{D}_{0ij}, \text{if } \mathbf{D}_{0ij} = 1,$$

W subproblem

$$\min_{\mathbf{W}} \|\mathbf{X}\mathbf{W} - \mathbf{F}\|_F^2 + \lambda \|\mathbf{W}\|_F^2 \,,$$

We extend the above model to a kernel-based non-linear version. Let $\phi(\cdot)$: $\mathbb{R}^d \rightarrow \mathbb{R}^h$ denote the feature transformation that maps the origin feature space X to a higher dimensional Hilbert space $\phi(X)$.

According to the Representer Theorem, W can be expressed as a linear combination of the input features, i.e. $W = \phi(X)^T H$, where $H \in \mathbb{R}^{m\times q}$ stores the combination weights.

Then, we have $\phi(X)W = KH$, where $K = \phi(X)\phi(X)^T \in \mathbb{R}^{m\times m}$ is the kernel matrix and each element $K_{ij} = \mathcal{K}(x_i, x_j)$. Finally, the nonlinear version is represented as:

$$\min_{\mathbf{H},\mathbf{b}} \left\|\mathbf{K}\mathbf{H} + \mathbf{1}_m \mathbf{b}^\mathsf{T} - \mathbf{F}\right\|_F^2 + \lambda \mathrm{Tr}(\mathbf{H}^\mathsf{T}\mathbf{K}\mathbf{H}),$$

In the experiments, we use the RBF kernel as the kernel function, i.e., $\mathcal{K}(x_i, x_j) = \exp(-||x_i - x_j||_2^2/\sigma^2)$, for our method and the compared ones. When the first derivatives of H and b reach 0,

$$\mathbf{H} = \left(\mathbf{K} + \lambda \mathbf{I}_{m\times m} - \frac{\mathbf{1}_m \mathbf{1}_m^\mathsf{T} \mathbf{K}}{m}\right)^{-1} \left(\mathbf{F} - \frac{\mathbf{1}_m \mathbf{1}_m^\mathsf{T} \mathbf{F}}{m}\right), \mathbf{b} = \frac{1}{m}\left(\mathbf{F}^\mathsf{T}\mathbf{1}_m - \mathbf{H}^\mathsf{T}\mathbf{K}^\mathsf{T}\mathbf{1}_m\right),$$

F subproblem

$$\min_{\mathbf{F}} \|\mathbf{F}-\mathbf{P}\|_F^2 + \alpha \left\|\mathbf{A} \odot \mathbf{F}\mathbf{F}^{\mathrm{T}}\right\|_1$$
$$\text{s.t. } \mathbf{F}\mathbf{1}_q = \mathbf{1}_m, \mathbf{0}_{m\times q} \le \mathbf{F} \le \mathbf{Y},$$

where $P = KH + 1_m b^T \in \mathbb{R}^{m\times q}$ is the output matrix of the model. Eq. (12) can be formulated as a standard quadratic programming (QP) problem, and solved by any QP tools.

D subproblem

$$\min_{\mathbf{D}} \beta \mathrm{Tr}(\mathbf{D}\mathbf{L}\mathbf{D}^{\mathrm{T}}) + \frac{\mu}{2}\left\|\mathbf{D}-\mathbf{A}+\frac{\mathbf{\Phi}}{\mu}\right\|_F^2 .$$

Eq. reaches the minimum when its first-order derivative with respect to D vanishes, leading to

$$\mathbf{D} = (\mu\mathbf{A} - \mathbf{\Phi})(2\beta\mathbf{L} + \mu\mathbf{I}_{m\times m})^{-1}.$$

A subproblem

$$\min_{\mathbf{A}} \alpha \left\| \mathbf{A} \odot \mathbf{F}\mathbf{F}^{\mathsf{T}} \right\|_1 + \frac{\mu}{2} \left\| \mathbf{D} - \mathbf{A} + \frac{\mathbf{\Phi}}{\mu} \right\|_F^2$$
$$\text{s.t. } \mathbf{0}_{m\times m} \leq \mathbf{A} \leq \mathbf{1}_{m\times m}, \mathbf{A}_{ij} = \mathbf{D}_{0ij}, \text{if } \mathbf{D}_{0ij} = 1.$$

Eq. can solved element-wisely, i.e.,

$$\mathbf{A} = \mathcal{T}\left(\mathcal{T}_1\left(\mathcal{T}_0\left(\frac{\mu\mathbf{D} + \mathbf{\Phi} - \alpha\mathbf{F}\mathbf{F}^{\mathsf{T}}}{\mu}\right)\right)\right),$$

where $\mathcal{T}, \mathcal{T}_0, \mathcal{T}_1$ are three thresholding operators in elementwise, i.e., $\mathcal{T}(C_{ij}) = 1, if\ D_{0ij} = 1$, $\mathcal{T}_1(C_{ij}) \coloneqq \min(1, C_{ij})$, $\mathcal{T}_0(C_{ij}) \coloneqq \max(0, C_{ij})$.

Finally, the Lagrangian multiplier matrix and μ are updated by

$$\begin{cases} \mathbf{\Phi} \leftarrow \mathbf{\Phi} + \mu(\mathbf{D} - \mathbf{A}) \\ \mu \leftarrow \min(1.1\mu, \mu_{\max}), \end{cases}$$

where $\mu_{max} = 10^{10}$ is a predefined upper bound for μ.

Method

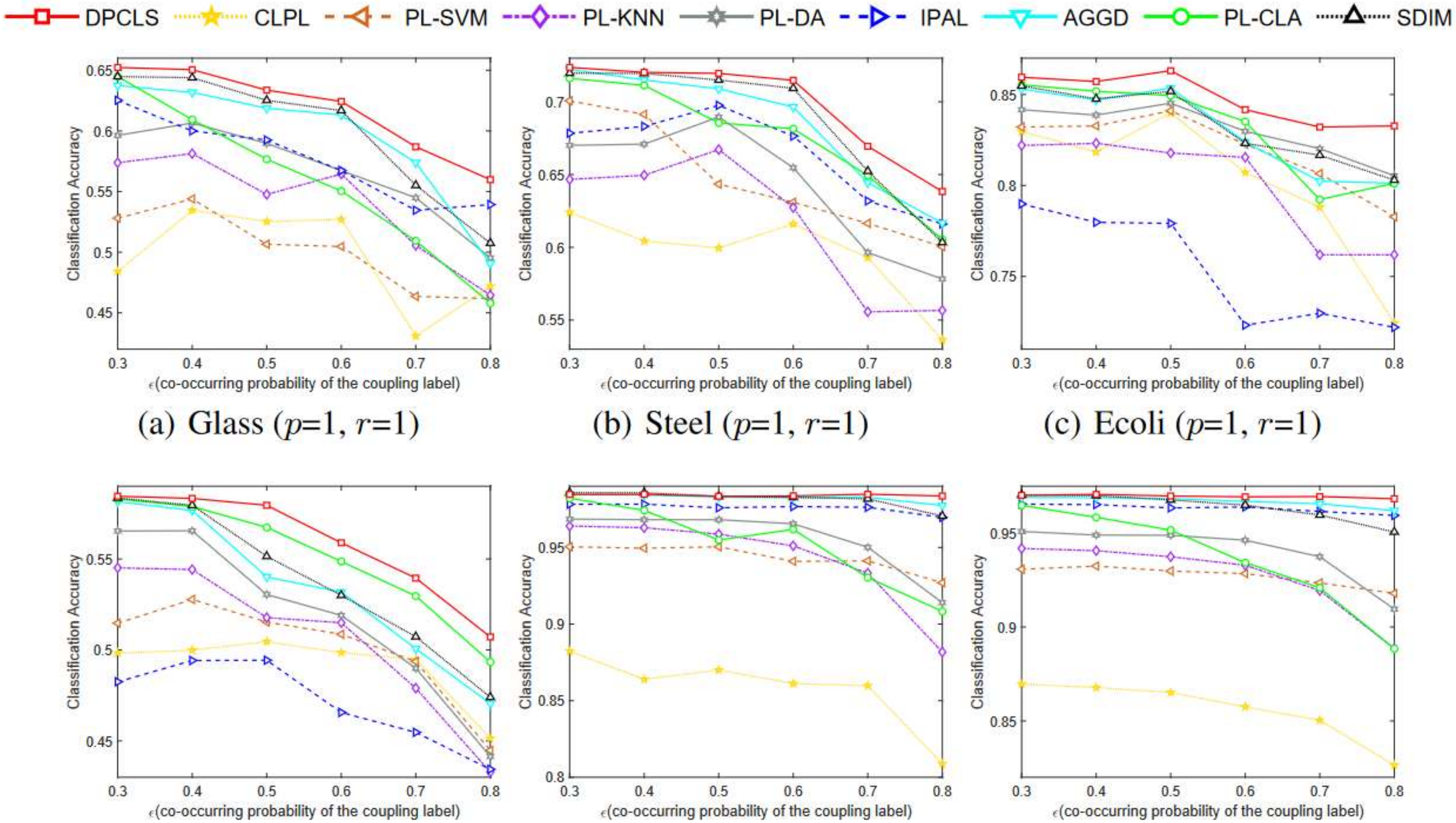
Algorithm 1 The Pseudo Code of the Proposed Method

Input: $\mathcal{D}$: the partial label training set; $\lambda, \alpha, \beta, k$: the parameters of model; $\widehat{x}$: an unseen test sample
Output: $\widehat{y}$: the predicted label for sample $\widehat{x}$

```
Construct the dissimilarity matrix D_0 according to Eq. (2) and the kernel matrix K = [𝒦(x_i, x_j)]_{m×m}
Initialize D = A = Φ = 0_{m×m}, μ = 10^{-4}
while not converged do
    Update H and b by Eq. (11)
    Update F by solving Eq. (12)
    Update D by Eq. (14)
    Update A by Eq. (16)
    Update Φ, μ by Eq. (17)
    Check the convergence condition ||D − A||_∞ < 10^{-8}
end while
Return the predicted label ŷ according to Eq. (18).
```

Experiment

ParNeC 模式识别与神经计算研究组
PAttern Recognition and NEural Computing



(a) Glass (p=1, r=1) (b) Steel (p=1, r=1) (c) Ecoli (p=1, r=1)

(d) Yeast (p=1, r=1) (e) Optdigits (p=1, r=1) (f) Usps (p=1, r=1)

Figure 1: The classification accuracy of each algorithm as ϵ increases from 0.1 to 0.7 with p=1, r=1.

Experiment

Table 1: Classification accuracy (mean±std) of each comparing algorithm on four synthetic data sets as r increases from 3 to 6. •/○ indicates whether the accuracy of DPCLS is statistically superior/inferior to the compared algorithm according to the pairwise t-test at 0.05 significance level.

Data set	r	DPCLS	CLPL	PL-SVM	PL-KNN	PL-DA	IPAL	AGGD	PL-CLA	SDIM
Isolet	3	.912±.008	.604±.023•	.668±.034•	.695±.025•	.782±.020•	.793±.018•	.898±.011•	.880±.007•	**.917±.007**
	4	**.899±.011**	.594±.023•	.577±.057•	.661±.026•	.768±.022•	.772±.017•	.880±.013•	.855±.011•	.891±.015
	5	**.880±.012**	.535±.015•	.543±.070•	.668±.018•	.751±.016•	.790±.010•	.855±.017•	.845±.018•	.849±.017•
	6	**.858±.017**	.525±.020•	.452±.062•	.645±.020•	.733±.020•	.776±.011•	.824±.015•	.824±.015•	.824±.018•
Orl	3	**.807±.040**	.780±.037	.262±.034•	.247±.033•	.347±.061•	.715±.064•	.764±.054	.735±.055•	.744±.044•
	4	**.776±.049**	.743±.031	.196±.052•	.216±.030•	.287±.042•	.665±.065•	.730±.065	.673±.066•	.692±.067•
	5	**.736±.041**	.704±.032	.158±.029•	.189±.027•	.247±.040•	.594±.045•	.677±.067•	.607±.049•	.622±.056•
	6	**.669±.042**	.660±.045	.114±.024•	.159±.025•	.218±.029•	.536±.047•	.646±.073	.545±.050•	.556±.053•
Amazon	3	**.062±.007**	.058±.007	.041±.006•	.021±.004•	.022±.004•	.060±.009	.056±.010	.055±.008•	.047±.006•
	4	**.057±.007**	.055±.007	.043±.005•	.020±.005•	.021±.005•	.055±.009	.054±.009	.053±.007	.045±.005•
	5	**.055±.009**	.050±.007	.043±.005•	.020±.005•	.022±.005•	.050±.009	.047±.008•	.044±.009•	.037±.007•
	6	**.048±.009**	.042±.007	.029±.007•	.020±.003•	.020±.003•	.046±.009	.043±.008	.041±.006•	.032±.006•
Bookmark	3	**.337±.007**	.248±.007•	.208±.013•	.115±.017•	.178±.014•	.286±.012•	.333±.009	.291±.009•	.305±.010•
	4	**.326±.008**	.244±.009•	.190±.020•	.121±.028•	.170±.013•	.282±.010•	.325±.007	.284±.011•	.300±.011•
	5	**.323±.009**	.240±.010•	.169±.019•	.116±.025•	.189±.051•	.278±.009•	.319±.005	.278±.010•	.298±.011•
	6	**.322±.009**	.240±.010•	.148±.013•	.103±.029•	.186±.043•	.270±.009•	.316±.007	.276±.009•	.295±.007•

Experiment

Table 2: Classification accuracy (mean±std) of each algorithm on real-world partial label data sets. •/○ indicates whether the accuracy of DPCLS is statistically superior/inferior to the compared algorithm according to the pairwise t-test at 0.05 significance level. "S" and "D" indicate shallow and deep PLL methods respectively.

Type	Method	FG-NET	FG-NET3	FG-NET5	Lost	MSRCv2	BirdSong	Malagasy	Soccer Player	Yahoo! News
S	DPCLS	**.077±.009**	**.436±.017**	**.586±.011**	**.770±.024**	**.557±.014**	**.751±.009**	**.676±.004**	**.532±.002**	.626±.003
	CLPL	.058±.009•	.383±.016•	.538±.017•	.665±.019•	.371±.010•	.610±.012•	.675±.016	.497±.002•	.544±.004•
	PL-SVM	.052±.010•	.357±.022•	.511±.026•	.578±.078•	.310±.060•	.682±.023•	.564±.061•	.500±.002•	.546±.006•
	PL-KNN	.038±.005•	.287±.022•	.433±.019•	.334±.021•	.391±.023•	.657±.014•	.573±.007•	.493±.002•	.383±.003•
	PL-DA	.042±.004•	.166±.050•	.255±.070•	.309±.069•	.416±.022•	.690±.013•	.606±.008•	.495±.003•	.397±.004•
	IPAL	.052±.006•	.347±.015•	.510±.016•	.610±.020•	.494±.024•	.722±.006•	.621±.017•	.530±.005	.618±.007•
	AGGD	.075±.010	.423±.016	.568±.018•	.702±.024•	.477±.019•	.722±.014•	.593±.050•	.527±.003•	.616±.004•
	PL-CLA	.074±.011	.424±.020	.571±.015•	.696±.021•	.470±.016•	.722±.012•	.654±.005•	.525±.003•	.606±.004•
	SDIM	.073±.009	.423±.022	.568±.019•	.736±.023•	.475±.016•	.724±.012•	.643±.007•	.524±.003•	.607±.004•
D	RC	.072±.009	.391±.012•	.488±.020•	.740±.026•	.446±.019•	.715±.007•	.664±.004•	**.532±.004**	.620±.003•
	PRODEN	.071±.009	.415±.016•	.567±.025•	.712±.032•	.430±.019•	.704±.013•	.665±.017•	.528±.004•	.620±.003•
	CAVL	.071±.006	.365±.020•	.488±.021•	.747±.060•	.444±.013•	.695±.017•	.668±.039	.510±.004•	**.628±.004**

Table 4: Ablation study of our method on the real-world partial label data sets.

	FG-NET	Lost	MSRCv2	BirdSong	Malagasy	Soccer Player	Yahoo! News
DPCLS	**.077±.009**	**.770±.024**	**.557±.014**	**.751±.009**	**.676±.004**	**.532±.002**	**.626±.003**
DPCLS-LM	.067±.009•	.652±.023•	.357±.009•	.577±.012•	.587±.014•	.492±.002•	.447±.004•
DPCLS-KE	.068±.009•	.701±.023•	.388±.014•	.595±.014•	.674±.009	.495±.002•	.463±.004•
DPCLS-DP	.073±.010	.687±.027•	.466±.018•	.721±.014•	.612±.011•	.524±.003•	.604±.004•

Thanks