软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2020,31(4):1162−1172 [doi: 10.13328/j.cnki.jos.005801]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

面向众包数据清洗的主动学习技术*

叶 晨, 王宏志, 高 宏, 李建中

(哈尔滨工业大学 计算机科学与技术学院,黑龙江 哈尔滨 150001)

通讯作者: 王宏志, E-mail: wangzh@hit.edu.cn

摘 要: 传统方法多数采用机器学习算法对数据进行清洗.这些方法虽然能够解决部分问题,但存在计算难度大、缺乏充足的知识等局限性.近年来,随着众包平台的兴起,越来越多的研究将众包引入数据清洗过程,通过众包来提供机器学习所需要的知识.由于众包的有偿性,研究如何将机器学习算法与众包有效且低成本结合在一起是必要的.提出了两种支持基于众包的数据清洗的主动学习模型,通过主动学习技术来减少众包开销,实现了对给定的数据集基于真实众包平台的数据清洗,最大程度减少成本的同时提高了数据的质量.在真实数据集上的实验结果验证了所提模型的有效性.

关键词: 众包;数据清洗;主动学习;机器学习;领域专家

中图法分类号: TP311

中文引用格式: 叶晨,王宏志,高宏,李建中.面向众包数据清洗的主动学习技术.软件学报,2020,31(4):1162−1172. http://www.jos.org.cn/1000-9825/5801.htm

英文引用格式: Ye C, Wang HZ, Gao H, Li JZ. Active learning approach for crowdsourcing-enhanced data cleaning. Ruan Jian Xue Bao/Journal of Software, 2020,31(4):1162−1172 (in Chinese). http://www.jos.org.cn/1000-9825/5801.htm

Active Learning Approach for Crowdsourcing-enhanced Data Cleaning

YE Chen, WANG Hong-Zhi, GAO Hong, LI Jian-Zhong

(School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China)

Abstract: Traditional methods usually adopt machine learning algorithms for data cleaning. Although these methods can solve some problems, there still are computational difficulties, lack of sufficient knowledge, and other limitations. In recent years, with the rise of the crowdsourcing, more and more research has introduced crowdsourcing into the process of data cleaning, providing the extra knowledge needed for machine learning. Since workers on the crowdsourcing platforms require to be paid, it is essential to study how to effectively combine machine learning algorithms with crowdsourcing on a limited budget. This study proposes two active learning models to support crowdsourcing-enhanced data cleaning. By using active learning technology to reduce crowdsourcing cost, data cleaning based on real crowdsourcing platform is realized for given data sets, which can reduce cost and improve data quality at the same time. Experimental results on the real-world datasets show the effectiveness of the proposed methods.

Key words: crowdsourcing; data cleaning; active learning; machine learning; domain expert

当今信息化时代,互联网的兴起和产业的数字化令各种数据量急剧增长.大数据产生的同时也带来了大量劣质数据,这些劣质数据的存在可能给企业带来时间上、金钱上的浪费以及客户信任降低等不可估量的损失.对于企业,用户信息录入错误,信息过时,企业合并等都会导致劣质数据的出现,从而影响数据的质量.我们给出

* 基金项目: 国家自然科学基金(U1509216, U1866602, 61472099, 61602129); 国家重点研发计划(2016YFB1000703); 黑龙江省留学归国人员科学基金(LC2016026)

Foundation item: National Natural Science Foundation of China (U1509216, U1866602, 61472099, 61602129); National Key Research and Development Program (2016YFB1000703); Scientific Research Foundation for the Returned Overseas Chinese Scholars of Heilongjiang Provence (LC2016026)

收稿时间: 2018-07-20; 修改时间: 2018-10-08; 采用时间: 2018-12-18

几组数据[1]来展现劣质数据所产生的影响.(1) 据估计,由于地址的错误,被美国邮政部门标记为无法投递的退税支票每年超过175 000张;(2) 电话公司由于数据错误(包括50%的电话账单错误)每年损失600万英磅;(3) 糟糕的数据质量也是导致医疗健康产业中每年 98 000 人死亡的原因之一.从上述例子中我们不难看出劣质数据所造成的糟糕影响.因此,研究如何处理劣质数据,利用数据清洗提高数据质量一直是人们关注的课题.

目前关于数据清洗的研究主要包括单数据源上的数据不一致性检测与修复、缺失值填充以及多数据源上的实体识别、真值发现.尽管在各个方面都已存在众多基于机器学习算法的研究[2−14],例如贝叶斯推理被广泛地应用在数据缺失的填充上[12],现有的方法主要存在两个问题.

(1) 缺少足够的知识.在很多情况下,没有额外知识的加入,规则定义得不够全面,都将导致数据清洗的结果精确度不够高.即使在有些研究[9]中使用了专家的指导,当数据集非常大时,昂贵的代价也使得这种方法变得不可行.

(2) 复杂的数学计算.多数的机器学习方法都涉及到各类复杂的数学计算,它们之中大部分都是NP-hard难题或者是根本不可计算的.对此,这些机器学习方法的解决方法通常是提出一个近似算法或者启发式算法,然而这往往不能满足我们对数据精度的要求.

将众包引入数据清洗的过程则可以克服以上两个问题.一方面,其凭借较少的花费便能在众包平台上收集到高质量的反馈信息,从而可以代替专家的指导为机器学习提供额外的知识.另一方面,利用众包反馈来直接获取信息则可以在一定程度上避免复杂的数学计算,降低机器学习模型的复杂度.然而,将众包引入数据清洗也带来了一系列的挑战.若全部利用众包反馈进行劣质数据的挖掘与修复,其开销将是一笔不小的数目.而且,要获得大批量的众包反馈也会增加不必要的等待时间.我们希望利用众包对少量的劣质数据进行清洗就可以最大限度地提升整体数据清洗的效果,比如选择那些机器学习算法很难确定的记录或是标记后能最大限度地提升机器学习算法精确度的记录进行众包.而主动学习技术[15−18]的特点正是只选择一小部分更有价值的样本进行训练就可以使机器学习算法达到很高的精度.因此,研究如何将主动学习技术与众包相结合进行数据清洗是必要的.目前的研究主要集中在利用主动学习技术与众包相结来完成某个特定分类任务[19,20],比如图像搜索[21,22]、语义分析[22]、实体识别[22−24]等.在数据清洗领域,尽管存在主动学习与众包相结合的研究[25,26],这些研究往往存在一定的局限性[25].采用的众包为专家众包,即未考虑众包给出错误反馈的情况[26].将数据建模为不确定元组进行清洗,其针对的对象仅仅是单数据源上的元组.与以上基于主动学习技术的研究所不同的是,设计一个支持基于真实众包平台的数据清洗的主动学习技术需要满足以下条件.

(1) 通用性.数据清洗领域涉及的任务不单是在单个数据源上的不一致记录的检测和修复以及对不完整记录填充缺失值,还涉及到多数据源上的真值发现和实体识别等等.例如,判断两条记录是否代表现实生活中的同一实体和判断给定不完整记录的缺失属性值可能会需要两种不同的分类器.因此,为了能够处理不同的数据清洗任务,我们设计的主动学习技术需要能够适应不同的分类器,同时不对任何分类器的内部参数进行修改,即在对分类器的选择上要有通用性的特点.

(2) 批量性.传统的主动学习技术有些应用在时间流场景中[15−17],在这些场景中,通常在一个时间节点只能产生一条记录.如果主动学习算法决定要对当前记录进行标记,则需要等待该数据的产生而不能同时选择下一个待标记记录.而数据清洗领域的数据量通常很大,一个个对待标记记录进行选择和等待众包反馈是不现实的.因此我们设计的主动学习技术需要能够支持批量选择待众包记录,以及批量地对待众包记录进行众包反馈.

(3) 容错性.尽管我们可以从众包平台上的工人反馈中获取额外的知识,但这些知识并不一定都是正确且有效的.这是因为,在真实的众包平台上,回答问题的工人不一定是专家,不同工人的可信度并不相同,甚至存在恶意工人故意给出错误反馈的情况.因此,为了保证众包反馈的准确性,我们设计的主动学习模型需要容错机制,即在众包反馈存在噪声的情况下,依然可以找到正确的反馈值.

本文提出了满足以上 3 个条件的支持基于众包的数据清洗的主动学习技术,将主动学习与众包相结合对劣质数据进行数据清洗.本文立足于设计支持基于众包的数据清洗的主动学习模型,通过众包手段进行数据清洗以保证一定的精确度,同时结合主动学习模型来减少众包的开销.本文的主要贡献如下.

(1) 我们设计了两种满足不同需求数据清洗的主动学习模型,将以往多应用在模式识别领域加快分类的主动学习技术应用在数据清洗领域.

(2) 我们提出了利用主动学习机制结合真实众包平台来进行数据清洗的方法,在加快机器学习模型修复速度的同时提高了修复的正确性.据我们所知,目前还没有主动学习技术与真实众包平台相结合应用在数据清洗领域多个任务的研究.

(3) 在实验中,我们分别从多数据源上的真值发现和单数据源上的缺失填充两方面,对提出算法的性能进行了测试.实验结果表明,将众包与主动学习相结合在很大程度上提高了劣质数据修复的准确性.

本文第 1 节给出问题定义.第 2 节详细阐述将主动学习与众包相结合的两个机器学习算法.第 3 节展示实验部分对算法性能的验证.第 4 节得出结论.

1 问题定义

由于机器学习模型通常处理分类问题,因此我们考虑将数据清洗问题定义为分类问题,即将关系表中每个对象的每个属性值作为最小单位,利用机器学习模型判断其是否为真.这样定义的好处是不但能够通过机器学习算法对其进行判断,也方便众包上的工人对其进行直观反馈.我们将问题定义如下.

给定一个关系表 T,包含 n 个对象 $O_1,...,O_n$,每个对象由 m 个属性 $A_1,...,A_m$ 组成.每个元组由 $u=\langle O_i,A_j,v\rangle$表示,该元组表示对象 O_i 在属性 A_j 上的值为 v.若存在多个数据源提供对象 O_i 在属性 A_j 上的值,则该对象 O_i 在属性 A_j 的候选真值集为$\langle O_i,A_j,v_1,...,v_n\rangle$,其中,$v_1,...,v_n$ 为不同数据源提供的冲突值.若对象 O_i 在属性 A_j 的值缺失,则该对象 O_i 在属性 A_j 的候选真值集为$\langle O_i,A_j,dom(A_j)\rangle$,其中,$dom(A_j)$为属性 A_j 可能的取值范围.我们的目标是在对象 O_i 和其属性 A_j 给定的情况下,得到其正确值 v^*,即对于关系表中每个对象 $O_i\in\{O_1,...,O_n\}$ 的每个属性 $A_j\in\{A_1,...,A_m\}$, 返回其正确三元组$\langle O_i,A_j,v^*\rangle$.

由于我们设计的主动学习模型要满足通用性、批量性和容错性,因此问题的解决需要满足:(1) 将机器学习模型看作是一个黑盒,不对其内部进行任何处理;(2) 由于主动学习是以少量的有标记样本作为初始训练集来标记大量的无标记样本,因此需要有一个评估策略对每条元组进行信息评估,从而根据该评估策略给出的分数选择出价值更高的一类元组,利用众包平台批量进行人工标记;(3) 对于众包平台上人工标记的元组,我们从不同工人的反馈中选出最优反馈结果作为众包反馈结果.我们将问题分成 3 步来解决.

(1) 首先通过初始少量训练数据集训练一个机器学习模型 M.

(2) 使用该机器学习模型 M 对每个元组 $u=\langle O_i,A_j,v\rangle$进行确认,计算该元组的信息价值度 $Score(u)$.

(3) 主动学习根据这些元组的信息价值度选择一部分元组利用众包平台进行确认,从不同工人的反馈中选出最优反馈结果作为众包反馈结果,从而利用机器学习模型确认的结果和众包反馈的结果产生最终的修复结果.

我们将在下一节详细阐述如何在主动学习算法中通过上述 3 个步骤解决具体的数据清洗问题.

2 支持基于众包的数据清洗的主动学习技术

在主动学习过程中,我们选择信息价值更高的一类记录,利用众包平台进行人工标记.由于众包平台上工人的可信度各不相同,考虑到众包反馈结果的准确性,我们根据获得的众包反馈结果是否加入初始训练集进行再次训练,将基于众包的主动学习算法分为两种,直接主动学习算法和交互主动学习算法.下面我们将在第 2.1 节和第 2.2 节分别阐述这两种不同的学习算法及其应用场景,并在第 2.3 节讨论一些算法的实际应用问题.

2.1 直接主动学习算法

直接法的基本思想是只采用初始训练集来训练机器学习模型,训练一次即停止.然后不断地通过信息价值评估策略选择待众包元组,最后返回的修复结果集是众包反馈的元组集和机器学习算法确认过的元组集之和.由于算法在每次训练记录集时挑选最有价值的元组集送往众包,而最有价值的记录集中通常包含大部分机器

学习模型最不确定的元组,因此在初始训练集一定的情况下机器学习模型标记越来越少、越来越精确的元组集,准确度通常会越来越高.

整个算法的过程如图 1 所示,直接主动学习算法伪代码如算法 1 所示,其过程如下.

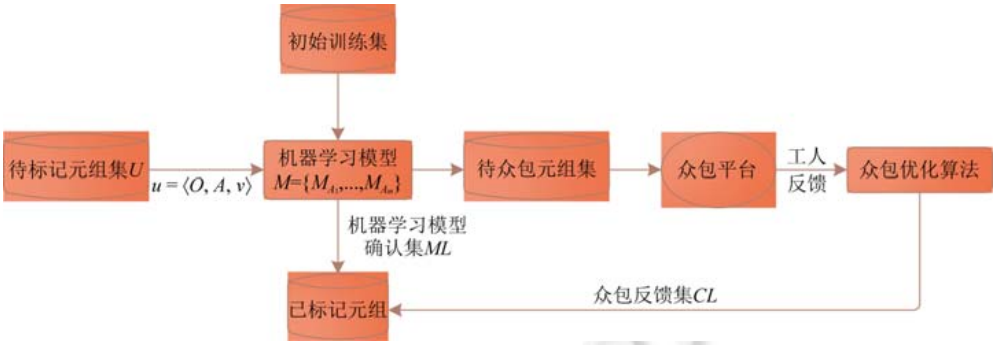


Fig.1 Direct active learning model

图 1 直接主动学习模型

算法 1. 直接主动学习算法.

输入:初始训练集 L_0,待标记元组集 U,机器学习模型 $M=\{M_{A_1},...,M_{A_m}\}$, 批量待众包元组个数 n,众包优化算法 φ,众包反馈集 CL,机器学习模型确认集 ML,准确率 Q,众包预算 B;

输出:清洗完成的数据集 L.

1: $CL\leftarrow\varnothing, ML\leftarrow\varnothing$

2: $L\leftarrow M^{L_0}(U)$

3: **while** 准确率未达到 Q or 众包预算 B 未被用光 do

4: 从待标记元组集 U 中选择信息价值最高的 n 条元组组成待众包元组集 U'

5: 众包平台标记待众包元组集 U'获得工人反馈结果集 $R^{U'}$

6: 获得本轮众包反馈最优结果集 $L'\leftarrow\varphi(R^{U'})$

7: $CL\leftarrow CL\cup L'$

8: $U\leftarrow U-U'$

9: $ML\leftarrow M^{L_0}(U)$

10: **done**

11: $L\leftarrow ML\cup CL$

12: **return** L

(1) 学习模型初始化.首先确定我们的研究对象是一个三元组 $u=\langle O_i,A_j,v\rangle$,表示元组 O_i 在属性 A_j 中的一个修复值 v.我们针对关系表中的每个属性 $\{A_1,...,A_m\}$ 训练了一系列分类器模型 $\{M_{A_1},...,M_{A_m}\}$, 这些分量分类器构成了我们的总体分类器 M.给定一个三元组 $u=\langle O_i,A_j,v\rangle$,模型 M_{A_j} 用于预测该修复的正确性.这个阶段主要是通过初始训练集 L_0 中的少量记录对各个分量分类器进行训练,从而得到一个初始的机器学习模型 M(1~2 行).

(2) 选择待众包元组.在这个阶段,我们首先利用初始机器学习模型对待标记元组集 U 进行预测.对每个待标记元组 $u=\langle O_i,A_j,v\rangle$,相应的分量分类器 M_{A_j} 给出其预测结果以及信息价值度,最后选择信息价值较大的元组进行人工标记.这个阶段主要是利用待标记元组对应的分量分类器反馈的信息价值度来排序,将信息价值最大的 n 个元组组成待众包记录集 U'(3~4 行).

(3) 众包反馈.在这个阶段,众包平台上的工人对机器学习模型挑选出来的待众包元组集 U'进行标记,工人对每一个待众包元组 $u=\langle O_i,A_j,v\rangle$给出反馈 $R\in\{$确认,拒绝$\}$.针对同一个记录的多个工人反馈,我们采用众包优化算法 φ选出最可能的真值并返回.得到众包的反馈结果后,机器学习模型将进行重新训练,去除掉那些已经得到

标记的元组,在剩下的元组产生待众包元组集合.由于阶段 2 中已经选择了分类器最不确定的元组进行标记,所以在待标记元组集合中的会是分类器不确定度相对低的元组,因而在下次迭代中学习模型的准确率会越来越高(5~10 行).

(4) 结果反馈.重复阶段 1～阶段 3,直到分类结果已经达到一定准确率 Q 或众包预算 B 被用光,合并众包反馈集 CL 和机器学习模型确认集 ML,从而产生最终的修复结果 L,数据集的修复完成(11~12 行).

直接主动学习算法在每次机器学习模型标记元组集时挑选最不确定的元组送往众包.因此,该学习模型标记剩下的元组集的准确度通常比上一次的要高.算法的时间复杂度为 $O(m\times T(|U|))$,其中,m 为机器学习模型中分量分类器的个数,$T(\cdot)$代表分量分类器的运行时间,$|U|$是每一次迭代待标记元组的个数.当分类的准确性很高时,一次迭代就可以完成全部数据修复.

该算法由于在训练机器学习模型时只考虑了初始训练集,因此适用于一些初始训练集信息量就已经非常有效的情况,以及对精度要求非常高而使训练集的元组只能是正确元组的情况.另外,该算法在等待众包平台反馈的同时,可以获得通过机器学习模型确认后的这部分正确元组,因此也适用于用户希望先看到一部分清洗后的元组的情况.然而,由于其初始训练集的个数有限,其精度会依赖于初始训练集的数量及分布,不能够充分利用众包数据集的价值,提升整体精确度的空间有限.

2.2 交互主动学习算法

上文我们讨论了直接主动学习的优缺点.当初始训练集的个数不够多时,训练得不够充分,机器学习模型的准确度是有限的.为了解决这个问题,我们提出了交互主动学习算法,考虑将众包标记过的元组增加到训练集,对机器学习模型进行重新训练.然后,用重新训练后更加精确的机器学习模型对剩余的待标记元组进行标记.在交互主动学习过程中,机器学习模型不断地筛选待众包元组,将得到的众包反馈作为新增的训练集对自身进行重新训练.在众包反馈的准确率和效率有保证的情况下,这种方法能够在一定程度上提高机器学习模型的标记精度,从而提高最终的修复结果的准确率.

交互主动学习的整个过程如图 2 所示,伪代码如算法 2 所示,算法分为 4 个步骤.

(1) 学习模型初始化.这一阶段和直接主动学习模型是一致的(1~2 行).

(2) 选择待众包记录.在这个阶段,我们首先利用初始机器学习模型对待标记元组集 U 进行预测.对每个待标记元组 $u=\langle O_i,A_j,v\rangle$,相应的分量分类器 M_{A_j} 给出其预测结果以及信息价值度,最后选择信息价值较大的元组进行人工标记.这个阶段主要是利用待标记元组对应的分量分类器反馈的信息价值度来排序,将信息价值最大的 n 个元组组成待众包元组集 U'(3~4 行).

(3) 结果反馈和学习模型重训练.在这个阶段,众包平台上的工人对学习模型挑选出来的待众包元组进行标记,通过优化算法整合众包平台的反馈,从而得到众包最优反馈结果.机器学习模型重新对待标记元组集合进行确认,再一次产生待众包元组集合.由于阶段 2 中选择了信息价值最大的元组进行标记,因此在下一次迭代中众包反馈结果加入到初始训练集中进行再训练后,机器学习模型的分类正确性将得到最大加强(5~10 行).

(4) 循环训练.重复阶段 1~阶段 3,直到已经达到一定准确率 Q 或众包预算 B 被用光,则数据集的修复完成(11~12 行).

利用更新训练集使机器学习模型重训练来提高其分类的准确性.其时间复杂度为

$$O(m\times T(|U|+|L_0\cup CL|),$$

其中,m 为机器学习模型中分量分类器的个数,$T(\cdot)$代表分量分类器的运行时间,$|U|$是每一次迭代待标记元组的个数,$|L_0\cup CL|$是每轮训练集包含的元组个数,是初始训练集个数和该轮众包反馈的元组集个数之和.

该算法由于每次将众包反馈的结果加入到初始训练集进行重新训练,当众包反馈的精确度可以保证时,更多的训练元组的加入将使机器学习模型对剩余待标记元组的标记更加准确.然而,当众包的反馈不够准确时,可能导致机器学习模型的精确度有所下降.另外,该算法由于每次要等待众包反馈的结果进行重训练,总体的等待时间比直接主动学习算法所需时间要长.因此,该算法比直接主动学习算法更适用于众包反馈质量高且可利用的时间充足的场景.

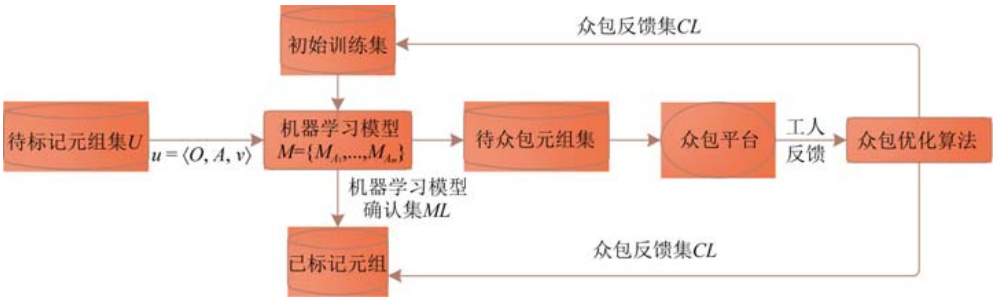


Fig.2 Mutual active learning model

图 2 交互主动学习模型

算法 2. 交互主动学习算法.

输入:初始训练集 L_0,待标记元组集 U,机器学习模型 $M=\{M_{A_1},...,M_{A_m}\}$,批量待众包元组个数 n,众包优化算法 φ,众包反馈集 CL,机器学习模型确认集 ML,准确率 Q,众包预算 B;

输出:清洗完成的记录集 L.

1: $CL\leftarrow\varnothing, ML\leftarrow\varnothing$

2: $L\leftarrow M^{L_0}(U)$

3: **while** 准确率未达到 Q or 众包预算 B 未被用光 do

4: 从待标记元组集 U 中选择信息价值最高的 n 条元组组成待众包元组集 U'

5: 众包平台标记待众包元组集 U' 获得工人反馈结果集 $R^{U'}$

6: 获得本轮众包反馈最优结果集 $L'\leftarrow\varphi(R^{U'})$

7: $CL\leftarrow CL\cup L'$

8: $U\leftarrow U-U'$

9: $ML\leftarrow M^{L_0\cup CL}(U)$

10: **done**

11: $L\leftarrow ML\cup CL$

12: **return** L

2.3 讨 论

在本小节,我们讨论一些上述算法的实际应用问题,分别为初始训练集生成、机器学习模型创建与分类以及众包优化算法选择.

初始训练集生成.我们的主动学习模型需要一小部分已标记的元组作为初始训练集来训练机器学习模型 M,这部分已标记的元组可以来自专家确认或领域知识.其中,分量分类器 M_{A_j} 构造的训练集形式为 $\langle O_i, A_j, v_1, ..., v_n, v^*, R\rangle$, $v_1,...,v_n$ 为对象 O_i 在属性 A_j 上所有可能的取值,v^* 是真值, $R\in\{$确认,拒绝$\}$ 是标记结果.

机器学习模型创建与分类.算法中的机器学习模型可以是贝叶斯模型、决策树或者支持向量机等等.在我们的实验中,每一个分量分类器 M_{A_j} 都是由 k 个决策树组成的随机森林.假设训练集中的元组个数为 Z,每一个决策树的训练方法如下:随机地从原始数据中取样大小为 $S<Z$ 的训练集 L.通过训练集 L 训练出一棵决策树.为了进一步降低决策树与决策树之间的相关性,在每个属性 A_j 分裂时,随机取属性集 $A'\subseteq\{A_1,...,A_{j-1},A_{j+1},...,A_m\}$ 作为其特征.当初始的机器学习模型创建好后,对于一个三元组 $u=\langle O_i,A_j,v\rangle$, M_{A_j} 中的每一棵决策树分别对 u 进行分类,而在获得的反馈结果集合$\{R_1,...,R_k\}$中,投票数最大的分类结果就是分量分类器 M_{A_j} 最终的分类结果.在衡量每个三元组的分类结果的信息价值程度时,本文采用了最大熵策略,即将反馈结果最不一致的元组视为信息价值最大的元组.当然,其他衡量信息价值的方法,例如边缘误差策略或最不确定策略等等也可以用在我们的主动模型中.

众包优化算法选择.针对众包可能返回错误结果的情况,我们需要众包优化算法来过滤众包的错误结果.常用的众包优化算法可采用投票算法,或通过估计工人的可信度来决定最优反馈[11].本文采用文献[11]中提出的方法来决定最优反馈,即工人的可信度越高,其提供的反馈越可能是正确的.

3 实 验

本节我们将在真值发现和缺失填充两个代表性数据清洗任务上验证提出的两个模型的高效性和准确性.首先给出实验采用的数据集和实验设置,然后分别从主动学习的有效性、最大信息价值元组评估机制、众包反馈结果的有效性这 3 个方面来验证直接主动学习和交互主动学习的准确率.

数据集.

Stock.该数据集来自于纽约州立大学宾汉姆顿分校计算机研究室,其收集了 2011 年 7 月的所有工作日约 1 000 只股票的信息,其中包括分别来自 55 个数据源的 16 个属性,并且提供了真值.我们的目标是从 55 个数据源提供的数据中发现真值,并通过与给定真值的对比得出准确率.我们用该数据集来验证所提出的模型在真值发现应用上的有效性.对于每个股票的每个属性,我们将全部 55 个数据源提供的不同值作为其候选真值.

Wine、Chess.这两个数据集来自于加州大学欧文分校机器学习数据库(UCI),它们分别包括 200 条记录,13 个属性;3 196 条记录,37 个属性.我们对于每一条记录随机置空其某个属性值作为缺失值.对于每个缺失值,将该对应属性在所有记录上的全部取值作为其候选真值.我们用这两个数据集来验证所提出的模型在缺失填充应用上的有效性.

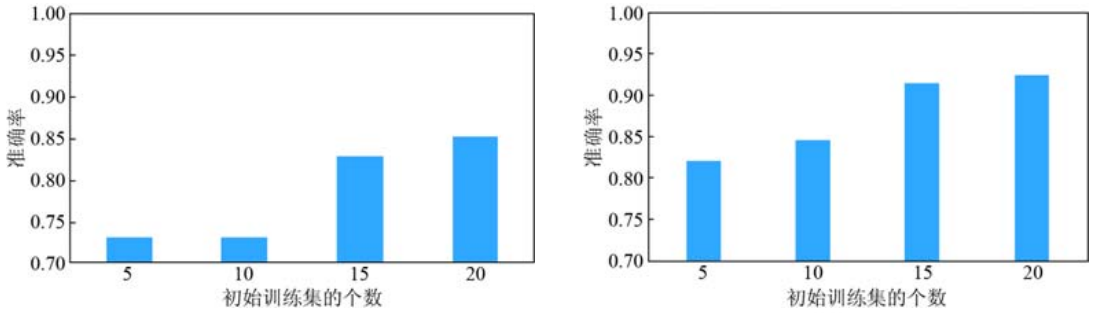
实验设置.

所有实验均在 Intel(R) Core(TM) i5-2400 (3.10 GHz) CPU 和 8GB 内存的机器上运行,我们多次重复实验并记录每个实验结果的平均值.所有的算法用 Python 实现.我们选择决策树作为组成机器学习模型的分类器,并选择信息熵最大的元组,即最不确定的元组作为待众包元组.在 Stock 数据集上,我们将信息价值大于 0.8 的元组视为待众包元组,并在真实众包平台 Amazon Mechanical Turk(AMT)上采集来自不同工人对待众包元组的反馈.我们为完成每个 HIT 任务的工人提供 0.01 美元的报酬.为了减少开销,将每 20 个待众包元组放入一个 HIT 任务.为了保证众包反馈的准确度,每个 HIT 被分发给 3 个工人,并采用众包优化算法[11]的结果作为最终的众包反馈结果.在 Wine 和 Chess 数据集上,我们以不同概率生成错误的众包反馈结果来测试模型的性能.

3.1 主动学习的有效性

首先我们对初始训练集的个数对机器学习模型准确率的影响进行实验,实验结果如图 3 所示.由于主动学习的特点是在初始训练集个数很少的情况下对待标记元组有选择地加以标记,因此我们将初始训练集的个数设定为 5、10、15、20,对分类器分别训练.初始训练集的个数对机器学习模型准确率的影响如图 3 所示,由图中可以得知,机器学习模型的准确率随着初始训练集个数的增大而增大.由图 3(a)可以看出,当初始训练集内的元组少于 10 条时,机器学习模型的准确率在 73%左右.当训练集内的元组超过 15 条时,机器学习模型的准确率超过 80%.而由图 3(b)可以看出,当训练集内的元组超过 15 条时,机器学习模型的准确率已可达到 90%,这是因为我们的主动学习技术每次都选择价值最大的元组进行标记.因此,能够很快地提升机器学习模型的准确度.

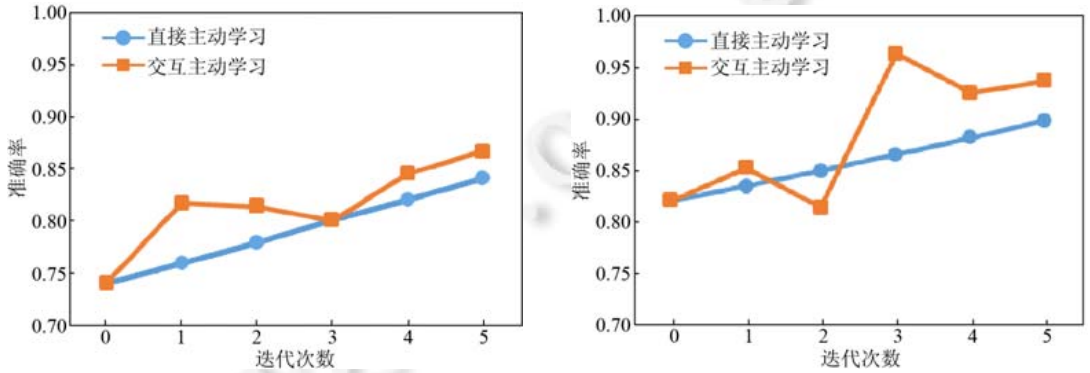
下面我们对主动学习模型的有效性进行验证.当训练集个数为 20 个、待标记元组集的个数为 200 个时,每一次迭代的准确性变化如图 4 所示.在直接主动学习模型中,由于每一次都选择机器学习模型不确定的元组进行众包,因此机器学习模型标记错误的元组个数会一直在减少.因此,如图 4(a)和图 4(b)所示,直接主动学习算法的准确性都随着迭代次数的增加而不断提升.而在交互主动学习模型中我们发现,每次迭代将已标记的元组加入到初始训练集的做法并不总是能够提高机器学习模型的准确率,这是因为,其中存在众包反馈错误和过拟合的影响.尽管如此,直接主动学习和交互主动学习模型凭借主动学习技术依然可以用很小的训练集在几次迭代后达到 80%以上的准确率.



(a) Stock (b) Wine

Fig.3 The effectiveness of the number of tuples in the training set

图 3 初始训练集的个数对分类器准确率的影响



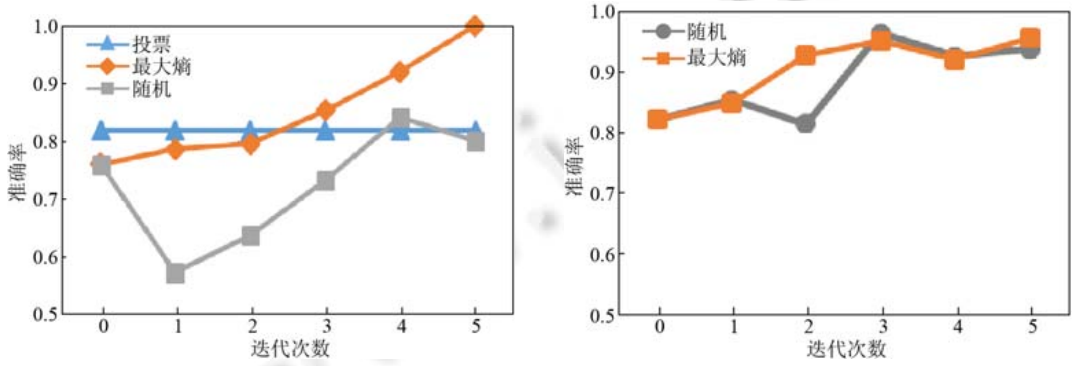
(a) Stock (b) Wine

Fig.4 The effectiveness of the number of iterations

图 4 不同学习方法的迭代次数对分类器准确率的影响

3.2 最大信息价值元组评估机制

本节将测试选择最不确定的元组作为待众包元组的有效性,我们将其与随机算法和投票算法进行对比(由于 Wine 数据集不存在多个数据源提供数据,故只在 Stock 上应用投票算法).在随机算法中,我们将随机选择元组进行众包标记.在投票算法中,我们选择占比最大的候选值作为真值.在 Stock 数据集上实验采用直接主动学习模型实现,在 Wine 数据集上采用交互主动学习模型实现,实验结果如图 5 所示.



(a) Stock (b) Wine

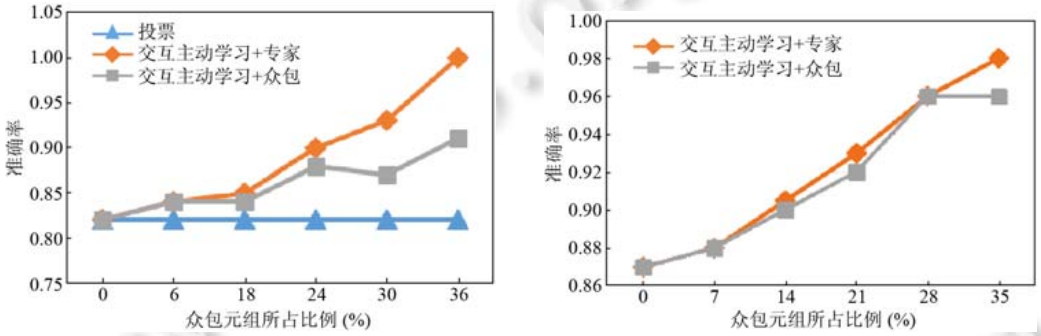
Fig.5 The effectiveness of crowd tuples selection

图 5 选择待众包元组的方法对准确率的影响

从图中我们可以看出,随机算法的随机性非常大,每次记录的选择对准确度都有较大的影响.如图 5(a)所示,投票算法的准确率在 0.82 左右,本文提出的最大熵方法在第 3 次迭代时其准确率就超过了投票算法,而且准确率随着迭代次数的增加稳步上升,可以看出,我们采用最大熵方法选择元组进行众包对比投票算法和随机算法具有很大的优势.如图 5(b)所示,依然可以看出我们的最大熵方法相较于随机方法更稳定,有更高的准确率.

3.3 众包反馈结果的有效性

本节我们测试众包反馈结果的有效性,众包元组个数占全部待标记元组个数的比例对准确率的影响如图 6 所示,实验采用交互主动学习算法进行实验.Stock 数据集上的众包反馈来自真实的众包平台 AMT,Chess 数据集上的众包反馈以 10%的错误概率自动生成.为了更好地展现引入众包模块的效果,我们将其和交互主动学习与专家指导结合的方法作对比.我们将专家指导的准确率设为 1.从图 6(a)可以看出,我们的交互主动学习算法在没有众包参与时就已经可以达到和投票算法差不多的准确率,随着众包元组所占比例的升高,准确率也在逐步上升.当众包元组所占比例达到 36%时,结合众包反馈的模型准确率即可达到 90%以上,结合专家反馈的模型准确率接近于 1.尽管众包反馈的准确率没有专家指导的高,但其也远远好于投票算法.如图 6(b)所示,当众包元组所占比例达到 14%时,结合众包反馈和专家反馈的模型准确率都可达到 90%以上.由此可以看出,将众包与主动学习相结合应用于真值发现和缺失填充领域都能起到很好的效果.

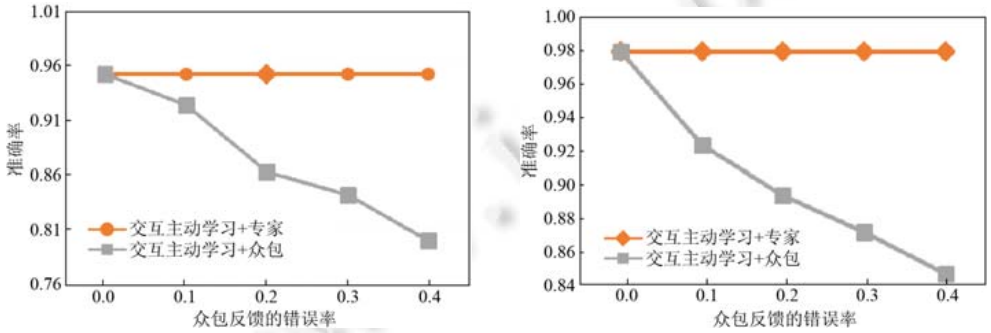


(a) Stock (b) Chess

Fig.6 The effectiveness of the proportion of crowd tuples

图 6 众包记录所占比例对准确率的影响

下面我们测试众包反馈的错误率对主动学习模型的准确率的影响.实验采用交互主动学习算法进行,众包元组所占比例设为 20%.Wine 和 Chess 数据集上的众包反馈以不同错误率自动生成.我们依然采用交互主动学习与专家指导的方法作对比,其中,专家指导的准确率设为 1.实验结果如图 7 所示.



(a) Wine (b) Chess

Fig.7 The effectiveness of crowd feedback

图 7 众包反馈的错误率对模型准确率的影响

可以看到,随着众包反馈错误率的升高,模型的准确率在这两个数据集上都呈下降趋势.当众包反馈的错误率达到 40%时,我们的主动学习模型结合众包反馈的准确率有较大程度的下降,这说明错误的反馈影响了我们模型的准确度,而当众包反馈的错误率在 20%之内时,我们的模型依然有较好的表现.

3.4 小 结

通过以上实验,有效地验证了本文提出的两种支持基于众包的数据清洗的主动学习算法.我们发现:(1) 直接主动学习算法和交互主动学习算法都能够有效地提升基于众包的数据清洗的精确性.(2) 直接主动学习算法对精确度的提升更稳定也更有限.(3) 交互主动学习算法对精确度的提升很大程度上取决于众包反馈的质量,反馈质量越高,可提升的准确度就越高.

4 结 论

由于劣质数据在互联网时代普遍存在,为了提高数据质量,我们需要对脏数据进行数据清洗.本文提出的主动学习模型即为利用机器学习技术改善数据质量的一种尝试.

本文的研究是在主动学习和众包平台相结合的基础上展开的,对如何将主动学习与众包结合在一起提出了两种模型,直接主动学习模型和交互主动学习模型,适用于不同的数据应用场合.将主动学习与众包结合起来,在解决准确性问题的同时减少了开销,从而使模型更加健壮,适用范围更加广泛.

未来的主要研究方向包括更好地评估众包平台上的工人可信度,在主动学习模型中加入容错机制等等.

References:

[1] Chen Y. Research on key technologies of data cleaning based on crowdsourcing [MS. Thesis]. Harbin: Harbin Institute of Technology, 2015 (in Chinese with English abstract).

[2] Fan WF, Geerts F. Foundations of data quality management. Synthesis Lectures on Data Management, 2012,4(5):1–217.

[3] Cong G, Fan WF, Geerts F, *et al*. Improving data quality: Consistency and accuracy. Proc. of the VLDB Endowment, 2007, 315–326.

[4] Wang JN, Krishnan S, Franklin MJ, *et al*. A sample-and-clean framework for fast and accurate query processing on dirty data. In: Proc. of the SIGMOD Int'l Conf. on Management of Data. ACM, 2014. 469–480.

[5] Batini C, Cappiello C, Francalanci C, *et al*. Methodologies for data quality assessment and improvement. ACM Computing Surveys, 2009,41(3):16.

[6] Tong Y, Cao CC, Zhang CJ, *et al*. Crowdcleaner: Data cleaning for multi-version data on the Web via crowdsourcing. In: Proc. of the Int'l Conf. on Data Engineering. IEEE, 2014. 1182–1185.

[7] Galhardas H, Florescu D, Shasha D, *et al*. AJAX: An extensible data cleaning tool. ACM SIGMOD Record, 2000,29(2):590.

[8] Raman V, Hellerstein JM. Potter's wheel: An interactive data cleaning system. Proc. of the VLDB Endowment, 2001,1:381–390.

[9] Jeffery SR, Franklin MJ, Halevy AY. Pay-as-you-go user feedback for dataspace systems. In: Proc. of the SIGMOD Int'l Conf. on Management of Data. ACM, 2008. 847–860.

[10] Rahm E, Do HH. Data cleaning: Problems and current approaches. Database Engineering Bulletin, 2000,23(4):3–13.

[11] Ye C, Wang HZ, Gao H, *et al*. Truth discovery based on crowdsourcing. In: Proc. of the Int'l Conf. on Web-age Information Management. Springer Int'l Publishing, 2014. 453–458.

[12] Ye C, Wang HZ, Li JZ, *et al*. Crowdsourcing-enhanced missing values imputation based on Bayesian network. In: Proc. of the Int'l Conf. on Database Systems for Advanced Applications. Springer Int'l Publishing, 2016. 67–81.

[13] Ye C, Li Q, Zhang HT, *et al*. AutoRepair: An automatic repairing approach over multi-source data. Knowledge and Information Systems, 2019,61(1):227–257.

[14] Li JZ, Wang HZ, Gao H. State-of-the-art of research on big data usability. Ruan Jian Xue Bao/Journal of Software, 2016,27(7): 1605–1625 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/5038.htm [doi: 10.13328/j.cnki.jos.005038]

[15] Beygelzimer A, Dasgupta S, Langford J. Importance weighted active learning. In: Proc. of the Int'l Conf. on Machine Learning. ACM, 2009. 49–56.

[16] Beygelzimer A, Langford J, Tong Z, *et al*. Agnostic active learning without constraints. In: Advances in Neural Information Processing Systems. MIT Press, 2010. 199–207.

[17] Dasgupta S, Monteleoni C, Hsu DJ. A general agnostic active learning algorithm. In: Advances in Neural Information Processing Systems. MIT Press, 2007. 353–360.

[18] Settles B. Active learning literature survey. Computer Sciences Technical Report, 1648, University of Wisconsin-Madison, 2009.

[19] Yan Y, Rosales R, Fung G, *et al*. Active learning from crowds. In: Proc. of the Int'l Conf. on Machine Learning. ACM, 2011,11: 1161–1168.

[20] Zhong J, Tang K, Zhou ZH. Active learning from crowds with unsure option. In: Proc. of the Int'l Joint Conf. on Artificial Intelligence. AAAI Press, 2015. 1061–1067.

[21] Parameswaran A, Sarma AD, Garcia-Molina H, *et al*. Human-assisted graph search: It's okay to ask questions. Proc. of the VLDB Endowment, 2011,4(5):267–278.

[22] Mozafari B, Sarkar P, Franklin M, *et al*. Scaling up crowd-sourcing to very large datasets: A case for active learning. Proc. of the VLDB Endowment, 2014,8(2):125–136.

[23] Wang J, Kraska T, Franklin MJ, *et al*. Crowder: Crowdsourcing entity resolution. Proc. of the VLDB Endowment, 2012,5(11): 1483–1494.

[24] Whang SE, Lofgren P, Garcia-Molina H. Question selection for crowd entity resolution. Proc. of the VLDB Endowment, 2013,6(6): 349–360.

[25] Chu X, Morcos J, Ilyas IF, *et al*. KATARA: A data cleaning system powered by knowledge bases and crowdsourcing. In: Proc. of the SIGMOD Int'l Conf. on Management of Data. ACM, 2015. 1247–1261.

[26] Zhang CJ, Chen L, Tong Y, *et al*. Cleaning uncertain data with a noisy crowd. In: Proc. of the Int'l Conf. on Data Engineering. IEEE, 2015. 6–17.

附中文参考文献:

[1] 叶晨.基于众包的数据清洗关键技术的研究[硕士学位论文].哈尔滨:哈尔滨工业大学,2015.

[14] 李建中,王宏志,高宏.大数据可用性的研究进展.软件学报,2016,27(7):1605–1625. http://www.jos.org.cn/1000-9825/5038.htm [doi: 10.13328/j.cnki.jos.005038]

叶晨(1992－),女,浙江乐清人,硕士,CCF 学生会员,主要研究领域为数据质量,劣质数据清洗,真值发现与模式发现.

高宏(1966－),女,博士,教授,博士生导师,CCF 杰出会员,主要研究领域为复杂结构数据管理,无线传感器网络.

王宏志(1978－),男,博士,教授,博士生导师,CCF 杰出会员,主要研究领域为数据质量管理,海量数据管理,知识图谱,XML 数据管理,工业大数据.

李建中(1950－),男,博士,教授,博士生导师,CCF 会士,主要研究领域为数据库系统实现技术,数据仓库,半结构化数据,传感器网络,压缩数据库技术,Web 数据集成,数据挖掘,计算生物学.