

基于信任模型的鲁棒众包数据分析方法

孙杰¹, 陈敏², 焦玉全²

(1. 南京邮电大学 计算机技术、软件工程、空间安全学院, 江苏 南京 210003;

2. 南京邮电大学, 江苏 南京 210003)

摘要:众包是互联网大发展趋势下衍生的一种非常流行的新型商业模式,企业将过去由员工执行的任务分配出去,以自由自愿的形式外包给非特定的(通常是大型的)大众志愿者来完成。但是,在这一过程中,有些志愿者并没有认真地完成任务,为了骗取佣金,使利益最大化,往往会提供虚假数据,造成任务结果数据虚实混杂,雇主无法得到高准确度的众包任务结果。针对这个问题,在对现有众包数据分析方法进行分析之后,提出了一种新的众包数据分析方法。在考虑工作者历史信誉度的基础上,对部分任务结果采用投票一致性规则来分析,将工作者历史信誉度和在本次任务中提交的结果数据与贝叶斯算法模型相结合,得出工作者在本次任务中提交结果数据的最终准确度信息。实验结果表明,使用该方法筛选出的工作者可靠性更高。

关键词:众包;准确度;信任模型;投票一致性规则;贝叶斯算法

中图分类号:TP181

文献标识码:A

文章编号:1673-629X(2021)05-0026-05

doi:10.3969/j.issn.1673-629X.2021.05.005

Robust Crowdsourcing Data Analysis Method Based on Trust Model

SUN Jie¹, CHEN Min², JIAO Yu-quan²

(1. School of Computer Science, Nanjing University of Posts and Telecommunications, Nanjing 210003, China;

2. Nanjing University of Posts and Telecommunications, Nanjing 210003, China)

Abstract: Crowdsourcing is a quite popular new business model derived from the great development trend of the Internet. Enterprises assign tasks previously performed by employees to non-specific (usually large) public volunteers in a free and voluntary manner. However, in this process, some volunteers did not complete the task seriously. In order to cheat the commission and maximize the benefits, they often provided false data, resulting in mixed actual and actual task result data, so the employers could not obtain high-accuracy crowdsourcing task results. To solve this problem, after analyzing the existing crowdsourcing data analysis methods, we propose a new crowdsourcing data analysis method. On the basis of considering the historical credibility of workers, we use voting consistency rules to analyze some task results. The historical credibility of workers and the result data submitted in this task are combined with Bayesian algorithm model to obtain the final accuracy information of the result data submitted by workers in this task. The experiment shows that the workers screened by this method have higher reliability.

Key words: crowdsourcing; accuracy; trust model; rule of voting consistency; Bayesian algorithm

0 引言

随着通信基础设施和网络信息技术的迅速发展,依托于网络环境的众包^[1]研究也在逐步深入。互联网具备的覆盖面广、无地域限制等优势,也使得过去只能面向专业机构的外包可以通过互联网扩展到全社会,国内外众多网络化众包服务市场日趋完善。

大量的志愿者得益于众包提供的工作模式,获得了不小的报酬。但是,在这一过程中,有些志愿者并没有认真地完成任务,为了骗取佣金,使利益最大化,往

往会提供虚假数据^[2],导致众包任务结果准确度不高。针对这一问题,现有的解决方式多为采用基于黄金标准数据策略^[3-4]的研究方法,在任务开始之前,对工作者的工作能力进行检测评估,筛选出符合要求的工作者来完成任务。但是这种方式局限性比较大,对于恶意工作者^[5-7]的预防效果比较低,最终所得结果准确度不高,无法满足雇主的需求。

该文提出一种新的方法,在考虑众包工作者历史信誉度的基础上,对部分任务结果采用投票一致性策

收稿日期:2020-06-03

修回日期:2020-10-09

基金项目:国家自然科学基金面上项目(61571238)

作者简介:孙杰(1993-),男,硕士研究生,研究方向为机器学习、贝叶斯算法等。

略^[8-9]进行分析,然后将二者结合于贝叶斯模型^[10-12],推算出工作者在此次任务中提交任务结果的验后准确度。下面分别对信任模型构建、计算过程、具体实施流程以及实验结果进行介绍。

1 贝叶斯信任模型构建

该文提出贝叶斯信任模型(Bayesian trust model, BTM),主要将工作者历史信誉信息^[13]和此次任务结果结合于贝叶斯算法来分析,提高对众包任务结果质量判别的准确度。贝叶斯信任模型的流程图如图1所示。

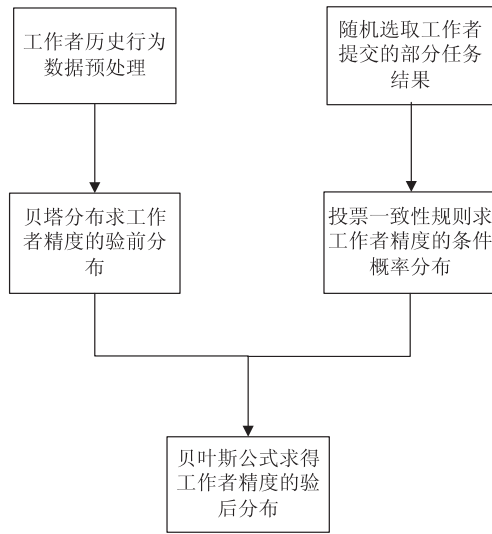


图1 贝叶斯信任模型流程图

模型构建主要分为三步:

首先,评估工作者的历史信誉度。众包工作者的工作态度无法直接通过观察来了解,但是,雇主可以通过分析每次任务审核方(例如教育众包平台^[14])的反馈信息来分析该名众包工作者的历史任务完成满意度,得出该名工作者对待任务的态度,即众包工作者的可靠性。因为工作者的工作状态往往在某一段时间内比较稳定,所以模型中选取众包工作者最近完成的 k 次任务结果准确度来计算工作者的历史信誉值。

其次,分析此次任务结果。众包工作者在提交任务结果之后,雇主对任务结果进行预处理,然后随机选取其中部分任务结果,采用投票一致性规则对选取的任务结果进行检测分析,得出工作者精度的条件概率分布。

最后,在求得工作者精度的验前分布和条件概率分布之后,将二者结合于贝叶斯算法模型,得出工作者本次任务中的工作精度。

2 贝叶斯信任模型计算

2.1 工作者精度验前分布

工作者工作精度 a ,表示工作者完成任务的准确

度。 a 的验前分布 $g(a_m), m = (1, 2, \dots, m)$,表示工作者总数为 m 。此处,采用贝塔分布来计算众包工作者工作精度的验前分布。

贝塔分布的概率密度函数为:

$$f(x; \alpha, \beta) = \frac{x^{\alpha-1} (1-x)^{\beta-1}}{\int_0^1 u^{\alpha-1} (1-u)^{\beta-1} du} = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1} (1-x)^{\beta-1}, x \in (0, 1) \quad (1)$$

其中, α 、 β 分别为贝塔分布的两个参数, Γ 为伽马函数,其中:

$$\Gamma(\alpha) = \int_0^1 t^{\alpha-1} e^{-t} dt \quad (2)$$

$$\Gamma(\alpha + \beta) = \alpha\Gamma(\alpha), \alpha > 0 \quad (3)$$

当 α 为正整数时:

$$\Gamma(n) = (n-1)! \quad (4)$$

贝塔分布的均值为:

$$\mu = E(x) = \frac{\alpha}{\alpha + \beta} \quad (5)$$

方差为:

$$\text{Var}(x) = E(x - \mu)^2 = \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)} \quad (6)$$

因为众包工作者的工作状态大多在某段时间内相似,所以在算法中取众包工作者最近完成的 k 次任务结果精度来计算工作者历史信誉值,任务完成精度用 $a = (a_1, a_2, \dots, a_k)$ 表示,由此可得:

$$\bar{a} = \frac{1}{k} \sum_{i=1}^k (a_i) = \frac{\alpha}{\alpha + \beta} \quad (7)$$

$$D(X) = \frac{1}{k-1} \sum_{i=1}^k (a_i - \bar{a})^2 = \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)} \quad (8)$$

通过式(7)和式(8),可以得到 α 和 β 的值分别为:

$$\alpha = \bar{a} \left(\frac{(1 - \bar{a})\bar{a}}{D(X)} - 1 \right) \quad (9)$$

$$\beta = (1 - \bar{a}) \left(\frac{(1 - \bar{a})\bar{a}}{D(X)} - 1 \right) \quad (10)$$

求得 α 和 β 的值以后,即可以求得众包工作者工作精度的验前分布。

2.2 工作者精度条件概率分布

以往,雇主在发放众包任务之前,会先提供一些黄金标准数据任务来让众包工作者完成(任务结果已知),通过众包工作者在黄金标准数据任务中的表现来评估工作者的业务能力,达到筛选可靠众包工作者的目的。但是,这种筛选方法存在局限性,有些不良工作者为了达到接任务的目的,在做黄金测评时很认真,

甚至,有些能力不足的工作者会通过其他手段来通过测评。但是,一旦接到众包任务,在接下来的任务完成过程中会出现对待任务态度消极的现象,更有甚者会出现恶意捣乱,草草了事等情况,这些情况会导致工作者提交的任务结果不准确,无法满足雇主的要求。所以,针对这一情况,对以往在任务开始之前对工作者进行黄金标准数据测评的方法进行改良,改为在工作者提交任务结果之后,抽取部分任务结果来检查,评估所抽取任务结果的准确度。对于抽检部分,采用投票一致性规则来进行检测分析。由于完成一个任务或者回答一个问题并不能完全体现工作者的工作精度,对众包工作者工作情况的判断存在局限性,所以在文章中,将多个任务分给多名工作者来完成,然后通过投票一致性策略来聚合所有的投票结果,达到更加精确地计算工作者的工作精度的目的。

具体过程如下:

假设将 n 份任务发放给 m 名工作者去完成 (m, n 均已知),工作者提交的任务结果用 $R_{n,m} = (r_{1,1}, r_{1,2}, \dots, r_{1,m}; r_{2,1}, r_{2,2}, \dots, r_{2,m}; \dots; r_{n,1}, r_{n,2}, \dots, r_{n,m})$ 表示,其中 n 表示问题数量, m 表示众包工作者数量, $r_{1,1}, r_{1,2}, \dots, r_{1,m}$ 表示所有众包工作者提交第一个问题的答案。

根据众包工作者提交的任务结果,可以得到所有问题结果的一致性数据为:

$$\bar{r}_n = \frac{1}{m}(r_{n,1} + r_{n,2} + \dots + r_{n,m}), n = (1, 2, \dots, n) \quad (11)$$

式(11)表示 m 名工作者提交第 n 个问题结果的均值,即工作者提交第 n 个问题结果的一致性数据。

$$U_{m,r_n} = \begin{cases} 1, & |r_{n,m} - \bar{r}_n| < \omega \\ 0, & \text{其他} \end{cases}$$

$$m = (1, 2, \dots, m), n = (1, 2, \dots, n) \quad (12)$$

式(12)为判别公式,判断第 m 名工作者对第 n 个问题的作答结果是否满足雇主的要求。其中, ω 为判断任务精度的阈值。如果 $r_{n,m}$ 和 \bar{r}_n 的差距小于雇主设定的阈值,即 U_{m,r_n} 的值等于 1,表示该名工作者提交的任务结果满足雇主要求,可以被采纳。反之,不可以被采纳。

由式(12)已知可得,每位工作者提供任务结果的准确度为:

$$V_m = \frac{\sum_{n=1}^n (U_{m,r_n})}{n}, (m = 1, 2, \dots, m) \quad (13)$$

由式(1)和式(13)可以求得条件概率为:

$$f(R | a_m) = \exp\left(\frac{-|g(a_m) - V_m|}{\tau}\right) \quad (14)$$

$$m = (1, 2, \dots, m)$$

其中, $0 < \tau < 1$ 是预置参数。 $g(a_m)$ 表示参与任务的 m 名工作者信誉度,即根据历史行为数据所求出的综合信誉度信息。

2.3 工作者精度验后分布

求得工作者精度的验前分布和条件概率之后,由贝叶斯公式可得,工作者精度验后分布为:

$$g(a_m | R) = \frac{f(R | a_m)g(a_m)}{\int f(R | a_m)g(a_m)} \quad (15)$$

在求得工作者精度的验后分布之后,雇主可以根据自己对任务准确度的需求,设定阈值,筛选出工作者精度符合要求的工作者提交的任务结果数据。

3 实验

为验证介绍的信任模型数据分析结果的准确度,选用 Dog and Cat Recognition (DCR) 数据集^[15] 和 Comment Sense Question (CSQ) 数据集^[16] 进行仿真实验。两种数据集都是二元类问题数据集,问题答案均为是或否。并将实验结果与黄金标准数据策略检测方法(简称 Gold 方法)作对比。

3.1 实验参数设置

实验采用的编程语言为 Matlab,采用的编辑工具为 Matlab R2017(a)。

实验中选择数据集前 20% 的问题答案作为评估数据来获得工作者精度验前分布参数 (α, β)。DCR 数据集中共包含 300 名工作者提供的 1 000 张图片答案,选择前 200 张图片答案,分为 10 组,每组 20 张图片答案,表示众包工作者前 10 次工作完成准确度。CSQ 数据集中包含 164 名工作者提交的 164 个问题答案,选择前 80 个问题答案,分为 10 组,每组 8 个问题答案,表示工作者前 10 次工作完成的准确度。

在数据集剩下的 80% 数据中,选择 20% 作为抽检数据,对工作者的当前表现情况进行评估,剩下的 60% 数据作为正式任务评估该方法的有效性。另外,两个数据集中所有问题的答案均为已知。

对于 ω 的确定:因为问题的结果都已转化为二元类问题答案,所以根据投票一致性规则,在式(12)中, r 的均值肯定接近 0 或者接近 1,最不可能出现的情况为 0.5,所以选择 0.5 作为阈值,在此处应为最佳,式(14)中, $\tau = 0.1$ 。

3.2 实验结果

在实际任务中,雇主对任务精度的需求各不相同。有的雇主对任务精度的需求比较低,例如为 70%,只要工作者提交的任务结果准确度达到 70% 就可以被采纳使用。但是有的雇主对任务结果的精度要求比较高,例如为 90%,他们的任务比较特殊,任务精度越高

越好,只有达到 90% 才可以满足需要。因此,分别计算出基于信任模型的鲁棒众包数据分析方法(简称

Trust 方法)和黄金标准数据策略(简称 Gold 方法)在不同精度需求时候筛选出的人数,如图 2 和图 3 所示。

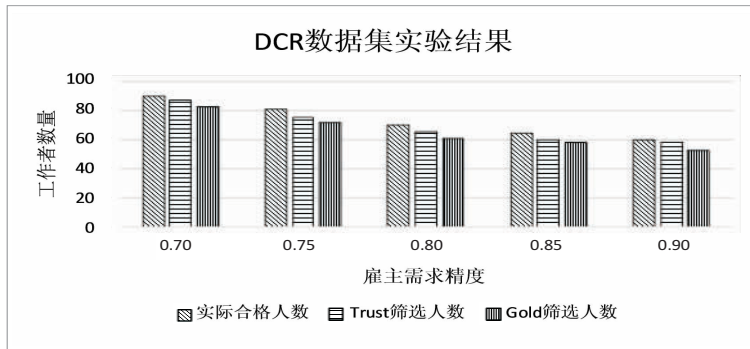


图 2 DCR 数据集中两种方法筛选的工作者与实际合格工作者数量的比较

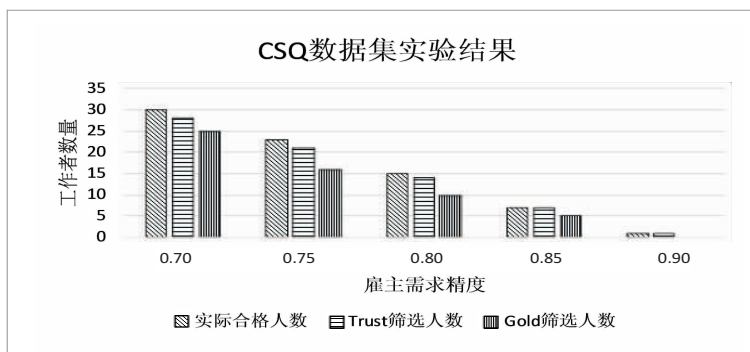


图 3 CSQ 数据集中两种方法筛选的工作者与实际合格工作者数量的比较

通过分析两种方法在不同数据集中的实际效果,可以发现,选用 Trust 方法筛选的工作者数量比使用 Gold 方法筛选的工作者数量更多,更接近于数据集中实际合格的人数,并且,雇主需求的精度越高,使用

Trust 方法筛选出的工作者数量越接近实际工作者数量。

图 4 和图 5 中将给出两种方法所选人数的准确度信息。

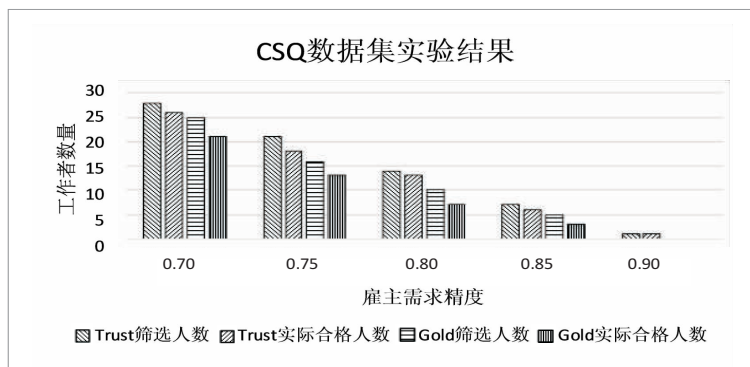


图 4 CSQ 数据集中两种方法选择的工作者数量和实际合格数量的比较

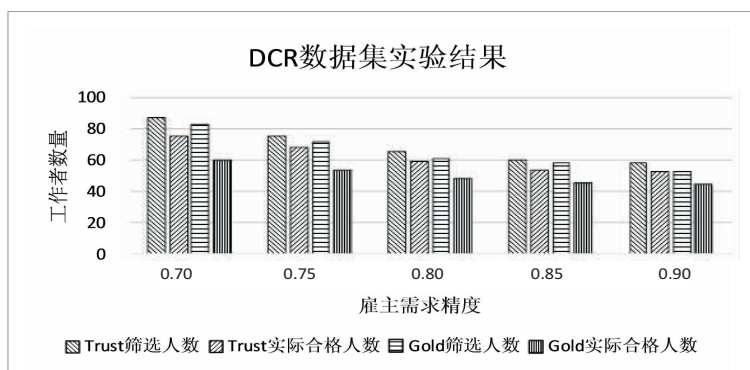


图 5 DCR 数据集中两种方法选择的工作者数量和实际合格数量的比较

通过图 4 和图 5 可以发现,使用 Trust 方法筛选出实际合格工作者数量高于使用 Gold 方法筛选出的实际合格工作者数量,且筛选结果准确度更高。为了更好地展示实验效果,表 1 和表 2 中详细列举了在不同数据集中使用两种方法筛选出的工作者准确度情况。

表 1 DCR 数据集中两种方法实验结果比较

雇主需求精度	0.70	0.75	0.80	0.85	0.90
Trust 筛选人数	88	76	66	60	58
Trust 实际合格人数	76	68	59	54	53
Trust 选人准确度	0.86	0.89	0.89	0.90	0.91
Gold 筛选人数	83	72	61	58	53
Gold 实际合格人数	60	54	48	46	45
Gold 选人准确度	0.72	0.75	0.79	0.79	0.85

表 2 CSQ 数据集中两种方法实验结果比较

雇主需求精度	0.70	0.75	0.80	0.85	0.90
Trust 筛选人数	28	21	14	7	1
Trust 实际合格人数	26	18	13	6	1
Trust 选人准确度	0.93	0.86	0.93	0.86	1.00
Gold 筛选人数	25	16	10	5	0
Gold 实际合格人数	21	13	7	3	0
Gold 选人准确度	0.84	0.81	0.70	0.60	0.00

如表 1 和表 2 所示,可以直观地发现,使用 Trust 方法对众包数据结果进行分析筛选出的结果要比使用 Gold 方法筛选出的结果更好,筛选的合格工作者数量更多,质量更高。使用 Trust 方法筛选出工作者提供的数据鲁棒性要高于使用 Gold 方法筛选出工作者提供的数据。

4 结束语

主要讨论了众包数据分析方法,针对现有数据分析方法存在的不足,提出了基于信任模型的鲁棒众包数据分析方法,将工作者历史信誉信息和此次任务结果数据结合于贝叶斯信任模型,提高了数据筛选的准确性。对于众包行业的稳定发展有十分长远的意义。

参考文献:

[1] 刘 锋. 威客(witkey)的商业模式分析[D]. 北京:中国科学院研究生院,2006.
 [2] 张志强,逢居升,谢晓芹,等. 众包质量控制策略及评估算法研究[J]. 计算机学报,2013,36(8):1636-1649.

[3] LE J, EDMONDS A, HESTER V, et al. Ensuring quality in crowdsourced search relevance [C]//Workshop on crowdsourcing for search evaluation at SIGIR'10. [s. l.]: [s. n.], 2010:21-26.
 [4] BERNSTEIN M, BRANDT J, MILLER R, et al. Crowds in two seconds: enabling realtime crowd-powered interfaces [C]//Proceedings of the 24th annual ACM symposium on user interface software and technology. Santa Barbara, California, USA: ACM, 2011:33-42.
 [5] MASON W, WATTS D. Financial incentives and the performance of crowds[J]. ACM SIGKDD Explorations Newsletter, 2010, 11(2):100-108.
 [6] 唐 思. 众包质量控制算法及评价框架研究[D]. 杭州:浙江大学, 2015.
 [7] 严 俊, 库少平, 喻 楚. 基于活跃度的众包工作者信誉模型[J]. 计算机应用, 2017, 37(7):2039-2043.
 [8] LIU X, LU M, OOI B, et al. Cdas: a crowdsourcing data analytics systems[J]. VLDB Endowment, 2012, 5(10):1040-1051.
 [9] TIAN Tian, JUN Zhu, YOU Qiaoben. Max-margin majority voting for learning from crowds[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019, 41(10):33-36.
 [10] LI Binbin, AU Siu-Kui. An expectation-maximization algorithm for Bayesian operational modal analysis with multiple (possibly close) modes[J]. Mechanical Systems and Signal Processing, 2019, 132:490-511.
 [11] TAGHIZADEH L, KARIMI A, PRESTERL E, et al. Bayesian inversion for a biofilm model including quorum sensing [J]. Computers in Biology and Medicine, 2020, 117:103582.
 [12] 李腾飞. 基于多项式朴素贝叶斯算法的垃圾邮件过滤器设计与实现[J]. 科技资讯, 2018, 16(33):1-2.
 [13] 阮闪闪, 王小平, 薛小平. 基于证据理论信任模型的众包质量监控[J]. 计算机应用, 2015, 35(8):2380-2385.
 [14] 赵少卡, 林晓明, 郭永宁, 等. 校企深度合作构建 IT 类创新创业教育平台的探索与实践[J]. 计算机教育, 2019(6):4-8.
 [15] 孙 彦, 丁学文, 雷雨婷, 等. 基于 SSD_MobileNet_v1 网络的猫狗图像识别[J]. 天津职业技术师范大学学报, 2020, 30(1):38-44.
 [16] GAO J, LIU X, OOI B C, et al. An online cost sensitive decision-making method in crowdsourcing systems [C]//Proceedings of the 2013 ACM SIGMOD international conference on management of data. New York, NY: ACM, 2013:217-228.