ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 38,No. 3,May 2023,pp. 629－642
DOI:10. 16337/j. 1004-9037. 2023. 03. 012

http://sjcj. nuaa. edu. cn
E-mail:sjcj@nuaa. edu. cn
Tel/Fax:＋86-025-84892742

基于多特征融合的无监督真值发现方法

陈华凤[1]，董永权[2]，杨昊霖[2]，张国玺[2]

(1. 江苏师范大学信息化建设与管理处，徐州 221116；2. 江苏师范大学计算机科学与技术学院，徐州 221116)

摘　要：真值发现是数据集成领域具有挑战性的研究热点之一。传统的方法利用数据源与观测值之间的交互关系推断真值，缺乏足够的特征信息；基于深度学习的方法可以有效地进行特征抽取，但其性能依赖于大量手工标注，而在实际应用中很难获取到大量高质量的真值标签。为克服以上问题，本文提出一种基于多特征融合的无监督真值发现方法(Unsupervised truth discovery method based on multi-feature fusion，MFOTD)。首先，利用集成学习无监督标注“真值”标签；然后，分别使用预训练模型Bert和独热编码获取观测值的语义特征和交互特征；最后，融合观测值多种特征并使用其“真值”标签构建初始训练集，通过自训练方式训练真值预测模型。在两个真实数据集上的实验结果表明，与已有方法相比，本文所提出的方法具有更高的真值发现准确性。

关键词：Web数据集成；半监督学习；数据清洗；真值发现；数据源质量

中图分类号：TP391.1　　**文献标志码：**A

Unsupervised Truth Discovery Method Based on Multi-feature Fusion

CHEN Huafeng[1]，DONG Yongquan[2]，YANG Haolin[2]，ZHANG Guoxi[2]

(1. Department of Information Construction and Management，Jiangsu Normal University，Xuzhou 221116，China；2. College of Computer Science and Technology，Jiangsu Normal University，Xuzhou 221116，China)

Abstract：Truth discovery is one of the challenging research hotspots in the field of data integration. Traditional methods use the interaction between data sources and values to infer the truth，which lack sufficient feature information. Deep learning-based methods can effectively perform feature extraction，but their performance depends on a large number of manual annotations，and it is difficult to obtain a large number of high-quality truth labels in practical applications. To overcome these problems，this paper proposes an unsupervised truth discovery method based on multi-feature fusion(MFUTD). First，ensemble learning is used to label truth without supervision. Then，the pre-training Bert model and the one-hot coding method are used to obtain the semantic features and interactive features of the values. Finally，the initial training set is constructed by fusing multiple features of the values and using their “truth” labels to train the truth prediction model by self-training. Experimental results on two real data sets show that the proposed method has the higher truth discovery accuracy than the existing methods.

Key words: Web data integration; semi-supervised learning; data cleaning; truth discovery; data source quality

基金项目：国家自然科学基金(61872168)；江苏省研究生科研与实践创新项目(KYCX20_2382)。
收稿日期：2022-06-24；**修订日期：**2022-07-19

引 言

随着互联网的广泛应用和移动终端设备的普及,信息的获取变得更加便捷和有效,与此同时,Web上的信息质量问题也日益突出。数据量的爆炸式增长,导致低质量的数据充斥着整个Web,大量过时、错误、虚假和片面的信息,极大地影响了数据的使用价值,其中不同数据源为同一对象提供冲突信息的问题尤为突出。这种冲突信息对数据准确性造成了前所未有的挑战。相关研究表明,多源环境下的数据冲突问题是造成数据质量低下的重要因素[1]。如何完成数据的价值提纯是大数据时代亟待解决的难题。真值发现能够从多个数据源收集的冲突数据中找出最真实、准确的信息,已经成为提高数据质量的重要手段之一,受到学界和业界的广泛关注。

早期的真值发现方法包括基于迭代的方法[2-8]、基于优化的方法[9-12]和基于概率图的方法[13-20]。基于迭代的方法主要利用数据源可靠度和观测值可信度之间的相互影响,将真值推理和数据源权重计算设计为一个迭代过程,直到满足预设的停止准则。Yin等[5]利用数据源的可靠度和观测值的可信度以及观测值之间的相互影响,提出TruthFinder迭代算法。Pasternack等[6]在TruthFinder基础上引入先验知识,将通用常识推理和用户已有的知识用约束不等式表示,应用整数线性规划方法在真值发现过程中施加约束,同时提出了3个真值发现方法:AverageLog、Investment和PooledInvestment。此类方法的缺点是假定所有数据源都具有相同的可靠性,这在直观上具有不合理性,导致较差的性能。基于优化的方法一般假设可靠度越高的数据源提供的观测值越可信,同时,经常提供可信观测值的数据源也越可靠,据此设计目标函数,利用优化方法学习能够反映出数据源可靠度和观测值可信度之间关系的相关参数,具体的推断过程仍然采用迭代方法。典型的算法有Li等[12]提出的CRH模型,将真值和数据源可靠度定义为两组未知变量,通过最小化真值和多源观测值之间的总体加权偏差进行真值发现。该模型的优点是能够采用各种损失和正则化函数来有效地表征不同的数据类型和权重分布,适用于广泛存在的异构数据集。基于概率图的方法假设数据源的可靠度与其提供的观测值可信度之间存在概率关系,贝叶斯网络模型[20]常用于捕捉这种关系,通过采样和参数估计的方法推断真值。由于概率图模型的灵活性,除了数据源可靠度之外,还引入了多种相关因素进行建模。例如,观测值之间的相似性[7]、某些观测值的获取难度[19]、数据源权威性和公平性[21]以及数据源独立性[22-23]等。此类方法大多数都假定数据源在所提供的整个数据上具有相同的可靠性,但忽略了其在不同属性、对象和对象组也可能存在不同可靠性的问题。文献[13]提出概率图模型TFAR、TFOR和TFGR,通过引入数据源可靠性差异的概念,估计多源异构数据环境中,数据源在不同属性、对象和对象组的不同可靠度。

虽然早期的这些方法在特定的应用中已经显示出各自的有效性,但是这些算法大都使用人工设计的简单函数对数据源和观测值之间的交互特征进行建模。事实上,数据源可靠度和观测值可信度的依赖关系先验未知且十分复杂,依赖人工设计的简单函数难以捕获充分的信息,将会导致不理想的结果。近年来,深度学习模型在计算机视觉、自然语言处理等场景下取得巨大成功,部分学者将其应用于真值发现问题[24-30]。Marshall等[24]使用前馈神经网络来建模数据源可靠度和观测值可信度之间的关系。Lyu等[25]通过构造异构网络,从数据源和观测值之间的相互作用中自动学习观测值的特征表示,但该方法仅考虑了数据源和观测值的交互信息,忽略了观测值自身蕴含的深层语义信息。Li等[26]使用长短期记忆神经网络模型来学习数据源可靠度向量和观测值可信度向量,其中观测值可信度向量是对象名、属性名和属性值词向量的拼接向量,数据源可靠度向量与观测值可信度向量相乘,通过误差反向传播进行更新。尽管该模型将(对象,属性,值)三元组作为彼此的上下文来获取观测值可信度向量,但其使用word2vec作为编码器,忽略了上下文语境信息。这些方法显示出了深度学习的强大表征能力,但仍然存在一些局限性。一方面,这些模型都属于监督学习,更多地依赖于大量人工标注的真值标签,但在

真值发现的一般场景中,通常很难获取到足够的真值标签来学习监督模型,从而导致该方法在真值发现中无法普及和有效利用;另一方面,模型只考虑了数据源和观测值的交互特征,忽视了观测值自身的深层语义特征。据文献[20]中所述,标注训练数据的质量和获取成本问题已经成为制约深度学习方法在真值发现任务上继续深入发展的因素之一。

针对以上方法存在的问题,本文提出一个基于多特征融合的无监督真值发现方法。首先设计一种基于集成学习的标注方法,获取"真值"标签,无需任何人工参与。但是这种策略产生的标注信息不可避免地存在噪声标签,噪声数据带来的误差会在模型训练过程中逐渐积累,从而影响模型的性能。为了解决这一问题,本文提出了一种融合观测值语义特征和交互特征的自训练真值发现方法,有效地降低了初始标注集中噪声数据带来的消极影响。

1 问题定义

1.1 相关概念

在正式定义本文研究的真值发现问题之前,先给出以下相关概念及其解释,并使用航班领域中的数据示例(表1)来说明这些概念。

表1 航班UA-2708的属性信息

Table 1 Attribute information for flight UA-2708

Website	Flight	Scheduled departure time	Actual departure time	Departure gate
orbitz	UA-2708	08:45	08:55	D37
flightstats	UA-2708		09:00	D37
flights	UA-2708	08:45	08:50	E5
ifly	UA-2708	09:15		A7
flightaware	UA-2708	08:45	08:55	D37

定义1 对象(Object)表示一个能在真实世界中被识别的、唯一的实体;属性(Attribute)用于描述具体对象的特征,一个对象可能有多个属性。

示例1 "航班UA-2708"是一个对象;"Scheduled departure time"是航班UA-2708的一个属性。

定义2 数据源(Source)表示可以收集有关对象属性信息的数据来源。

示例2 网站"flights"是一个数据源。

定义3 观测值(Value)表示描述某一对象的某个属性的值。观测值根据具体属性种类可分为分类型和连续型两种。

示例3 航班UA-2708的Scheduled departure time中的08:45是一个连续型观测值,Departure gate中的D37是一个分类型观测值。

定义4 真值(Truth)是用于描述某一对象某个属性的真实、准确的观测值。真值的数量存在一个或者多个,分别称为单真值和多真值。本文主要研究单真值发现问题。

本文中使用的变量定义如表2所示。

1.2 问题描述

根据1.1节中的相关概念和定义,本文主要研究的问题可描述为:给定数据源集合S,对象集O及属性集A,其所有的观测值集$V=\{V_{ij},1\leqslant i\leqslant|O|,\ 1\leqslant j\leqslant|A|\}$,其中$V_{ij}=\{v_{ij}^1,\cdots,v_{ij}^{|S|}\}$,$v_{ij}^k(1\leqslant k\leqslant|S|)$表示第$k$个数据源提供的关于对象$O_i$在属性$A_j$上的观测值。对象$O_i$在属性$A_j$上的真值发现是从观测值集$V_{ij}$中推断出其真值$v_{ij}^*$。

表2 符号描述

Table 2 Description of the symbols

符号	描述
$O=\{O_i\}_{i=1}^{\lvert O\rvert}$	对象集合，O_i表示第i个对象
$A=\{A_j\}_{j=1}^{\lvert A\rvert}$	属性集合，A_j表示第j个属性
$S=\{s_k\}_{k=1}^{\lvert S\rvert}$	数据源集合，s_k表示第k个数据源
$V=\{V_{ij},1\leqslant i\leqslant\lvert O\rvert,1\leqslant j\leqslant\lvert A\rvert\}$	观测值集合
$V_{ij}=\{v_{ij}^1,v_{ij}^2,\cdots,v_{ij}^{\lvert S\rvert}\}$	对象O_i在属性A_j上的观测值集合
v_{ij}^k	对象O_i在属性A_j上的第k个观测值
v_{ij}^*	对象O_i在属性A_j上预测的真值
$\boldsymbol{h}_{ij}^k$	观测值v_{ij}^k的语义特征向量
$\boldsymbol{r}_{ij}^k$	观测值v_{ij}^k的交互特征向量
$\boldsymbol{x}_{ij}^k$	观测值v_{ij}^k的融合特征向量

2 本文提出的方法

2.1 模型框架

本文提出了一个基于多特征融合的无监督真值发现方法(Unsupervised truth discovery method based on multi-feature fusion,MFUTD),其整体框架如图1所示,主要包含4个步骤:(1)基于集成学习的无监督真值标注方法,基于集成学习思想,使用多种基线方法预测对象O_i的属性A_j的伪真值,从中选取少量一致度高的观测值作为“真值”标签进行标注,从而获取标注的观测值集;(2)基于Bert编码的观测值语义特征表示,针对对象O_i的属性A_j的第k个观测值v_{ij}^k,利用Bert编码构建其语义特征向量$\boldsymbol{h}_{ij}^k$;(3)基于独热编码的观测值交互特征表示,为观测值v_{ij}^k构建其交互特征向量$\boldsymbol{r}_{ij}^k$;(4)基于自训练学习的真值预测模型,通过将v_{ij}^k的多特征融合向量表示为$\boldsymbol{x}_{ij}^k=\left[\boldsymbol{h}_{ij}^k,\boldsymbol{r}_{ij}^k\right]$,使用第一步获取到的小部分标注观测值集与其多特征融合向量集作为初始训练集,通过自训练学习方式生成最后的真值发现预测模型,并选取预测为真概率最高的观测值作为真值输出。

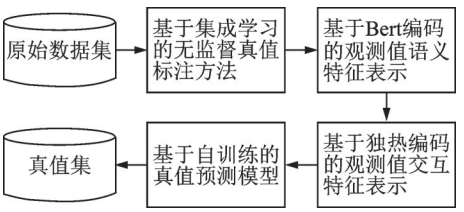

图1 本文整体框架

Fig.1 Overall framework of the proposed method

2.2 基于集成学习的无监督真值标注方法

监督学习模型通常依赖大量标注数据,但在真值发现实际应用中,人工标注真值标签需要耗费巨大的人力与时间代价,这也使得大多数真值发现方法都是以无监督方式进行。然而,随着数据规模与日俱增,数据种类与形式多样化特征日趋明显,现有真值发现方法的准确性和可适用性都受得了前所未有的挑战。目前,为了解决真值发现标签稀疏问题,有研究者提出了基于弱监督[31]和半监督[32]的真值发现方法,实验结果发现可以显著提高真值发现结果的质量,然而这类方法无法在获取不到真值标签的场景中使用。本文首先提出一种基于集成学习的无监督真值标注方法,在没有任何人工参与的情况下,仍可以获取到较高质量的标注集。具体流程如算法1所示。

算法1 基于集成学习的无监督真值标注方法

Input: 所有观测值集 $V=\{V_{ij},1\leqslant i\leqslant\lvert O\rvert,1\leqslant j\leqslant\lvert A\rvert\}$, m个基线真值发现方法 $TD_1,\cdots,TD_m$

Output: 标注集 V_L

(1) $V_L=\varnothing$;
(2) for each V_{ij} in V;
(3) $t_{ij}=$ getPseTruthsByBaseTD($V_{ij}, TD_1, \cdots, TD_m$);
(4) for each t'_{ij} in t_{ij};
(5) if count(t'_{ij}) == m then;
(6) $L_{ij}=$ getLabelDataByVoting(t'_{ij}, V_{ij});
(7) end if;
(8) end for;
(9) $V_L=V_L\cup L_{ij}$;
(10) end for;
(11) return V_L。

首先初始化标注集 V_L 为空集(第1行),接着使用getPseTruthsByBaseTD函数获取所有的观测值集 V 中各对象不同属性的伪真值集(2~10行)。具体地,在对象 O_i 及其属性 A_j 给定的情况下,根据观测值集 V_{ij},利用 m 种基线方法 $TD_1, \cdots, TD_m$ 得到伪真值集,记作 $t_{ij}=\left\{t_{ij}^1, \cdots, t_{ij}^m\right\}$。然后,根据伪真值集 t_{ij},使用getLabelDataByVoting函数实现观测值的标注,用以获取标注集 L_{ij}。具体的,对 t_{ij} 中的值进行投票,选出票数为 m 的伪真值作为对象 O_i 在属性 A_j 上的初始真值,根据初始真值标注 V_{ij} 中的每一个观测值,与初始真值相同的标注为1,否则标注为0,从而获取"真值"标签。标注后的$|S|$个观测值组成标注集 L_{ij},如果没有票数为m的伪真值,则不进行标注。最后,遍历所有观测值集 $V_{ij}\in V$,重复以上步骤,获取最终标注集 V_L。与使用基于单一方法或者投票策略的真值标注方法相比,集成多种基线方法可以有效减少错误标注数据。

2.3 基于Bert编码的观测值语义特征表示

2.3.1 Bert模型

Bert模型于2018年由Google公司提出,在分类、问答、翻译等11项不同的NLP任务中均达到最优性能,其结构如图2所示。其中 $\boldsymbol{E}_i$ 表示嵌入向量,$\boldsymbol{T}_i$ 表示编码向量。Bert主要由Transformer的编码器[33]部分构成。Transformer的编码器(encoder)部分主要包括多头自注意力层和前馈网络层2个部分。其中多头自注意力机制能够有效解决长距离信息丢失的问题,并能充分获取到上下文的语义信息。Bert模型的基础base模型内部是由Transformer的编码器部分堆叠了12层构成,因此其在语义特征提取、长距离信息捕获、句法特征提取等方面都具有一定的优势。Bert结构支持单文本输入和文本对输入。单文本输入需要将符号[CLS]和[SEP]分别放在文本序列的首部和尾部;文本对输入则需要在2个文本序列之间添加符号[SEP]作为分隔符。

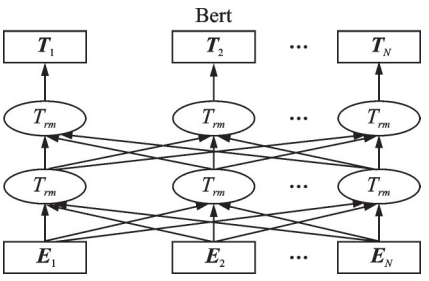

图2 Bert模型结构
Fig. 2 Bert model architecture

2.3.2 观测值语义特征编码

定义5 观测值模式语句。根据观测值描述的对象和属性,结合相关自然语言,包括对象、对象类型、属性、观测值和标点符号等,组合成完整的句子模式,在特定的语境中充当一个完整的语义单位。观测值 v_{ij}^k 的模式语句记为 s_{ij}^k,由于选取的自然语言不唯一,本文构造观测值模式语句示例如下:

‘The’+[属性名]+‘for 对象类型’+[对象名]+‘is’+[观测值].。例如,表1中包含的观测值08:45的观测值模式语句表示为“ The scheduled departure time for flight UA-2708 is 08:45. ”。

为所有观测值集 V 中的每一个观测值构造观测值语句的目的是为了获取真值发现数据中连贯的语义信息。与文献[4]中直接将(对象名,属性名,观测值)作为彼此的上下文来学习观测值语义特征不同,本文通过构建完整的模式语句来捕获连贯的语义信息。由于真值发现输入数据存在不同的结构,如带有时间戳的数据,构建的“模式语句”长度不等,本文使用Bert预训练模型提取观测值的语义特征,可以用固定维度的向量表示不同长度的“模式语句”,具体过程如算法2中所示。

算法2 观测值语义特征表示

Input: $v_{ij}^k, v_{ij}^k \in V_{ij}, V_{ij} = \left\{ v_{ij}^1, v_{ij}^2, \cdots, v_{ij}^{|S|} \right\}$

Output: v_{ij}^k 的语义特征向量 $\boldsymbol{h}_{ij}^k$

(1) begin;

(2) $\boldsymbol{E}_{ij} = \varnothing$;

(3) for each v_{ij}^k in V_{ij};

(4) $s_{ij}^k =$ getPatternSentence (v_{ij}^k);

(5) $e_{ij}^k =$ getPSVectorByBert(s_{ij}^k);

(6) $\boldsymbol{E}_{ij} = \boldsymbol{E}_{ij} \cup \boldsymbol{e}_{ij}^k$;

(7) end for;

(8) if $V_{ij} \in$ categorical then;

(9) $e_{ij}^* = \text{mean}(E_{ij})$;

(10) for each v_{ij}^k in V_{ij};

(11) $d_{ij}^k = \text{sim}\left(e_{ij}^k, e_{ij}^*\right)$;

(12) end for;

(13) end if;

(14) if $V_{ij} \in$ continual then;

(15) $v_{ij}^* = \text{mean}(V_{ij})$;

(16) for each v_{ij}^k in V_{ij};

(17) $d_{ij}^k = \text{dist}\left(v_{ij}^k, v_{ij}^*\right)$;

(18) end for;

(19) end if;

(20) return $\boldsymbol{h}_{ij}^k = [e_{ij}^k, d_{ij}^k]$。

首先为描述对象 O_i 的属性 A_j 的观测值集 V_{ij} 初始化句向量集 $\boldsymbol{E}_{ij}$ 为空集,接着获取 V_{ij} 中每一个观测值的句向量,组成 $\boldsymbol{E}_{ij}$。具体地,使用getPatternSentence函数获取每一个观测值 v_{ij}^k 的模式语句 s_{ij}^k,使用getPSVectorByBert函数获取每一个模式语句 s_{ij}^k 的句向量 $\boldsymbol{e}_{ij}^k$,这里Bert预训练模型可以学习到观测值模式语句中的连贯语义信息,得到的句向量 $\boldsymbol{e}_{ij}^k$ 作为观测值 v_{ij}^k 的语义特征向量。然后,为了适应不同数据类型的特征差异性,在观测值语义特征表示的基础上,根据观测值类型,为分类型和连续型观测值分别构造特征。

具体地,如果 V_{ij} 属于分类型,在获取到句向量集 $\boldsymbol{E}_{ij}$ 之后,计算所有句向量的平均值作为描述对象 O_i 在属性 A_j 上的“真实”语义向量 $\boldsymbol{e}_{ij}^* = \text{mean}(\boldsymbol{E}_{ij})$,然后计算每一个观测值语义特征向量与其的相似度,

如式(1)所示。在此前的工作中,Truthfinder方法[5]通过计算不同观测值之间的距离来度量观测值之间的相互影响,包括正相关和负相关,受此启发,如果相似度越大表示这个观测值的语义与“真实”语义之间的差异越小,为真的概率也越大。

$$\mathrm{sim}\left(\boldsymbol{e}_{ij}^{k},\boldsymbol{e}_{ij}^{*}\right)=\cos\left(\frac{\boldsymbol{e}_{ij}^{k}\cdot\boldsymbol{e}_{ij}^{*}}{\left\|\boldsymbol{e}_{ij}^{k}\right\|\left\|\boldsymbol{e}_{ij}^{*}\right\|}\right) \tag{1}$$

如果V_{ij}属于连续型,计算观测值集V_{ij}的平均值,作为描述对象O_i在属性A_j上的“真实”值$v_{ij}^{*}=\mathrm{mean}(V_{ij})$,然后计算每一个观测值与其的距离,以表示每一个观测值和“真实”值的差异性,如式(2)中所示。如果距离越大,表示这个观测值与“真实”值之间的差异越大,为真的概率越小。

$$\mathrm{dist}\left(v_{ij}^{k},v_{ij}^{*}\right)=\frac{\left(v_{ij}^{k}-v_{ij}^{*}\right)^{2}}{\mathrm{std}\left(v_{ij}^{1},v_{ij}^{2},\cdots,v_{ij}^{|S|}\right)} \tag{2}$$

最后将观测值v_{ij}^{k}的句向量$\boldsymbol{e}_{ij}^{k}$和$\boldsymbol{d}_{ij}^{k}$进行拼接作为其最终的语义特征向量$\boldsymbol{h}_{ij}^{k}$。

2.4 基于独热编码的观测值交互特征表示

定义6 观测值交互特征向量。每个分量表示不同的数据源,如果该数据源提供了这个观测值,该分量为1,反之,为0。

在真值发现的一般场景中,观测值除了自身蕴含的语义特征之外,还包括与数据源之间交互特征,受文献[10]的启发,由大量可靠数据源提供的观测值为真值的概率更高,而由少量低质量数据源提供的观测值为真值的概率较低。为了获取观测值的这种交互特征,本文使用独热编码方式构建观测值交互特征向量,具体过程如算法3所示。

算法3 观测值交互特征表示

Input: 数据源集合$S=\{s_k\}_{k=1}^{|S|}$,$v_{ij}^{k},v_{ij}^{k}\in V_{ij}$,$V_{ij}=\left\{v_{ij}^{1},v_{ij}^{2},\cdots,v_{ij}^{|S|}\right\}$;

Output: v_{ij}^{k}的交互特征向量$\boldsymbol{r}_{ij}^{k}$;

(1) begin;

(2) Let $\boldsymbol{\theta}_{ij}^{k}=\left[\theta_1,\theta_2,\cdots,\theta_{|S|}\right]$ be the $|S|$ dimensional source vector of zeros for v_{ij}^{k};

(3) for each $\theta_{k'}$ in $\boldsymbol{\theta}_{ij}^{k}$;

(4) for each $s_{k'}$ in S;

(5) $\theta_{k'}=\mathrm{getOnehotBySource}(v_{ij}^{k},s_{k'})$;

(6) end for;

(7) end for;

(8) return $\boldsymbol{r}_{ij}^{k}=\boldsymbol{\theta}_{ij}^{k}$。

首先,根据数据源个数定义向量维度,为每一个观测值v_{ij}^{k}构造相同维度的0向量$\boldsymbol{\theta}_{ij}^{k}$,每一维分量表示对应位置的数据源。然后根据向量下标遍历数据源集S,使用getOnehotBySource函数获取$\boldsymbol{\theta}_{ij}^{k}$每一分量的值。具体的,如果相同位置的数据源提供的观测值与v_{ij}^{k}相同,则该分量值为1。最后得到维度与数据源个数相同的向量$\boldsymbol{\theta}_{ij}^{k}$作为观测值$v_{ij}^{k}$的交互特征向量。

在基于独热编码的观测值交互特征生成方法中,由于没有先验知识,特征的生成是根据每个数据源是否提供该数据项得到的,这与基于投票的思想一致,这种基于统计的独热编码方式可以获取到真值发现数据中的大部分观测值的共性特征。

2.5 基于自训练的真值预测模型

以上介绍了基于集成学习的无监督真值标注方法和观测值特征表示方法，为实现使用深度学习模型预测真值的目标，提出一种自训练方法生成真值预测模型来推断真值。首先需要构建真值预测模型的初始训练集：根据2.2节获取到标注集 V_L，真值标签用 Y 表示，则未标注集为 $V-V_L$。根据2.3和2.4节获取到的所有观测值集 V 中所有观测值的特征，记作 $\boldsymbol{X}=[h,r]$，其中 h 为观测值语义特征，r 表示观测值交互特征。为适应真值发现的一般场景且降低模型初始训练的时间代价，对标注集 V_L 随机采样 $n\%$，然后在 $\boldsymbol{X}$ 中提取对应的观测值的特征 $\boldsymbol{X}'$，$\boldsymbol{X}'\in\boldsymbol{X}$，标签 Y'，$Y'\in Y$，组成最后的训练集 $Z=(X',Y')$。剩下未进行训练的全部观测值重新组成未标注集 $V'=V-V_L\times n\%$，在 X 中获取对应的观测值特征得到最后的待预测数据集 L。在获取到已标注的训练集 Z 之后，具体的自训练过程如算法4所示。

算法4 无监督自训练真值发现方法

Input: 已标注训练集 Z，未标注的待预测数据集 L，置信度阈值 $\rho_{\max}$，最大迭代次数 itermax

Output: 训练完成的分类器 cls

(1) Initialize cls;

(2) $i=0$, $Z_i=Z$, $L_i=L$;

(3) while $L_i!=\varnothing$ and $i<$ itermax do;

(4) cls = train_cls(Z_i);

(5) S_i = select($\rho_{\max}$, i);

(6) $L_i=L_i-S_i$;

(7) $Z_i=Z_i\cup S_i$;

(8) $i=i+1$;

(9) end while;

(10) $Z=Z_i$;

(11) cls = train_cls(Z);

(12) return classifier cls。

算法4跟传统的自训练算法不同，是本文提出的适用于真值发现领域的一种新型自训练算法。具体来说，传统的自训练算法每轮迭代时，选择置信度最高的 k 个伪正例样本和 k 个伪负例样本加入训练集 Z。在实验中发现，传统的取置信度最高的 k 个样本的 自训练算法效果不好，因为在增加样本时，传统自训练算法往往会加入与 Z 中数据相似度较高的数据，真值少，虚假值过多，造成 Z 中数据分布的不均衡，导致训练的深度学习模型效果变差。因此，本文设计了一种新的伪标注样本选择函数 $\text{select}(\rho_{\max})$：从置信度超过 $\rho_{\max}$ 的各个类别的样本中分别选取 $\lambda=\min(c_{\mathrm{T}},c_{\mathrm{F}})$ 个样本，c_{T} 表示满足阈值条件中的正样本个数，c_{F} 表示满足阈值条件中的负样本个数。在第 i 轮迭代时，S_i 为第 i 次迭代选出的伪标注样本集，$S_i=\text{select}(\rho_{\max},i)$。通过选择合适的 $\rho_{\max}$，该策略能使得伪标注样本分类准确率较高的同时数据分布较为均匀，从而保证训练得到的模型较为准确且有泛化能力。阈值的选取会对当前分类器产生一定的影响，阈值选取不合适可能会导致错误被不断放大。可以通过验证集来选取合适的参数，保证预测的准确性。

真值推断过程，使用学习到的分类器 cls 来预测 V' 中每个观测值为真的概率，选取预测为真概率最高的观测值作为真值输出。例如，V' 中的描述对象 O_i 在属性 A_j 的观测值集 $V_{ij}=\left\{v_{ij}^1,\cdots,v_{ij}^{|S|}\right\}$，cls 预测的每一个观测值的为真的概率为 $p_{ij}=\left\{p_{ij}^1,\cdots,p_{ij}^{|S|}\right\}$，则真值 $v_{ij}^*=V_{ij}\left[\text{argmax}\left(p_{ij}\right)\right]$。

3 实验与分析

3.1 实验设置

3.1.1 数据集

股票数据集[34]:该数据集包含在1个月内从16个网站收集到的关于1 000个股票的16种属性信息。在本节的实验中,属性volume、shares outstanding和market cap数据被视为连续型,而其他属性数据看作分类型。

航班数据集[34]:该数据集包含在1个月内,从38个网站收集到的关于1 200个航班的6种属性信息。在本节的实验中,属性scheduled departure time, actual departure time, scheduled arrival time, actual arrival time被视为连续型属性,departure gate和arrival gate看作分类型属性。

表3显示了两个真实数据集的统计信息。

表3 两个真实数据集统计信息

Table 3 Statistics of the two real data sets

数据集	Source	Value	Object	Attribute	Truth
股票数据集	55	12 137 940	1 000	16	29 207
航班数据集	38	2 804 595	1 200	6	16 051

3.1.2 评估标准

本文的实验中,在两种不同类型的数据上使用与文献[12]相同的评估指标。对于两种不同的度量标准,值越小,方法表现越好。两种评估指标为:

(1) 错误率(Error rate):使用错误率作为分类型数据的度量标准,计算为错误预测的百分比,错误预测是指与真值不同的模型的输出。

(2) 平均归一化绝对距离(Mean normalized absolute distance, MNAD):使用MNAD作为连续型数据的度量标准,计算每种方法的预测的真值与实际真值之间的绝对距离。由于描述同一对象相同属性的观测值集具有不同的规模,通过其方差进行规范化,然后计算其平均值,即为MNAD[12, 35]。

3.1.3 基线方法

为了证实本文所提方法的有效性,将MFUTD方法和如下方法进行比较:

(1) Mean and median:均值和中位数是用于连续类型数据的方法,分别计算观测值的平均值和中位数做为连续数据最终预测的真实值。

(2) GTM[18]:高斯真值发现模型是一个主要用于连续数据的方法。该方法使用贝叶斯概率模型,对数据源、观测值和真值进行建模。

(3) Voting:将出现次数最多的观测值作为真值,用于分类型数据。

(4) Investment[6]:在该方法中,数据源统一“invest”其提供的观测值的可信度。

(5) Pooledinvestment[6]:Invest和Poolinginvestment之间的差异是Poolinginvestment使用线性函数来度量观测值的可信度。

(6) Truthfinder[5]:该方法被广泛用于各个领域,方法假设数据源的可靠度与观测值的可信度之间存在概率关系,引入imp描述观测值之间的相互影响,以推导出真实值。

(7) Accusim[7]:该算法考虑了观测值之间的相似性,并使用贝叶斯分析进行真值发现。

(8) CRH[12]:CRH是解决异构数据冲突的优化框架,由两个过程组成:真值计算和数据源权值计算。

(9) CASE[25]:基于表征学习的真值发现模型,从数据源和观测值之间的相互作用中自动学习观测值的特征表示。

(10) MFUTD-S:仅使用观测值的语义信息作为真值发现的主要特征,该方法用于验证模型基于Bert构造观测值语义特征向量的效果。

(11) MFUTD-I: 仅使用观测值与数据源之间的交互信息作为真值发现的主要特征,该方法用于验证模型基于one-hot构造观测值交互特征向量的效果。

(12) MFUTD-R: 不使用自训练方式生成真值预测模型,在获取到初始标注集之后,将其作为最终训练集训练真值预测模型。

3.1.4 实验参数设置

本文实验采用Python实现所有算法,实验的运行环境为NVIDIA1080 GPU, 16 GB内存,Ubuntu18.04LTS操作系统。基于集成学习的无监督真值标注方法中,使用的基线方法包括CRH、Investment、PooledInvestment、TruthFinder和Accusim,这些基线方法在两种类型数据上都适用。初始采样数据的比例$n\%$设为5%,最大迭代次数itermax设置为10,置信度阈值$\rho_{\max}$设置为0.95。在真值推断阶段,使用scikit-learn库中DNN分类器作为初始分类器,且使用默认参数。由于本文实验使用与文献[12]相同的数据集、数据处理方法、评估标准和基线方法,实验结果直接与该文献中报告的结果进行比较,其他实验方法重复实验5次取平均值作为其最终的实验结果,实验中CASE方法使用文献[25]的开源代码进行复现,且使用文献报告的最优参数。

3.2 实验结果分析

3.2.1 对比实验结果

为了评估本文所提方法MFUTD的效果,本节实验对比了以上13个真值发现方法,在分类型数据和连续型数据的指标分别为错误率和MNAD。表4给出了所有方法在两个不同数据集上的对比实验结果。

表4 两个真实数据集的性能比较

Table 4 Performance comparison on two real data sets

方法	股票数据集		航班数据集	
	Error rate	MNAD	Error rate	MNAD
MFUTD	**0.063 8**	**0.035 4**	**0.018 7**	**1.894 0**
MFUTD-I	0.069 9	0.278 9	0.038 2	1.913 0
MFUTD-S	0.088 5	0.341 8	0.090 0	2.604 1
MFUTD-R	0.108 7	0.592 1	0.099 2	2.879 2
CASE	0.502 9	0.062 5	0.092 1	2.057 4
CRH	0.070 0	2.644 5	0.082 3	4.861 3
Mean	NA	7.185 8	NA	8.289 4
Median	NA	3.933 4	NA	7.847 1
GTM	NA	3.425 3	NA	7.670 3
Voting	0.081 7	NA	0.085 9	NA
Investment	0.098 3	2.808 1	0.091 9	6.415 3
Pooledinvestment	0.099 0	2.794 0	0.092 5	5.856 2
Truthfinder	0.119 4	2.714 0	0.095 0	8.135 1
AccuSim	0.072 6	2.850 3	0.088 1	7.320 4

由表4可以看出,本文提出的MFUTD方法优于现有的最优基线方法CRH以及最新的基于表征学习的真值发现方法CASE。在现有的基线方法上,CRH在两个数据集的不同类型数据上都取得最优的结果。两个数据集的分类数据上,尽管CRH已经取得很好的效果,但本文提出的MFUTD方法进一步降低了错误率,实现真值发现性能的进一步提升。在两个数据集上的连续数据上,MFUTD性能提升

尤为明显,相比CRH模型,MNAD分别从2.644 5降至0.035 4,从4.861 3降至1.894 0。基于多特征融合的无监督方法MFUTD在连续数据上显著优于现有的方法。与CASE模型相比,MFUTD更具有一般性和适用性,CASE模型仅学习观测值与数据源之间的交互信息作为可信度表征,忽视了观测值自身的语义特征,导致真值发现结果较差且不稳定。MFUTD-S和MFUTD-I方法效果较差的原因在于仅使用观测值单一维度特征,但仍可发现观测值的交互特征蕴含着相对重要的信息,在此基础上,融合观测值的语义信息可以进一步提升真值发现结果的质量。MFUTD-R因为没有使用自训练学习方式训练模型,初始标注训练集的噪声对模型造成较大的影响,导致效果不佳。这从侧面验证了MFUTD在自训练方式下,可以减少初始标注训练集中噪声数据的影响,充分吸收观测值数据中的语义信息和交互信息,在两个真实数据集上都取得了最优的真值发现结果,证明了本文所提MFUTD方法的优越性和良好的泛化能力。

3.2.2 集成学习标注方法的有效性

为了评估本文提出的基于集成学习的无监督标注方法的有效性,将MFUTD方法与使用单一基线方法标注"真值"标签方法的实验结果进行对比。具体地,分别使用集成学习中使用的五种基线方法预测全部数据的真值,然后标注全部数据"真值"标签,从有标注数据集中随机选取5%的数据作为初始训练集来训练模型,固定其他参数,对比实验结果如表5所示。

表5 集成标注与单一标注方法性能比较

Table 5 Performance comparison of integrate marks with single mark

标注方法	股票数据集		航班数据集	
	Error rate	MNAD	Error rate	MNAD
CRH	0.309 7	0.051 6	0.083 1	2.357 2
Investment	0.192 0	0.061 4	0.098 7	1.994 6
Pooledinvest	0.125 0	0.107 7	0.057 4	1.951 1
Truthfinder	0.118 2	0.089 2	0.043 2	1.992 5
Accusim	0.234 9	0.241 4	0.062 6	1.982 1
Integration	**0.063 8**	**0.035 4**	**0.018 7**	**1.894 0**

从表5中可以看出,使用单种基线标注"真值"标签的方法效果明显低于使用集成标注方法的效果,这是因为使用单种基线方法标注"真值"标签会引人较多的噪声数据,即预测的"真值"错误个数过多,即使在模型训练阶段使用自训练学习方式,初始标注集中的噪声对模型造成无法忽视的影响。实验结果证实了MFUTD方法中使用基于集成学习的真值标注方法的有效性和必要性。

3.2.3 自训练学习的有效性

为了验证本文提出的自训练方法的有效性,本节针对航班数据集上的连续数据进行实验,验证模型MFUTD、MFUTD-S、MFUTD-I和MFUTD-R在不同规模有标注数据上训练的效果。从有标注数据集中随机选取$n\%$(n=0.5,1,1.5,2,…,5)的数据作为训练集来训练模型,对比实验结果如图3所示。

当从原始标注数据集中随机采样0.5%的数据(58个样本)对模型以自训练学习方式优化时,MFUTD的错误率已经低至0.064 6,超出性能第一的CRH方法,同时超出了MFUTD-S和MFUTD-R使用5%有标注训练数据时的性能,该结果有力证明了本文所提出的方法MFUTD在极少标注数据上的优越性,原因在于基于集成学习的无监督真值标注方法减少了初始样本中的噪声,并且基于自训练的模型学习方式可以通过筛选高置信度样本有效地实现训练集的扩充,减少噪声数据的影响。随着训

练数据比例的增加,MFUTD和MFUTD-I方法的错误率平稳降低,而MFUTD-R和MFUTD-S观察到较小的波动。通过对比4种方法的实验结果,验证了本文使用观测值多特征融合方法的有效性和观测值交互特征的重要性。同时,证实了在观测值交互特征的基础上融合其语义特征,可以进一步提升真值发现性能。

表6中报告了采样不同规模数据时的错误真值个数。MFUTD-R方法使用常规训练方式学习模型,其效果较差且不稳定。主要因为基于集成的初始标注数据上仍然存在不容忽视的噪声,并且采样的样本过少,随着数据规模的增加(3%之后),MFUTD-R性能反而下降,结合表6,可以注意到标注数据中的错误真值个数增加,噪声对模型MFUTD-R产生了较大影响。MFUTD-R直接将含有较多噪声的标注集作为训练集使用,噪声会对模型的训练过程产生严重干扰,同时从反面证明了MFUTD使用自训练方法训练模型的必要性。

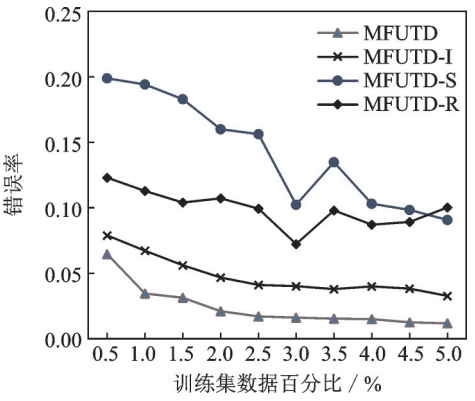

图3 不同规模训练集的实验结果

Fig.3 Experimental results of different training sets

表6 不同规模训练集中的错误真值个数

Table 6 Number of error truth values in training sets with different sizes

训练集规模n%	错误真值个数	训练集规模n%	错误真值个数
0.5	3	3.0	6
1.0	1	3.5	12
1.5	2	4.0	7
2.0	5	4.5	10
2.5	2	5.0	12

3.2.4 概率阈值的影响

在基于多特征融合的无监督真值发现方法MFUTD中,概率阈值$\rho_{\max}$用于筛选自训练迭代过程的高置信度预测样本,以扩充训练集,本节实验对参数$\rho_{\max}$分别取值为0.80, 0.83, 0.85, 0.88, 0.90, 0.93和0.95,观察方法MFUTD、MFUTD-I和MFUTD-S在航班数据集上的错误率变化情况,结果如图4所示。从图4中可以发现3种方法的错误率随着参数$\rho_{\max}$的增加而降低,当参数$\rho_{\max}$增加到一定程度($\rho_{\max}\geqslant 0.85$)算法的错误率变化较为平缓。为了实验的稳定性并且在自训练过程中有效筛选出预测正确的样本,实验应选取较高的概率阈值$\rho_{\max}$。本文实验最后将阈值$\rho_{\max}$设置为0.95。

4 结束语

现有的基于深度学习的真值发现方法都需要使用大量的人工标注数据训练分类器,但人工标注数据获取成本高昂限制了训练数据规模,从而制约了深度学习模型在真值发现上的性能。本文提出一种基于多特征融合的无监督真值发现方法MFUTD,该方法使用集成学习无监督标注“真值”标签,可以获取到较高质量的训练数据;接着,融合观测值的语义信息和交互信息作为其可信度表征,使用自训练方式训练模型,可以充分学习标注数据的潜在信息同时削弱噪声样本带来的负面影响。在真实股票数据集和航班数据集上的实验结果充分验证了本文所提出方法的有效性和优越性。更为重要的是,本文所提MFUTD方

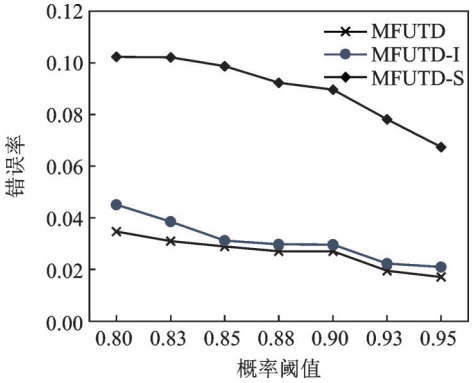

图4 不同概率阈值下错误率的实验结果

Fig. 4 Experimental results of error rate under different probability thresholds

法在没有任何人工参与的情况下取得了显著优于同类算法的性能,有效解决缺乏大规模标注数据这一制约深度学习模型用于真值发现领域的瓶颈问题。

参 考 文 献:

[1] 李建中, 王宏志, 高宏 . 大数据可用性的研究进展[J]. 软件学报, 2016, 27(7): 1605-1625.
LI Jianzhong, WANG Hongzhi, GAO Hong. Research progress in usability of big data[J]. Journal of Software, 2016, 27(7): 1605-1625.

[2] 马如霞, 孟小峰 . 基于数据源分类可信性的真值发现方法研究[J]. 计算机研究与发展, 2015, 52(9): 1931-1940.
MA Ruxia, MENG Xiaofeng. Research on truth discovery method based on data source classification credibility[J]. Journal of Computer Research and Development, 2015, 52(9): 1931-1940.

[3] 张志强, 刘丽霞, 谢晓芹, 等 . 基于数据源依赖关系的信息评价方法研究[J]. 计算机学报, 2012, 35(11): 2392-2402.
ZHANG Zhiqiang, LIU Lixia, XIE Xiaoqin, et al. Research on information evaluation method based on data source dependency[J]. Chinese Journal of Computers, 2012, 35(11): 2392-2402.

[4] 考明军, 张炜, 高宏 . 冲突数据中的真值发现算法[J]. 计算机研究与发展, 2010, 47(1): 188-192.
KAO Mingjun, ZHANG Wei, GAO Hong. Truth discovery algorithm in conflicting data[J]. Journal of Computer Research and Development, 2010, 47(1): 188-192.

[5] YIN X, HAN J, YU P S. Truth discovery with multiple conflicting information providers on the Web[J]. IEEE Transactions on Knowledge and Data Engineering, 2008, 20(6): 796-808.

[6] PASTERNACK J, ROTH D. Knowing what to believe (when you already know something) [C]//Proceedings of the 23rd International Conference on Computational Linguistics. USA: Association for Computational Linguistics, 2010: 877-885.

[7] DONG X L, BERTI-EQUILLE L, SRIVASTAVA D. Integrating conflicting data: The role of source dependence[J]. Proceedings of the VLDB Endowment, 2009, 2(1): 550-561.

[8] 艾静, 王仲远, 孟小峰 . C-Rank: 一种 Deep Web 数据记录可信度评估方法[J]. 计算机科学与探索, 2009, 3(6): 585-593.
AI Jing, WANG Zhongyuan, MENG Xiaofeng. C-Rank: A reliability evaluation method for Deep Web data records[J]. Journal of Frontiers of Computer Science and Technology, 2009, 3(6): 585-593.

[9] WANG D, ABDELZAHER T, KAPLAN L. On truth discovery in social sensing: A maximum likelihood estimation approach [C]//Proceedings of the 11th International Conference on Information Processing in Sensor Networks. New York, United States: ACM, 2012: 233-244.

[10] LI Q, LI Y, GAO J, et al. A confidence-aware approach for truth discovery on long-tail data[J]. Proceedings of the VLDB Endowment, 2014, 8(4): 425-436.

[11] LI Y, LI Q, GAO J, et al. On the discovery of evolving truth[C]// Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, United States:ACM, 2015: 675-684.

[12] LI Y, LI Q, GAO J, et al. Conflicts to harmony: A framework for resolving conflicts in heterogeneous data by truth discovery [J]. IEEE Transactions on Knowledge and Data Engineering, 2016, 28(8): 1986-1999.

[13] TIAN H, SHENG W, SHEN H, et al. Truth finding by reliability estimation on inconsistent entities for heterogeneous data sets[J]. Knowledge-Based Systems, 2020, 187: 12-26.

[14] ZHANG L, QI G, ZHANG D, et al. Latent dirichlet truth discovery: Separating trustworthy and untrustworthy components in data sources[J]. IEEE Access, 2018, 6: 1741-1752.

[15] ZHANG H, LI Q, MA F, et al. Influence-aware truth discovery[C]//Proceedings of the International Conference on Information and Knowledge Management. New York, United States:ACM, 2016: 851-860.

[16] PASTERNACK J, ROTH D. Latent credibility analysis[C]//Proceedings of the International Conference on World Wide Web. New York, United States:ACM, 2013: 1009-1020.

[17] 张永新, 李庆忠, 彭朝晖 . 基于 Markov 逻辑网的两阶段数据冲突解决方法[J]. 计算机学报, 2012, 35(1): 101-111.
ZHANG Yongxin, LI Qingzhong, PENG Zhaohui. Two-stage data conflict resolution method based on Markov Logic network [J]. Chinese Journal of Computers, 2012, 35(1): 101-111.

[18] ZHAO B, HAN J. A probabilistic model for estimating real-valued truth from conflicting sources[C]//Proceedings of the International Workshop on Quality in Databases. [S.l.]:[s.n.], 2012.

[19] GALLAND A, ABITEBOUL S, MARIAN A, et al. Corroborating information from disagreeing views[C]//Proceedings of the ACM International Conference on Web Search and Data Mining. New York, United States:ACM, 2010: 131-140.

[20] YANG J, TAY W P. An unsupervised Bayesian neural network for truth discovery in social networks[J]. IEEE Transactions on Knowledge and Data Engineering, 2021. DOI: 10.1109/TKDE.2021.3054853.

[21] LI Y, SUN H, WANG W. Towards fair truth discovery from biased crowdsourced answers[C]//Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, United States:ACM, 2020: 599-607.

[22] QI G, AGGARWAL C, HAN J, et al. Mining collective intelligence in diverse groups[C]//Proceedings of the 22nd international Conference on World Wide Web. New York, United States:ACM, 2013: 1041-1052.

[23] DONG X L, HU Y, BERTI-EQUILLE L, et al. Solomon: Seeking the truth via copying detection[J]. Proceedings of the VLDB Endowment, 2010, 3(2): 1617-1620.

[24] MARSHALL J, ARGUETA A, WANG D. A neural network approach for truth discovery in social sensing[C]//Proceedings of the 14th International Conference on Mobile Ad Hoc and Sensor Systems(MASS). Orlando, FL, USA: IEEE, 2017: 343-347.

[25] LYU S, OUYANG W, WANG Y, et al. Truth discovery by claim and source embedding[J]. IEEE Transactions on Knowledge and Data Engineering, 2021, 33(3): 1264-1275.

[26] LI L, QIN B, REN W, et al. Truth discovery with memory network[J]. Tsinghua Science and Technology, 2017, 22(6): 609-618.

[27] CHEN H, DONG Y, GU Q, et al. An end-to-end deep neural network for truth discovery[C]//Proceedings of the International Conference in Web Information Systems and Applications. Cham:Springer, 2020: 377-387.

[28] 曹建军, 常宸, 翁年凤, 等 . 基于神经网络编码的真值发现[J]. 计算机工程与科学, 2021, 43(9): 1546-1557.
CAO Jianjun, CHANG Chen, WENG Nianfeng, et al. Truth discovery based on neural network coding[J]. Computer Engineering and Science, 2021, 43(9): 1546-1557.

[29] 常宸, 曹建军, 吕国俊, 等 . 基于 Bi-GRU 并包含注意力机制的文本数据真值发现[J]. 中文信息学报, 2020, 34(2): 46-55.
CHANG Chen, CAO Jianjun, LYU Guojun, et al. Truth discovery of text data based on BI-GRU and including attention mechanism[J]. Journal of Chinese Information Processing, 2020, 34(2): 46-55.

[30] CHANG C, CAO J, ZHENG Q, et al. An unsupervised approach of truth discovery from multi-sourced text data[J]. IEEE Access, 2019, 7: 143479-143489.

[31] YANG Y, BAI Q, LIU Q. On the discovery of continuous truth: A semi-supervised approach with partial ground truths[C]//Proceedings of the Web Information Systems Engineering. Cham:Springer, 2018: 424-438.

[32] YIN X, TAN W. Semi-supervised truth discovery[C]//Proceedings of the 20th International Conference on World Wide Web. New York, United States:ACM , 2011: 217-226.

[33] JAWAHAR G, SAGOT B, DJAME S. What does BERT learn about the structure of language[C]//Proceedings of the 57th Annual Meeting of the ACL. Florence, Italy: Computational Linguistics, 2019: 3651-3657.

[34] LI X, DONG X L, LYONS K, et al. Truth finding on the deep web: Is the problem solved?[J]. Proceedings of the VLDB Endowment, 2012, 6(2): 97-108.

[35] LI Q, LI Y, GAO J, et al. Resolving conflicts in heterogeneous data by truth discovery and source reliability estimation[C]//Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data. New York, United States:ACM , 2014: 1187-1198.

作者简介:

陈华凤(1996-),女,硕士研究生,研究方向:真值发现、深度学习,E-mail: s_dimple@163.com。

董永权(1979-),**通信作者**,男,教授,硕士生导师,研究方向:数据集成、数据挖掘, E-mail: tomdyq@163.com。

杨昊霖(1997-),男,硕士研究生,研究方向:多真值发现、机器学习,E-mail: yhl@jsnu.edu.cn。

张国玺(1997-),男,硕士研究生,研究方向:数据挖掘、深度学习,E-mail: zgx116@jsnu.edu.cn。

(编辑:刘彦东)