Submodular Feature Selection for Partial Label Learning

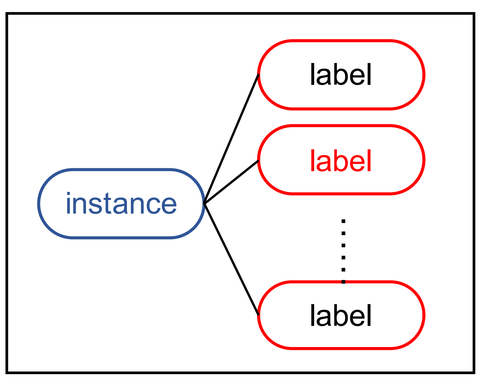
Wei-Xuan Bao, Jun-Yi Hang, Min-Ling Zhang*

School of Computer Science and Engineering, Southeast University, China

Key Laboratory of Computer Network and Information Integration (Southeast University), MOE, China

Introduction

Partial Label (PL) Learning



- Each instance is associated with multiple candidate labels
- Only one of candidate labels is true but unknown

PL Feature Selection

- In PL learning, most existing works focus on manipulating the label space while the task of manipulating the feature space by dimensionality reduction has been rarely investigated.
- Feature selection is a common strategy to perform dimensionality reduction, which helps remove irrelevant and redundant features, increase classification accuracy and enhance learning comprehensibility. However, it is challenge in PL learning framework due to the concealed ground-truth label.

Notations

$\mathcal{X}$: d-dimensional feature space $\mathbb{R}^d$

$\mathcal{D} = \{(x_i, S_i) | 1 \leqslant i \leqslant m\}$: PL training set with m examples

y_i : the ground-truth label of x_i, $y_i \in S_i$

A_p : the selected feature subset in the pth greedy step, which is initialized as $A_0 = \phi$

$I(\cdot,\cdot)$: the mutual information of the random variables

$\mathcal{Y}$: label space with q class labels $\{l_1, l_2, ..., l_q\}$

S_i : the candidate label set of x_i

$F = \{f_1, ..., f_d\}$: the original feature set

$H(\cdot)$: the entropy of the random variable

The SAUTE Approach

SAUTE performs feature selection via iteratively maximizing the dependency between selected feature variables and the latent label variable, which is evaluated by mutual information.

To fulfill the alternative procedure, we construct the labeling confidence matrix $\mathbf{Y} = [\mathbf{Y}(i,j)]_{m \times q}$ where each element $\mathbf{Y}(i,j)$ denotes the estimated confidence of l_j being the ground-truth label for $\boldsymbol{x}_i$ and initialize it as follows:

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

Dependency Maximization

The original objective function of MI-based dependency maximization is formulated as:

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

This problem is NP-Hard. However, $I(A;c)$ is a non-decreasing, non-negative submodular function under weak conditional independence assumption. As a result, the solution of the above problem can be approximated by a tailored greedy algorithm according to the properties of submodularity:

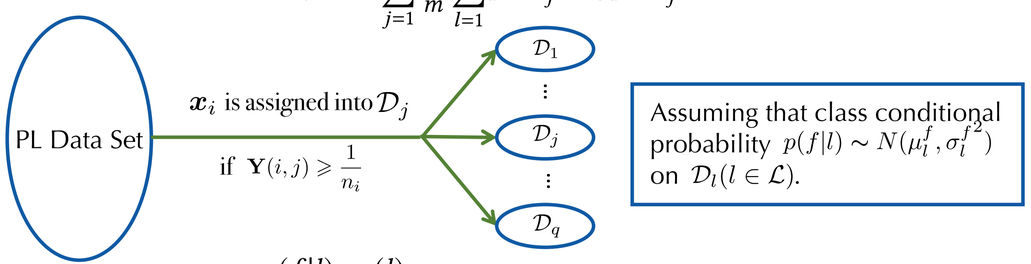
$$f_p^* = \underset{f \in F \setminus A_{p-1}}{\arg\max}\ I(f;c)$$

In order to further eliminate the influence of redundant features, we revise the greedy policy as:

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

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

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



$$\hat{p}(l|f) = \frac{p(f|l) \cdot p(l)}{\sum_{u \in \mathcal{L}} p(f|u) \cdot p(u)}, \text{ where } p(u) = \frac{\sum_{i=1}^{m} \mathbf{Y}(i,u)}{m} (u \in \mathcal{L})$$

The CENDA Approach (Cont.)

Latent Label Inference

According to the selected feature subset A_p, we construct the lower-dimensional PL training set $\mathcal{D}' = \{(\boldsymbol{x}_i', S_i) | 1 \leqslant i \leqslant m\}$.

The labeling confidence matrix is updated by resorting to kNN aggregation in the lower-dimensional feature space.

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

The intermediate matrix: $\mathbf{Y}' = (1-\alpha) \cdot \mathbf{Y} + \alpha \cdot \mathbf{L}$

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

Experiments

Synthetic Data Sets

We generate the synthetic PL data set from multi-class data set with controlling parameter r which denotes the number of false positive labels in candidate label set.

Win/tie/loss counts (pairwise t-test at 0.05 significance level) between $\mathcal{A}$-SAUTE and $\mathcal{A}$

	$\mathcal{A}$-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**

- Against $\mathcal{A}$, $\mathcal{A}$-SAUTE wins in 71.2% cases and loses only in 4.4% cases.

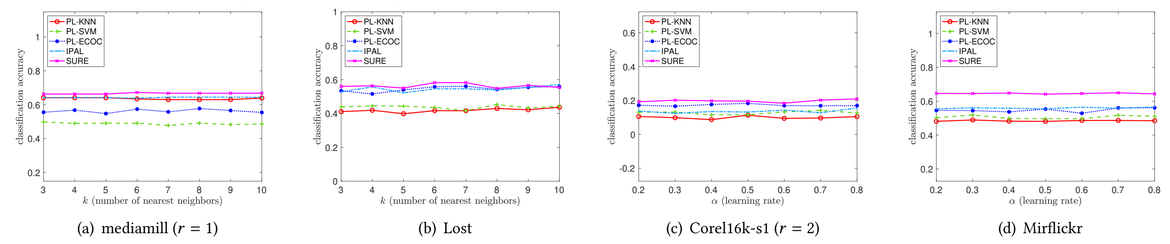
Real-World Data Sets

Win/tie/loss counts (pairwise t-test at 0.05 significance level) between $\mathcal{A}$-SAUTE and $\mathcal{A}$-baselines

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**

- Against $\mathcal{A}$-baselines, $\mathcal{A}$-SAUTE wins in 91.7% cases and never losses.

Sensitivity Analysis



(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`.

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 Y derived from SAUTE may bring further improvement of predictive performance for specific partial label learning algorithms with proper utilization.

MLA 2022, Nanjing