Disambiguation Enabled Linear Discriminant Analysis for Partial Label Dimensionality Reduction

Jing-Han Wu

[†]School of Computer Science and Engineering, Southeast University, Nanjing 210096, China
[‡]Key Laboratory of Computer Network and Information Integration (Southeast University), Ministry of Education, China
wujh915@seu.edu.cn

Min-Ling Zhang[*]

[†]School of Computer Science and Engineering, Southeast University, Nanjing 210096, China
[‡]Key Laboratory of Computer Network and Information Integration (Southeast University), Ministry of Education, China
zhangml@seu.edu.cn

ABSTRACT

Partial label learning is an emerging weakly-supervised learning framework where each training example is associated with multiple candidate labels among which only one is valid. Dimensionality reduction serves as an effective way to help improve the generalization ability of learning system, while the task of partial label dimensionality reduction is challenging due to the unknown ground-truth labeling information. In this paper, the first attempt towards partial label dimensionality reduction is investigated by endowing the popular linear discriminant analysis (LDA) techniques with the ability of dealing with partial label training examples. Specifically, a novel learning procedure named Delin is proposed which alternates between LDA dimensionality reduction and candidate label disambiguation based on estimated labeling confidences over candidate labels. On one hand, the projection matrix of LDA is optimized by utilizing disambiguation-guided labeling confidences. On the other hand, the labeling confidences are disambiguated by resorting to kNN aggregation in the LDA-induced feature space. Extensive experiments on synthetic as well as real-world partial label data sets clearly validate the effectiveness of Delin in improving the generalization ability of state-of-the-art partial label learning algorithms.

CCS CONCEPTS

• **Computing methodologies → Supervised learning**; **Machine learning algorithms**.

KEYWORDS

Partial label learning; Dimensionality reduction; Linear discriminant analysis

[*]Corresponding author

KDD '19, August 4–8, 2019, Anchorage, AK, USA

ACM ISBN 978-1-4503-6201-6/19/08…$15.00
https://doi.org/10.1145/3292500.3330901

ACM Reference Format:
Jing-Han Wu and Min-Ling Zhang. 2019. Disambiguation Enabled Linear Discriminant Analysis for Partial Label Dimensionality Reduction. In *The 25th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '19), August 4–8, 2019, Anchorage, AK, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3292500.3330901

1 INTRODUCTION

Partial label learning is one of the emerging weakly-supervised learning frameworks with ambiguous labeling [40], where each training example is associated with multiple *candidate* class labels simultaneously among which only one corresponds to the ground-truth label [7, 36]. The task of partial label learning is to learn a multi-class classification model from the partial label training examples, which can assign proper class label for the unseen instance in prediction phase. The task of learning from examples with candidate label sets naturally arises under many real-world scenarios, such as web mining [15], multimedia content analysis [4, 6, 19, 34], ecoinformatics [3, 37], natural language processing [39], etc.

It is well-known that dimensionality reduction serves as an effective way to help improve the generalization ability of learning system, and exploring dimensionality reduction mechanism for partial label learning is even more desirable as the generalization performance of partial label classification model is usually less satisfactory due to the limited supervision information available from training set. Existing works on partial label learning mainly focus on classification model induction by disambiguating the candidate label set [4, 5, 7, 10, 13, 19, 21, 32, 37], while the usefulness of dimensionality reduction for partial label learning hasn't been well investigated. Here, the major challenge for designing supervised dimensionality reduction techniques lies in that the ground-truth label of each partial label training example is not directly accessible to the learning algorithm.

In this paper, the first attempt towards partial label dimensionality reduction is investigated where a novel dimensionality reduction procedure for partial label examples named Delin, i.e. *Disambiguation Enabled LINear discriminant analysis*, is proposed. Briefly, Delin works by adapting the popular linear discriminant analysis (LDA) mechanism to accommodate the exploitation of partial label training examples. Specifically, an alternating procedure is employed to endow LDA with the ability of partial label dimensionality reduction based on estimating the labeling confidences over candidate labels. On one hand, LDA dimensionality reduction is performed by optimizing the projection matrix via the utilization of disambiguation-guided labeling confidences. On the other hand, candidate label disambiguation is performed by resorting to kNN aggregation in the feature space induced by LDA projection matrix. Comprehensive experiments conducted over synthetic and real-world partial label data sets show that the generalization performance of state-of-the-art partial label learning algorithms can be significantly improved by incorporating Delin for dimensionality reduction.

The rest of this paper is organized as follows. Section 2 presents technical details of the proposed Delin approach. Section 3 reports experimental results of comparative studies. Section 4 briefly discusses related works. Finally, Section 5 concludes.

2 THE PROPOSED APPROACH

Let $\mathcal{X} = \mathbb{R}^d$ denote the d-dimensional instance space and $\mathcal{Y} = \{l_1, l_2, \ldots, l_q\}$ denote the label space with q class labels. Given the partial label training set $\mathcal{D} = \{(\boldsymbol{x}_i, S_i) \mid 1 \leq i \leq m\}$ where $\boldsymbol{x}_i \in \mathcal{X}$ is a d-dimensional feature vector $(x_{i1}, x_{i2}, \ldots, x_{id})^\top$ and $S_i \subseteq \mathcal{Y}$ is the candidate label set associated with $\boldsymbol{x}_i$. In partial label learning, the key assumption lies in that the ground-truth label y_i for $\boldsymbol{x}_i$ resides in its candidate label set S_i (i.e. $y_i \in S_i$) which is not directly accessible to the learning algorithm. The task of partial label learning is to derive a *multi-class* classification model $f : \mathcal{X} \mapsto \mathcal{Y}$ from the training set $\mathcal{D}$.

For partial label dimensionality reduction, the task here is trying to find a projection matrix $\mathbf{W} = [\boldsymbol{w}_1, \boldsymbol{w}_2, \ldots, \boldsymbol{w}_{d'}] \in \mathbb{R}^{d\times d'}$ ($d' \ll d$) which maps the training examples $\mathbf{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_m] \in \mathbb{R}^{d\times m}$ into the projected d'-dimensional feature space $\mathbf{X}' = \mathbf{W}^\top\mathbf{X}$. Correspondingly, Delin adapts the linear discriminant analysis mechanism to learn $\mathbf{W}$ via an iterative procedure alternating between *LDA dimensionality reduction* and *candidate label disambiguation*. The alternating procedure is fulfilled by utilizing the estimated labeling confidences $\mathrm{Y} = [Y_{ij}]_{m\times q}$ which is initialized as:

$$\forall\ 1 \leq i \leq m,\ 1 \leq j \leq q:\ Y_{ij} = \begin{cases} \frac{1}{|S_i|}, & \text{if } l_j \in S_i \\ 0, & \text{otherwise} \end{cases} \tag{1}$$

Here, the constraints $\sum_{j=1}^{q} Y_{ij} = 1$ ($1 \leq i \leq m$) will be ensured to hold for each iteration of Delin.

Thereafter, technical details of the two alternating steps of Delin are scrutinized.

2.1 LDA Dimensionality Reduction

To enable multi-class LDA [9, 20] for partial label examples, the key adaptation lies in the derivation of *between-class* scatter matrix $\mathrm{S}_b \in \mathbb{R}^{d\times d}$ and *within-class* scatter matrix $\mathrm{S}_w \in \mathbb{R}^{d\times d}$ which are used to optimize the projection matrix as follows:

$$\begin{aligned} &\arg\max_{\mathbf{W}}\ \mathrm{tr}\left(\mathbf{W}^\top \mathrm{S}_b \mathbf{W}\right) \\ &\text{s.t.}:\ \boldsymbol{w}_h^\top \mathrm{S}_w \boldsymbol{w}_h = 1 \quad (1 \leq h \leq d') \end{aligned} \tag{2}$$

Given the current labeling confidence matrix Y, the global mean vector $\boldsymbol{\mu} \in \mathbb{R}^{d\times 1}$ and the class-wise mean vector $\boldsymbol{\mu}_j \in \mathbb{R}^{d\times 1}$ ($1 \leq j \leq q$) can be specified as:

$$\boldsymbol{\mu} = \frac{\sum_{i=1}^{m} \boldsymbol{x}_i}{m} \tag{3}$$

$$\boldsymbol{\mu}_j = \frac{\sum_{i=1}^{m} Y_{ij} \cdot \boldsymbol{x}_i}{\sum_{i=1}^{m} Y_{ij}} \tag{4}$$

Accordingly, the total scatter matrix $\mathrm{S}_t \in \mathbb{R}^{d\times d}$ and within-class scatter matrix S_w are derived as:

$$\begin{aligned} \mathrm{S}_t &= \sum_{i=1}^{m} (\boldsymbol{x}_i - \boldsymbol{\mu})(\boldsymbol{x}_i - \boldsymbol{\mu})^\top \\ &= \bar{\mathrm{X}}^\top \bar{\mathrm{X}} \end{aligned} \tag{5}$$

$$\mathrm{S}_w = \sum_{j=1}^{q}\sum_{i=1}^{m} Y_{ij} \cdot (\boldsymbol{x}_i - \boldsymbol{\mu}_j)(\boldsymbol{x}_i - \boldsymbol{\mu}_j)^\top \tag{6}$$

Here, $\bar{\mathrm{X}} = \mathrm{X} - \boldsymbol{\mu}\boldsymbol{e}^\top$ corresponds to the centralized training examples with $\boldsymbol{e} = [1, 1, \ldots, 1]^\top$ being the d-dimensional unit vector. Then, it is not difficult to show that the between-class scatter matrix S_b can be derived as:

$$\begin{aligned} \mathrm{S}_b &= \mathrm{S}_t - \mathrm{S}_w \\ &= \sum_{j=1}^{q}\left(\sum_{i=1}^{m} Y_{ij}\right) \cdot (\boldsymbol{\mu}_j - \boldsymbol{\mu})(\boldsymbol{\mu}_j - \boldsymbol{\mu})^\top \\ &= \bar{\mathrm{X}}^\top \mathrm{Y}\mathrm{C}^{-1}\mathrm{Y}^\top \bar{\mathrm{X}} \end{aligned} \tag{7}$$

Here, $\mathrm{C} = \mathrm{diag}[c_1, c_2, \ldots, c_q]$ corresponds to the $q \times q$ diagonal matrix with diagonal element $c_j = \sum_{i=1}^{m} Y_{ij}$ ($1 \leq j \leq q$).

By introducing Lagrange multipliers λ_h ($1 \leq h \leq q$) to Eq.(2), the Lagrange function for each projection vector $\boldsymbol{w}_h$ ($1 \leq h \leq d'$) in $\mathbf{W}$ corresponds to:

$$L(\boldsymbol{w}_h, \lambda_h) = \boldsymbol{w}_h^\top \mathrm{S}_b \boldsymbol{w}_h - \lambda_h(\boldsymbol{w}_h^\top \mathrm{S}_w \boldsymbol{w}_h - 1) \tag{8}$$

By setting $\frac{\partial L(\boldsymbol{w}_h, \lambda_h)}{\partial \boldsymbol{w}_h} = \mathbf{0}$, we can have the necessary condition for the optimal solution of $\boldsymbol{w}_h$:

$$\left(\mathrm{S}_w^{-1}\mathrm{S}_b\right)\boldsymbol{w}_h = \lambda_h \boldsymbol{w}_h \tag{9}$$

In other words, λ_h and $\boldsymbol{w}_h$ should be an eigenvalue and its corresponding eigenvector of $\mathrm{S}_w^{-1}\mathrm{S}_b$. Therefore, Delin chooses the eigenvectors w.r.t. the top d' eigenvalues of $\mathrm{S}_w^{-1}\mathrm{S}_b$ to form the LDA projection matrix $\mathbf{W}$.

2.2 Candidate Label Disambiguation

Based on the LDA projection matrix, the original partial label training examples can be mapped into the LDA-induced feature space $\mathcal{D}' = \{(\boldsymbol{x}_i', S_i) \mid \boldsymbol{x}_i' = \mathbf{W}^\top \boldsymbol{x}_i, 1 \leq i \leq m\}$. Thereafter, the labeling confidence matrix will be updated to $Y' = [Y_{ij}']_{m\times q}$ by utilizing kNN-based candidate label disambiguation.

For each instance $\boldsymbol{x}_i' \in \mathbb{R}^{d'}$, its k nearest neighbors identified in $\mathcal{D}'$ is denoted as $\mathcal{N}(\boldsymbol{x}_i')$. A weighted voting matrix $\mathbf{Z} = [Z_{ij}]_{m\times q}$ is calculated by aggregating the labeling assignment of each neighboring example in $\mathcal{N}(\boldsymbol{x}_i')$:

$$\begin{aligned} &\forall 1 \leq i \leq m,\ 1 \leq j \leq q: \\ &Z_{ij} = \sum_{(\boldsymbol{x}_a', S_a)\in\mathcal{N}(\boldsymbol{x}_i')} Y_{aj} \cdot [\![l_j \in S_a]\!] \cdot \omega_a \end{aligned} \tag{10}$$

For any predicate π, $[\![\pi]\!]$ returns 1 if π holds and 0 otherwise. Furthermore, for the a-th nearest neighbor ($1 \leq a \leq k$), the voting

Table 1: The pseudo-code of Delin.

Inputs:

$\mathcal{D}$: the partial label training set $\{(\boldsymbol{x}_i, S_i) \mid 1 \leq i \leq m\}$ $(\mathcal{X} = \mathbb{R}^d, \mathcal{Y} = \{l_1, l_2, \ldots, l_q\}, \boldsymbol{x}_i \in \mathcal{X}, S_i \subseteq \mathcal{Y})$

d': the number of retained features after dimensionality reduction

k: the number of nearest neighbors used for candidate label disambiguation

Outputs:

$\mathbf{W}$: the $d \times d'$ projection matrix learned by the proposed approach

Process:

1: Initialize the $m \times q$ labeling confidence matrix $\mathbf{Y}$ according to Eq.(1);
2: Specify the global mean vector $\boldsymbol{\mu}$ according to Eq.(3);
3: **repeat**
4: Specify the class-wise mean vector $\boldsymbol{\mu}_j$ $(1 \leq j \leq q)$ according to Eq.(4);
5: Derive the total scatter matrix S_t and within-class scatter matrix S_w according to Eq.(5) and Eq.(6) respectively;
6: Derive the between-class scatter matrix S_b according to Eq.(7);
7: Form the LDA projection matrix $\mathbf{W} = [\boldsymbol{w}_1, \boldsymbol{w}_2, \ldots, \boldsymbol{w}_{d'}]$ with $\boldsymbol{w}_h$ $(1 \leq h \leq q)$ set to be the eigenvector w.r.t. the top-h eigenvalue of
8: $\mathrm{S}_w^{-1}\mathrm{S}_b$ satisfying $\boldsymbol{w}_h^\top \mathrm{S}_w \boldsymbol{w}_h = 1$;
9: Derive the partial label training set in LDA-induced feature space $\mathcal{D}' = \{(\boldsymbol{x}'_i, S_i) \mid \boldsymbol{x}'_i = \mathbf{W}^\top \boldsymbol{x}_i, 1 \leq i \leq m\}$;
10: **for** i=1 to m **do**
11: Identify the k-nearest neighbors of $\boldsymbol{x}'_i$ in $\mathcal{D}'$ as $\mathcal{N}(\boldsymbol{x}'_i)$;
12: **end for**
13: Calculate the $m \times q$ weighted voting matrix $\mathbf{Z}$ via kNN aggregation according to Eq.(10);
14: Calculate the $m \times q$ counting matrix $\mathbf{V}$ according to Eq.(11);
15: Specify the updated labeling confidence matrix $\mathbf{Y}'$ according to Eqs.(12)-(13);
16: Let $\mathbf{Y} = \mathbf{Y}'$;
17: **until** convergence
18: Return the learned partial label LDA projection matrix $\mathbf{W}$.

weight is set as $\omega_a = k - a + 1$ [13, 35]. Meanwhile, a counting matrix $\mathbf{V} = [V_{ij}]_{m \times q}$ is specified as:

$$\forall 1 \leq i \leq m,\ 1 \leq j \leq q: \quad V_{ij} = \sum_{(\boldsymbol{x}'_a, S_a) \in \mathcal{N}(\boldsymbol{x}'_i)} [\![l_j \in S_a]\!] \tag{11}$$

Here, V_{ij} stores the number of k nearest neighbors of $\boldsymbol{x}'_i$ which take l_j as their candidate label.

Among the set of candidate labels S_i for $\boldsymbol{x}'_i$, the one with largest weighted voting is denoted as l_{j^*}:[1]

$$l_{j^*} = \arg\max_{l_j \in S_i} Z_{ij} \tag{12}$$

Then, the updated labeling confidence matrix $\mathbf{Y}'$ will be set as:

$$\forall\ 1 \leq i \leq m,\ 1 \leq j \leq q: \quad Y'_{ij} = \begin{cases} [\![j = j^*]\!], & \text{if } |S_i| = 1 \\ \frac{V_{ij^*}}{k}, & \text{if } |S_i| > 1 \text{ and } j = j^* \\ \left(1 - \frac{V_{ij^*}}{k}\right) / (|S_i| - 1), & \text{if } |S_i| > 1 \text{ and } j \neq j^* \end{cases} \tag{13}$$

[1]In case that there are more than one class label which have the same largest weighted voting, one of them will be randomly selected to instantiate l_{j^*}.

Table 1 summarizes the complete procedure of Delin. Firstly, the labeling confidence matrix is initialized based on the candidate label assignment (Step 1) and the global mean vector is specified by averaging all training examples (Step 2). After that, an iterative procedure alternating between LDA dimensionality reduction (Steps 4-8) and candidate label disambiguation (Steps 9-16) is conducted. Finally, the resulting LDA projection matrix $\mathbf{W}$ is returned (Step 18). Here, the iterative procedure terminates if $\mathbf{W}$ does not change or the maximum number of iterations is reached.[2]

3 EXPERIMENTS

3.1 Experimental Setup

To evaluate the effectiveness of the proposed partial label dimensionality reduction approach, Delin is coupled with state-of-the-art partial label learning algorithms for performance evaluation. Given the partial label learning algorithm $\mathcal{A}$, its coupling version with Delin is denoted as $\mathcal{A}$-Delin which learns from partial label training examples in the LDA-induced feature space. Accordingly,

[2]In this paper, the maximum number of iterations is set to be 75 which suffices to yield stable performance for the proposed approach.

Table 2: Characteristics of the synthetic experimental data sets.

Data Set	# Examples	# Features	# Class Labels	# False Positive Labels (r)	Task Domain
mediamill	2,854	120	10	$r = 1, 2, 3$	*video semantic detection* [26]
tmc2007	8,670	981	18	$r = 1, 2, 3$	*text anomaly detection* [28]
slashdot	3,142	1,079	19	$r = 1, 2, 3$	*text classification* [18]
amazon	1,500	1,326	50	$r = 1, 2, 3$	*authorship identification* [8]
DeliciousMIL	1,409	1,389	20	$r = 1, 2, 3$	*sentence labeling* [27]
bookmark	2,500	1,413	57	$r = 1, 2, 3$	*automatic tag suggestion* [17]
sports	9,120	1,738	19	$r = 1, 2, 3$	*human activity recognition* [1]
sector	6,412	6,104	105	$r = 1, 2, 3$	*text classification* [25]

Table 3: Classification accuracy (mean±std) of each comparing algorithm on controlled synthetic data sets (with one false positive candidate label [$r = 1$]). For partial label learning algorithm $\mathcal{A} \in \{$Pl-knn, Pl-svm, Pl-ecoc, Ipal$\}$, the performance of $\mathcal{A}$-Delin is compared against that of $\mathcal{A}$ where the better performance is shown in boldface.

Data Set	Comparing Algorithm							
	Pl-knn	Pl-knn-Delin	Pl-svm	Pl-svm-Delin	Pl-ecoc	Pl-ecoc-Delin	Ipal	Ipal-Delin
mediamill	0.637±0.034	**0.688±0.027**	0.495±0.042	**0.600±0.035**	0.592±0.037	**0.666±0.037**	**0.642±0.020**	0.640±0.037
tmc2007	0.402±0.012	**0.654±0.013**	0.645±0.021	**0.666±0.013**	0.635±0.016	**0.669±0.013**	0.598±0.019	**0.610±0.019**
slashdot	0.163±0.022	**0.698±0.033**	0.595±0.018	**0.717±0.029**	0.528±0.033	**0.719±0.027**	0.417±0.023	**0.694±0.027**
amazon	0.025±0.014	**0.609±0.048**	0.120±0.026	**0.558±0.038**	0.065±0.021	**0.608±0.046**	0.105±0.023	**0.610±0.048**
DeliciousMIL	0.033±0.039	**0.464±0.043**	0.036±0.017	**0.354±0.043**	0.072±0.038	**0.464±0.042**	0.062±0.017	**0.463±0.044**
bookmark	0.170±0.026	**0.536±0.036**	0.279±0.029	**0.543±0.037**	0.325±0.039	**0.550±0.033**	0.309±0.030	**0.550±0.027**
sports	0.288±0.015	**0.865±0.014**	0.677±0.019	**0.709±0.013**	0.697±0.031	**0.851±0.013**	**0.905±0.009**	0.880±0.011
sector	0.014±0.005	**0.530±0.034**	0.070±0.012	**0.496±0.035**	0.058±0.012	**0.527±0.033**	0.144±0.015	**0.531±0.034**

the performance of $\mathcal{A}$-Delin is compared against that of $\mathcal{A}$ to verify whether the proposed dimensionality reduction techniques do help improve generalization ability of the learning system.

In this paper, the following state-of-the-art partial label learning algorithms are utilized to instantiate $\mathcal{A}$ with parameter configuration suggested in respective literatures:

- Pl-knn [13]: an instance-based partial label learning approach which makes prediction for unseen instance by employing the kNN rule with weighted voting [suggested configuration: k=10].
- Pl-svm [21]: a maximum-margin partial label learning approach which learns the predictive model by maximizing the classification margin over candidate and non-candidate class labels [suggested configuration: regularization parameter pool with $\{10^{-3}, \ldots, 10^3\}$].
- Pl-ecoc [36]: a transformation-based partial label learning approach which learns the predictive model by decomposing the original partial label learning problem into a number of binary learning problems via error-correcting output codes (ECOC) [suggested configuration: ECOC coding length $\lceil 10 \cdot \log_2(q) \rceil$].
- Ipal [32]: another instance-based partial label learning approach which makes prediction for unseen instance by employing graph-based disambiguation with label propagation [suggested configuration: balancing parameter $\alpha = 0.95$].

As shown in Table 1, the parameters d' and k for Delin are set to be $\lceil thr \cdot \min(q, d) \rceil$ with $thr = 0.6$ and $k = 8$ respectively.

Furthermore, comparative studies are conducted on both synthetic and real-world data sets in this paper. On each data set, ten-fold cross-validation is performed and the mean predictive accuracy as well as standard deviation are recorded.

3.2 Synthetic Data Sets

Following the widely-used experimental protocol in partial label learning [4, 5, 7, 10, 19, 32, 36], synthetic partial label data set can be generated from multi-class data set with controlling parameter r. Here, r specifies the number of false positive labels in the candidate label set (i.e. $|S_i| = r + 1$). Specifically, for any multi-class example $(\boldsymbol{x}_i, y_i)$, a partial label training example $(\boldsymbol{x}_i, S_i)$ is generated by randomly adding r class labels from $\mathcal{Y}$ into S_i.

Table 2 summarizes characteristics of the synthetic data sets used for experimental studies with $r \in \{1, 2, 3\}$, which are roughly ordered according to the dimensionality of each data set.[3] Accordingly, Tables 3 to 5 report the detailed experimental results of each comparing algorithm with $r = 1, 2, 3$ respectively. Given partial label learning algorithm $\mathcal{A} \in \{$Pl-knn, Pl-svm, Pl-ecoc, Ipal$\}$, $\mathcal{A}$-Delin is compared against $\mathcal{A}$ where the better predictive performance is shown in boldface.

[3] Most data sets in Table 2 are derived from multi-label benchmark data sets [41] by retaining examples with only one relevant label.

Table 4: Classification accuracy (mean±std) of each comparing algorithm on controlled synthetic data sets (with two false positive candidate label [$r = 2$]). For partial label learning algorithm $\mathcal{A} \in$ {Pl-knn, Pl-svm, Pl-ecoc, Ipal}, the performance of $\mathcal{A}$-Delin is compared against that of $\mathcal{A}$ where the better performance is shown in boldface.

Data Set	Comparing Algorithm							
	Pl-knn	Pl-knn-Delin	Pl-svm	Pl-svm-Delin	Pl-ecoc	Pl-ecoc-Delin	Ipal	Ipal-Delin
mediamill	0.623±0.023	**0.665±0.036**	0.490±0.041	**0.608±0.016**	0.514±0.036	**0.598±0.039**	0.592±0.023	**0.597±0.030**
tmc2007	0.379±0.016	**0.650±0.013**	0.631±0.039	**0.668±0.016**	0.584±0.027	**0.653±0.017**	0.583±0.009	**0.606±0.018**
slashdot	0.160±0.020	**0.668±0.018**	0.575±0.029	**0.687±0.024**	0.428±0.035	**0.688±0.023**	0.402±0.025	**0.664±0.023**
amazon	0.021±0.009	**0.466±0.021**	0.073±0.021	**0.438±0.023**	0.040±0.016	**0.466±0.022**	0.088±0.020	**0.468±0.021**
DeliciousMIL	0.027±0.014	**0.258±0.042**	0.035±0.019	**0.220±0.038**	0.063±0.034	**0.253±0.039**	0.052±0.011	**0.258±0.042**
bookmark	0.162±0.012	**0.486±0.033**	0.261±0.019	**0.504±0.030**	0.284±0.035	**0.495±0.033**	0.304±0.018	**0.499±0.038**
sports	0.290±0.015	**0.842±0.018**	0.640±0.015	**0.686±0.015**	0.601±0.037	**0.818±0.013**	**0.901±0.008**	0.863±0.013
sector	0.015±0.007	**0.392±0.022**	0.054±0.011	**0.373±0.022**	0.036±0.009	**0.390±0.022**	0.136±0.009	**0.392±0.022**

Table 5: Classification accuracy (mean±std) of each comparing algorithm on controlled synthetic data sets (with three false positive candidate label [$r = 3$]). For partial label learning algorithm $\mathcal{A} \in$ {Pl-knn, Pl-svm, Pl-ecoc, Ipal}, the performance of $\mathcal{A}$-Delin is compared against that of $\mathcal{A}$ where the better performance is shown in boldface.

Data Set	Comparing Algorithm							
	Pl-knn	Pl-knn-Delin	Pl-svm	Pl-svm-Delin	Pl-ecoc	Pl-ecoc-Delin	Ipal	Ipal-Delin
mediamill	0.598±0.017	**0.656±0.022**	0.471±0.039	**0.602±0.031**	0.101±0.024	**0.231±0.113**	0.525±0.024	**0.564±0.026**
tmc2007	0.364±0.011	**0.627±0.013**	0.619±0.035	**0.659±0.018**	0.568±0.021	**0.576±0.033**	0.557±0.016	**0.593±0.013**
slashdot	0.165±0.030	**0.642±0.033**	0.562±0.038	**0.667±0.035**	0.373±0.039	**0.645±0.035**	0.373±0.030	**0.639±0.038**
amazon	0.021±0.008	**0.347±0.027**	0.055±0.019	**0.309±0.030**	0.031±0.017	**0.346±0.026**	0.084±0.024	**0.349±0.027**
DeliciousMIL	0.043±0.022	**0.198±0.035**	0.038±0.020	**0.158±0.032**	0.063±0.036	**0.188±0.032**	0.044±0.015	**0.197±0.036**
bookmark	0.140±0.012	**0.437±0.037**	0.247±0.028	**0.452±0.040**	0.203±0.043	**0.443±0.036**	0.293±0.042	**0.447±0.032**
sports	0.292±0.021	**0.824±0.012**	0.603±0.019	**0.641±0.021**	0.492±0.043	**0.762±0.022**	**0.892±0.009**	0.840±0.019
sector	0.017±0.005	**0.295±0.018**	0.047±0.008	**0.273±0.019**	0.020±0.007	**0.293±0.017**	0.133±0.013	**0.294±0.017**

Table 6: Win/tie/loss counts (pairwise t-test at 0.05 significance level) between $\mathcal{A}$-Delin and $\mathcal{A}$ in terms of different number of false positive labels ($r = 1, 2, 3$).

	$\mathcal{A}$-Delin against $\mathcal{A}$			
	$\mathcal{A}$=Pl-knn	$\mathcal{A}$=Pl-svm	$\mathcal{A}$=Pl-ecoc	$\mathcal{A}$=Ipal
$r = 1$	8/0/0	8/0/0	8/0/0	6/1/1
$r = 2$	8/0/0	8/0/0	8/0/0	6/1/1
$r = 3$	8/0/0	8/0/0	8/0/0	7/0/1
In Total	**24/0/0**	**24/0/0**	**24/0/0**	**19/2/3**

Pairwise t-test at 0.05 significance level is further conducted to show whether the performance difference between $\mathcal{A}$ and $\mathcal{A}$-Delin is significant, where the resulting win/tie/loss counts are reported in Table 6. Based on these results, it is impressive to observe that:

- For Pl-knn, the performance improvement of Pl-knn-Delin against Pl-knn is moderate on `mediamill` which corresponds to the synthetic data set with smallest number of features. On the rest seven data sets in Table 2 with larger number of features, the predictive performance of Pl-knn has been greatly improved by incorporating the proposed dimensionality reduction techniques. Specifically, for `tmc2007` on which Pl-knn has the highest predictive accuracy, the classification accuracy has been improved with Delin by 25.2%, 27.1% and 26.3% for $r = 1$, 2 and 3 respectively. For `sector` on which Pl-knn has the lowest predictive accuracy, the performance improvement with Delin is even more pronounced by an increase of 51.6%, 37.7% and 27.8% for $r = 1$, 2 and 3 respectively.
- For Pl-svm and Pl-ecoc, the performance of both algorithms have been significantly improved on all the eight synthetic data sets. On the five data sets with more than 1,300 features (i.e. `amazon`, `DeliciousMIL`, `bookmark`, `sports` and `sector`), out of the 30 statistical comparisons (2 algorithms x 5 data sets x 3 configurations of r), the classification accuracy has been improved with Delin by more than 20.0% in 22 cases. These results indicate that the benefits brought by Delin would be more significant when the dimensionality of the feature space is high.

Table 7: Characteristics of the real-world experimental data sets.

Data Set	# Examples	# Features	# Class Labels	average # Candidate Labels	Task Domain
Lost	1,122	108	16	2.23	*automatic face naming* [7]
Yahoo! News	22,991	163	219	1.91	*automatic face naming* [11]
FG-NET	1,002	262	78	7.48	*facial age estimation* [22]
Soccer Player	17,472	279	171	2.09	*automatic face naming* [34]
Mirflickr	2,780	1,536	14	2.76	*web image classification* [12]

- For IPAL, the predictive performance of IPAL-DELIN is outperformed by IPAL on `sports` which corresponds to the synthetic data set with largest number of examples. On the other hand, on the two data sets with smallest number of examples (i.e. `amazon`, `DeliciousMIL`), the classification accuracy has been significantly improved with DELIN by more than 40.0%, 20.0% and 15.0% for $r = 1$, 2 and 3 respectively. These results indicate that the benefits brought by DELIN would be more significant when the number of available training examples is insufficient.

3.3 Real-World Data Sets

Table 7 summarizes characteristics of the real-world partial label data sets from different task domains, including `FG-NET` [22] for facial age estimation, `Lost` [7], `Soccer Player` [34] and `Yahoo! News` [11] for automatic face naming from images or videos, and `Mirflickr` [12] for web image classification.[4] For *facial age estimation*, human faces with landmarks are represented as instances while ages annotated by crowdsourced labelers are regarded as candidate labels. For *automatic face naming*, faces cropped from an image or video frame are represented as instances while names extracted from the associated captions or subtitles are regarded as candidate labels. For *web image classification*, web images are represented as instances while annotations extracted from the web environment are regarded as candidate labels.

[4]Data sets available at: http://palm.seu.edu.cn/zhangml/Resources.htm#partial_data

Table 8: Win/tie/loss statistic (pairwise t-test at 0.05 significance level) between $\mathcal{A}$-DELIN and $\mathcal{A}$ on each real-world partial label data set.

	$\mathcal{A}$-DELIN against $\mathcal{A}$			
	$\mathcal{A}$=PL-KNN	$\mathcal{A}$=PL-SVM	$\mathcal{A}$=PL-ECOC	$\mathcal{A}$=IPAL
Lost	win	win	win	win
Yahoo! News	win	tie	win	win
FG-NET	win	win	win	win
Soccer Player	tie	win	win	win
Mirflickr	win	win	win	win
In Total	4/1/0	4/1/0	5/0/0	5/0/0

Figure 1 illustrates the predictive accuracy of each partial label training algorithm before and after employing the proposed dimensionality reduction techniques. Furthermore, Table 8 reports the win/tie/loss statistics based on pairwise t-test at 0.05 significance level on each real-world experimental data set. From the above results, it is also impressive to observe that:

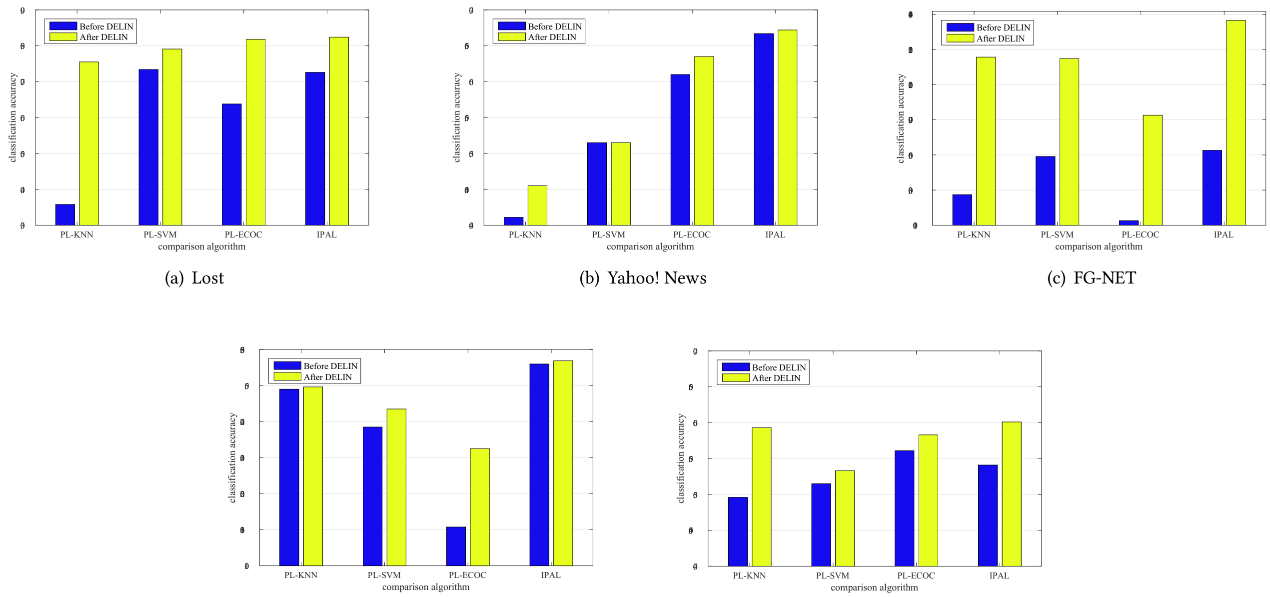
- Out of the 20 statistical comparisons (4 algorithms x 5 data sets), the predictive performance has been significantly improved by employing DELIN in 18 cases. There are only two ties on data sets `Soccer Player` (PL-KNN-DELIN against PL-KNN) and `Yahoo! News` (PL-SVM-DELIN against PL-SVM) which have the largest number of class labels among the real-world partial label data sets.
- As shown in Fig. 1(c), the relative performance improvement is rather pronounced on `FG-NET` which is most difficult to learn with smallest number of examples but largest average number of candidate labels. Specifically, the classification accuracy of each partial label learning algorithm has at least been doubled on `FG-NET`. These results indicate that the benefits brought by DELIN would be more significant under difficult learning scenarios.

3.4 Sensitivity Analysis

As shown in Table 1, the number of retained features after dimensionality reduction (i.e. d') serves as the key parameter for DELIN. Following the common practice of applying LDA for multi-class classification [9, 20], we set $d' = \lceil thr \cdot \min(q, d) \rceil$ with $thr \in (0, 1)$ which is less than the number of class labels.

Table 9 reports the predictive accuracy of applying DELIN to partial label learning algorithm on all real-world data sets with varying number of retained features. Here, thr increases from 0.5 to 0.9 with an interval of 0.1 and the best performance across different values of thr is shown in boldface. As shown in Table 9, the performance of each partial label learning algorithm fluctuates moderately by incorporating DELIN for dimensionality reduction as the value of thr changes. Furthermore, there is no single configuration of thr which can yield best performance in most cases. Therefore, the value of thr is fixed to be 0.6 in this paper while DELIN may lead to further performance improvement by fine-tuning parameter thr on the training set.

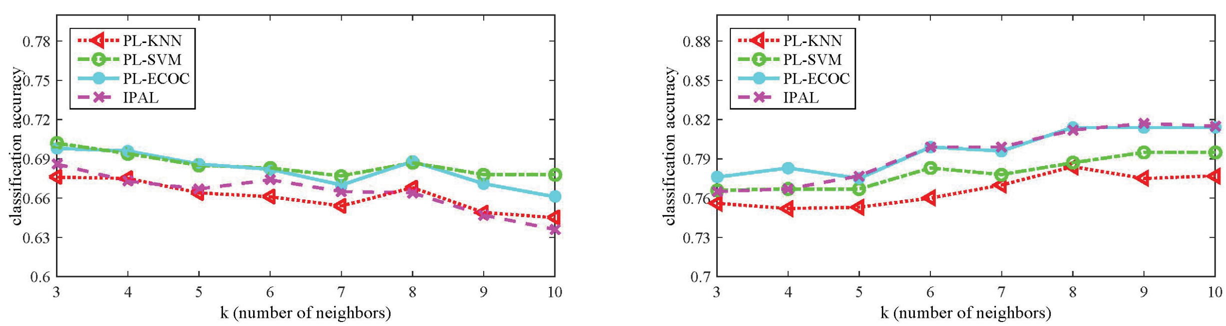
In addition to d', Figure 2 illustrates how the predictive performance of each algorithm changes w.r.t. the other parameter k, i.e. the number of nearest neighbors used for candidate label disambiguation. Here, k increases from 3 to 10 with an interval of 1. As



(a) Lost (b) Yahoo! News (c) FG-NET

(d) Soccer Player (e) Mirflickr

Figure 1: Comparison of the classification accuracy of each partial label learning algorithm on real-world data sets before (*blue* bar) and after (*yellow* bar) employing Delin.



(a) `slashdot` (r=2) (b) `Lost`

Figure 2: Predictive accuracy of $\mathcal{A}$-Delin ($\mathcal{A}$ ∈{Pl-knn, Pl-svm, Pl-ecoc, Ipal}) changes as the number of nearest neighbors used for candidate label disambiguation (i.e. k) increases from 3 to 10 with an interval of 1. Left: synthetic data set `slashdot` with r = 2; Right: real-world data set `Lost`.

shown in Figure 2, on either the synthetic data set `slashdot` (r = 2) or the real-world data set `Lost`, the performance of each partial label learning algorithm by incorporating Delin is relatively stable as the value of k changes. Therefore, the value of k is fixed to be 8 in this paper.

4 RELATED WORKS

As a *weakly-supervised* learning framework [40], partial label learning deals with *implicit* supervision information where the ground-truth label is concealed in the candidate label set of each training example. Partial label learning is related to other well-established weakly-supervised learning frameworks including *semi-supervised learning*, *multi-instance learning* and *multi-label learning*. The differences between partial label learning and other related learning frameworks lie in the form of weak supervision information to be dealt with. Specifically, semi-supervised learning deals with unlabeled examples with *blind* supervision information [42], multi-instance learning deals with bag-of-instances examples with *ambiguous* supervision information [2], and multi-label learning deals with multi-label examples with *non-unique* supervision information [41].

Table 9: Predictive accuracy of $\mathcal{A}$-Delin ($\mathcal{A}$ ∈{Pl-knn, Pl-svm, Pl-ecoc, Ipal}) changes as the number of retained features ($d' = \lceil thr \cdot \min(q, d) \rceil$) varies with thr increases from 0.5 to 0.9 with an interval of 0.1. On each data set, the best performance across different values of thr is shown in boldface. Furthermore, the predictive accuracy of $\mathcal{A}$ on the original feature space is also shown in the lower part of the table for reference purpose (after the dashed line).

Data Set	thr	# Retained Features	Pl-knn-Delin	Pl-svm-Delin	Pl-ecoc-Delin	Ipal-Delin
	0.5	8	0.790±0.050	0.790±0.051	0.794±0.046	0.792±0.049
	0.6	10	0.784±0.031	0.787±0.035	0.814±0.046	0.812±0.046
Lost	0.7	12	0.808±0.046	0.813±0.046	0.842±0.050	0.833±0.051
	0.8	13	**0.823±0.045**	**0.822±0.044**	**0.845±0.043**	**0.858±0.051**
	0.9	15	0.790±0.027	0.790±0.032	0.819±0.039	0.823±0.039
	0.5	82	**0.475±0.006**	0.509±0.008	**0.639±0.007**	0.671±0.005
	0.6	98	0.455±0.009	0.515±0.010	0.635±0.007	**0.672±0.006**
Yahoo! News	0.7	115	0.437±0.007	0.517±0.011	0.628±0.007	0.671±0.004
	0.8	131	0.424±0.004	**0.518±0.009**	0.621±0.008	0.667±0.007
	0.9	147	0.413±0.005	**0.518±0.009**	0.615±0.009	0.666±0.005
	0.5	39	**0.128±0.032**	0.115±0.030	0.076±0.028	0.143±0.036
	0.6	47	0.120±0.011	0.119±0.027	**0.082±0.035**	**0.144±0.037**
FG-NET	0.7	55	0.090±0.031	0.116±0.036	0.067±0.029	0.114±0.040
	0.8	63	0.090±0.023	**0.122±0.031**	0.079±0.032	0.128±0.016
	0.9	71	0.074±0.025	0.119±0.031	0.066±0.029	0.132±0.019
	0.5	86	**0.497±0.013**	0.445±0.027	0.323±0.062	**0.556±0.015**
	0.6	103	**0.497±0.012**	0.448±0.033	**0.360±0.054**	0.555±0.012
Soccer Player	0.7	120	0.494±0.014	0.449±0.043	0.288±0.072	0.554±0.013
	0.8	137	0.493±0.013	**0.450±0.039**	0.297±0.065	0.554±0.013
	0.9	154	0.494±0.014	0.435±0.049	0.287±0.074	0.552±0.013
	0.5	7	0.579±0.077	0.504±0.159	0.507±0.132	0.538±0.099
	0.6	9	**0.593±0.011**	0.533±0.134	**0.583±0.118**	**0.601±0.115**
Mirflickr	0.7	10	0.523±0.117	0.543±0.100	0.526±0.113	0.534±0.105
	0.8	12	0.501±0.120	0.554±0.097	0.512±0.126	0.513±0.122
	0.9	13	0.499±0.106	**0.555±0.085**	0.523±0.101	0.513±0.106
		# Original Features	Pl-knn	Pl-svm	Pl-ecoc	Ipal
Lost	-	108	0.358±0.029	0.734±0.004	0.638±0.051	0.726±0.041
Yahoo! News	-	163	0.411±0.005	0.515±0.001	0.610±0.009	0.667±0.005
FG-NET	-	262	0.030±0.019	0.055±0.024	0.013±0.015	0.059±0.019
Soccer Player	-	279	0.492±0.014	0.408±0.043	0.186±0.064	0.548±0.014
Mirflickr	-	1,536	0.496±0.127	0.515±0.127	0.561±0.013	0.541±0.129

To learn from partial label examples, the major strategy is trying to disambiguate the candidate label set so as to recover the ground-truth labeling information. One way towards disambiguation is to treat the ground-truth label as latent variable whose value is estimated via iterative optimization procedure such as EM. The objective function can be instantiated based on the maximum likelihood criterion where the likelihood is defined as the probability of observing each partial label training example over its candidate label set [16, 19], or the maximum margin criterion where the classification margin is defined over the predictive difference between candidate labels and non-candidate labels of each partial label training example [21, 32].

Another way towards disambiguation is to treat all candidate labels in an equal manner and make final prediction by averaging their modeling outputs. For discriminative models, the averaged output from all candidate labels is distinguished from the outputs from non-candidate labels [7, 30]. For instance-based models, the predicted class label for unseen instance is determined by the voting among candidate labels of its neighboring examples [10, 13, 35]. Note that for the proposed Delin approach, kNN techniques have also been utilized to help disambiguate the candidate label set by further exploiting the estimated labeling confidences over candidate labels.

The task of dimensionality reduction for data associated with multiple valid class labels have been well studied [14, 23, 24, 29, 31, 33, 38], while to the best of our knowledge the same task for data associated with multiple candidate labels has not been well investigated. Other than performing transformation in the feature space with dimensionality reduction, there have been some works which perform transformation in the label space by decomposing the partial label learning problem into binary classification [7, 36], multi-class classification [5], or regression [37] problems.

5 CONCLUSION

In this paper, the problem of dimensionality reduction for partial label learning is investigated. Accordingly, a novel partial label dimensionality reduction approach is proposed which works in an iterative manner by alternating between LDA dimensionality reduction and candidate label disambiguation. Comparative experiments over a number of synthetic and real-world data sets show that state-of-the-art partial label learning algorithms can significantly benefit from the proposed dimensionality reduction approach in improving their generalization performance.

ACKNOWLEDGMENTS

The authors wish to thank the anonymous reviewers for their helpful comments and suggestions, and the Big Data Center of Southeast University for providing the facility support on the numerical calculations in this paper. This work was supported by the National Key R&D Program of China (2018YFB1004300), the National Science Foundation of China (61573104), and partially supported by the Collaborative Innovation Center of Novel Software Technology and Industrialization.

REFERENCES

[1] K. Altun and B. Barshan. 2010. Human activity recognition using inertial/magnetic sensor units. In *Proceedings of the 1st International Conference on Human Behavior Understanding*. Istanbul, Turkey, 38–51.

[2] J. Amores. 2013. Multiple instance classification: Review, taxonomy and comparative study. *Artificial Intelligence* 201 (2013), 81–105.

[3] F. Briggs, X. Z. Fern, and R. Raich. 2012. Rank-loss support instance machines for MIML instance annotation. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Beijing, China, 534–542.

[4] C.-H. Chen, V. M. Patel, and R. Chellappa. 2018. Learning from ambiguously labeled face images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 40, 7 (2018), 1653–1667.

[5] Y.-C. Chen, V. M. Patel, R. Chellappa, and P. J. Phillips. 2014. Ambiguously labeled learning using dictionaries. *IEEE Transactios on Information Forensics and Security* 9, 12 (2014), 2076–2088.

[6] T. Cour, B. Sapp, C. Jordan, and B. Taskar. 2009. Learning from ambiguously labeled images. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. Miami, FL, 919–926.

[7] T. Cour, B. Sapp, and B. Taskar. 2011. Learning from partial labels. *Journal of Machine Learning Research* 12, May (2011), 1501–1536.

[8] D. Dheeru and E. Karra Taniskidou. 2017. UCI Machine Learning Repository. http://archive.ics.uci.edu/ml

[9] K. Fukunaga. 2013. *Introduction to Statistical Pattern Recognition* (2nd edition ed.). Academic Press, Cambridge, MA.

[10] C. Gong, T. Liu, Y. Tang, J. Yang, J. Yang, and D. Tao. 2018. A regularization approach for instance-based superset label learning. *IEEE Transactions on Cybernetics* 48, 3 (2018), 967–978.

[11] M. Guillaumin, J. Verbeek, and C. Schmid. 2010. Multiple instance metric learning from automatically labeled bags of faces. In *Lecture Notes in Computer Science 6311*, K. Daniilidis, P. Maragos, and N. Paragios (Eds.). Springer, Berlin, 634–647.

[12] M. J. Huiskes and M. S. Lew. 2008. The MIR Flickr Retrieval Evaluation. In *Proceedings of the 1st ACM International Conference on Multimedia Information Retrieval*. Vancouver, Canada, 39–43.

[13] E. Hüllermeier and J. Beringer. 2006. Learning from ambiguously labeled examples. *Intelligent Data Analysis* 10, 5 (2006), 419–439.

[14] S. Ji and J. Ye. 2009. Linear dimensionality reduction for multi-label classification. In *Proceedings of the 21st International Joint Conference on Artificial Intelligence*. Pasadena, TX, 1077–1082.

[15] L. Jie and F. Orabona. 2010. Learning from candidate labeling sets. In *Advances in Neural Information Processing Systems 23*, J. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. S. Zemel, and A. Culotta (Eds.). MIT Press, Cambridge, MA, 1504–1512.

[16] R. Jin and Z. Ghahramani. 2003. Learning with multiple labels. In *Advances in Neural Information Processing Systems 15*, S. Becker, S. Thrun, and K. Obermayer (Eds.). MIT Press, Cambridge, MA, 897–904.

[17] I. Katakis, G. Tsoumakas, and I. Vlahavas. 2008. Multilabel Text Classification for Automated Tag Suggestion. In *Proceedings of the ECML/PKDD 2008 Discovery Challenge*. Antwerp, Belgium.

[18] J. Leskovec, K. Lang, A. Dasgupta, and M. Mahoney. 2009. Community Structure in Large Networks: Natural Cluster Sizes and the Absence of Large Well-Defined Clusters. *Internet Mathematics* 6, 1 (2009), 29–123.

[19] L. Liu and T. Dietterich. 2012. A conditional multinomial mixture model for superset label learning. In *Advances in Neural Information Processing Systems 25*, P. Bartlett, F. C. N. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger (Eds.). MIT Press, Cambridge, MA, 557–565.

[20] G. J. McLachlan. 2004. *Discriminant Analysis and Statistical Pattern Recognition*. John Wiley & Sons, Inc., Hoboken, NJ.

[21] N. Nguyen and R. Caruana. 2008. Classification with partial labels. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Las Vegas, NV, 381–389.

[22] G. Panis and A. Lanitis. 2015. An overview of research activities in facial age estimation using the FG-NET aging database. In *Lecture Notes in Computer Science 8926*, C. Rother L. Agapito, M. M. Bronstein (Ed.). Springer, Berlin, 737–750.

[23] R. B. Pereira, A. Plastino, B. Zadrozny, and L. H. C. Merschmann. 2018. Categorizing feature selection methods for multi-label classification. *Artificial Intelligence Review* 49, 1 (2018), 57–78.

[24] B. Qian and I. Davidson. 2010. Semi-supervised dimension reduction for multi-label classification. In *Proceedings of the 24th AAAI Conference on Artificial Intelligence*. Atlanta, GA, 569–574.

[25] J. D. M. Rennie and R. Rifkin. 2001. *Improving multiclass text classification with the support vector machines*. Technical Report AIM-2001-026. Artificial Intelligence Laboratory, Massachusetts Institute of Technology.

[26] C. G. M. Snoek, M. Worring, J. C. van Gemert, J.-M. Geusebroek, and A. W. M. Smeulders. 2006. The challenge problem for automated detection of 101 semantic concepts in multimedia. In *Proceedings of the 14th ACM International Conference on Multimedia*. Santa Barbara, CA, 421–430.

[27] H. Soleimani and D. J. Miller. 2016. Semi-supervised Multi-Label Topic Models for Document Classification and Sentence Labeling. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*. Indianapolis, IN, 105–114.

[28] A. N. Srivastava and B. Zane-Ulman. 2005. Discovering recurring anomalies in text reports regarding complex space systems. In *Proceedings of the 2005 IEEE Aerospace Conference*. Big Sky, MT.

[29] L. Sun, S. Ji, and J. Ye. 2013. *Multi-label Dimensionality Reduction*. Chapman and Hall/CRC, Boca Ration, FL.

[30] C.-Z. Tang and M.-L. Zhang. 2017. Confidence-rated discriminative partial label learning. In *Proceedings of the 31st AAAI Conference on Artificial Intelligence*. San Francisco, CA, 2611–2617.

[31] H. Wang, C. Ding, and H. Huang. 2010. Multi-label linear discriminant analysis. In *Lecture Notes in Computer Science 6316*, K. Daniilidis, P. Maragos, and N. Paragios (Eds.). Springer, Berlin, 126–139.

[32] F. Yu and M.-L. Zhang. 2017. Maximum margin partial label learning. *Machine Learning* 106, 4 (2017), 573–593.

[33] K. Yu, S. Yu, and V. Tresp. 2005. Multi-label informed latent semantic indexing. In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. Salvador, Brazil, 258–265.

[34] Z. Zeng, S. Xiao, K. Jia, T.-H. Chan, S. Gao, D. Xu, and Y. Ma. 2013. Learning by associating ambiguously labeled images. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. Portland, OR, 708–715.

[35] M.-L. Zhang and F. Yu. 2015. Solving the partial label learning problem: An instance-based approach. In *Proceedings of the 24th International Joint Conference on Artificial Intelligence*. Buenos Aires, Argentina, 4048–4054.

[36] M.-L. Zhang, F. Yu, and C.-Z. Tang. 2017. Disambiguation-free partial label learning. *IEEE Transactions on Knowledge and Data Engineering* 29, 10 (2017), 2155–2167.

[37] M.-L. Zhang, B.-B. Zhou, and X.-Y. Liu. 2016. Partial label learning via feature-aware disambiguation. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. San Francisco, CA, 1335–1344.

[38] Y. Zhang and Z.-H. Zhou. 2010, Article 14. Multi-label dimensionality reduction via dependency maximization. *ACM Transactions on Knowledge Discovery from Data* 4, 3 (2010, Article 14).

[39] D. Zhou, Z. Zhang, M.-L. Zhang, and Y. He. 2018. Weakly supervised POS tagging without disambiguation. *ACM Transactions on Asian and Low-Resource Language Information Processing* 17, 4 (2018), Article 35.

[40] Z.-H. Zhou. 2018. A brief introduction to weakly supervised learning. *National Science Review* 5, 1 (2018), 44–53.

[41] Z.-H. Zhou and M.-L. Zhang. 2017. Multi-label learning. In *Encyclopedia of Machine Learning and Data Mining*, C. Sammut and G. I. Webb (Eds.). Springer, Berlin, 875–881.

[42] X. Zhu and A. B. Goldberg. 2009. Introduction to semi-supervised learning. In *Synthesis Lectures to Artificial Intelligence and Machine Learning*, R. J. Brachman and T. G. Dietterich (Eds.). Morgan & Claypool Publishers, San Francisco, CA, 1–130.