本科毕业论文（设计）

人在环路的混合型错误数据清洗技术研究

RESEARCH ON HYBRID ERROR DATA CLEANING TECHNOLOGY FOR HUMAN IN THE LOOP

李昊轩

哈尔滨工业大学

2023 年 6 月

密级：公开

本科毕业论文（设计）

人在环路的混合型错误数据清洗技术研究

本　　科　　生：李昊轩

学　　　　号：1190202423

指　导　教　师：丁小欧

专　　　　业：计算机科学与技术

学　　　　院：计算学部

答　辩　日　期：2023 年 5 月 30 日

学　　　　校：哈尔滨工业大学

摘　要

海量数据产生并被应用于人们的生产生活中，数据已经成为不可或缺的资源。数据质量决定着数据被分析和使用的价值。近年来，多种识别和修复数据质量问题的方法被提出，其中人工参与数据清洗得到重视。然而现阶段人工参与的数据清洗技术存在修复算法低效、人工成本高等问题。因此，本文旨在构建高效的人工参与清洗框架提高数据质量。

本文提出了人在环路的混合型错误数据清洗框架。该框架包括对缺失点和离群值两类常见的数据质量问题的检测和修复方法。算法首先判断能否自动修复异常值。如果不满足置信度则与人类专家交互，使其参与修复的同时记录修复意见。本研究使人类专家可以持续调整数据的动态合理取值区间，以完成对异常点更精确的检测与修复。在提高数据清洗质量的同时降低人工投入成本。本文对提出的清洗框架进行不同噪声数据比例下的分类准确度实验和合理取值环境变化后的回归准确度实验。考虑多种变量变化时算法的分类准确率和预测值回归准确率的变化。实验表明在变量满足要求时本框架在提高数据质量方面有着较好的应用价值。

关键词：数据质量；数据清洗；混合型数据；缺失值；离群值

Abstract

Massive data is generated and applied in people's production and life, and data has become an indispensable resource. The quality of data determines the value of its analysis and use. In recent years, various methods for identifying and repairing data quality issues have been proposed, with human involvement in data cleaning receiving attention. However, at present, there are issues with inefficient repair algorithms and high labor costs in the data cleaning technology that involves manual participation. Therefore, this article aims to construct an efficient manual cleaning framework to improve data quality.

This article proposes a hybrid error data cleaning framework for human in the loop. The framework includes detection and repair methods for missing points and outlier, two common data quality problems. The algorithm first determines whether outlier can be automatically repaired. If the confidence level is not met, interact with human experts to participate in the repair while recording repair opinions. This study enables human experts to continuously adjust the dynamic and reasonable value range of data to achieve more accurate detection and repair of outliers. Reduce labor costs while improving data cleaning quality. This article conducts classification accuracy experiments on the proposed cleaning framework under different noise data ratios and regression accuracy experiments after reasonable environmental changes. Consider the changes in the classification accuracy and predictive value regression accuracy of the algorithm when considering changes in multiple variables. The experiment shows that this framework has good application value in improving data quality when the variables meet the requirements.

Keywords： data quality, data cleaning, mixed data, missing value, outlier

目　录

第 1 章 绪　论

1.1 课题来源及研究目的和意义

1.1.1 课题来源

国家自然科学基金项目 62232005，U1866602

1.1.2 研究目的和意义

如今，海量的数据产生并被应用于人们的生产和生活中。移动设备中的应用软件产生了社交活动、消费记录、位置信息等大量个人数据，被广泛用于制定营销策略、改善产品服务等商业和市场研究。各类工业传感器和物联网设备产生的监测数据被用于科学实验研究，改善工业技术、提高生产效率。无论是个人生活、商业活动还是科学研究，数据已经成为不可或缺的资源。

数据可靠无误才能准确地反映现实状况，有效地支持组织决策[1]。数据质量决定着数据价值，是数据得以被分析、应用的关键因素[2]。美国检察官办公室估计，14%的医疗保健支出因重复或不完整等数据质量问题而被浪费[3]。无论是人工输入数据时的拼写错误、传感器监测时由于环境干扰产生的错误，还是数据集成时的格式、结构不一致导致的错误，数据不可避免地存在着各类质量问题。影响数据质量的核心因素有数据的完整性、准确性、一致性和时效性[4]。本文旨在分析同时存在多种数据类型的混合型数据中存在的质量问题并构建有效的技术框架提高数据质量。

近年来，学术界和工业界提出了多种识别和修复数据质量问题的方法，即数据清洗。由于数据被输入方式不同、记录的事件不同，数据类型种类多样[5]。因此，数据清洗方法需要明确其适合清洗的数据类型，才能更高效的提高数据质量。例如，针对文本数据的清洗方法需要考虑字符、词义；针对图像数据的清洗方法需要考虑去噪、平滑处理、压缩和尺寸调整；针对数值数据的清洗方法需要考虑缺失值、异常值等。数据清洗是数据库界长期存在的研究问题[6]，是提高数据质量、确保分析准确性的关键技术。

在众多的数据清洗技术中，人们提出了很多基于条件函数依赖和完整性检测的清洗框架[7]。然而，部分数据的条件依赖会随着环境的改变而发生变化。例如，在金融市场领域，数据量和数据分布随着交易类型、订单数量和买卖商品的变化而变化[8]。在数据清洗过程中，与提取数据相关领域的人类专家可以

根据积累的经验和技术了解到周围环境的变化过程，进而为可能随时间流动而发生变化的数据质量评价标准提供动态的指导性意见。使人工参与到数据清洗过程对高质量、高效率的数据清洗目标有着重要意义。但与此同时，人力成本高、时间消耗大、人为错误和疏漏等问题不容忽视。因此，人工参与的数据清洗技术在以下三个方面亟待解决：

（1）人工投入提高数据清洗质量。不同领域的数据在收集方式、数据特征方面存在诸多差异。某一种特定的数据清洗算法难以完成高数据质量、高效率的清洗任务。在数据清洗流程中，人工参与可以动态调整函数依赖关系、丢弃有误导风险的数据、提供合理的数据取值范围，指导性地提高数据清洗质量。

（2）避免错误、无效的人工操作降低数据清洗质量。相较于机器执行的清洗算法，人工参与可能出现由疲劳、注意力不集中导致的人为错误和疏漏。严谨的人工参与数据清洗算法应做到利用人类专家经验的同时规避人为纰漏。

（3）降低人工投入成本。人工的反应速度和反馈效率及其有限，不但会产生经济雇佣成本，还会降低数据的时效性价值。因此，设计高效的人工参与数据清洗算法十分关键。

本文旨在解决上述人工参与的数据清洗技术问题，目标如下：

（1）分析混合型数据普遍存在的数据错误类型。设计可用的数据清洗框架的前提是明确适用的数据类型及出现的错误种类。本文将提出相关数据类型的定义和函数依赖关系。

（2）提出人在环路的数据清洗框架。实现人工参与的数据清洗框架，完成针对混合型数据的清洗目标并实验分析清洗质量。

（3）描述提高人工参与清洗效率的算法。构造合理的清洗算法，在避免错误人工操作的同时提高人机交互效率。

1.2 相关领域研究现状

目前，在人工参与的数据清洗技术方面主要分为两种类型：提供依赖关系和修复反馈验证。在提供依赖关系方面，清洗框架以人类专家主动提供的特征依赖关系作为检测、修复数据错误的依据。人类专家通常标记某（些）特征对另一（些）特征的依赖关系，给予算法启发式学习以构建约束集，也可以基于算法已生成的约束集给予更精确的约束或修正偏差的约束关系。在修复反馈验证方面，清洗框架通过自有的检测和修复错误数据算法提供其认定的潜在数据错误，并向人类专家反馈求证错误的真实性。人类专家基于常识、经验和环境变化给出反馈结果[9]。此外，人类专家也可以主动提供错误数据及其错误原因。

清洗框架在多次与人类专家的交互迭代训练中不断修正并适应环境的改变。

1.2.1 基于依赖关系的人工参与数据清洗技术研究现状

多个相关领域算法让用户以不同的角度向清洗框架提供特征依赖关系。Alexander 等[10]在 Snorkel 算法中让用户编写标签函数来表达任意启发式，使用户挑选可用的由弱监督算法生成的标签。Snorkel 相较于以往的启发式算法大幅提高了预测性能。Vijayshankar Raman 等[11]展示了集成转换和差异检测的交互式数据清理系统 Potter's Wheel，用户被允许定义可以强制执行的自定义域约束，并在有较好阅读性的界面上添加或撤销转换，逐步构建转换以清除数据。

与数据相关领域专家制定的特征依赖关系通常是可靠且适应环境变化的，但它们是不可伸缩、难以复用的。为不同输入方式、不同事件记录、不同类型的数据集提供高准确率、高覆盖率的特征依赖关系具有较高的挑战性，这些约束关系的制定在带来高成本的同时降低了数据的时效性价值，在当今社会数据爆发性增长的背景下大大降低了实用性。而机器算法自动生成的弱监管规则，虽然可以灵活的适应多样的数据特征和很大程度的覆盖数据集，但是这些规则可能非常嘈杂，降低数据清洗的准确率[12]。

1.2.2 基于修复反馈验证的人工参与数据清洗技术研究现状

基于修复反馈验证的清洗框架被较多采用于人工参与的清洗技术中。学者们以不同角度提出了多个人类专家与清洗框架的交互方式。Ziawasch 等[13]在 DataXFormer 算法着重于数据表示形式的转换，特别是在多列输入数据转换、作为其他组合的间接转换、公共数据知识库的转换等情况引入基于置信度评分的人工反馈。Xu 等[14]提出了基于知识库和群组驱动的 KATARA 算法，人类专家不必主动提供修复范围，而是被动选择待修复项。Ahmad 等[15]在 DANCE 中考虑元组之间的依赖关系，并通过丢弃可疑的元组缩减搜索空间，降低人类专家检查的成本。Mohamed 等[16]在 GDR 中将用户反馈用于训练机器学习组件，将决策理论和主动学习相结合推理修复方法。

在基于修复反馈验证方面的算法研究中，人类专家可以多种方式与清洗框架交互：在错误检测方面报告给定元组或特征的错误，在错误修复方面提供准确且适应环境的数据修复错误，在验证方面验证机器算法自动修复的结果并给予评价训练。相较于完全自动化的清洗算法，人工提供修复反馈的清洗框架牺牲不同程度成本和时效提高了数据清洗的质量。

1.2.3 国内外研究现状分析

纵观国内外研究现状可以发现，人工参与的数据清洗技术是近年来数据清洗领域的学者具有较高研究兴趣的方向，各个清洗框架针对相关适用类型的数据集和理想的人类专家前提下对数据清洗领域有着相当程度的贡献。但现有的研究工作仍有不同方面的不足与局限：

（1）人工与机器算法缺乏构建约束规则的协同性。现有的人工参与构建约束规则未能较好的与机器提取的规则交互协调。针对多样性的数据特征和庞大的数据集，机器可以迅速灵活的提取适用于特定数据集的约束，而人工参与可以相对精确的给出约束的范围。将人工给出的约束范围合理的利用在机器生成的候选规则可以提高数据的清洗质量。

（2）现有研究的人工参与验证计算修复算法较为低效。在上述多个相关研究中，人类专家只能粗略的给出数据的整体合理性，并不能对某一具体规则进行修改。同时，人类专家的反馈意见也极其受限于清洗框架提供的反馈方式，人工参与缺乏主动性和多样性。此外，部分算法未能规避人工对于某些数据元组不能提供准确反馈的缺失情况。

（3）人工参与具有高成本并降低数据时效价值。部分上述算法利用群组、众包等方式实现人工参与的数据清洗技术，这在带来高成本投入的同时延长了数据清洗的时间，降低了数据应用的时效价值，使相关算法无法被切实地应用于工业界的清洗过程也无法满足商业市场的实际清洗需求。

1.3 本文贡献和研究内容归纳

1.3.1 本文贡献

以高效利用人工参与提高数据清洗质量为目标，本文主要贡献如下：

（1）分析了国内外在有关人工参与的数据清洗技术研究现状。依据人机交互方式将相关研究分为提供依赖关系和修复反馈验证两方面。总结了现有算法对数据集适用类型和人工参与方式的差异及其贡献与不足之处。

（2）描述混合型数据集特征和数据错误类型的形式化定义。在明确本研究面向的数据类型和人工参与方式的前提下，分别提出针对不同数据错误类型的相应检测与修复算法。利用机器学习相关技术提高清洗算法的可移植性并降低人工操作成本。

（3）提出人在环路的混合型错误数据清洗框架。嵌入针对不同数据错误类

型的检测与修复算法，维护人工贡献的指导性意见并输入机器学习训练数据集实现良性循环反馈系统，构建具有实用性的清洗程序并通过实验证明提高数据清洗质量的可行性。

1.3.2 本文组织结构

本文共分为五个章节，具体结构如下：

第一章为绪论。介绍了本文的研究目的和研究意义，归纳总结了国内外在人工参与数据清洗领域的研究现状，分析现有算法的贡献及不足之处。概述了本文贡献，并提供本文组织结构。

第二章为研究基础。确定本研究面向的数据类型和人工参与方式。给出混合型数据集特征和数据错误类型的形式化定义。为后续章节提供理论依据。

第三章提出针对不同数据错误类型的相应检测与修复算法。引入机器学习相关算法并描述其与人类专家的指导性意见结合的方法，详细阐述了利用机器学习相关技术提高清洗算法的可执行性并降低人工操作成本的原因。

第四章将嵌入第三章所述的多个算法，归纳构建具有实用清洗能力的人在环路的混合型错误数据清洗框架。详细描述算法的整体清洗流程和人机交互的过程。

第五章记录了实验结果，分析了程序清洗的准确率和人工、机器花费成本，分析并验证本文所述清洗框架的可行性。

最后在结论中总结了本文的研究内容，展望需要进一步开展的研究工作。

1.4 本章小结

本章介绍了本文的研究目的及意义，分析相关领域现有研究算法的贡献及不足之处，概述本文贡献，提供本文组织结构。

第 2 章 研究基础

2.1 引言

由于不同领域的数据在提取方式、数据特征、应用场景方面都存在较大差异，在构建数据清洗框架时必须明确面向的数据类型和清洗方向。本章将以形式化定义的方式规范清洗框架面向的数据类型，明确数据质量评价标准，确定人类专家的专业要求及其与清洗算法交互的方式。为后续章节算法的提出奠定理论基础。

2.2 规范数据类型

从数据的生成角度分类，当今大数据的两种主要类型是结构化数据和非结构化数据。虽然它们都用于描述和组织信息，但是二者在数据管理和分析中具有不同的特征和用途。

结构化数据是指可以使用二维表结构表示和存储的数据，具有易于输入、存储、查询和分析的特点[1]。结构化数据中每个数据字段的类型、长度和格式都是事先定义好的，例如整数、字符串、日期等。结构化数据以表格的形式进行组织，其中每一行代表一个数据记录，每一列代表一个数据字段。

非结构化数据是不以明确的模式存储和组织的数据，它可以是文本文档、电子邮件、音频、视频、图像等。非结构化数据不遵循固定的模式或格式，其结构和组织方式可能会因数据类型和来源的不同而有所变化。

由于当今市场对高质量数据的需求主要为结构化数据，当前该领域研究也主要着重清洗结构化数据，因此本文将结构化数据作为清洗算法面向的数据类型。

定义 1（结构化数据） 结构化数据是以明确定义的格式存储和组织的数据。

通常结构化数据集都描述了事物的多个方面，本文将描述事物的每个方面称为数据集的一个特征。对于包含字符型和数据型数据的数据集，可以将每一列视为数据的一个特征。本文将描述事物所有方面的特征总和称为特征集合。

定义 2（特征） 特征 f 是数据可供识别的特殊的属性或标记。

定义 3（特征集合） 特征集合 F 是数据的所有特征 f 组成的集合。

本文明确混合型数据有着明确定义的存储格式和组织方式，混合型数据归属于结构化数据。而进一步规范，混合型数据是包含多个不同特征的数据集合。

定义 4（混合型数据） 混合型数据 D 是包含多种不同特征 f 的结构化数据。本文将混合型数据 D 的特征集合记为 F_D

$$F_D = \left\{ f_1, f_2, \cdots f_n \mid n > 1, \forall i, j : i \neq j \Rightarrow f_i \neq f_j \right\}$$

例 1： 如表 2-1 所示为某公司工资数据集 D_Salary。该数据集共有 6 个特征，其特征集合 $F_{D\_Salary} = \left\{ f_1 = FirstName, f_2 = LastName, \cdots f_6 = City \right\}$。我们可以看到，$f_1 = FirstName$、$f_2 = LastName$、$f_6 = City$ 为表意不同的字符特征，$f_3 = EmployeeID$、$f_4 = Salary$、$f_5 = ZipCode$ 为表意不同的整数特征。因此，数据集 D_Salary 为包含多种不同特征的结构化数据，即混合型数据。

表 2-1　某公司员工信息数据集

LNR	FirstName	LastName	EmployeeID	Salary	ZipCode	City
1	James	Brown	107	70	50210	Miami
2	Craig	Rosberg	107	50	50210	Miami
3	James	Brown	308	70	21100	Atlanta
4	Monica	Johnson	308	60	21100	Houston
5	James	Brown	401	80	65300	NY
6	Monica	Johnson	401	100	65300	NY
7	Craig	Rosberg	401	80	65300	Boston
8	Mark	Douglas	401	130	65300	NY

2.3 明确数据质量评价标准

数据质量是反映数据能否被分析、应用的关键因素，也为数据清洗奠定了基础和方向，反映了清洗算法的优劣。相关研究领域的学者从不同角度给出了对数据质量的定义。文献[17]认为数据质量是一组固有特性满足要求的程度。Naumann 等[18]认为数据质量是数据适合被使用的程度。为了使数据质量可以被定量的评价以适应清洗算法的构建与改进，本文规定混合型数据的数据质量定义如下：

定义 5（数据质量） 混合型数据的数据质量是特征集合 F 满足一组固有特性要求的程度。

为了形式化的描述本文涉及的名词定义和算法解释，我们在此定义混合型数据中的某一数据点。由于混合型数据普遍为二维表结构表示的数据，我们以数据点所处的行号 LNR，特征 f 和取值 x 定义数据点。

定义 6（数据点） 数据点 $x_i^{f_j}$ 表示在第 i 行，特征为 f_j，取值为 x 的数据点。

由于不同领域对数据的需求不同，数据的单一特性难以全面准确地评价数据质量。本文查找了相关领域研究学者对数据质量的评价标准。McGilvray[19]等认为数据质量是满足数据完整性、准确性、服务性和可用性的程度。Laudon[20]等从准确性、完整性、一致性、及时性的角度评价数据质量。在众多相关文献中，准确性和完整性是被提及最多的衡量数据质量的特性指标。结合其适合对混合型数据质量的评价，因此本文采用完整性和准确性作为数据质量评价标准。

为了给出完整性和准确性的形式化定义，本文首先描述混合型数据中常见的质量问题。

收集数据时，由于数据采集过程中的遗漏、数据源主观意愿不回答特定问题、技术故障或设备错误、数据转换或处理过程中的错误等原因，常常会出现数据缺失的情况。缺失值常常被认为是最常见的数据质量问题，也是提高混合型数据质量必须要处理的质量问题。我们认为缺失值仍为一个数据点 $x_i^{f_j}$，其仍有对应的行号 i 和特征 f，但是其值 x 为空缺 $NULL$。

定义 7（缺失值） 缺失值是有对应的行号 i 和特征 f_j，其值 x 为空缺 $NULL$ 的数据点。$m = x_i^{f_j}, i \neq NULL, f_j \neq NULL, x = NULL$。缺失值集合

$$X_{Null} = \left\{ x_i^{f_j} \mid i \neq NULL, f_j \neq NULL, x = NULL \right\}$$

例 2：如表 2-2 所示为存在数据质量问题的某公司员工信息数据集。在该表行号为 5，特征 $f = LastName$ 的数据点取值空缺。根据上述定义，我们认为这是一个缺失值。$m = x_5^{f=LastName}, x = NULL$

表 2-2　存在数据质量问题的某公司员工信息数据集

LNR	FirstName	LastName	EmployeeID	Salary	ZipCode	City
1	James	Brown	107	70	50210	Miami
2	Craig	Rosberg	107	50	50210	Miami
3	James	Brown	308	70	21100	Atlanta
4	Monica	Johnson	308	60	21100	Houston
5	James		401	80	65300	NY
6	Monica	Johnson	401	100	65300	NY
7	Craig	Rosberg	401	800000	65300	Boston
8	Mark	Douglas	401	130	65300	NY

在未被清洗的数据集中可能会出现与其周围观测值明显不同的异常值。它们可能是由于测量误差、数据录入错误或周围环境波动等因素引起的。本文将

这样的数据点称为离群值。离群值的存在可能会对数据统计、数据分析或机器学习算法等应用产生影响较大的干扰。并扭曲数据的统计性质和模型的准确性。与缺失值的定义模式类似，我们认为离群值也为一个数据点 $x_i^{f_j}$，其有对应的行号 i 和特征 f，但是其数据值 x 与同特征的其他近邻行数据值有明显差异。

定义 8（离群值） 离群值是有对应的行号 i 和特征 f_j 的数据点，其取值 x 相较于同特征 f_j 近邻行平均取值 μ 以及常数 k 和标准差 σ 满足：$x > \mu + k\sigma$ 或 $x < \mu - k\sigma$。$o = x_i^{f_j}, i \neq NULL, f_j \neq NULL, |x - \mu| > k\sigma$。离群值集合

$$X_{Outlier} = \left\{ x_i^{f_j} \mid i \neq NULL, f_j \neq NULL, |x - \mu| > k\sigma \right\}$$

例 3： 在上述表 2 中，在行号为 7，特征 $f = Salary$ 的数据点取值与其相邻多行同特征的数据值取值有明显差异，并且在 $\mu = 100069$，$k = 2$，$\sigma = 282657$ 的前提下满足 $x = 800000 > 665384$ ：。因此，我们认为该值为离群值。$o = x_7^{f=Salary}, |800000 - 100069| > 565314$

在描述了混合型数据集中常见的数据质量错误后，我们可以形式化的定义数据质量评价标准完整性和准确性。

完整性描述了数据集中数据信息的缺失程度，是评价数据质量的重要标准。较高的完整性表示数据在生成和转移中被较少的丢失或损坏，更高比例的必要信息得到了全面的记录和保存。数据具有良好的完整性为向工业界和商业界提供高质量数据实现了重要保障，让数据拥有更高的利用价值。本文将完整性定义为在混合型数据中，非缺失值占所有数据点的比例。并为了后续算法的构建与修正，将某特征的完整性定义为该特征的非缺失值占该特征所有数据点的比例。

定义 9（完整性） 混合型数据 D 的完整性 *Integrity* 表示非缺失值集合 $X - X_{Null}$ 占所有数据点 X 的比例。$Integrity_D = \dfrac{X - X_{Null}}{X}$，$X = \left\{ x_i^{f_j} \mid i \neq NULL, f_j \neq NULL \right\}$，$X_{Null} = \left\{ x_i^{f_j} \mid i \neq NULL, f_j \neq NULL, x = NULL \right\}$

准确性描述了数据与事实相符的程度。数据具有良好的准确性对于企业利用数据进行有效的决策制定、可靠的分析和良好的业务运营至关重要。在数据从生成到传输、应用等流程中，数据的准确性常常被破坏。因此，提高脏数据集的准确性是评价清洗算法的关键标准。在混合型数据中，本文将准确性定义为非离群值占所有数据点的比例。同样为了后续算法的构建与修正，将某特征的准确性定义为该特征的非缺失值占该特征所有数据点的比例。

定义 10（准确性） 混合型数据的准确性 *Accuracy* 表示非离群值集合 $X-X_{Outlier}$ 占所有数据点 X 的比例。 $Accuracy_D=\frac{X-X_{Outlier}}{X}$ 。 $X=\left\{x_i^{f_j}\mid i\neq NULL, f_j\neq NULL\right\}$， $X_{Outlier}=\left\{x_i^{f_j}\mid i\neq NULL, f_j\neq NULL, \left|x-\mu\right|>k\sigma\right\}$

2.4 确定人员需求与参与方式

在人工参与的数据清洗框架研究领域，多个相关算法让用户以不同角度与机器交互完成清洗任务。本文在 1.2 节已提到，部分算法利用群组、众包等方式完成人机交互，较多的人员参与在带来高成本的同时延长了数据清洗的时间，降低了数据应用的时效价值。另一些相关算法由于其特定的人机交互方式让人类专家只能粗略的给出数据的整体合理性，并不能对某一具体规则修改，这样的人工参与方式缺乏主动性和多样性，造成清理成本的浪费。

在本文提出的人在环路的混合型错误数据清洗框架中，我们以人员选择修改方式多样性、提供修复数据主动性，同时降低人员成本和时间花销为目的实现人机交互。因此，本文将在人员需求和参与方式两方面阐述人机交互过程。

2.4.1 人员需求

人员需求为与清洗数据领域相关的单一人类专家。因此需要一位完全了解待清洗数据属性与变化的人类专家作为人机交互对象。在清洗过程中，我们需要人员主动指出算法认为存在潜在错误的数据值，这需要人员基本了解并给出某缺失值的取值（范围）、某疑似离群值的取值（范围），或是任何违反完整性或准确性的数据值是否可以被丢弃。同时，考虑到对于个别数据值，人员无法给出准确回复，算法允许人员忽略该此询问或删掉该行数据。除此之外，人员还应该根据其长期积累的经验了解到数据可能由于环境因素的变化，其合理取值范围发生了改变，并由此向算法反馈提高算法清洗的准确性。

满足此类要求的人类专家并不罕见，他们分布广泛。同时，由于只需要一位人类专家，该清洗框架不需要高经济成本的投入，也提高了数据应用的实际价值。例如，市场调研员了解消费者行为数据：市场调研员通常负责收集和分析消费者行为数据，以了解他们的购买习惯、偏好和趋势；医院的流行病学家了解疾病流行数据：流行病学家研究疾病在人群中的传播和流行情况，以了解疾病的传播模式和风险因素；政府经济学家了解国家经济数据：政府经济学家负责收集和分析国家的经济数据，以了解经济增长趋势、劳动力市场状况和通

货膨胀压力。

2.4.2 人员参与方式

对于任一次清洗，满足要求的人类专家向算法指出开始清洗的特征名称和清洗的数据质量问题类型，例如离群值和缺失值。算法在遇到疑似不满足完整性或准确性的数据值时向人类专家报告该情况，报告中包括该数据值所在行的整行数据。人类专家选择修复方式，包括给定修复值、删除异常数据值所在行、忽略该次错误。当算法完成指定特征的所有行数据的清洗后，人类专家可以选择针对其他特征的清洗或退出清洗程序。

通过上述人机交互方式，我们利用了人类专家的选择修改方式多样性和提供修复数据主动性。也通过忽略错误、删除错误数据值所在行的方式给予人类专家一定的容错性，最大程度的发挥了人员参与数据清洗过程的作用。

2.5 本章小结

本章详细叙述了本文的研究基础。本章规范了数据类型，形式化的定义了普遍出现的数据质量问题和相关领域常用的数据质量评价标准。详细阐述了算法对人员的需求和人机交互方式。为后续章节的算法提出规范了符号表示并奠定了理论基础。

第 3 章 数据错误检测与修复算法

3.1 引言

本章将介绍针对数据质量问题的检测与修复算法。数据错误检测与修复算法目的是着重提高数据质量评价标准中的完整性和准确性。由第二章定义可知，混合型数据的完整性表示为非缺失值集合 $X-X_{Null}$ 占所有数据点 X 的比例，即 $Integrity_D=\frac{X-X_{Null}}{X}$。混合型数据的准确性表示为非离群值集合 $X-X_{Outlier}$ 占所有数据点 X 的比例，即 $Accuracy_D=\frac{X-X_{Outlier}}{X}$。因此，本章将在 3.1 节介绍缺失值检测与修复算法，在 3.2 节介绍离群值检测与修复算法。

3.2 缺失值检测与修复算法

表 3-1　缺失点操作算法

算法 1	**MissingDataOperation**(X , F_X)
Input:	混合型数据集 X ， X 的特征集合 F_X
Output:	无
1	rows = *X*.getRow　// 获取行数
2	columns = *X*.getColumn　// 获取列数
3	*X*.isnull = *X*.missingDataDetection //缺失值检测，返回 true/false 二维数组
4	**for** row **in** rows:
5	**for** column **in** columns:
6	**if** *X*.isnull[row,column] 存在缺失值：
7	confidence = AutomaticCorrectData(*X*) // 判断能否自动修复
8	**if** !confidence :　// 如果不满足自动修复条件
9	MissingDataReport(*X*.row,column) //请求用户修复并记录
10	End

缺失值检测与修复算法的输入数据是混合型数据集 X 和它的特征集合 F_X。该算法并无输出，算法结束即完成了对某一特征 $f\in F_X$ 缺失值的检测与修复。但是算法中包含的子算法会在每次完成对一缺失值 $x_i^{f_j}\in X_{Null}$ 的修复后更新修复后的数据集，同时更新记录用户修改数据方式和修复值的数据集。缺失值检

测与修复算法如上述算法 1 所示，该算法主要包括以下 4 个步骤：

Step1: 读取数据集 X 及其特征集合 F_X，获取行数 rows 和列数 columns

Step2: 如图 3.1 所示，检测 X 中是否存在缺失值。假设 X 的行数为 m，列数为 n。则该步骤会返回一个 m 行 n 列的二维数组 X.isnull。对于任一 i 行 j 列的数据值，如果是缺失值则 $X.isnull(i,j)=true$。如果不是缺失值则 $X.isnull(i,j)=false$。该步骤具体算法在 3.1.1 节描述。

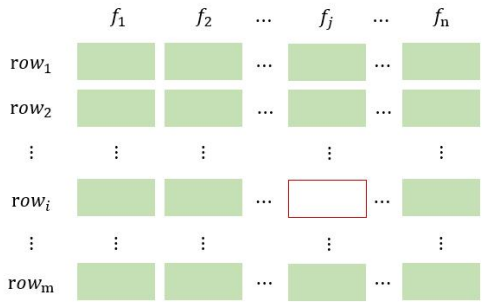

图 3-1　检测 X 中是否存在缺失值

Step3: 如图 4.2 所示，对任一行 $i \in m$，列 $j \in n$，如果 $X.isnull(i,j)=false$ 则忽略。如果 $X.isnull(i,j)=true$ 则判断能否满足自动修复置信度。如果满足则自动修复并更新修复后的数据集 X_repair。该步骤具体算法在 3.1.2 节描述。

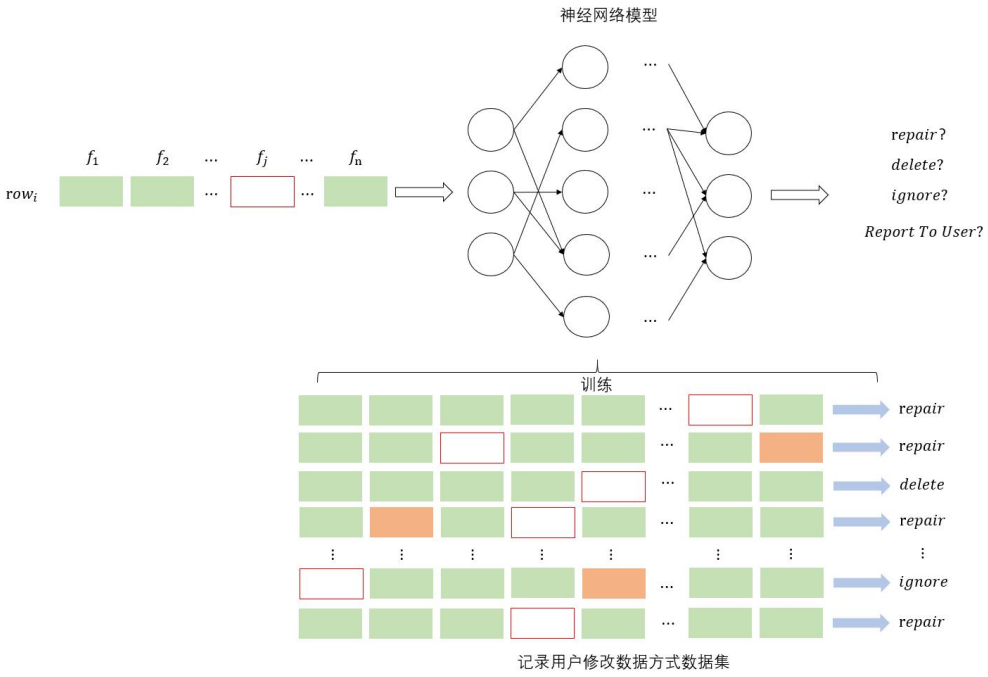

图 3-2　判断能够满足缺失值自动修改置信度

Step4: 如图 3.3 所示，如果在上一步骤中不满足自动修复置信度，则请求用户修复。在更新修复后的数据集 X_repair 的同时更新记录用户修改数据方式和修复值的数据集，该数据集被用作机器学习的训练数据集，帮助算法进行修复分类预测和修复值预测，本文将在第四章具体说明。该步骤具体算法在 3.1.3 节描述。

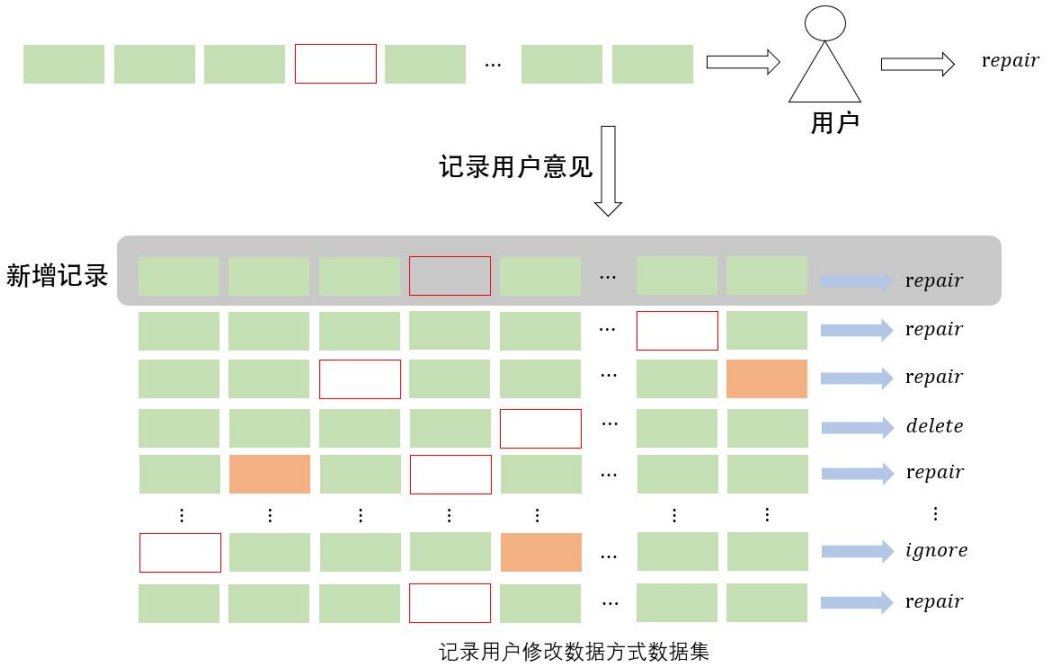

图 3-3　请求用户修复并记录修改意见

3.2.1 缺失值检测

该步骤会返回一个与数据集 X 相同行列的二维数组。对于任一 i 行 j 列的数据值，如果是缺失值则 $X.isnull(i,j)=true$。如果不是缺失值则 $X.isnull(i,j)=false$。

例 3.1 假设 X 是一个 4 行 3 列的二维数据集，X 的特征集合 $F_X=\{Name, EmployeeID, Salary\}$。如下展示了 X 和 $X.isnull$。

$$X=\begin{bmatrix} Brown & 107 & 70 \\ Rosberg & 308 & 50 \\ Johnson & & 130 \\ Douglas & 401 & 800000 \end{bmatrix} \qquad X.isnull=\begin{bmatrix} False & False & False \\ False & False & False \\ False & True & False \\ False & False & False \end{bmatrix}$$

图 3-4　缺失值检测示例

3.2.2 缺失值自动修复置信度判断

表 3-2　自动检测错误算法

算法 2	**AutomaticCorrectData** (X , row, column，L_classify)
Input:	数据集 X ， 缺失值所在行列，记录用户修复方式数据集
Output:	是否满足自动修复置信度
1	RecordedData = L_classify[:, :-1]　// 获取记录修复方式数据
2	RecordedLabel = L_classify[:, -1]　// 获取记录修复方式标签
3	nn = Model(input = L0, outputs = Ln(…(L1)))　// 构建神经网络模型
4	one_hot_labels = to_categorical(RecordedLabel) //将标签转换为独热编码
5	nn.compile(loss,optimizer) // 训练模型
6	nn.fit(RecordedData, one_hot_labels)
7	predictResult = nn.predict(X, row, column) // 预测数据
8	c_delete, c_repair, c_ignore = predictResult　// 获取预测结果
9	**if** confidence_delete > confidence:　// 删除置信度满足要求
10	X.delete_row　// 删除异常数据所在行
11	**return** true
12	**elseif** confidence_repair > confidence:　// 修复置信度满足要求
13	predictRepairValue(X, row, column) //回归模型得到修复值并修复数据
14	**return** true
15	**elseif** confidence_ignore > confidence:　// 忽略置信度满足要求
16	**return** true
17	**else**
18	**return** false

由算法 2 所示，该步骤通过构建神经网络判断能否满足自动修复置信度。算法首先读取记录用户修复方式的数据集，将其分类为数据和标签并用于训练神经网络模型。神经网络模型需构建合理的输入层、隐藏层和输出层。其中隐藏层的数量要根据数据的复杂程度构建，使其对待预测数据有着较好的分类效果。将构建好的神经网络模型进行训练。完成训练后输入缺失值及其所在的行、列和特征完成预测。预测结果为满足各分类置信度的比例。将该比例与各个置信度进行比较。如果预测结果满足删除置信度，则算法自动删除缺失值所在行。如果预测结果满足修复置信度，则调用在下文算法 3 中讲述的回归模型，构建回归模型并进行训练，再将训练好的模型进行本次异常值的预测后得到修复值并修复数据。如果预测结果满足忽略置信度则忽略该次异常。如果上述三种置信度均不满足则返回 false 请求用户修复。

表 3-3　预测修复值算法

算法 3	**predictRepairValue**(X, row, column, L_value)
Input:	数据集 X，　缺失值所在行列，记录用户修复取值数据集
Output:	无
1	RecordedData = L_value[:, :-1]　// 获取记录修复取值数据
2	RecordedLabel = L_value[:, -1]　// 获取记录修复取值标签
3	knn = KNeighborsRegressor()　// 构建 k 近邻回归模型
4	hnn.fit(RecordedData, RecordedLabel)　// 训练模型
5	PredictedValue = knn.predict(X, row, column)　// 预测修复值
6	X.changeValue　// 将预测的修复值写入文件

在算法 3 中，通过构建 K 近邻回归模型获取修复取值。算法首先读取记录用户修复取值的数据集，该数据集记录了用户提供的历次修复取值信息，将其分类为数据和标签并训练 K 近邻回归模型。近邻回归模型的 K 值选取也应根据数据集的数据量选取合适的 K 值。K 值对数据的适合程度将直接影响到算法预测值的准确率。将构建好的 K 近邻回归模型进行训练，并将完成训练后输入缺失值及其所在的行、列和特征完成预测。并将预测的修复值写入文件，即更新 X 为 X.repair。

例 3.2 仍以在例 3.1 中提出的 4 行 3 列的二维数据集 X 为例。通过 Step2 可知 $X.isnull(3,2)=true$。算法会判断能否满足自动修复置信度。假设在算法 2 的神经网络模型判断中 confidence_repair =75 > confidence=70，即修复置信度满足要求，并且删除置信度 confidence_delete =10 < confidence=70 和忽略置信度 confidence_ignore =15 < confidence=70 均不满足要求。由于修复置信度满足要求，则并不需要询问用户意见，算法 2 调用算法 3，算法 3 通过 k 近邻回归模型预测取值为 280，则更新 X 为 X.repair。

$$X=\begin{bmatrix} Brown & 107 & 70 \\ Rosberg & 308 & 50 \\ Johnson & & 130 \\ Douglas & 401 & 800000 \end{bmatrix} \qquad X.repair=\begin{bmatrix} Brown & 107 & 70 \\ Rosberg & 308 & 50 \\ Johnson & 280 & 130 \\ Douglas & 401 & 800000 \end{bmatrix}$$

图 3-5　缺失值修复示例

3.2.3 人机交互修复缺失值

表 3-4　用户操作算法

算法 4	**UserOperation**(*X*, row, column, L_classify, L_value)
Input:	数据集 *X*， 缺失值所在行列, 记录用户修复方式和修复值的数据集
Output:	无
1	missing_operation = input()　// 请求用户提供修复方式
2	**if** missing_operation == “1”:　// 如果用户要求删除数据
3	delete_missing_data(*X*, row, column)　// 删除缺失值所在行
4	RecordUserClassify (L_classify, delete, row, column) //记录用户操作
5	**elseif** missing_operation == “2”:　// 如果用户要求修改数据
6	value = input()　// 请求用户提供修复值
7	repair_missing_data(*X*, row, column, value)　// 替换缺失值为修复值
8	RecordUserClassify(L_classify, repair, row, column) //记录用户操作
9	RecordUserValue (L_value, repair, row, column) //记录用户取值
10	**else** missing_operation == “3”:　// 如果用户要求忽略缺失
11	RecordUserClassify(L_classify, ignore, row, column) //记录用户操作

由算法 4 所示，当自动修复置信度不满足要求时，则请求用户修复。首先请求用户提供修复方式，如果用户要求删除数据，则删除缺失值所在行。如果用户要求修改数据，则要求用户提供修复值，并将缺失值替换为修复值。如果用户要求忽略本次缺失值，则忽略错误。无论用户提供的修复方式是什么，都将更新入记录用户修复方式数据集里，为之后的自动修复提供更准确、更适应环境变化的修复决策。

例 3.3 仍以在例 3.1 中提出的 4 行 3 列的二维数据集 X 为例。通过 Step2 可知 $X.isnull(3,2)=true$。算法会判断能否满足自动修复置信度。假设在算法 2 的神经网络模型判断中 confidence_repair =55 < confidence=70，即修复置信度不满足要求。则算法 2 调用算法 4，算法 4 通过询问用户意见，得知用户要求修复缺失值，并给出修复值 205，则更新数据集 X 为 $X.repair$，并更新记录用户修复方式数据集和记录用户修复取值数据集。

$$X=\begin{bmatrix} Brown & 107 & 70 \\ Rosberg & 308 & 50 \\ Johnson & & 130 \\ Douglas & 401 & 800000 \end{bmatrix} \qquad X.repair=\begin{bmatrix} Brown & 107 & 70 \\ Rosberg & 308 & 50 \\ Johnson & 205 & 130 \\ Douglas & 401 & 800000 \end{bmatrix}$$

$$L\_classify=\begin{bmatrix} \vdots & \vdots & \vdots & \vdots \\ MissError & 3 & 2 & repair \end{bmatrix} \qquad L\_value=\begin{bmatrix} \vdots & \vdots & \vdots & \vdots \\ MissError & 3 & 2 & 205 \end{bmatrix}$$

图 3-6　记录用户修复意见示例

3.3 离群值检测与修复算法

表 3-5　离群值操作算法

算法 5	**OutlierDataOperation**(X , F_X)
Input:	混合型数据集 X ， X 的特征集合 F_X
Output:	无
1	rows = *X*.getRow　// 获取行数
2	columns = *X*.getColumn　// 获取列数
3	**while** True:
4	column_name = input()　// 获取要检查的列名
5	row = OutlierDataDetection(*X*, column_name, rows)　// 检查离群点
6	confidence = AutomaticCorrectData(*X*, row, column)　//判断自动修复
7	**if** !confidence :　// 如果不满足自动修复条件
8	OutlierDataReport(*X*.row,column) //请求用户修复并记录
9	**if** ! continue_check = input()　// 是否检查其他离群点
10	**break**
11	End

与缺失值检测类似，离群点检测与修复算法的输入数据也是混合型数据集 X 和它的特征集合 F_X 。该算法无输出，该算法根据用户输入的特征 f ，检测特征 f 相应列的离群点。完成对某一特征离群点的检测后会询问用户是否检测其他列。算法中包含的子算法会在每次完成对一离群值 $x_i^{f_j} \in X_{Outlier}$ 的修复后更新修复后的数据集，同时更新记录用户修改数据方式和修复值的数据集。离群值检测与修复算法如上述算法 5 所示，该算法主要包括以下 4 个步骤：

Step1: 读取数据集 X 及其特征集合 F_X ，获取行数 rows 和列数 columns。

Step2:检测 X 中特征为 f 的列是否存在离群值。本文使用箱线图检查离群点。获取箱体图特征，标记相应特征值并划定非异常范围。将非异常范围之外的数据标记为离群点。该步骤具体算法在 3.2.1 节描述。

Step3:由 Step2 获取离群值所在行号，首先判断是否满足自动修复置信度。如果满足要求，自动修复离群值并更新修复后的 X_repair 。如果不满足要求，则请求用户修复。该步骤具体算法在 3.2.2 节描述。

Step4: 当 Step3 返回不满足自动修复置信度时，请求用户修复。向用户展示离群点所属特征和所在行的所有数据值，得到用户的修复反馈后更新数据集 X_repair ，同时更新记录用户修改数据方式和修复值的数据集。该数据集与

3.1 节所述修复缺失值的数据集为同一数据集，用作机器学习的训练数据集，帮助算法进行修复分类预测和修复值预测。该步骤具体算法在 3.2.3 节描述。

3.3.1 离群值检测

表 3-6　离群值检测算法

算法 6	**OutlierDataDetection**(*X*, column_name, rows)
Input:	数据集 *X*，待检测的列名，数据集的行数
Output:	离群点所在行号
1	percentile=np.percentile(column,(25,50,75),interpolation)//箱体图特征
2	Q1 = percentile[0]　// 上四分位数
3	Q2 = percentile[2]　// 下四分位数
4	IQR = Q3 – Q1　// 四分位距
5	ulim = Q3 + 1.5 * IQR　// 上限，非异常范围内的最大值
6	llim = Q1 – 1.5 * IQR　// 下限，非异常范围内的最小值
7	**for** row **in** rows:
8	**if** check_column.iloc[row] < llim **or** check_column.iloc[row] > ulim:
9	**return** row

在算法 6 中，使用箱线图检查数据集 *X* 的某特征列是否存在离群点。获取箱线图上四分位数 Q1、中位数 Q2、下四分位数 Q3 的特征值，并确定四分位距 IQR。本文规定箱线图的上界，即非异常范围内的最大值为 Q3 + 1.5 * IQR。箱线图的下界，即非异常范围内的最小值为 Q1 – 1.5 * IQR。检查 *X* 中选定特征列的所有行，若某行取值属于异常范围则返回该行行号。

例 3.4 假设数据集 *X* 是一个 n 行 3 列的二维数组。*X* 的特征集合 $F_X=\{Name, EmployeeID, Salary\}$。假设用户选择检查特征 $f=Salary$ 的离群值。由算法 6 求得上四分位数 Q1=130，下四分位数 Q3=40。四分位距 IQR=70。经计算得知 800 属于异常范围，因此算法返回 $\{Douglas, 401, 800\}$ 所在行。

$$X=\begin{bmatrix} \vdots & \vdots & \vdots \\ Brown & 107 & 70 \\ Rosberg & 308 & 50 \\ Douglas & 401 & 800 \\ \vdots & \vdots & \vdots \end{bmatrix}$$

图 3-7　离群值数据示例

3.3.2 离群值自动修复置信度判断

与 3.1.2 节所述算法 2 相似，通过构建神经网络判断能否满足自动修复置信度。因记录用户修复方式数据集中其中一个属性为异常原因，即异常数据为缺失值或离群点。因此神经网络的预测模型依然可以调用与 3.1 节所述相同的数据集进行分类训练并预测。

例 3.5 以在例 3.4 中提出的 n 行 3 列的二维数据集 X 为例。假设在算法 2 的神经网络模型判断中 confidence_repair =85 > confidence=70，即自动修复置信度满足要求。则算法 2 调用算法 3，算法 3 通过 k 近邻回归模型预测取值为 67，则更新 X 为 X.repair。

$$X=\begin{bmatrix} \vdots & \vdots & \vdots \\ Brown & 107 & 70 \\ Rosberg & 308 & 50 \\ Douglas & 401 & 800 \\ \vdots & \vdots & \vdots \end{bmatrix} \qquad X.repair=\begin{bmatrix} \vdots & \vdots & \vdots \\ Brown & 107 & 70 \\ Rosberg & 308 & 50 \\ Douglas & 401 & 67 \\ \vdots & \vdots & \vdots \end{bmatrix}$$

图 3-8　离群值自动修复示例

3.3.3 人机交互修复离群值

当自动修复置信度不满足要求时，则请求用户修复。与 3.1.3 节所述算法 4 类似，根据用户提供的修复方式完成相应的操作。将用户提供的修复方式和可能提供的修复值更新入记录用户修复方式数据集和记录用户修复取值数据集，为之后的自动修复提供更准确、更适应环境变化的修复决策。

例 3.6 以在例 3.4 中提出的 n 行 3 列的二维数据集 X 为例。假设在算法 2 的神经网络模型判断中 confidence_repair =55 < confidence=70，confidence_delete =20 < confidence=70，confidence_ignore =25 < confidence=70，即修复置信度、删除置信度和忽略置信度均不满足要求。则算法不能对该次异常值进行自动修改。则算法 2 调用算法 4，算法 4 通过询问用户意见，得知用户要求修复缺失值，并给出修复值 80，则更新数据集 X 为 $X.repair$，并更新记录用户修复方式数据集和记录用户修复取值数据集。

$$X=\begin{bmatrix} \vdots & \vdots & \vdots \\ Brown & 107 & 70 \\ Rosberg & 308 & 50 \\ Douglas & 401 & 800 \\ \vdots & \vdots & \vdots \end{bmatrix} \qquad X.repair=\begin{bmatrix} \vdots & \vdots & \vdots \\ Brown & 107 & 70 \\ Rosberg & 308 & 50 \\ Douglas & 401 & 80 \\ \vdots & \vdots & \vdots \end{bmatrix}$$

图 3-9　用户提供修复意见示例

3.4 本章小结

本章分别介绍了出现缺失值和离群点数据质量问题时完成的检测与清洗算法。二者在检测方法时有所差异。在检测某数据值符合异常数据时，均调用自动修复方法，通过构建神经网络模型训练用户记录修复方式数据集判断是否满足自动修复置信度，并在满足修改取值置信度的条件下构建 K 近邻回归模型训练用户记录修复取值数据集得到修复值，完成本次修复。若不满足则请求用户修复，将用户提供的修复方法和修复取值记录相应数据集为之后的自动修复提供更准确、更适应环境变化的修复决策。

第 4 章 清洗框架与人机交互

4.1 引言

本章将介绍人在环路的混合型数据清洗整体算法框架。该框架包括了第三章提出的针对两类数据质量问题的检测与修复算法，同时介绍了人类专家在清洗流程中与算法交互的方式和算法学习人类专家经验的过程。

4.2 人在环路的清洗算法整体框架

表 4-1 人在环路的数据清洗算法

算法 7	**HumanInTheLoopCleaning**(*X*)
Input:	待检测和清洗的数据集 *X*
Output:	无
1	**while** True:
2	operation = input() // 选择清洗种类
3	**if** operation == “Missing”: // 如果选择清洗缺失值
4	*X*.isnull = *X*.missingDataDetection //缺失值检测
5	**for** row **in** rows:
6	**for** column **in** columns:
7	**if** *X*.isnull[row,column] 存在缺失值：
8	confidence = AutomaticCorrectData(*X*) // 自动修复
9	**if** !confidence : // 如果不满足自动修复条件
10	MissingDataReport(*X*.row,column) //用户修复
11	RecordUserClassify(L_classify, operation, r, c)
12	**if** operation == “Outlier”: // 如果选择清洗离群值
13	**while** True:
14	column_name = input() // 获取要检查的列名
15	row = OutlierDataDetection(*X*, column_name, rows)
16	confidence = AutomaticCorrectData(*X*, row, column)
17	**if** !confidence : // 如果不满足自动修复条件
18	OutlierDataReport(*X*.row,column) //请求用户修复记录
19	RecordUserClassify(L_classify, operation, row, column)
20	**if** ! continue_check = input() // 是否检查其他离群点
21	**break**
22	**if** ContinueOperation == “false” // 选择是否进行下一次清洗
23	break

算法 7 为人在环路的混合型错误数据清洗整体算法框架。如下图 4.1 所示，首先由用户选择要检测并修复的数据质量问题。获取用户的选择后算法分别通过判断单元格是否为空和箱线图的方式检测数据集中存在的缺失值和离群值。对于每个检测到的异常值，算法会先通过构建机器学习神经网络模型判断是否满足自动修复要求。如果满足要求，则算法自动修复并更新数据集，完成对一个异常值的清洗。如果不满足自动修复置信度，则将异常值所在行及其所属特征告知用户，请求用户提供修复方式和修复值。用户可能会根据自己的经验和周围环境的变化给出正确范围的修复值，但如果用于也对该异常没有合适的修改建议，可以选择删除异常值所在行。如果用户认为该值并非异常值，可以选择忽略本次异常。当算法收到用户的反馈后，会更新数据集，同时记录用户提供的修复方式和修复值，并将它们用于训练神经网络模型和 K 近邻回归模型，作为更新环境变化的方式和学习的过程，用于后续异常值的自动修复。

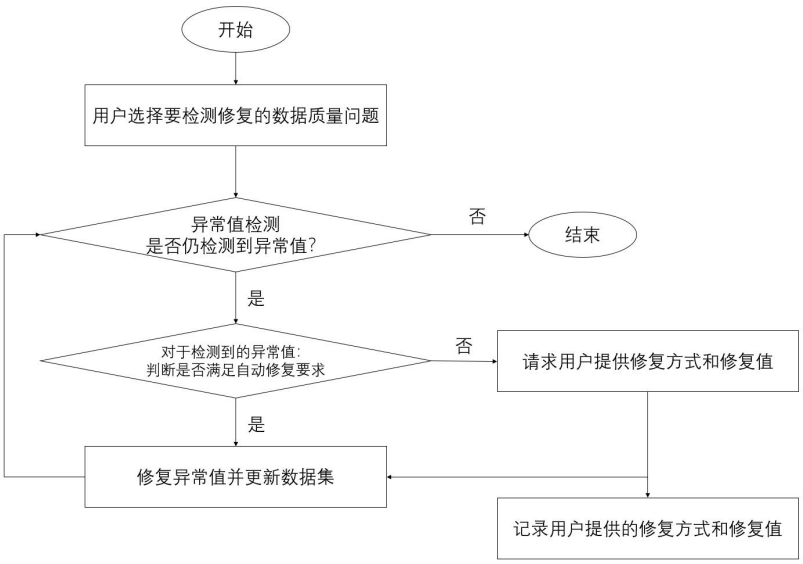

图 4-1　人在环路的混合型错误数据清洗整体算法框架

4.3 人机交互方式与算法学习过程

在整体清洗流程中，用户与算法共有三次交互过程。第一次是选择检测数据存在的质量问题类别，第二次是针对某一异常值给出修复方式，如果第二次给出的修复方式是替换值，则第三次交互是给出要替换的值。

本研究的贡献之一则是发挥人类专家对数据相关领域长期积累的经验和对环境变化的感知的优势，进而调整持续生成的数据会发生动态变化的合理取值区间。在整体清洗框架中，我们把人类专家的修改建议以二维表的形式存储在

记录用户修复方式数据集和记录用户修复取值数据集。两个数据集如下表所示，均有 5 个属性，前四个属性相同，分别为：数据异常原因、原数据值、异常数据值所在列号和行号。最后的属性为标签即目标变量，分别为用户修改方式和用户修改的数据值。

表 4-2　记录用户修复方式数据集

	参数一	参数二	参数三	参数四	标签/目标变量
名称	数据异常原因	原数据值	异常数据值所在列	异常数据值所在行	用户修改方式
类型	0 – 数据缺失 1 – 数据离群	Null –数据值缺失 其他：整数/浮点数	整数	整数	0 – 数据被删除 1 – 数据被修改 2 – 数据忽略异常

表 4-3 记录用户修复取值数据集

	参数一	参数二	参数三	参数四	标签/目标变量
名称	数据异常原因	原数据值	异常数据值所在列	异常数据值所在行	用户修改数据值
类型	0 – 数据缺失 1 – 数据存在	Null –数据值缺失 其他：整数/浮点数	整数	整数	整数/浮点数

记录用户修复方式数据集会被用于构建神经网络的训练数据集，前四个参数为训练特征，最后的参数为训练标签。当使用该表的数据完成训练后即算法完成了最新的学习过程，读入给定异常值的相关参数即可预测用户修改方式（用 0，1，2）表示。记录用户修复取值数据集会被用于构建 K 近邻回归模型的训练数据集。同样，前四个参数为训练特征，最后的参数为训练标签。当完成训练后，读入给定异常值的相关参数即可预测修复后的数据值（用整数或浮点数表示）。

4.4 本章小结

本章介绍了人在环路的混合型数据清洗整体算法框架和人机交互方式。给出记录用户修复方式和取值的数据集的结构化定义。

第 5 章 实验评估

5.1 引言

本章将对上文提出的人在环路的混合型错误数据清洗算法进行实验评估，分析其检测并修复异常值，提高数据质量的能力。本章将首先描述实验前的准备工作，包括实验环境、选用的数据集和采用的度量标准。之后会介绍不同噪声数据比例下的分类准确度实验和合理取值环境变化后的回归准确度实验的结果，并对结果进行分析，归纳总结算法的清洗能力。

5.2 实验设置

5.2.1 实验环境

本实验运行在 Intel(R) Core(TM) i7-10875H CPU @ 2.30GHz 的 CPU 和 16GB 内存的计算机上。使用了 Windows 10 操作系统和 Pycharm 2022.2 集成开发环境。本实验使用的编程语言为 Python，解释器版本为 Python3.7。

5.2.2 数据集

本实验使用的数据集是芝加哥市犯罪数据。该数据集反映了从 2001 年至今在芝加哥市发生的犯罪报告事件。该数据集是以明确定义的格式存储和组织的二维表形式的结构化数据。该数据集共有 22 个特征，其中有 10 个整形特征、3 个浮点型特征、5 个字符型特征、2 个布尔型特征和 2 个日期格式的特征。符合本文对混合型数据的定义。

5.2.3 度量标准

本文的第一个实验是验证在人类提供的数据存在噪声时，算法的分类准确率。我们认为，当使用该算法进行实用的数据清洗过程时，人类专家并不总会提供正确、统一的分类建议。由于人类专家对于异常数据的理解存在偏差或操作存在疏忽等情况，他们会在异常数据的所有参数均相似，即可以被归类为相似错误、应该采用相同的分类方式时，有时会提供误导性的分类意见，本文亦称为噪声意见。为此，本文开展第一个实验，验证在人类专家提供不同比例的噪声意见时算法分类的准确率。针对该实验的详细结果在 5.2.1 节讲述。

分类准确率是本文实验的第一个度量标准。见式（5-1），我们假设C为所有的分类方式，$c_{correct} \in C$为相似错误应采用的统一分类方式，$c_{noise} = \{x \in C \mid x \neq c_i\}$为错误采用的噪声分类方式。则分类准确率可以被形式化定义为：

$$Accuracy_{classify} = \frac{c_{correct}}{\frac{|C|}{c_{correct} + \sum_{i=1} c_{noise-i}}} \tag{5-1}$$

本文的第二个实验是测试环境改变后算法修复的准确率。在实际的清洗过程中，由于环境因素的变化，有时数据的合理取值范围也会发生变化。清洗算法本身难以注意到此类变化，因此可能对数据进行错误的检测和修复，使其与事实有较大偏差。而引入人类专家提供指导性意见则是对此类问题的其中一个较好的解决方案。为此，本文开展第二个实验，验证环境改变后，算法根据人工提供的指导性意见对数据采取适应环境改变的修复的准确率。针对该实验的详细结果在 5.2.2 节讲述。

适应环境改变的修复准确率是本文实验的第二个度量标准。根据本文第二章的定义，数据点为$x_i^{f_j}$。同时，我们假设T_{before}为环境变化前的时间段，在该时间段内，数据的合理取值范围为R_{before}。与此类似，假设T_{after}为环境变化后的时间段，在该时间段内，数据的合理取值范围为R_{after}。见式（5-2），则适应环境改变的修复准确率可以形式化描述为：

$$Accuracy_{value} = \frac{x_i^{f_j} \in \left(R_{before} \cap T_{before}\right) + x_i^{f_j} \in \left(R_{after} \cap T_{after}\right)}{x_i^{f_j} \in \left(T_{before} + T_{after}\right)} \tag{5-2}$$

5.3 实验结果

考虑到实际应用本算法进行数据清洗时存在的变量，对于下述两个实验我们均设置如下两个变量：

（1）与预测数据相关的数据占所有数据的比例。人类专家提供的建议并不会全部与某次预测有关，与预测无关的建议可能会对预测产生噪声影响。因此，我们设置与预测数据相关的数据占所有数据的比例分别为 5%，15%，25%。

（2）用户提供修复方法数据集总数据量。人类专家提供的修复方法建议的数量会影响算法学习和构建模型的准确性。因此，我们设置该数据量分别为 10000，100000，1000000。

5.3.1 不同噪声数据比例下的分类准确度实验

在本实验中，我们分析了芝加哥市犯罪数据集，对于警员 ID 一列特征，其正常取值为 12940000-13050000。该特征包含多个异常数据值即离群点。我们假设人类专家对离群点的正确统一修复方式为 $c_{correct}=repair$，噪声意见为 $c_{noice}=delete\cup ignore$。在上述前提下，我们假设正确非噪声意见分别占用户提供的总意见比例为 30%、40%、50%、60%、70%、80%、90%、100%。实验验证在人类专家提供不同比例的噪声意见时算法分类的准确率 $Accuracy_{classify}$。

（1）用户提供修复方法数据集总数据量为1×10^5时：

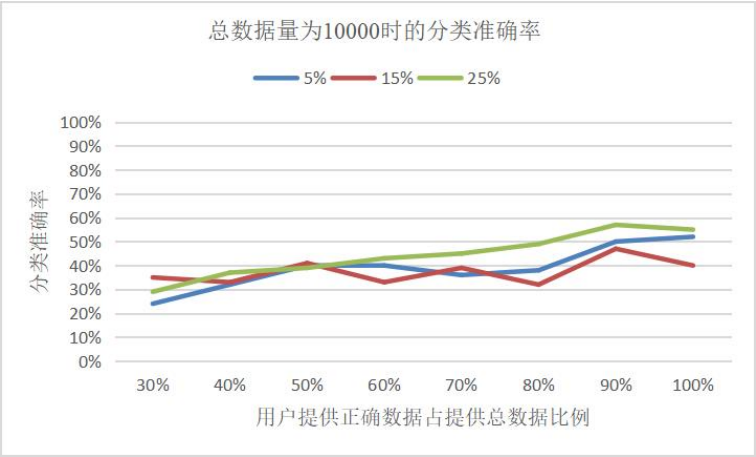

图 5-1　总数据量为 10000 时的分类准确率

（2）用户提供修复方法数据集总数据量为1×10^6时：

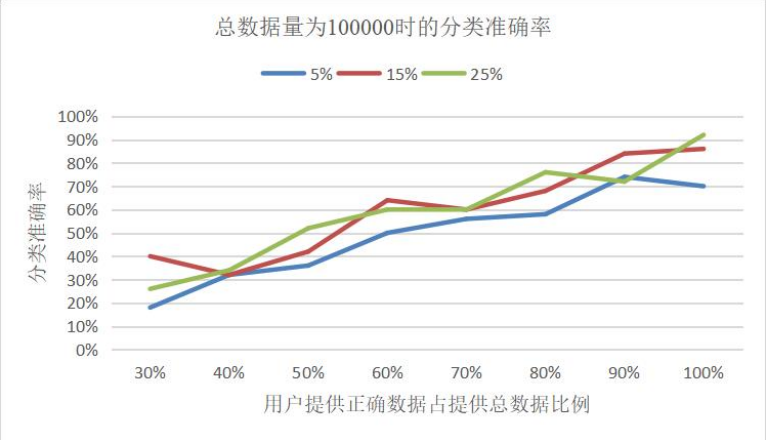

图 5-2　总数据量为 100000 时的分类准确率

（3）用户提供修复方法数据集总数据量为1×10^7时：

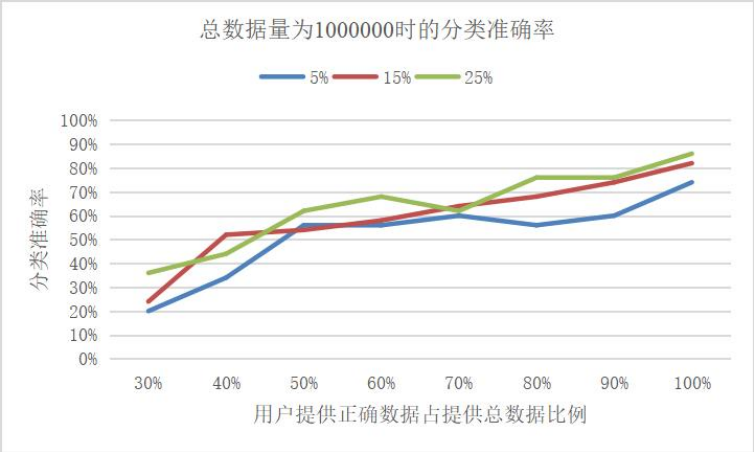

图 5-3　总数据量为 1000000 时的分类准确率

5.3.2 合理取值环境变化后的回归准确度实验

在本实验中，我们假设属于警员 ID 特征的某些异常数据值被算法要求采用修复的清洗方式 $c_{correct}=repair$。而由于环境的变化，人类专家给出的指导意见在总数据量 10%-25%阶段的合理取值区间为 13000000 – 13000100，即当 $T_{before}\in(|D|\times10\%\sim|D|\times10\%)$ 时，$R_{before}\in(13000000,13000100)$。在总数据量 75%-90% 阶段的合理取值区间为 13000900 – 13001000，即当 $T_{after}\in(|D|\times75\%\sim|D|\times90\%)$ 时，$R_{after}\in(13000900,13001000)$。在上述前提下，我们假设用户提供的环境变化前数据量分别占用户提供的总数据量比例为 10%、20%、30%、40%、50%、60%、70%、80%、90%。实验验证在合理取值环境变化后算法预测回归值的准确率 $Accuracy_{value}$。

（1）用户提供修复方法数据集总数据量为 1×10^5 时：

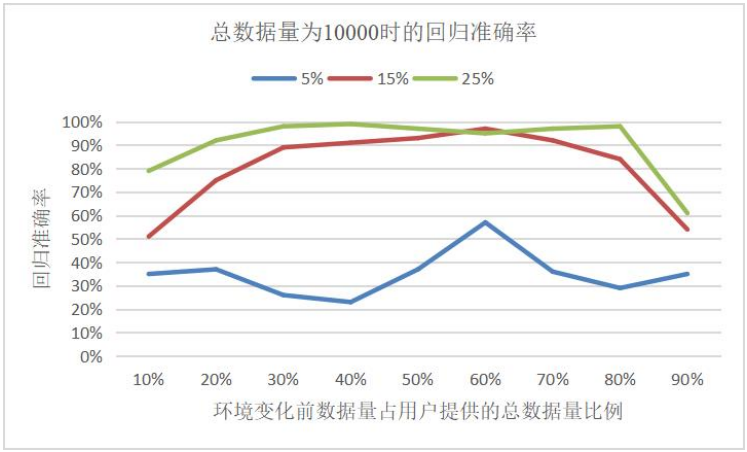

图 5-4　总数据量为 10000 时的回归准确率

（2）用户提供修复方法数据集总数据量为1×10^6时：

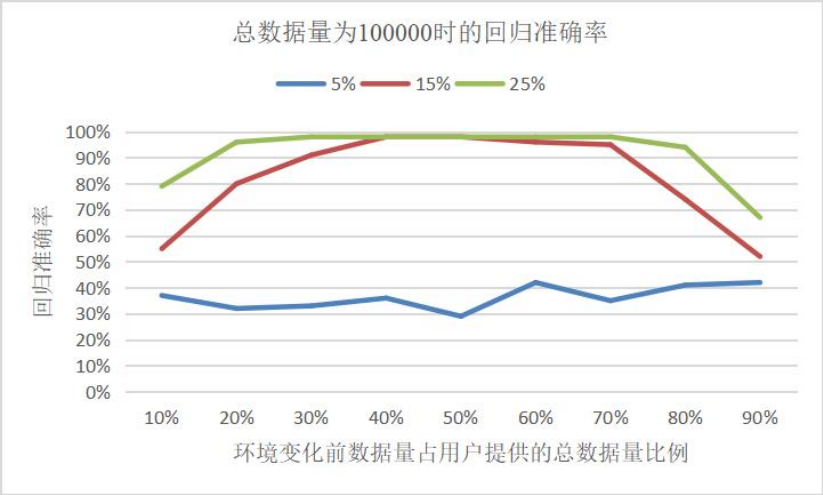

图 5-5　总数据量为 100000 时的回归准确率

（3）用户提供修复方法数据集总数据量为1×10^7时：

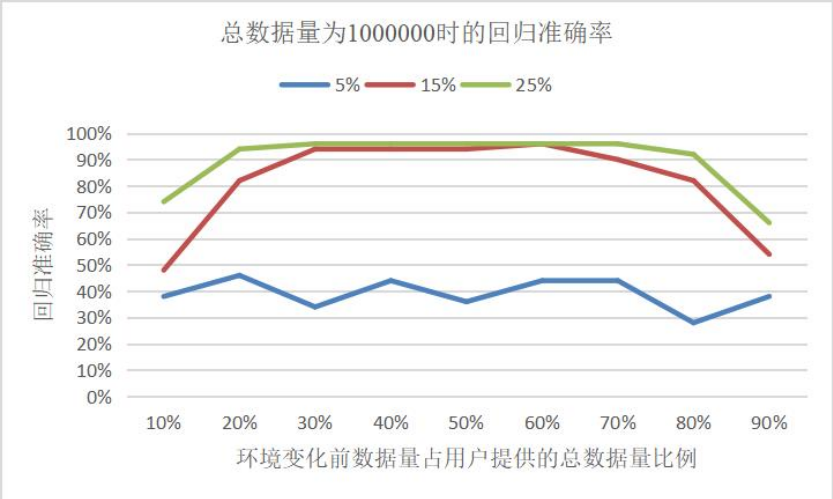

图 5-6　总数据量为 1000000 时的回归准确率

5.4 结果分析

5.4.1 不同噪声数据比例下的分类准确度实验结果分析

分析 5.2.1 节展示的针对不同噪声数据比例下的分类准确度实验，可以得出如下结论：

（1）与预测数据相关的数据占所有数据的比例对算法的分类准确率影响较小。当与预测数据相关的数据占所有数据的比例提升时，算法的分类准确率普遍会提高。不过提高的幅度极其有限。当相关数据占比提升 10%时，分类准确率的提高普遍要小于 10%。虽然从整体数据变化趋势上看相关数据比例的增高会给算法准确率带来正面的影响，但是从图像中依然可以看出数据的波动较大。此外，与预测数据相关的数据占比越高，越会挤压其他数据的空间，进而导致人类专家提供意见的多样性减少。因此，综合来看相关数据占比增高不会给算法在实际清洗中带来积极影响。在实际清洗时，相关数据占比为 5%左右时，人类专家提供意见的多样性会相对最高，对清洗大范围的异常数据更有利。

（2）用户提供数据的正确率与算法的分类准确率有密切的关系。从实验结果图示中可以看出，用户提供数据的正确率与算法的分类准确率基本呈线性正相关的关系。在数据总量满足一定要求的前提下例如1×10^6以上时，当正确数据占比只有 30%时，算法的分类准确率仅在 30%左右，并不能满足实际情况的清洗需求。随着用户提供数据的正确率的提升，算法的分类准确率显著增高。在用户提供数据的正确率达到 90%，算法的分类准确率已经在 75%左右。这是因为机器学习的神经网络模型对训练数据集的正确率要求较高。如果错误率过高会产生较严重的噪声数据干扰。因此，该实验表明本研究的人在环路数据清洗算法对人类专家的分类意见准确率有较高的要求，所挑选参与人机交互的人类专家要充分了解数据的特征和合理分布范围。

（3）总数据量对算法的分类准确率有较大影响。根据实验结果图示可知，在总数据量为1×10^5时，无论与预测数据相关数据的占比如何变化，或是用户提供的正确率如何变化，算法的分类准确率均低于 60%左右。这并不能满足实际清洗需求。而当总数据量在1×10^6及以上时，数据量已不再成为限制算法分类准确率的影响因素。算法的分类准确率会随着用户提供数据的正确率的增长而线性增长。因此，当人类专家提供的修复意见数据量在1×10^6以上时，本算法才可以发挥高质量的清洗能力。

5.4.2 合理取值环境变化后的回归准确度实验结果分析

分析 5.2.2 节展示的针对合理取值环境变化后的回归准确度实验，可以得出如下结论：

（1）算法的预测修复值回归准确率与预测数据相关的数据占所有数据的比例有较大关系。当预测数据相关数据仅占总数据的 5%左右时，算法的回归准确率保持在 30%-40%左右，此时算法不满足实际清洗需求。而当预测数据相关数

据占总数据的比例达到 15%左右时，算法的回归准确率有较高的提升，保持在 80%-100%左右，此时算法在实际清洗过程中可以给出普遍符合要求的修复值。当预测数据相关数据占比进一步提升，达到 12%左右时，对于环境变化前数据量和变化后数据量有明显差异时，对算法回归准确率有 25%-30%左右的提升效果，但是在二者数据量相当时对算法的回归准确率提升程度不大，算法在此时已经达到 95%以上的准确率。本文认为，之所以算法的预测修复值回归准确率与预测数据相关的数据占所有数据的比例有较大关系的原因是与预测数据无关的数据对机器学习 K 近邻模型的学习产生了较大影响，即成为噪声数据。因此，在实际清洗过程中预测数据相关数据占所有数据的比例需要保持在 15%以上。

（2）环境变化前后的数据量比例对算法的回归准确率也有较大的影响。当预测数据相关数据占所有数据比例在 10%以上时，环境变化前后的数据量如果比例接近，即在 3:7 – 7:3 之间时，算法可以达到 95%以上的准确率。但如果二者的比值较为悬殊，比如在 1:9 或 9:1 左右时，算法的回归准确率将下降到 50%-80%左右。因此，在实际清洗中，人类专家应调控在环境变化前后给出的指导数据量比例不能过于悬殊，最好保持在 3:7 – 7:3 之间。

（3）总数据量对算法的回归准确率影响极其有限。在实验中，总数据量的提升通常可以将算法的回归准确率提升 5%左右。在与预测数据相关数据占所有数据比例较低、环境变化前后数据量比例有较大差异时，提高总数据量会在实际清洗时带来一定的帮助，但是在上述两个变量有较好的效果时，总数据量可以保持较低，如1×10^5的水平。在实验中，总数据量与算法运行时间基本保持线性的正相关关系，当总数据量有 10 倍的提升时，算法也有约 10 倍左右的运行时间消耗。因此，为了保持人机交互的高效率，在上述两个变量有较好效果时可以保持较低的总数据量。

5.5 本章小结

本章对本文提出的人在环路的混合型错误数据清洗算法进行实验评估，分析了算法的清洗性能和为了进行高质量清洗对数据所需的各方面要求。本章首先介绍了实验环境、采用的数据集和度量标准。提出不同噪声数据比例下的分类准确度实验和合理取值环境变化后的回归准确度实验。并考虑与预测数据相关的数据占所有数据的比例、用户提供修复方法数据集总数据量、正确非噪声意见分别占用户提供的总意见比例和用户提供的环境变化前数据量分别占用户提供的总数据量比例这些变量变化时算法的分类准确率和预测值回归准确率的改变。在结果分析部分总结归纳了算法如何保持高质量清洗性能。

结　论

海量数据产生并被应用于人们的生产生活中，数据质量决定着数据被分析和使用的价值。近年来，学术界和工业界提出了多种识别和修复数据质量问题的方法，其中人工参与数据清洗得到了众多学者的重视。然而现阶段人工参与的数据清洗技术存在缺乏构建约束规则的协同性、参与验证计算修复算法低效、人工成本高等问题。

以提高数据清洗质量、避免无效人工操作、降低人工投入成本为目的，本文提出并实现人在环路的混合型错误数据清洗技术。主要研究内容及贡献如下：

（1）对国内外有关人工参与的数据清洗技术的研究现状进行了分析。根据人机交互方式，将相关研究划分为提供依赖关系和修复反馈验证方面。总结了现有算法在适用数据集类型和人工参与方式上的差异，以及它们的优点和不足。

（2）以形式化定义的方式规范清洗框架面向的数据类型，明确数据质量评价标准，确定人类专家要求及与清洗算法交互方式。为提出算法奠定理论基础。

（3）提出了人在环路的混合型错误数据清洗框架。该框架包括对缺失点和离群值两类常见的数据质量问题的检测和修复方法。将清洗算法与人类专家交互，发挥人类专家对数据相关领域长期积累经验和对环境变化感知的优势，进而调整持续生成的数据会发生动态变化的合理取值区间，对异常点进行更精确的检测与修复。

（4）对本文提出的算法进行实验分析。以多变量的不同取值展开实验，分析为了进行高质量清洗对数据所需的各方面要求，展示了算法的分类准确率和预测值回归准确率。

本文实现的人在环路混合型错误数据清洗技术在有效的利用人类专家对算法清洗给予积极指导意见的同时降低了人工投入成本，显著提高数据清洗质量并保留数据的时效价值。

本文在人工参与的清洗技术领域展开研究，并提出了有一定创造性的清洗模型。然而，随着研究的进展，我们认识到未来仍然需要开展以下方面的工作：

（1）当前研究对人类专家提供的意见数量有较高要求。由于不能保证人类专家总能提供大量意见，未来的研究应致力于减少对提供意见数量的依赖。

（2）在本研究开展的过程中，生成式 AI 模型取得进展并在众多领域广泛应用。未来的研究工作应考虑将生成式 AI 模型与人工参与的数据清洗技术结合，优化人机交互对话方式并考虑使其分担部分人类专家工作以提高清洗效率。

参考文献

[1] 郝爽，李国良，冯建华，王宁. 结构化数据清洗技术综述. 清华大学学报(自然科学版), 2018, 58(12): 1037-1050.

[2] Sarah E, Justin L, Jennifer C, et al. Ten practical questions to improve data quality, Rangelands, Volume 44, Issue 1. In 2022.

[3] SHALLCROSS S.2 Reasons why your data is lying to you. (2016-09-15).

[4] Solomon M, Addise M, Tassew B, et al. Data quality assessment and associated factors in the health management information system among health centers of Southern Ethiopia. PLOS ONE 16(10): e0255949. In 2021.

[5] Hosseinzadeh M, Azhir E, Ahmed O.H, et al. Data cleansing mechanisms and approaches for big data analysics: a systematic study. Journal of Ambient Intelligence and Humanized Computing 14, 99-111. In 2023.

[6] Fakhitah R, Wan M, et al. A Review on Data Cleansing Methods for Big Data, Procedia Computer Science, Volume 161, ISSN 1877-0509. In 2019.

[7] 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.

[8] M. Volkovs, F. Chiang, J. Szlichta and R. J. Miller, "Continuous data cleaning," 2014 IEEE 30th International Conference on Data Engineering, Chicago, IL, USA, 2014, pp. 244-255, doi: 10.1109/ICDE.2014.6816655.

[9] Feliks P. Sejahtera Surbakti, Wei Wang, et al. Factors influencing effective use of big data: A research framework. Information & Management. In 2020.

[10]Alexander Ratner, Stephen H. Bach, et al. Snorkel: Rapid Training Data Creation with Weak Supervision. VLDB Endowment. In 2017.

[11]Vijayshankar Raman, Joseph M. Hellerstein, et al. Potter's Wheel:An Interactive Data Cleaning System. Proceedings of the 27th VLDB Conference. In 2001.

[12]Ju Fan, Guoliang Li. Human-in-the-loop Rule Learning for Data Integration. Bulletin of the IEEE Computer Society Technical Committee on Data Engineering.

[13]Ziawasch Abedjan, John Morcos, et al. DataXFormer: A Robust Transformation Discovery System. In ICDE. 2016.

[14]Xu Chu, John Morcos, et al. KATARA: A Data Cleaning System Powered by Knowledge Bases and Crowdsourcing. In SIGMOD Conference. 2015.

[15]Ahmad Assadi, Tova Milo, et al. DANCE: Data Cleaning with Constraints and Experts. IEEE 33rd International Conference on Data Engineering(ICDE). In 2017:1409-1410.

[16]Mohamed Yakout, Ahmed K. Elmagarmid, et al. Guided Data Repair. VLDB Endowment. In 2011.

[17]GB/T19000 – 2000 idt ISO 9000:2000[s] 质量管理体系——基础和术语.

[18]Naumann, F., & Rolker, C. (2000). Assessment Methods for Information Quality Criteria. *IQ*.

[19]Danette McGilvray. 数据质量工程实践：获取高质量数据和可信信息的十大步骤[M]. 电子工业出版社, 2010.

[20]Laudon, Kenneth C. Data quality and due process in large interorganizational record systems[J]. Communications of the ACM, 1986, 29(1):4-11.

哈尔滨工业大学本科毕业论文（设计）

原创性声明和使用权限

本科毕业论文（设计）原创性声明

本人郑重声明：此处所提交的本科毕业论文（设计）《人在环路的混合型错误数据清洗技术研究》，是本人在导师指导下，在哈尔滨工业大学攻读学士学位期间独立进行研究工作所取得的成果，且毕业论文（设计）中除已标注引用文献的部分外不包含他人完成或已发表的研究成果。对本毕业论文（设计）的研究工作做出重要贡献的个人和集体，均已在文中以明确方式注明。

作者签名：李昊轩　　日期：2023 年 5 月 28 日

本科毕业论文（设计）使用权限

本科毕业论文（设计）是本科生在哈尔滨工业大学攻读学士学位期间完成的成果，知识产权归属哈尔滨工业大学。本科毕业论文（设计）的使用权限如下：

（1）学校可以采用影印、缩印或其他复制手段保存本科生上交的毕业论文（设计），并向有关部门报送本科毕业论文（设计）；（2）根据需要，学校可以将本科毕业论文（设计）部分或全部内容编入有关数据库进行检索和提供相应阅览服务；（3）本科生毕业后发表与此毕业论文（设计）研究成果相关的学术论文和其他成果时，应征得导师同意，且第一署名单位为哈尔滨工业大学。

保密论文在保密期内遵守有关保密规定，解密后适用于此使用权限规定。

本人知悉本科毕业论文（设计）的使用权限，并将遵守有关规定。

作者签名：李昊轩　　日期：2023 年 5 月 28 日

导师签名：丁小欧　　日期：2023 年 5 月 28 日

致　谢

四年的本科生涯转瞬即逝。回首望去，学习了很多知识，结识了众多挚友，请教了诸多良师，收获了许多成长。这必将是对我人生道路有着深远影响的四年，感谢在这宝贵的时光里遇到的每一个人。

首先感谢我的导师丁小欧讲师。相比于老师，我更习惯小欧姐这个称呼。自 2021 年相识已两年有余，小欧姐是我学术道路上的启蒙人，我也很荣幸见证了你从学生到教师身份的转变。在小欧姐的引领下，我了解了数据清洗领域的前沿知识，在学习和实践中构建了科研思维，提升了学术素养。感谢你的悉心教诲和对我的鼓励与宽容。我还要感谢王宏志教授，感谢你接纳我成为海量数据计算研究中心的一员，也感谢你在我学业成长中至关重要的支持。

感谢在本科四年教授过我的每位老师，你们卓越杰出的教学能力和耐心细致的指导使我学到了许多宝贵的知识，让我向着规格严格，功夫到家的哈工大人稳步迈进。感谢辅导员老师和教学秘书老师，你们专业负责的工作带给我重要的支持和帮助。感谢学校和学部领导对我的培养与关怀。

感谢在这四年里一起学习成长的伙伴们，很幸运能遇到众多的挚友，收获珍贵的友谊。我们一起并肩前行，在冰城留下了许多足迹。你们让我的本科四年时光充实欢愉，我会珍惜我们在一起度过的每一个片段，保存好这份回忆。

感谢父母在我成长道路上的支持，让我拥有接受良好教育的环境。你们的陪伴让我敢于面对人生道路上的一个个挑战，朝着自己向往的生活前行。

奈何时光飞逝，要朝着下一段人生旅程匆匆奔赴。感谢在本科四年里遇到的每一个人，我会留存好这段珍贵的回忆。谢谢！