基于成对约束的偏标记数据消歧算法

征 察 [1] 吉立新 [1] 高 超 [1] 李邵梅 [1] 吴翼腾 [1]

摘 要 偏标记数据消歧是利用偏标记数据进行机器学习的基础. 针对偏标记数据中广泛存在的数据不平衡问题, 以及现有消歧算法对样本间约束信息利用不足的问题, 本文提出一种基于成对约束的偏标记数据消歧算法. 首先, 基于低秩表示, 推导出数据不平衡条件下样本低秩表示系数和样本相似度之间的关系; 其次, 基于推导结果, 分别构建基于样本间正约束和负约束的图模型, 通过最小化图模型的能量函数求解偏标记数据的标签. 在 5 个公开数据集上的实验结果表明本文方法相对基准算法在消歧准确率上平均提高了 2.9 % ∼ 14.9 %.

关键词 偏标记数据, 消歧, 数据不平衡, 低秩表示, 成对约束

引用格式 征察, 吉立新, 高超, 李邵梅, 吴翼腾. 基于成对约束的偏标记数据消歧算法. 自动化学报, 2020, **46**(7): 1367−1377

DOI 10.16383/j.aas.c170522

Partial Label Data Disambiguation Algorithm Based on Pairwise Constraints

ZHENG Cha[1] JI Li-Xin[1] GAO Chao[1] LI Shao-Mei[1] WU Yi-Teng[1]

Abstract Partial label data disambiguation is the basis of machine learning using partial label data. In order to solve the data imbalance problem widely existing in partial label data, and the problem that the existing disambiguation algorithms have insufficient utilization of constraints between samples, a partial label data disambiguation algorithm based on pairwise constraints is proposed in this paper. Firstly, the relation between low-rank representation coefficients and sample similarities in unbalanced datasets is deduced by utilizing low-rank representation. Secondly, according to the deduced results, two graphs are created based on positive constraint and negative constraint respectively. Finally, the labels of partial label data samples are obtained by minimizing energy functions based on graphs. Experimental results on five open datasets indicate that the proposed algorithm outperforms benchmark algorithms by 2.9 % ∼ 14.9 % at disambiguation accuracy.

Key words Partial label data, disambiguation, imbalanced data, low-rank representation, pairwise constraints

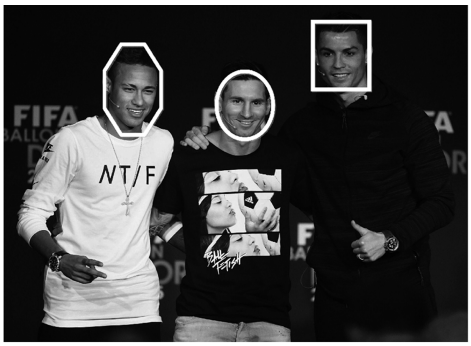
Citation Zheng Cha, Ji Li-Xin, Gao Chao, Li Shao-Mei, Wu Yi-Teng. Partial label data disambiguation algorithm based on pairwise constraints. *Acta Automatica Sinica*, 2020, **46**(7): 1367−1377

偏标记数据是一种常见的弱监督数据. 在这类数据中, 每个样本同时具备多个候选类别标签, 但只有一个标签是正确的. 图 1 展示了两例典型的偏标记数据, 将新闻标题中的人名作为新闻图像中人脸的姓名标签, 则一个人脸可能对应多个姓名标签[1]; 将诊断图像对应的可能病因作为医学图像的标签, 则图像可能对应多个病因标签[2]. 和带有唯一、正确标签的强监督数据集类似, 偏标记数据集也常具有高维、数据不平衡的特点. 但由于偏标记数据获取成本远低于传统监督学习所需的强监督数据, 如何利用偏标记数据进行弱监督学习已成为机器学习中的一个研究热点, 具有广阔的应用前景.

为利用偏标记数据进行学习, 文献 [3] 提出一种基于纠错输出编码的偏标记学习方法, 直接利用偏标记数据训练一个多分类器, 但该方法在训练过程中可能存在部分数据未被利用的情况. 为充分利用偏标记数据, 大多数偏标记学习算法[4−10] 首先对偏标记数据进行消歧, 确定每个偏标记样本的正确类别标签. 根据是否需要利用参数模型来假设样本分布, 现有的消歧方法可以分为两类: 1) 基于辨识 (Identification) 的消歧; 2) 基于平均 (Averaging) 的消歧.

基于辨识的消歧将偏标记样本的真实标签设为参数模型的隐变量, 并基于最大似然准则[2], 或最大间隔准则[5, 10] 建立目标函数, 之后采用迭代的方式优化目标函数求解隐变量实现消歧. 如文献 [6] 提出一种基于字典学习的消歧算法, 首先假设每类数据呈高斯混合分布, 然后迭代地对样本标签置信度矩阵 – 字典矩阵进行更新, 并根据最终的样本标签置信度矩阵来确定样本标签. 基于辨识的方法需要进

收稿日期 2017-09-13 录用日期 2018-04-16
Manuscript received September 13, 2017; accepted April 16, 2018
国家自然科学基金 (61601513) 资助
Supported by National Natural Science Foundation of China (61601513)
本文责任编委 王立威
Recommended by Associate Editor WANG Li-Wei
1. 国家数字交换系统工程技术研究中心 郑州 450002
1. National Digital Switching System Engineering & Technological R&D Center of China, Zhengzhou 450002



梅西、C 罗、内马尔等入选国际足联公布的 2017 年度世界足球先生候选人名单

(a) 关联多个姓名的人脸图像
(a) Face images associated with multiple names

(b) 关联多种病因的症状
(b) Symptom associated with multiple diseases

图 1 典型的偏标记数据

Fig. 1 Examples of typical partial label data

行合理的模型假设, 错误的模型假设将对消歧带来不利影响[4]. 基于平均的消歧通过赋予偏标记样本的各个候选标签相同的权重, 综合学习模型在各候选标签上的输出实现消歧[7−9]. 在基于平均的消歧算法中, 图模型因无需模型假设, 以及便于描述样本间的标签关系而得到广泛应用[7, 9]. 它是根据一定规则在样本间建立一个有权图 $G = (V, E, W)$, 通过分析图模型上节点候选标签间的关系来消歧. 其中 V 代表样本集合, E 为样本间的边集合, W 为相应的边权重集合. 文献 [7] 采用近邻消歧算法, 根据弱监督学习中的流形假设[11] 构建图模型: 假设邻近样本具有相同的标签, 令样本 $\boldsymbol{x}$ 和其近邻样本建立连边, 通过对近邻样本的候选标签集加权投票来确定 $\boldsymbol{x}$ 的标签. 文献 [9] 改进了文献 [7], 提出基于实例的偏标记学习算法[9] (Instance-based partial label learning, IPAL). 在构建图模型后, 该方法采用迭代的标签传播算法进行消歧.

虽然现有的基于图模型的消歧算法具有无需模型假设的优势, 但仍存在问题. 首先, 偏标记数据通常具有较高的维度, 而欧氏距离等一些常用于度量相似度的方法在高维空间中通常难以奏效. 近年来, 在半监督学习和聚类领域, 低秩表示[12] 因擅于表达高维数据结构而在构建图模型时取得良好效果[13−15]. 然而, 这些方法都是针对数据平衡的数据集, 而偏标记数据常面临数据不平衡问题, 因此无法直接将低秩表示用于偏标记数据消歧. 其次, 现有消歧算法中, 建立图模型时只利用一种样本间约束, 即更相似样本之间边权重越大, 标签相同的可能性越大. 但在半监督、聚类领域, 有两种约束得到广泛使用, 其中一种被称为正约束 (Must-link), 即部分样本必定属于同一类, 另一种为负约束 (Cannot-link), 即部分样本必定属于不同类. 这两种约束通常共同使用, 因此被合称为成对约束. 研究表明利用成对约束能有效提高聚类效果[16−17]. 受此启发, 本文将成对约束的概念迁移至偏标记数据消歧中, 将 “相似样本应具有相同标签” 定义为正约束, 将 “差异较大的样本应具有不同标签” 定义为负约束, 采用成对约束对偏标记数据消歧.

综上, 本文提出一种基于成对约束的偏标记数据消歧算法 (Partial label data disambiguation algorithm based on pairwise constraints, PLDPC), 其创新之处在于: 1) 针对偏标记数据中广泛存在的数据不平衡问题, 研究该条件下低秩表示系数和样本相似度的关系; 2) 在考虑数据不平衡后, 利用低秩表示构建两个分别基于正、负约束的图模型, 并基于定义在图模型上的能量函数[18], 结合类块标准化 (Class mass normalization) 准则[19] 进行消歧.

本文主要分为 5 个部分: 第 1 节简要介绍低秩表示算法; 第 2 节和第 3 节是本文的主要工作: 第 2 节对数据不平衡条件下低秩表示系数和样本相似度的关系进行分析, 第 3 节基于第 2 节中的结论提出基于成对约束的偏标记数据消歧算法; 第 4 节是仿真实验及结果; 第 5 节对本文进行总结.

1 低秩表示

低秩表示是一种广泛应用的子空间分割方法. 假设在 D 维欧几里得空间中, 存在一组向量 $X = [\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_n]$ (每列为一个样本), 这些样本分布在 k 个线性子空间 $\{S_i\}_{i=1}^{k}$ 上, 子空间具有低秩特性且相互独立. 理想情况下, 低秩表示可以将 X 分割到这 k 个线性子空间中, 使每个空间上的样本对应一个类. 具体地, 低秩表示将数据矩阵 X 自身作

为字典矩阵, 求解 X 在字典矩阵下的低秩表示系数矩阵 Z. 问题的优化目标如式 (1) 所示:

$$\arg_Z \min ||Z||_* \quad \text{s.t.} \quad X = XZ \tag{1}$$

其中, $||Z||_*$ 为 Z 的核范数, 是 Z 秩的凸近似, 定义为 $||Z||_* = \sum_{i=1}^{\text{rank}(Z)} \delta_i$, δ_i 为 Z 第 i 个奇异值. 针对 X 中含有噪声的情况, 可通过加入噪声项 E 来增加鲁棒性, 如式 (2) 所示:

$$\arg_{Z,E} \min ||Z||_* + \lambda ||E||_{1,1} \quad \text{s.t.} \quad X = XZ + E \tag{2}$$

式中, 参数 $\lambda > 0$, $||E||_{1,1} = \sum_i \sum_j |E_{i,j}|$, 用于控制噪声的稀疏性. 为较好地平衡精度和效率, 式 (2) 的求解通常采用非精确增广拉格朗日乘子法[12]. 文献 [12] 证明, 当 $X = [X_1, X_2, \cdots, X_k]$ (X_i 为 X 中第 i 类的所有样本组成的矩阵) 中每个子空间采样充足, 且噪声稀疏、有限时, 求解式 (2) 得到的 Z 近似于式 (3) 所示的块对角矩阵, 即同一类别样本被分割到同一子空间中, 相互表示系数绝对值较大, 而不同类别样本被分割到不同的子空间中, 相互表示系数接近于 0. 因此, 低秩表示系数可以很好地描述数据全局结构[13]. 文献 [12] 使用 Z 中元素的绝对值构建样本间的相似度矩阵, 在数据平衡的 Extended Yale Database B 数据集上, 相比 4 种基准子空间聚类算法获得了最好的聚类效果.

$$Z = \begin{pmatrix} Z_1^* & & & & \\ & Z_2^* & & & \\ & & Z_3^* & & \\ & & & \ddots & \\ & & & & Z_k^* \end{pmatrix}_{n \times n} \tag{3}$$

2　基于低秩表示系数的样本相似度分析

记 Z 的第 i 列为 $\boldsymbol{z}_i$, 第 i 列中第 j 个元素为 z_{ji}. 由第 1 节可知, 从数据的全局结构角度分析, 当 $|z_{ji}|$ 越大时, 样本 $\boldsymbol{x}_i$ 和 $\boldsymbol{x}_j$ 处于同一子空间的可能性越大, 因此可使用 $|z_{ji}|$ 来表示 $\boldsymbol{x}_i$ 和 $\boldsymbol{x}_j$ 的相似度, 称为基于全局结构的相似度. 然而除了数据全局结构, 低秩表示系数还能够在一定程度上反映数据局部结构. 在第 2.1 节中, 本文从数据局部结构的角度, 分析了由低秩表示得到的基于局部结构的相似度. 此外, 现有低秩表示的工作主要是在数据平衡条件下进行的, 未考虑偏标记数据中常见的数据不平衡问题对低秩表示的影响. 为了将低秩表示用于偏标记数据消歧, 第 2.2 节在第 2.1 节的基础上分析了样本不平衡时低秩表示系数和样本相似度的关系.

2.1　基于数据局部结构的样本相似度分析

现有工作普遍认为低秩表示系数矩阵 Z 仅能反映数据全局结构, 但实际上 Z 也可一定程度上描述数据局部结构. 具体分析如下: 在求解 Z 前, 本文和文献 [9] 一致, 首先利用 L2 范数归一化法将每个样本的表征向量归一化为方向不变的单位向量. 根据式 (1) 求出 Z 后, 可得 $\boldsymbol{x}_i = \sum_{j \in J} z_{ji} \boldsymbol{x}_j$, $J = \{j | 1 \le j \le n\}$. 设有 $a, b \in J$, $\boldsymbol{x}_i$ 与 $\boldsymbol{x}_a$ 的余弦相似度 $\cos(\boldsymbol{x}_i, \boldsymbol{x}_a)$ 和 $\boldsymbol{x}_i$ 与 $\boldsymbol{x}_b$ 的余弦相似度 $\cos(\boldsymbol{x}_i, \boldsymbol{x}_b)$ 之差为

$$\begin{aligned} &\cos(\boldsymbol{x}_i,\boldsymbol{x}_a) - \cos(\boldsymbol{x}_i, \boldsymbol{x}_b) = \\ &\quad \boldsymbol{x}_i \cdot \boldsymbol{x}_a - \boldsymbol{x}_i \cdot \boldsymbol{x}_b = \\ &\quad (z_{ai}\boldsymbol{x}_a + z_{bi}\boldsymbol{x}_b + \widetilde{\boldsymbol{x}}_i) \cdot \boldsymbol{x}_a - \\ &\quad (z_{ai}\boldsymbol{x}_a + z_{bi}\boldsymbol{x}_b + \widetilde{\boldsymbol{x}}_i) \cdot \boldsymbol{x}_b = \\ &\quad [1 - \cos(\boldsymbol{x}_a, \boldsymbol{x}_b)](z_{ai} - z_{bi}) + \widetilde{\boldsymbol{x}}_i \cdot (\boldsymbol{x}_a - \boldsymbol{x}_b) \end{aligned} \tag{4}$$

其中, $\widetilde{\boldsymbol{x}}_i = \sum_{j \in J \setminus \{a,b\}} z_{ji} \boldsymbol{x}_j$. 因为无法凭借先验知识确定 $\widetilde{\boldsymbol{x}}_i$ 的方向, 本文根据最大熵原则设 $\widetilde{\boldsymbol{x}}_i$ 和 $\boldsymbol{x}_a - \boldsymbol{x}_b$ 之间的夹角服从 $[0, \pi]$ 之间的均匀分布, 可得 $\cos(\boldsymbol{x}_i, \boldsymbol{x}_a) > \cos(\boldsymbol{x}_i, \boldsymbol{x}_b)$ 的概率如式 (5):

$$\begin{aligned} &\text{P}[\cos(\boldsymbol{x}_i, \boldsymbol{x}_a) > \cos(\boldsymbol{x}_i, \boldsymbol{x}_b)] = \\ &\quad \frac{1}{\pi} \arccos \left[\frac{(1 - \cos(\boldsymbol{x}_a, \boldsymbol{x}_b))(z_{bi} - z_{ai})}{|\widetilde{\boldsymbol{x}}_i||\boldsymbol{x}_a - \boldsymbol{x}_b|} \right] \end{aligned} \tag{5}$$

可以看出, 当 $z_{ai} > z_{bi}$ 时, $\text{P}[\cos(\boldsymbol{x}_i, \boldsymbol{x}_a) > \cos(\boldsymbol{x}_i, \boldsymbol{x}_b)] > \text{P}[\cos(\boldsymbol{x}_i, \boldsymbol{x}_a) < \cos(\boldsymbol{x}_i, \boldsymbol{x}_b)]$, 且 $z_{ai} - z_{bi}$ 越大, $\boldsymbol{x}_i$ 与 $\boldsymbol{x}_a$ 的余弦相似度越可能大于 $\boldsymbol{x}_i$ 与 $\boldsymbol{x}_b$ 的余弦相似度, $\boldsymbol{x}_a$ 与 $\boldsymbol{x}_i$ 也更可能具有相同的标签, 反之亦然.

综上可知, 在对样本进行 L2 范数归一化预处理的前提下, Z 可反映出数据局部结构. $\boldsymbol{z}_i$ 中越大的系数对应的样本和 $\boldsymbol{x}_i$ 应越相似, 反之相似度应越小, 差异越大. 因此可直接使用 z_{ji} 作为 $\boldsymbol{x}_i$ 和 $\boldsymbol{x}_j$ 基于局部结构的相似度. 当使用式 (2) 求解 Z 时, 可提高基于局部结构的相似度的抗噪声能力.

2.2　数据不平衡时低秩表示系数和样本相似度关系

将基于全局结构、局部结构的相似度进行比较可以发现, 若 $z_{ji} > 0$, 两者都反映 z_{ji} 越大时, $\boldsymbol{x}_i$ 和 $\boldsymbol{x}_j$ 相似度越高; 但当 $z_{ji} < 0$ 时, 前者反映 $|z_{ji}|$ 越大时, $\boldsymbol{x}_i$ 和 $\boldsymbol{x}_j$ 相似度越高, 而后者反映 $|z_{ji}|$ 越大时, $\boldsymbol{x}_i$ 和 $\boldsymbol{x}_j$ 相似度应越低. 理想情况下, 低秩表示可以完美地将不同类别的数据分割在不同的子空间中, 此时应采用基于全局结构的相似度.

但现实中, 由于噪声和离群点的影响, 不同类别的数据很难被完美分割, 此时应根据具体情况综合考虑两种相似度. 当数据不平衡时, 数据集中强势类 (样本偏多的类别) 样本数量远大于弱势类 (样本较少的类别) 样本数量, 这近似于数据集中含有大量噪

声和离群数据的情况. 研究表明这种情况会影响低秩表示系数矩阵 Z 描述数据全局结构的能力, 造成不同类别数据的子空间出现重叠[20]. 因此在这种情况下, 当 $z_{ji} < 0$ 时, 应采用基于数据局部结构的相似度, 即使用 z_{ji} 作为 $\boldsymbol{x}_i$ 和 $\boldsymbol{x}_j$ 相似度, $|z_{ji}|$ 越大时, $\boldsymbol{x}_i$ 和 $\boldsymbol{x}_j$ 相似度越低, 差异越大, 两者不属于同一类别的可能性越高. 第 4.3 节中的实验结果验证了数据不平衡时选择这种方式表示样本相似度的合理性.

3 基于成对约束的数据消歧

现有方法主要利用 "相似样本应具有相同的标签" 这一正约束消歧, 忽略了对 "差异大的样本应具有不同的标签" 这一负约束的利用, 但是综合利用成对约束能更真实地描述样本间的关系, 丰富样本标签的推理线索, 进而更加有效地消歧. 例如图 2 所示, 假设已度量出 $\boldsymbol{x}_1$ 和 $\boldsymbol{x}_2$ 具有非常高的相似度, 且都和 $\boldsymbol{x}_3$ 的相似度非常低, 分别利用正约束、负约束和成对约束可以构造出图 2 (a) ∼ (c) 所示图模型. 结合样本间的约束关系和样本候选标签, 容易看出: 由图 2 (a) 可推理出 $\boldsymbol{x}_1$ 和 $\boldsymbol{x}_2$ 的标签应为克里斯蒂亚诺 · 罗纳尔多, 但无法推理出 $\boldsymbol{x}_3$ 的标签; 由图 2 (b) 无法推理出任一样本的标签; 由图 2 (c) 则可以推理出所有样本的标签.

在求解出样本间的低秩表示系数后, 基于成对约束的消歧过程如图 3 所示, 可分为三步: 1) 建立基于样本间成对约束的有权图; 2) 基于能量函数最小化的样本标签求解; 3) 基于类块标准化准则的标签修正.

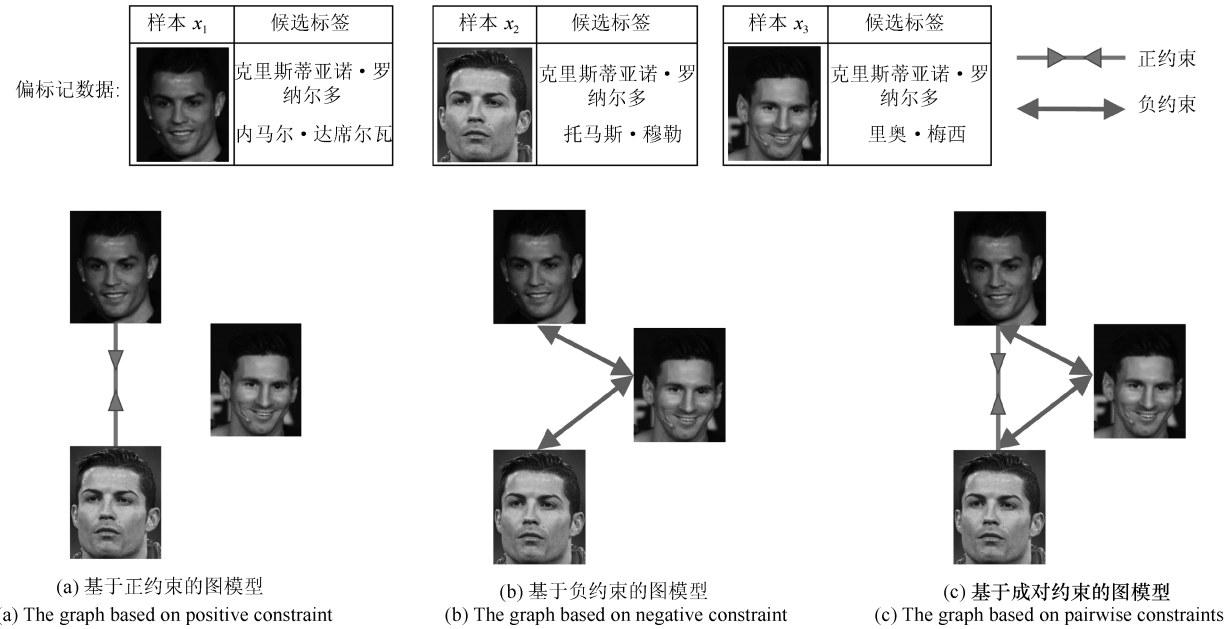


图 2 正负约束作用于消歧的效果

Fig. 2 The effects of positive and negative constraints on disambiguation

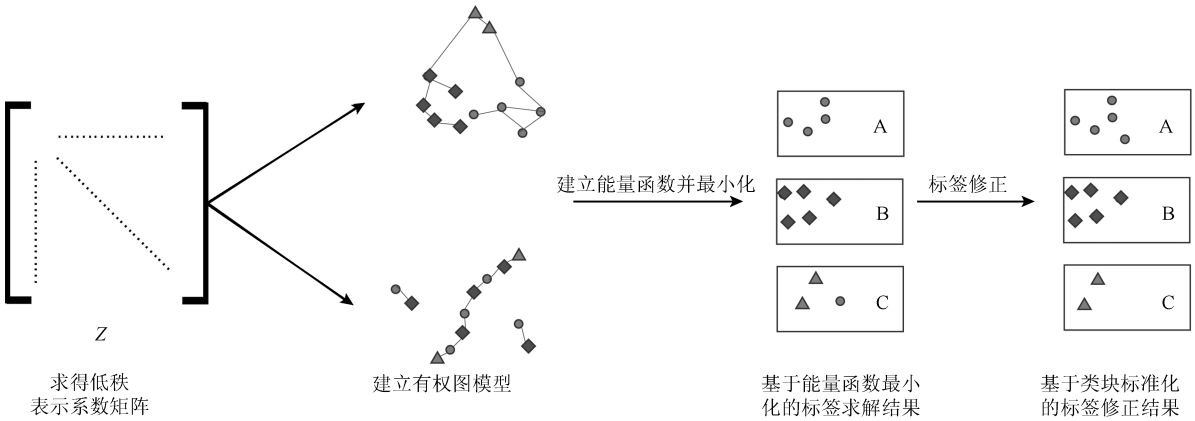


图 3 基于成对约束的偏标记数据消歧算法流程

Fig. 3 The main procedure of PLDPC

3.1 有权图建立

假设有偏标记数据集 $D=\{(\boldsymbol{x}_i,C_i)|1\le i\le n\}$, $X\in\mathbf{R}^{d\times n}$ 为样本数据矩阵, $\boldsymbol{x}_i$ 为 X 第 i 列, 是一个 d 维向量 $(x_{i1},x_{i2},\cdots,x_{id})^{\mathrm{T}}$, 而 $C_i\subseteq C$ 是样本 $\boldsymbol{x}_i$ 关联的候选标签集合, 且 $C=C_1\cup C_2\cup\cdots\cup C_n$. 在对 X 各列进行 L2 范数归一化并根据式 (2) 求得低秩表示系数矩阵 Z 后, 建立两个分别基于正负约束的有权图. 记基于正约束的图为 G_p, 基于负约束的图为 G_n. 在偏标记数据集中, 候选标签集交集为空的两个样本必定以概率 1 具有不同的标签, 因此在 G_p 中, 这样的样本之间不应存在连边. 在 G_n 中, 若要确保两个样本以概率 1 具有不同的标签, 需要在两者间建立权重值为无穷大的边, 这会抑制候选标签集交集不为空的样本间负约束的作用效果. 因此, 若 $C_i\cap C_j=\emptyset$, 本文将 G_p 中 i 至 j 的连边 $G_p(i,j)$ 和 j 至 i 的连边 $G_p(j,i)$, 以及 $G_n(i,j)$ 和 $G_n(j,i)$ 均设为 0.

对于 $C_i\cap C_j\neq\emptyset$ 的情况, 需根据具体情况讨论样本间的连边及权重设置. 根据第 2.2 节可知, $\boldsymbol{z}_i$ 中值越大的系数对应的样本和 $\boldsymbol{x}_i$ 应相似, 反之相似度应越小, 差异越大. 由此, 可将 G_p 和 G_n 分别定义如式 (6) 和式 (7) 所示:

$$G_p(i,j)=\begin{cases}z_{ji}, & \text{若 } z_{ji}>0,\ \boldsymbol{x}_j\in N_k(\boldsymbol{x}_i)\\ 0, & \text{其他}\end{cases}\tag{6}$$

$$G_n(i,j)=\begin{cases}-z_{ji}, & \text{若 } z_{ji}<0,\ \boldsymbol{x}_j\in R_k(\boldsymbol{x}_i)\\ 0, & \text{其他}\end{cases}\tag{7}$$

在式 (6) 中, $N_k(\boldsymbol{x}_i)$ 表示在集合 $\{\boldsymbol{x}_j|C_i\cap C_j\neq\emptyset,1\le j\le n,i\neq j\}$ 中, 和 $\boldsymbol{x}_i$ 最相似的 k 个样本所组成的集合. 式 (6) 的含义为在 G_p 中, 样本 $\boldsymbol{x}_i$ 和最相似的 k 个、且可能是相同标签的样本相连, 边权重设置为相应的表示系数, 并将负权重边和自环断开. 类似地, 在式 (7) 中, $R_k(\boldsymbol{x}_i)$ 表示在集合 $\{\boldsymbol{x}_j|C_i\cap C_j\neq\emptyset,1\le j\le n,i\neq j\}$ 中, 和 $\boldsymbol{x}_i$ 差异应最大的 k 个样本的集合. 式 (7) 的含义为在 G_n 中, 样本 $\boldsymbol{x}_i$ 和差异应最大的 k 个、且可能是相同标签的样本相连, 边权重设置为相应表示系数的相反数, 同时断开负权重边和自环. 为了满足样本间相似度和差异度的对称性, 采用文献 [12, 14] 中的方法, 将 G_P 和 G_R 对称化: $G_P=(G_P+G_P^T)$, $G_R=(G_R+G_R^T)$.

3.2 标签状态求解

设 $\boldsymbol{y}=(y_1,y_2,\cdots,y_n)^{\mathrm{T}}$ 为样本数据矩阵 X 的标签向量, 其中 $y_i\in C$ 为 $\boldsymbol{x}_i$ 的标签. 设 $H=[\boldsymbol{h}_1,\boldsymbol{h}_2,\cdots,\boldsymbol{h}_n]^{\mathrm{T}}$ 为 X 的候选标签矩阵, 其中 $\boldsymbol{h}_i=(h_{i1},h_{i2},\cdots,h_{ic},\cdots,h_{im})^{\mathrm{T}}\in\mathbf{R}^m$ 为 $\boldsymbol{x}_i$ 的候选标签向量, m 为偏标记数据集 D 中所有候选标签的类别数. $h_{ic}=1$ 表示 $\boldsymbol{x}_i$ 的候选标签集中含有 c, 反之 $h_{ic}=0$. $F=[\boldsymbol{f}_1,\boldsymbol{f}_2,\cdots,\boldsymbol{f}_n]^{\mathrm{T}}$ 为 X 的标签状态矩阵, 其中 $\boldsymbol{f}_i=(f_{i1},f_{i2},\cdots,f_{ic},\cdots,f_{im})^{\mathrm{T}}\in\mathbf{R}^m$ 为 $\boldsymbol{x}_i$ 的标签状态, f_{ic} 为 $\boldsymbol{x}_i$ 的标签为 c 的概率, 即 $f_{ic}=\mathrm{P}(y_i=c|\boldsymbol{x}_i)$, $\sum_{c\in C}f_{ic}=1$.

根据相似样本应具有相同标签, 而差异较大的样本之间应具有不同标签的假设, 可根据式 (8) 求解 F:

$$\begin{aligned}&\arg_F\min Q(F)\\ &Q(F)=\frac{1}{nk}\times\frac{1}{2}\sum_{i,j=1}^{n}G_p(i,j)|\boldsymbol{f}_i-\boldsymbol{f}_j|^2-\\ &\frac{\alpha}{nk}\times\frac{1}{2}\sum_{i,j=1}^{n}G_n(i,j)|\boldsymbol{f}_i-\boldsymbol{f}_j|^2+\\ &\frac{\beta}{n}\times\frac{1}{2}\sum_{i=1}^{n}|\boldsymbol{f}_i-\boldsymbol{p}_i|^2\end{aligned}\tag{8}$$

其中, $Q(F)$ 为能量函数, 参数 α, β 用于调节不同能量项的权重. $Q(F)$ 中第一项为 G_p 上的平均正约束势能. 可以看出, 当 G_p 上权重越大的边连接的样本的标签状态越一致时, 平均正约束势能越小. 因为 G_p 中至多存在 nk 个非零元素, 所以选用 $1/nk$ 作为平均系数. 第二项为 G_n 上的平均负约束势能, 可以看出, G_n 上权重越大的边连接的样本的标签状态差异越大时, 平均负约束势能越小. 和正约束势能一样, 选用 $1/nk$ 作为负约束势能的平均系数. 鉴于平均正约束势能和平均负约束势能形式相似, 将其合并为 $Q_G(F)$, 如式 (9) 所示:

$$\begin{aligned}Q_G(F)=&\frac{1}{nk}\times\frac{1}{2}\sum_{i,j=1}^{n}G(i,j)|\boldsymbol{f}_i-\boldsymbol{f}_j|^2=\\ &\frac{1}{nk}\times\mathrm{tr}(F^{\mathrm{T}}(\Delta-G)F)=\\ &\frac{1}{nk}\times\mathrm{tr}(F^{\mathrm{T}}LF)\end{aligned}\tag{9}$$

其中, $G=G_p-\alpha G_n$, $\Delta\in\mathbf{R}^{n\times n}$ 为对角矩阵, $\Delta_{ii}=\sum_{j=1}^{n}G(i,j)$. $L=\Delta-G$ 称为拉普拉斯矩阵, $\mathrm{tr}(\cdot)$ 表示求迹. $Q(F)$ 中第三项为 F 和初始标签状态矩阵 $P=[\boldsymbol{p}_1,\boldsymbol{p}_2,\cdots,\boldsymbol{p}_n]^{\mathrm{T}}$ 之间的差异能量, 其中 $\boldsymbol{p}_i=(p_{i1},p_{i2},\cdots,p_{ic},\cdots,p_{im})^{\mathrm{T}}\in\mathbf{R}^m$ 为 $\boldsymbol{x}_i$ 的初始标签状态. 第三项的目的是防止 F 过于偏离 P. P 的定义如式 (10), $\|\cdot\|_0$ 表示向量的 L0 范数.

$$p_{ic}=\begin{cases}\dfrac{1}{||\boldsymbol{h}_i||_0}, & \text{若 } h_{ic}=1\\ 0, & \text{其他}\end{cases}\tag{10}$$

结合式 (8) ∼ 式 (10), 式 (8) 可以被重写为式 (11):

$$Q(F)=\frac{1}{nk}\mathrm{tr}(F^{\mathrm{T}}LF)+\frac{\beta}{2n}\mathrm{tr}((F-P)(F-P)^{\mathrm{T}}) \tag{11}$$

为了求解式 (8), 求 $Q(F)$ 对 F 的导数如式 (12):

$$\frac{\mathrm{d}(Q(F))}{\mathrm{d}F}=\left(\frac{L+L^{\mathrm{T}}}{nk}+\frac{\beta}{n}I_n\right)F-\frac{\beta}{n}P \tag{12}$$

令 $\frac{\mathrm{d}(Q(F))}{\mathrm{d}F}=0$ 可得:

$$F=\beta k(L+L^{\mathrm{T}}+\beta kI_n)^{-1}P \tag{13}$$

其中, I_n 为 n 维单位矩阵.

3.3 标签修正

根据类块标准化准则, 消岐后不同类别样本的占比在一定程度上应接近类别先验分布. 因此在求得 F 后, 还需根据该准则, 利用候选标签矩阵 H 和初始标签状态矩阵 P 对 F 修正, 得到修正后的标签状态矩阵 $\widetilde{F}$. 首先根据式 (14), 将 F 中非候选标签的概率值置 0:

$$f_{ic}=\begin{cases}f_{ic}, & \text{若 } h_{ic}=1\\ 0, & \text{其他}\end{cases} \tag{14}$$

之后, 对 F 按行归一化, 使概率分布的和为 1, 具体为令 $f_{ic}=f_{ic}/\sum_{c'\in C_i}f_{ic'}$. 然后, 利用式 (15) 使不同类别样本的占比在一定程度上接近类别先验分布. 其中, 参数 μ 用于控制这种接近程度. 最终, 在得到 $\widetilde{F}$ 后, 根据消岐规则 $y_i=\arg_{c\in C_i}\max\widetilde{f}_{ic}$ 获得样本的标签.

$$\widetilde{f}_{ic}=\left(1+\mu\frac{n_c}{\hat{n}_c}\right)f_{ic} \tag{15}$$

其中, $n_c=\sum_{1\le i\le n}p_{ic}$, 即 P 第 c 列元素之和, $\hat{n}_c=\sum_{1\le i\le n}f_{ic}$, 即 F 第 c 列元素之和.

算法 1 总结了 PLDPC 算法的具体流程:

算法 1. PLDPC 算法流程

输入. 数据矩阵 X, 候选标签矩阵 H, 参数 k, α, β, μ, λ.

输出. 标签向量 $\boldsymbol{y}=(y_1,y_2,\cdots,y_n)^{\mathrm{T}}$.

1) 对 X 各列进行 L2 范数归一化.

2) 根据式 (2) 求得样本间的低秩表示系数矩阵 Z.

3) 根据式 (6) 和 (7) 建立有权图 G_p 和 G_n, 并对称化.

4) 根据式 (10) 建立初始标签状态矩阵 P.

5) 将标签状态矩阵 F 设为未知量, 根据式 (8), 基于 G_p, G_n 以及 P 建立能量函数.

6) 根据式 (11) ∼ (13) 最小化能量函数求得标签状态矩阵 F.

7) 根据式 (14) 将 F 中非候选标签上的概率值清零, 并将 F 按行进行归一化, 使概率分布之和为 1.

8) 基于类块标准化准则, 利用式 (15) 对标签矩阵进行进一步修正, 利用 μ 调节消岐后不同类别标签的样本比例接近于类别先验分布的程度, 求得修正后的标签状态矩阵 $\widetilde{F}$.

9) 根据如下方式求得样本的标签:

for $i=1:n$ do

$y_i=\arg_{c\in C_i}\max\widetilde{f}_{ic}$

end for

4 实验

4.1 实验设置

实验所采用的偏标记数据集包括人脸自动标注领域中的 Lost[8]、Soccer Player[21] 和 Yahoo!News[22] 数据集, 目标分类领域中的 MSRCv2[23] 数据集, 以及鸟鸣分类领域中的数据集 BirdSong[24]. 表 1 总结了每个数据集的具体信息. 图 4 展示了各个数据集中每个类别的样本数量分布. 从图表中可以看出, 所有的数据集都具有较高的特征维度, 并存在严重的数据不平衡现象.

本文实验是在 Ubuntu14.04 系统中的 MATLAB 2016a 中进行的, 计算机的配置为: 型号为 i7 6700 K 的 CPU, 16 GB 内存. 在实验中, 式 (2) 中的平衡参数 λ 根据经验取 0.05.

表 1 数据集信息

Table 1 The information of datasets

数据集	样本数量	特征维度	类别数量	平均候选标签数量	领域
Lost	1 122	108	16	2.23	人脸自动标注
MSRCV2	1 758	48	23	3.16	目标分类
BirdSong	4 998	38	13	2.18	鸟鸣分类
Soccer Player	17 472	279	171	2.09	人脸自动标注
Yahoo!News	22 991	163	219	1.91	人脸自动标注

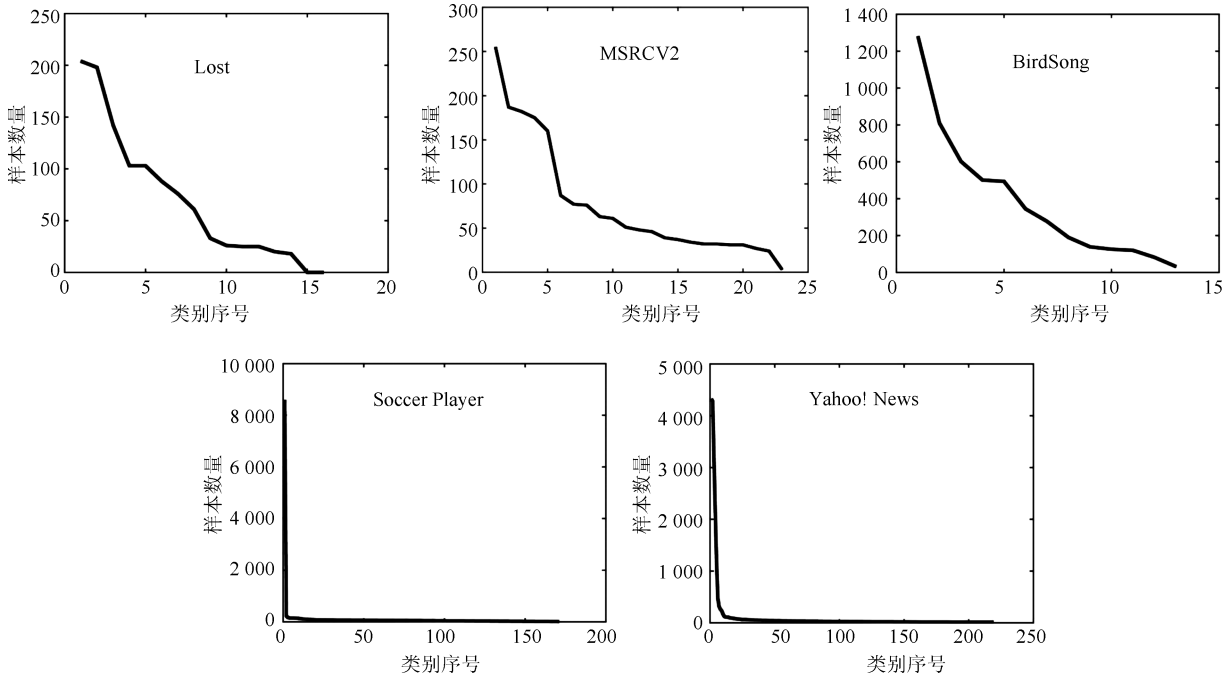


图 4　数据集中不同类别样本数量分布

Fig. 4　The distributions of different categories′ sample number in datasets

在第 4.2 节中, 介绍了 PLDPC 中关键参数的设置. 在第 4.3 节中, 为了验证第 2.2 节中数据不平衡时低秩表示系数表示样本相似度方式的合理性, 以及第 3 节中同时采用正负两种约束进行消歧的合理性, 将 PLDPC 和如下的三种变体进行比较:

PLDPC-abs: 流程和 PLDPC 基本一致, 区别在于在利用式 (6) 和式 (7) 建立有权图之前, 对 Z 进行取绝对值处理, 即采用基于数据全局结构的相似度构建图模型.

PLDPC-p: 流程和 PLDPC 基本一致, 区别在于在利用式 (6) 和式 (7) 建立有权图之前, 将 Z 中的负值置为 0, 即仅利用正约束.

PLDPC-n: 流程和 PLDPC 基本一致, 区别在于在利用式 (6) 和式 (7) 建立有权图之前, 将 Z 中的正值置为 0, 即仅利用负约束.

为了进一步验证 PLDPC 的有效性, 将其和 4 种偏标记消歧基准算法进行对比, 每个算法均使用原始文献中的推荐参数. 这 4 种基准算法的简介如下:

PL-KNN (Partial label learning based on k-nearest neighbors)[7]: 提出一种基于图模型的消歧算法, 每个偏标记样本的标签根据其 k 个最近邻样本的候选标签集加权投票确定 (推荐参数: 近邻样本数 $k = 10$).

MMS (Maximum margin set learning)[25]: 提出一种基于辨识的消歧算法. 在建立了一个线性分类模型后, 通过最大化模型在样本候选标签上和非候选标签上的置信度差异优化模型参数 (推荐参数: 模型结构风险项惩罚系数 $\lambda = 1/n$, n 为偏标记样本总数)

IPAL (Instance-based partial label learning)[9]: 提出一种基于图模型的消歧算法, 在确定样本的初始标签状态后, 采用迭代的标签传播算法求得样本的标签 (推荐参数: 近邻样本数 $k = 10$, 最高迭代次数 $T = 100$).

PL-LEAF (Partial label learning via feature-aware disambiguation)[26]: 提出一种融合方法, 首先通过分析样本在特征空间的流形结构获得初始的标签状态, 之后训练一个多类回归模型对样本的标签进行进一步确定 (推荐参数: 近邻样本数 $k = 10$, 回归模型: 支持向量回归模型).

4.2　参数设置

PLDPC 在消歧的过程中应用了如下 4 个参数: α, β, k 和 μ. 图 5 展示了在控制三个参数不变时, PLDPC 消歧准确率随另一参数变化趋势, 可以看出:

所有数据集最初随着 α 的增加而逐步提高, 说明成对约束的综合使用有助于消歧. 准确率在 0.5 附近取得最高值, 但在 0.6 之后几乎所有的数据集的准确率呈下降趋势, 说明在消歧过程中, 正约束的作用效果要强于负约束, 过多地依赖负约束进行消歧会降低消歧准确率. 为了平衡各个数据集的准确率, 将 α 设为 0.5.

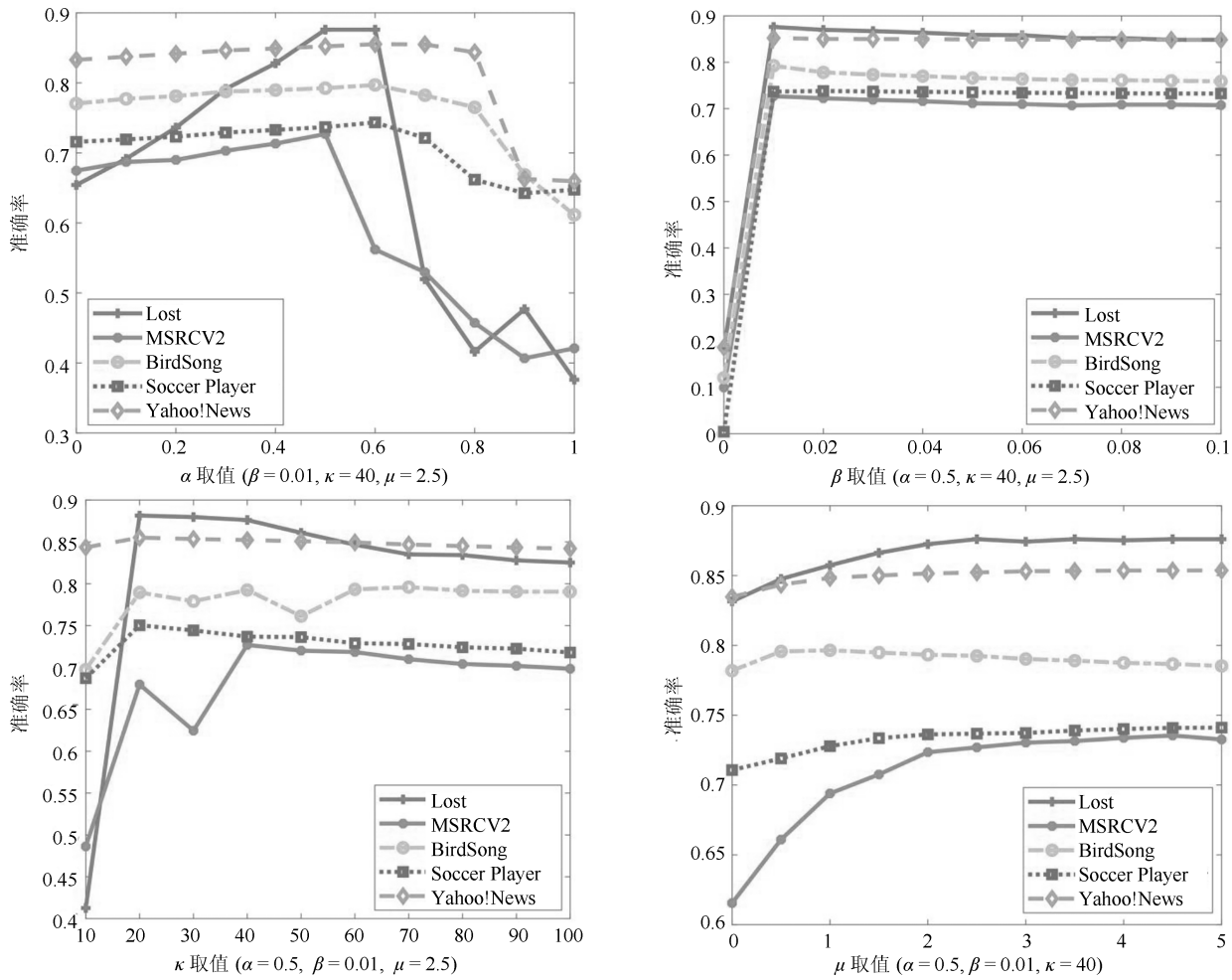


图 5 PLDPC 消歧准确率随不同参数的变化趋势

Fig. 5 The accuracy of disambiguation changes as different parameters varying

所有数据集在 $\beta = 0.01$ 附近取得最高的准确率, 之后随着 β 的增加而小幅度下降, 说明相比样本初始标签状态, 样本间成对约束在消歧过程中发挥主要作用. 因此将 β 设为 0.01.

所有数据集在 $k = 40$ 时取得较高准确率, 当到达 40 后, 各个数据集准确率有小幅波动, 但基本稳定, 为了平衡各个数据集的准确率, 将 k 设为 40.

各个数据集的准确率在 $\mu = 2.5$ 后基本稳定. 但不同数据集取得最高准确率时的 μ 值仍略有不同, 说明不同数据集中不同类别的样本数量比例和类别先验分布之间的近似程度略有不同. 为了平衡各数据集的准确率, 将 μ 设置为 2.5.

4.3 对比实验结果及分析

在对比实验中, 本文同时测试了消歧准确率和处理时间两项指标, 结果分别如表 2、表 3 所示. 在表 2 中粗体表示相应数据集上最好的消歧准确率. 可以看出: PLDPC-p 相对于 PLDPC-abs 取得了较高的准确率, 说明当数据不平衡时, 基于局部结构的相似度更适合度量样本间的相似度, 但 PLDPC-abs 利用了基于全局结构的相似度构建基于正约束的图模型, 导致大量相似度低的样本相连, 从而降低了消歧准确率. 这验证了第 2.2 节中数据不平衡条件下低秩表示系数表示样本相似度方式的合理性. PLDPC 和 PLDPC-p、PLDPC-n 相比准确率更高, 说明相对于正约束、负约束的单独利用, 成对约束的利用能显著提高消歧准确率.

从 PLDPC 和其他 4 种对比算法的结果中可以看出: 1) 相对于同样利用了图模型的 PL-KNN, IPAL, 和 PL-LEAF 算法, PLDPC 在所有的数据集上都优于这三种基准算法, 说明综合利用成对约束信息有助于消歧准确率的提升; 2) 消歧准确率方面, PLDPC 在 5 个数据集上的平均消歧准确率相对基准算法提高了 2.9 %～14.9 %. 具体地, PLDPC 在 3 个数据集上的结果都是最优的, 在其余两个数据集上排名第二, 逊于 MMS 算法, 但在这两个数据集上

表 2 各算法消歧准确率 (%)

Table 2 The disambiguation accuracy of each algorithm (%)

算法	Lost	MSRCV2	BirdSong	Soccer Player	Yahoo!News
PLDPC-abs	57.93	65.81	76.73	70.28	82.03
PLDPC-p	65.81	67.46	77.05	71.58	83.29
PLDPC-n	41.98	37.26	62.30	62.84	57.17
PL-KNN	64.53	58.25	70.99	57.76	72.32
IPAL	77.54	71.44	76.61	67.35	82.37
PL-LEAF	79.32	66.67	75.55	70.50	82.90
MMS	**91.71**	68.27	66.47	70.03	**87.32**
PLDPC	87.61	**72.70**	**79.25**	**73.68**	85.22

表 3 各算法消歧处理时间 (秒 (s)、分钟 (min) 、天 (d))

Table 3 The processing time of each algorithm (second (s), minute (min), day (d))

算法	Lost	MSRCV2	BirdSong	Soccer Player	Yahoo!News
PLDPC-abs	2.13 s	2.39 s	11.62 s	4 min	6 min
PLDPC-p	2.05 s	2.30 s	11.07 s	4 min	6 min
PLDPC-n	2.12 s	2.69 s	11.71 s	4 min	6 min
PL-KNN	0.06 s	0.08 s	0.10 s	59.27	69.78 s
IPAL	0.51 s	0.63 s	1.56 s	73.75 s	94.62 s
PL-LEAF	56.04 s	4 min	35 min	>1 d	>1 d
MMS	57.02 s	1 min	2 min	34 min	35min
PLDPC	2.16 s	2.45 s	11.61 s	4 min	6 min

MMS 仅比 PLDPC 平均高出 3.1 %, 而在 PLDPC 获胜的 3 个数据集上, PLDPC 的准确率平均高出 MMS 约 7 %, 这说明 PLDPC 在准确率上的平均表现是最佳的; 3) 虽然 PLDPC 的消歧处理效率不是最高的 (处于对比算法的中间水平) 但 PLDPC 相比于速度较快的算法, 具有最高的消歧准确率, 而且相对于消歧准确率相当的 MMS 算法有明显的效率优势. 所以, PLDPC 能够在保证效率的情况下具有较高的消歧准确率, 适用于对消歧准确率要求较高的应用场景.

5 结语

基于图模型的偏标记数据消歧是近年来的研究热点, 而采用何种方式构建图模型是该问题的关键. 低秩表示作为一种效果优异的子空间分割算法, 在基于图模型的聚类、半监督学习等领域得到了广泛应用. 然而这些领域在利用低秩表示来分析样本间的相似度时, 未考虑数据不平衡对低秩表示的影响, 以及数据不平衡时低秩表示系数表示样本间相似度的合理方式. 而在偏标记数据中, 数据不平衡是一种普遍存在的问题, 因此本文详细研究了数据不平衡时低秩表示系数表示样本相似度的合理方式, 并在实验中验证了研究结论的正确性. 此外, 针对现有方法仅利用样本间正约束, 忽略了负约束的问题, 本文综合利用正负约束来设计图模型, 并通过最小化基于图模型的能量函数求解出样本的标签. 实验结果表明, 相比 PL-KNN、IPAL、PL-LEAF 等基准算法, 本文方法在所有数据集上都有更高的准确率; 相比 MMS 算法, 本文方法尽管在 Lost、Yahoo!News 两个数据集上准确率略低, 但平均消歧准确率优于 MMS, 且效率平均提高了约 10 倍, 说明本文方法能够在保证效率的情况下具有更高的消歧准确率, 适用于对消歧准确率要求较高的应用场景.

References

1 Su X P, Peng J Y, Feng X Y, Wu J. Labeling faces with names based on the name semantic network. *Multimedia Tools and Applications*, 2016, **75**(11): 6445−6462

2 Jin R, Ghahramani Z. Learning with multiple labels. In: Proceedings of the 15th International Conference on Neu-

ral Information Processing Systems. Cambridge, MA: MIT Press, 2002. 921−928

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

4 Zhang Min-Ling. Research on partial label learning. *Journal of Data Acquisition and Processing*, 2015, **30**(1): 77−87
(张敏灵. 偏标记学习研究综述. 数据采集与处理, 2015, **30**(1): 77−87)

5 Nguyen N, Caruana R. Classification with partial labels. In: Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Las Vegas, Nevada, USA: ACM, 2008: 551−559

6 Chen Y C, Patel V M, Chellappa R, Phillips P J. Ambiguously labeled learning using dictionaries. *IEEE Transactions on Information Forensics and Security*, 2014, **9**(12): 2076−2088

7 Hüllermeier E, Beringer J. Learning from ambiguously labeled examples. In: Proceedings of the 6th International Symposium on Intelligent Data Analysis. Madrid, Spain: Springer, 2005. 168−179

8 Cour T, Sapp B, Taskar B. Learning from partial labels. *Journal of Machine Learning Research*, 2011, **12**: 1501−1536

9 Zhang M L, Yu F. Solving the partial label learning problem: an instance-based approach. In: Proceedings of the 24th International Conference on Artificial Intelligence. Buenos Aires, Argentina: AAAI Press, 2015. 4048−4054

10 Yu F, Zhang M L. Maximum margin partial label learning. *Machine Learning*, 2017, **106**(4): 573−593

11 Olson C C, Judd K P, Nichols J M. Manifold learning techniques for unsupervised anomaly detection. *Expert Systems with Applications*, 2018, **91**: 374−385

12 Liu G C, Lin Z C, Yu Y. Robust subspace segmentation by low-rank representation. In: Proceedings of the 27th International Conference on Machine Learning. Haifa, Israel: ICML, 2010. 663−670

13 Wang Wei-Wei, Li Xiao-Ping, Feng Xiang-Chu, Wang Si-Qi. A survey on sparse subspace clustering. *Acta Automatica Sinica*, 2015, **41**(8): 1373−1384
(王卫卫, 李小平, 冯象初, 王斯琪. 稀疏子空间聚类综述. 自动化学报, 2015, **41**(8): 1373−1384)

14 Li Bo, Lu Chun-Yuan, Leng Cheng-Cai, Jin Lian-Bao. Robust low rank subspace clustering based on local graph Laplace constraint. *Acta Automatica Sinica*, 2015, **41**(11): 1971−1980
(李波, 卢春园, 冷成财, 金连宝. 基于局部图拉普拉斯约束的鲁棒低秩表示聚类方法. 自动化学报, 2015, **41**(11): 1971−1980)

15 Hou X, Yao G J, Wang J. Semi-supervised classification based on low rank representation. *Algorithms*, 2016, **9**(3): Article No. 48

16 Pasteris S, Vitale F, Gentile C, Herbster M. On pairwise clustering with side information [Online], available http://arxiv.org/abs/1706.06474, December 9, 2017

17 Xu Ming-Liang, Wang Shi-Tong, Hang Wen-Long. A semi-supervised affinity propagation clustering method with homogeneity constraint. *Acta Automatica Sinica*, 2016, **42**(2): 255−269
(徐明亮, 王士同, 杭文龙. 一种基于同类约束的半监督近邻反射传播聚类方法. 自动化学报, 2016, **42**(2): 255−269)

18 Zhu X J, Ghahramani Z, Lafferty J. Semi-supervised learning using Gaussian fields and harmonic functions. In: Proceedings of the Twentieth International Conference on Machine Learning. Washington DC, USA: ICML, 2003. 912−919

19 Zhu X J, Goldberg A B. *Introduction to Semi-Supervised Learning*. San Rafael: Morgan and Claypool Publishers, 2009. 1−130

20 You Cong-Zhe. Novel Subspace Clustering Algorithms and Applications [Ph. D. dissertation], Jiangnan University, China, 2017
(由从哲. 子空间聚类分析新算法及应用研究 [博士学位论文], 江南大学, 中国, 2017)

21 Zeng Z N, Xiao S J, Jia K, Chan T H, Gao S H, Xu D, et al. Learning by associating ambiguously labeled images. In: Proceedings of the 2013 IEEE Computer Vision and Pattern Recognition. Portland, OR, USA: IEEE, 2013. 708−715

22 Guillaumin M, Verbeek J, Schmid C. Multiple instance metric learning from automatically labeled bags of faces. In: Proceedings of the 11th European Conference on Computer Vision. Heraklion, Crete, Greece: Springe, 2010. 634−647

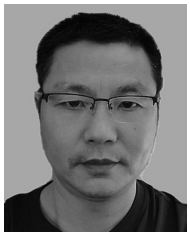
23 Liu L P, Dietterich T G. A conditional multinomial mixture model for superset label learning. In: Proceedings of the 25th International Conference on Neural Information Processing Systems. Lake Tahoe, Nevada: Curran Associates Inc., 2012. 548−556

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

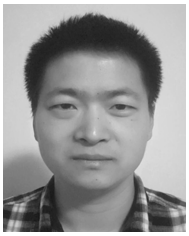
25 Luo J, Orabona F. Learning from candidate labeling sets. In: Proceedings of the 23rd International Conference on Neural Information Processing Systems. Vancouver, British Columbia, Canada: Curran Associates Inc., 2010. 1504−1512

26 Zhang M L, Zhou B B, Liu X Y. Partial label learning via feature-aware disambiguation. In: Proceedings of the 22th International Conference on Knowledge Discovery and Data Mining. San Francisco, California, USA: ACM, 2016. 1335−1344

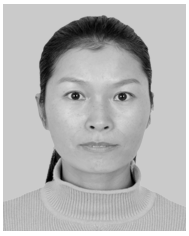
征 察 国家数字交换系统工程技术研究中心硕士研究生. 主要研究方向为机器学习, 计算机视觉. 本文通信作者.
E-mail: zcpi31415926@163.com
(**ZHENG Cha** Master student at the China National Digital Switching System Engineering and Technological Research and Development Center. His research interest covers machine learning and computer vision. Corresponding author of this paper.)



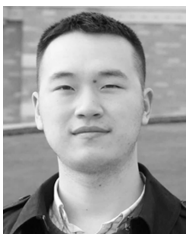
吉立新 国家数字交换系统工程技术研究中心研究员. 主要研究方向为电信网信息关防, 信息安全.
E-mail: jlx@ndsc.com.cn
(**JI Li-Xin** Professor at the China National Digital Switching System Engineering and Technological Research and Development Center. His research interest covers telecom network information gateway, and information security.)



高 超 国家数字交换系统工程技术研究中心助理研究员. 主要研究方向为计算机视觉. E-mail: chaosndsc@163.com
(**GAO Chao** Assistant professor at the China National Digital Switching System Engineering and Technological Research and Development Center. His main research interest is computer vision.)



李邵梅 国家数字交换系统工程技术研究中心副研究员. 主要研究方向为计算机视觉.
E-mail: lishaomei_may@126.com
(**LI Shao-Mei** Associate professor at the China National Digital Switching System Engineering and Technological Research and Development Center. Her main research interest is computer vision.)



吴翼腾 国家数字交换系统工程技术研究中心博士研究生. 主要研究方向为网络大数据分析.
E-mail: wuyiteng1992@163.com
(**WU Yi-Teng** Ph. D. candidate at the China National Digital Switching System Engineering and Technological Research and Development Center. His main research interest is network big data analysis.)