软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software, 2023,34(3):1049−1064 [doi: 10.13328/j.cnki.jos.006791]

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

基于多视角的多类型错误全面检测方法*

彭锦峰, 申德荣, 寇 月, 聂铁铮

(东北大学 计算机科学与工程学院, 辽宁 沈阳 110169)

通信作者: 彭锦峰, E-mail: pengjinfeng11@163.com

摘 要: 随着信息化社会的发展, 数据的规模越发庞大, 数据的种类也越发丰富. 时至今日, 数据已经成为国家和企业的重要战略资源, 是科学化管理的重要保障. 然而, 随着社会生活产生的数据日益丰富, 大量的脏数据也随之而来, 数据质量问题油然而生. 如何准确而全面地检测出数据集中所包含的错误数据, 一直是数据科学中的痛点问题. 尽管已有许多传统方法被广泛用于各行各业, 如基于约束与统计的检测方法, 但这些方法通常需要丰富的先验知识与昂贵的人力和时间成本. 受限于此, 这些方法往往难以准确而全面地检测数据. 近年来, 许多新型错误检测方法利用深度学习技术, 通过时序推断、文本解析等方式取得了更好检测效果, 但它们通常只适用于特定的领域或特定的错误类型, 面对现实生活中的复杂情况, 泛用性不足. 基于上述情况, 结合传统方法与深度学习技术的优点, 提出了一个基于多视角的多类型错误全面检测模型 CEDM. 首先, 从模式的角度, 结合现有约束条件, 在属性、单元和元组层面进行多维度的统计分析, 构建出基础检测规则; 然后, 通过词嵌入捕获数据语义, 从语义的角度分析属性相关性、单元关联性与元组相似性, 进而基于语义关系, 从多个维度上更新、扩展基础规则; 最终, 联合多个视角对多种类型的错误进行全面检测. 在多个真实数据集与合成数据集上进行了实验, 结果表明, 该方法优于现有的错误检测方法, 并且能够适用于多种错误类型与多种领域, 具有更高的泛用性.

关键词: 数据质量; 错误检测; 多视角; 数据语义

中图法分类号: TP311

中文引用格式: 彭锦峰, 申德荣, 寇月, 聂铁铮. 基于多视角的多类型错误全面检测方法. 软件学报, 2023, 34(3): 1049–1064. http://www.jos.org.cn/1000- 825/6791.htm

英文引用格式: Peng JF, Shen DR, Kou Y, Nie TZ. Comprehensive Error Detection Method for Multiple Types Errors Based on Multiple Views. Ruan Jian Xue Bao/Journal of Software, 2023, 34(3): 1049−1064 (in Chinese). http://www.jos.org.cn/1000-9825/6791.htm

Comprehensive Error Detection Method for Multiple Types Errors Based on Multiple Views

PENG Jin-Feng, SHEN De-Rong, KOU Yue, NIE Tie-Zheng

(School of Computer Science and Engineering, Northeastern University, Shenyang 110169, China)

Abstract: With the development of the information society, the scale of data has become larger and the types of data have become more abundant. Nowadays, data have become important strategic resources, which are the vital guarantees for scientific management for countries and enterprises. Nevertheless, with the increasing of data generated in social life, a large amount of dirty data come along with it, and data quality issue ensues. In the field of data science, it has always been a pain point that how to detect errors in an accurate and comprehensive manner. Although many traditional methods based on constraints or statistics have been widely used, they are usually limited by prior knowledge and labor cost. Recently, some novel methods detect errors by utilizing deep learning model to inference time series data or analyze context data and achieve better performance. However, these methods tend to be only applicable to specific areas or specific types of errors, which are not general enough for complex reality cases. Based on above observations, this study takes advantages

* 基金项目: 国家自然科学基金(62172082, 62072084, 62072086); 中央高校基本科研业务费(N2116008)
本文由“大数据治理的理论与技术”专题特约编辑杜小勇教授、杨晓春教授和童咏昕教授推荐.
收稿时间: 2022-05-15; 修改时间: 2022-07-29; 采用时间: 2022-09-23; jos 在线出版时间: 2022-10-27

of both traditional methods and deep learning model to propose a comprehensive error detection method (CEDM), which can deal with multiple type errors in multiple views. Firstly, under the view of patterns, basic detection rules can be constructed based on the statistical analysis with constraints from multiple dimensions, including attributes, cells, and tuples. After this, under the semantic view, data semantics are captured by word embedding and attribute relevance, cell dependency, and tuple similarity are analyzed. And hence, the basic rules can be extended and updated based on the semantic relations in different dimensions. Finally, the errors of multiple types could be detected comprehensively and accurately in multiple views. Extensive experiments on real and synthetic datasets demonstrate that the proposed method outperforms the state-of-the-art error detection methods and has higher generalization ability that can be applicable to multiple areas and multiple error types.

Key words: data quality; error detection; multiple views; data semantics

随着社会的发展, 信息技术水平不断提升, 新型的数据采集技术和应用不断出现, 数据的种类与规模也越发庞大. 高质量的数据具有巨大的潜在价值, 大至城镇规划、疫情防控, 小至旅游出行、商品推荐, 数据在社会生活的诸多方面中都具有重要作用, “数据即资产”的观念已经得到了国家和企业的一致认可. 然而, 由于人工错误、环境干扰以及多源数据融合等因素的影响, 数据中往往存在着大量的拼写错误、数据冲突、数据缺失、数据重复等诸多问题, 严重影响了后续的数据管理与分析任务的质量, 并导致了严重的后果. 据统计, 这些错误数据每年给美国经济造成 3.1 万亿美元的损失[1]. 此外, 由于数据集中地址信息的错误, 美国邮政部门每年无法投递的支票超过 175 000 张, 同时, 医疗数据集中的错误数据也是导致健康产业中每年约 98 000 人死亡的重要原因[2]. 因此, 如何全面且准确地检测出数据中所包含的错误, 已经成为世界范围内的一个痛点问题.

时至今日, 已经有众多的研究聚焦于数据中的错误检测工作. 为了有效地评估和检测数据质量, 专家们提出了多项数据质量评估指标, 其中, 数据一致性、实体同一性和数据完整性是尤为关键的 3 个方面. 影响数据一致性的因素主要包括数据录入时产生的拼写错误与数据采集时环境干扰导致的数值异常, 现有方法通常使用约束规则和统计信息进行错误检测, 检测数据间的冲突、规则的违反与数据中的离群值; 影响实体同一性的因素主要有单源数据的重复录入与多源数据的低质集成, 现有方法通常使用属性值匹配与实体解析的方式检测重复数据; 影响数据完整性的因素主要为数据采集与传输时的信息丢失, 现有方法通常通过检测数据中存在的空值, 确定数据缺失情况. 尽管已有很多方法已经取得了良好的成果, 但这些方法往往只适合检测单一或少数特定的领域与错误类型. 与此同时, 虽然基于统计与约束的方法能够广泛用于多方领域, 但受限于人力成本与知识限制, 仍具有较高的局限性[3]. 针对不同的错误类型, 当前错误检测研究的挑战主要包括以下几个方面.

(1) 检测拼写错误与异常值时, 基于约束规则的方法需要大量的领域知识或专家参与, 受限于人力成本与知识限制, 生成的约束规则通常不够全面, 且较为死板; 基于统计信息的离群点检测在数值型数据上有良好的表现, 但随着教育、医疗及电商等领域的发展, 数据中诸如地址信息、疾病诊断和用户信息的文本类型信息逐渐增多, 对于这些数据, 通常难以通过简单的统计信息发掘错误与异常.

(2) 检测数据中的重复元组时, 传统一般采用基于度量的方法, 通过属性值匹配, 比较两个元组之间的相似程度, 从而判断是否存在重复数据. 然而在实际的生产生活中, 由于书写不规范以及多源数据融合等情况, 经常会出现数据库存在同一个实体具有多种不同的表示, 如“新冠”“新型冠状病毒”“COVID-19”等. 这就会导致属性值匹配相似度不高但实际上是重复数据的漏判情况.

(3) 检测数据中的缺失值时, 传统思想通常认为存在空值即为数据缺失. 然而在实际生活中, 信息采集时存在部分空值是正常现象, 如体检数据中, 部分指标仅有少部分人进行检测, 这就会使得大部分人在指标上存在空值. 将这种数据判别为错误数据, 后续进行丢弃或缺失值填充都是不合理的.

综上所述, 尽管传统的错误检测方法已经在多个方面取得了一定的成果, 但仍具有较高的局限性. 集中体现于仅从模式的角度, 凭借有限且死板的约束规则与统计信息进行属性值级别的检测, 容易出现漏判和误判的情况. 而与此同时, 随着人工智能技术的兴起, 深度学习在图像、音频和自然语言处理等方面都取得了重大的成功, 其语义感知能力也为错误检测技术带来了新的发展. 但由于同一个词汇在不同的领域中可能表达

不同的语义, 现有的检测方法往往只适用于单一领域, 并针对其领域特性进行针对性的优化, 如对论文数据集进行作者与论文的信息匹配、社区数据集的用户识别等. 这些方法尽管在特定领域能够取得良好的效果, 但其不具备通用性, 难以应对现实生活中不同领域下的复杂情况. 同时, 对于部分小众的领域, 当缺少丰富的语料库时, 方法性能会大幅下降.

因此, 基于上述考察, 本文提出了一个基于多视角的多类型错误全面检测模型 CEDM, 从功能的角度出发, 可以分为 3 个功能模块.

(1) 异常值检测模块: 基于现有知识与约束, 通过统计数据构建基础规则, 通过对数据集进行自训练的词嵌入, 捕获数据语义; 然后, 基于数据间的语义关系对原始规则进行补充与扩展, 对拼写错误与数值异常等异常值进行全面检测.

(2) 缺失值检测模块: 从模式的角度识别数据集中非关键属性, 然后从语义的角度衡量属性间相似度, 并将具有相同语义的属性相融合, 排除其干扰后, 对数据集中真正缺失值准确地进行检测.

(3) 重复值检测模块: 综合数据元组中属性值的匹配、不同单元间语义上的关联与不同元组间整体语义相似度度量, 对数据中的重复元组进行精准检测.

此外, 本模型也可以从维度的角度进行划分: (1) 首先, 我们在数据库中, 在单元的粒度上进行表征学习, 学习不同单元之间语义关系, 扩展单元粒度上的约束规则; (2) 同时, 我们从元组的角度出发, 综合元组中的单元构成与元组整体, 衡量不同元组之间模式与语义的关系; (3) 并且, 我们还从属性的角度出发, 排除同义属性与无关属性对结果造成的影响. 值得注意的是, 不同维度上的处理是并行的, 单元之间的语义关联可以用于元组整体的评估, 属性上的关系也会影响单元分析与元组匹配的结果. 本模型通过结合数据模式与数据语义, 考虑属性、单元和元组等多个维度, 从多个视角下全面地进行分析, 使其能够适用于不同领域下的不同类型数据, 实现对多种类型错误的全面检测, 具有较高的泛用性. 模型整体结构如图 1 所示.

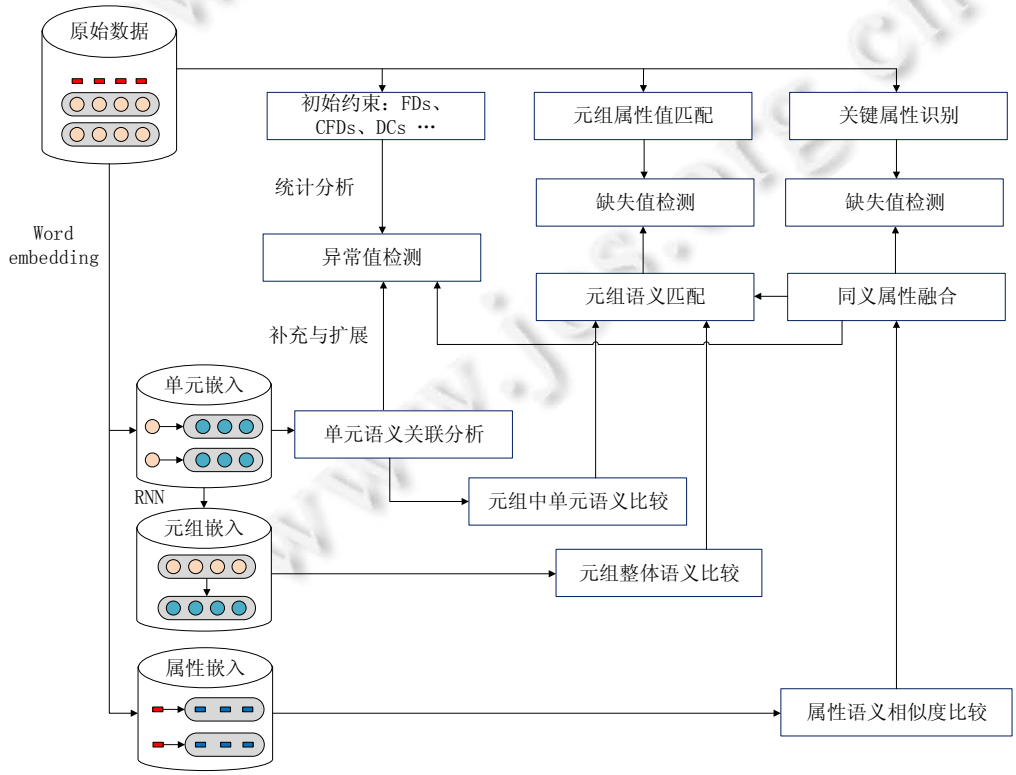


图 1 基于多视角的多类型错误全面检测模型

综上所述, 本文的主要贡献如下.

(1) 提出了一种多视角下的异常数据检测方法, 捕获数据间的语义关系, 对原始规则进行补充与扩展, 提升了规则的灵活性与规则对数据的覆盖率.

(2) 提出了一种多视角下的缺失值检测方法, 结合模式上的统计信息与语义上的属性相似度分析, 识别数据集中非关键属性与同义属性, 提升了检测方法对现实中复杂情况的鲁棒性与泛用性.

(3) 提出了一种多视角下的重复值检测方法, 综合衡量了数据属性、单元与元组整体间的差异, 避免了数据集中由于书写规范不同造成的大量漏判.

(4) 提出了一个面向多错误类型的多视角全面错误检测方法 CEDM, 结合数据模式与数据语义, 在属性、单元、元组等不同维度上, 综合检测数据中的多种错误. 并通过在多个不同领域的真实与合成数据上开展实验, 验证了所提方法的有效性. 实验结果表明, 与现有的常用错误检测方法相对比,本文的方法具有更高的准确率与召回率, 且泛用性更强.

本文第 1 节介绍错误检测的相关工作. 第 2 节给出本文研究相关的基础问题定义. 第 3 节提出本文的多视角全面检测模型, 并对各个模块进行详细的描述. 第 4 节在多个数据集上进行实验, 展示模型的性能. 第 5 节对本文工作进行分析与总结.

1 相关工作

错误检测是数据科学中的一个经典问题, 目前已经有大量的工作聚焦于错误检测技术的研究. 根据错误类别的不同, 错误检测可大致分为 3 类: (1) 针对拼写错误与数值异常的异常值检测; (2) 针对数据重复录入与数据集成错误的重复值检测; (3) 针对信息丢失与数据破损的缺失值检测. 其中, 拼写错误与数值异常是实际数据中最常见的错误类型. 针对这类错误, 使用最广泛的是基于约束的检测方法. 其中, 完整性约束通常由专家或外部知识库进行构建, 通过先验知识, 结合数据的统计信息定义多种约束规则[4], 如函数依赖(functional dependency, FD)、条件函数依赖(conditional functional dependency, CFD)、拒绝约束(denial constraint, DC)等. 将这些规则与数据集中的元组进行匹配, 不满足约束的数据即可被认为是错误数据. 如, Bohannon 等人[5]基于推理系统与一致性分析定义条件函数依赖, 通过检查数据集中违反约束的情况, 检测错误数据; Chu 等人[6]让用户使用拒绝约束来指定质量规则, 基于规则寻找错误数据; Fan 等人[7]提出了匹配依赖(MD), 即在约束规则的左侧引入相似度度量, 将上述严格的相等关系放宽为相似关系. 然而, 在现实中有时难以通过人为定义的规则准确地约束数据. 为此, 很多研究通过统计与学习的方法对数据分布上看起来“异常”的数据值进行检测, 从而发现数据集中的错误数据[8]. 如, Rao 等人[9]通过迭代的 *k*-means 聚类寻找数据中的离群点作为异常数据; Yan 等人[10]通过扩展的局部离群因子 LOF 检测异常数据; Mariet 等人[11]提出了一个新型的集成异常值检测框架 dBoost, 通过直方图统计、高斯分布检测和多元高斯混合等方式联合检测离群值. 然而, 这些方法的准确性与规则覆盖率取决于用户提供的统计参数与制定的规则数量, 需要高昂的人力成本. 由于数据采集时的重复录入和多源数据的低质集成, 数据重复也是真实数据集中的常见问题. 针对数据重复, 最普遍的检测方法是对记录设计一个相似度度量, 相似度越高, 数据重复的可能性越大. 当相似度超过设定的阈值时, 认为其为重复数据[12]. 常用的相似度度量包括有编辑距离、Jaccard 相似度、余弦相似度、欧式距离等[13].此外, 实体解析技术也常用于重复数据检测方法中. 实体解析也称记录连接, 判断两条记录是否指向同一实体对象, 通常可分为基于规则、概率和学习这 3 类. 如, Li 等人[14]从样本数据集中挖掘匹配规则, 然后在整个数据集上应用匹配规则完成记录匹配; Schurle 等人[15]基于 Fellegi-Sunter 模型, 通过匹配两个条件概率的比率,结合约束判定记录是否属于同一实体; Bilenko 等人[16]提出了 MARLIN 模型, 通过支持向量机和决策实现实体匹配, 匹配操作需要利用相似度度量计算记录间的相似度, 从而判断它们是否属于同一个实体. 与此同时, 在数据采集与传输过程中产生的缺失错误也需要准确地识别. 通常来说, 现有的检测方法通常认为存在空值即为缺失. 然而, 缺失值还存在两种特殊情况, 即伪缺失与非错误缺失. 其中, 伪缺失由占位符代替空值造成,这种错误可以通过异常值检测的方式进行识别, 如 FAHES[17]; 非错误缺失是指非核心属性数据所允许的缺失, 如备注等, 该

属性下的缺失不应被记为错误数据, 通常由人工的方式在预处理阶段进行指定与识别. 然而, 这些方法都是对数据进行模式上的统计、匹配与识别, 难以处理真实数据集中记录不同却具有相近语义的数据单元, 遇到如缩写、代号等情况时常常会产生误判或漏判的情况.

与此同时, 尽管上述传统方法已广泛用于多个领域, 但其往往需要大量的先验知识和人力成本. 现实生活中, 知识和人力都是有限的, 进而导致生成的检测规则较为死板, 且对数据的覆盖率不足. 随着人工智能技术的兴起, 通过机器学习学习数据分布进行概率推断, 使用深度学习学习数据表征, 结合数据单元在环境中的语义关系进行检测的方式, 为错误检测技术带来了新的发展方式. 如, Ihab 等人提出了 Holistic[18], 将函数依赖结合上下文进行扩展, 提升了违约检测的准确性与查全率; Shiue 等人[19]通过双向长短期记忆网络检测汉语文本中的错误; Li 等人[20]结合隔离林与深度学习划分数据并训练语义特征, 以检测电力运维数据中存在的错误; Wang 等人[21]通过深度强化学习, 在社交网络数据中匹配重复的用户实体. 但这些方法往往对数据的领域、错误的类型有较大限制, 面对实际生活中不同领域下的多种数据类型中的多种错误, 仍有较大的局限性. 这是由于同一个单词在不同领域或时间下往往具有不同的语义, 因此需要对数据集加以限制或通过社区划分与时间窗口等方式对数据加以划分. 而这就使得这些方法对数据集所处领域、数据类型和错误类型等有较高要求, 在满足要求的特定环境中具有较高的性能, 但对于面对复杂的现实环境难以广泛应用, 泛用性较差.

通过上述观察, 我们可以看出, 现有的数据检测方法大多是针对特定类型的错误, 通过分析其特点与性质, 进行针对性的检测. 虽然能够取得良好的结果, 但难以满足现实情况下需要同时检测多种错误类型的需求. 近年来, 很多研究开始聚焦于更加通用的多类型错误全面检测. 如, NADEEF[22]通过用户指定 FD, MD 和相似度匹配等多种类型的质量规则以识别多类型的错误; Abedjan 等人[23]将多个检测单一错误的方法进行组合, 用于检测多类型的错误, 然而他们也指出, 该方法在捕获真实数据集中的数据错误时只有较低的查全率; He 等人则提出了 Auto-Validate[24]和 Uni-Detect[25], 利用数据湖或大规模语料库学习数据分布模式与规则, 从而对数据进行全面验证. 但与目标数据相匹配的数据湖和语料库等先验知识, 要求并不总是能够满足的.

区别于上述方法, 本文提出的模型以自身为语料库进行学习, 不受限于外界知识; 同时, 深入多种视角, 综合考虑了多种类型错误对数据检测的影响, 能够更加全面且准确地修复数据, 且具有更高的泛用性. 具体来说, 本文的模型不仅使用了多种模式上的检测规则应对多种类型的错误, 还考虑了语义视角下数据间不同纬度上的差异性与关联性, 基于语义扩展并完善检测规则. 一方面, 学习得到的语义关系使模式规则更加全面且灵活, 减少由于规则不足与死板导致的数据漏判与错判; 另一方面, 模式规则约束数据语义关系, 在多个维度上综合检测多种类型的错误, 使得该模型既具有传统方法的泛用性, 又通过数据语义提升了错误检测的准确率与查全率.

2 问题描述与定义

数据质量是度量数据可靠性和可用性的重要指标, 其评估标准包括: 数据一致性、数据完整性、数据精确性、数据时效性和实体同一性[26]. 其中, 数据一致性评估数据记录、格式、内容等方面的一致情况, 若数据集中存在两个单元之间的数据相互矛盾, 则认为数据存在异常; 数据完整性要求所有满足信息需求的数据都必须是存在且可用的, 若数据集中未能包含足够的信息以支持合理的查询和应用, 则认为数据存在缺失; 数据精确性评估数据集中的数据准确表述物理世界信息的能力; 数据时效性要求数据能够反映当前目标的情况与信息, 若数据过于陈旧, 则认为数据可靠性不足; 实体同一性要求同一实例在数据库中的描述是统一的, 即同一实例在数据集中有且仅有一个元组, 若数据库中存在表示相同实例的多个元组, 则认为数据存在重复数据.

本文主要从数据一致性、数据完整性、实体同一性这 3 个方面检测错误数据, 检测数据中存在的异常值、缺失值与重复值. 而对于数据精确性和数据时效性, 尽管数据精度不足和数据老旧可能导致数据可靠性不足, 但其本身并非错误数据, 本文暂不加以考虑. 下面给出错误检测的形式化定义.

定义 1(错误检测). 为了检测数据库中的错误数据, 通常使用数据质量规则或完整性约束作为描述合法或

正确数据实例的声明性方法. 即给定一个数据规则集合Σ和完整性约束Ω, 如果数据库实例 D 违背了Σ中的任何一条规则或Ω中的任何一个约束条件, 则称 D 相对于(Σ,Ω)不一致. 任何不符合已定义规则与完整性约束的数据, 都被认为是错误数据.

其中, 数据规则集合Σ一般是基于知识或统计信息构建的基础约束条件, 如数量约束、等值约束等. 完整性约束Ω是通过数据单元的关联关系描述合法或正确数据实例的一种声明性方法, 包括函数依赖(FD)、条件函数依赖(CFD)、拒绝约束(DC)等[4]. 这些约束规则的形式化定义如下所示.

定义 2(函数依赖). 对于一个包含 m 个属性的关系 R, $Attrs(R)=(A_1,\dots,A_m)$表示 R 上的属性集合, $Dom(A)$表示属性 A 的域. 令 I 表示关系 R 的一个实例, 包含$|I|$个元组, 各元组均属于域 $Dom(A_1)\times\dots\times Dom(A_m)$. 对于元组 $t\in I$, 记为 $t[A]$, 表示元组 t 在 A 属性上的取值. 则 R 上的一个函数依赖 FD 可表示为 $X\to Y$, 其中, $X,Y\subseteq R$. 一个 FD 在数据库实例 I 上成立, 当且仅当$\forall t,t'\in R$, $t[X]=t'[X]\Rightarrow t[Y]=t'[Y]$.

定义 3(条件函数依赖). 关系 R 上的一个条件函数依赖 CFD 可表示为 $R(X\to Y,t_p)$, 其中, $R(X\to Y)$是 R 上的一个标准函数依赖, t_p 是一个具有 X 和 Y 中所有属性的模式表. 一个 CFD 在数据库实例 I 上成立, 当且仅当:

$$\forall t,t'\in R,\ t[X]=t'[X]\asymp t_c[X]\Rightarrow t[Y]=t'[Y]\asymp t_c[Y].$$

定义 4(拒绝约束). 拒绝约束 DC 表示对于关系 R 上的任何一种元组组合, 一组谓词不能同时为真. 其形式化表示为$\varphi:t_\alpha,t_\beta,\dots\in R:\neg(p_1\wedge\dots\wedge p_n)$, $p_i:t_x.A\theta t_y.B\|t_x.A\theta c$. 其中, $x,y\in\{\alpha,\beta,\dots\}$, $A,B\in R$, c 是一个常数, θ是一组有限谓词.

以一个地址信息数据集为例, 其包括省份、城市、地区和邮编等多个属性. 其中, “城市⇒省份”可以看作是一条 FD, 表示对于该数据集中所有具有相同城市信息的元组一定也具有相同的省份信息. 类似地, 地区也可以确定城市, 但又存在部分重名的情况, 则使用 CFD, 如“浑南区, 邮编为 110170⇒沈阳”. 其中, “邮编为 110170”是依赖关系“浑南区⇒沈阳”成立的条件. 而 DC 则排除一个元组不同单元之间有明显冲突的情况, 如对于一个体检数据集, ¬(血糖为低∧高血糖)表示不可存在血糖指标为低而疾病结论为高血糖的情况. 错误检测即是检测数据集中存在违法上述多种约束规则的元组, 并将其记为错误数据的过程.

3 基于多视角的多类型错误全面检测模型

本模型从多个视角下对数据进行全面错误检测, 宏观上包括两个大的方面, 即数据模式与数据语义. 在检测时, 我们考虑多个检测粒度, 包括属性、单元与元组. 在检测过程中, 我们从不同角度分析数据之间的模式关联性与语义相关性, 包括: 一个数据集中不同属性之间的关系、一个元组中不同单元之间的关系、同一个属性下不同单元之间的关系. 此外, 我们还考虑了数值型与字符型数据在检测时的差异. 基于多视角下的联合分析与检测, 我们构建了一个面向多类型错误的全面检测模型. 其整体流程如图 1 所示.

3.1 多视角下的异常值检测

传统方法在检测数据中拼写错误与异常数据时, 最常用的方法是基于约束与统计信息构建检测规则, 检测不满足约束条件与离群的数据. 然而, 人的知识与时间都有限的, 面对日益增长的海量数据与复杂的现实情况, 传统方法获得的检测规则通常较为死板且不够全面, 只有较低的数据覆盖率. 因此, 我们在通过初始约束条件结合统计分析得到初始检测规则的基础上, 从语义的角度分析数据单元之间的语义相关性. 从语义的角度对检测规则进行补充与扩展, 使得规则更加丰富且灵活, 从而提高规则对数据的覆盖率. 下面, 我们介绍具体的检测流程, 并以一个医保数据集为例, 具体地进行解释说明.

给定一个待检测的数据集, 其属性集合为$(A_1,\dots,A_m)$, $t[A]$表示元组 t 在 A 属性上的取值. 我们首先根据现有的信息, 包括 FD、CFD 和 DC 等多种先验知识, 结合统计信息构建初始检测规则. 如根据函数依赖的确定性关系“疾病编码⇒疾病名称”与“疾病名称, 化验指标⇒药物名称”, 然后通过遍历检索、挖掘频繁项集等统计方法指定具体的约束规则, 如, “J42.X02→慢性支气管炎”与“轻度糖尿病, 7.0<血糖<8.4→阿卡波糖片/…”. 但这种规则过于死板, 且仅能覆盖率少量数据, 可能会导致误判与漏判的情况产生. 因此, 接下来我们将整个数

据集为语料，输入每个 cell 自身的数据单元，训练词嵌入模型，捕获单元之间的语义相关性，通过语义信息，扩展并完善检测规则. 本文中，我们采用 Transformer 结构获取数据词嵌入，数据单元输入编码器 Encoder 后，首先经过一个自注意力(self-attention)层进行编码求和与归一化，然后输入一个前馈神经网络(feed forward)并得到下一层编码结果，重复数个 Encoder 后，将结果输入解码器 Decoder，解码后，将结果通过一个全连接神经网络线性层(linear)给出 softmax 得分，训练至收敛后，即可得到对应单元的词嵌入.

值得注意的是，本文中的单元嵌入不同于通常的词嵌入训练，没有采用词典与通用语料库，如维基百科语料进行训练，而基于待检测的数据库本身的数据单元进行自训练. 这是由于在一个数据集中，语义相近的词往往具有相近的上下文. 如，在医疗数据集中，对于患有糖尿病的患者，常用的药物包括格列本脲或瑞格列奈等，这就使得“格列本脲”和“瑞格列奈”在数据集总是对应着相似的疾病名称、医疗项目与患者指标. 换言之，它们在数据集的语料库中往往具有相似的上下文. 因此，在对其进行单元嵌入后，这两种药物就具有了相近的语义向量. 与之相对的，糖尿病 I 型和糖尿病 II 型，可能看似具有相近的语义，但由于疾病的产生机制不同，用药和身体指标都大不相同. 因此，它们在单元嵌入后就只具有较低的语义相似度. 这种自训练的嵌入方式使得获得的语义向量更加符合数据集自身的内容，同时也避免了由于某些较为专业的领域语料不足导致的语义偏差，既提高了嵌入的准确性，也使得方法具有更高的泛用性. 此外，这种表征学习得到的嵌入向量是通过数据库中不同单元的上下文关系确定的，与数据语言的种类无关. 因而，本文的模型能同时应用于中英文等多语言的数据集.

具体地，如图 2 所示，对于满足关系 R 的数据库实例 D，我们将元组 t_1 到 t_n 作为语料，通过 Transformer 模型进行训练，得到对应的单元嵌入. 训练完成后，我们按属性进行分块，对比在不同元组 t_α和 t_β同一属性 A_i 下的单元之间的语义相似度，相似度计算为对应词向量 $\overrightarrow{t_\alpha[A_i]}$和 $t_\beta[A_i]$ 之间的余弦相似度，记为 $sim=\cos\langle\overrightarrow{t_\alpha[A_i]},\overrightarrow{t_\beta[A_i]}\rangle$. 通过比对相似度与语义相似度阈值 T_O，我们可以得到目标单元的语义近似单元，从而扩展初始规则，得到近似函数依赖 AFD. 扩展包括两个方面：一方面，我们补充规则的结果部分(RHS)，如“J42.X02→慢性支气管炎/支气管炎(慢性)”，通过对同义结果的补充，我们能够使规则更加灵活，减少规则的误判率；另一方面，我们扩展规则的推断部分(LHS)，如“糖尿病/轻度糖尿病, 7.0<血糖<8.4→阿卡波糖片/拜唐苹/…”，通过对同义推断部分的扩展，我们能够使规则更加完善，提高规则覆盖率，从而减少规则的漏判率.

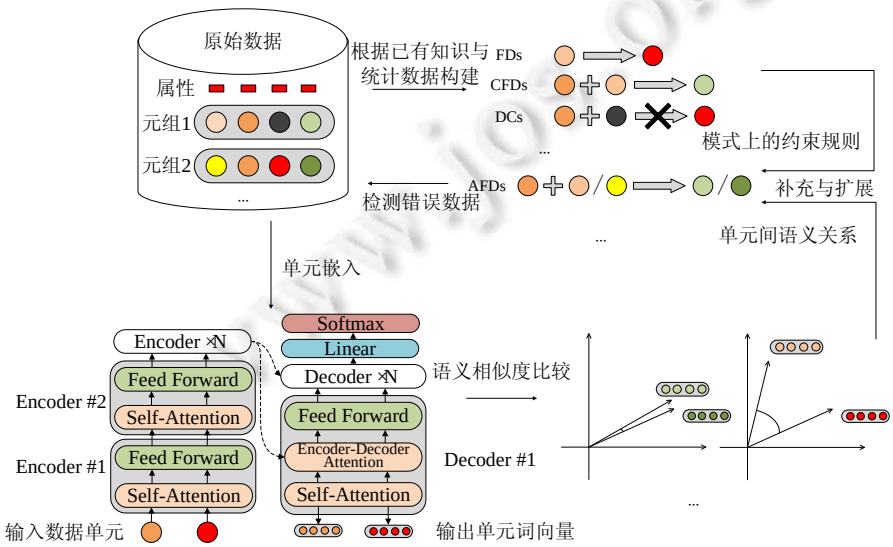


图 2 基于多视角的异常值检测模型

3.2 多视角下的缺失值检测

传统方法在进行缺失值检测时，通常是检测数据集中存在的空值：如果某个 cell 中值为空，即认为该条数据存在缺失情况，并标记为错误数据．然而在实际的信息采集时，并非所有信息都必须填写，这会使得部分人在某些属性上存在数据缺失．此外，多源数据融合也可能导致产生新的缺失值，如图 3(c)所示．将含有这种缺失的数据判别为错误，进行丢弃或缺失值填充都是不合理的．因此，本文不光从模式上属性值的角度检测缺失值，还考虑了属性之间的语义关系以及属性本身的重要程度．

给定一个待检测数据集 D，我们首先对数据集中属性名称(A_1,…,A_m)进行词嵌入．不同于上文中的单元嵌入，获取与一个数据集中全部数据单元相匹配的语料资源是困难的，但属性名称往往是更加通用且常见的．因此，我们基于通用语言模型 TaBert，在本地进行微调，获得属性名称所对应的属性嵌入向量 $(\vec{a}_1,...,\vec{a}_m)$，捕获属性语义．然后，计算不同属性 A_i 与 A_j 之间的语义相似度，记为 $sim_{i,j}=\cos\langle\vec{a}_i,\vec{a}_j\rangle$，进行同义属性融合．如果属性 A_i 与 A_j 之间的语义相似度高于阈值 T_M，则认为 A_i 与 A_j 具有较高概率表达着相同的信息，将其合并．

我们以一个集成体检数据集为例，介绍上述过程．如图 3 所示：图 3(a)、图 3(b)为不同地区的部分体检数据样例；图 3(c)为前两者的集成数据，也是本次错误检测的输入数据．首先，我们对数据集中属性名称(患者编号,体检指标(白蛋白),...,体检指标(AFP))进行词嵌入．在图 3(d)中，我们可以看出，“葡萄糖”与“白蛋白”、“AFP”等属性名称语义相似度较低，而“白蛋白”与“Alb”具有极高的语义相似度，因此将其合并，使得目标数据集从图 3(c)这种具有较高缺失率的稀疏数据集转变成为一个相对密集的数据集，如图 3(e)所示．

患者编号	体检指标(白蛋白)	体检指标(葡萄糖)
B201	75	5.5
B202	45	6.0

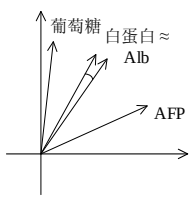
(a) A 组体检数据

患者编号	体检指标(Alb)	体检指标(葡萄糖)	体检指标(AFP)
B203	43		
B204	43	12.7	6.5

(b) B 组体检数据

患者编号	体检指标(白蛋白)	体检指标(Alb)	体检指标(葡萄糖)	体检指标(AFP)
B201		75	5.5	
B202		45	6.0	
B203	43			
B204	45		12.7	6.5

(c) 集成体检数据



(d) 属性嵌入，比较语义相似性

患者编号	体检指标(白蛋白)	体检指标(葡萄糖)	体检指标(AFP)
B201	75	5.5	
B202	45	6.0	
B203	43		
B204	45	12.7	6.5

(e) 语义融合后数据

患者编号	体检指标(白蛋白)	体检指标(葡萄糖)
B201	75	5.5
B202	45	6.0
B203	43	
B204	45	12.7

(f) 消除非关键属性后数据

图 3 基于多视角的缺失值检测方法

从语义视角下合并近似属性后，我们再从模式的角度排除非关键属性带来的影响．在如体检数据、调查问卷数据等数据集中，往往会存在部分属性自身就具有较高的缺失率，将该属性数值缺失的数据判断为错误数据，后续进行丢弃或补全都是不合理的．如图 3(e)中，“体检指标(AFP)”为甲种胎儿蛋白，男性体检时通常不检测该项指标，是体检数据中的非关键数据，因此我们设计了一个关键度度量，用以排除非关键属性的影响．假设属性语义合并后数据集中共有 row 行，col 列，n 个空值，其中，第 i 列中的空值个数为 n_i，对应属性缺失率为 $mr_i=n/row$，全数据集缺失率为 $mr=n/row\times col$，若 $mr_i>\log(col)\times m$，则认为该数据集中第 i 个属性为非关键属性，予以排除，不检测该属性中的缺失值．最后，在合并语义近似属性并排除非关键属性数据后，检测剩余数据中的空值标记为缺失数据，如图 3(f)所示．下面给出缺失值检测的整体流程．

算法 1. 多视角下的缺失值检测.

输入: 数据集实例 D.

输出: 标记的缺失数据.

1) **For each** $A_i \in Dom(A)$ **do**
2) 通过 word2vec 计算属性 A_i 的属性嵌入向量 $\vec{a}_i$
3) **For each** $A_i, A_j \in Dom(A)$ **do**
4) 计算 A_i 与 A_j 之间的语义相似度 $sim_{i,j}$
5) **If** $sim_{i,j} < T_M$ **then**
6) 合并 A_i 与 A_j 的数据
7) 统计属性语义合并后的数据集信息: row 行, col 列, n 个空值
8) 计算全数据集平均缺失率: $mr=n/row \times col$
9) **For each** $A_i \in Dom(A)$ **do**
10) 计算第 i 个属性的缺失率: $mr_i=n/row$
11) **If** $mr_i > \log(col) \times mr$ **then**
12) 在后续的缺失值检测中, 不再检测该属性的缺失值
13) 在合并语义近似属性并排除非关键属性数据后, 检测剩余数据中存在的空值

3.3 多视角下的重复值检测

传统方法在检测数据集中的重复数据时, 一般采用基于度量的方法, 通过属性值匹配, 判断是否存在重复数据. 然而在现实的生产生活中, 由于书写不规范或多源数据融合, 经常会出现同一实体具有多种不同的表示, 如“新冠”“新型冠状病毒”“COVID-19”等, 而这类重复数据通常难以被传统方法所识别. 因此, 本文从语义的角度对传统方法进行优化, 考虑多个维度, 准确地识别重复数据元组.

首先, 依然要考虑数据之间模式上的相似程度. 对于两个元组 t_α和 t_β, 我们第 1 步计算比较其相同属性下单元之间的相似度. 元组 t_α和 t_β在第 i 个属性上的相似度记为 $sim(t_\alpha[A_i], t_\beta[A_i])$. 而由于元组是由多个单元组成, 每个单元间的相似度最终决定了元组是否重复. 因此, 我们将元组上所有单元之间的平均相似度作为元组的模式相似度, 元组 t_α和 t_β之间的整体相似度计算如下:

$$sim(t_\alpha, t_\beta) = \frac{1}{k}\sum_{i=1}^{k} sim(t_\alpha[A_i], t_\beta[A_i]) \tag{1}$$

对于不同类型的数据, 本文采用了不同的度量方法计算单元相似度. 其中, 对于数值型数据 v_i, v_j,二者的相似度通过其数值之差与其中最大值的比例进行衡量, 计算公式如下:

$$sim(v_i, v_j) = 1 - \frac{|v_i - v_j|}{\max(|v_i|, |v_j|)} \tag{2}$$

对于字符型数据, 本文采用编辑距离进行度量. 编辑距离是指通过插入、删除和替换的方式, 将一条字符串转换为另一条字符串所需的用最少操作次数, 是字符串之间距离的经典度量, 广泛应用于记录匹配任务中, 距离越小, 表示字符串越相似. 编辑距离通常采用动态规划的方式进行计算, 本文中给定字符串 s_i, s_j, 记 $d[n,m]$为 s_i 中前个字符与 s_j 中前 m 个字符的编辑距离, 其计算公式如下:

$$d[n,m] = \min \begin{cases} d[n,m-1]+1 \\ d[n-1,m]+1 \\ d[n-1,m-1]+c(s_i[n], s_j[m]) \end{cases}, \quad c(s_i[n], s_j[m]) = \begin{cases} 1, & s_i[n] \neq s_j[m] \\ 0, & s_i[n] = s_j[m] \end{cases} \tag{3}$$

其中, $s_i[n]$为字符串 s_i 的第 n 个字符. 当 $n \times m=0$ 时, $d[n,m]=n+m$. 基于该编辑距离, 字符串相似度 $sim(s_i,s_j)$计算如下:

$$sim(s_i,s_j)=1-\frac{d(s_i,s_j)}{\max(|s_i|,|s_j|)} \tag{4}$$

然而，考虑到尽管书写方式不同，但其可能表达同一实体，即使模式上两个单元具有较大区别，也无法断定其并非重复. 因此，从模式的角度下计算相似度后，我们考虑元组 t_α和 t_β之间的语义相似度，对其进行补充与优化. 对于语义相似度，我们关注 2 个维度的比较，即元组中每个单元的语义相似度与元组整体的语义相似度. 首先，我们比较两个元组中，相同属性A_i下，每个单元之间的语义相似度，$t_\alpha[A_i]$和$t_\beta[A_i]$之间的相似度记为 $sim'(t_\alpha[A_i],t_\beta[A_i])=\cos\langle\overrightarrow{t_\alpha[A_i]},\overrightarrow{t_\beta[A_i]}\rangle$，并以此优化相似度度量. 我们着重考虑了两种特殊情况.

(a) 模式相似度小，而语义相似度大. 这种情况可能是由于书写不规范，如上文提及的“新冠”和“COVID-19”，我们认为，其具有较高相似度.

(b) 模式相似度大，而语义相似度小. 这种情况往往出现于数值型数据，如“40.5”与“40.51”. 我们同样认为其具有较高相似度.

因此，我们度量 $t_\alpha[A_i]$和 $t_\beta[A_i]$之间的相似度，为其模式相似度与语义相似度之间的最大值.

除了单元本身以外，元组整体相似度也是其是否重复的重要指标. 此时，我们不关注每个单元对元组的影响,而是直接比较元组整体的语义相似性. 我们采用 Seq2Seq 模型中的 encoder 部分，将多层神经网络中最后一个输入的隐含状态进行变换，得到语义向量 $\vec{t}_\alpha$ 和 $\vec{t}_\beta$，计算其余弦相似度 $sim'(t_\alpha,t_\beta)=\cos\langle\vec{t}_\alpha,\vec{t}_\beta\rangle$，作为元组整体的语义相似度.

最终，如图 4 所示，我们将模式相似度与语义相似度相结合，计算最终的元组相似度. 对于具有 k 个属性单元的两个元组 t_α和 t_β，其整体相似度定义如下:

$$sim(t_\alpha,t_\beta)=\frac{1}{2}\left(\frac{1}{k}\sum_{i=1}^{k}\max(sim_k,sim'_k)+sim'(t_\alpha,t_\beta)\right) \tag{5}$$

其中，sim_k代表元组 t_α和 t_β在第 k 个属性上的数值或字符串相似度，通过公式(2)或公式(4)计算得到，sim'_k 表示二者在第 k 个属性上的语义相似度，$sim'(t_\alpha,t_\beta)$表示两个元组间整体的语义相似度. 若两个元组的整体相似度大于阈值 T_D，则将该组元组标记为重复数据.

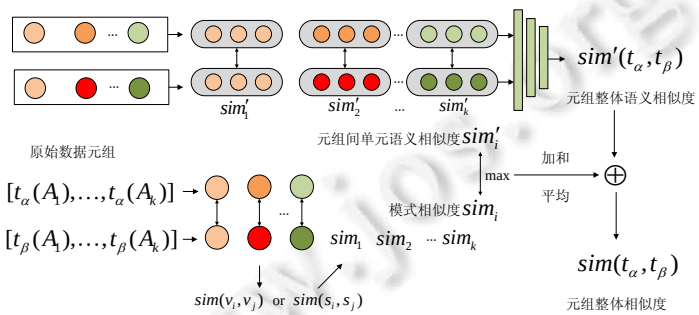


图 4 基于多视角的重复值检测模型

3.4 多类型错误检测的联合优化

传统错误检测方法中，不同类型的错误检测方法之间往往是独立的，互相之间几乎不产生影响. 这是由于错误检测的本质是对错误数据的标注，并不改变数据本身，在错误比例不高的情况下，对数据分布的统计分析也影响较小. 但与模式统计不同，数据语义更侧重于捕获数据之间的关联. 在对语义向量的学习过程中，如果目标单元的上下文中存在缺失值和异常值，会对生成的词向量产生较大影响. 而低质的语义向量又会导致单元和元组的语义相似度比较出现偏差，从而进一步影响异常值与重复值的检测.

为此，本文设计了一个多类型错误检测的联合优化模型，分别将上文中的多视角下异常值检测方法、多视角下缺失值检测方法和多视角下重复值检测方法作为异常检测模块 Model_O、缺失检测模块 Model_M、重复检测模块 Model_D. 首先，基于原始数据的统计信息和先验知识中的规则与约束构建基础检测规则，此时

的规则是死板且不完善的. 然后, 通过属性、单元、元组不同级别的嵌入表示, 获取数据的多维度语义, 基于数据语义扩展基础规则, 此时的规则更加灵活且拥有较高的覆盖率, 但由于错误数据的影响, 存在一定的误差. 接下来, 基于扩展后的规则并行使用 3 个检测模块, 在检测过程中, Model_M 模块检测到的同义属性会进行数据合并, 并传递给 Model_O 模块和 Model_D 模块; Model_O 模块计算的单元相似度会共享给 Model_D 模块; Model_D 模块检测到的重复元组会标记给 Model_O 模块以修正检测结果. 此外, 3 个模块共享错误数据计数, 当数据集中错误比例超过阈值 T 时, 认为此时错误比例较高, 需修正数据语义. 因此, 基于未标记为错误的数据重新学习语义向量并修改扩展规则. 然后, 继续基于修改后的扩展规则进行 3 个模块的检测. 重复上述流程, 直到没有新的错误出现. 流程结构如图 5 所示.

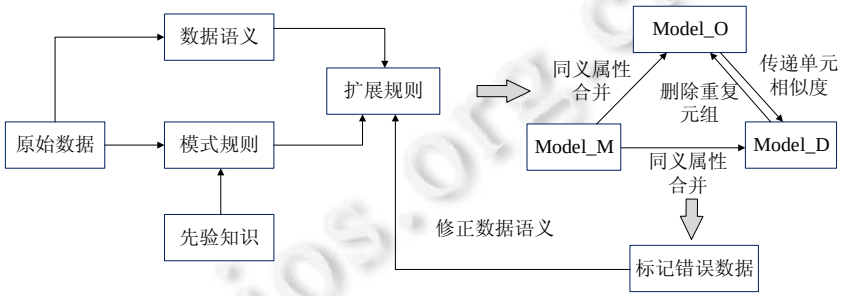


图 5 多类型错误联合检测流程

在联合优化过程中, 异常值检测模块为整体排除了错误数据的干扰, 缺失值检测模块整合了同义属性并排除了非核心属性数据缺失的影响, 重复值检测模块排除了重复数据对检测过程的影响. 三者之间的协作是在多个维度上同时进行的, 检测单元之间的关联会影响元组整体, 标记重复元组时, 也会修正单元之间的评估结果, 属性层级上的合并与删除, 会同时影响异常检测与重复检测的结果. 并且, 对于某模块已完成检测的元组, 如果经由其他模块发生变化, 可实时更新结果, 不同模块之间的结果是共通的. 相比于传统方法的集成往往存在检测一致性问题, 即方法使用的先后会对结果造成较大影响, 且已造成的错误影响无法被消除, 但又由于底层设计不同无法并行使用, 本文的联合优化模型能够很好地结合不同检测模块, 具有更好的检测性能.

4 实 验

实验采用了多个领域下的真实数据集与合成数据集, 包括医院信息真实数据集、体检信息真实数据集、体检信息合成数据集以及地址信息合成数据集. 实验结果的检测指标包括错误检测的准确率、召回率、$F1$ 值. 对比方法包括面向单独错误类型的异常检测方法、重复值检测方法、缺失值检测方法、面向多错误类型的整体检测方法以及单类型错误的集成方法. 最后, 通过本文方法的消融实验证明了模型的有效性.

4.1 数据集

为了验证本文方法的性能, 我们采用了不同领域下的多个数据集进行了实验. HOSP 数据集为美国卫生部门公布的真实医院数据集, 其中 2 万条字符型数据, 由 10 个属性组成, 已知 12 个不同的约束规则. UIS 数据集为合成的地址信息数据集, 由 Mauricio 制作的地址生成器生成, 其中数据为随机生成的美国邮寄地址的列表, 包含有 2 万条字符型数据, 由 11 个属性组成, 已知 6 个不同的约束规则. Wdbc 是 UCI 资源库中公开的体检数据集, 包含有 31 个属性, 共 570 条数值型数据, 约束规则包括多个空值约束与数值区间约束, 属性之间无关联规则. JB1+JB2 为两个医疗数据集合成的集成数据集, 包含 10 个属性, 共 1 000 条数据, 数据类型包括字符、数值、时间等多种, 已知 10 个不同的约束规则. 其中, HOSP 和 UIS 数据集是用以验证数据清洗质量的 benchmark 数据集, 用以判别常规的数据检测能力. Wdbc 相比于前者, 具有更多的属性与更加复杂的数据类型, 用以对比各种方法在更复杂情况下的性能变化. JB1+JB2 则通过合并两个不同来源的医疗数据集, 模拟现

实中的多源集成数据集. 表 1 展示了各数据集的数据类型与其对应规则的类型与样例.

表 1 实验数据集

数据集名称	数据集类型	类别	规则类型	规则样例
HOSP	医院信息	真实数据	FD, CFD, DC	*PhoneNumber*→*County*; *Measure_ID*→*Measure_Name*
UIS	地址信息	合成数据	FD, CFD	*Zip*→*State*, *City*;
Wdbc	体检数据	真实数据	Check	*PID* should not be *null*;
JB1+JB2	集成医疗数据	合成数据	FD, CFD, DC, Check	空腹血糖≥0.7→糖尿病; 患者编号不能为空;

此外, 对于错误数据, 我们采用数据清洗领域中常用的错误生成方法[27], 即将上述原始数据作为正确数据, 通过随机替换数据、增加数据、删除数据作为错误数据. 在进行异常检测时, 生成的错误包括拼写错误、字符串替换、数值异常等; 重复检测时, 生成的重复数据包括复制产生的元组与对元组单元进行同义词替换产生的代表同一实体的元组; 缺失检测时, 对数据中的有效属性进行随机破碎, 记为缺失数据. 在进行单一错误类型检测时, 不生成其他类型的错误数据. 在进行整体检测时, 各种类别的错误占比相同, 整体错误率为 20%.

4.2 评价指标

错误检测可根据实际情况与检测结果将实验结果分为以下 4 类.

- True positives (TP): 实际为错误数据且算法检测为数据的实例数.
- False positives (FP): 实际为正常但算法检测为错误的实例数.
- False negatives (FN): 实际为错误数据但算法检测为正常的实例数.
- True negatives (TN): 实际为正常数据且算法检测也为正常的实例数.

以准确率(precision)、召回率(recall)和 *F*1 值(*F*1-*score*)作为主要评价指标, 验证方法的有效性. 计算方法如公式(6)–(8)所示.

$$Precision = \frac{TP}{TP + FP} \tag{6}$$

$$Recall = \frac{TP}{TP + FN} \tag{7}$$

$$F1\text{-}score = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{8}$$

4.3 对比方法

(1) KATARA、dBoost、TAMR、Null detect

首先, 我们将本文方法的各个模块与多种单类型错误检测进行性能比较. 针对异常值检测, 我们对比 KATARA[28]和 dBoost 两种经典方法, 其中, KATARA 是一个模式发现与检测方法, 利用知识库、约束规则和数据关联关系进行异常检测; dBoost 是一个集成的离群点检测框架, 通过将直方图、高斯和多元高斯混合等多个最广泛应用的离群点检测算法集成使用, 检测数据中的异常值. 针对重复值检测, 我们对比一个工业上实用的重复数据检测方法 TAMR[29]. 它通过结合相似度度量和专家知识匹配重复记录, 进行重复值检测. 针对缺失值检测, 我们对比直接检测空值的方法 Null detect.

(2) Union all、NADEEF

接下来, 我们将本文的全面检测方法与现有的多类型错误检测的方法进行对比. 首先, 我们将公式(1)中的单类型错误检测方法联合使用检测多种类型错误, 记为 Union. 此外, 我们还对比了 NADEEF, 它是一个可扩展的、通用的数据清理平台, 集成了如 Holistic 等多种知名检测方法, 通过用户指定多种类型的数据质量规则检测对应的多类型错误.

(3) 本文方法的消融实验

最后, 针对本文方法的多个视角, 我们进行消融实验, 分别去除属性语义、单元语义、元组语义、整体语

义视角和联合优化的部分, 在证明模型有效性的同时, 分析各项操作对检测性能提升的贡献.

4.4 实验结果与分析

首先, 我们将本文方法的异常检测模块 Model_O、重复值检测模块 Model_D 和缺失值检测模块 Model_M 分别与 KATARA、dBoost、TAMR、Null detect 这些单类型的错误检测方法进行性能比较. 对于每种错误检测方法, 我们都给定相同的约束规则和统计信息作为条件. 重复多次实验, 选择 *F*1 值最高的结果作为最终结果, 其准确率(*P*)、召回率(*R*)和 *F*1 值(*F*1)见表 2.

表 2 不同错误类型下错误检测性能比较

Algorithm		HOSP			UIS			Wdbc			JB1+JB2		
		P	*R*	*F*1	*P*	*R*	*F*1	*P*	*R*	*F*1	*P*	*R*	*F*1
Outlier detection	Model_O	1.0	0.68	**0.81**	0.85	0.43	**0.57**	0.87	0.42	0.57	0.94	0.55	**0.69**
	KATARA	0.96	0.67	0.79	0.77	0.25	0.48	0.86	0.43	0.57	0.81	0.31	0.57
	dBoost	0.49	0.18	0.26	0.31	0.17	0.22	0.82	0.71	**0.76**	0.56	0.45	0.50
Duplicate detection	Model_D	0.95	0.88	**0.91**	0.92	0.85	**0.88**	0.93	1.0	**0.96**	0.93	0.92	**0.93**
	TAMR	0.89	0.52	0.66	0.91	0.27	0.42	0.93	1.0	0.96	0.91	0.37	0.53
Missing value detection	Model_M	1.0	1.0	1.0	1.0	1.0	1.0	0.94	1.0	**0.96**	0.96	1.0	**0.98**
	Null detect	1.0	1.0	1.0	1.0	1.0	1.0	0.61	1.0	0.76	0.55	1.0	0.71

从表 2 的结果可以看出:

- 对于异常检测, 本文的方法和 KATARA 相比, 在 HOSP 和 Wdbc 上表现相近, 而在 UIS 和 JB1+JB2 数据集上更优. 这是由于 HOSP 和 Wdbc 分别是较为规范的信息数据和数值型数据, 语义扩充对该数据下的规则影响较小. 而 UIS 和 JB1+JB2 中存在着大量约束规则没有直接覆盖的数据, 语义扩充大大提升了规则的覆盖率, 同时也减少了规则死板导致的误判, 进而提升了检测方法的准确率与召回率.
- 与 dBoost 相比, 本文的方法仅在 Wdbc 上表现略低, 在 HOSP、UIS 和 JB1+JB2 上均表现更优. 这是由于 dBoost 采用的离群值检测方法对于体现指标这类存在正常范围的数值型数据非常适用, 但面对如字符型关系数据和医疗诊断的文本型数据等其他数据类型表现不佳.

对于重复检测, 本文的方法在 Wdbc 上和 TAMR 方法表现相近, 在 HOSP、UIS 和 JB1+JB2 上均高于 TAMR 方法. 这是因为 Wdbc 中数据为体验指标, 语义匹配作用较小, 而其他数据集中存在大量关系型数据. 而由于 TAMR 仅从模式的角度构建约束规则, 其检测标准较为死板, 对于有书写差异但实际相同的元组难以区分为不同数据. 因此, 在关系型数据集上, 该方法虽然准确率较高, 但召回率很低. 而本文的方法通过多维度语义匹配, 能够大幅度提升对重复数据识别的精度, 从而提升了检测的召回率.

对于缺失检测, 由于 HOSP 和 UIS 分别是医院信息和地址数据, 其属性不存在近似的情况, 且均为重要属性, 可以认为空值即为缺失值. 因此, 在该组数据上, 准确率与召回率均为 1. 而对于数据集 Wdbc 和 JB1+JB2, 本文的方法优于空值检测的方法. 这是因为 Wdbc 和 JB1+JB2 中包含多项体检指标, 其中部分指标为非关键数据. 因此, 直接判断空值为缺失会存在误判的情况. 此外, JB1+JB2 的集成数据中还存在着同义属性, 本文的方法通过属性融合与属性关键性判断, 能够大幅度降低误判的情况, 从而提升了检测的准确率.

综上, 可以看出不同的错误检测方法适合于不同的情况. 这些方法通常在部分数据集上表现良好, 但面对复杂的现实情况, 存在较大的局限性. 而本文的模型从宏观上结合了模式和语义, 又从不同维度对数据进行了分析与检测, 大大提升了检测的泛用性, 面对不同领域的多种数据集均有良好的表现.

接下来, 我们将 KATARA、dBoost、TAMR 和 Null detect 这 4 种方法联合使用, 并与本文的全面错误检测方法 CEDM 进行不同数据集上的性能对比, 结果如图 6 所示.

从图 6 可以看出, 对于 Union 方法, 尽管通过联合使用多种错误检测方法提升了错误检测的查全率, 但由于不同检测方法侧重的方向不同, 对于不同领域数据的适用性也不同, 但只要某一个检测方法产生了误判, 其产生的错误就会保留下来. 经实验, 通过只将被多种检测方法所标记且被标记比例超过一定阈值的错误保留下来, 能够提升其准确率, 但与此同时, 召回率也会降低, 最终无法获得 *F*1 值更优的结果. Abedjan 等人[21]同样证明了该结果.

而对于 NADEEF 方法，尽管其能够通过主键约束保证数据唯一性、通过 FD、CFD 等约束保证数据一致性、通过非空约束保证数据完整性，然而基于这些函数依赖进行错误检测，只能检测到模式上的违规，这种方法对于具有较强约束关系的 HOSP 数据集有较好的效果，但对于约束不全的 UIS 数据集、缺少语义关联的体检数据 Wdbc 和包含多种类型数据的集成数据 JB1+JB2，都难以取得良好的结果.

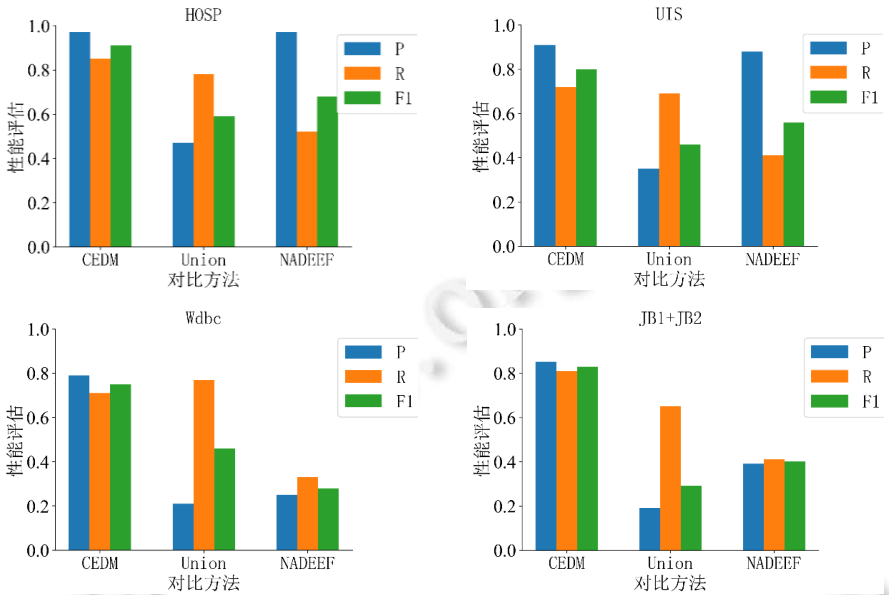


图 6　多类型错误检测方法性能比较

根据对比结果可以认为，本文的多类型错误全面检测方法 CEDM 优于多种单类型错误检测方法联合 Union 和 NADEEF 方法.

最后，我们进行本文方法的消融实验，分别去除属性语义、单元语义、元组语义、联合优化部分和整体语义视角，在证明模型有效性的同时，分析各项操作对检测性能提升的贡献. 实验结果见表 3.

表 3　不同数据集下的消融实验结果

Algorithm	HOSP			UIS			Wdbc			JB1+JB2		
	P	*R*	*F*1	*P*	*R*	*F*1	*P*	*R*	*F*1	*P*	*R*	*F*1
CEDM	0.97	0.85	0.91	0.91	0.72	0.80	0.79	0.71	0.75	0.85	0.81	0.83
去除属性语义	0.97	0.85	0.91	0.91	0.72	0.80	0.79	0.71	0.75	0.59	0.65	0.62
去除单元语义	0.91	0.67	0.77	0.68	0.59	0.63	0.79	0.71	0.75	0.69	0.66	0.67
去除元组语义	0.95	0.79	0.86	0.88	0.63	0.73	0.79	0.71	0.75	0.78	0.71	0.74
去除联合优化	0.91	0.78	0.84	0.83	0.59	0.69	0.68	0.62	0.65	0.68	0.63	0.65
去除整体语义视角	0.87	0.63	0.73	0.61	0.52	0.56	0.79	0.71	0.75	0.52	0.49	0.50

从表 3 中可以看出，不同维度下的数据语义在不同情况下贡献不同. 其中，贡献最大的是单元语义. 去除掉单元语义后，除数值型体检数据集外，其他数据上检测性能都有明显下降. 这是由于单元语义捕获对于异常检测、重复检测都有重大作用，是扩展检测规则的关键部分. 属性语义也是多个数据上影响检测性能的因素，但其贡献主要作用于重复检测，作用略小于单元语义. 属性语义对于独立的数据影响较小，去掉属性语义后，本文所选的数据集中只有集成数据集 JB1+JB2 检测结果存在变化. 这是由于其他数据集中属性之间相似度较少，同义属性往往存在于集成数据当中. 但对于存在同义属性的数据集，属性语义的影响甚至超过了单元语义. 此外，联合优化贡献也具有较大贡献，其优化方法对所有数据集均有影响. 最后，去除整体语义的变化展示了模式与语义相结合的有效性.

综合以上实验可以看出，本文提出的 CEDM 与现有方法相比，具有更高的准确率与召回率；同时，面对

现实情况中不同领域下的复杂数据类型与错误类型, 有着更好的泛用性.

5 总 结

错误检测是提升数据质量、进行有效数据管控的前提. 而现有的错误检测方法或是受限于知识与人力, 检测规则死板且覆盖率不足; 或者只能针对某些特定的领域或错误类型进行针对性检测, 不具备泛用性. 为此, 本文提出了一个多视角的多类型错误检测模型 CEDM. 通过结合模型与语义扩展检测规则, 通过分析与检测属性、单元、元组多个维度上的信息, 实现多类型错误的检测, 通过多种数据类型的不同策略, 使模型具有更高的泛用性, 并通过多模块的联合优化进一步提升了检测的性能. 经实验, 本文的方法在面对不同领域的多种错误类型时都有良好的表现, 优于现有错误检测模型.

References:

[1] Ilyas IF, Chu X. Trends in cleaning relational data: Consistency and deduplication. Foundations and Trends in Databases, 2015, 5(4): 281−393.

[2] 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 with English abstract). http://www.jos.org.cn/1000-9825/5801.htm [doi: 10.13328/j.cnki.jos.005801]

[3] Rekatsinas T, Chu X , Ilyas IF, *et al*. HoloClean: Holistic data repairs with probabilistic inference. Proc. of the VLDB Endowment, 2017, 10(11): 1190−1201.

[4] Chu X, Ilyas IF, Krishnan S, *et al*. Data cleaning: Overview and emerging challenges. In: Proc. of the Int'l Conf. on Management of Data. 2016. 2201−2206.

[5] Bohannon P, Fan W, Geerts F, *et al*. Conditional functional dependencies for data cleaning. In: Proc. of the IEEE 23rd Int'l Conf. on Data Engineering. IEEE, 2007. 746−755.

[6] Chu X, Ilyas IF, Papotti P. Discovering denial constraints. Proc. of the VLDB Endowment, 2013, 6(13): 1498−1509.

[7] Fan W, Jia X, Li J, *et al*. Reasoning about record matching rules. Proc. of the VLDB Endowment, 2009, 2(1): 407−418.

[8] Heidari A, McGrath J, Ilyas IF, *et al*. HoloDetect: Few-shot learning for error detection. In: Proc. of the Int'l Conf. on Management of Data. 2019. 829−846.

[9] Rao AR, Garai S, Clarke D, *et al*. A system for exploring big data: An iterative *k*-means searchlight for outlier detection on open health data. In: Proc. of the Int'l Joint Conf. on Neural Networks. IEEE, 2018. 1−8.

[10] Yan Y, Cao L, Kulhman C, *et al*. Distributed local outlier detection in big data. In: Proc. of the 23rd ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. 2017. 1225−1234.

[11] Mariet Z, Harding R, Madden S. Outlier detection in heterogeneous datasets using automatic tuple expansion. 2016. https://dspace.mit.edu/bitstream/handle/1721.1/101150/MIT-CSAIL-TR-2016-002.pdf?sequence=1&isAllowed=y

[12] Kushagra S, Ben-David S, Ilyas I. Semi-supervised clustering for de-duplication. In: Proc. of the 22nd Int'l Conf. on Artificial Intelligence and Statistics. PMLR, 2019. 1659−1667.

[13] Kushagra S, Saxena H, Ilyas IF, *et al*. A semi-supervised framework of clustering selection for de-duplication. In: Proc. of the IEEE 35th Int'l Conf. on Data Engineering. IEEE, 2019. 208−219.

[14] Li L, Li J, Gao H. Rule-based method for entity resolution. IEEE Trans. on Knowledge and Data Engineering, 2014, 27(1): 250−263.

[15] Schürle J. A method for consideration of conditional dependencies in the Fellegi and Sunter model of record linkage. Statisti cal Papers, 2005, 46(3): 433−449.

[16] Bilenko M, Mooney RJ. Adaptive duplicate detection using learnable string similarity measures. In: Proc. of the 9th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. 2003. 39−48.

[17] Qahtan AA, Elmagarmid AK, Ouzzani M, *et al*. FAHES: Detecting disguised missing values. In: Proc. of the 34th IEEE Int'l Conf. on Data Engineering. 2018. 1609−1612.

[18] Chu X, Ilyas IF, Papotti P. Holistic data cleaning: Putting violations into context. In: Proc. of the IEEE 29th Int'l Conf. on Data Engineering. IEEE, 2013. 458−469.

[19] Shiue YT, Huang HH, Chen HH. Detection of Chinese word usage errors for non-native Chinese learners with bidirectional LSTM. In: Proc. of the 55th Annual Meeting of the Association for Computational Linguistics. 2017. 404−410.

[20] Li XN, Cai Y, Zhu WH. Power data cleaning method based on isolation forest and LSTM neural network. In: Proc. of the Int'l Conf. on Cloud Computing and Security. Cham: Springer, 2018. 539−550.

[21] Wang Y, Feng C, Chen L, *et al*. User identity linkage across social networks via linked heterogeneous network embedding. World Wide Web, 2019, 22(6): 2611−2632.

[22] Dallachiesa M, Ebaid A, Eldawy A, *et al*. NADEEF: A commodity data cleaning system. In: Proc. of the ACM SIGMOD Int'l Conf. on Management of Data. 2013. 541−552.

[23] Abedjan Z, Chu X, Deng D, *et al*. Detecting data errors: Where are we and what needs to be done? Proc. of the VLDB Endowment, 2016, 9(12): 993−1004.

[24] Song J, He Y. Auto-validate: Unsupervised data validation using data-domain patterns inferred from data lakes. In: Proc. of the Int'l Conf. on Management of Data. 2021. 1678−1691.

[25] Wang P, He Y. Uni-detect: A unified approach to automated error detection in tables. In: Proc. of the Int'l Conf. on Management of Data. 2019. 811−828.

[26] Ding XO, Wang HZ, Zhang XY, Li JZ, Gao H. Association relationships study of multi-dimensional data quality. Ruan Jian Xue Bao/Journal of Software, 2016, 27(7): 1626−1644 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/5040.htm [doi: 10.13328/j.cnki.jos.005040]

[27] Fan W, Li J, Ma S, *et al*. Towards certain fixes with editing rules and master data. The VLDB Journal, 2012, 21(2): 213−238.

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

[29] Stonebraker M, Bruckner D, Ilyas IF, *et al*. Data curation at scale: The data tamer system. In: Proc. of the 6th Biennial Conf. on Innovative Data Systems Research. 2013. https://cs.brown.edu/courses/csci2270/archives/2017/papers/data-tamer.pdf

附中文参考文献:

[2] 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 [doi: 10.13328/j.cnki.jos.005801]

[26] 丁小欧, 王宏志, 张笑影, 李建中, 高宏. 数据质量多种性质的关联关系研究. 软件学报, 2016, 27(7): 1626−1644. http://www.jos.org.cn/1000-9825/5040.htm [doi: 10.13328/j.cnki.jos.005040]

彭锦峰(1992－), 男, 博士生, CCF 学生会员, 主要研究领域为数据质量, 人工智能.

寇月(1980－), 女, 博士, 副教授, CCF 高级会员, 主要研究领域为推荐系统, 实体识别.

申德荣(1964－), 女, 博士, 教授, 博士生导师, CCF 高级会员, 主要研究领域为 Web 数据处理, 分布式数据库.

聂铁铮(1980－), 男, 博士, 副教授, CCF 高级会员, 主要研究领域为数据质量, 数据集成.