计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.191100086

基于小样本置信区间的众包答案决策方法

张光园　王　宁
北京交通大学计算机与信息技术学院　北京 100044
(17120441@bjtu.edu.cn)

摘　要　众包工人的水平良莠不齐,质量控制是众包面临的挑战之一。目前的研究大多通过评估工人质量来保证最终答案的有效性,但是常常忽略众包任务中普遍存在的长尾现象。因此,综合考虑不同任务类型、长尾现象的特点以及工人完成任务的情况,提出构造小样本置信区间来估计工人质量,以解决工人完成任务数量普遍较少情况下的答案决策问题。首先依据黄金标准答案策略对工人质量进行预评估,根据工人质量分布分别对数值型任务和单项选择型任务采用不同的真值初始化方法;然后构造小样本置信区间以准确评估工人质量;最后进行任务答案决策并迭代更新工人质量。为了验证提出方法的有效性,实验在5个真实数据集上进行,与现有方法相比,所提方法能很好地解决长尾现象。特别是在工人完成任务数量普遍较少的情况下,提出的方法在单项选择型任务数据集中的平均准确率高达93%,相比现有方法的最好表现高出16%,且在数值型任务数据集中的*MAE*值和*RMSE*值均低于现有方法。

关键词:众包;长尾现象;小样本置信区间;工人质量估计;答案决策
中图法分类号　TP391.1

Truth Inference Based on Confidence Interval of Small Samples in Crowdsourcing

ZHANG Guang-yuan and WANG Ning
School of Computer and Information Technology,Beijing Jiaotong University,Beijing 100044,China

Abstract　Crowdsourcing is an increasingly important area of computer applications,because it can address problems that difficult for computer to handle alone. For the openness of crowdsourcing,quality control becomes one of the important challenges. In order to ensure the effectiveness of truth inference,current researches leverage answers of trustful workers to infer truths by evaluating worker quality generally. However,most existing methods ignore the long-tail phenomena in crowdsourcing,and there is a lack of researches on the truth inference when the number of tasks completed by workers is generally small. Considering the characteristics of different task types,long-tail phenomenon and worker answers,this paper constructs the confidence interval of small samples to solve truth inference when the number of tasks completed by workers are generally small. Firstly,worker quality is pre-estimated according to the gold standard answer strategy,and different truth initialization methods are adopted according to the result of pre-estimated. Then,the confidence interval of small samples is constructed to evaluate worker quality accurately. Finally,task truths are inferred and worker quality is updated iteratively. In order to verify the effectiveness of the proposed method,5 real datasets are selected to conduct experiments. Compared with the existing methods,the proposed method can solve the problem of the long tail phenomenon effectively,especially the number of tasks completed by each worker is generally small. The average accuracy of the proposed method for the single-choice tasks is as high as 93%,and higher than 16% of the best performance of the existing methods. Meanwhile,the values of *MAE* and *RMSE* of the proposed method for the numerical tasks are lower than that of the existing methods.

Keywords　Crowdsourcing,Long-tail phenomenon,Small sample confidence interval,Worker quality estimation,Truth inference

1　引言

众包即请求者将难以由计算机单独处理的任务发布到互联网,利用大众智慧来解决问题。众包是一种广泛应用于自然语言处理[1]、数据清洗[2]等领域的新型分布式问题处理方式。由于众包工人的背景和知识水平不同,以及欺骗类型工人的存在,工人的答案不是完全可信的。如何评估工人质量,并聚合工人答案得到高质量的结果是众包面临的挑战之一。

众包任务具有不同的形式[3],常见的有数值型任务和单项选择型任务。数值型任务要求工人提供一个数值答案,最

到稿日期:2019-11-11　返修日期:2020-04-27　本文已加入开放科学计划(OSID),请扫描上方二维码获取补充信息。
基金项目:国家重点研发计划项目(2018YFC0809800)
This work was supported by the National Key R&D Program of China (2018YFC0809800).
通信作者:王宁 (nwang@bjtu.edu.cn)

直接的方法是将所有工人答案的均值或中值作为每个任务的真值。Sheng 等考虑冗余策略与估计答案相关误差之间的关系，分别为所有工人的答案分配相同或者不同的权重[4]。Zhi 等针对动态变化的数值型任务，提出一种数值数据动态真值发现模型[5]。与数值型任务不同，单项选择型任务要求工人从多个候选答案中选择一个作为答案提交。少数服从多数投票法[6]是最直接的答案决策方法。期望最大化方法[7]将答案决策分为 E 和 M 两个步骤，重复迭代并更新各参数的概率，直到候选答案后验概率收敛。Li 等提出的 CRH 模型[8]是一种同时适用于数值型任务和单项选择型任务的真值推理方法，利用答案和估计真值之间的不同距离函数来识别各种数据类型的特征。

然而，目前多数答案决策方法的准确率在很大程度上依赖于工人提交答案的数量，而忽略了众包中长尾现象的存在。众包长尾现象即大多数工人完成任务的数量很少，这对于工人的质量估计是不利的[9-10]，进而影响答案决策。针对众包中的长尾现象，目前的主要方法有以下几种。

1）移除回答数量少的工人的方法[11]。此类方法适合任务量大的众包任务。然而，若存在较多此类工人，移除工人会导致众包答案数量稀疏，无法验证答案决策模型的有效性。

2）伪计数法[12]。当工人完成任务数量少时，其质量主要由伪计数主导；当工人完成足够多任务时，其质量更接近真实质量。此类方法时效高，但如何合理设置伪计数有待验证。

3）工人质量置信区间估计法[9-10]。CATD 方法通过工人答案与真值之间的差异构建置信区间来对工人质量建模；ETCIBoot 方法通过 Bootstrap 抽样方法削弱异常值对答案决策的影响，并采取迭代策略对工人质量更新。现有的置信区间估计法能有效评估完成任务数量少的工人质量，然而模型在真值初始化时平等看待每个工人，置信区间的构建受初值影响大。此外，置信区间的构建主要针对数值型问题，对单项选择型任务中工人质量的置信区间的构建缺乏针对性。

目前缺乏对工人整体完成情况都较少的答案决策的研究，现有方法对回答数量少的工人的质量所构建的置信区间跨度大，从而低估了工人质量，不利于答案决策。

针对上述问题，本文在 CATD 方法的基础上提出更为有效的解决方案，在工人完成任务数量普遍较少的情况下，分别构造适合数值型任务和单项选择型任务的小样本置信区间以对工人质量进行估计，并结合不同的答案决策方法进行答案决策。由于置信区间的构建受任务真值的影响，为了合理地初始化任务真值，首先依据黄金标准答案策略对工人质量进行预评估，根据完成每项任务的一组工人质量分布判断对该项任务采取何种真值初始化方法；其次，利用工人答案与任务真值之间的差异构造小样本置信区间，据此评估工人质量；最后，迭代推断任务真值并动态更新工人质量。

本文的主要贡献如下：1）通过工人质量预判定流程合理地初始化任务真值，为准确地推断任务真值奠定基础；2）构建小样本置信区间来对工人质量建模；3）针对两种任务类型，分别在真实数据集上将所提方法与现有方法进行对比实验，结果表明，本文提出的方法在工人完成任务数量普遍较少的情况下的答案决策性能更优，并且具有良好的时间效率。

本文第 2 节给出小样本众包答案决策的问题定义；第 3 节阐述基于小样本置信区间的答案决策算法；第 4 节对实验结果进行分析和评估；最后总结全文。

2 问题定义

本文重点研究在工人完成任务数量普遍较少的情况下，如何通过构建适合数值型任务和单项选择型任务的小样本置信区间来估计工人质量，进而推断任务真值。下面给出相关的模型定义。

定义 1（任务） 任务集合 $T=\{t_1,t_2,\cdots,t_n\}$ 包含 n 个任务，每个任务可以由多个工人完成。

定义 2（工人） 工人集合 $W=\{w_1,w_2,\cdots,w_m\}$ 包含 m 个工人。工人 $w_j\in W$ 的质量表示为 q_j，q_j 越大表明工人 w_j 越可信。记工人的质量集合为 $Q=\{q_j\mid 1\leqslant j\leqslant|W|\}$。

定义 3（答案） 每个任务 $t_i\in T$ 由 W 中的工人来回答，任务 t_i 收集到的所有答案表示为集合 $U^{t_i}=\{u_{ij}\mid 1\leqslant j\leqslant|W|\}$，其中 u_{ij} 表示工人 w_j 对任务 t_i 的答案。此外，工人 w_j 回答的所有任务记为 $T^{w_j}=\{t_g\mid 1\leqslant g\leqslant|T|\}$。

定义 4（真值） 对于任务 $t_i\in T$，存在一个推断的真值 r_i 和一个真实的真值 r_i^*。一般而言，r_i^* 是未知的。

定义 5（答案决策） 根据每个任务 $t_i\in T$ 收集到的所有答案的集合 U^{t_i} 以及完成该项任务的工人质量集合 Q，构建答案决策算法来推断任务 t_i 的真值 r_i，这个过程称为答案决策。

基于上述定义，本文将需要解决的问题定义为：给定任务集合 T 以及工人集合 W，根据工人的回答情况以及任务的类型，构建小样本置信区间以估计工人的质量集合 Q，最后根据 Q 和工人的答案集合 U^{t_i} 推断出每个任务 $t_i\in T$ 的真值 r_i。

本文提出的答案决策 SCTI（Small Sample Confidence Interval Truth Inference）算法的框架如图 1 所示，主要分为 3 个步骤：1）真值初始化，利用黄金标准答案策略对工人的质量进行预评估，根据评估结果以及任务类型选择真值初始化方式；2）工人质量估计，根据不同类型任务的特点，构建小样本置信区间以估计工人质量；3）答案决策，根据工人质量和答案推断任务真值，迭代更新任务真值和工人质量。

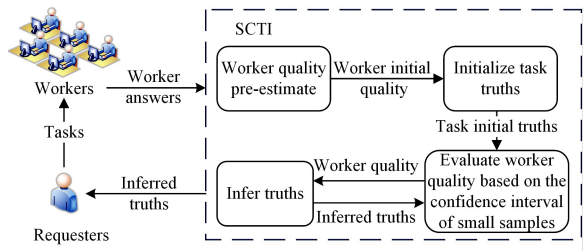


图 1 基于小样本置信区间的答案决策方法框架

Fig. 1 Framework of truth inference based on confidence interval of small samples

3 基于小样本置信区间的答案决策方法

基于小样本置信区间的答案决策方法的基本思想是通过工人答案与任务真值的差异构建小样本置信区间，以估计工人质量，并针对数值型任务和单项选择型任务进行答案决策。

3.1 任务真值初始化

现有的工作在真值初始化过程中平等地看待每一个工人，即将工人提交答案的平均值或中值作为数值型任务的初

始真值，或是通过少数服从多数方法选取工人选择最多的候选答案作为单项选择任务的初始真值[9-10]。由于置信区间的构建受初始真值的影响，若工人答案中存在异常值或者工人质量差别较大，不加区分地处理工人的答案会影响初始真值的质量，进而影响下一步的建模。为了降低初始真值受到的影响，我们设计了工人质量预判定流程，通过黄金标准答案策略获取工人质量，根据质量离散程度与阈值 ξ 的比较结果来决定采取何种真值初始化方式。若质量离散程度小于 ξ（多次实验表明，ξ 取 1.5 较优），则采取直接决策的方法，即数值型任务采取中位数方法；单项选择型任务采取少数服从多数投票方法。反之，数值型任务采取加权中值方法，单项选择型任务采取加权投票方法。

假设任务 t_i 收集 k 个来自工人集合 W 的答案，记 t_i 的初始真值为 $r_i^{(0)}$。我们首先通过黄金标准答案策略初始化工人质量，得到回答任务 t_i 的工人的初始质量集合为 $Q_i^{(0)}=\{q_j^{(0)}\mid 1\leqslant j\leqslant|W|\}$。令 $\delta\{\cdot\}$ 表示克罗内克函数，当输入的判定为真时，$\delta\{\cdot\}$ 的输出为 1；反之，输出为 0。记 T^G 为已知真值的任务集合，则工人的初始质量计算如下：

$$q_j^{(0)}=\frac{\sum_{0\leqslant i\leqslant|T^G|}\delta\{u_{ij}=r_i^*\}}{|T^G|} \tag{1}$$

若工人之间的质量相差过大，则表明回答该任务的工人质量分布不均衡，真值初始化需要对工人的答案区别对待。我们利用信息熵获得工人质量的离散程度：

$$indicator=-\sum_{1\leqslant j\leqslant|W|}q_j^{(0)}\log_2 q_j^{(0)} \tag{2}$$

例 1　设有 A,B 两组工人，其初始质量分别为 $\{0.14,0.02,0.5\}$，$\{0.28,0.8,0.63\}$，由式(2)可以得到两组工人的初始质量分布分别为 2.67 和 1.08。

第一组工人质量的离散程度 2.67 大于 ξ，若仅通过少数服从多数法初始化真值，则低质量工人的答案占主导，初始答案不可靠，会影响工人质量置信区间的构建。针对大于离散程度阈值的不同任务类型，不同的真值初始化方法如下。

1）数值类型任务。针对加权平均方法对异常值存在敏感的问题，本文采取加权中值法初始化真值[8]，满足式(3)两个条件的 u_g 即为最终任务的初始化真值。

$$\sum_{j:u_j<u_g}q_j<\frac{1}{2}\sum_{j=1}^{|W|}q_j \text{ and } \sum_{j:u_j>u_g}q_j\leqslant\frac{1}{2}\sum_{j=1}^{|W|}q_j \tag{3}$$

针对例 1 中的 A 组工人，假设工人对一项数值任务 t_i 提交的答案为 $\{0.42,0.89,0.58\}$，任务真值为 $r_i^*=0.6$。根据加权中值法推断得到 $r_i^{(0)}=0.58$，而加权投票计算所得 $r_i^{(0)}=\frac{0.14\cdot0.42+0.02\cdot0.89+0.5\cdot0.58}{0.14+0.02+0.5}=0.55$，可知加权中值法的结果更接近真值 0.6。

2）单项选择类型任务。利用工人的初始质量，本文采用加权投票法初始化真值：

$$r_i^{(0)}=\arg\max_{a_{ic}\in A^{t_i}}\sum_{j\in|W|}q_j\cdot\delta\{u_{ij}=a_{ic}\} \tag{4}$$

其中，$A^{t_i}=\{a_{i1},a_{i2},\cdots,a_{iz}\}$ 为任务 t_i 的候选答案集合。

针对例 1 中的 A 组工人，假设工人对一项选择任务 t_i 提交的答案为 $\{A,A,C\}$，任务真值 $r_i^*=C$。依据质量的加权投票法推断得到 $r_i^{(0)}=C$，而采取少数服从多数投票法则可得 $r_i^{(0)}=A$，可知加权投票更准确。

3.2　工人质量估计

假设某一组任务完成或某一任务周期结束后，工人完成的任务集合为一组样本值，存在一组任务的任务总数较少或是工人在一个任务周期完成的任务数量较少的情况，则工人完成的任务集合为小样本。本文着重解决小样本情况下的众包答案决策问题，在现有方法的基础上，根据不同类型的任务特点，利用不同的分布来描述工人的行为。此外，为了使置信区间更集中，本文构建小样本置信区间估计工人质量。

1）数值类型任务。假设工人 w_j 的误差系数（即 w_j 提交的答案与真值之间的差异）是连续值，服从正态分布，记为 $\varepsilon_j\sim(v_j,\sigma_j^2)$。由于工人完成的任务集的总体方差 σ_j^2 未知，当工人完成任务的数量 $|T^{w_j}|$ 较少时，工人 w_j 的误差系数均值经过标准化后的随机变量服从自由度为 $(|T^{w_j}|-1)$ 的 t 分布：

$$\frac{\bar{v}_j-v_j}{s_j/\sqrt{|T^{w_j}|}}\sim t(\sqrt{|T^{w_j}|}-1) \tag{5}$$

其中，利用目前工人完成任务所提交的答案与真值之间的差异的方差 $s_j^2=\frac{\sum_{0\leqslant i\leqslant|T^{w_j}|}(u_{ij}-r_i^{(0)})^2}{|T^{w_j}|}$ 来估计 σ_j^2。$\bar{v}=\frac{\sum_{i\in T^{w_j}}(u_{ij}-r_i^{(0)})}{|T^{w_j}|}$ 为工人 w_j 目前完成任务的答案的差异的均值。给定 $\alpha=0.05$，工人 w_j 的总体误差系数均值 v_j 在 $1-\alpha$ 置信水平下的置信区间为：

$$\left(\bar{v}_j-t_{\alpha/2}(\sqrt{|T^{w_j}|}-1)\frac{s_j}{\sqrt{|T^{w_j}|}},\ \bar{v}_j+t_{\alpha/2}(\sqrt{|T^{w_j}|}-1)\frac{s_j}{\sqrt{|T^{w_j}|}}\right) \tag{6}$$

考虑最糟糕的场景，本文使用 $1-\alpha$ 置信区间的上限作为 v_j 的估计量来反映工人的质量 θ_j，记为 $\theta_j=\bar{v}_j+t_{\alpha/2}(\sqrt{|T^{w_j}|}-1)\frac{s_j}{\sqrt{|T^{w_j}|}}$。$\theta_j$ 越大，即工人 w_j 的误差系数的均值越大，表明 w_j 提交错误答案的概率越大。随着工人 w_j 完成任务数量 $|T^{w_j}|$ 的增加，置信区间变窄，θ_j 估计更准确。

2）单项选择类型任务。对于单项选择任务，工人的回答要么正确，要么错误，是离散值。设工人 w_j 正确回答一项单项选择任务 t_i 的概率为 q_j，该工人回答 t_i 的结果 x_i 服从以 q_j 为参数的二点分布，记为 $x_i\sim b(1,q_j)$。

设 $x_1,x_2,\cdots,x_n$ 是工人 w_j 目前完成任务所提交的答案，它们均服从二点分布 $b(1,q_j)$。其中，q_j 可以利用工人 w_j 在目前已完成任务中的校正准确率 $\hat{q}_j=\frac{2+\sum_{i=1}^{n}\delta\{x_i=r_i\}}{n+4}$ 进行估计。给定 $\alpha=0.05$，根据 Wald 校正区间方法构建小样本置信区间[13]，工人 w_j 的质量 q_j 的置信水平为 $1-\alpha$ 的置信区间可以表示为：

$$\left(\hat{q}_j-z_{(1-\alpha/2)}\sqrt{\frac{\hat{q}_j(1-\hat{q}_j)}{n+4}},\ \hat{q}_j+z_{(1-\alpha/2)}\sqrt{\frac{\hat{q}_j(1-\hat{q}_j)}{n+4}}\right) \tag{7}$$

同样地，本文使用 $1-\alpha$ 置信区间的下限作为工人 w_j 的质量，即 $q_j=\hat{q}_j-z_{(1-\alpha/2)}\sqrt{\frac{\hat{q}_j(1-\hat{q}_j)}{n+4}}$。随着工人 w_j 完成任

务数量 n 的增加，置信区间变窄，q_j 估计更准确。

3.3 真值推断

本文根据不同任务类型的特点，构建不同的答案决策模型来推断任务的答案，并结合上述工人质量模型动态更新工人的质量。

1）数值类型任务。假设一项数值类型任务 t_i 由一组工人 $W^{t_i}=\{w_j \mid w_j \in W\}$ 完成，根据本文提出的方法初始化任务 t_i 的真值并利用小样本置信区间进行工人质量估计，得到工人初始的误差系数为 $O=\{\theta_j \mid 0 \leqslant \theta_j \leqslant |W^{t_i}|\}$。本文通过最小化误差系数的加权平均策略推断任务真值，任务 t_i 的真值 r_i 的计算式如下：

$$r_i=\arg\min_{\theta_j \in O} \frac{\sum_{1 \leqslant j \leqslant |W^{t_i}|} u_{ij} \cdot \theta_j}{\sum_{\theta_j \in O} \theta_j} \tag{8}$$

2）单项选择类型任务。给定一项单项选择类型任务 t_i、一组候选答案 $A^{t_i}=\{a_{i1},a_{i2},\cdots,a_{ik}\}$、任务 t_i 收集的答案集合 U^{t_i}，以及基于小样本计算的工人质量集合 $Q=\{q_j \mid 1 \leqslant j \leqslant |W|\}$，本文采用贝叶斯决策方法计算任务 t_i 的后验概率，选取后验概率最大的候选答案作为任务 t_i 的真值 r_i。记候选答案的后验概率集合为 $\rho^{t_i}=\{a_{i1},a_{i2},\cdots,a_{ik}\}$。由于先验知识未知，将每个候选答案先验概率设为 $\frac{1}{k}$（k 为候选答案的个数），则任务 t_i 的候选答案 a_{ic}（$a_{ic} \in A^{t_i}$）的后验概率计算如下：

$$\rho_{ic}=P(r_i=a_{ic} \mid U^{t_i})=\frac{P(U^{t_i} \mid r_i=a_{ic})P(r_i=a_{ic})}{P(U^{t_i})}$$

$$=\frac{\prod_{u_{ij} \in U^{t_i}} (q_j)^{\delta\{u_{ij}=r_i\}} \left(\frac{1-q_j}{k}\right)^{\delta\{u_{ij} \neq r_i\}}}{\sum_{a_{ib} \in A^{t_i}} \prod_{u_{ij} \in U^{t_i}} (q_j)^{\delta\{u_{ij}=r_i\}} \left(\frac{1-q_j}{k}\right)^{\delta\{u_{ij} \neq r_i\}}} \tag{9}$$

根据式(9)依次得到每个候选答案的后验概率后，选取后验概率最大的候选答案作为答案。

$$r_i=\arg\max_{\rho_{ic} \in \rho^{t_i}} \{\rho_{ic}\} \tag{10}$$

本文采取迭代方式动态更新工人质量，即当一组任务决策完成，则根据推断的任务答案更新工人的质量，直至满足迭代条件。

4 实验结果与分析

4.1 实验设置

1）实验环境：硬件环境为 Intel(R) Core(TM) i5-3337U CPU@1.80 GHz，4.00 GB 内存，500 GB 硬盘；软件环境为 Windows10 专业版 64 位操作系统。

2）数据集：针对数值任务和单项选择任务，分别使用不同的真实数据集验证本文所提方法的性能。其中，数值任务使用的数据集为 Emotion[1] 以及 Movie Rating[2]；单项选择任务使用的数据集为具有两个以上候选答案的数据集 AdultContent2[3] 以及决策类型数据集 rteStandardized[4] 和 step[5]。

3）评价指标：数值任务使用的评价指标为平均绝对误差（Mean Absolute Error，MAE）和均方根误差（Root Mean Square Error，RMSE）。这两项指标的值越低，证明方法的有效性越高。单项选择任务在具有两个以上候选答案的数据集上使用的评价指标为准确率（Accuracy），在决策类型数据集上的评价指标为准确率以及 F1 分数（F1-score）。这两项指标的值越高，证明方法的有效性越高。具体的评价函数如下：

$$MAE=\frac{\sum_{t_i \in T} |r_i-r_i^*|}{|T|} \tag{11}$$

$$RMSE=\sqrt{\frac{\sum_{t_i \in T} (r_i-r_i^*)^2}{|T|}} \tag{12}$$

$$Accuracy=\frac{\sum_{t_i \in T} \delta\{r_i=r_i^*\}}{|T|} \tag{13}$$

$$F1\text{-}score=\frac{2 \cdot \sum_{t_i \in T} \delta\{r_i=1\} \cdot \delta\{r_i^*=1\}}{\sum_{t_i \in T} (\delta\{r_i=1\}+\delta\{r_i^*=1\})} \tag{14}$$

4）阈值：经过多次实验，工人质量离散度的阈值 ξ 最终确定为 1.5。

4.2 对比方法

本文在不同数据集上将所提算法与现有的数值类型和选择类型的答案决策算法进行了性能比较。具体对比算法有：针对数值任务的中位数法 Median[3] 和平均值法 Mean[3]，针对选择任务的少数服从多数投票法 MV(Majority Voting)[6] 和期望最大化算法 EM(Expectation-Maximization)[7]，以及可以同时用于数值任务和选择任务的答案决策算法 CRH (Conflict Resolution on Heterogeneous Data)[8] 和 CATD (Confidence-aware Truth)[9]。CATD 考虑了长尾现象，但是所构建的置信区间跨度大，容易低估工人的质量。本文方法综合考虑了不同任务类型的长尾现象和小样本数据集的特点，构造了适合小样本数据集的置信区间以对工人质量进行估计，改善了答案决策的质量。

4.3 算法性能评估

本文算法主要解决工人回答普遍较少的情况下的众包答案决策问题。比较实验分为两部分：首先，控制每个工人完成任务的数量在 5～30 项的范围内；其次，为了验证方法的适应性，将增加工人完成任务的数量（大于 30 项），为了体现长尾问题，将该部分实验的场景设为存在部分工人完成任务数量较少的情况。

4.3.1 数值类型任务的有效性评估

本文将所提方法（SCTI）与其他 4 种处理数值任务的方法（Mean，Median，CRH 和 CATD）进行对比，表 1 和表 2 分别展示了不同方法在 Emotion 和 Movie Rating 数据集上的 *MAE* 和 *RMSE* 值，其中每一列代表不同方法在对应工人完成任务数量上的实验结果。

可以观察到，Mean 及 Median 方法得出的答案与真值差距最大，本文提出的方法在两个数据集上的表现均远远优于其他算法。值得注意的是，虽然在答案决策中，CATD 同样考虑了长尾现象并使用了置信区间估计工人质量，但本文方法使用了小样本置信区间，使得估计的工人质量的置信区间更窄，能够更准确地捕捉工人质量与工人完成情况之间的内在

1) http://dbgroup.cs.tsinghua.edu.cn/ligl/crowddata/

2) https://grouplens.org/datasets/movielens/

3) https://github.com/ipeirotis/Get-Another-Label/tree/master/data

4) http://ir.ischool.utexas.edu/square/data.html

5) https://github.com/DMhouping/ETCIBoot

关系，因此，SCTI的答案决策效果远超CATD。CRH在数值类型任务中使用了异常值不敏感的加权中位数法，其答案决策的效果优于CATD，而SCTI综合考虑了工人质量、任务真值初始化以及小样本，因此获得了最佳性能。此外，在这两个数据集中，随着工人完成任务数量的增加，工人的积极性随之下降，提交了低质量的答案。SCTI算法的*MAE*值和*RMSE*值出现小幅上升，但上升幅度逐渐变小（当任务数从25增加到30时，*MAE*值和*RMSE*值分别仅增加了0.018和0.008）；相对而言，CRH算法使用了异常值不敏感的加权中位数法，工人答案质量的变化对其影响较小，*MAE*值和*RMSE*值即使随工人回答的任务数增加略有下降，但下降幅度逐渐变小（当任务数从25增加到30时，*MAE*值和*RMSE*值分别仅下降了0.003和0.008）。总体而言，SCTI在数值型任务上的表现与其他算法相比仍然是最优的。

表1 不同方法在Emotion数据集上的实验结果

Table 1 Experimental results of different methods on the Emotion dataset

Method	*MAE*						*RMSE*					
	5	10	15	20	25	30	5	10	15	20	25	30
Mean	17.759	17.795	17.759	17.722	17.766	17.831	28.169	28.136	28.186	28.201	28.180	28.181
Median	17.673	17.749	17.673	17.749	17.859	17.934	28.051	27.986	27.909	27.944	27.913	27.912
CRH	10.307	10.300	10.307	9.280	8.736	8.620	19.425	17.451	19.278	17.772	16.878	16.220
CATD	17.951	17.951	17.951	17.924	17.91	17.91	28.302	28.302	28.302	28.254	28.240	28.240
SCTI	**1.256**	**0.856**	**1.256**	**1.461**	**1.779**	**2.067**	**3.887**	**5.388**	**6.960**	**7.339**	**7.937**	**8.077**

表2 不同方法在Movie Rating数据集上的实验结果

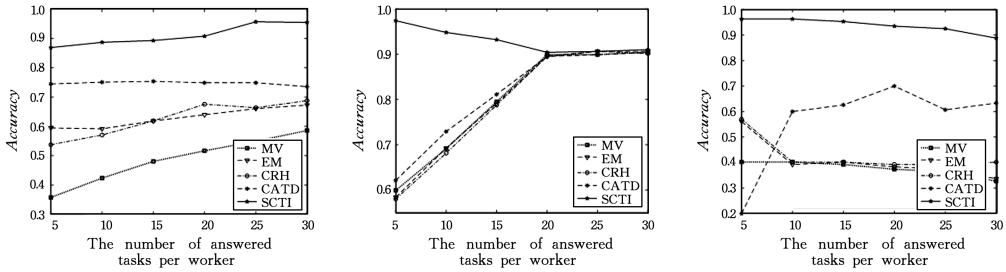
Table 2 Experimental results of different methods on Movie Rating dataset

Method	*MAE*						*RMSE*					
	5	10	15	20	25	30	5	10	15	20	25	30
Mean	2.672	2.196	1.898	1.713	1.566	1.457	2.948	2.596	2.366	2.216	2.089	1.990
Median	2.664	2.185	1.883	1.693	1.544	1.430	2.946	2.594	2.361	2.209	2.082	1.982
CRH	0.718	0.692	0.642	0.619	0.594	0.591	0.940	0.935	0.862	0.848	0.827	0.819
CATD	3.348	3.336	3.326	3.316	3.307	3.299	3.390	3.378	3.368	3.358	3.349	3.341
SCTI	**0.181**	**0.304**	**0.346**	**0.371**	**0.390**	**0.408**	**0.487**	**0.651**	**0.657**	**0.670**	**0.679**	**0.687**

在工人完成任务数量普遍较少的情况下，本文方法在两个数值类型任务数据集上的平均绝对误差值均低于9，均方根误差值均低于2，误差远低于对比方法，显示了本文方法在数值型任务上使用小样本置信区间的优越性。

4.3.2 单项选择类型任务的有效性评估

本实验同样将每个工人完成任务的数量设置在5～30项的范围内，比较各个答案决策方法在选择类型数据集上的有效性。图2(a)－图2(e)分别给出了5种方法在AdultContent2，rteStandardized以及step数据集上的实验结果。本文方法的准确率和*F*1分数在3个数据集上均能达到0.95以上，始终高于其他所有方法，且具有较高的稳定性。实验结果表明，SCTI在单项选择任务和决策任务上的实验效果均远远优于其他比较方法，这是因为该方法通过真值初始化的预处理，给工人的质量估计提供了有效可靠的输入，并结合小样本置信区间，可以更准确地估计工人的质量，从而能在工人完成任务数量普遍较少的情况下，提高答案决策的有效性。



(a) Accuracy on AdultContent dataset (b) Accuracy on rteStandardized dataset (c) Accuracy on step dataset

(d) F1-score on rteStandardized dataset (e) F1-score on step dataset

图2 不同方法在3个单项选择数据集上的实验结果

Fig. 2 Experimental results of different methods on 3 single-choice datasets

4.3.3 适应性评估

为了验证方法的适应性，在保留部分完成任务数量较少的工人的情况下，我们将工人完成任务的数量设为40～100，分别得到不同方法在5个不同数据集上相应的评价指标值，

最后取平均值作为实验结果。

表 3 列出了不同方法在多数工人完成任务数量较多情况下的实验结果。可以看出，SCTI 方法在 5 个数据集上的表现仍然优于其他方法。综合上述实验，SCTI 在工人完成任务数量普遍较少或者部分较少的情况下，得到答案决策结果均优于其他方法。

表 3 适应性评估结果

Table 3 Adaptability assessment results

Method	AdultContent2	rteStandardized		step		Emotion		Movie Rating	
	Accuracy	*Accuracy*	*F1-score*	*Accuracy*	*F1-score*	*MAE*	*RMSE*	*MAE*	*RMSE*
Mean	—	—	—	—	—	16.858	26.732	1.035	1.561
Median	—	—	—	—	—	17.091	26.110	0.999	1.548
MV	0.676	0.910	0.910	0.437	0.476	—	—	—	—
EM	0.701	0.910	0.908	0.555	0.622	—	—	—	—
CRH	0.724	0.906	0.905	0.625	0.628	9.779	18.302	0.547	0.773
CATD	0.727	0.912	0.911	0.658	0.577	17.715	28.003	3.255	3.296
SCTI	**0.946**	**0.916**	**0.915**	**0.773**	**0.685**	**5.235**	**13.027**	**0.450**	**0.711**

4.3.4 运行时间评估

我们使用运行时间平均值来对时间效率进行评估。由表 4 可知，Mean，Median，MV 方法的运行时间较短，但它们不能有效处理长尾问题。由于算法的迭代性，EM 方法的平均运行时间均高于其他算法。可以看到，本文方法不仅能处理长尾现象，且有较高的时间效率。

表 4 运行时间对比结果

Table 4 Comparison results of runing time

Method	Average Runtime				
	Adult-Content2	rteStan-dardized	step	Emotion	Movie Rating
Mean	—	—	—	0.0952	0.203
Median	—	—	—	0.0283	0.149
MV	1.069	0.0403	0.014	—	—
EM	70.989	2.9810	0.611	—	—
CRH	2.620	0.1202	0.030	0.2867	1.371
CATD	18.091	0.7503	0.117	0.4233	4.905
SCTI	**4.906**	**0.1652**	**0.026**	**0.1507**	**0.513**

结束语 本文研究长尾现象下众包的答案决策问题。通过分析众包数据中的长尾现象，结合数值类型任务和单项选择类型任务的特点，综合考虑了真值初始化、工人完成情况以及工人质量，设计了一个基于小样本置信区间的工人质量估计算法，从而提高了答案决策质量。未来的研究将尝试应用本文的框架和方法来处理更复杂的任务，考虑更多的影响答案决策的因素，如任务难度和任务领域等。

参 考 文 献

[1] ZAIDAN O F, CALLISON-BURCH C. Crowdsourcing translation: Professional quality from non-professionals[C]// Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, 2011: 1220-1229.

[2] CHU X, MORCOS J, ILYAS I F, et al. Katara: A data cleaning system powered by knowledge bases and crowdsourcing[C]// Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data. ACM, 2015: 1247-1261.

[3] ZHENG Y, LI G, LI Y, et al. Truth inference in crowdsourcing: Is the problem solved? [J]. Proceedings of the VLDB Endowment, 2017, 10(5): 541-552.

[4] SHENG K, GU Z, MAO X, et al. Answer inference for crowdsourcing based scoring[C]// 2014 IEEE Global Communications Conference. IEEE, 2014: 2733-2738.

[5] ZHI S, YANG F, ZHU Z, et al. Dynamic Truth Discovery on Numerical Data[C]// 2018 IEEE International Conference on Data Mining (ICDM). IEEE, 2018: 817-826.

[6] PARAMESWARAN A G, PARK H, GARCIA-MOLINA H, et al. Deco: declarative crowdsourcing[C]// Proceedings of the 21st ACM International Conference on Information and Knowledge Management. ACM, 2012: 1203-1212.

[7] DAWID A P, SKENE A M. Maximum likelihood estimation of observer error-rates using the EM algorithm[J]. Journal of the Royal Statistical Society: Series C (Applied Statistics), 1979, 28(1): 20-28.

[8] 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. ACM, 2014: 1187-1198.

[9] 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.

[10] XIAO H, GAO J, LI Q, et al. Towards confidence in the truth: A bootstrapping based truth discovery approach[C]// Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2016: 1935-1944.

[11] HUNG N Q V, TAM N T, TRAN L N, et al. An evaluation of aggregation techniques in crowdsourcing[C]// International Conference on Web Information Systems Engineering. Heidelberg: Springer, 2013: 1-15.

[12] LI Y, LIU C, DU N, et al. Extracting medical knowledge from crowdsourced question answering website[J]. IEEE Transactions on Big Data, 2016: 1-1.

[13] BROWN L D, CAI T T, DASGUPTA A. Interval estimation for a binomial proportion[J]. Statistical Science, 2001, 16(2): 101-117.

ZHANG Guang-yuan, born in 1994, M.S., is a member of China Computer Federation. Her main research interests include data quality and crowdsourcing.

WANG Ning, born in 1967, Ph.D, professor, is a member of China Computer Federation. Her main research interests include web data integration, big data management and crowdsourcing.