基于概率推断的质量控制智能体

徐耀丽 李战怀

(西北工业大学计算机学院 西安 710072)

(西北工业大学大数据存储与管理工业和信息化部重点实验室 西安 710129)

摘 要 实体解析(Entity Resolution,ER)是数据集成和清洗领域的基础问题,而不一致性消歧(Inconsistency Reconciliation,IR)通过对现存的不同ER算法产生的不一致记录对进行消歧,进一步提升解析效果。但是现有的IR方法有一个局限,即消歧结果没有质量保障。对此,首次提出了一个基于概率推断的质量控制智能体,记为QCAgent。该智能体不需要训练数据集,能够在满足给定查准率的约束条件下输出查全率最大的消歧结果。它的核心思想是:首先,使用异常点检测模型来估算不一致记录对匹配的概率,并依据这些概率估算查准率和查全率,再将计算出的查准率和查全率作为环境端的反馈;其次,使用二分搜索算法,选择满足查准率要求且查全率最大的翻转方案,作为QCAgent的下一次行动;然后,用更新后的一致结果训练异常点模型,并估算查准率和查全率。按此循环,当新估计的查准率满足约束条件时,该迭代过程停止。在真实的数据集上,实验结果表明:QCAgent能够有效解决消歧结果的质量控制问题。

关键词 质量控制,实体解析,不一致性消歧,智能体,查准率

中图法分类号 TP391 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2019.04.002

Quality Control Agent Based on Probability Inference

XU Yao-li LI Zhan-huai

(School of Computer Science and Engineering,Northwestern Polytechnical University,Xi'an 710072,China)

(Key Laboratory of Big Data Storage and Management,Northwestern Polytechnical University,Ministry of Industry and Information Technology,Xi'an 710129,China)

Abstract Entity resolution (ER) is the fundamental problem of data integration and cleaning,while inconsistency reconciliation(IR) further improves the resolution performance through reconciling inconsistent pairs resolved by existing diverse ER approaches. However,previous IR approaches have a limitation that the reconciliation solution has no quality guarantee. To solve this problem,this paper firstly proposed a quality control agent based on probability inference,denoted as QCAgent. QCAgent does not require any manually labeled pair,and can automatically output reconciliation result with the highest recall on the premise of satisfying the given precision threshold. Its core idea is as follows. Firstly,the outlier detection model is utilized to estimate the matching probability for each inconsistent pair,and then the estimated precision and recall are regarded as the environmental feedback according to these probabilities. Next,the binary search algorithm is used to select a flipping solution as the next action of QCAgent,which can make flipped reconciliation result satisfy the precision requirement with the highest recall. Then the outlier detection model is retrained by using the new consistent pairs,and the recall and precision of flipped reconciliation result are estimated. The iterative process terminates until the newest estimated precision meets the constraints. On the real data set,the experimental results show that QCAgent can effectively solve the quality control problem of reconciliation result.

Keywords Quality control,Entity resolution,Inconsistency reconciliation,Agent,Precision

1 引言

实体解析(ER)也称重复记录检测,是指从数据集中推理出描述现实世界同一个实体的记录对。它是数据集成和清洗系统的基础问题,因此被广泛研究,并有大量的模型或算法[1-16,29]被提出。这些算法可分为无监督的(如基于距离的[3,10,14,16]、基于规则的[5,9]和基于图理论的[11])和有监督的(如基于深度学习模型的[13,15]和基于主动学习的[7-8,12])。这

到稿日期:2018-12-04 返修日期:2019-01-26 本文受中国科技部国家重点研发计划(2016YFB1000703),国家自然科学基金重点项目(61732014,61332006),国家自然科学基金面上项目(61472321,61672432),国家自然科学基金青年项目(61502390),陕西省自然科学基础研究计划(2018JM6086),西北工业大学中央高校基本科研业务费项目(3102017jg02002)资助。

徐耀丽(1987-),女,博士生,CCF学生会员,主要研究方向为数据修复、实体解析和不一致性消歧;李战怀(1961-),男,博士,主要研究方向为数据管理和数据质量,E-mail:lizhh@nwpu.edu.cn(通信作者)。

些算法各有优缺点，例如：基于规则的方法能充分利用领域专家的先验知识；基于距离的方法能有效地解析某类数据集，尤其当距离度量能很好地区分数据集中匹配或不匹配的记录对时；基于图理论的方法能分析记录对之间的相似度，以及 token 和记录之间的关系。有监督的方法则可以充分利用机器学习、深度学习和主动学习等技术进行实体解析。然而，对于一个给定的实体解析任务，不同模型的输出结果可能包含大量的不一致记录对集合C^{inp}、少量的一致匹配记录对集合C^{cpp}和大量的一致不匹配记录对集合C^{cnp}，其中 inp 表示不一致，cpp 表示一致且匹配，cnp 表示一致且不匹配。随着模型数目的增多，一致记录对的数目减少，而不一致记录对的数目增多。为了充分利用现有的解析模型，用户必然面临如何有效地解决这些不一致记录对的问题。已有研究者[1]提出了针对不一致记录对的消歧算法，但该算法只提供了尽力而为的消歧结果，并没有给出任何质量保证。本文研究的问题是如何对消歧结果进行概率推断，估算出查准率和查全率，挑选出错误的可能性高的记录对并修正，使得解析结果能满足用户给定的查准率要求，且最大化查全率。

自 DeepMind 公司的 AlphaGo 战胜李世石和柯洁后，深度强化学习便得到学术界和工业界的广泛关注。部分学者[17]提出了基于深度强化学习的可视化问答对话智能体。受深度强化学习的启发，本文把质量控制过程建模为一个交互式的强化学习过程。其核心思想是把质量控制模块看作环境，把不一致记录对中误判结果的纠错操作看作一个智能体对环境的反馈。经过多轮交互后，消歧结果的查准率逐渐提升。环境的质量估计模块和智能体的翻转策略选择模块是基于异常点检测的概率推理。

本文的创新点如下：1）弥补了现有消歧框架的不足，新增了满足查准率约束且最大化查全率的消歧框架；2）首次提出了一个类强化学习的基于概率推断的质量控制智能体；3）在真实的数据集上进行实验，并通过实验结果验证了该方法的有效性和可扩展性。

本文第 2 节介绍了与本研究相关的工作；第 3 节形式化地定义了研究问题；第 4 节描述了基于概率推断的质量控制智能体的系统框架，并详细阐述了质量估计模块和翻转策略选择模块；第 5 节验证了算法的有效性和可扩展性；最后总结全文，并探讨了未来值得研究的方向。

2　相关工作

强化学习（Reinforcement Learning，RL）是研究软件智能体如何依据环境的反馈来调整自身的策略，从而获得最大的累积收益。考虑到当问题的状态和动作很多或者它们的值域属于连续空间时，存储所有与状态和动作组合相对应的策略不现实，且策略的搜索成本高，有些学者把强化学习和深度学习相结合，并提出一些策略搜索算法，如 DeepMind 公司提出的 DQN（Deep Q Network）[18]。但强化学习主要应用在计算机博弈、视频游戏、机器人[18]和对话系统[17]等领域[19-20]，在实体解析和不一致性消歧领域还没有得到深入的研究和应用。

现有的实体解析算法和模型可划分为两类：带质量约束的和无质量约束的。大部分的实体解析算法属于无质量约束的，也就是说仅能提供尽力而为的解析结果。这部分算法又可进一步细分为基于距离的[3,10,14,16]、基于规则的[5,9]、基于图的[11]和基于学习的[13,15]。基于距离的解析算法首先使用字符串的相似度（或者距离）量化记录对匹配的可能程度，然后选择合适的阈值将候选记录对划分为两类：匹配和不匹配。这类算法的不足是无法有效地处理某些解析任务，例如在会议的全写和简写记录中识别表示同一个会议的记录对。这类解析任务需要借助文本的上下文或者语义信息。基于规则的方法是依据领域知识或者置信度高的训练样本集，对各类距离度量和阈值进行组合，从而得到匹配规则。基于图理论的最新无监督解析算法综合了 token 的区分能力和记录对的匹配概率来进行实体解析。基于学习的算法则运用了 SVM 或深度学习相关技术，如循环神经网络和注意力策略等。这些算法虽无法满足质量约束的需求，但可作为消歧算法的个体方法。

近年来，具有质量约束保证的实体解析算法[7-8,12,21]得到了学者们的广泛关注。这类算法把实体解析问题建模为一个二分类问题。一般而言，对于给定的实体解析任务，匹配和不匹配的记录对是严重类不平衡的。与传统的准确率相比，查全率和查准率更能有效地评估解析结果的质量。Arasu 等[12]利用查准率的单调性假设，提出了一个满足查准率约束且最大化查全率的主动学习分类器。Bellare 等[7-8]针对 Arasu 提出的算法在最坏情况下需要 $O(n)$ 个标记量的问题（其中 n 是输入样本的数目），提出了一种在某些分布假设下标记量的复杂度为 $O(\log(n))$ 的算法。Chen 等[21]提出了一种满足查准率和查全率约束且最小化人工量的算法。这类算法依赖于人的参与。考虑到人力资源的成本，本文主要研究如何从弱样本集中多次迭代估算记录对的匹配概率和解析结果的质量估计。所谓弱样本集，是指一致记录对集合 C^{cp}。我们把 C^{cp} 看作一个有一定错误率的训练集合。

目前，不一致性消歧问题已得到部分学者的关注，针对该问题，这些学者提出了无标签数据场景下不一致性消歧算法 GL-RF[1]。与 GL-RF 不同的是，本文提出的 QCAgent 能够保证输出的消歧结果满足查准率的约束。

3　问题描述

本节首先定义了实体解析和不一致性消歧，然后引出了不一致性消歧的质量控制问题，并给出了该问题的形式化描述。由于已有工作[22-25]研究了实体解析的 blocking 技术，本文假设实体解析问题的输入是经过 blocking 处理后的候选记录对集合。所谓 blocking 技术，就是通过数据分析过滤掉那些明显不匹配的记录对。

定义 1（实体解析，ER）　给定某实体解析的候选记录对集合 $C=\{c_1,c_2,\cdots,c_n\}$，其中 c_i 代表第 i 个候选记录对，实体解析 ER 就是预测这些记录对是否匹配。每个候选记录对 c_i 由两个记录 r_i 和 r_j 构成。每个记录由 m 个属性描述。如表 1 和表 2 所列，rID 是记录的唯一标识，而 cID 是候选记录对的唯一标识。例如，候选记录对 c_2 由 r_1 和 r_7 构成，且 r_1 和 r_7 由 authors，title 和 address 这 3 个属性组成。其中，“—”表示数据缺失。

表 1 CORA 数据样例

Table 1 Data examples in CORA

rID	authors	title	address
r_1	n. cesa-bianchi, y. freund, d. p. helmbold, and m. warmuth.	on-line prediction and conversion strategies.	oxford, 1994.
…	…	…	…
r_7	cesa-bianchi, n. , freund, y. , helmbold,	on-line prediction and conversion strategies. in proceedings of the first eurocolt workshop.	—

表 2 IR 问题的数据样例

Table 2 Data examples for IR problem

cID	rID	rID	m_1	m_2	m_3	m_4	$\widetilde{S}$
c_1	r_1	r_2	P	P	P	P	P
c_2	r_1	r_7	P	N	P	P	P
c_3	r_2	r_4	P	N	N	P	P
c_4	r_3	r_5	P	P	N	P	N
c_5	r_5	r_6	P	N	N	P	P
c_6	r_6	r_7	N	N	N	N	N

假设 $S=\{s_1,s_2,\cdots,s_n\}$ 为某实体解析任务 C 的解决方案，其中 s_i 表示 c_i 的状态。当 $s_i=P$ 时，表示匹配，反之则表示不匹配，即 $s_i=N$。$\widetilde{S}=\{\widetilde{s}_1,\widetilde{s}_2,\cdots,\widetilde{s}_n\}$ 是 C 中记录对的真实标签。$C_{tp}=\{c_i\mid s_i=P\wedge\widetilde{s}_i=P\}$ 是 S 中真阳性的记录对集合；$C_{fp}=\{c_i\mid s_i=P\wedge\widetilde{s}_i=N\}$ 是 S 中假阳性的记录对集合。类似可得，$C_{tn}=\{c_i\mid s_i=N\wedge\widetilde{s}_i=N\}$ 是 S 中真阴性的记录对集合，$C_{fn}=\{c_i\mid s_i=N\wedge\widetilde{s}_i=P\}$ 是 S 中假阴性的记录对集合。S 的查准率 *precision* 和查全率 *recall* 公式如下：

$$precision(C,S)=\frac{|C_{tp}|}{|C_{tp}+C_{fp}|} \tag{1}$$

$$recall(C,S)=\frac{|C_{tp}|}{|C_{tp}+C_{fn}|} \tag{2}$$

定义 2（不一致性消歧，IR） 给定某实体解析任务 C 和来自 k 个个体方法的解析结果 $M=\{m_1,m_2,\cdots,m_k\}$，可得到两类记录对：一致记录对 C^{cp} 和不一致记录对 C^{inp}。不一致性消歧是指预测不一致记录对是否匹配。C^{cp} 可进一步划分为两类：一致匹配记录对 C^{cpp} 和一致不匹配记录对 C^{cnp}。

例 1 如表 1 和表 2 所列，C 包括 6 个候选记录对。m_i 列表示第 i 个个体方法的解析结果，而 $\widetilde{S}$ 列表示候选记录对真实的状态。c_1 和 c_6 分别属于 C^{cpp} 和 C^{cnp}，而 c_2-c_5 则属于 C^{inp}。IR 是指预测 C^{inp} 中记录对是否匹配。

ER 和 IR 有两点不同：1）两者解析的记录对不同，ER 处理的是 C，而 IR 处理的是更具有挑战性的不一致记录对集合 C^{inp}，因为 C^{inp} 是存在分歧的记录对的集合，是已有的分析技术可能判断错误的部分；2）IR 可利用解析结果相对可信的 C^{cpp} 和 C^{cnp}。

定义 3（质量控制，QC） 给定 C^{inp} 的解析结果 S 和查准率的阈值 τ，QC 通过识别并纠正 S 中误判的记录对，输出消歧结果 S^*，使得 S^* 的查准率大于 τ 同时查全率最大。它的形式化定义如下：

$$\begin{aligned}&\text{maximize } recall(C^{inp},S^*)\\&\text{subject to } precision(C^{inp},S^*)\geqslant\tau\end{aligned} \tag{3}$$

4 基于概率推断的质量控制智能体

本节首先概述基于概率推断的质量控制智能体（QCAgent）的整个处理流程，然后详细介绍它的两个核心模块，即质量估计模块和翻转策略选择模块。

4.1 系统框架

本文把式(3)定义的最优化问题建模为一个质量控制智能体与环境的交互过程，其处理流程如下。1）利用现存的个体方法集 M 处理 C，并得到 C^{cpp}，C^{cnp} 和 C^{inp}。2）利用消歧方法 m_{IR} 预测 C^{inp} 中的记录对是否匹配，并输出 S^0。3）环境端利用异常点检测模型来估算 C^{inp} 中每个记录对 c_i 的匹配概率，并使用标记信息和概率信息混合的方法来估算查准率 $\hat{p}$ 和查全率 $\hat{r}$；将这些估算的信息（包括查准率、查全率和记录对的匹配概率）作为环境端的反馈信息。4）智能体利用环境端的反馈信息和查准率阈值 τ，搜索出满足查准率约束且查全率最大的翻转策略，并得到第 t 步的消歧结果 S^t。若 $\hat{p}\geqslant\tau$，则整个交互过程停止，最后一轮交互的输出结果就是质量控制 QC 的结果 S^*；若 $\hat{p}<\tau$，则使用智能体的消歧结果 S^t，对环境端的概率推断模块更新训练样本。由于智能体的翻转策略是保留匹配概率高的记录对标签，S^t 的查准率要高于 S^{t-1} 的，因此利用 S^t 的样本可优化概率推断模块。

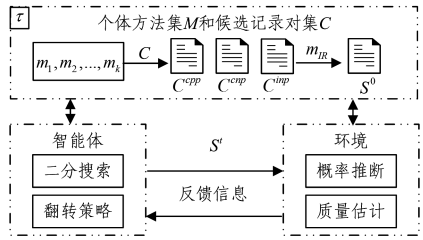

图 1 QCAgent 的系统框架图

Fig. 1 System architecture of QCAgent

4.2 基于概率推断的质量估计

基于概率推断的质量估计是环境端的核心功能。在没有 C^{inp} 的真实标签下，估算一个消歧结果的查准率和查全率是非常具有挑战性的。但在消歧场景下，相比于 C^{inp}，C^{cpp} 存在相当少的误判标签，且隐含了真正匹配记录对的分布特征。另一方面，统计模型的中间输出结果（如类标签的可能性）隐含了该模型对预测结果的置信程度。也就是说，该置信程度可作为估算质量指标的有用信息。注意：当选择概率推断模型时，不要使用与消歧模型的工作原理类似的概率模型作为质量控制部分。其原因是：使用类似的概率模型进行质量估计，更容易输出被高估的质量指标。

鉴于此，本文提出了一个基于概率推断的质量估计算法，该算法包括概率推断部分和质量估计部分。概率推断部分的输入是 C^{cpp} 和 C^{cnp}，输出是不一致记录对的匹配概率；而质量估计部分的输入是概率推断部分的匹配概率，输出是质量指标。

概率推断部分使用了异常点检测算法 OneClassSVM[26] 来推断 C^{inp} 中不一致记录对的匹配概率。它的核心思想是在 p 维空间中，从观测样本集（如 C^{cpp}）中学习到一个非线性软边界。该边界描述的子空间近似逼近观测样本集所隐含的概率分布，其中 p 是样本的特征数目。对于一个新样本点 $c\in C^{inp}$，如果 c 位于该软边界描述的子空间，那么 c 的标签是 P，同时用 c 与该软边界的距离表示该预测正确的可能性。类似

地，如果 c 位于子空间的外部，c 的标签是 N，同时用 c 与该软边界的距离表示预测正确的可能性。在最坏情况下，One-ClassSVM[27]的时间复杂度是 $O(n_f \times \#C^{cpp^3})$，其中 n_f 是记录的属性数目。本文的概率推断模型 Q 将 C^{cpp} 作为观测样本集，将 OneClassSVM 作为概率推断算法，其输出是不一致记录对的匹配概率。该匹配概率是使用 sigmoid 函数来转换记录对与软边界之间的距离而得到的。

若直接用 Q 模型的概率值进行估算，则 $\#C_{tp}^{inp}$ 和 $\#C_{tn}^{inp}$ 的估计值会过低；若用模型的预测标签作为真实的标签，则可能出现高估查准率和查全率的情况。$C_{tp}^{inp}=\{c_i \mid s_i=P \wedge \tilde{s}_i=P, c_i \in C_p^{inp}\}$ 是指消歧标签为 P 且真实标签也为 P 的记录对，其中 s_i 和 $\tilde{s}_i$ 分别是 c_i 的消歧标签和真实标签。$\#C_{tp}^{inp}$ 是指 C_{tp}^{inp} 的元素数目。类似地，C_{tn}^{inp} 是指消歧标签为 N 且真实标签也为 N 的记录对。为了综合利用 Q 模型的概率推断和标签信息，本文设计了一个混合的质量估计算法 QC，如算法 1 所示。它的主要思想是对消歧结果 S 和 Q 的预测标签一致的部分使用概率信息；而对不一致的部分使用标签信息。

假设当前迭代次数为 t，QC 的处理流程如算法 1 所示：1）首先得到 Q^t，以及第 t 次迭代中 C^{inp} 的概率值 P^t 和标签值 L^t，如算法 1 第 1 行和第 2 行所示。Q^t 的输入是 $C_{cp}^{t-1} \cup C^{cpp}$，其中 C_{cp}^{t-1} 是第 $t-1$ 次迭代中，S^{t-1} 和 Q^{t-1} 预测标签一致且标签为 P 的记录对。注意：当 $t=1$ 时，$C_{cp}^{t-1}=\emptyset$。2）依据 S^t 和 L^t，得到 C_{cp}^t，C_{cn}^t 和 $C_{inc}^t=C^t-(C_{cp}^t \cup C_{cn}^t)$，如算法 1 第 3 行所示。$C_{cn}^t$ 是第 t 次迭代中，S^t 和 Q^t 预测标签一致且标签为 N 的记录对。3）依据式(4)估算 C_{cp}^t 中真正匹配的记录对数目 $\#\hat{C}_{tcp}^t$，其中 $p(c)$ 是 Q^t 中 c 的匹配概率，进而可得 C_{cp}^t 中误判的记录对数目 $\#C_{fcp}^t$ 的估计值 $\#\hat{C}_{fcp}^t=\#C_{cp}^t-\#\hat{C}_{tcp}^t$。类似地，依据式(5)，估算 C_{cn}^t 中真正不匹配的记录对数目 $\#\hat{C}_{tcn}^t$，进而可得 C_{cn}^t 中误判的记录对数目 $\#\hat{C}_{fcn}^t$。对于 C_{inc}^t，本文把 Q^t 中 C_{inc}^t 的标签当成真实的标签，依据式(6)，计算 C_{inc}^t 的 $\hat{tp}^t$，$\hat{fp}^t$，$\hat{tn}^t$ 和 $\hat{fn}^t$，其中 $S^t(c)$ 和 $L^t(c)$ 分别是在 S^t 和 L^t 中 c 的相应标签，$I(\cdot)$ 是指示函数。4）最后，依据式(7)计算第 t 轮的查准率和查全率估计值 $\hat{p}^t$ 和 $\hat{r}^t$，如算法 1 第 4 行所示。

$$\#\hat{C}_{tcp}^t=\lceil \sum_{c \in C_{cp}^t} p(c) \rceil \tag{4}$$

$$\#\hat{C}_{tcn}^t=\lceil \sum_{c \in C_{cn}^t} 1-p(c) \rceil \tag{5}$$

$$\begin{aligned}
\hat{tp}^t &= \sum_{c \in C_{inc}^t} I(S^t(c)=P \wedge L^t(c)=P) \\
\hat{fp}^t &= \sum_{c \in C_{inc}^t} I(S^t(c)=P \wedge L^t(c)=N) \\
\hat{tn}^t &= \sum_{c \in C_{inc}^t} I(S^t(c)=N \wedge L^t(c)=N) \\
\hat{fn}^t &= \sum_{c \in C_{inc}^t} I(S^t(c)=N \wedge L^t(c)=P)
\end{aligned} \tag{6}$$

$$\hat{p}^t=\frac{\#\hat{C}_{tcp}^t+\hat{tp}^t}{\#\hat{C}_{tcp}^t+\hat{tp}^t+\#\hat{C}_{fcp}^t+\hat{fp}^t} \tag{7}$$

$$\hat{r}^t=\frac{\#\hat{C}_{tcp}^t+\hat{tp}^t}{\#\hat{C}_{tcp}^t+\hat{tp}^t+\#\hat{C}_{fcn}^t+\hat{fn}^t} \tag{8}$$

算法 1　基于概率推断的质量估计算法 QC

输入：S^t，C^{inp}，C^{cpp}，C_{cp}^{t-1}

输出：$\hat{p}^t$，$\hat{r}^t$，C_{cp}^t，sorted C^{inp}

1. Using $C_{cp}^{t-1} \cup C^{cpp}$ to train Q^t
2. P^t，$L^t \leftarrow Q^t(C^{inp})$
3. C_{cp}^t，C_{cn}^t，$C_{inc}^t \leftarrow$ split(S^t，L^t)
4. $\hat{p}^t$，$\hat{r}^t \leftarrow$ estimate(C_{cp}^t，C_{cn}^t，C_{inc}^t，S^t，P^t，L^t)
5. Sorted C^{inp} according to P^t
6. Return $\hat{p}^t$，$\hat{r}^t$，C_{cp}^t，sorted C^{inp}

定理 1（基于概率推断的质量估计算法的时间复杂度）给定 C^{cpp}，C^{inp}，C_{cp}^{t-1} 和 n_f，则 QC 的时间复杂度是 $O(n_f \times (\#C^{cpp}+\#C_{cp}^{t-1})^3+\#C^{inp^2})$。

证明：如算法 1 所示，最耗时的操作是训练过程（即第 1 行）和排序过程（即第 5 行），其中训练过程使用 OneClass-SVM，而排序过程使用快速排序。由于在最坏情况下，One-ClassSVM 的时间复杂度是 $O(n_f \times (\#C^{cpp}+\#C_{cp}^{t-1})^3)$[27]，快速排序的时间复杂度是 $O(\#C^{inp^2})$[28]，则 QC 的时间复杂度是 $O(n_f \times (\#C^{cpp}+\#C_{cp}^{t-1})^3+\#C^{inp^2})$。证毕。

4.3　基于二分搜索的翻转策略选择

基于二分搜索的翻转策略选择是智能体的核心功能。本文的研究问题是搜索满足给定查准率阈值 τ 且查全率最大的消歧结果，该问题等价于在所有查准率大于或等于 τ 的消歧结果集 $\mathbb{S}=\{S_1, S_2, \cdots, S_n\}$ 中，搜索标记为 P 的记录对数目最多的 S^*，然后把上一轮的消歧结果中每个记录对的标签翻转为 S^* 中相应的标签。

直观的算法是，首先组合出所有的翻转方案，然后逐一估算这些翻转方案的查准率和查全率，并选择被标记为 P 的记录对数目最多的方案。该算法的时间复杂度为 $O(2^{|C^{inp}|})$，为此，本文提出了一个近似的翻转策略搜索算法 flip。该算法首先把 C^{inp} 中的候选记录对按照匹配概率值递增排序，这样就可以构建出 $\#C^{inp}$ 个翻转方案，其中第 k 个方案是指把前 k 个记录对的标签设置为 P，而把其他的设置为 N。然后计算这些方案的查准率。容易验证，随着 k 值的增大，第 k 个候选方案的查准率估计值 $\hat{p}$ 是单调递减的，如图 2 所示。最后使用二分搜索算法找到最后一个查准率大于或等于 τ 的方案。最后一个查准率大于或等于 τ 的方案有最多的标记为 P 的记录对，即拥有最大的查全率。由于二分搜索算法假设搜索元素的键值是单调的，因此本文简化查准率的计算公式为计算第 k 个方案中标记为 P 的记录对的匹配概率均值。由于最坏情况下二分搜索的时间复杂度为 $O(\log(\#C^{inp}))$，因此最坏情况下，flip 算法的时间复杂度为 $O(\log(\#C^{inp}))$。

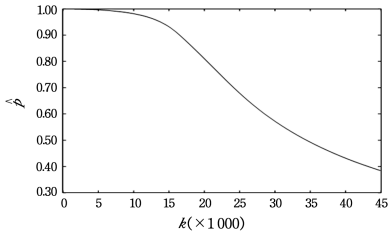

图 2　$\hat{p}$ 和 k 之间的函数关系

Fig. 2　Functional relationship between $\hat{p}$ and k

定理 2(基于概率推断的质量控制算法的时间复杂度) 给定 C^{cpp},C^{inp},τ,n_f 和 $maxT$,QCAgent 的时间复杂度是 $O(maxT\times(\log(\#C^{inp})+n_f\times(\#C^{inp}+\#C^{cpp})^3+\#C^{inp^2}))$。

证明:如算法 2 所示,QCAgent 的关键操作是质量估计算法 QC 和翻转算法 flip。由定理 1 可知,在最坏情况下,QC 的时间复杂度是 $O(n_f\times(\#C^{inp}+\#C^{cpp})^3+\#C^{inp^2})$。flip 的时间复杂度为 $O(\log(\#C^{inp}))$,且整个交互过程最多进行 $maxT$ 次。因此 QCAgent 的时间复杂度是 $O(maxT\times(\log(\#C^{inp})+n_f\times(\#C^{inp}+\#C^{cpp})^3+\#C^{inp^2}))$。证毕。

算法 2 基于概率推断的质量控制算法 QCAgent

输入:S^0,C^{inp},C^{cpp},τ,maxT

输出:S^*

1. $\hat{p}^1$,$\hat{r}^1$,C^1_{cp},sorted $C^{inp}\leftarrow QC(S^0,C^{inp},C^{cpp},\varnothing)$
2. t=1
3. While $\hat{p}^t<\tau\wedge t<maxT$
4. $S^t\leftarrow flip(\hat{p}^t,\hat{r}^t,\text{sorted } C^{incp},\tau)$
5. $\hat{p}^t,\hat{r}^t,C^t_{cp},\text{sorted } C^{inp}\leftarrow QC(S^t,C^{inp},C^{cpp},C^{t-1}_{cp})$
6. Return $S^*\leftarrow S^t$

5 实验验证

本节概述了实验的运行环境,并在真实数据集上验证了算法的有效性和可扩展性。

实验环境的配置为 Intel(R) Core(TM) i7-4710MQ 2.50 GHz 处理器、16 GB 内存和 Ubuntu16.04 64 位的操作系统。编程语言是 Python 3。服务器端数据库是 MongoDB。

实验所用数据集为 Cora[1),该数据集是一个文献数据,包含 1 295 条记录,每条记录隶属于 112 个实体的某一个。每条记录由 12 个属性描述。不一致记录对是本文处理的对象。本文使用了 10 个个体方法,即 5 个无监督的解析方法(基于规则的方法 Rule、基于距离的方法 Distance、基于 K-mean 的 Cluster、基于高斯混合模型的 GMM、基于狄利克雷过程的变分贝叶斯高斯混合模型 DPBGM)和 5 个基于学习的解析方法(基于支持向量机的 SVM、基于决策树模型的 CART、基于随机森林的 ERT、基于高斯朴素贝叶斯模型的 GNB 和基于多层感知器的 MLP)。经 blocking 技术处理后,得到的候选记录对数目为 837 865。对于 Cora 数据而言,不一致记录对的数目为44 909,一致匹配对的数目为1 015,而一致不匹配对的数目为 791 941。由于个体方法无法提供质量控制保证,本文方法不与个体方法进行对比。

基于概率推断的质量控制智能体 QCAgent 是第一个在没有标签数据的场景中,输出满足查准率约束且最大化查全率的消歧算法。本文设计并实现了基于匹配概率推断的消歧算法 QCProb 和基于预测标签的消歧算法 QCLabel。QCLabel 的思想是先用 OneClassSVM 模型预测 C^{inp} 中记录对的标签,然后将这些标签看作真实的标签,最后对 C^{inp} 的消歧结果进行质量估计。QCProb 则是用 OneClassSVM 模型输出的记录对匹配概率估算消歧结果的质量指标。

我们采用实体解析文献[1,21]广泛使用的查准率 *precision* 来评价算法的有效性。

5.1 有效性

本节对比了在不同查准率阈值 τ 下,QCAgent,QCProb 和 QCLabel 估算的查准率 $\hat{p}$ 与真实的查准率 p_{gt} 之间的差异。τ 的取值从 0.75 变化到 0.99,如表 3 所列。实验结果表明:1)混合标签和概率的算法 QCAgent 可高效地输出满足用户查准率需要的消歧结果,这是因为 QCAgent 输出的 $\hat{p}$ 和 p_{gt} 均大于阈值 τ,且迭代次数远小于最大迭代次数 $maxT$,实验中 $maxT$ 为 100。2)基于标签的算法 QCLabel 无法输出满足用户查准率要求的消歧结果,这是因为 QCLabel 估算的 $\hat{p}$ 大于相应的阈值,但实际获得的查准率 p_{gt} 却小于 τ。3)完全基于概率的算法 QCProb 在大部分情况下输出的 p_{gt} 都满足阈值 τ,但在 τ 为 0.75 和 0.99 时,p_{gt} 小于 τ,无法输出满足用户查准率约束的消歧结果;另外,与 QCAgent 相比,QCProb 需要更多的迭代次数,如 τ 为 0.8 时,QCProb 需要迭代 68 次,而 QCAgent 仅需要迭代 12 次。

表 3 迭代次数、查准率的估计值和真实值

Table 3 Interation number, estimated values and real values of precision

τ	QCAgent			QCLabel			QCProb		
	$\#iter$	$\hat{p}$/%	p_{gt}/%	$\#iter$	$\hat{p}$/%	p_{gt}/%	$\#iter$	$\hat{p}$/%	p_{gt}/%
0.75	20	76.09	93.07	1	81.12	73.81	1	79.84	73.81
0.80	12	81.42	95.87	1	81.12	73.81	68	80.04	96.88
0.85	12	92.22	97.85	6	88.34	79.87	36	85.02	96.59
0.90	10	92.11	98.11	5	95.55	79.30	18	90.43	97.23
0.95	7	96.68	97.24	3	95.88	79.99	8	95.86	97.36
0.99	5	99.71	99.65	2	99.45	88.17	4	99.65	98.66

5.2 可扩展性

本节研究了在不同的不一致记录对数目 $\#C^{inp}$ 下,QCAgent 运行时间的变化。尽管由定理 2 可知,最坏情况下,QCAgent 的时间复杂度与数据量 $\#C^{inp}$ 是三次方关系,但在真实数据集上,实验结果表明:QCAgent 所需要的运行时间与数据量近似呈线性关系。如图 3 所示,当 $\#C^{inp}$ 从 2 000 变化到 16 000 时,QCAgent 的运行时间随着 $\#C^{inp}$ 的增加而近似线性增长。

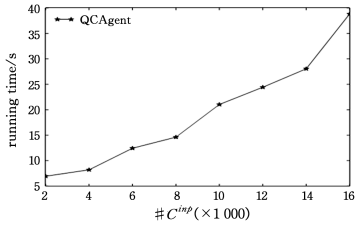

图 3 可扩展性

Fig. 3 Scalability

1) http://www.cs.utexas.edu/users/ml/riddle/data/cora.tar.gz

结束语　本文首次提出了一个基于概率推断的质量控制智能体 QCAgent。该智能体不需要训练数据集，能够输出满足给定查准率约束条件且查全率最大的消歧结果。实验结果表明，该算法能够有效地返回满足查准率要求的消歧结果。由于没有真实标签的辅助，消歧结果的查全率还不够好，在后续的研究工作中，将考虑加入人的反馈，并设计有效的人机协作质量控制算法。

参 考 文 献

[1] XU Y, LI Z, CHEN Q, et al. GL-RF: A Reconciliation Framework for Label-free Entity Resolution [J]. Frontiers of Computer Science, 2018, 12(5): 1035-1037.

[2] LI G. Human-in-the-loop data integration [J]. Proceedings of the VLDB Endowment, 2017, 10(12): 2006-2017.

[3] FAN F F, LI Z H, CHEN Q, et al. An outlier-detection based approach for automatic entity matching [J]. Chinese Journal of Computers, 2017, 40(10): 2197-2211. (in Chinese)
樊峰峰，李战怀，陈群，等. 一种基于离群点检测的自动实体匹配方法[J]. 计算机学报，2017，40(10)：2197-2211.

[4] EFTHYMIOU V, STEFANIDIS K, CHRISTOPHIDES V. Minoan ER: Progressive Entity Resolution in the Web of Data[C]// Proceedings of the 19th International Conference on Extending Database Technology. 2016: 670-671.

[5] LI L, LI J, GAO H. Rule-Based Method for Entity Resolution [J]. IEEE Transactions on Knowledge & Data Engineering, 2015, 27(1): 250-263.

[6] WHANG S E, MARMAROS D, GARCIA-MOLINA H. Pay-as-you-go entity resolution [J]. IEEE Transactions on Knowledge and Data Engineering, 2013, 25(5): 1111-1124.

[7] BELLARE K, IYENGAR S, PARAMESWARAN A, et al. Active Sampling for Entity Matching with Guarantees [J]. ACM Transactions on Knowledge Discovery from Data, 2013, 7(3): 1-24.

[8] BELLARE K, IYENGAR S, PARAMESWARAN A G, et al. Active sampling for entity matching[C]// Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM: New York, 2012: 1131-1139.

[9] WANG J, LI G, YU J X, et al. Entity matching: how similar is similar [J]. Proceedings of the VLDB Endowment, 2011, 4(10): 622-633.

[10] MONGE A E, ELKAN C. The Field Matching Problem: Algorithms and Applications[C]// Proceedings of the Second International Conference on Knowledge Discovery and Data Mining. AAAI Press: California, 1996: 267-270.

[11] ZHANG D, GUO L, HE X, et al. A Graph-Theoretic Fusion Framework for Unsupervised Entity Resolution[C]// Proceedings of the 34th IEEE International Conference on Data Engineering. IEEE Computer Society, 2018: 713-724.

[12] ARASU A, GÖTZ M, KAUSHIK R. On active learning of record matching packages[C]// Proceedings of the 2010 ACM SIGMOD International Conference on Management of Data. ACM: New York, 2010: 783-794.

[13] MUDGAL S, LI H, REKATSINAS T, et al. Deep Learning for Entity Matching: A Design Space Exploration[C]// Proceedings of the 2018 International Conference on Management of Data. ACM: New York, 2018: 19-34.

[14] COHEN W, RAVIKUMAR P, FIENBERG S. A comparison of string metrics for matching names and records[C]// Proceedings of the KDD Workshop on Data Cleaning and Object Consolidation. 2003: 73-78.

[15] EBRAHEEM M, THIRUMURUGANATHAN S, JOTY S, et al. Distributed representations of tuples for entity resolution[J]. Proceedings of the VLDB Endowment, 2018, 11(11): 1454-1467.

[16] COHEN W W. Data integration using similarity joins and a word-based information representation language [J]. ACM Transactions on Information Systems, 2000, 18(3): 288-321.

[17] DAS A, KOTTUR S, MOURA J M F, et al. Learning Cooperative Visual Dialog Agents with Deep Reinforcement Learning [C]// Proceedings of the IEEE International Conference on Computer Vision. 2017: 2970-2979.

[18] MNIH V, KAVUKCUOGLU K, SILVER D, et al. Human-level control through deep reinforcement learning [J]. Nature, 2015, 518(7540): 529-533.

[19] LIU Q, ZHAI J W, ZHANG Z Z, et al. A Survey on Deep Reinforcement Learning [J]. Chinese Journal of Computers, 2018, 41(1): 1-27. (in Chinese)
刘全，翟建伟，章宗长，等. 深度强化学习综述 [J]. 计算机学报，2018，41(1)：1-27.

[20] ZHAO X Y, DING S F. Research on Deep Reinforcement Learning [J]. Computer Science, 2018, 45(7): 1-6. (in Chinese)
赵星宇，丁世飞. 深度强化学习研究综述 [J]. 计算机科学，2018，45(7)：1-6.

[21] CHEN Z, CHEN Q, FAN F, et al. Enabling quality control for entity resolution: A human and machine cooperation framework [C]// Proceedings of the 2018 IEEE 34th International Conference on Data Engineering. IEEE: New Jersey, 2018: 1156-1167.

[22] EFTHYMIOU V, PAPADAKIS G, PAPASTEFANATOS G, et al. Parallel meta-blocking for scaling entity resolution over big heterogeneous data [J]. Information Systems, 2017, 65: 137-157.

[23] WANG Q, CUI M, LIANG H. Semantic-aware blocking for entity resolution [J]. IEEE Transactions on Knowledge and Data Engineering, 2016, 28(1): 166-180.

[24] SIMONINI G, BERGAMASCHI S, JAGADISH H. BLAST: a loosely schema-aware meta-blocking approach for entity resolution [J]. Proceedings of the VLDB Endowment, 2016, 9(12): 1173-1184.

[25] PAPADAKIS G, KOUTRIKA G, PALPANAS T, et al. Meta-Blocking: Taking Entity Resolution to the Next Level [J]. IEEE Transactions on Knowledge & Data Engineering, 2014, 26(8): 1946-1960.

[26] SCHÖLKOPF B, PLATT J C, SHAWE-TAYLOR J, et al. Estimating the support of a high-dimensional distribution [J]. Neural computation, 2001, 13(7): 1443-1471.

[27] PEDREGOSA F, VAROQUAUX G, GRAMFORT A, et al. Scikit-learn: Machine learning in Python [J]. Journal of Machine Learning Research, 2011, 12: 2825-2830.

[28] CORMEN T H, LEISERSON C E, RIVEST R L, et al. 算法导论 [M]. 殷建平，徐云，王刚，等译. 北京：机械工业出版社，2013.

[29] KÖPCKE H, THOR A, RAHM E. Evaluation of entity resolution approaches on real-world match problems [J]. Proceedings of the VLDB Endowment, 2010, 3(1-2): 484-493.