Submodular Feature Selection for Partial Label Learning

Wei-Xuan Bao
[†]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
baowx@seu.edu.cn

Jun-Yi Hang
[†]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
hangjy@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 induces a multi-class classifier from training examples each associated with a candidate label set where the ground-truth label is concealed. Feature selection improves the generalization ability of learning system via selecting essential features for classification from the original feature set, while the task of partial label feature selection is challenging due to ambiguous labeling information. In this paper, the first attempt towards partial label feature selection is investigated via mutual-information-based dependency maximization. Specifically, the proposed approach SAUTE iteratively maximizes the dependency between selected features and labeling information, where the value of mutual information is estimated from confidence-based latent variable inference. In each iteration, the near-optimal features are selected greedily according to properties of submodular mutual information function, while the density of latent label variable is inferred with the help of updated labeling confidences over candidate labels by resorting to kNN aggregation in the induced lower-dimensional feature space. Extensive experiments over synthetic as well as real-world partial label data sets show that the generalization ability of well-established partial label learning algorithms can be significantly improved after coupling with the proposed feature selection approach.

CCS CONCEPTS

• **Computing methodologies** → **Machine learning algorithms**; *Learning paradigms.*

KEYWORDS

Partial Label Learning, Feature Selection, Submodular Function, Mutual Information

ACM Reference Format:
Wei-Xuan Bao, Jun-Yi Hang, and Min-Ling Zhang. 2022. Submodular Feature Selection for Partial Label Learning. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '22), August 14–18, 2022, Washington, DC, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3534678.3539292

*Corresponding author

KDD '22, August 14–18, 2022, Washington, DC, USA

ACM ISBN 978-1-4503-9385-0/22/08…$15.00
https://doi.org/10.1145/3534678.3539292

1 INTRODUCTION

As an emerging weakly-supervised learning framework, partial label (PL) learning aims to learn a multi-class classifier from ambiguous examples where each instance is associated with a set of candidate labels, among which only one is valid [10, 34, 60]. Owing to the ability of directly dealing with inaccurate supervision information [63], partial label learning has been successfully applied in many real-world application domains where collecting accurately labeled data is difficult and costly, such as web mining [23], multimedia content analysis [8, 57], ecoinformatics [5, 50], natural language processing [62], etc.

Although learning from ambiguously labeled examples greatly reduces the cost of data annotation, the generalization performance of partial label classification model is usually less satisfactory due to the limited supervision information retrieved from training set. Endowed with the strength of improving the generalization ability of learning system, dimensionality reduction mechanisms are expected to be incorporated into partial label learning. Dimensionality reduction can be generally divided into two categories: feature transformation and feature selection. Contrasting to feature transformation which maps the original high-dimensional feature vector into a meaningful representation in the induced lower-dimensional feature space [15, 16, 29, 54], feature selection performs dimensionality reduction via identifying the most informative feature subset of the observed data, which is capable of removing irrelevant and redundant features, increasing learning accuracy and enhancing learning comprehensibility [7, 19, 26]. To the best of our knowledge, DELIN [53, 58] and CENDA [2] are the only two available feature-transformation-based partial label dimensionality reduction approaches which induce the lower-dimensional feature space by adapting the *Linear Discriminant Analysis* (LDA) technique and the *Hilbert-Schmidt Independence Criterion* (HSIC) respectively, while the problem of selecting the most informative feature subset from partial label examples has not been well investigated.

In this paper, we propose a novel partial label feature selection method named SAUTE, i.e. *SubmodulAr featUre selecTion for partial labEl learning*. SAUTE performs feature selection via maximizing the dependency between selected feature variables and the latent label variable, which is evaluated by mutual information. Since the ground-truth label is not accessible during the learning

procedure, the density of latent label variable which is essential for the computation of mutual information is estimated with the help of iteratively-updated labeling confidences over candidate labels. In each iteration, superior features are selected by a greedy scheme according to the properties of submodular function, while the density of latent label variable is further estimated from updated labeling confidences by resorting to kNN aggregation in the lower-dimensional space induced by selected features. Comprehensive experiments over synthetic and real-world partial label data sets show that SAUTE serves as an effective feature selection approach to improve the generalization ability of well-established partial label learning algorithms.

The rest of this paper is organized as follows. Section 2 briefly reviews related works on partial label learning. Section 3 presents technical details of the proposed SAUTE approach. Section 4 reports experimental results over a broad range of partial label data sets. Finally, section 5 concludes this paper.

2 RELATED WORKS

Partial label learning induces a multi-class classifier from ambiguously labeled training examples each associated with a candidate label set, where the ground-truth label is concealed. To learn from partial label examples, most existing works adopt the strategy of candidate label disambiguation to reveal the ground-truth labeling information. Identification-based disambiguation treats the ground-truth label as latent variable and utilizes iterative optimization procedure to estimate the value of latent variable, where the optimization objective can be instantiated with different methods such as maximum likelihood criterion [24, 30, 31] or maximum margin criterion [6, 33, 56]. Averaging-based disambiguation treats all candidate labels equally and yields the final prediction via modifying their modeling outputs according to different averaging strategies, such as distinguishing the averaged modeling outputs from candidate labels between the modeling outputs from non-candidate labels for discriminative models [10, 44, 48] , or aggregating the votes among candidate labels of the unseen instance's neighboring examples for instance-based models [17, 22, 59].

As the fundamental approach to alleviating the issue of *curse of dimensionality*, dimensionality reduction [20, 38, 40, 49] has been studied extensively and is expected to significantly improve the generalization ability of the learning system. A number of advanced feature-transformation-based and feature-selection-based dimensionality reduction techniques have been introduced into weakly-supervised learning frameworks such as semi-supervised learning [41, 51], multi-instance learning [47, 52] and multi-label learning [45, 46, 61] to improve their less satisfactory generalization performance caused by limited supervision information retrieved from training set. Nevertheless, for partial label learning, most existing works focus on classification model induction by disambiguating the candidate label set while the task of manipulating the feature space by dimensionality reduction has been rarely investigated.

To the best of our knowledge, there are only two available feature-transformation-based partial label dimensionality reduction methods, namely DELIN [53, 58] and CENDA [2], while the application of feature selection [7, 19, 26] which not only facilitates removing irrelevance and redundancy in the feature space, but also brings about the advantages of interpretability and efficiency, has not been well studied in partial label learning framework. DELIN utilizes the LDA technique to maximize the inter-class separability in the projected feature space, whose dimensionality is upper-bounded by the number of class labels due to the intrinsic properties of LDA. CENDA adapts HSIC to assist maximizing the dependence between the projected feature information and the confidence-based labeling information. The above methods both assume the existence of a meaningful and computable distance metric in the input space, which brings extra bias to the learning procedure and might lead to suboptimal performance with inappropriate metric assumption.

3 THE PROPOSED APPROACH

Let $\mathcal{X} = \mathbb{R}^d$ and $\mathcal{L} = \{l_1, l_2, ..., l_q\}$ denote the d-dimensional instance space and the label space with q class labels respectively. Given the partial label training set $\mathcal{D} = \{(\boldsymbol{x}_i, S_i) | 1 \leqslant i \leqslant m\}$, where $\boldsymbol{x}_i \in \mathcal{X}$ is a d-dimensional feature vector $(x_{i_1}, x_{i_2}, ..., x_{i_d})^\top$ and $S_i \subseteq \mathcal{L}$ is the candidate label set associated with $\boldsymbol{x}_i$, partial label learning aims to derive a multi-class classifier $h : \mathcal{X} \to \mathcal{L}$ from the training set $\mathcal{D}$.

Let $F = \{f_1, ..., f_d\}$ denote the original feature set and latent variable c denote the unknown ground-truth label of the instance. The task of partial label feature selection is trying to select a subset $A(|A| = d', d' \ll d)$ from original features, i.e., $A \subseteq F$, which is recognized as the essential features of the instances. These essential features commonly have the maximal statistical dependency on the target class c [36]. Therefore, SAUTE performs feature selection via maximizing the dependency between selected features A and labeling information represented by random variable c, which is evaluated by mutual information in this paper, as mutual information is widely employed to define the dependency of random variables [4, 13]. Besides, maximizing the mutual information $I(A; c)$ also guarantees minimizing the lower bound of the misclassification probability of classifier according to Fano's inequality [11]. To tackle ambiguous labeling information, SAUTE operates in an iterative manner by alternating between mutual-information-based dependency maximization and density estimation of latent label variable. The two-stage alternating procedure is fulfilled by constructing labeling confidence matrix $\mathbf{Y} = [\mathbf{Y}(i, j)]_{m \times q}$ where each element $\mathbf{Y}(i, j)$ represents the estimated confidence of l_j being the ground-truth label for x_i. The matrix is initialized as Eq.(1) and the constraints $\sum_{j=1}^{q} \mathbf{Y}(i, j) = 1 (1 \leqslant i \leqslant m)$ hold for each iteration of SAUTE.

$$\forall\, 1 \leqslant i \leqslant m,\ 1 \leqslant j \leqslant q :\ \mathbf{Y}(i, j) = \begin{cases} \frac{1}{|S_i|}, & \text{if } l_j \in S_i \\ 0, & \text{otherwise} \end{cases} \tag{1}$$

In order to obtain a compact set of d' superior features, we expect the selected features have the maximal dependency on the concealed labeling information. For the stage of mutual-information-based dependency maximization, we formulate the objective function as:

$$A^* = \underset{A \subseteq F, |A| = d'}{\arg\max}\ g(A) = \underset{A \subseteq F, |A| = d'}{\arg\max}\ I(A; c) \tag{2}$$

The above problem is NP-Hard in spite of its simple expression [37]. It is difficult and costly to search the best d' features exhaustively. Nevertheless, the optimization goal $g(A) = I(A; c)$ is

a non-decreasing, non-negative submodular function under weak conditional independence assumption [27] with $g(\phi) = 0$ by definition. One of the most popular consequences of submodularity is that the maximum value of a non-negative and monotone submodular function can be effectively approximated with a tailored greedy algorithm [32, 55]. Therefore we can obtain a near-optimal subset of original features, i.e., the solution of Eq.(2), with theoretical performance guarantees via a greedy incremental scheme according to the properties of submodular function. In this scheme, supposing that we already have the feature subset $A_{p-1}(1 \leqslant p \leqslant d')$ with $p-1$ selected features which is initialized as $A_0 = \phi$, the pth feature is selected from $F \setminus A_{p-1}$ according to Eq.(3):

$$f_p^* = \underset{f \in F \setminus A_{p-1}}{\arg\max}\ I(A_{p-1} \cup \{f\}; c) \tag{3}$$

The final selected feature subset A_{greedy} satisfies the theoretical performance guarantee [32] that:

$$g(A_{\text{greedy}}) \geqslant (1 - \frac{1}{e}) \max_{|A|=d'} g(A) \tag{4}$$

In each greedy step, the computation of mutual information $I(A_{p-1} \cup \{f\}; c)$ involves the estimation of multivariate density $p(f_{s_1}, f_{s_2}, ..., f_{s_{p-1}}, f)$ and $p(f_{s_1}, f_{s_2}, ..., f_{s_{p-1}}, f, c)$. Nevertheless, in high-dimensional space the number of samples is usually insufficient for accurate multivariate density estimation. Moreover, computing the inverse of the high-dimensional covariance matrix which is needed for density estimation is time-consuming and usually an ill-posed problem. In order to avoid the problems mentioned above, we further assume that features are independent. Then we obtain the modified greedy policy for the pth $(1 \leqslant p \leqslant d')$ step as:

$$\begin{aligned} f_p^* &= \underset{f \in F \setminus A_{p-1}}{\arg\max}\ I(A_{p-1} \cup \{f\}; c) \\ &= \underset{f \in F \setminus A_{p-1}}{\arg\max}\ \big(H(A_{p-1} \cup \{f\}) - H(A_{p-1} \cup \{f\}|c)\big) \\ &\overset{①}{=} \underset{f \in F \setminus A_{p-1}}{\arg\max}\ \big(\big(H(A_{p-1}) + H(f)\big) - \big(H(A_{p-1}|c) + H(f|c)\big)\big) \\ &\overset{②}{=} \underset{f \in F \setminus A_{p-1}}{\arg\max}\ (H(f) - H(f|c)) \\ &= \underset{f \in F \setminus A_{p-1}}{\arg\max}\ I(f; c) \end{aligned} \tag{5}$$

where $H(\cdot)$ denotes the entropy of random variable. Here, equality ① is derived from the independence assumption. Furthermore, equality ② is derived from the fact that $H(A_{p-1})$ and $H(A_{p-1}|c)$ are constants for the pth step.

The above derivation reduces the computational complexity from calculating multivariate mutual information to calculating bivariate mutual information so as to improve the calculation accuracy and efficiency. Eq.(5) indicates that the criterion of dependency maximization is equivalent to the criterion of relevance maximization given the independence assumption, i.e., the scheme only needs to select the feature that has the maximal relevance with labeling information in each greedy step to maximize the dependency between eventually selected features and labeling information.

Nevertheless, features generally are not independent of each other in machine learning tasks. The above greedy policy effectively eliminates irrelevant features while redundant information between features is not well handled. Therefore, we attempt to make up for deficiencies of the independence assumption and revise the greedy policy as:

$$f_p^* = \underset{f \in F \setminus A_{p-1}}{\arg\max} \left(I(f; c) - \frac{1}{|S|} \sum_{f_i \in A_{p-1}} I(f; f_i) \right) \tag{6}$$

The second term in parentheses indicates that the newly selected feature f_p in each step should have minor relevance with features already selected in A_{p-1}, which facilitates removing redundant information in the induced feature space. Considering the fact that $H(c)$ is a constant, we further simplify Eq.(6) as:

$$f_p^* = \underset{f \in F \setminus A_{p-1}}{\arg\max} \left(-H(c|f) - \frac{1}{|S|} \sum_{f_i \in A_{p-1}} I(f; f_i) \right) \tag{7}$$

Implementation Issues. For partial label examples, it is infeasible to directly calculate the value of entropy corresponding to latent variable c due to the concealed ground-truth label. In this paper, we make the first attempt to estimate conditional entropy $H(c|f)$ in partial label learning framework.

For each partial label example $(\boldsymbol{x}_i, S_i)(|S_i| = n_i)$, if $\mathrm{Y}(i, j) \geqslant \frac{1}{n_i}(1 \leqslant j \leqslant q)$, $\boldsymbol{x}_i$ will be put into the set $\mathcal{D}_j$. In order to calculate $H(c|f)(\forall f \in F)$, we assume that class-conditional probability $p(f|l) \sim N(\mu_l^f, {\sigma_l^f}^2)$ on $\mathcal{D}_l(l \in \mathcal{L})$ where μ_l^f and σ_l^f denote the derived mean value and standard derivation respectively corresponding to feature f. Then $p(l|f)$ can be estimated by:

$$\hat{p}(l|f) = \frac{p(f|l) \cdot p(l)}{\sum_{u \in \mathcal{L}} p(f|u) \cdot p(u)} \tag{8}$$

where $p(u) = \frac{\sum_{i=1}^{m} \mathrm{Y}(i,u)}{m}(u \in \mathcal{L})$.

The class has discrete values while the input features are usually continuous variables. As a result, conditional entropy $H(c|f)$ is defined by:

$$H(c|f) = -\int_{\mathcal{X}_f} p(f) \sum_{l=1}^{q} p(l|f) \log p(l|f)\, df \tag{9}$$

We replace the integration with a summation of m training samples and suppose each sample has the same probability [28], then $H(c|f)$ is estimated as:

$$\hat{H}(c|f) = -\sum_{j=1}^{m} \frac{1}{m} \sum_{l=1}^{q} \hat{p}(l|x_j^f) \log \hat{p}(l|x_j^f) \tag{10}$$

where x_j^f is the value of the jth training sample corresponding to feature f.

For terms $I(f; f_i)(f_i \in A_{p-1})$ in Eq.(7), in order to avoid complicated integrals, we simply discretize each feature variable into five intervals according to Eq.(11) to estimate the value of mutual information between features [36]:

$$\hat{x}_i^f = \begin{cases} -2, & \text{if } x_i^f \leqslant \mu_f - 2 \cdot \sigma_f \\ -1, & \text{if } \mu_f - 2 \cdot \sigma_f < x_i^f \leqslant \mu_f - \sigma_f \\ 0, & \text{if } \mu_f - \sigma_f < x_i^f \leqslant \mu_f + \sigma_f \\ 1, & \text{if } \mu_f + \sigma_f < x_i^f \leqslant \mu_f + 2 \cdot \sigma_f \\ 2, & \text{if } x_i^f > \mu_f + 2 \cdot \sigma_f \end{cases} \tag{11}$$

Table 1: The pseudo-code of SAUTE.

Inputs:

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

d' : the cardinality of selected feature subset

α : the learning rate in Eq.(13)

k : the number of exploited nearest neighbors

Outputs:

$\mathcal{D}'$: the induced lower-dimensional PL training set $\{(\boldsymbol{x}'_i, S_i) \mid 1 \leqslant i \leqslant m\}$

Process:

1: Initialize the $m \times q$ labeling confidence matrix $\mathbf{Y}$ according to Eq.(1);
2: **repeat**
3: Initialize $A_0 = \phi$;
4: **for** p=1 to d' **do**
5: Calculate $\hat{H}(c|f)$ for $\forall f \in F \setminus A_{p-1}$ according to Eq.(10);
6: Calculate $\sum_{f_i \in A_{p-1}} I(f; f_i)$ for $\forall f \in F \setminus A_{p-1}$ by discretization;
7: Find f_p^* according to Eq.(7);
8: $A_p = A_{p-1} \cup \{f_p^*\}$;
9: **end for**
10: Construct the lower-dimensional PL training set $\mathcal{D}' = \{(\boldsymbol{x}'_i, S_i) | 1 \leqslant i \leqslant m\}$ where $\boldsymbol{x}'_i$ is derived from $\boldsymbol{x}_i$ in accordance with the selected feature subset;
11: Identify the k nearest neighbors $\mathcal{N}(x'_i)$ for $\forall \boldsymbol{x}'_i (1 \leqslant i \leqslant m)$;
12: Calculate the learning matrix $\mathbf{L}$ according to Eq.(12);
13: Calculate the intermediate matrix $\mathbf{Y}'$ according to Eq.(13);
14: Calculate the updated labeling confidence matrix $\mathbf{Y}_{\text{new}}$ according to Eq.(14);
15: $\mathbf{Y} = \mathbf{Y}_{\text{new}}$;
16: **until** convergence
17: Construct the lower-dimensional PL training set $\mathcal{D}'$ according to selected feature subset A_p;
18: Return $\mathcal{D}'$

where μ_f and σ_f respectively denote the mean value and standard deviation of each feature $f \in F$ derived from training set $\mathcal{D}$.

After determining the selected feature subset, we construct a lower-dimensional PL training set $\mathcal{D}' = \{(\boldsymbol{x}'_i, S_i) | 1 \leqslant i \leqslant m\}$ where $\boldsymbol{x}'_i$ is derived from $\boldsymbol{x}_i$ in accordance with selected features. Thereafter, the density estimation of latent label variable is refined via updating the labeling confidence matrix by resorting to kNN aggregation in the lower-dimensional feature space.

For each instance $\boldsymbol{x}'_i \in \mathbb{R}^{d'}$, the probability of each candidate label being its ground-truth label is re-estimated via exploiting labeling information of its k nearest neighbors. The learning matrix $\mathbf{L} = [\mathbf{L}(i, j))]_{m \times q}$ is defined as:

$$\mathbf{L}(i, j) = \sum_{\boldsymbol{x}'_{i_a} \in \mathcal{N}(\boldsymbol{x}'_i)} \mathbf{Y}(i_a, j) \times \omega_a \tag{12}$$

where $\mathcal{N}(\boldsymbol{x}'_i)$ denotes the k nearest neighbors of $\boldsymbol{x}'_i$ and the voting weight is set as $\omega_a = k - a + 1 (1 \leqslant a \leqslant k)$ for the ath nearest neighbor [22, 59].

Afterwards, the labeling confidence matrix is updated by:

$$\mathbf{Y}' = (1 - \alpha) \cdot \mathbf{Y} + \alpha \cdot \mathbf{L} \tag{13}$$

where the learning rate is set as $\alpha = 0.6 (0 < \alpha < 1)$ in this paper.

In order to ensure that the confidences of labels outside the candidate label set are zero and the constraints $\sum_{j=1}^{q} \mathbf{Y}(i, j) = 1 (1 \leqslant i \leqslant m)$ are satisfied, we make further adjustments to matrix $\mathbf{Y}'$ and obtain $\mathbf{Y}_{\text{new}}$ by:

$$\mathbf{Y}_{\text{new}}(i, j) = \begin{cases} \frac{\mathbf{Y}'(i,j)}{\sum_{b \in S_i} \mathbf{Y}'(i,b)} & \text{if } j \in S_i \\ 0 & \text{otherwise} \end{cases} \tag{14}$$

Table 1 summarizes the complete procedure of SAUTE. Firstly, the labeling confidence matrix is initialized (step 1) based on the assignment of the training data set. After that, an iterative procedure alternating between mutual-information-based dependency maximization (step 3-9) and density estimation of latent label variable (step 10-15) is conducted. The iterative procedure terminates if the selected feature subset does not change or the maximum number of iteration is reached. [1] Finally, the lower-dimensional PL training set is constructed according to the selected feature subset.

4 EXPERIMENTS

4.1 Experimental Setup

In this section, SAUTE is coupled with state-of-the-art partial label learning algorithms to evaluate the effectiveness of the proposed partial label feature selection approach. Given the partial label learning algorithm $\mathcal{A}$, its coupling version with SAUTE is denoted as $\mathcal{A}$-SAUTE. The performance of $\mathcal{A}$-SAUTE is compared against that of $\mathcal{A}$ to verify the effectiveness of the proposed partial label feature selection approach in improving the generalization ability of the learning system.

In this paper, we utilize five well-established partial label learning algorithms with suggested parameter configuration in respective literatures to instantiate $\mathcal{A}$:

- PL-KNN [22]: An averaging-based partial label learning approach which makes prediction on unseen instance by employing weighted kNN voting strategy [suggested configuration: k=10].
- PL-SVM [33]: An identification-based partial label learning approach which learns the predictive model by maximizing the classification margin over candidate label set and non-candidate label set [suggested configuration: regularization parameter pool with $\{10^{-3}, \ldots, 10^{3}\}$].
- PL-ECOC [60]: A transformation-based partial label learning approach which learns the predictive model by decomposing the PL learning problem into a group of binary learning problems via adapting the error-correcting output codes (ECOC) techniques [suggested configuration: ECOC coding length $\lceil 10 \cdot \log_2(q) \rceil$].
- IPAL [59]: An instance-based partial label learning approach which learns the predictive model by adapting label propagation for graph-based disambiguation [suggested configuration: balancing parameter $\alpha = 0.95$].
- SURE [14]: A self-training partial label learning approach which learns the desired model and performs pseudo-labeling jointly by solving a tailored convex-concave optimization problem [suggested configuration: regularization parameters $\lambda = 0.3, \beta = 0.05$].

As is shown in Table 1, the parameters α and k are set to be 0.6 and 8 respectively. The cardinality of the selected feature subset is

[1] In this paper, the maximum number of iterations is set to be 20 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* [42]
Corel16k-s1	1,075	417	87	$r = 1, 2, 3$	*matching words and pictures* [3]
amazon	1,500	1,326	50	$r = 1, 2, 3$	*authorship identification* [12]
DeliciousMIL	1,409	1,389	20	$r = 1, 2, 3$	*sentence labeling* [43]
bookmark	2,500	1,413	57	$r = 1, 2, 3$	*automatic tag suggestion* [25]
sports	9,120	1,738	19	$r = 1, 2, 3$	*human activity recognition* [1]

Table 3: Classification accuracy (mean±std) of each comparing algorithm on controlled synthetic data sets ($r \in \{1, 2, 3\}$). Given partial label learning algorithm $\mathcal{A} \in \{$PL-KNN, PL-SVM, PL-ECOC, IPAL, SURE$\}$, the performance of $\mathcal{A}$-SAUTE is compared against that of $\mathcal{A}$ where the best performance on each data set is shown in boldface.

Comparing Algorithms	Data Set					
	mediamill	**Corel16k-s1**	**amazon**	**DeliciousMIL**	**bookmark**	**sports**
	$r = 1$ (*one false positive label*)					
PL-KNN	**0.637±0.024**	0.016±0.017	0.025±0.025	0.033±0.039	0.170±0.026	0.288±0.015
PL-KNN-SAUTE	0.630±0.023	**0.108±0.053**	**0.044±0.019**	**0.156±0.022**	**0.346±0.017**	**0.409±0.018**
PL-SVM	0.485±0.049	0.100±0.016	0.105±0.105	0.036±0.015	0.280±0.023	**0.673±0.023**
PL-SVM-SAUTE	**0.487±0.042**	**0.142±0.015**	**0.571±0.038**	**0.195±0.041**	**0.417±0.031**	0.500±0.017
PL-ECOC	**0.604±0.042**	0.192±0.088	0.069±0.069	0.065±0.040	0.330±0.041	0.680±0.030
PL-ECOC-SAUTE	0.558±0.044	**0.199±0.079**	**0.354±0.050**	**0.209±0.022**	**0.414±0.029**	**0.703±0.023**
IPAL	0.642±0.027	0.154±0.054	0.105±0.043	0.062±0.020	0.309±0.040	0.887±0.010
IPAL-SAUTE	**0.645±0.029**	**0.155±0.064**	**0.452±0.033**	**0.263±0.026**	**0.445±0.031**	**0.924±0.007**
SURE	**0.691±0.032**	0.185±0.061	0.153±0.072	0.116±0.031	0.388±0.029	0.755±0.013
SURE-SAUTE	0.668±0.032	**0.187±0.064**	**0.649±0.037**	**0.290±0.032**	**0.478±0.028**	**0.891±0.014**
	$r = 2$ (*two false positive labels*)					
PL-KNN	0.622±0.023	0.021±0.014	0.021±0.009	0.027±0.014	0.162±0.012	0.290±0.015
PL-KNN-SAUTE	**0.625±0.019**	**0.094±0.53**	**0.040±0.009**	**0.127±0.035**	**0.338±0.018**	**0.485±0.017**
PL-SVM	0.488±0.038	0.070±0.034	0.081±0.019	0.031±0.020	0.261±0.021	**0.638±0.011**
PL-SVM-SAUTE	**0.489±0.024**	**0.124±0.052**	**0.435±0.030**	**0.200±0.039**	**0.402±0.027**	0.560±0.016
PL-ECOC	**0.500±0.037**	0.156±0.073	0.043±0.011	0.040±0.026	0.288±0.038	0.603±0.033
PL-ECOC-SAUTE	0.493±0.043	**0.171±0.069**	**0.211±0.036**	**0.187±0.033**	**0.400±0.025**	**0.687±0.026**
IPAL	0.585±0.029	0.141±0.050	0.088±0.047	0.053±0.034	0.304±0.017	**0.874±0.008**
IPAL-SAUTE	**0.586±0.029**	**0.143±0.060**	**0.425±0.036**	**0.227±0.041**	**0.436±0.016**	0.939±0.005
SURE	**0.667±0.026**	0.158±0.012	0.102±0.043	0.115±0.034	0.374±0.018	0.711±0.011
SURE-SAUTE	**0.667±0.026**	**0.184±0.016**	**0.605±0.021**	**0.261±0.035**	**0.474±0.017**	**0.911±0.009**
	$r = 3$ (*three false positive labels*)					
PL-KNN	**0.598±0.017**	0.018±0.015	0.020±0.008	0.043±0.022	0.140±0.026	0.292±0.021
PL-KNN-SAUTE	**0.598±0.021**	**0.095±0.051**	**0.044±0.013**	**0.083±0.024**	**0.292±0.033**	**0.427±0.022**
PL-SVM	0.479±0.046	0.065±0.059	0.063±0.015	0.029±0.017	0.252±0.030	**0.601±0.022**
PL-SVM-SAUTE	**0.504±0.042**	**0.138±0.051**	**0.317±0.059**	**0.182±0.034**	**0.369±0.033**	0.555±0.019
PL-ECOC	**0.095±0.014**	0.126±0.086	0.031±0.012	0.063±0.035	0.200±0.044	0.503±0.039
PL-ECOC-SAUTE	0.087±0.025	**0.160±0.067**	**0.114±0.030**	**0.149±0.043**	**0.353±0.027**	**0.535±0.015**
IPAL	0.511±0.026	0.139±0.061	0.084±0.043	0.044±0.041	0.293±0.041	0.863±0.013
IPAL-SAUTE	**0.513±0.034**	**0.148±0.048**	**0.387±0.061**	**0.228±0.019**	**0.413±0.033**	**0.927±0.010**
SURE	0.649±0.021	0.163±0.011	0.073±0.048	0.116±0.048	0.370±0.040	0.671±0.010
SURE-SAUTE	**0.651±0.021**	**0.197±0.013**	**0.559±0.047**	**0.274±0.029**	**0.461±0.045**	**0.873±0.011**

set to be 15% of the number of original features for each data set, i.e., $d' = \lceil 15\% \cdot d \rceil$.

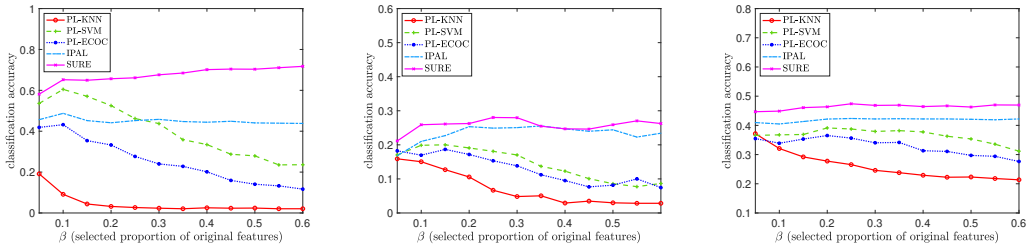
In the rest of this section, comparative studies are conducted on both synthetic and real-world partial label data sets with ten-fold cross-validation where detailed experimental results of each data set are presented subsequently.

4.2 Synthetic Data Sets

Following the conventional experimental protocol in partial label learning [8–10, 17, 30, 56, 60], we generate synthetic partial label data sets from multi-class data sets with controlling parameter r which specifies the number of false positive labels in the candidate label set (i.e., $|S_i| = r + 1$). Given a multi-class example $(\boldsymbol{x}_i, y_i)$, r false positive class labels $\Delta_r \subseteq \mathcal{Y} \setminus \{y_i\} (|\Delta_r| = r)$ are randomly selected to form the candidate label set along with the ground-truth label y_i, i.e., $S_i = \Delta_r \cup \{y_i\}$, and the partial label example $(\boldsymbol{x}_i, S_i)$ is obtained consequently. Table 2 summarizes characteristics of the synthetic data sets ($r \in \{1, 2, 3\}$) which are roughly ordered according to the dimensionality of each data set. [2]

Table 3 reports detailed experimental results of each comparing algorithm over various synthetic data sets. Given partial label learning algorithms $\mathcal{A} \in \{$PL-KNN, PL-SVM, PL-ECOC, IPAL, SURE$\}$, $\mathcal{A}$-SAUTE is compared against $\mathcal{A}$ where the best classification performance is shown in boldface. In addition, pairwise t-test at 0.05 significance level is conducted to show whether the performance difference between $\mathcal{A}$-SAUTE and $\mathcal{A}$ is significant, where the resulting win/tie/lose counts are reported in Table 4.

[2]Most data sets presented in Table 2 are derived from multi-label benchmark data sets [64] by retaining examples with only one relevant label.



(a) amazon($r = 1$) (b) delicious($r = 2$) (c) bookmark($r = 3$)

Figure 1: Trend of classification accuracy of $\mathcal{A}$-SAUTE ($\mathcal{A} \in$ {PL-KNN, PL-SVM, PL-ECOC, IPAL SURE}) where the number of selected features is set as $d' = \lceil \beta \cdot d \rceil$. The coefficient β increases from 0.05 to 0.6 with step-size 0.05 in (a) `amazon` ($r = 1$), (b) `delicious` ($r = 2$) and (c) `bookmark` ($r = 3$).

Table 4: Win/tie/loss counts (pairwise t-test at 0.05 significance level) between $\mathcal{A}$-SAUTE and $\mathcal{A}$ in terms of different number of false positive labels ($r = 1, 2, 3$).

	$\mathcal{A}$-SAUTE against $\mathcal{A}$				
	$\mathcal{A}$=PL-KNN	$\mathcal{A}$= PL-SVM	$\mathcal{A}$=PL-ECOC	$\mathcal{A}$=IPAL	$\mathcal{A}$=SURE
$r = 1$	5/1/0	4/1/1	3/2/1	4/2/0	4/2/0
$r = 2$	5/1/0	4/1/1	4/2/0	4/2/0	5/1/0
$r = 3$	5/1/0	4/1/1	4/2/0	4/2/0	5/1/0
In Total	**15/3/0**	**12/3/3**	**11/6/1**	**12/6/0**	**14/4/0**

Table 5: 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* [10]
Yahoo! News	22,991	163	219	1.91	*automatic face naming* [18]
FG-NET	1,002	262	78	7.48	*facial age estimation* [35]
Soccer Player	17,472	279	171	2.09	*automatic face naming* [57]
Mirflickr	2,780	1,536	14	2.76	*web image classification* [21]
Malagasy	5,303	384	44	8.35	*POS tagging* [62]

Table 6: Win/tie/loss statistics (pairwise t-test at 0.05 significance level) between $\mathcal{A}$-SAUTE and $\mathcal{A}$, $\mathcal{A}$-baselines on real-world data sets.

Data Set	$\mathcal{A}$-SAUTE against $\mathcal{A}$ and $\mathcal{A}$-baselines ($\mathcal{A}$ = PL-KNN)				$\mathcal{A}$-SAUTE against $\mathcal{A}$ and $\mathcal{A}$-baselines ($\mathcal{A}$ = PL-ECOC)			
	$\mathcal{A}$(Ori)	$\mathcal{A}$-RS	$\mathcal{A}$-MJE	$\mathcal{A}$-MR	$\mathcal{A}$(Ori)	$\mathcal{A}$-RS	$\mathcal{A}$-MJE	$\mathcal{A}$-MR
Lost	win	win	win	win	win	win	win	win
Yahoo! News	win	win	win	win	win	win	win	win
FG-NET	win	win	win	win	win	win	win	tie
Soccer Player	tie	win	win	win	win	win	win	win
Mirflickr	tie	win	win	win	win	win	win	win
Malagasy	win	win	win	win	tie	win	win	win
In Total	**4/2/0**	**6/0/0**	**6/0/0**	**6/0/0**	**5/1/0**	**6/0/0**	**6/0/0**	**5/1/0**

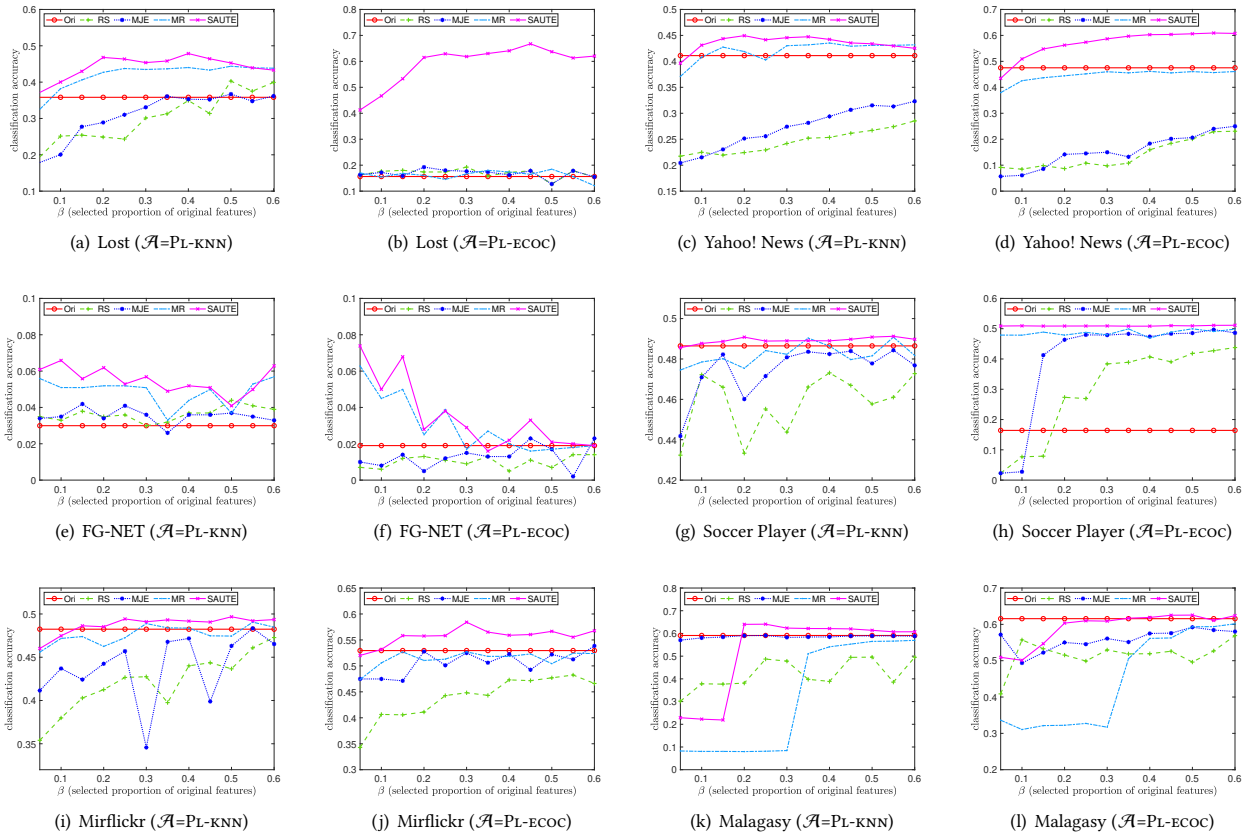
In order to explore the influence of parameter d' on the performance of the proposed algorithm SAUTE, we further conduct a series of experiments with $d' = \lceil \beta \cdot d \rceil$ where β varies from 0.05 to 0.6 with step-size 0.05. Owing to the limited length of the paper, only parts of experimental results are depicted in Fig. 1.

Based on the above experimental results over synthetic data sets, we can draw following conclusions:

- The performance improvement of $\mathcal{A}$-SAUTE against $\mathcal{A}$ is moderate on `mediamill` which corresponds to the smallest number of features (Table 3). On the three data sets with more than 1300 features and relatively small number of examples (i.e., `amazon`, `DeliciousMIL` and `bookmark`), $\mathcal{A}$-SAUTE achieves better performance than $\mathcal{A}$ in **all** 45 cases (Table 4), and the classification accuracy has been improved with SAUTE by more than **0.1** in **80%** cases. These results demonstrate that the benefits brought by SAUTE are even more pronounced under challenging circumstances of high dimensionality and insufficient training examples.
- As is shown in Fig. 1, the classification accuracy of each partial label learning algorithm coupled with SAUTE fluctuates moderately as the value of d' changes. The evaluation results do not monotonously increase or decrease with the number of selected features in all curves. There is no one single value of d' which can consistently lead to the best performance, although $d' = \lceil 0.15 \cdot d \rceil$ is a reasonable default setting in this paper. Further performance improvement could be achieved through fine-tuning the value of d' for different data sets and learning algorithms.

4.3 Real-World Data Sets

Table 5 summarizes characteristics of the real-world partial label data sets collected from different task domains, including `Lost` [10], `Soccer Player` [57] and `Yahoo! News` [18] for automatic face



(a) Lost ($\mathcal{A}$=PL-KNN) (b) Lost ($\mathcal{A}$=PL-ECOC) (c) Yahoo! News ($\mathcal{A}$=PL-KNN) (d) Yahoo! News ($\mathcal{A}$=PL-ECOC)

(e) FG-NET ($\mathcal{A}$=PL-KNN) (f) FG-NET ($\mathcal{A}$=PL-ECOC) (g) Soccer Player ($\mathcal{A}$=PL-KNN) (h) Soccer Player ($\mathcal{A}$=PL-ECOC)

(i) Mirflickr ($\mathcal{A}$=PL-KNN) (j) Mirflickr ($\mathcal{A}$=PL-ECOC) (k) Malagasy ($\mathcal{A}$=PL-KNN) (l) Malagasy ($\mathcal{A}$=PL-ECOC)

Figure 2: Classification accuracy of each base classifier $\mathcal{A} \in \{$PL-KNN, PL-ECOC$\}$ before (denoted by `Ori` in the legend) and after employing PL feature selection methods (including SAUTE, RS, MJE, MR). The number of selected features is set as $d' = \lceil \beta \cdot d \rceil$ where the coefficient β increases from 0.05 to 0.6 with step-size 0.05 in (a, b) `Lost` , (c, d) `Yahoo! News`, (e, f) `FG-NET`, (g, h) `Soccer Player`, (i, j) `Mirflickr` and (k, l) `Malagasy`.

naming from images or videos, `FG-NET` [35] for facial age estimation, `Mirflickr` [21] for web image classification and `Malagasy` [62] for part-of-speech (POS) tagging. [3] In the data set of *automatic face naming*, instances denote faces cropped from images or video frames while candidate labels are derived from names extracted from the associated captions or subtitles. In the data set of *facial age estimation*, instances denote human faces with landmarks while candidate labels are derived from ages denoted by crowdsourced labelers. In the data set of *web image classification*, instances denote web images while candidate labels are derived from annotations extracted from the web environment. In the data set of *POS tagging*, instances denote the target words with contextual features while candidate labels are derived from the POS tags that the target words may have.

In this subsection, two base classifiers ($\mathcal{A} \in \{$PL-KNN, PL-ECOC$\}$) are coupled with SAUTE and other three naive partial label feature selection approaches which are constructed as comparing algorithms:

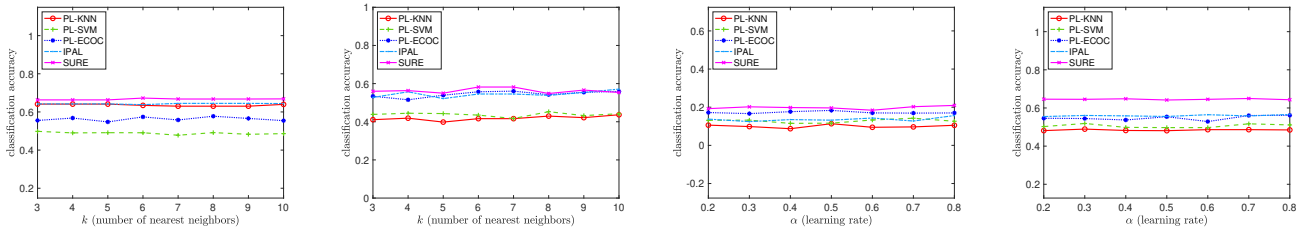
- **Random Selection** *(RS)*: Construct the feature subset A_{RS} by randomly selecting d' features from the original feature set.
- **Maximum Joint Entropy** *(MJE)*: Entropy is commonly used to measure the quantity of information [39]. In order to achieve the most informative feature subset, *MJE* constructs the feature subset A_{MJE} by solving the optimization problem $A_{\text{MJE}} = \arg\max_{A \subseteq F, |A|=d'} H(A)$.
- **Maximum Relevance** *(MR)*: Construct the feature subset A_{MR} by solving the optimization problem Eq.(2) with independence assumption, i.e., greedily select the near-optimal feature in each step according to Eq.(5).

Fig. 2 illustrates the predictive accuracy of each base classifier before (denoted by `Ori` in the legend of the figure) and after employing the proposed feature selection technique SAUTE and three baseline methods on each real-world data set. Furthermore, pairwise t-test at 0.05 significance level is conducted to show whether the performance differences between $\mathcal{A}$-SAUTE and $\mathcal{A}$, $\mathcal{A}$-baselines are significant. The resulting win/tie/loss statistics are reported in Table 6.

From the above experimental results on real-world data sets, we can observe that:

- As is shown in Fig. 2, the performance improvement of each base classifier can be achieved after being coupled with SAUTE through fine-tuning the value of d' for each data

[3] Data available at: http://palm.seu.edu.cn/zhangml/



(a) mediamill ($r = 1$) (b) Lost (c) Corel16k-s1 ($r = 2$) (d) Mirflickr

Figure 3: Trend of classification accuracy of $\mathcal{A}$-SAUTE ($\mathcal{A} \in$ {PL-KNN, PL-SVM, PL-ECOC, IPAL SURE}). The number of exploited nearest neighbors (i.e. k) increases from 3 to 10 with step-size 1 in (a) synthetic data set `mediamill` ($r = 1$) and (b) real-world data set `Lost`; the number of learning rate (i.e. α) increases from 0.2 to 0.8 with step-size 0.1 in (c) synthetic data set `Corel16k-s1` ($r = 2$) and (d) real-world data set `Mirflickr`.

set. It is worth mentioning that the classification accuracy of each base classifier has at least been doubled on `FG-NET`, which corresponds to the real-world data set with smallest number of examples but large average number of candidate labels. These impressive results indicate that the benefits brought by SAUTE would be more significant under challenging circumstances of insufficient training examples and high rate of false positive labels.

- Out of the 36 statistical comparisons (6 data sets $\times$ 3 baselines $\times$ 2 base classifiers), the performance of $\mathcal{A}$-SAUTE is significantly superior to that of $\mathcal{A}$-baselines in 35 cases (Table 6). These results indicate that mutual information is an appropriate evaluation indicator of dependency in partial label learning framework and the proposed partial label feature selection approach SAUTE could significantly improve the performance of base classifiers via effectively removing irrelevant and redundant features.

4.4 Sensitivity Analysis

As is shown in Table 1, d' serves as an essential parameter for SAUTE which determines the cardinality of the selected feature subset. The influence of parameter d' on the performance of SAUTE has been shown in Fig. 1 and Fig. 2. Overall, the proposed feature selection approach behaves smoothly as the value of d' changes within a certain range. The classification accuracy of partial label learning algorithms coupled with SAUTE could achieve further improvement by fine-tuning the value of d', although $d' = \lceil 0.15 \cdot d \rceil$ is a reasonable default setting in this paper.

Apart from d', the learning rate α and the number of exploited nearest neighbors k also serve as critical parameters for SAUTE. Fig. 3 illustrates how the predictive performance of each partial label learning algorithm coupled with SAUTE changes as α increases from 0.2 to 0.8 with an interval of 0.1 and k increases from 3 to 10 with an interval of 1 respectively. As is shown in Fig. 3, the performance of each partial label learning algorithm coupled with SAUTE is relatively stable as the value of α or k changes. Therefore, the value of α and k is fixed to be 0.6 and 8 respectively in this paper.

5 CONCLUSION

In this paper, we make the first attempt towards partial label feature selection problem. Accordingly, a novel approach named SAUTE is proposed which performs partial label feature selection by maximizing the mutual-information-based dependency between selected features and labeling information in an iterative manner. In each iteration, the near-optimal features are selected greedily according to properties of submodular function, while the density of latent label variable is estimated from updated labeling confidences over candidate labels by resorting to kNN aggregation in the induced lower-dimensional feature space. Comprehensive experiments over synthetic as well as real-world partial label data sets show that SAUTE is an effective partial label feature selection approach to improve the performance of state-of-the-art partial label learning algorithms. It is worth mentioning that the labeling confidence matrix $\mathbf{Y}$ derived from SAUTE may bring further improvement of predictive performance for specific partial label learning algorithms with proper utilization.

ACKNOWLEDGMENTS

The authors wish to thank the anonymous reviewers for their helpful comments and suggestions. This work was supported by the National Science Foundation of China (62176055). We thank the Big Data Center of Southeast University for providing the facility support on the numerical calculations in this paper.

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] W.-X. Bao, J.-Y. Hang, and M.-L. Zhang. 2021. Partial label dimensionality reduction via confidence-based dependence maximization. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*. Virtual Event, 46–54.
[3] K. Barnard, P. Duygulu, D. A. Forsyth, N. de Freitas, D. M. Blei, and M. I. Jordan. 2003. Matching words and pictures. *Journal of Machine Learning Research* 3 (2003), 1107–1135.
[4] L. Batina, B. Gierlichs, E. Prouff, M. Rivain, F.-X. Standaert, and N. Veyrat. 2011. Mutual information analysis: A comprehensive study. *Journal of Cryptology* 24, 2 (2011), 269–291.
[5] 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.
[6] J. Chai, I. W. Tsang, and W. Chen. 2020. Large margin partial label machine. *IEEE Transactions on Neural Networks and Learning Systems* 31, 7 (2020), 2594–2608.
[7] G. Chandrashekar and F. Sahin. 2014. A survey on feature selection methods. *Computers & Electrical Engineering* 40, 1 (2014), 16–28.
[8] 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.

[9] 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.

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

[11] T. M Cover. 1999. *Elements of Information Theory*. John Wiley & Sons.

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

[13] A. Dionisio, R. Menezes, and D. A. Mendes. 2004. Mutual information: A measure of dependency for nonlinear time series. *Physica A: Statistical Mechanics and its Applications* 344, 1-2 (2004), 326–329.

[14] L. Feng and B. An. 2019. Partial label learning with self-guided retraining. In *Proceedings of the 33th AAAI Conference on Artificial Intelligence. Honolulu, Hawaii, USA*. 3542–3549.

[15] K. Fukumizu, F. R. Bach, and M. I. Jordan. 2004. Dimensionality reduction for supervised learning with reproducing kernel hilbert spaces. *Journal of Machine Learning Research* 5 (2004), 73–99.

[16] A. Globerson and N. Tishby. 2003. Sufficient dimensionality reduction. *Journal of Machine Learning Research* 3 (2003), 1307–1331.

[17] 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.

[18] M. Guillaumin, J. Verbeek, and C. Schmid. 2010. Multiple instance metric learning from automatically labeled bags of faces. In *Proceedings of the 11th European Conference on Computer Vision*. Heraklion, Greece, 634–647.

[19] I. Guyon and A. Elisseeff. 2003. An introduction to variable and feature selection. *Journal of Machine Learning Research* 3, Mar (2003), 1157–1182.

[20] R. Hadsell, S. Chopra, and Y. LeCun. 2006. Dimensionality reduction by learning an invariant mapping. In *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Vol. 2. 1735–1742.

[21] 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.

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

[23] L. Jie and F. Orabona. 2010. Learning from candidate labeling sets. In *Advances in Neural Information Processing Systems 23*. Cambridge, MA, 1504–1512.

[24] R. Jin and Z. Ghahramani. 2003. Learning with multiple labels. In *Advances in Neural Information Processing Systems 15*. Cambridge, MA, 897–904.

[25] I. Katakis, G. Tsoumakas, and I. Vlahavas. 2008. Multilabel text classification for automated tag suggestion. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Antwerp, Belgium, 5.

[26] K. Kira and L. A. Rendell. 1992. A practical approach to feature selection. In *Proceedings of the 9th International Workshop on Machine Learning, Scotland, UK*. 249–256.

[27] A. Krause and C. Guestrin. 2005. Near-optimal nonmyopic value of information in graphical models. In *Proceedings of the 21st Conference in Uncertainty in Artificial Intelligence, Edinburgh, Scotland*. 324–331.

[28] N. Kwak and C.-H. Choi. 2002. Input feature selection by mutual information based on parzen window. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 24, 12 (2002), 1667–1671.

[29] T. Li and Y. Dou. 2021. Representation learning on textual network with personalized PageRank. *Science China Information Sciences* 64, 11 (2021), 1–10.

[30] L. Liu and T. Dietterich. 2012. A conditional multinomial mixture model for superset label learning. In *Advances in Neural Information Processing Systems 25*. Cambridge, MA, 557–565.

[31] J. Lv, M. Xu, L. Feng, G. Niu, X. Geng, and M. Sugiyama. 2020. Progressive identification of true labels for partial-label learning. In *Proceedings of the 37th International Conference on Machine Learning*. Virtual Conference, 6500–6510.

[32] G. L. Nemhauser, L. A. Wolsey, and M. L. Fisher. 1978. An analysis of approximations for maximizing submodular set functions—I. *Mathematical programming* 14, 1 (1978), 265–294.

[33] 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.

[34] P. Ni, S.-Y. Zhao, Z.-G. Dai, H. Chen, and C.-P. Li. 2021. Partial label learning via conditional-label-aware disambiguation. *Journal of Computer Science and Technology* 36, 3 (2021), 590–605.

[35] G. Panis and A. Lanitis. 2014. An overview of research activities in facial age estimation using the FG-NET aging database. In *Proceedings of the 13th European Conference on Computer Vision*. Zurich, Switzerlan, 737–750.

[36] H. Peng, F. Long, and C. Ding. 2005. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 27, 8 (2005), 1226–1238.

[37] C. Qian, Y. Yu, K. Tang, X. Yao, and Z.-H. Zhou. 2019. Maximizing submodular or monotone approximately submodular functions by multi-objective evolutionary algorithms. *Artificial Intelligence* 275 (2019), 279–294.

[38] M. L. Raymer, W. F. Punch, E. D. Goodman, L. A. Kuhn, and A. K. Jain. 2000. Dimensionality reduction using genetic algorithms. *IEEE Transactions on Evolutionary Computation* 4, 2 (2000), 164–171.

[39] A. Rényi. 1961. On measures of entropy and information. In *Proceedings of the 4th Berkeley Symposium on Mathematical Statistics and Probability*. Berkeley, California, 547–561.

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

[41] R. Sheikhpour, M. A. Sarram, S. Gharaghani, and M. A. Z. Chahooki. 2017. A survey on semi-supervised feature selection methods. *Pattern Recognition* 64 (2017), 141–158.

[42] 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.

[43] 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 Conference on Information and Knowledge Management*. Indianapolis, IN, 105–114.

[44] K. Sun, Z. Min, and J. Wang. 2019. PP-PLL: Probability propagation for partial label learning. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Würzburg, Germany, 123–137.

[45] L. Sun, S. Ji, and J. Ye. 2013. *Multi-label dimensionality reduction*. Boca Raton, Florida, CRC Press.

[46] Y.-P. Sun and M.-L. Zhang. 2021. Compositional metric learning for multi-label classification. *Frontiers of Computer Science* 15, 5 (2021), 1–12.

[47] Y.-Y. Sun, M. K. Ng, and Z.-H. Zhou. 2010. Multi-instance dimensionality reduction. In *24th AAAI Conference on Artificial Intelligence*. Atlanta, GA., 11–15.

[48] 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.

[49] L. Van Der Maaten, E. Postma, and J.p Van den Herik. 2009. Dimensionality reduction: A comparative. *Journal of Machine Learning Research* 10, 66-71 (2009), 13.

[50] D.-B. Wang, L. Li, and M.-L. Zhang. 2019. Adaptive graph guided disambiguation for partial label learning. In *Proceedings of the 25th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. Anchorage, AK, 83–91.

[51] H. Wu and S. Prasad. 2018. Semi-supervised dimensionality reduction of hyperspectral imagery using pseudo-labels. *Pattern Recognition* 74 (2018), 212–224.

[52] J. Wu, Z. Hong, S. Pan, X. Zhu, Z. Cai, and C. Zhang. 2014. Exploring features for complicated objects: Cross-view feature selection for multi-instance learning. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*. Shanghai, China, 1699–1708.

[53] J.-H. Wu and M.-L. Zhang. 2019. Disambiguation enabled linear discriminant analysis for partial label dimensionality reduction. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. Anchorage, AK, 416–424.

[54] S. Yan, D. Xu, B. Zhang, H.-J. Zhang, Q. Yang, and S. Lin. 2007. Graph embedding and extensions: A general framework for dimensionality reduction. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 29, 1 (2007), 40–51.

[55] R. Yang, D. Xu, L. Guo, and D. Zhang. 2022. Regularized two-stage submodular maximization under streaming. *Science China Information Sciences* 65, 4 (2022), 1–10.

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

[57] 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.

[58] M.-L. Zhang, J.-H. Wu, and W.-X. Bao. 2022. Disambiguation enabled linear discriminant analysis for partial label dimensionality reduction. *ACM Transactions on Knowledge Discovery from Data* 16, 4 (2022), 1–18.

[59] 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.

[60] 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.

[61] Y. Zhang and Z.-H. Zhou. 2010. Multilabel dimensionality reduction via dependence maximization. *ACM Transactions on Knowledge Discovery from Data* 4, 3 (2010), 1–21.

[62] 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), 1–19.

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

[64] Z.-H. Zhou and M.-L. Zhang. 2017. Multi-label learning. In *Encyclopedia of Machine Learning and Data Mining*. Springer, Berlin, 875–881.