人工参与的迭代式数据清洗方法研究

刘一达，丁小欧，王宏志，杨东华
哈尔滨工业大学计算机科学与技术学院，黑龙江 哈尔滨 150001

摘要
数据采集技术的进步导致了数据集规模的飞速上涨，由于数据的大规模和高复杂性引起了严重的数据质量问题，数据清洗是数据活动中必要且重要的环节。为了在保证清洗准确率的情况下有效地降低人工标注成本，提出了一种人工参与的迭代式的数据清洗方法（IDCHI）。该方法在检测模块中提出了数据选择优化方法，使分类器在初始阶段就拥有较高的准确度；并进一步提出了待人工标注数据选择方法，有效地降低人工标注的数据量。实验结果表明该方法可有效且高效地清洗错误数据。

关键词
数据清洗；人工参与；迭代式；小批量梯度下降

中图分类号：TP311 文献标志码：A doi: 10.11959/j.issn.2096-0271.2023048

Research on iterative data cleaning of human-computer interaction

LIU Yida, DING Xiaoou, WANG Hongzhi, YANG Donghua
School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China

Abstract
The advancement of data collection technology has led to a rapid increase in the size of datasets. Due to the big scale and high complexity of the data volume, serious data quality issues arise. Therefore, data cleaning is a necessary and important step in data activities. To effectively reduce human annotation costs while ensuring the accuracy of cleaning, an iterative data cleaning method (IDCHI) with human participation was proposed. This method proposed a data selection optimization method in the detection module, which enables the classifier to have high accuracy in the initial stage; and further proposed a method for selecting data to be manually annotated, effectively reducing the amount of data to be manually annotated. The experimental results show that the proposed method is effective and efficient in cleaning erroneous data.

Key words
data cleaning, human_in_loop, iteration, mini-batch gradient descent

0 引言

随着数据采集技术的不断进步，数据集的规模飞速上涨，针对数据的处理成为不同程序中一个重要的任务。数据由不同的传感器产生，这些数据总量庞大。但是数据量的庞大以及数据来源的复杂也导致出现错误数据的概率增加，因为数据之间存在关联关系，这种关联性会导致新的错误类型产生[1]。为了解决这些错误数据引发的问题，如果直接删除错误的数据会破坏数据之间关系的完整性，降低数据的价值。因此不仅需要对错误数据进行检测，更需要采取方法将其清洗为干净的数据。

数据的复杂性和不确定性使自动化修复算法很难以100%的置信度和准确率修复错误数据。以领域专家和用户为代表的人工参与方法是通用数据清洗和持续数据清洗的重要部分[2-4]。近年来，范举等人[5]提出了人在回路的数据准备概念，归纳了在数据提取、标注、集成、清洗等数据准备过程[6]中的人工参与方法和人工任务。相比于自动化修复算法，人工修复具有修复准确率高、可靠性强，且对特定领域的数据修复效果好的优势[7]，但同时也具有修复成本高的问题[8]。因此，如何在提高数据清洗效果的同时减少人工标注的成本，是数据清洗中面临的主要挑战。

本文的研究着重于通过人机结合进行迭代式的数据清洗。本文的贡献主要在于以下3个方面。

- 本文提出了人工参与的迭代式清洗算法（iterative data cleaning of human-computer interaction，IDCHI）。本方法结合人工参与以及迭代式的方法，利用规则依赖将数据集分为符合规则的数据以及低质量数据，并对低质量数据进行迭代式的人工修复，实现了较高的清洗效果。
- 本文提出了一种针对人工标注阶段数据的挑选方法。本文通过优化分类器模型和计算违反分数的方式，将针对单一维度内错误数据的违反分数计算、针对单一数据同一维度综合的违反分数计算、针对不同数据不同维度之间错误数据的违反分数计算这3种不同的违反分数结合，提高了违反分数的代表性，让违反分数较高的数据更可能成为错误数据，提高了挑选参与人工标注的错误数据的准确度，使数据可以通过较少的人工标注工作量得到较高的准确率。
- 本文在数据集上展开了大量的测试。人工参与的迭代式数据清洗方法可以通过较少的样本规模实现较高质量的错误检测任务，可以显著地减少达到一定准确度的数据清洗所需要的标注数量，进而减少人工标注的负担。

1 数据清洗工作的研究现状

目前，国内外对于时间序列上错误数据的研究往往集中于基于数据统计特征进行数据清洗的方法和基于规则依赖等先验知识进行数据清洗的方法。

1.1 基于统计特征的数据清洗方法

基于统计特征的数据清洗方法通常根据目前已知序列的分布，通过计算数据本身的统计量和统计指标，使用聚类等方法将具有接近相似系数的数据进行聚类，从而对错误数据进行清洗[9-12]。2016年Krishnan等人[13]提出的ActiveClean算法通过判断数据在对应模型中成为劣质数据的可能性来挑选要清洗的数据样本。近年来有学者提出了使用自动编码器[14]的深度学习方法。该方案将数据转换到低维空

间，并通过解码器进行重构，提取出数据中的特征，正确的数据将会得到较好的重构，而错误数据的重构将会出现问题。Le等人[15]通过计算数据的违反分数，结合逆最近邻（INN）算法，计算数据的幅值分数、相关分数和方差分数3类违反分数，进行决策树的构建，并最终利用决策树上的相关系数进行聚类，根据聚类结果通过人工修复来进行数据清洗。

1.2 基于规则依赖的数据清洗方法

基于规则依赖的数据清洗方法通过现有先验知识或者其学习的方式从已清洗过的数据中推断规则依赖的方式，减少人工参与的成本，通过这些规则依赖对其余数据进行一定的清洗[16-18]。Charfi等人[19]通过将数据分为时空上的不同粒度，对不同时空粒度的数据采用不同粒度的约束进行对应的处理，实现了较为精细的数据清洗。范举等人[5]提出了人在回路的数据准备概念，通过基于众包的数据准备技术结合大量众包工作者来提升计算能力，从而支持数据准备的基本任务，对清洗质量和清洗成本进行控制。相比于自动化修复算法，人工修复具有修复准确率高、可靠性强，且对特定领域的数据修复效果好的优势[20]，但同时有修复成本高的问题。

2 研究问题介绍

2.1 问题定义

定义1：多元序列。对于输入的待清洗的原始数据，可以将其定义为多元序列$X=\{X_1,X_2,\cdots,X_n\}$，其中每个元组的特征集合为$X_t=\{f_1,f_2,\cdots,f_n\}$，每个$f$表示元组$X$的一个特征。同时定义多元序列$X_t$表示时间戳为$t$时，多元序列$X$上所有的数据集合。

定义2：规则依赖。令R表示一个关系，它包含m个属性$\mathrm{Attrs}(R)=(A_1,\cdots,A_m)$表示$R$上的属性集合，$\mathrm{Dom}(A)$表示一个给定属性$A$的域。令$I$表示关系$R$的一个实例，包含若干元组，各元组均属于域$\mathrm{Dom}(A_1)\times\cdots\times\mathrm{Dom}(A_m)$。令$\mathrm{Dom}I(A)$表示属性$A$的空间，它包括所有出现在实例$I$中的$A$属性值。假设$I$中的每个元组均有一个标识符，即使元组的其他属性都发生变更，该标识符也不会改变。令$\mathrm{TIDs}(A)$表示在实例I中的所有元组的标识符的集合。令$t[A]$表示元组t的一个单元，其中，$A\in\mathrm{Attrs}(R)$，$t\in\mathrm{TIDs}(I)$。每一个单元$t[A]$由元组以及属性来确定。

在R上定义一个函数依赖集合，包含多个函数依赖。对于两个属性集合X和Y，它们均属于$\mathrm{Attrs}(R)$。基于实例I的一个函数依赖，$X\rightarrow Y$被表示为$I \trianglelefteq X \rightarrow Y$。换言之，对于实例$I$中的任意两个元组$t_1$和$t_2$，如果$t_1[X]=t_2[X]$成立，则$t_1[Y]=t_2[Y]$必然成立。令$\Sigma$表示基于关系$R$的函数依赖集合。本文假设$\Sigma$是正则最小化的。每个函数依赖均可以被描述为如下的形式：$X\rightarrow A$。其中，$X\in\mathrm{Attrs}(R)$，且$A\in\mathrm{Attrs}(R)$。

定义3：错误数据。假设正确的序列为X_{true}，得到的原始数据中的序列为X_{normal}。如果$X_{\mathrm{true}}=X_{\mathrm{normal}}$，那么就可以称这个数据点为正确的数据点，如果$X_{\mathrm{true}}!=X_{\mathrm{normal}}$，那么就可以称这个数据点为一个错误数据。

结合规则依赖的数据清洗指基于给定的规则依赖集合Σ对劣质数据集进行清洗，将检测出其中的错误数据，并将其通过人工或者自动化的方式进行修复。

2.2 方法框架

本文的方法通过计算违反分数的方法对数据进行计算，从中找出违反分数较

高的数据。如**图1**所示，首先针对原始数据集，基于现有的规则依赖对数据进行初步检测，将原始数据集分为违反规则的数据和符合规则的数据，其中符合规则的数据会对接下来的分类器模型进行初始化，之后这两部分数据集共同组成劣质数据集。接下来通过更新后的分类器对劣质数据集进行违反分数的计算，从中挑选出高违反分数的错误数据，在人工修复部分对错误数据进行修复，再将这些修复过后的数据集传给分类器，对分类器模型进行更新，修复后的数据构成干净数据。反复迭代以上步骤，就可以提高训练模型的精度。

3 人工参与的迭代式数据清洗

3.1 检测模块

数据清洗的第一步就是基于现有的规则依赖对原始数据集进行检测，并对原始数据集进行分类，将原始数据集X分为违反规则的数据X_{vio}和符合规则的数据X_{acc}，其中符合规则的数据X_{axx}将对分类器进行初始化，违反规则的数据X_{vio}和符合规则的数据X_{acc}这两部分数据都会组成劣质数据集，并在之后的迭代中使用。

虚构数据举例见**表1**。

假设存在规则依赖：学号→姓名，年龄，专业。即在确定学号的基础上就能确定对应的姓名、年龄和专业。但是**表1**中第1行数据和第3行数据在学号相同的情况下专业不同，因此第1行数据和第3行数据就是违反规则的数据。

3.2 分类器的更新与数据的自动修复

本文方法中的分类器输出对劣质数据的预测值，该预测值可作为下一步计算违反分数的基础。在本文中，采用小批量梯度下降的方法作为分类器的模型。

本方法中的分类器主要分为两步：初始化步骤中根据符合规则的数据X_{acc}对分类器进行初始化；在每一轮迭代中对劣质数据集X_{dir}进行预测，预测值作为下一步违反分数的计算基础。

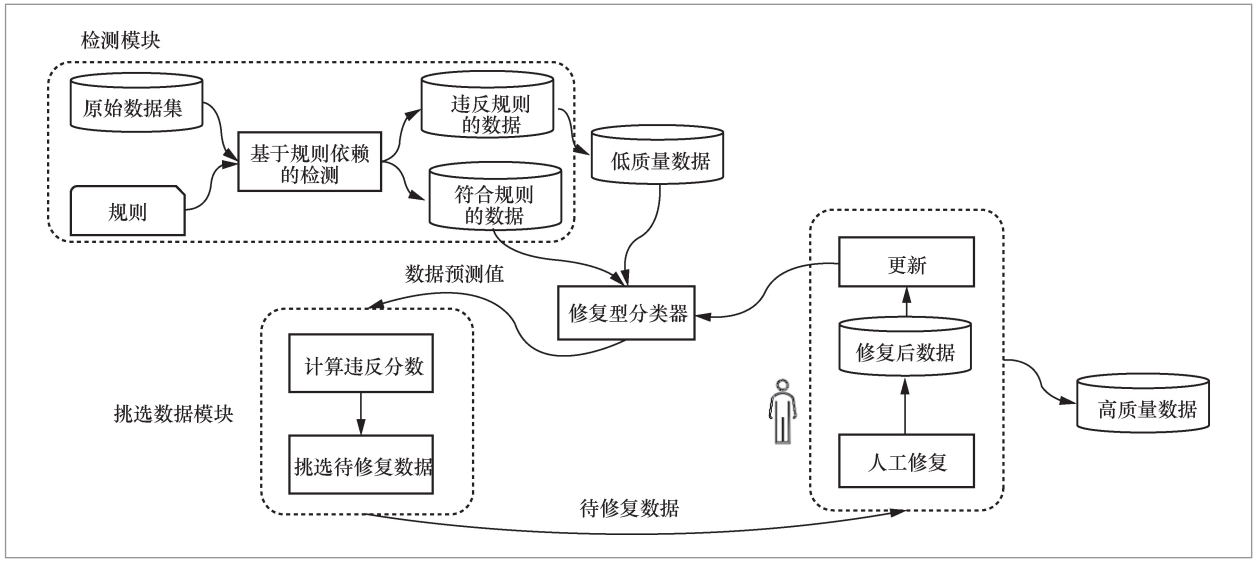

图1 人工参与的迭代式数据清洗

在对分类器的初始化步骤中，将原始数据集X分为违反规则的数据X_{vio}和符合规则的数据X_{acc}，采用符合规则的数据X_{acc}对分类器模型进行更新，之后将违反规则的数据X_{vio}作为劣质数据集X_{dir}，得到劣质数据集作为以后的待清洗数据。对于每一个劣质数据，需要计算其所有维度的综合梯度，并计算这些劣质数据的平均梯度，利用劣质数据点的平均梯度对小批量梯度下降模型的系数θ进行初始化。

在每一轮迭代过程中，采用被人工修复后的数据X_{rep}对分类器进行更新。具体到本方法中，在小批量梯度下降算法的更新中，每轮迭代采用固定数量的数据进行更新，对每轮更新列表中的数据逐行进行梯度计算，利用更新数据点的平均梯度对小批量梯度下降模型的系数θ进行更新。然后对未标注数据进行筛选，已经清洗过的数据不需要再次筛选和清洗，最后得到待清洗数据。

算法1：分类器的更新

输入：传入的系数θ、数据集x、数据集结果y、迭代最大轮数n_epochs

输出：更新后的参数θ

（1）初始化更新序列

（2）for epoch in n_epochs do

（3）for 遍历更新序列中元组 do

（4）计算在该元组时的梯度

（5）计算所有更新序列的梯度之和

（6）end

（7）通过梯度和计算平均梯度

（8）更新模型的系数=更新模型的系数-学习率 *平均梯度

（9）筛选下一轮的更新序列

（10）对筛选出的序列进行人工修复

（11）end

在如上所示算法中，首先需要在第一轮对分类器模型涉及的更新数据元组进行初始化，从中挑选出需要更新的元组。之后进行n_epochs轮迭代，每一轮中，首先计算更新序列中的元组所处位置的平均梯度，之后通过平均梯度计算更新模型的系数，并筛选下一轮的更新序列；对这些更新序列进行人工修复，之后开始下一轮的迭代。

表1 虚构数据举例

学号	姓名	年龄	专业
22780974	潘达	23	计算机
22781074	黄天	22	计算机
22780974	潘达	23	数据科学
22798774	谢凡	24	生物

3.3 挑选数据模块

为了挑选参与人工修复的错误数据，需要计算数据点对应的违反分数Score。在这一步要从数据点中找出违反分数较高的数据进行人工修复，需要利用小批量梯度下降模型中对数据的预测值。对数据违反分数的计算分为3类，分别是对单一维度内数据错误的违反分数$Score_{sin}$、对单一数据不同维度综合的违反分数$Score_{sinmul}$以及不同数据不同维度之间结合比值法的违反分数$Score_{mul}$的计算。之后，对这3种违反分数进行求和，求取违反分数$Score=Score_{sin}+Score_{sinmul}+Score_{mul}$。之后将数据按照违反分数Score从大到小排序，从中挑选违反分数较高的数据进行人工修复。

（1）对单一维度内错误数据的违反分数$Score_{sin}$进行计算

单一维度的违反分数主要考察其值与平均值的差。对于单一维度的计算，需要计算其最大值X_{max}、最小值X_{min}，以及此维度所有数据的平均值X_{avg}，违反分数为$|X-X_{avg}|/|X_{max}-X_{min}|$。例如，对于给定的数据

(5, 3, 2, 3, 2)，这5个数字的最大值为5，最小值为2，平均值为3，因此对其中的每一个数据，可以计算它的比值，5个数据的单一维度的违反分数为$\left(\frac{2}{3},\frac{0}{3},\frac{1}{3},\frac{0}{3},\frac{1}{3}\right)$。

(2)对单一数据不同维度综合的违反分数$\text{Score}_{\text{sinmul}}$进行计算

对违反分数计算方法进行判断，将得到的数据初始值与它的预测值进行比较，差距越大，就说明这个数据的违反分数越低。对于一个数据X以及它的预测值X_{predict}，可以定义这个违反分数为：

$$\frac{\text{abs}(X)+1}{\text{abs}(X_{\text{predict}}-X)+1} \quad (1)$$

(3)对不同数据不同维度之间的违反分数$\text{Score}_{\text{mul}}$进行计算

对于多维度的错误数据，通过同一元组的不同数据的比值是否超出范围来判定数据是不是错误数据。对于一个数据、它的预测值X_{predict}、在规则中存在的相关关系的数据X'以及它的预测值X'_{predict}，可以按照比值$\frac{X'_{\text{predict}}}{X_{\text{predict}}}$进行排序，如果该关系中含有多个比值需要判断，那可以于违反规则的数据的$\text{Score}_{\text{mul}}$求平均值并作为违反分数。

3.4 人工修复与更新

在数据修复步骤中，对于现有的约束$X,Z\rightarrow Y$，即数据X在满足规则Z的情况下必然存在数据Y的形式，笔者对两种修复情况进行讨论。

(1)对Y进行修复

更改右侧属性Y，使Y符合约束。但是，如果存在多条约束$X_1,Z_1\rightarrow Y$、$X_2,Z_2\rightarrow Y$涉及数据Y，且对数据Y提出了不同的要求，就会对传统的自动化清洗造成一定的困难，需要人工修复对数据Y给出正确的修复值。

(2)对于X进行修复

通常来说，本方法假定在依赖中涉及的X数据都是正确的，但如果$X,Z_1\rightarrow Y_1$、$X,Z_2\rightarrow Y_2$这两个涉及数据X的依赖都被违反，就需要考虑可能需要对X进行人工修复。

在修复阶段中，对于需要修复的数据X_1，需要进行人工修复并传回结果。在人工修复步骤中，传入的数据为劣质数据X_{dir}，X_{dir}通常是一段或者几段时间内连续的数据。人工修复既需要指出这段数据是否错误，也要返回正确的结果X_{true}。人工修复的结果传入分类器模型中，利用小批量梯度下降方法来更新模型。

4 实验分析

4.1 实验设定

(1)数据集

本节采用引风机数据集进行实验，该数据集共有2 087条数据，能够真实反映引风机组的运行变化过程。本文涉及的实验运行在Intel(R) Core(TM) i5-1035G1 CPU @ 1.00GHz 1.19 GHz的CPU和16 GB内存的PC上。

(2)对比算法

本文的实验目标是验证前文所述的人工参与的迭代式清洗算法的性能，选择ActiveClean算法作为基准算法，进行性能对比实验。

ActiveClean算法[13]是一种渐进式清理方法，其中模型是增量更新的，而不是重新训练的，其通过构建一定的挑选模型来优先清理那些可能影响结果的记录。

(3)度量标准

本文的实验任务是进行二分类问题，

即对正确数据与错误数据进行分类，正例（positive）为正常数据、反例（negative）为错误数据，可以将数据分为4类。

● TP（true positive）：实际为正常数据、算法结果为正常数据的数据。

● FP（false positive）：实际为错误数据、算法结果为正常数据的数据。

● TN（true negative）：实际为正常数据、算法结果为错误数据的数据。

● FN（false negative）：实际为错误数据、算法结果为错误数据的数据。

之后，本文通过准确率（P）与召回率（R）来衡量算法性能：

$$P=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}} \tag{2}$$

$$R=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} \tag{3}$$

在测试阶段，本文的时间轴为模型迭代的轮数，将预测的准确率与召回率作为评价指标。本文在训练中将数据分为训练集和测试集，其中前80%是训练集，后20%是测试集。每轮选取10组数据进行标注。

4.2 方法有效性计算

本文分别测试序列维数总数、错误数据总数和训练集规模对上述2种算法检测性能的影响。

（1）序列维数总数

在针对序列维数总数部分对训练方法性能的测试中，本文采用3~7个维度分别对两种方法进行测试，对比本文方法和基准方法之间的准确率差别。从图2可以看出，在序列维数较少的情况下，IDCHI算法的准确率和召回率明显高于ActiveClean算法，在序列维数增加至5维左右时，由于数据维度增高后数据中含有的信息增加，因

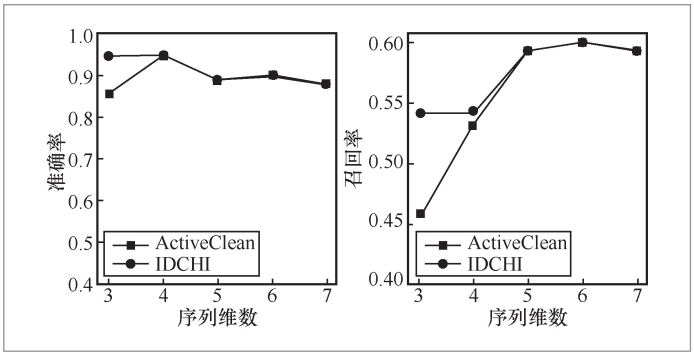

图 2 序列维数对于性能的影响

此主动学习部分针对数据信息的提取优势减小，两者的召回率和准确率接近。总体来说，IDCHI算法的准确率和召回率普遍高于ActiveClean算法，说明IDCHI算法可以高效地完成数据清洗任务。

（2）错误数据比例

在针对错误数据比例对训练方法性能的测试中，本文采用0.1、0.12、0.14、0.16、0.18、0.2 6个不同比例的错误数据的训练集分别对两种方法进行测试，对比本文方法和基准方法之间的准确率差别。从图3可以看出，随着错误数据比例的增加，两者的准确率和召回率都出现了不同程度的下滑，说明随着错误数据比例的增大，对错误数据的判断难度也加大。两者相比，IDCHI算法保持着相对较高的准确率，说明IDCHI算法能在复杂数据的情况下完成数据清洗任务。

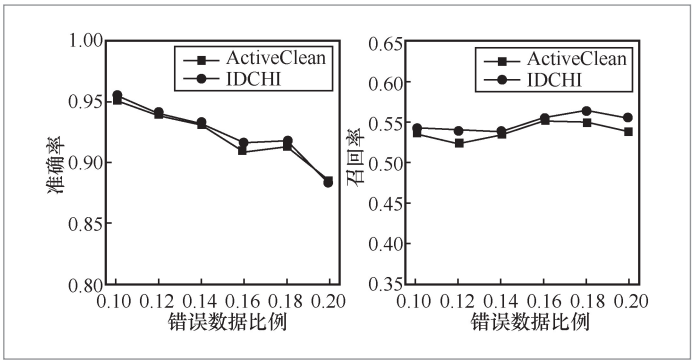

图 3 错误数据比例对于性能的影响

(3) 训练集的规模

在针对训练集规模对训练方法性能的测试中，本文采用100、150、200、250个数据组作为训练集分别对两种方法进行测试，对比本文方法和基准方法之间的准确率差别。从**图4**可以看出，在准确率和召回率两个指标上，IDCHI算法都显著好于ActiveClean算法。在训练集规模较少的情况下，IDCHI算法的召回率和准确率显著优于ActiveClean算法，但是在训练集规模增大的情况下，主动学习在少量训练集上的优势下降，ActiveClean算法的准确率和IDCHI算法接近，例如200和250规模的训练集情况下，IDCHI算法和ActiveClean算法的准确率和召回率接近。实验结果说明，本文提出的IDCHI算法可以通过较少的样本规模完成较高质量的错误数据清洗任务。

5 结束语

本文提出了一种结合了人工参与的迭代式数据清洗方法。该方法结合了人工以及规则依赖，通过检测模块迭代式地对分类器进行更新，提高了训练模型的精度。对比实验证明了该方法的准确性高于现有方法，能够在较少数据样本的情况下得到高质量的数据修复结果，其不仅能通过较少的数据样本完成数据清洗任务，而且能在复杂数据的情况下高效地对数据进行清洗。

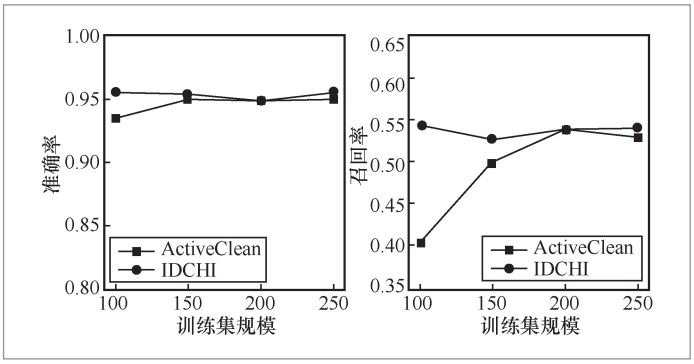

图 4 训练集规模对于性能的影响

参考文献:

[1] LIANG Z, WANG H Z, DING X O, et al. Industrial time series determinative anomaly detection based on constraint hypergraph[J]. Knowledge-Based Systems, 2021, 233: 107548.

[2] BERGMAN M, MILO T, NOVGORODOV S, et al. Query-oriented data cleaning with oracles[C]//Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data. New York: ACM Press, 2015: 1199-1214.

[3] SIDDIQUI T, KIM A, LEE J, et al. Effortless data exploration with zenvisage: an expressive and interactive visual analytics system[J]. arXiv preprint, 2016, arXiv: 1604.03583.

[4] SHOU C L, SHUKLA A. Arachnid: generalized visual data cleaning[C]// Proceedings of the 2019 International Conference on Management of Data. New York: ACM Press, 2019: 1850-1852.

[5] 范举, 陈跃国, 杜小勇. 人在回路的数据准备技术研究进展[J]. 大数据, 2019, 5(6): 1-18. FAN J, CHEN Y G, DU X Y. Progress on human-in-the-loop data preparation[J]. Big Data Research, 2019, 5(6): 1-18.

[6] QIN X D, LUO Y Y, TANG N, et al. Making data visualization more efficient and effective: a survey[J]. The VLDB Journal, 2020, 29(1): 93-117.

[7] DING X O, LIU Y D, WANG H Z, et al. SNN-AAD: active anomaly detection method for multivariate time series with

sparse neural network[C]//International Conference on Database Systems for Advanced Applications. Cham: Springer, 2023: 253-269.

[8] SATYANARAYAN A, MORITZ D, WONGSUPHASAWAT K, et al. Vega-lite: a grammar of interactive graphics[J]. IEEE Transactions on Visualization and Computer Graphics, 2017, 23(1): 341-350.

[9] HEER J, AGRAWALA M, WILLETT W. Generalized selection via interactive query relaxation[C]//Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. New York: ACM Press, 2008: 959-968.

[10] ABEDJAN Z, CHU X, DENG D, et al. Detecting data errors[J]. Proceedings of the VLDB Endowment, 2016, 9(12): 993-1004.

[11] HANRAHAN P. VizQL: a language for query, analysis and visualization[C]// Proceedings of the 2006 ACM SIGMOD International Conference on Management of Data. New York: ACM Press, 2006: 721.

[12] DING X O, SONG Y C, WANG H Z, et al. Cleanits-MEDetect: multiple errors detection for time series data in cleanits[C]// International Conference on Database Systems for Advanced Applications. Cham: Springer, 2023: 674-678.

[13] KRISHNAN S, WANG J N, WU E, et al. ActiveClean[J]. Proceedings of the VLDB Endowment, 2016, 9(12): 948-959.

[14] BEYGELZIMER A, DASGUPTA S, LANGFORD J. Importance weighted active learning[C]//Proceedings of the 26th Annual International Conference on Machine Learning. New York: ACM Press, 2009: 49-56.

[15] LE K H, PAPOTTI P. User-driven error detection for time series with events[C]//Proceedings of 2020 IEEE 36th International Conference on Data Engineering (ICDE). Piscataway: IEEE Press, 2020: 745-757.

[16] YAKOUT M, ELMAGARMID A K, NEVILLE J, et al. Guided data repair[J]. arXiv preprint, 2011, arXiv: 1103.3103.

[17] BOSTOCK M, OGIEVETSKY V, HEER J. D3 data-driven documents[J]. IEEE Transactions on Visualization and Computer Graphics, 2011, 17(12): 2301-2309.

[18] XIAO H, BIGGIO B, BROWN G, et al. Is feature selection secure against training data poisoning?[C]//Proceedings of the 32nd International Conference on International Conference on Machine Learning. New York: ACM Press, 2015: 1689-1698.

[19] CHARFI M, GRIPAY Y, PETIT J M. Spatio-temporal functional dependencies for sensor data streams[C]//International Symposium on Spatial and Temporal Databases. Cham: Springer, 2017: 182-199.

[20] LUO Y Y, CHAI C L, QIN X D, et al. VisClean[J]. Proceedings of the VLDB Endowment, 2020, 13(12): 2821-2824.

作者简介

刘一达(2000-),男,哈尔滨工业大学计算机科学与技术学院博士生,主要研究方向为数据清洗、数据依赖松弛方向。

丁小欧（1993– ），女，博士，哈尔滨工业大学计算机科学与技术学院助理教授，主要研究方向为数据清理、时间数据质量管理、时间数据挖掘、工业数据清理和多元时间序列数据中的异常行为挖掘。在数据库领域的国际会议和期刊上发表14篇学术论文。

王宏志（1978– ），男，博士，哈尔滨工业大学计算机科学与技术学院教授、博士生导师，主要研究方向为大数据管理、数据质量、图形数据管理和Web数据管理，发表论文100多篇。

杨东华（1976– ），男，哈尔滨工业大学计算机科学与技术学院副教授、博士生导师，主要研究方向为数据库、大数据管理与分析等，发表论文30余篇，SCI/EI检索30余次。主持国家自然科学基金面上项目2项、青年项目1项，以主要成员参与国家重点基础研究发展计划项目1项、国家重点研发项目1项、国家自然科学基金重点项目1项。主持中国博士后科学基金、黑龙江省博士后科学基金、黑龙江省自然科学基金等项目5项。

收稿日期： 2023-02-28

通信作者： 杨东华，yang.dh@hit.edu.cn

基金项目： 国家重点研发计划资助项目（No.2021YFB3300502）；国家自然科学基金资助项目（No.62202126，No.62232005）；中国博士后科学基金项目（No.2022M720957）；黑龙江省博士后面上资助项目（No.LBH-Z21137）

Foundation Items: The National Key Research and Development Program of China (No.2021YFB3300502), The National Natural Science Foundation of China (No.62202126, No.62232005), China Postdoctoral Science Foundation (No.2022M720957), Heilongjiang Postdoctoral Financial Assistance (No.LBH-Z21137)