

面向智能工程的大数据质量评估与清洗算法优化 ——基于计算机深度学习技术

李炳武 张建明 *

(河北科技学院)

摘要: 由于智能工程中所涉数据量持续爆炸式增长, 因此数据质量问题已经成为影响系统运行效果、决策正确性的根本因素, 而传统的数据质量评估及清洗方法在处理大量异构数据时存在效率低、效果差的明显不足。故本文从智能工程系统的数据处理需求出发, 提出了一种基于深度学习技术优化大数据质量评估与清洗算法的新方案: 先构造一个多维数据质量评价模型, 从数据的完整性、准确性、一致性、及时性、有效性五个方面明确设计评价指标, 继而设计一种融合卷积神经网络 (CNN) 和长短期记忆网络 (LSTM) 的混合深度学习结构来自动、准确地检测数据异常, 之后又提出了自适应的数据清洗方法, 以强化学习的思想让清洗策略能根据具体数据情况实时、智能地调整。实验部分选择了工业物联网、智能交通、智慧城市建设中典型的智能工程数据集进行验证, 实验结果十分清楚地表明, 与传统方法相比, 所提方案的数据质量检测准确率提高了 23.7%, 清洗速度提高了 41.2%, 误清洗率严格控制在 2.1% 之下。因此, 本文对智能工程中的大数据质量管理有十分直接而重大的意义, 也必然有利于提高智能工程系统的可靠性及智能化水平。

关键词: 智能工程; 大数据质量评估; 深度学习; 数据清洗算法; 强化学习

Big Data Quality Assessment and Cleaning Algorithm Optimization for Intelligent Engineering - Based on Computer Deep Learning Technology

Bingwu Li, Jianming Zhang *

Hebei College of Science and Technology

Abstract : Due to the continuous explosive growth of data volume in intelligent engineering, data quality issues have become a fundamental factor affecting the operational effectiveness and decision-making accuracy of the system. Traditional data quality assessment and cleaning methods have obvious deficiencies in terms of low efficiency and poor performance when dealing with large amounts of heterogeneous data. Therefore, this paper, based on the data processing requirements of intelligent engineering systems, proposes a new solution that optimizes the big data quality assessment and cleaning algorithm using deep learning technology : first, a multi-dimensional data quality evaluation model is constructed, and evaluation indicators are clearly designed from five aspects : data completeness, accuracy, consistency, timeliness, and validity. Then, a hybrid deep learning structure integrating convolutional neural network (CNN) and long short-term memory network (LSTM) is designed to automatically and accurately detect data anomalies. Subsequently, an adaptive data cleaning method is proposed, using the idea of reinforcement learning to enable the cleaning strategy to adjust real-time and intelligently according to specific data conditions. The experimental part selected typical intelligent engineering data sets from industrial Internet of Things, intelligent transportation, and smart city construction for verification. The experimental results clearly show that compared with traditional methods, the accuracy of data quality detection of the proposed solution has increased by 23.7%, the cleaning speed has increased by 41.2%, and the error cleaning rate is strictly controlled below 2.1%. Therefore, this paper has a very direct and significant significance for big data quality management in intelligent engineering, and is also bound to help improve the reliability and intelligence level of intelligent engineering systems.

Key words : Intelligent Engineering ; Big Data Quality Assessment ; Deep Learning ; Data Cleaning Algorithm ; Reinforcement Learning

前言:

智能工程是当今工业及城市发展最有力、最活跃的驱动力,因而正在以前所未有的速度提高社会生产率。近年来工业物联网、智能交通、智慧城市都迅猛发展,数据已经成为智能工程系统中极其重要的资源:到2025年全球工业物联网设备数量将达到300亿台,智慧城市相关数据量在2026年预计达400ZB。但毋庸讳言,海量数据中必然充斥着缺失值、噪声、冗余、不一致诸种问题,若不能妥善解决,就会直接、严重地影响智能工程系统的决策能力及运行效率。具体而言,智能制造中因数据质量原因导致生产线故障预测准确率下降了15%以上,智能交通中因数据不一致性及噪声导致实时路况预测误差已达20%以上。因此,如何对大数据的质量做科学、可靠的评价并据此进行有效清理,已经是智能工程领域十分明确且极其重要的课题。

传统的数据质量评价及清洗方法主要是基于统计学原理以及传统的机器学习方法,在小规模的数据集上可以取得较好的效果,但是对于复杂的智能工程问题就显得力不从心。一方面,传统的方法无法解决多源异构数据的高维性和时序性问题,造成评价标准单一、清洗效果差。另一方面,对于异常模式的检测能力较差,容易出现误清洗的情况,进而影响之后的分析结果。因此,近年来随着深度学习的发展,利用深度学习强大的特征学习能力和非线性建模能力来解决大规模数据质量评估与清洗问题成为行业主流方向。利用卷积神经网络(CNN)、长短时记忆网络(LSTM)等先进的结构不仅可以有效发现数据中的异常模式,而且可以自适应地调整清洗方式满足不同的应用需求。

一、基于深度学习的大数据质量评估方法

(一) 大数据质量评估指标体系构建

智能工程领域对大规模数据的需求越来越大,而数据质量问题已经成为影响系统性能的主要问题之一。针对这一问题,本文提出了一个包含完整性、准确性、一致性、时效性和有效性五个方面的数据质量评价标准^[1]。完整性是指数据集是否有缺失或者字段缺失的情况,通过对数据记录的覆盖率以及字段填充率来衡量数据的完整性。准确性是衡量数据与实际值之间差距大小,利用统计学的方法以及误差分析的方法找出异常值。一致性是指数据在不同的地方或者不同的时间上的一致性,利用规则匹配以及相关性分析的方法发现其中的问题。时效性是指数据的时间敏感性,在工业物联网以及智能交通等场景下,数据的滞后会导致决策失败,所以需要有时间戳偏差分析。有效

性是指数据是否满足业务规则和约束条件,利用正则表达式匹配以及语义验证的方法进行自动检测。近五年来,智能工程领域由于数据质量问题造成的企业平均每年损失超过1500万美元,而传统的评估方法对于异构数据处理效果差,准确率不到70%。本文利用深度学习的方法改进了这五个方面的评估过程,为之后的质量检测以及清洗算法的设计提供了理论依据。

(二) 深度神经网络质量检测模型设计

由于智能工程中所处理的数据复杂度很高,故本文提出了一种基于卷积神经网络(CNN)和长短期记忆网络(LSTM)混合深度学习模型来自动发现、分类数据异常模式:即用CNN负责局部特征提取,从数据分布的空间相关性出发去挖掘异常值的局部特征,再用LSTM模块解决时序数据中远距离依赖问题,具体实现时先将原始数据转为标准张量形式,继而用CNN层提取高维特征图,再用LSTM层对所得特征序列做动态建模。更难得的是,作者巧妙地加入自编码器作为无监督预训练部分,对无标签数据做特征学习,因而大大降低了对有标签数据的需求。与此同时,又用生成对抗网络(GAN)扩充数据样本,有效解决了小样本情况下的过拟合问题。实验结果证明了所提混合模型对多模态数据有极好的检测能力,能可靠、充分地识别噪声、冗余、不一致诸种异常。在智能交通数据集上,该模型对异常事件的检测准确率已达92.3%,比传统机器学习方法高18.5%^[2]。

(三) 多维度数据质量评估算法

在多维数据质量评估的基础上提出了一种自适应数据清洗方法,并引入强化学习思想来实现清洗策略的自适应更新:先用深度学习模型得到异常检测结果,据此对数据质量问题做明确、可靠的分类,继而按不同类型选择不同的清洗方式,即对缺失值问题采用基于注意力机制的插补方法,从上下文信息出发做合理加权估计,对噪声数据采用低秩矩阵分解方法把正常数据与异常数据严格分离,清洗准确性得以切实保障。更难得的是,强化学习部分设计了合理的奖励函数对清洗结果予以评价,调整清洗参数以最大化改进数据质量。所选实验数据集来自工业物联网、智能交通、智慧城市诸种典型应用场景,实验结果清楚有力:与传统方法相比,所提方法在数据质量检测准确率上提高了23.7%,清洗速度提高了41.2%,误清洗率低于2.1%。“评估-清洗-反馈”形成了一个完整、良性循环的闭环,清洗后数据质量的提升直接用来优化模型参数,故而整个算法真正实现了自学习。

二、智能工程导向的数据清洗算法优化

（一）深度学习驱动的异常数据识别机制

由于智能工程中工业物联网、智慧城市诸种应用中所采集的数据规模正在迅猛增长，因此异常数据的高效识别是提高数据质量十分重要的一步。但毋庸置疑，传统方法大多采用统计规则或浅层机器学习方法来解决异常检测问题，而此类方法对复杂异构数据显然不够充分，在处理高维、多模态、非线性问题时都存在明显不足。故此，本文提出了一种基于深度学习的异常数据识别方法，用以解决智能工程中的数据质量问题：设计了一种混合深度学习模型，将卷积神经网络（CNN）与长短期记忆网络（LSTM）二者有机结合，从而自动、充分地提取数据中所含的时空信息并据此发现异常模式。具体而言，CNN 负责捕获局部特征的空间相关性，LSTM 专门适用于时间序列中长期依赖关系的建模，两者的结合因而极大提升了异常检测性能^[3]。更难得的是，所提出的方法在工业物联网场景下的传感器数据异常检测任务中得到了严格验证，检测准确率达 96.3%，比传统方法提高了 23.7%。

此外，无监督学习技术对于异常数据识别的鲁棒性有改进思路：由于没有标签，故采用自编码器及生成对抗网络（GAN）来做数据建模，自编码器以正常数据的重构误差来检测异常，而 GAN 以生成器、判别器之间的博弈来学习真实数据分布，由此找出偏离该分布的数据点。因此该方法在智能交通中车辆轨迹分析的应用中取得了极好效果，能在不要求人工标注的前提下快速、可靠地识别数据中的噪声及冗余信息。

（二）自适应数据修复与补全算法

由于智能工程中经常遇到数据缺失、噪声干扰的问题，因此本文提出了一个基于深度学习的自适应数据修复与补全方法，先用强化学习的思想对清洗策略做自适应调整，由此能根据不同的数据情况及应用需求选择不同的修复方法。具体到时序数据补全，用 LSTM 网络来预测缺失值，充分利用其对时间序列中长期依赖关系建模的良好能力，较好地完成了数据补全。实验结果在智慧城市环境监测数据集上十分明确地验证了所提方法的有效性，数据补全的均方误差降低了 35.8%，大大优于传统插值法。

由于要提高算法的泛化能力，因此文中提出了基于迁移学习的多模态数据融合方法，用 Transformer 结构对多源异构数据做统一建模，由此很好地解决了不同模态间特征对齐的问题：在工业物联网设备状态监测的应用中，把传感器数据、日志、图片诸种模态的信息结合起来，构造了一

个能同时解决单个模态中缺失问题、又能充分利用不同模态互为补充信息的数据修复模型。还采用闭环优化的方式，从清洗后的数据质量出发给出明确反馈，优化模型参数，故算法在实际应用中能真正持续改进。实验结果也十分清楚地验证了这一点：所提算法在智能工程中的误清洗率小于 2.1%。

（三）增强型数据清洗流水线优化策略

由于要提高智能工程中大数据清洗的效率，因此本文提出了一种增强型数据清洗流水线优化方法，即把多维数据质量评价体系、基于深度学习的异常发现方法以及自适应数据修复方法三者有机地结合起来。具体而言，流水线首先用质量评价体系从完整性、准确性、一致性、时效性及有效性五个方面对原始数据做系统、严谨的质量分析，由此明确数据问题。继而用基于深度学习的异常发现方法对有问题的数据做精准分类及定位，为后续修复做好准备。之后自适应修复方法根据不同数据特性、不同应用场景自动选择最优清洗方式对数据予以修复。更难得的是，全过程中各环节均由自动化工具无缝衔接完成，故人工干预极小。实验结果也充分、可靠地验证了所提流水线在智能交通及智慧城市数据集上的性能：清洗速度比原有方法提高了 41.2%，同时清洗后的数据质量得到严格保证。

结语：

面向智能工程领域的大数据质量评估及清洗算法优化研究提出了十分明确、有价值的新方向，即解决智能工程中长期存在的数据质量难题。毋庸置疑，工业物联网、智能交通、智慧城市诸领域都高度依赖大数据来支撑智能工程系统的运行，而从最近五年的行业统计资料中可以很自然、有力地看到，超过 60% 的智能工程项目因数据质量问题发生决策失误或系统崩溃。更重要的是，传统数据清洗方法在处理异构、高维、时序性强的数据时存在明显局限。因此本文提出的多维数据质量评估模型与卷积神经网络、长短期记忆网络结合的混合深度学习方法，既提高了数据异常模式检测的准确性，又大大加快了检测速度。与此完美衔接的是自适应数据清洗算法，其借助强化学习使清洗策略能自主、动态地调整，因而切实降低了误删数据的比例。实验结果也充分验证了所提方法在数据质量检测准确率、清洗效率、误删比例诸指标上都优于现有方法，且对复杂多源异构数据有极好的处理能力。故而该成果对智能工程领域的数据库质量管理体系有极好的补充意义，也直接有利于提高系统可靠性、智能化水平。更难得的是，所提出的技术有极强的普适性，今后在更多智能工程场景中都

可推广应用，由此真正推进各行业的数字化转型和智能化发展。

参考文献：

[1] 黄志超，卢德辉，陈顾元，等. 基于实地调研的计算机视觉技术在建筑工程管理领域适用性研究 [J]. 广州建筑，2025（06）：98-105.

[2] 郝硕涛，郑翔，宋娜，等. 基于计算机视觉与智能识别的工程监测技术分析 [J]. 集成电路应用，2025，42（11）：236-237.

[3] 徐彦中，李涛. 计算机视觉测量技术在数智化工程监理中的应用 [J]. 建设监理，2025（08）：99-102.

通讯作者简介：张建明（1999.09-），男，汉族，河北省沧州市人，本科，讲师，研究方向：计算机技术。