候选标记信息感知的偏标记学习算法

陈鸿昶 谢 天* 高 超 李邵梅 黄瑞阳

(国家数字交换系统工程技术研究中心 郑州 450002)

摘 要：在偏标记学习中，示例的真实标记隐藏在由一组候选标记组成的标记集中。现有的偏标记学习算法在衡量示例之间的相似度时，只基于示例的特征进行计算，缺乏对候选标记集信息的利用。该文提出一种候选标记感知的偏标记学习算法(CLAPLL)，在构建图的阶段有效地结合候选标记集信息来衡量示例之间的相似度。首先，基于杰卡德距离和线性重构，计算出各个示例的标记集之间的相似度，然后结合示例相似度和标记集的相似度构建相似度图，并通过现有的基于图的偏标记学习算法进行学习和预测。3个合成数据集和6个真实数据集上实验结果表明，该文方法相比于基线算法消歧准确率提升了0.3%～16.5%，分类准确率提升了0.2%～2.8%。

关键词：偏标记学习；弱监督学习；消歧；杰卡德距离；线性重构

中图分类号：TP18 **文献标识码**：A **文章编号**：1009-5896(2019)10-2516-09

DOI: 10.11999/JEIT181059

Candidate Label-Aware Partial Label Learning Algorithm

CHEN Hongchang XIE Tian GAO Chao LI Shaomei HUANG Ruiyang

(*National Digital Switching System Engineering & Technological R&D Center*, *Zhengzhou* 450002, *China*)

Abstract: In partial label learning, the true label of an instance is hidden in a label-set consisting of a group of candidate labels. The existing partial label learning algorithm only measures the similarity between instances based on feature vectors and lacks the utilization of the candidate labelset information. In this paper, a Candidate Label-Aware Partial Label Learning (CLAPLL) method is proposed, which combines effectively candidate label information to measure the similarity between instances during the graph construction phase. First, based on the jaccard distance and linear reconstruction, the similarity between the candidate labelsets of instances is calculated. Then, the similarity graph is constructed by combining the similarity of the instances and the label-sets, and then the existing graph-based partial label learning algorithm is presented for learning and prediction. The experimental results on 3 synthetic datasets and 6 real datasets show that disambiguation accuracy of the proposed method is 0.3%～16.5% higher than baseline algorithm, and the classification accuracy is increased by 0.2%～2.8%.

Key words: Partial label learning; Weakly supervised learning; Disambiguation; Jaccard distance; Linear reconstruction

1 引言

偏标记数据是一类常见的弱监督数据。在偏标记数据中，每一个示例的标记由多个标记组成的候选标记集表示，其中只有一个标记是该示例的真实标记[1]。例如：网络新闻中的每个人脸图像，在标题中可能存在多个对应的姓名文本。此外偏标记数据还在医疗诊断[2]、在线标注[3]、人机交互[4]等场景中广泛存在。由于偏标记数据的获取难度远低于准确标记数据，因此偏标记学习具有重要的研究价值和广阔的应用场景。

由于偏标记数据的真实标记隐藏在候选标记集中，因此标记具有复杂性模糊性，人们无法直接访问示例的真实标记。因此，现有的偏标记学习算法大都采用消歧算法来为每个偏标记数据在候选标记集中选取一个最可能的真实标记作为偏标记数据的唯一标记。根据采用的消歧策略的不同，消歧算法可以分为两种：基于辨识的消歧和基于平均的消歧。

基于辨识的消歧采用参数化模型进行消歧，该类方法将示例的真实标记作为参数模型的隐变量，通过最大间隔法[5]、最大似然法[6]等方法构建目标函数，采用迭代优化的方式来对模型的隐变量进行

收稿日期：2018-11-20；改回日期：2019-04-21；网络出版：2019-05-16
*通信作者： 谢天 xietianxt@foxmail.com
基金项目：国家自然科学基金(61601513)
Foundation Item: The National Natural Science Foundation of China (61601513)

更新求解。该方法需要合理的模型假设以及参数配置，否则消歧效果会受到很大影响。基于平均的消歧则不需要模型假设，该方法为所有候选标记设置相同的置信度，并结合近邻示例的候选标记进行消歧，其中基于图模型的方法应用最为广泛[7–9]，它的消歧过程主要分为两步：(1)根据示例特征向量之间的欧氏距离来衡量示例之间的相似度，并为示例之间建立一个相似度图；(2)根据相似度图，利用示例在相似度图中的各个近邻示例的候选标记集中的标记，通过采用加权投票，标签传播等方法来为当前示例进行消歧。

由于偏标记数据的标记具有模糊性，人们无法直接访问偏标记示例的真实标记，因此现有的基于图模型的偏标记学习算法在过构建相似度图的阶段没有利用候选标记信息，仅通过示例的特征向量间的欧氏距离来衡量示例之间的相似度[7,9]来构建图，然后第2步时在图模型的基础上才利用候选标记集信息来进行消歧。在构建图模型的过程中，只采用了示例的特征信息，因此现有方法在构建图模型的过程中是一种无监督的方法，而在根据图进行消歧过程中采用了候选标记信息，因此现有的方法整体上属于弱监督学习算法中的不准确监督算法[10,11]。目前在构建相似度图的过程中，还不存在采用了候选标记信息的方法。因此，本文提出一种在构建相似度图的过程中，考虑了示例候选标记集相似度的偏标记学习算法 (Candidate Label Aware Partial Label Learning, CLAPLL)，以下简称CAP，在人工UCI数据集和6个真实数据集上的实验上表明，本文算法可以提高基于图模型的偏标记学习算法的消歧和分类表现，获得了优于现有算法的结果。

2　本文方法

2.1　问题引入

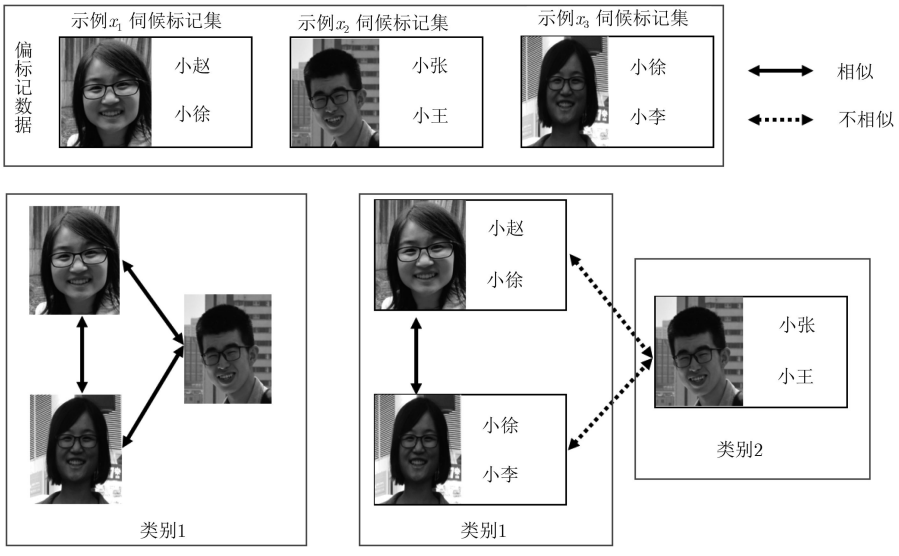
偏标记学习的特点是示例的真实标记隐藏在候选标记集中，算法无法直接访问示例的真实标记，现有的偏标记学习算法在衡量示例之间的相似性时，都仅考虑到了示例特征之间的相似性，即“特征相似的示例应该属于同一类”，并通过计算示例空间中的欧式距离来衡量示例特征之间的相似度，没有使用到标记空间中的信息。通过结合候选标记集信息，可以更好地来衡量示例之间的相似度，丰富训练信息。

如图1所示, 示例x_1, x_2和x_3在示例特征空间已经具有很高的相似度，然而在标记空间，x_1与x_2的标记集中无相同的标记，那么x_1与x_2必然属于不同的类别，只利用特征空间的信息算法会将x_1, x_2和x_3混淆为一类(如图1(a))，通过结合标记空间的信息，可以成功地区分x_1, x_3和x_2(如图1(b))。

因此，本文提出了CLAPLL算法，简称CAP算法。该方法通过构建示例特征相似度图$G_i(V,E)$和示例标记集相似度图$G_c(V,E)$来综合考虑示例的特征相似度和标记集相似度来衡量偏标记数据间的相似度关系。

2.2　基于候选标记感知的偏标记学习算法

假设$D=\{(\boldsymbol{X}_i,S_i)|1\le i\le m\}$是偏标记训练集，其中$\boldsymbol{X}_i\in\boldsymbol{X}$，$\boldsymbol{X}_i=(x_{i1},x_{i2},\cdots,x_{id})^{\mathrm{T}}$是示例$x_i$的$d$维特征矢量，$\boldsymbol{X}\in\mathbf{R}^{m\times d}$是训练数据的特征矩阵，$S_i\in Y$表示与$x_i$对应的候选标记集合，且满足



(a) 不采用候选标记集信息　　(b) 采用候选标记集信息

图 1 采用候选标记集信息的消歧效果

$Y=S_1\cup S_2\cup\cdots\cup S_m$。$x_i$的真实标记$y_i$未知，但满足$y_i\in S_i$。

为了排除偏标记数据各个特征采用了不同的量纲的影响，首先对特征矩阵进行Z-score归一化，将数据量化服从标准的正态分布，再对数据进行L2范数归一化以构建示例之间的相似度图$G_i(V,E)$。$V=\{x_i|1\le i\le m\}$表示每个示例，$N(x_i)$表示训练集中示例x_i在欧式距离下最近邻的k个示例，则相似度图G_i的边可以表示为$E_i=\{(x_i,x_j)|x_i\in N(x_j),1\le i\ne j\le m\}$，即如果示例$x_i$在示例$x_j$的$k$近邻中，那么存在一条连接$x_i$, x_j的边。对于边集E可以定义权重矩阵$\boldsymbol{W}_{ij}=[w_{ij}]_{m\times m}$，当存在一条连接$x_i$在示例$x_j$的边时$w_{ij}>0$，否则$w_{ij}=0$。$\boldsymbol{w}_j=[w_{i_1,j},w_{i_2,j},\cdots,w_{i_k,j}]$表示示例$x_j$与各个邻接节点的连接权重，$w_j$可以通过求解如式(1)的非线性最小二乘优化问题得到

$$\min_{w_j}\left\|x_j-\sum_{a=1}^{k}(w_{i_a,j}\cdot x_{i_a})\right\|+\frac{\mathrm{C}_1}{k}\sum_{a=1}^{k}(w_{i_a,j})^2,\quad \text{s.t. } w_{i_a}>0,\forall i_a\in N(i_j) \tag{1}$$

即利用x_j的k个最近邻样本对x_j进行线性重构的权重作为边的权重，其中$\sum_{a=1}^{k}(w_{i_a,j})^2$为正则化项，用于防止过拟合，参数$\mathrm{C}_1$用于调节正则化项的权重。

然后，再根据示例的标记集构建相似度图$G_\mathrm{c}=(V_\mathrm{c},E_\mathrm{c})$。为了构建标记集的相似度图，需要一种相似度度量来衡量两个示例标记集的相似程度。Jaccard距离是用来衡量集合之间差异性的一种指标，在聚类分析中有很好的应用效果。考虑采用Jaccard距离来构建标记集间的相似度图。示例x_i和示例x_j标记集的Jaccard距离可以表示为

$$J(S_i,S_j)=\frac{|S_i\cup S_j|-|S_i\cap S_j|}{|S_i\cup S_j|}\in[0,1] \tag{2}$$

$E_\mathrm{c}=\{(x_i,x_j)|x_i\in N_\mathrm{c}(x_j),1\le i\ne j\le m\}$, $N_\mathrm{c}(x_j)$表示示例x_j的标记集在Jaccard距离下的k个最近邻标记集。设边的权重矩阵为$\boldsymbol{U}_{ij}=[u_{ij}]_{m\times m}$，当存在一条连接$S_i$和$S_j$的边时$u_{ij}>0$，否则$u_{ij}=0$。$\boldsymbol{u}_j=[u_{i_1,j},u_{i_2,j},\cdots,u_{i_k,j}]$表示示例$S_j$与各个邻接节点的连接权重。由于Jaccard距离越小的标记集越相似，可以将$u_{i_a,j}$表示为

$$u_{i_a,j}=1-J(S_{i_a},S_j),\forall i_a\in N_\mathrm{c}(x_j) \tag{3}$$

此外，也可以考虑采用线性重构的方式来衡量标记集之间的相似程度，为了方便表示，将示例的候选标记集表示为$\boldsymbol{L}_i=[l_{i,1},l_{i,2},\cdots,l_{i,n}]$，当$l_{i,n}\in S_i$时$l_{i,n}=1$，否则$l_{i,n}=0$。利用示例$x_j$的最近邻$k$个示例的标记集来线性重构$S_j$，并将重构系数作为图边的权重。通过求解式(4)的优化问题来计算u_j

$$\min_{w_j}\frac{1}{|Y|}\left\|L_j-\sum_{a=1}^{k}(u_{i_a,j}\cdot L_{i_a})\right\|+\frac{\mathrm{C}_2}{k}\sum_{a=1}^{k}(u_{i_a,j})^2,\quad \text{s.t. } u_{i_a}>0,\forall i_a\in N(i_j) \tag{4}$$

其中，$\sum_{a=1}^{k}(u_{i_a,j})^2$为正则化项，用于防止过拟合，参数$\mathrm{C}_2$用于调节正则化项的权重。

设置信度矩阵为$\boldsymbol{F}=[f_{i,\mathrm{c}}]_{m\times n}$，每一行$\boldsymbol{F}_i=[f_{i,1},f_{i,2},\cdots,f_{i,n}]$，表示各个标记作为示例$x_i$的真实标记的置信度。那么可以通过求解式(5)得到

$$\arg\min_{F}\frac{(1-\alpha)}{mn}\cdot\sum_{i=1}^{m}\sum_{j=1,j\ne i}^{m}G_i(i,j)\,\|F_i-F_j\|+\frac{\alpha}{mn}\sum_{i=1}^{m}\sum_{j=1,j\ne i}^{m}G_\mathrm{c}(i,j)\,\|F_i-F_j\| \tag{5}$$

即同类示例之间候选标记的置信度的平均差异应当尽可能的小。合并同类项，式(5)可以写成

$$\arg\min_{F}\frac{1}{mn}\cdot\sum_{i=1}^{m}\sum_{j=1,j\ne i}^{m}G(i,j)\,\|F_i-F_j\| \tag{6}$$

其中，$G(i,j)=(1-\alpha)\,G_i(i,j)+\alpha G_\mathrm{c}(i,j)$为综合了示例和标记集相似度的相似度图。考虑到当$S_i\cap S_j=\varnothing$时，$x_i$, x_j属于不同类，不应存在连接x_i, x_j的边，故将$G(i,j)$修正为

$$G(i,j)=(1-\alpha)\,G_i(i,j)\kappa(G_\mathrm{c}(i,j)>0)+\alpha G_\mathrm{c}(i,j) \tag{7}$$

其中，$\kappa(\cdot)$为示性函数，当条件满足时取1，否则取0。

基于相似度图的偏标记算法都是根据相似度图来进行消歧，可以用$G(i,j)$替换现有基线算法的相似度图来进行消歧，得到消歧后的训练数据集$\hat{D}=\{(X_i,\hat{y}_i)|1\le i\le m\}$, $\hat{y}_i$为消歧后示例x_i的标记，其伪代码如表1所示。

分类阶段，对于未见示例x^*，根据它在消歧后的的训练集中的k个最近邻示例的标记通过加权的方式对x^*的标记进行预测。权重$\boldsymbol{w}_i^*=[w_{i_1}^*,w_{i_2}^*,\cdots,w_{i_k}^*]$, $(i_a\in N(x^*),1\le a\le k)$可以通过求解和式(1)相同的优化问题得到，那么，$x^*$的预测标记为

$$y^*=\arg\max_{y\in Y}\left(\sum_{a=1}^{k}\kappa(y=L_{i_a})\cdot w_{i_a}^*\right) \tag{8}$$

设偏标记数据集包含n个示例，特征向量的维数为d，标记集标记数量为s，每个示例与最近的k个示例存在连边。可以分析得到在构建相似度图

表 1　候选标记信息感知的偏标记学习算法伪代码

输入：偏标记数据集$D=\{(X_i,S_i)|1\le i\le m\}$，最近邻样本数$k$，标记相似度权重$\alpha$

训练阶段：

1 对特征矩阵$\boldsymbol{X}\in\mathbf{R}^{m\times d}$进行Z-score归一化；

2 根据式(1)求$\boldsymbol{w}_j$；

3 根据$\boldsymbol{w}_j$构建相似度图$G_i(V,E)$；

4 switch v；

　case Jaccard：根据式(3)计算$\boldsymbol{u}_j$，并构建候选标记集相似度图$G_c(i,j)$, (CAP-J算法)；

　case linear：根据式(4)计算$\boldsymbol{u}_j$，并构建候选标记集相似度图$G_c(i,j)$, (CAP-L算法)；

　end switch

5 根据式(7)计算最终相似度图$G(i,j)$；

6 结合现有图模型偏标记学习算法进行消歧，得到消歧结果$\hat{D}=\{(X_i,\hat{y}_i)|1\le i\le m\}$；

测试阶段：

7 对于未见示例x^*，根据式(8)计算得分类结果；

输出：消歧结果$\hat{D}=\{(X_i,\hat{y}_i)|1\le i\le m\}$和分类结果$y^*$。

时基线算法和本文算法CAP的复杂度，如表2所示。其中，CAP-J表示采用Jaccard距离来衡量标记集之间的相似度，CAP-L表示采用线性重构衡量

表 2　基线算法和本文算法复杂度比较

	算法复杂度	实际复杂度
基线算法	$O(d^2n^3\lg(n))$	$O(d^2n^3\lg(n))$
本文算法(CAP-J)	$O(d^2n^3\lg(n)+(s+1)k^2)$	$O(d^2n^3\lg(n))$
本文算法(CAP-L)	$O(d^2n^3\lg(n)+(sk+1)k^2)$	$O(d^2n^3\lg(n))$

标记集之间的相似度。对于表3真实偏标记数据集，样本数量远远大于样本的特征维数，即$n>>s$，同时显然$n>>k$，因此可以忽略掉本文算法的复杂度中的第2项，即在实际情况中，本文算法的复杂度和基线算法相同。

3　实验及结果分析

3.1　实验设置

为了彻底评估本文提出的方法的有效性，在4个基于UCI数据集[12]的合成数据集和6个真实数据集上进行了实验。表3和表4分别列出了合成数据集和真实数据集的基本特征。

采用了偏标记学习研究中广泛使用的方法来合成数据集，通过控制参数p和r可以将多类分类的UCI数据集合成为偏标记数据集，其中p表示合成数据集中带有多个标记的示例的比率，r控制多个标记的示例的平均候选标记数，参数取值如表3所示。

表 3　真实偏标记数据集的特征

数据集	样本数	特征数	类别标记数	候选标记数		
				平均	最小	最大
Lost	1122	108	16	2.23	1	3
Birdsong	4998	38	13	2.18	1	4
MSRSCv2	1758	48	23	3.16	1	7
FG-NET	1002	262	78	7.48	2	11
Yahoo! News	22991	163	219	1.91	1	5
Soccer Player	17472	279	171	2.09	1	11

实验采用的真实数据集来自于不同的应用场景，例如来自于人脸标注领域的Lost[9], Soccer Player[13]和Yahoo！ News数据集[14]；来自于人脸年龄估计的FG-NET[15]数据集；来自鸟鸣分类领域的Birdsong[16]数据集；来自目标检测领域的MSRCv2[17]数据集。每个数据集的候选标记数目如表4所示。

表 4　合成偏标记数据集的特征

数据集	样本数	特征数	类别标记数	参数设置
Ecoli	336	7	8	p={0.1, 0.2, 0.3, 0.4,0.5, 0.6, 0.7, 0.8} r={1, 2, 3, 4, 5}
Movement	360	90	15	
CTG	2126	21	10	

为了验证本文方法的有效性，在以下3个基于图模型的偏标记学习算法上进行了修改和对比，这3种方法在构建相似度图的过程中均没有使用候选标记信息。3个基准算法分别为：

PLKNN[9]：一种基于平均的消歧方法，采用示例的k个最近邻示例标记加权投票的方式确定每个偏标记示例的标记，近邻样本数设置为k=10。

IPAL[7]：一种基于平均的消歧方法，在为示例的标记设置初始置信度后，采用标签传播的方式来迭代更新置信度，求得示例标记，近邻样本数设置为k=10，最大迭代次数设置为T=100。

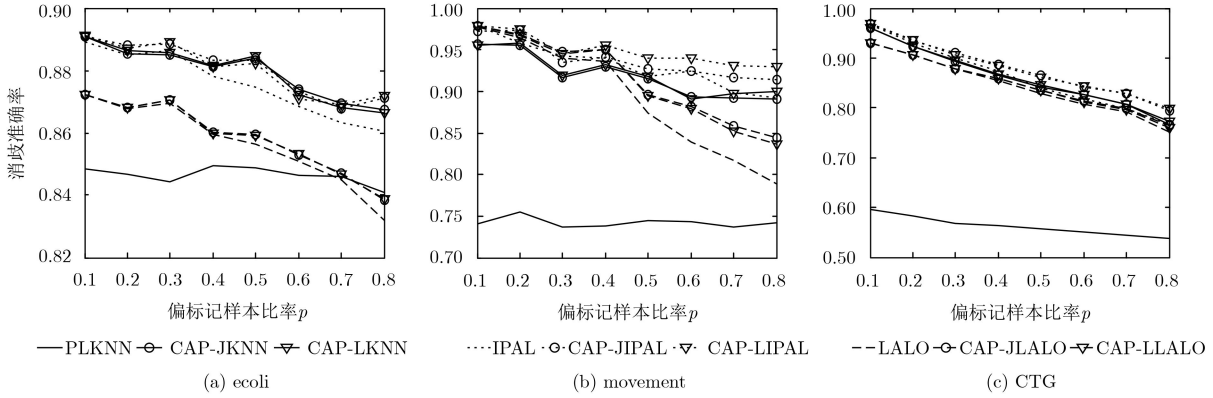
LALO[8]：一种基于辨识的消歧算法，通过参数模型估计候选标记置信度的隐分布，并采用标签

传播的方式来更新置信度，近邻样本数设置为k=10，最大迭代次数设置为T=100。

为了验证结合候选标记信息进行偏标记学习的CAP算法的合理性，本文分别在以上3种基线算法基础上进行修改，提出了在构建相似度图的过程中采用Jaccard距离衡量候选标记集相似度的CAP-J算法：CAP-JKNN, CAP-JIPAL以及CAP-JLALO算法，与采用线性重构衡量候选标记集相似度的CAP-L算法：CAP-LKNN, CAP-LIPAL以及CAP-LLALO算法。分别与对应的基线算法PLKNN, IPAL, LALO进行对比。同时，在真实数据集实验中添加了采用样本真实标记来构建相似度图的对照组，用来对比本文与采用真实标记(强监督信息)之间的差距。需要说明的是，实际场景中，并不能获得偏标记数据的真实标记，因此该方法只是用于对照，并不符合实际情况。本文采用了两种偏标记学习研究中广泛使用的评价指标，消歧准确率[18]和分类准确率[19]，即正确消歧的样本占所有偏标记样本的比例和正确分类的样本占所有测试样本的比例。

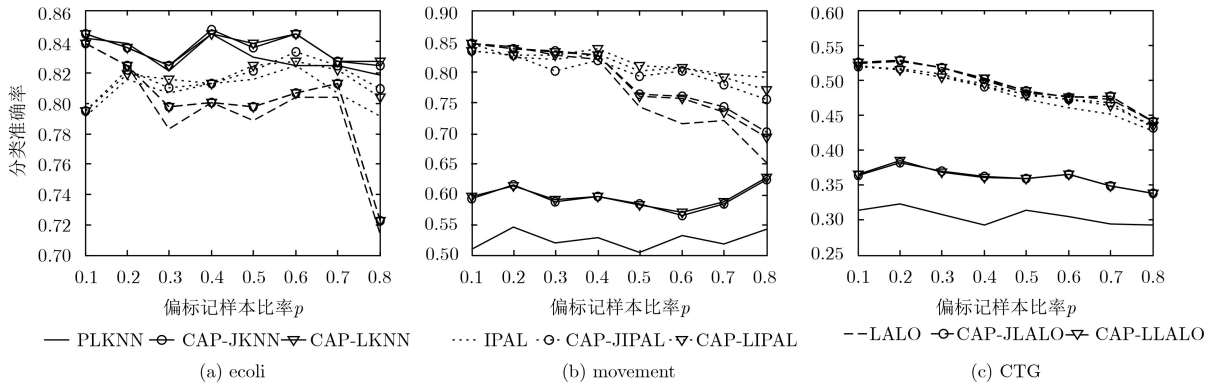
3.2 UCI合成数据集实验及参数敏感性分析

图2—图9展示了候选标记信息感知的偏标记学



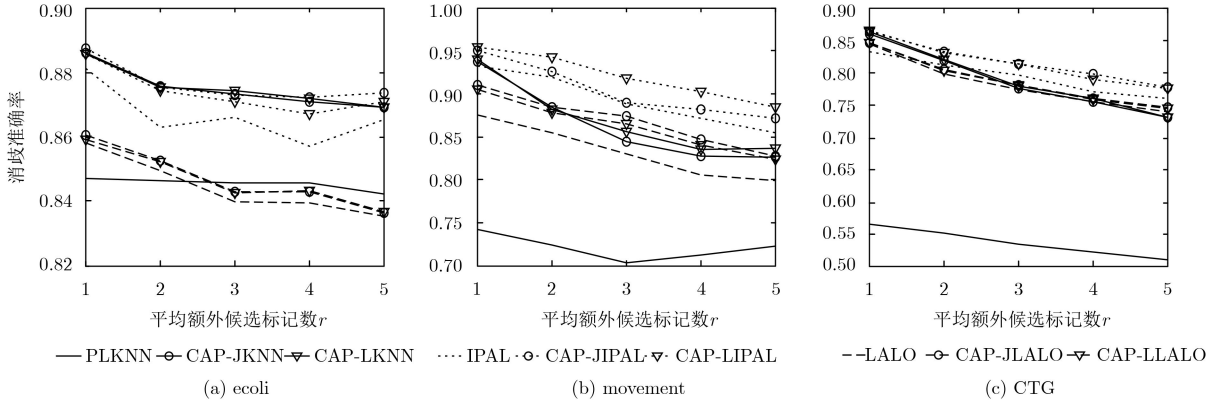
(a) ecoli (b) movement (c) CTG

图 2 消歧准确率随参数p的变化



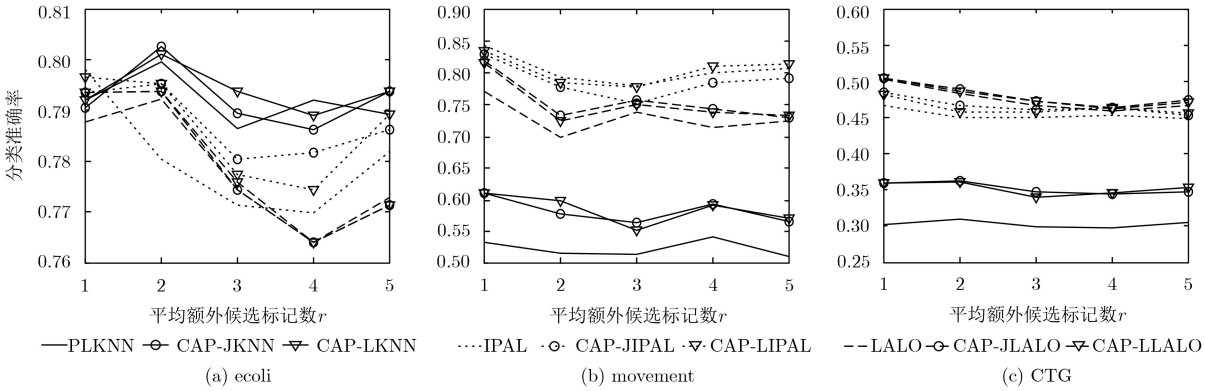
(a) ecoli (b) movement (c) CTG

图 3 分类准确率随参数p的变化



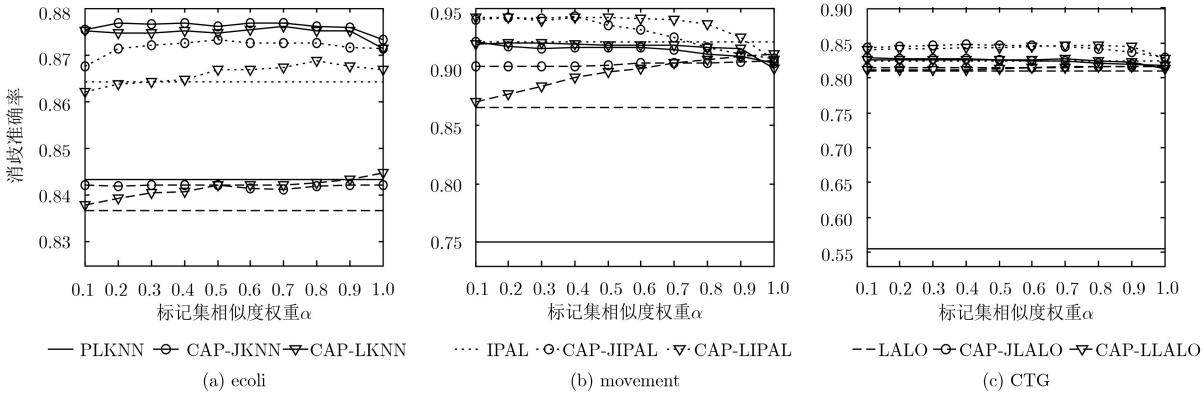
(a) ecoli (b) movement (c) CTG

图 4 消歧准确率随参数r的变化



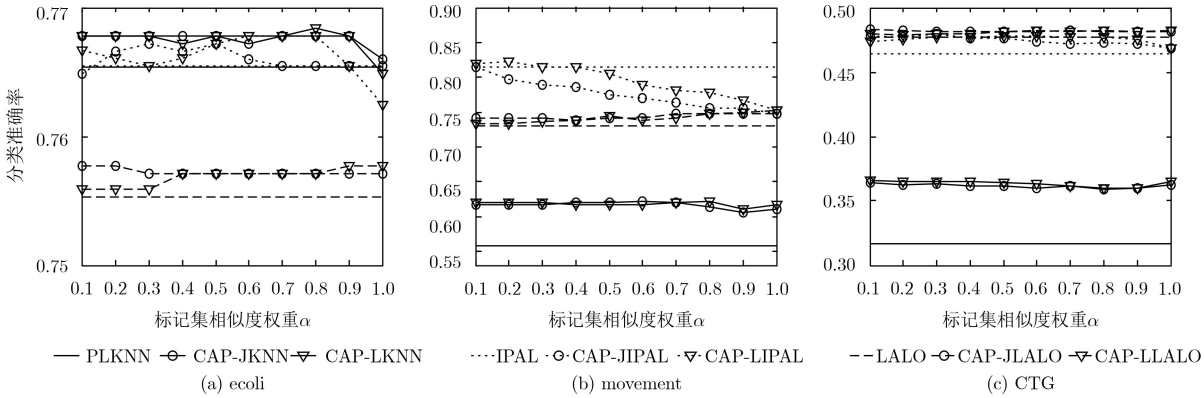
(a) ecoli　(b) movement　(c) CTG

图 5 分类准确率随参数r的变化



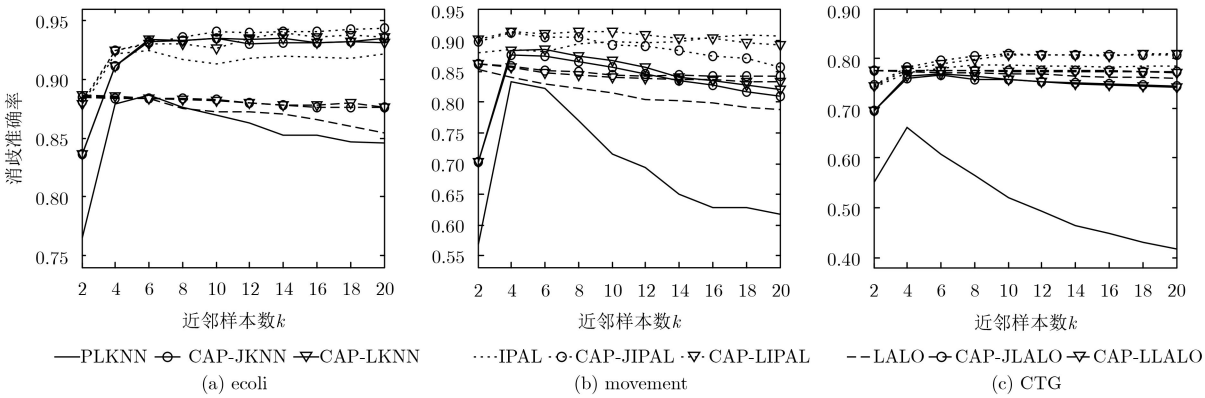
(a) ecoli　(b) movement　(c) CTG

图 6 消歧准确率随参数α的变化



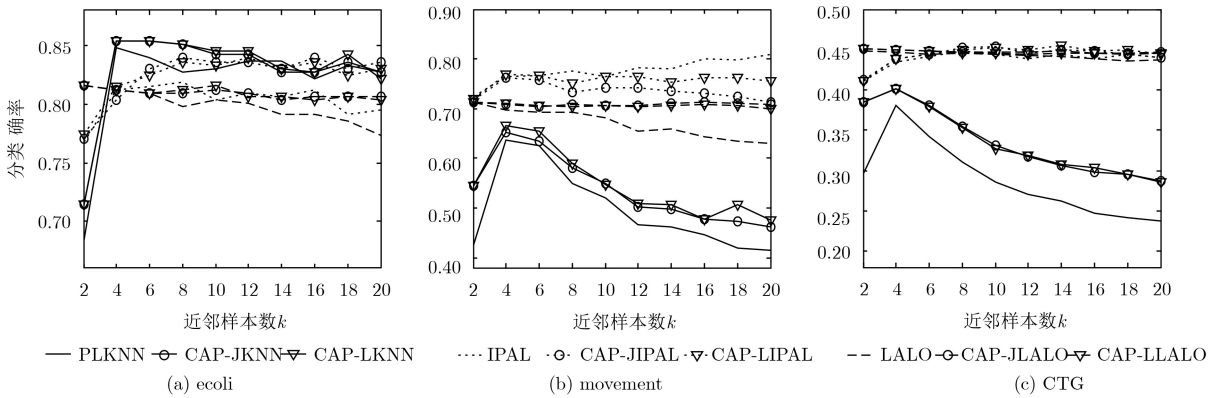
(a) ecoli　(b) movement　(c) CTG

图 7 分类准确率随参数α的变化



(a) ecoli　(b) movement　(c) CTG

图 8 消歧准确率随参数k的变化



(a) ecoli (b) movement (c) CTG

图 9 分类准确率随参数k的变化

习算法CAP-J, CAP-L与基线算法的消歧准确率和分类准确率效果对比，分别对应改变参数p(偏标记样本比率)，r(平均候选标记数量)，α(候选标记集相似度权重)，k(近邻样本数)时算法准确率的变化。分别表示算法在3个UCI合成数据集上消歧准确率和分类准确率的表现。

从图2—图5可以发现，采用了候选标记集相似度信息后，大多数情况下算法的消歧准确率和分类准确率均有显著提升，随着参数p和r的增大，算法的消歧准确率和分类准确率均有降低趋势，但考虑了标记集相似度信息的算法降低比率要明显小于基线算法。根据图6和图7，按不同权重采用标记集相似度信息后算法准确率有显著提高。图8的结果显示随着近邻样本的增多，基线算法的消歧准确率明显下降，但考虑了候选标记集相似度的算法下降不明显，说明考虑了候选标记集信息的偏标记学习算法对多近邻样本的情况有很好的消歧表现。

3.3 真实数据集实验

本文基于10倍交叉验证进行了20组实验，并采用了消歧准确率和分类准确率两项评价指标，计算出准确率的均值和标准差，结果如表5和表6所示，最优的算法结果在表中用粗体表示。

由表5可以看出，本文的结合候选标记集相似度信息的偏标记学习算法CAP-J和CAP-L在绝大多数数据集上均取得了优于基线算法的消歧表现，仅在FG-NET数据集上差于基线算法LALO。相比于采用了强监督信息的对照组，本文算法在消歧阶段取得了良好的表现。

由表6可以看出，本文算法CAP-J, CAP-L在绝大多数数据集上均取得了优于基线算法的分类表现。受实验电脑内存限制，LALO和CAP-JLALO, CAP-LLALO算法在Soccer Player和Yahoo! News数据集上的结果没有给出。相比于采用了强监督信息的对照组，本文算法准确率仅低了0.06%～5.19%，这是由于偏标记数据相比强监督数据具有更少的可用信息。本文方法通过采用弱监督信息也得以达到较为理想的效果。

表 5 不同算法在真实偏标记数据集上的消歧准确率(%)

数据集	消歧准确率(mean±std.)					
	Lost	MSRCv2	BirdSong	FG-NET	Soccer Player	Yahoo! News
PLKNN	67.54±0.09	51.00±0.09	68.69±0.04	11.06±0.13	52.60±0.02	66.06±0.02
CAP-JKNN	**73.60±0.10**	**62.19±0.08**	**77.14±0.04**	14.71±0.15	**69.55±0.01**	**80.00±0.02**
CAP-LKNN	73.38±0.13	61.88±0.09	76.67±0.04	**14.81±0.17**	69.22±0.02	79.78±0.05
PLKNN(监督)	84.93±0.04	73.07±0.02	84.29±0.14	14.94±0.05	90.65±0.03	91.21±0.03
IPAL	84.01±0.15	70.58±0.15	83.61±0.04	15.28±0.19	67.65±0.03	84.99±0.05
CAP-JIPAL	**85.58±0.17**	**71.25±0.20**	**84.22±0.04**	**15.40±0.19**	**67.94±0.02**	**85.33±0.04**
CAP-LIPAL	85.39±0.24	70.92±0.12	84.40±0.05	14.86±0.17	67.89±0.07	85.21±0.03
IPAL(监督)	85.43±0.32	76.43±0.22	85.92±**0.10**	15.53±**0.18**	71.43±0.05	86.43±0.06
LALO	75.05±1.24	59.42±0.89	78.14±0.75	**15.92±0.69**	–	–
CAP-JLALO	76.80±1.11	59.48±1.09	78.02±0.81	15.69±0.75	–	–
CAP-LLALO	**80.22±1.08**	**59.72±0.82**	**78.24±0.64**	15.76±0.94	–	–
LALO(监督)	84.53±1.53	60.04±1.14	79.25±0.88	16.13±0.62	–	–

表 6　不同算法在真实偏标记数据集上的分类准确率(%)

数据集	消歧准确率(mean±std.)					
	Lost	MSRCv2	BirdSong	FG-NET	Soccer Player	Yahoo! News
PLKNN	61.48±0.78	44.12±0.36	64.66±0.23	5.58±0.42	49.55±0.04	58.30±0.06
CAP-JKNN	**64.01±0.65**	**46.35±0.38**	**66.01±0.26**	**6.24±0.38**	**50.77±0.09**	**61.18±0.05**
CAP-LKNN	63.58±0.72	46.14±0.48	65.88±0.21	5.74±0.56	50.43±0.09	60.50±0.12
PLKNN(监督)	69.26±0.48	51.33±0.30	68.49±0.13	6.98±0.21	54.26±0.05	61.53±0.08
IPAL	73.18±0.79	53.08±0.33	71.09±0.33	5.28±0.55	54.84±0.10	65.88±0.14
CAP-JIPAL	**73.95±0.68**	**53.35±0.50**	71.34±0.30	5.45±0.60	**55.00±0.10**	**66.02±0.16**
CAP-LIPAL	73.44±0.68	52.61±0.71	**71.60±0.26**	**5.89±0.57**	54.46±0.18	66.02±0.18
IPAL (监督)	75.04±0.82	55.71±0.46	72.05±0.27	5.95±0.62	55.38±0.13	66.83±0.15
LALO	72.15±3.04	50.13±2.03	72.99±1.54	6.11±1.61	–	–
CAP-JLALO	73.02±2.88	49.23±2.10	73.00±1.62	5.96±1.19	–	–
CAP-LLALO	**74.84±2.20**	**50.27±3.19**	**73.37±1.50**	**6.76±1.64**	–	–
LALO(监督)	76.68±2.19	52.31±2.49	74.87±1.26	7.03±1.29	–	–

4　结束语

为了解决现有基于图模型的偏标记学习算法忽略了候选标记集信息的问题，本文提出了一种候选标记信息感知的偏标记学习方法CLAPLL(简称CAP)，通过Jaccard距离和线性重构两种方法来衡量候选标记集之间的相似度，从而有效地解决了如何在构建相似度图的过程中利用候选标记信息的问题。在UCI合成数据集和真实数据集上的实验结果表明，结合候选标记集信息可以有效地提高偏标记学习算法的消歧和分类表现。与在构建相似度图时忽略了候选标记集信息的基线算法相比，本文候选标记感知的偏标记学习算法在实际情况中，可以在不增加算法复杂度的同时取得更加优越的结果。本文的实验结果表明，采用不同方法衡量候选标记集之间的相似度可以获得不同的消歧和分类效果，寻找更好的衡量相似度的方法值得进行一步研究。

参 考 文 献

[1] HÜLLERMEIER E and BERINGER J. Learning from ambiguously labeled examples[J]. *Intelligent Data Analysis*, 2006, 10(5): 419–439. doi: 10.3233/IDA-2006-10503.

[2] SONG Jingqi, LIU Hui, GENG Fenghuan, *et al*. Weakly-supervised classification of pulmonary nodules based on shape characters[C]. The 14th International Conference on Dependable, Autonomic and Secure Computing, The 14th International Conference on Pervasive Intelligence and Computing, The 2nd International Conference on Big Data Intelligence and Computing and Cyber Science and Technology Congress, Auckland, New Zealand, 2016: 228–232.

[3] TANG Caizhi and ZHANG Minling. Confidence-rated discriminative partial label learning[C]. The 31st AAAI Conference on Artificial Intelligence, San Francisco, USA, 2017: 2611–2617.

[4] TODA T, INOUE S, and UEDA N. Mobile activity recognition through training labels with inaccurate activity segments[C]. The 13th International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services, Hiroshima, Japan, 2016: 57–64.

[5] YU Fei and ZHANG Minling. Maximum margin partial label learning[J]. *Machine Learning*, 2017, 106(4): 573–593. doi: 10.1007/s10994-016-5606-4.

[6] LUO Jie and ORABONA F. Learning from candidate labeling sets[C]. The 23rd International Conference on Neural Information Processing Systems, Vancouver, Canada, 2010: 1504–1512.

[7] ZHANG Minling and YU Fei. Solving the partial label learning problem: An instance-based approach[C]. The 24th International Conference on Artificial Intelligence, Buenos Aires, Argentina, 2015: 4048–4054.

[8] FENG Lei and AN Bo. Leveraging latent label distributions for partial label learning[C]. The Twenty-Seventh International Joint Conference on Artificial Intelligence, Stockholm, Sweden, 2018: 2107–2113.

[9] COUR T, SAPP B, and TASKAR B. Learning from partial labels[J]. *Journal of Machine Learning Research*, 2011, 12: 1501–1536.

[10] ZHOU Zhihua. A brief introduction to weakly supervised learning[J]. *National Science Review*, 2018, 5(1): 44–53. doi: 10.1093/nsr/nwx106.

[11] TOLDO R and FUSIELLO A. Robust multiple structures estimation with J-linkage[C]. The 10th European

Conference on Computer Vision, Marseille, France, 2008: 537–547.

[12] DUA D and TANISKIDOU E K. UCI machine learning repository[EB/OL]. http://archive.ics.uci.edu/ml, 2017.

[13] ZENG Zinan, XIAO Shijie, JIA Kui, *et al*. Learning by associating ambiguously labeled images[C]. 2013 IEEE Conference on Computer Vision and Pattern Recognition, Portland, USA, 2013: 708–715.

[14] GUILLAUMIN M, VERBEEK J, and SCHMID C. Multiple instance metric learning from automatically labeled bags of faces[C]. The 11th European Conference on Computer Vision, Heraklion, Greece, 2010: 634–647.

[15] ZHANG Minling, ZHOU Binbin, and LIU Xuying. Partial label learning via feature-aware disambiguation[C]. The 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, USA, 2016: 1335–1344.

[16] BRIGGS F, FERN X Z, and RAICH R. Rank-loss support instance machines for MIML instance annotation[C]. The 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Beijing, China, 2012: 534–542.

[17] LIU Liping and DIETTERICH T G. A conditional multinomial mixture model for superset label learning[C]. The 25th International Conference on Neural Information Processing Systems, Lake Tahoe, USA, 2012: 548–556.

[18] ZHANG Minling, YU Fei, and TANG Caizhi. Disambiguation-free partial label learning[J]. *IEEE Transactions on Knowledge and Data Engineering*, 2017, 29(10): 2155–2167. doi: 10.1109/TKDE.2017.2721942.

[19] ZHANG Minling and YU Fei. Solving the partial label learning problem: An instance-based approach[C]. The 24th International Conference on Artificial Intelligence, Buenos Aires, Argentina, 2015: 4048–4054.

陈鸿昶：男，1964年生，教授，博士生导师，研究方向为通信与信息系统，大数据处理分析.

谢 天：男，1994年生，硕士生，研究方向为机器学习.

高 超：男，1982年生，博士，研究方向为计算机视觉，机器学习.

李邵梅：女，1982年生，博士，研究方向为计算机视觉，机器学习.

黄瑞阳：男，1986年生，博士，研究方向为网络大数据分析.