文章编号：2095-2163(2023)08-0001-10 中图分类号：TP391 文献标志码：A

一种基于确定度的交互式迭代数据清洗方法

孙辞海[1,2]，王洪亚[1]，郭开彦[1]，程炜东[1]
（1 东华大学 计算机科学与技术学院，上海 201620；2 上海对外经贸大学 统计与信息学院，上海 201620）

摘 要：自动化的数据清洗技术可以极大地提升数据清洗的效率，但会导致一定的错误率和不可靠的结果，通过引入人的参与，对建议修改值进行检查可避免错误的修改，同时对最终结果的可靠性有直观的评估。基于上述考虑，本文提出了一种基于确定度的交互式迭代清洗方法，该方法利用主动学习技术，将基于统计方法的数据清洗技术和人的参与相结合，在迭代过程中不断提升清洗模型的清洗能力和数据质量，同时最小化人的参与度。具体地，此方法包含一个基于确定度的自动清洗模型，对数据是否需要修改的必要性进行度量，可有效减少错误的修复；此外，本文还定义了确定度增益，表示数据是保留、还是修改的分歧程度，将分歧最大的建议修改值交与人查看，以减小人的参与度。最终，本文在多个实验数据上验证了方法的有效性。

关键词：数据清洗；主动学习；确定度；交互式迭代

An interactive iterative data cleaning method based on certainty

SUN Cihai[1,2], WANG Hongya[1], GUO Kaiyan[1], CHENG Weidong[1]
(1 College of Computer Science and Technology, Donghua University, Shanghai 201620, China;
2 School of Statistics and Information, Shanghai University of International Business and Economics, Shanghai 201620, China)

【Abstract】 Automated data cleaning technology can greatly improve the efficiency of data cleaning, but it will lead to a certain error rate and unreliable results. By introducing people's participation, it can avoid the wrong modification by checking the recommended modification value, and the reliability of the final result can be evaluated intuitively. Based on the above consideration, this paper proposes a cleaning interactive iteration method based on certainty, using active learning techniques, this method will apply data cleaning technology based on the statistical methods in combination with the participation of people, and in the process of iteration enhance cleaning ability of the cleaning model and data quality, thereafter minimize the engagement of the people at the same time. Specifically, this method includes an automatic cleaning model based on the certainty, and measures the necessity of whether the data needs to be modified, which can effectively reduce the error repair. In addition, this paper also defines the certainty gain, indicating the degree of divergence between data retention and data modification, and submits the suggested modified values with the largest divergence to people for review, so as to reduce engagement. Finally, the validity of the method is verified by several experimental data.

【Key words】 data cleaning; active learning; certainty; interactive iterative

0 引 言

检测和修复脏数据是数据分析中的挑战之一，低质量的数据将导致分析不准确和决策不可靠。更多的数据来源和更多的数据量意味着数据质量问题的多样性和复杂性更大，以及以成本效益的方式来保持数据质量的复杂性更高。因此，各种数据清洗方法相继被提出，以便自动地或半自动地识别错误，并在可能的情况下对其加以纠正。

在过去几年里，出现了大量基于完整性约束[1-4]、统计[5]或机器学习[6]的数据清理方法。尽管这些方法具有适用性和通用性，但却无法确保修改数据的正确性。为了提高这些方法的准确性，常用的方法有引入表格主数据[7]和领域专家[8-10]等。然而这些方法需要的资源是稀缺的，通常也很昂贵。为了解决这些问题，结合知识库[11]和众包的方法被

基金项目：国家自然科学基金(61370205)；上海市自然科学基金项目(13ZR1400800)。

作者简介：孙辞海(1985-)，男，博士研究生，主要研究方向：数据库、集合相似连接；王洪亚(1976-)，男，博士，教授，主要研究方向：数据库系统内核分析与优化。

通讯作者：王洪亚 Email：hywang@dhu.edu.cn

收稿日期：2022-09-22

提出,而知识库的构建、存储、维护以及众包的使用仍需要一定的成本。为了结合以上方法的优点,规避其缺点,实现高效率、低成本、结果有一定保证的清洗方法,本文应用了主动学习技术,在使用机器学习的数据清洗方法基础上,部分利用用户交互,仅将最不确定的预测值交予用户检查清洗。与其他修复方法类似,本文在清洗时遵循了保守修复原则,通过引入确定度指标,将建议修改数据和原始数据在确定程度上进行比较来决定是否修复,此方法可以避免错误修改对数据的破坏。此外,为了提升方法的通用性,本文还在多个属性上建模,适用于多个属性上存在错误的情况。

本文主要贡献如下:

(1)运用主动学习的方法,在使用机器学习的数据清洗方法基础上,部分利用用户交互,在迭代的清洗过程中不断提升数据质量。首先本文构建一个预测模型用于提供建议修改数据,经过筛选后将一部分交付人工检查清洗,然后将人工清洗干净的数据反馈给预测模型重新建模,提升预测能力,在不断迭代中,提升模型清洗能力,实现高效率、低成本、且清洗结果有一定保证的清洗方法。

(2)提出了基于确定度的预测模型。此模型以概率分类器为基础(本文使用朴素贝叶斯分类器),然后在分类器上应用 BvSB 准则[12],计算预测结果确定度,用以表示分类器对其预测结果的正确性的确定程度;同时模型还对原始数据应用 BvSB 准则,计算原始数据确定度,用以表示原始数据对其自身的正确性的确定程度。当预测值与原数据不同时,从两者中选择确定度大的作为模型输出。这样,在确定度的指导下,分类器的预测结果有一定的概率保证,且对分类器建议修改加上一个限制条件,需要修改数据的确定度高于保留原始数据的确定度时才能做修改,从而减少错误的修改事件。

(3)提出了基于确定度增益的筛选规则。上述 2 种确定度越接近,是否修改的分歧也就越大,出错的概率也就越高,把这部分数据交付人工查看可避免错误的修改。因此对 2 种确定度做差,求得确定度增益,有效地反映了这种分歧。

(4)在多属性错误下构建了基于确定度的清洗模型,保证了清洗的通用性。在各数据集上进行了大量的实验,并与相关的清洗技术进行了对比,表明了本方法的有效性。

1 研究目的和相关工作

纠正错误数据是一个耗时、耗力且十分乏味的过程。为了提升清洗的效率,很多修复脏数据的技术采用基于约束的修复方法[13-14],通过检测数据是否满足一系列的约束(完整性约束、条件函数依赖等),以此有效地识别脏数据,然而这些方法在纠正脏数据上却有所欠缺,甚至在纠正过程中引入新的错误数据[7]。众包和知识库的应用可以弥补修复的不确定性,提升修复的质量。

1.1 使用统计方法进行数据清洗

为了提升数据清洗的效率和正确性,出现了多种使用统计方法的技术,机器学习即是其中之一。

文献 ERACER[5] 提出了 2 种统计的、数据驱动的方法来推断关系数据库中缺失的数据值。一种是使用卷积或回归的新颖的近似框架;另一种是使用贝叶斯网络的基线精确方法。在传感器网络文献[15-16]中也有一些统计的异常值检测和修复技术的例子。文献 SCARE[6] 使用机器学习方法在高效清洗的同时,提高了清洗的可靠性。

SCARE 方法的创新点有:

(1)使用概率分类器技术对干净数据集的概率分布进行建模,模型被用于计算脏数据集的似然。其核心思想是要在一定的修改(修改可能错误的值)次数内最大化似然,从而达到准确修复脏数据的目的。

(2)考虑到脏数据的多个属性上可能被认为存在脏值,所以构建了多属性概率分类器解决了此问题。

然而,SCARE 在实际应用中还存在一些不足:首先,在构建概率分类器时,干净数据集要么是与脏数据集同源(例如相同表的 2 个划分),要么使用现存的统计方法[17-18]得到,然而脏数据集很多时候没有同源的干净数据集,且使用统计方法得到的干净数据集不一定可靠。其次,在定义修改次数时,SCARE 假设已知脏数据集中每一个元组出错的概率,这是不容易实现的。缺乏人的参与也使得数据清理效果不够理想。

此外,为了进一步提升数据的可靠性,一些方法就使用主动学习技术[19-21],将需要处理的资源选择性地交付人工处理,在保证数据质量和清洗效率的同时,减少了人力资源的消耗。主动学习的基本思想就是将部分人处理过的数据作为基础数据,训练监督模型(如 SVM 和随机森林等),从众多待处理数据中选出更有价值的资源交付人工处理。如何从众多数据中筛选出最有价值的数据是使用主动学习进行数据清洗的重点。

综合前文分析可知，本文假设初始只有一个数据质量未知的脏数据集，应用主动学习技术，加入人工参与，由人工清洗得到干净数据集，在迭代清洗中使用尽可能少的干净数据集构建高质量的清洗模型，在减少人的参与度的同时可提升数据的可靠性。

1.2　主动学习算法

主动学习[22]是机器学习的子领域，其关键思想是：若机器学习算法可以选择用于训练的数据，那么就可以在更少的训练集上实现更高的准确率。因此，主动学习经常被用于标注问题。主动学习通过在庞大的未被标注的数据中，挑选合适的数据交予人工标注，使用尽可能少的标注数据，训练出高准确度的模型。

一般地，主动学习的框架如图 1 所示。首先，使用少量的已标注数据集构建一个分类器。接着，分类器从未标注数据中选取样例进行识别，若识别错误则筛选后进行人工标注。然后，将标注的数据补充到已标注数据集中，使用此新的数据集构建新的分类器。最后，迭代标注，直到达到设定的某个终止条件。

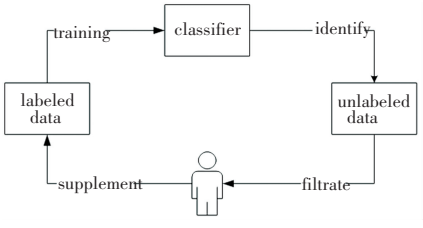

图 1　主动学习框架

Fig. 1　Active learning framework

与数据标注直接插入正确标记类似，数据清洗是先发现脏数据再替换为正确数据，两者都是为了得到正确值，且人工标注和人工清洗都是耗时耗力的，因此自然地想到将主动学习应用到数据清洗中，实现高效、低消耗的清洗。在应用机器学习方法进行数据清洗时，有 3 个问题：

（1）在初始仅有一个数据质量未知的脏数据时，构建初始分类器所需的训练集难以得到。

（2）机器学习具有一定的错误率，正确数据可能会被修改错误。

（3）机器学习清洗的结果有时是不可解释的。主动学习可以很好地与数据清洗结合并有效地解决以上问题。首先，主动学习有人参与，可人工清洗小部分数据用于构建初始模型。然后，主动学习应用高价值筛选方法，可以把模型容易出错的数据筛选出来交予人工检查清洗。最后，由人工对模型的清洗能力进行评估，可以验证模型的可靠性，对清洗质量把关。

2　问题定义

对数据集 D，具有关系型模式 G（表格、csv 文件等具有的模式），A 表示 G 的属性集合。在 A 中集合 $F=\{E_1,\cdots,E_K\}\in A$，表示脏属性集，$F$ 对应的值很可能由离散型的脏数据构成，因此可能被修改。A 的补集 $R=A-F=\{W_1,\cdots,W_L\}$ 为干净属性集，R 对应的值由干净数据构成。对于 A 中某个属性 E_i，其值域可表示为 $dom(E_i)$。这样，D 中某个元组 t 可分为 2 部分：干净部分 $t[R]=t[W_1,\cdots,W_L]$ 和可被修改部分 $t[F]=t[E_1,\cdots,E_K]$。$t[R]$ 和 $t[F]$ 将简写为 r 和 f，即 $t=<r,f>$。另外，D 在主动学习应用下，按照数据是否被人检查并清洗干净可被划分为 2 部分：检查为正确的或已经纠正的干净数据 $D_c\subset D$；未检查的或不可确认的数据 $D_e=D-D_c$。与其它清洗方法往往需要一个完整的、大量的干净数据用于寻找统计规律不同，本文的任务是在没有任何干净数据前提下，对一个干净程度未知的数据集进行清洗，即 $D=D_e$。具体地，首先从 D_e 中人工清洗少部分数据得到 D_c，然后对 D_c 构建清洗模型，最后针对可能存在错误的元组 $<r,f>\in D_e$，把 r 带入到模型中，尽可能预测出 f 可能正确的值 f'。

3　基于确定度的交互式迭代清洗方法

3.1　基于主动学习的交互式迭代清洗结构

为提高清洗效率和可靠性，减少资源消耗，本文结合了机器学习和人工参与，使用主动学习方法，在迭代过程中完成数据清洗。

基于主动学习的交互式迭代清洗结构如图 2 所示。其清洗过程可描述如下：

（1）划分：因为使用主动学习技术的数据清洗方法是迭代清洗的，所以将初始数据集 D 划分成多个小块是必要的。本文假设初始数据集 D 干净与否未知，也就是均为脏数据（$D=D_e$），整个数据都需要被清洗，且没有其它干净数据集作为参考，所以 D 会被划分为 2 部分。一部分为初始脏训练集 D_e^t，由人工清洗后用于构建预测模型；另一部分为若干脏数据块 $\{D_e^1,\cdots,D_e^N\}$，都是待清洗数据，将在迭代中被清洗。

（2）预清洗：模型的构建需要干净数据集支撑，预清洗将从初始脏数据集 D_e 中得到干净训练集 D_c^t。因为有人工参与，将划分出的初始脏训练集 D_e^t 交予人工进行清洗可得到一个干净训练集 D_c^t。得

益于主动学习技术,初始的 D_e^t 很小,所以人工清洗工作量也很小,但同时也导致初期的模型清洗能力不强,在基于主动学习的迭代清洗中,会不断将人工清洗的数据补充到 D_c^t。另外,在人工检查清洗 D_e^t 过程中,将统计样本中每个脏属性的正确值所占比例 P_{right}^i。P_{right}^i 作为数据集 D 的固有特征,近似地表示某个脏属性下值正确的概率,P_{right}^i 将被用于计算原始数据的确定度(在 3.2 节详细介绍)。

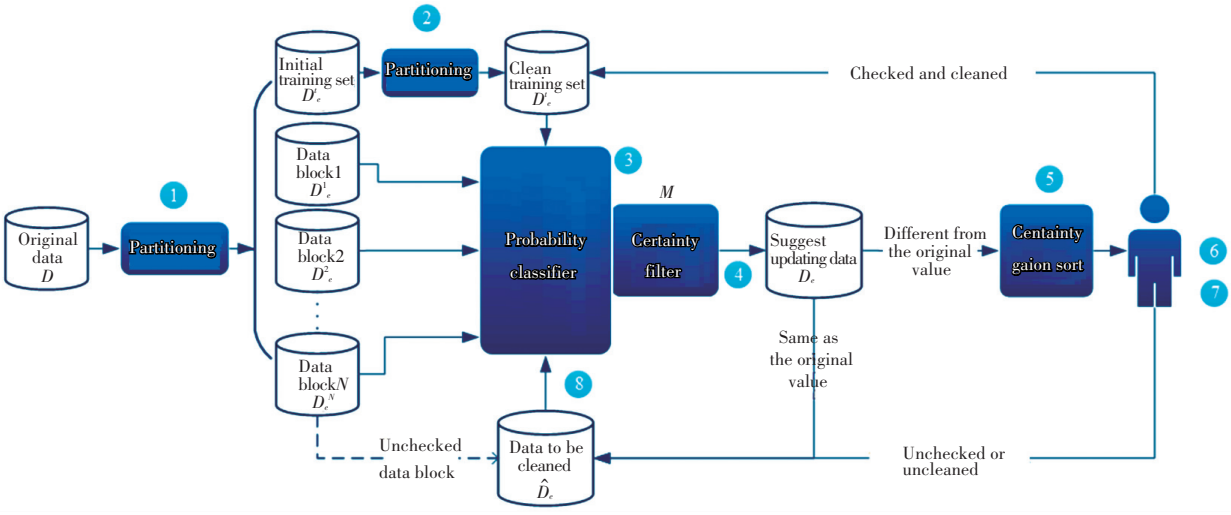

图 2 基于主动学习的交互式迭代清洗结构

Fig. 2 Interactive iterative cleaning structure based on active learning

(3)构建确定度预测模型:确定度预测模型 M 由 2 部分组成。一部分为一个概率分类器,从 D_c^t 中学习概率分布,用于计算每种结果被预测出的概率;另一部分为一个确定度筛选规则,基于 BvSB 准则,计算分类器预测结果的确定度,同时使用 P_{right}^i 计算原始数据的确定度,通过比较预测结果和原始数据的确定度,选出最确定的数据作为 M 的最终输出。由于采用了基于确定度的对比修复方法,使得数据修改的条件更严格,以此减少错误的修改操作。

(4)修改推荐:研究选择某一个数据块 $D_e^i(i=1,2,\cdots,N)$ 作为 M 的输入,对 D_e^i 中的每一个元组 $t=<r,f>$,在概率分类器中可得到一个建议修改元组 $t'=<r,f^i>$。根据确定度公式(在 3.2 节详细介绍)计算 t' 的确定度 C,同时原始元组 t 计算原始确定度 C_{init}。当 $t\neq t'$ 时,选择确定度最大的值作为 M 的输出。并将其插入到建议更新数据集 D_e^i 中。当 $t=t'$ 时,会将 t 直接插入待清洗的数据集 $\hat{D}_e$ 中,考虑到前期迭代时模型的推荐能力不强,即使预测值和原始值一致,也可能是错误的,因此将这部分数据汇总,在最后一轮清洗中,使用推荐能力更好的模型进行二次清洗。

(5)增益排序:对 D_e^i 的每一条数据 t',根据确定度增益公式(在 3.2 节详细讨论)计算其确定度增益 C_{gain}。然后,使用确定度增益将 D_e^i 进行升序排序,并依次交予人工查看。当人工认为 D_e^i 的干净程度已达到预期,会停止检查剩余数据。确定度增益是预测值确定度和原始值确定度的差值,代表了选择 2 个预选结果的分歧程度。因此,优先推荐增益小的数据交予人工查看,可辅助调解修改分歧,并对模型清洗能力做评估,验证清洗的有效性,保证修复数据的可靠性。

(6)人工检查:人工检查并清洗过的数据会反馈给干净训练集 D_c^t,这部分数据是分歧大的数据,加入到 D_c^t 可完善数据的概率分布,因此与单纯的扩充训练集相比,基于确定度增益的数据反馈方法可以更快地提升 M 的清洗能力;人工未检查或检查后无法给出正确值的数据会输出到 $\hat{D}_e$ 中,等待最后轮的二次清洗。到此,一轮清洗结束,当人工对 M 清洗能力不满意时,将继续清洗下一个脏数据块。

(7)迭代终止:在依次检查数据块的过程中,当人工认为 M 建议修改的多个数据块的干净程度已达到预期,即 M 已经符合清洗要求,则停止迭代,并把剩余未检查的数据块汇总于 $\hat{D}_e$ 中。此时,最终轮的准备完毕,包括一个已经清洗干净的 D_c^t,一个已经建模为人工满意的清洗模型 M,以及一个是人工

未检查或人工检查后不可修正的脏数据 $\hat{D}_e$。

(8)最终轮清洗：把 $\hat{D}_e$ 输入到高质量的模型 M 中，再次执行一遍整个清洗流程，直到人工对最后一批建议修改数据的干净程度表示满意，然后将人工检查清洗过的数据、人工未检查和不确定的数据、D_c^t 三者汇总，得到最终的清洗结果。

3.2　确定度和确定度增益

本文基于主动学习技术进行数据清洗，引入了确定度指标，用于构建预测模型。分别在分类器预测值和原始数据上都应用了确定度，这样数据修改更加谨慎，可减少错误的修改操作。

3.2.1　2 类确定度

概率分类器可以对每种结果给出一个分类概率，以表示结果出现的可能性，然而可能性最高的 2 个结果概率相近时，分类器对结果的判断是不确定的，最终结果将很难保证其正确性。在概率的基础上应用确定度，不仅对结果有一定的概率保证，还提供了评估结果正确性的依据。

分类器一般运用于二分类问题，而脏数据值域一般多于 2 类，且基于熵的确定度计算方法并不适用于多类分类问题[12]，因此本文选用 BvSB 准则计算预测值的确定度 C。设元组 $t=<r,e>$ 属于最优修改值和次优修改值的后验概率分别为 $P(e_{best}\mid r)$ 和 $P(e_{second-best}\mid r)$，其中 $e\in dom(E_i)$，选择 e_{best} 的确定度计算公式如下

$$C(e_{best}\mid r)=P(e_{best}\mid r)-P(e_{second-best}\mid r) \quad (1)$$

在数据清洗中，因为数据的修改是有风险的，所以人们希望尽可能少地修改数据。本文在分类器提供建议修改值时，考虑到 D 的错误率，给出原始数据的确定度 C_{init}。在人工清洗得到初始干净训练集 D_c^t 时，统计得到样本的脏属性对应值的干净程度 P_{right}^i。P_{right}^i 反映了 $t=<r,e>$ 中的 $e\in dom(E_i)$ 正确的概率，而 $dom(E_i)$ 中其它的某个值正确的概率为 $\dfrac{1-P_{right}^i}{count(dom(E_i))-1}$，根据 BvSB 准则可计算原始值 e 的确定度 C_{init}。原始数据的确定度仅与 E_i 的类别个数和样本的干净程度有关，所以 E_i 属性上所有值的原始确定度可表示为：

$$C_{init}(e\mid r)=C_{init}(E_i)=P_{right}-\frac{1-P_{right}}{count(dom(E_i))-1}$$
$$i=1,2,\cdots,K,\quad e\in dom(E_i) \quad (2)$$

3.2.2　数据元组最终确定度计算

本文不仅要采纳分类器提供的建议，还要考虑是否保留原始值。综合数据集自身的正确率、尽量少修改的原则以及分类器的预测，最终的建议修改值 e' 会从建议修改值 e_{best} 和原始值 e 中选择确定度最大的值。e' 筛选公式如下：

$$<r,e'>=argmax(C(e_{best}\mid r),C_{init}(e\mid r)) \quad (3)$$

3.2.3　确定度增益

本文还引入确定度增益指标，用于从模型建议修改数据中筛选出容易错误的数据，并将这部分数据交予人工检查。确定度可以作为筛选数据的依据，但本文除了使用建议修改的确定度，还使用了保持原始数据不变的确定度。因此，本文将 2 种确定度作差表示确定度增益 C_{gain}，那么更新元组为 $<r,e'>$ 的确定度增益为：

$$C_{gain}(<r,e'>)=C(e'\mid r)-C_{init}(e\mid r) \quad (4)$$

由式(4)可知，确定度增益表示了数据是否修改的分歧程度，建议修改值的确定度和原始数据的确定度越相近，确定度增益越小，对数据是否修改的分歧越大。本文通过确定度增益对建议修改数据进行排序，可以有效筛选分歧大的数据，在尽量少的人力资源消耗下，更快地提升模型清洗能力。

3.3　多属性建模和修改值预测

分类器一般是预测单属性的结果，但脏数据不会仅在一个属性上出现，因此需要在多属性上建模来清洗多属性上的错误。本文构建了基于确定度的多属性清洗模型。为了构建多属性模型，需要在每个属性上构建一个单属性模型，有 2 种选择。一种是直接构建干净属性集到每个脏属性集上的映射模型，可表示为 $M_i:R\rightarrow E_i$，$i=1,\cdots,K$；另一种是预测 E_i 时，考虑到 E_i 可能与 $\{E_1,\cdots,E_{i-1}\}$ 存在依赖关系，输入除了干净属性集还包括已预测的脏属性集 $\{E_1,\cdots,E_{i-1}\}$，此时 K 个模型就可表示为 $M_i:\{R,E_1,\cdots,E_{i-1}\}\rightarrow E_i$，$i=1,\cdots,K$。本文将使用第二种建模方法，充分利用属性间依赖关系，提升清洗能力。

对于元组 $t=<r,f>$，为了预测修改元组 $t'=<r,f'>$，需要构建 K 个模型 $\{M_1,\cdots,M_K\}$ 分别预测 K 个属性上的值。预测 $f'[E_1]$ 时，直接将 r 带入 M_1 即可。预测 $f'[E_2]$ 时，因为本文将原始值考虑在内，并非简单地将 $f'[E_1]$ 和 r 作为输入。当 $f'[E_1]=t[E_1]$ 时，因为预测结果和原始值相同，可直接将 $<r,f'[E_1]>$ 作为输入，而当 $f'[E_1]\neq t[E_1]$ 时，除了将 $<r,f'[E_1]>$ 作为输入，把 $f'[E_1]$

看作正确值,还会考虑原始值 $t[E_1]$ 为正确值,将 $<r, t[E_1]>$ 作为输入。对之后的属性进行预测也会考虑预测值与原始值是否相同而进行相应的处理。因此,多属性预测的过程可描述为一个 $K+1$ 层二叉树,如图 3 所示。从根节点出发,预测每一层的结果。对于第 $k-1$ 层的某节点,其左儿子节点为建议修改值 $f'[E_k]$,右儿子节点为原始值 $t[E_k]$,若 $f'[E_k]=t[E_k]$,则仅有一个左儿子节点。

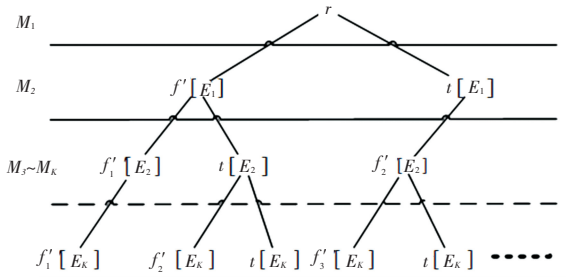

图 3 多属性预测过程

Fig. 3 Multi-attribute prediction process

在多属性上,每个属性上预测值的确定度将整合在一起,作为一条建议修改数据的确定度。本文定义建议修改值 f' 的确定度为所有 $f'[E_i]$ 的确定度的乘积,公式如下:

$$C(f' \mid r) = C(f'[E_1] \mid r) \prod_{i=2}^{K} C(f'[E_i] \mid r, f'[E_1 \cdots E_{i-1}]) \quad (5)$$

原始确定度同理可得:

$$C_{init}(f \mid r) = C_{init}(t[E_1] \mid r) \prod_{i=2}^{K} C_{init}(t[E_i] \mid r, t[E_1 \cdots E_{i-1}]) \quad (6)$$

由图 3 可知,r 作为确定度模型 M 的输入,最多有 2^K 种结果,然而由于很多时候建议修改值与原始值相同,实际的最终结果数会小于 2^K。假设对元组 $t=<r, f>$,有 N 个建议更新值($t=<r, f>, t_1'=<r, f_1'>, \cdots, t_{N-1}'=<r, f_{N-1}'>$)。在最确定优先原则下,最终预测结果为:

$$<r, f'> = \text{argmax}(C(f_1' \mid r), \cdots, C(f_{N-1}' \mid r), C_{init}(f \mid r)) \quad (7)$$

确定度增益计算公式为:

$$C_{gain}(<r, f'>) = C(f' \mid r) - C_{init}(f \mid r) \quad (8)$$

在多属性预测过程中,分类器预测值的确定度在 2 种特殊情况下会做特殊处理。

(1) $f'[E_i]=t[E_i]$ 时,即预测值和原始值相同时,本文遵循最确定优先的原则,选择确定度最大的作为 $f'[E_i]$ 的确定度。

$$C(f'[E_i] \mid r) = \max(C(f'[E_i] \mid r, f[E_1 \cdots E_{i-1}]), C_{init}(E_i)) \quad (9)$$

(2) $t[E_i] \notin D_c^t$ 时,即原始值不在干净训练集中时,分类器的结果将永远不会预测为 $t[E_i]$,因此 $t[E_i]$ 是尽可能要交予人工检查清洗的。此时设置 $f'[E_i]$ 确定度为一个极小值,表示预测结果是极不确定的,在分类器给出的建议中,包含 $t[E_i]$ 的结果将不会被优先选择。同时,减少原始值确定度,避免其值过高而被直接保留。

对一个元组 $t=<r, f>$,r 作为输入,多属性模型预测和确定度计算过程见算法 1,初始为 $getPredictions(M_1, r, 1.0, CI_1, t=rf)$。

算法 1 get Predictions (Classification Model M_i, input $r_i=<r, f[E_1], \cdots, f[E_{i-1}]>$, Certainty C, Initial Data Certainty CI_i, Database Tuple $t=rf$

```
if i > K then
    f' = r_i - r
    AllPredictions = AllPredictions ∪ (f', C)
    Return
end
f(E_i) = M_i(r_i)
r_s = < r_i, f[E_i] > {Adding the adviced E_i' s value to the next input}
if f[E_i] ≠ t[E_i] then
    C_s = C × C(f[E_i] | r_i)
end
else
    C_s = C × max(C(f[E_i] | r_i), CI_i) {Choosing the max certainty of E_i' s value
    When adviced E_i' s value is same as original E_i' s value}
end
get Predictions (M_{i+1}, r_s, C_s, CI_{i+1}, t)
if f[E_i] ≠ t[E_i] then
    r_s = < r_i, t[E_i] > {Adding the original E_i' s value to the next input}
if t[E_i] ∉ D_c^t then
    C_s' << CI_i
    CI_i = CI_s/3
end
else
    C_s' = C × CI_i
end
getPredictions (M_{i+1}, r_s', C_s', CI_{i+1}, t)
```

end

4　实验结果与分析

本文在 3 个数据集上进行了实验，主要与其他相关工作进行比较，在多个指标上多角度地验证了方法的有效性。

(1)数据集：本文使用 UCI 机器学习资源库上(http://archive.ics.uci.edu/ml/)的 3 个数据集。

① USCensus1990 数据集：美国 1990 年人口普查的部分数据，被用于评估确定度的有效性。

② Bank 数据集：葡萄牙一家银行机构的营销数据，被用于评估确定度增益的有效性。

③ Adult 数据集：美国成人收入普查数据，被用于评估不同人参与度对清洗效果的影响。实验所用的数据信息见表 1。

表 1　实验数据集信息

Tab. 1　The information of experimental data set

Data	#Rows	#Attributes	#Dirty attributes
USCensus1990	2 358 285	30	3
Bank	45 211	21	4
Adult	48 842	15	3

在实验中，数据会被划分为多块，其中一块(300 条数据)被清洗干净用于训练初始模型，其他数据块(100 条数据/块)在迭代中被清洗。

(2)评价指标：本文使用了以下 4 个评价指标用于评估本文清洗方法。

① 数据质量(Quality，简记为 Q)：正确记录占所有需要清洗的记录的比例，可表示为：

$$Q = \frac{\#recodes_{right}}{\#rows \times \#columns_{error}} \tag{10}$$

其中，$\#recodes_{right}$ 是清洗后正确的记录数；$\#rows$ 为数据行数；$\#columns_{error}$ 是错误数据所在列的个数。

② 参与度(Engagement，简记为 E)：人工检查的记录占需要清洗的记录的比例，可表示为：

$$E = \frac{\#recodes_{checked}}{\#rows \times \#columns_{error}} \tag{11}$$

其中，$\#recodes_{checked}$ 是被人工检查过的记录数。

③ 精度(Precision，简记为 P)：在所有修改过的记录中，修改正确的记录占有的比重，可表示为：

$$P = \frac{\#recodes_{update-right}}{\#recodes_{update}} \tag{12}$$

其中，$\#recodes_{update}$ 是被修改的记录数，$\#recodes_{update-right}$ 是被修改的记录中正确的记录数。

④ 召回率(Recall，简记为 R)：所有错误的记录中，被修改正确的记录占的比重，可表示为：

$$R = \frac{\#recodes_{update-right}}{\#recodes_{error}} \tag{13}$$

其中，$\#recodes_{error}$ 是数据中所有错误的记录数。

(3)默认参数：以下是本实验会使用的参数和说明。

① 初始正确率 $P_{right} \approx 70\%$：默认保留数据质量在 70%左右，即向数据插入 30%左右的脏数据。其中，15%通过在干净数据加上额外后缀(如 BvSB)生成脏数据，另一部分 15%则用域空间的其它值替换。

② 迭代终止的条件默认为：固定迭代 40 次。

③ 人满意度

(a)对每块数据清洗的满意度：对每一块数据，每检查 $\#columns_{error} \times 2$ 个记录，若脏记录所占比例小于 0.2，即表示用户对当前数据块清洗程度表示满意，则不再检查剩余数据。

(b)对模型清洗能力的满意度：对每一块数据，用户检查记录数小于 $\#columns_{error} \times 6$，表示为用户对每次迭代的模型满意；当连续 3 次迭代均满足上述条件，即连续 3 次迭代模型效果都令用户满意，则表示用户对模型的清洗能力满意，此时则终止迭代并准备最后一轮清洗。

(4)相关技术命名：以下是本实验将要进行比较的相关技术的命名和说明。

① ADC：本文方法，使用基于确定度的分类器预测正确数据，加上使用确定度增益的数据筛选方法。

② ADC_C：ADC 仅使用基于确定度的概率分类器进行预测。

③ SCARE_P：SCARE 使用概率分类器进行预测且未使用水平划分的方法，用于与 ADC_C 比较以评估基于确定度的分类器进行预测的效果。

④ ADC_R：ADC 未使用确定度增益排序，对分类器预测值采用随机推荐方式交予人工检查，用于与 ADC 比较以评估确定度增益对清洗效果的影响。

4.1　评估确定度

为了评估基于确定度的分类器预测方法的有效性，本文使用 USCensus1990 数据集进行实验，在每一次迭代中，观察分类器预测值的效果，通过比较 ADC_C 和 SCARE_P 在数据质量、精度、召回率上的效果，以评估应用确定度方法的有效性。使用确定度方法和使用似然方法的效果对比如图 4 所示。由

图 4 可看到,ADC_C 和 SCARE_ P 都能有效地提升数据质量。在迭代清洗中,对于每一块数据的召回率,ADC_C 均远高于 SCARE_ P,即 ADC_C 能修复更多的脏数据。在精度上,ADC_C 与 SCARE_ P 交错分布,总体上 ADC_C 是低于 SCARE_ P 的,所以 ADC_C 修改的错误率更高。2 种方法在召回率和精度上各有优劣,而对于数据质量指标,ADC_C 普遍高于 SCARE_ P,因为 30%的错误率使得 ADC_C 中原数据的确定度不高,建议修改值很多,所以修改覆盖的错误数据更多,召回率更高,同时更新错误的数据也更多,精度变低。ADC_C 与 SCARE_ P 相比,虽然精度总体偏低,却相差不大,而召回率是远高于 SCARE_ P 的,即 ADC_C 能找到并修改正确更多的脏数据,因此 ADC_C 能获得更高的数据质量。

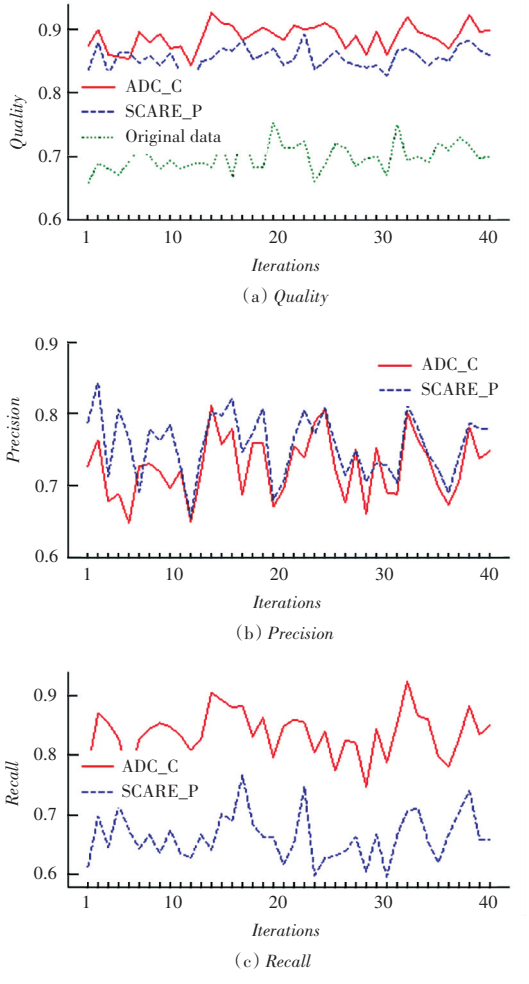

图 4 使用确定度方法和使用似然方法的效果比较

Fig. 4 Comparison of the effects between using the certainty method and using the likelihood method

4.2 评估确定度增益

本节将使用 Bank 数据集进行实验,从 2 方面评价确定度增益筛选方法。首先,评价确定度增益对清洗质量的影响,说明确定度增益筛选法可以有效提升数据质量。然后,评价确定度增益对模型提升的影响,说明确定度增益筛选法可以更有效地提升模型清洗能力。

4.2.1 评估对清洗质量的影响

为了评估确定度增益对清洗质量的影响,本文模拟真实的清洗过程,使用人工满意度替换固定迭代次数作为迭代终止条件,对比 ADC 与 ADC_R 在相同的人满意度下清洗的质量。表 2 展示了清洗结果。

由表 2 可知,ADC 的数据质量高于 ADC_R 的数据质量,说明了在以人工为主的清洗过程中确定度增益对提升数据质量很有效。然而,ADC 的人工参与度明显高于 ADC_R,即 ADC 需要更多的人力资源,这是因为确定度增益筛选法将分类器更容易出错的数据筛选出来优先交予人工检查,所以人工需要检查更多的建议修改数据才能会对当前这一轮清洗感到满意,这样清洗迭代次数也就变多。在精度和召回率方面,召回率两种方法相近,即找出的脏数据数量都差不多,而精度上 ADC 远高于 ADC_R,这是因为 ADC 使用的更多的人力资源把分类器容易预测错误的数据清洗了,减少了错误的修改,以此就避免了数据质量的下降。如果比较参与度差值(0.011 7)和数据质量差值(0.064 4),可以发现人工的参与能得到更多的数据质量提升,说明确定度增益筛选法可以很好地利用人力资源,从而达到事半功倍的效果。

表 2 基于确定度增益筛选法和随机筛选法的清洗效果对比

Tab. 2 Comparison of cleaning effect between certainty gain filter method and random filter method

Data	*Q*	*E*	*P*	*R*	*Iterations*
ADC	0.871 7	0.019 4	0.700 4	0.636 3	41
ADC_R	0.807 3	0.007 7	0.555 1	0.629 1	4

4.2.2 评估确定度增益对模型提升的影响

本文将一个固定 1 000 行的脏数据带入每一轮清洗中,对比观察 ADC 和 ADC_R 在每一次迭代后分类器的清洗能力。基于确定度增益筛选和随机筛选的模型提升对比如图 5 所示。由图 5 可知,ADC 和 ADC_R 在数据质量、精度和召回率上均处于上升趋势,说明增加训练集的数据量可提升分类器的预测能力。但是,ADC 的上升趋势明显高于 ADC_R,这说明了确定度增益筛选法可以更快地提升分类器的预测能力,因为筛选出的数据更有益于弥补

缺失的概率分布，使得分类器预测能力能得到更快提升。

4.3　评估不同的参与度对清洗效果的影响

为了评估不同的参与度对清洗效果的影响，本文中使用 adult 数据进行实验，并以多个迭代次数（10~90 次迭代）近似表示不同的参与度。不同参与度对清洗效果的影响如图 6 所示。由图 6 可知，随着人工参与度稳步上升，数据质量也在不断提升。但是，数据质量在快速提升后逐步放缓，这是因为模型的提升达到了瓶颈。通过计算相邻 2 次迭代数据质量的差值，得到数据质量增量，参与度增量同理可得。将 2 个增量进行比较可以发现，随着人工参与度不断提升，数据质量提升收益不断降低，在接近 70 次迭代时，人力资源消耗不能换回等价的数据质量提升。实验结果表明，投入过多的人力资源并不能线性提高数据品质，这也是所有自动清洗技术的瓶颈和面临的挑战。

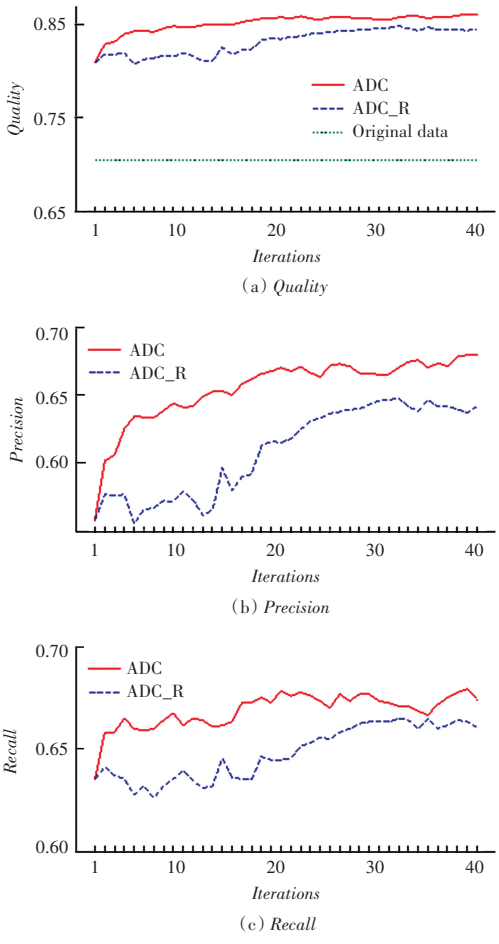

图 5　基于确定度增益筛选和随机筛选的模型提升对比

Fig. 5　Comparison of model enhancement between certainty gain filter method and random filter method

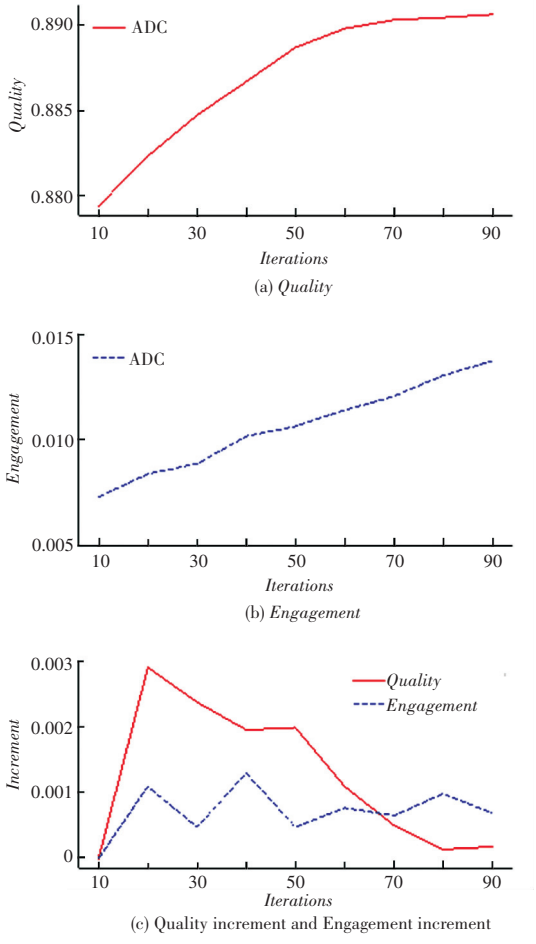

图 6　不同参与度对清洗效果的影响

Fig. 6　Influence of different engagement on cleaning effect

4.4　评估不同干净程度的清洗效果

本文在 3 个数据集上进行实验，用以评估本文方法在数据的不同干净程度下的清洗效果。不同干净程度的脏数据对清洗效果的影响见表 3。3 个数据集在清洗后数据质量的提升相当。通过统计方法进行清洗具有一定的错误率，受此影响，随着脏数据干净程度的提升，清洗难度将加大，除了要清洗固有错误数据，还要避免将干净数据修改错误。由表 3 可知，随着干净程度的增加，数据质量仍能提升，只是限于模型清洗能力的限制，提升的程度在不断减少。在参与度方面，USCensus1990 明显小于其他 2 个数据集，这是因为 USCensus1990 数据量更大，在检查相当的数据量时，USCensus1990 参与度自然变少，而数据质量提升却与其它 2 个数据集相当，这说明本文对更大数据量的数据能节省更多的人力资源，并收获很好的质量提升。

表 3 不同干净程度的脏数据对清洗效果的影响

Tab. 3 The influence of dirty data with different degree of cleanliness on cleaning effect

Data	0.7				0.8				0.9			
	E	*Q*	*P*	*R*	*E*	*Q*	*P*	*R*	*E*	*Q*	*P*	*R*
USCensus 1990	0.000 2	0.898 8	0.748 5	0.849 2	0.000 2	0.915 6	0.704 8	0.828 6	0.000 2	0.950 2	0.674 5	0.795 5
Bank	0.021 2	0.870 2	0.697 1	0.636 8	0.020 1	0.914 0	0.705 5	0.643 2	0.015 4	0.950 9	0.950 9	0.630 0
Adult	0.010 2	0.886 7	0.726 9	0.758 1	0.009 8	0.917 9	0.710 7	0.751 4	0.009 7	0.948 3	0.672 5	0.713 8

5 结束语

本文运用主动学习的方法,在使用机器学习的数据清洗方法基础上,部分利用用户交互,在高效的迭代清洗过程中提升数据质量,且清洗结果也有一定的可靠性。在基于主动学习的清洗框架上,本文提出确定度指标,构建了基于确定度的分类器,把建议修改值的确定度和原始数据的确定度进行对比,以此谨慎地对数据进行修改,减少错误修改事件的发生。同时,本文还提出了确定度增益指标,将最有分歧的建议修改数据筛选出来交予人工检查清洗,这不仅减少了人力资源消耗,还能更快地提升模型清洗能力,且修复了分类器容易预测错误的数据。最后,本文在多个数据集上进行实验,使用多个评价指标验证了本方法的有效性。

参考文献

[1] FEI C, MILLER R J. A unified model for data and constraint repair[C]// IEEE, International Conference on Data Engineering. Hannover, Germany:IEEE Computer Society, 2011:446-457.

[2] CHU Xu, ILYAS I F, PAPOTTI P. Holistic data cleaning: Putting violations into context[C]// IEEE International Conference on Data Engineering. Brisbane, QLD, Australia: IEEE Computer Society, 2013:458-469.

[3] GEERTS F, MECCA G, PAPOTTI P, et al. The LLUNATIC data-cleaning framework[C]// VLDB. Italy :dblp,2013:625-636.

[4] SONG Shaoxu, CHENG Hong, YU J X, et al. Repairing vertex labels under neighborhood constraints[J]. Proceedings of the VLDB Endowment, 2014, 7(11):987-998.

[5] MAYFIELD C, NEVILLE J, PRABHAKAR S. ERACER: A database approach for statistical inference and data cleaning[C]// ACM SIGMOD International Conference on Management of Data. New York, USA:ACM, 2010:75-86.

[6] YAKOUT M, ELMAGARMID A K. Don't be SCAREd: Use scalable automatic repairing with maximal likelihood and bounded changes[C]// ACM Conference on Management of Data. New York, USA:ACM, 2013:553-564.

[7] FAN Wenfei, LI Jianzhong, Ma S, et al. Towards certain fixes with editing rules and master data[J]. VLDB Journal, 2012, 21(2):213-238.

[8] RAMAN V, HELLERSTEIN J M. Potter's wheel: An interactive data cleaning system[C]// International Conference on Very Large Data Bases. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2001:381-390.

[9] VOLKOVS M, FEI C, SZLICHTA J, et al. Continuous data cleaning[C]// International Conference on Data Engineering. Chicago, IL, USA :IEEE, 2014:244-255.

[10] YAKOUT M, ELMAGARMID A K, NEVILLE J, et al. Guided Data Repair[J]. Proceedings of the VLDB Endowment, 2011, 4(5):1223-1226.

[11] CHU Xu, MORCOS J, ILYAS I F, et al. KATARA: Reliable data cleaning with knowledge bases and crowdsourcing[J]. Proceedings of the Vldb Endowment, 2015, 8(12):1952-1955.

[12] JOSHI A J, PORIKLI F, PAPANIKOLOPULOS N. Multi-class active learning for image classification[C]// IEEE Conference on Computer Vision and Pattern Recognition. Florida: IEEE, 2009: 2372-2379.

[13] FAN Wenfei, GEERTS F, JIA Xibei. Improving data quality: consistency and accuracy[C]// International Conference on Very Large Data Bases. Austria:VLDB Endowment, 2007: 315-326.

[14] LOPATENKO A, BRAVO L. Efficient approximation algorithms for repairing inconsistent databases[C]// IEEE International Conference on Data Engineering. Istanbul, Turkey: IEEE, 2007: 216-225.

[15] BALAZINSKA M, DESHPANDE A, FRANKLIN M J, et al. Data management in the worldwide sensor Web[J]. IEEE Pervasive Computing, 2007,6(2): 30-40.

[16] MADDEN S. Database abstractions for managing sensor network data[J]. Proceedings of the IEEE, 2010, 98(11):1879-1886.

[17] KOH J L Y , LEE M L , HSU W , et al. Correlation-based detection of attribute outliers[M]// KOTAGIRI R, KRISHNA P R, MOHANIA M, et al. Advances in Databases: Concepts, Systems and Applications. DASFAA 2007. Lecture Notes in Computer Science. Berlin/ Heidelberg : Springer, 2007, 4443: 164-175.

[18] ZHU Xinquan , WU Xindong . Class noise vs. attribute noise: a quantitative study of their impacts[J]. Artificial Intelligence Review,2004,22(3):177-210.

[19] HAAS D, WANG Jiannan, WU E, et al. CLAMShell: Speeding up crowds for low-latency data labeling[J]. Proceedings of the VLDB Endowment, 2015, 9(4):372-383.

[20] GOKHALE C , DAS S , DOAN A , et al. Corleone: Hands-off crowdsourcing for entity matching [C]// ACM Sigmod International Conference on Management of Data. NY, USA: ACM, 2014:601-612.

[21] MOZAFARI B , SARKAR P , FRANKLIN M , et al. Scaling up crowd-sourcing to very large datasets[J]. Proceedings of the VLDB Endowment, 2014, 8(2):125-136.

[22] SETTLES B. Active learning literature survey[J]. Science, 1995, 10(3): 237-304.