DOI:10.19344/j.cnki.issn1671-5276.2022.03.024

基于神经网络的工业时序数据质量管理方法

尤祺,袁堂晓,汪惠芬
(南京理工大学 机械工程学院,江苏 南京 210094)

摘　要:针对当前工业时序数据质量管理存在的缺乏有效管理方法、没有与工业领域知识相结合等问题,梳理数据质量问题的主要表现,引入风险评估机制以完善数据质量评价标准。在此基础上给出了工业时序数据质量管理方法,主要流程包括:多维度评价标准体系下的数据质量评价过程、基于数据典型应用场景和业务需求的数据质量分析过程以及知识与数据混合驱动的数据质量提升过程。提出一种基于 LSTM 神经网络的数据质量分析方法,通过实际数据集验证了数据质量管理和提升的效果。

关键词:工业大数据;时间序列;数据质量管理;神经网络
中图分类号:TP183　　**文献标志码**:A　　**文章编号**:1671-5276(2022)03-0096-04

Industrial Time Series Data Quality Management Method Based on Neural Network

YOU Qi, YUAN Tangxiao, WANG Huifen
(School of Mechanical Engineering, Nanjing University of Science and Technology, Nanjing 210094, China)

Abstract: In view of problems such as lack of effective management approaches and failure in relating to industrial knowledge in current industrial time series data quality management, the main manifestations of data quality problems are sorted out, and risk assessment mechanisms are introduced to improve data quality evaluation. On this basis, the industrial time series data quality management method is given. The main processes include data quality evaluation process under the multi-dimensional evaluation standard system, data quality analysis process based on typical data application scenarios and business requirements, and data quality standard improvement process driven by a mixture of knowledge and data. A data quality analysis method based on LSTM neural network is proposed. The effect of data quality management and improvement is verified through actual datasets.

Keywords: industrial big data; time series; data quality management; neural network

0　引言

在积累的工业大数据中,时间序列数据是最基本和最普遍的数据形式。对工业大数据进行信息提取和价值发现,前提是要拥有可靠准确的高质量数据。然而,由于数据来源的多样性、机器设备本身的局限性、工业现场环境因素的干扰等情况,工业数据可能存在异常或缺失,致使无法满足进一步分析应用的需要。因此,建立可行的数据质量评价、检测、治理与持续改善的管理机制,是工业大数据的重要研究方向。

在数据质量管理的研究领域,国外学者更关注管理框架和管理流程的研究。WANG R Y[1]提出的全面数据质量管理方法,通过定义、测量、分析和改进 4 个阶段实现数据质量的循环管理。JEUSFELD M A 等[2]提出的数据仓库质量方法,考虑到质量概念的主观性,根据不同的使用群体提供不同类别的质量目标。BATINI C 等[3-4]提出完全数据质量方法,可以应用于结构化、半结构化和非结构化数据。国内对于数据质量管理的研究更偏重于实际应用。方幼林等[5]提出了数据仓库中数据质量的度量和评价指标,并提出了数据质量成熟度模型。杨青云等[6]基于数据可信性和可用性提出了一个数据质量评估模型。颜宏文等[7]提出了一种基于云模型的电网统计数据质量评估方法,以避免传统方法的主观随意性。袁满等[8]针对数据质量维度与框架进行了对比分析,为具体应用提供了科学依据。周艳红[9]以数据生命周期为研究视角,基于层次分析法和专家打分法建立大数据质量评估模型。

虽然国内外研究学者针对数据质量管理提出了多种方法论和框架,强调数据清洗过程的自动化和一次成功率,但在实际应用中缺乏具体的执行手段;不同领域内数据质量问题存在差异,对于工业时序数据质量管理缺乏针对性的研究;数据清洗过程过于追求通用性,没有将工业领域知识与之融合。本文针对工业时序数据特点进行分析,对数据质量评价和控制方法进行集成与改进,给出了提升工业时序数据质量的管理方法,最后通过实际数据集验证了质量管理和提升的效果。

1　工业时序数据质量问题分析

1.1　工业时序数据质量问题的主要表现

工业时序数据主要来自于工业现场的物联网络、生产

基金项目:国家自然科学基金项目(51705256);江苏省研究生科研与实践创新计划项目(SJCX20_0104)
第一作者简介:尤祺(1994—),男,江苏泰州人,硕士研究生,研究方向为工业大数据、智能优化算法等。

制造装备和各类自动化系统等采集的数据,具有来源广泛、体量大、价值密度低等特点。由于器件系统故障、现场恶劣工况等影响,数据质量问题广泛存在,主要表现在以下几个方面[10]。

1)数据失真和失准。由于工业现场复杂环境因素的影响以及设备运维保养不当、缺乏有效的管理机制等原因,可能造成各类工业运行数据出现数据失真和失准问题。

2)时间序列周期异常。当供电出现故障时,元件功率的变化会影响数据采集频率,造成时间序列周期发生短暂变化。

3)数据错列。当数据采集器出现故障或是控制器发生收录错误时,会出现部分数据与其原本属性无法对应的错列问题。

此外,常见的工业时序数据质量问题还包括数据冗余、数据误采、数据不可识别、数据缺失、数据一致性差等。

1.2　工业时序数据质量问题的风险评估

不同的数据质量问题具有不同的严重性和发生的可能性,本文为这些数据质量问题建立了风险评估矩阵,如图1所示。该评估矩阵是在综合分析各类数据质量问题的出现频次、检测和修正难度以及对后续数据分析应用造成的影响的基础上设计的。需要指出的是,风险评估矩阵中质量问题的排列顺序是基于经验和判断,可能会因为案例或应用对象的不同而略有差异。

图1　工业时序数据质量问题风险评估矩阵

1.3　工业时序数据质量的评价标准

对数据质量维度进行定义和分析,是建立数据质量评价模型的前提和基础。根据工业时序数据的特点和存在的质量问题,结合相关研究[4],本文总结了适用于工业时序数据的数据质量维度,如表1所示。其中,时效性和及时性是与时间相关的主要维度,表征了数据在有效性、更新频率和稳定性等方面的表现;风险性则是依据风险评估矩阵对数据进行评价。

表1　工业时序数据质量维度

维度	定义
准确性	描述数据库中存储的值与实际值的符合程度
完整性	描述数据不存在缺失的程度
一致性	描述数据不存在违反约束与不合语义的错误以及关联逻辑关系相容的程度
唯一性	描述数据在同一记录中拥有非重复值的程度
时效性	描述数据对于处理当前任务的有效程度
及时性	描述数据更新的及时程度
风险性	描述数据存在质量问题的严重程度

1.4　工业时序数据质量问题的解决思路

1)交互式数据清洗。原始数据中往往存在多种异常,过于追求并依靠自动分析并不能很好地解决问题,由专业人员参与决策的交互式数据清洗模式才是符合实际的努力方向。

2)持续性数据管理。过于追求完美和一次成功率往往适得其反,原有的数据质量问题解决了,还会有新的问题出现。应当把数据质量管理视为数据生命周期内的一项经常性工作。

3)领域级数据修正。在数据质量提升环节,需要将数理知识与工业领域知识深度融合,依托工业知识推理决策进行离群值和异常值的修正。

2　工业时序数据质量管理方法

针对工业时序数据的特点,结合目前的数据质量管理架构和方法,本文给出如图2所示的工业时序数据质量管理方法,从定义、评价、分析、提升和监控5个流程环节持续改善数据质量。

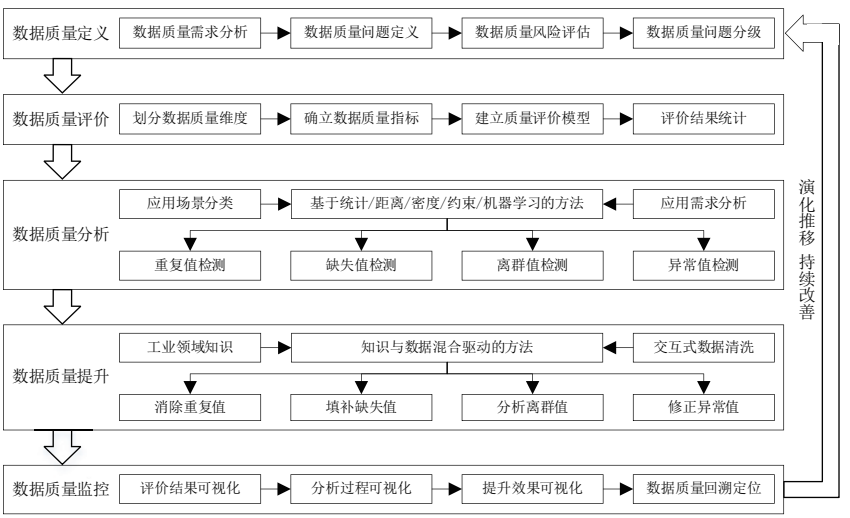

图2　工业时序数据质量管理方法

数据质量定义通过需求分析和问题定义,明确对数据的质量要求和检测标准,进行风险评估和问题分级,为评价环节提供方向和参考。数据质量评价对每个质量维度进行定义与分析,从而建立完整的评价模型,通过综合数据质量在各维度的计算值得到评价结果,评价结果是数据质量分析和提升的基础。数据质量分析针对不同特点和应用场景的工业时序数据选择合适的异常数据检测方法,在参考评价结果的基础上,检测出数据中的重复值、缺失值、离群值和异常值。数据质量提升通过与工业领域知识的深度融合,实现知识与数据混合驱动的全方位数据清洗。数据质量监控通过对各个环节的可视化呈现和质量问题的示踪定位,达到辅助决策的目的。在提升数据质量的过程中,随着时间推移和数据演化,可能会有新的数据质量问题出现,因此需要持续的数据质量管理,不断发现和解决数据中的问题。

2.1　工业时序数据质量评价

在数据质量评价过程中,数据质量维度权重的计算直接影响到评价模型的准确度以及最终的评价结果。本文提出了一种基于层次分析法和熵值法相结合的主客观组合赋权法,组合后的权重既能体现数据信息,又能反映专家意愿,兼顾了主观权重和客观权重的优点。计算过程如下。

1)基于熵值法计算客观权重:

$$a_i = \frac{1 - e_i}{\sum_{i=1}^{n}(1 - e_i)} \tag{1}$$

式中:n 为评价维度数;e_i表示第 i 个维度的熵值,计算公式为

$$e_i = -\frac{1}{\ln m}\sum_{j=1}^{m} p_{ij}\ln p_{ij} \tag{2}$$

式中:m 为待评价样本数;p_{ij}表示第 i 个维度中第 j 个样本值的比重,计算公式为

$$p_{ij} = \frac{x_{ij}}{\sum_{j=1}^{m} x_{ij}} \tag{3}$$

式中x_{ij}为第 i 个维度中第 j 个样本的数值。

2)基于层次分析法计算主观权重:

$$b_i = \frac{\left(\prod_{j=1}^{n} t_{ij}\right)^{1/n}}{\sum_{i=1}^{n}\left(\prod_{j=1}^{n} t_{ij}\right)^{1/n}} \tag{4}$$

式中t_{ij}表示维度 i 对维度 j 的重要度,使用 1~9 比率标度法进行定义。

3)对以上两种方法得出的权重计算综合权重,对于某一维度 i,其综合权重为

$$w_i = \frac{\sqrt{a_i\, b_i}}{\sum_{i=1}^{n}\sqrt{a_i\, b_i}} \tag{5}$$

最后结合数据在各维度的得分S_i,计算出该数据集的总得分:

$$S = \sum_{i=1}^{n} w_i\, S_i \tag{6}$$

2.2　工业时序数据质量分析

目前主流的时间序列异常检测方法包括:基于统计的方法、基于距离的方法、基于密度的方法、基于约束的方法以及基于机器学习的方法[11-12]。基于统计的方法根据数据在概率分布模型中的拟合情况来评估和提取时序趋势,但对于分布特征未知的数据,这种先验假设存在较大的局限性。基于距离的方法通过计算数据点之间的距离来检测孤立点,由于使用全局阈值,时间复杂度较高且不能处理不同密度区域的数据集。基于密度的方法克服了不同密度区域的数据集混合造成的检测错误,但也具有较高的时间复杂度。基于约束的方法根据数据变化的规则以及序列间的相关性来建立约束,通过约束规则来检测和修复异常点,但是对于规则变化较大的数据效果不理想。基于机器学习的方法将机器学习和深度学习的相关模型与工具应用于异常数据检测,模型复杂度更高,对于数据的预处理也有更高的要求。本文根据数据的不同应用场景及其业务需求,给出推荐的异常数据检测方法及其使用效果,见表 2。

表 2　工业时序数据异常检测方法及效果分析

应用场景	业务需求	推荐方法	使用效果
故障诊断	对异常检测的正确性要求较高	基于约束和机器学习的方法	鲁棒性更好,容错性更强
寿命预测	确保数据的准确性和完整性	基于统计和机器学习的方法	快速识别离群值并剔除异常值
设备协同	确保数据的准确性和时效性	基于距离和密度的方法	实现对动态数据的高精度检测
运维监控	对异常检测的处理速度要求较高	基于统计和密度的方法	快速识别离群值,可靠性更好

2.3　工业时序数据质量提升

原始数据在经过数据质量分析之后,除了异常数据被检测出来、正常数据被过滤出去以外,可能会出现如表 3 所示的正常数据被误测为异常的假异常以及异常数据被误测为正常的假正常的情况。此时,在算法提升效果有限的情况下,可以融入领域专家知识对以上两种情况进行纠正,以免造成数据质量的损失。

表 3　数据质量分析结果混淆矩阵

真实情况	检测结果	
	正常	异常
正常	真正常(TP)	假异常(FN)
异常	假正常(FP)	真异常(TN)

图 3 所示的是知识与数据混合驱动的数据质量提升过程。混合驱动模型包括两大部分:由专家知识的表达与约束规则界定融合成的领域知识库以及基于同类型历史数据进行特征抽取得到的数据特征库。当接收来自上游模块的异常数据时,对异常特征进行基于知识库的模式分

析和基于特征库的查找匹配,然后针对异常数据中的重复值、缺失值、离群值和异常值进行相应的处理操作,从而实现修正真异常和假正常、纠正假异常的目标。经过质量提升的数据将被抽取补充到该类数据的特征库中,使得混合驱动模型可以不断学习新的经验,以增强模型的数据处理能力。

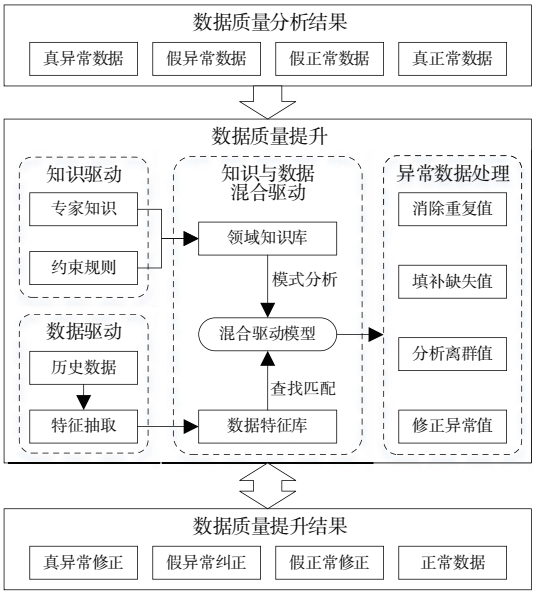

图 3　知识与数据混合驱动的数据质量提升过程

2.4　基于 LSTM 的数据质量分析方法

本文基于长短期记忆网络(long short-term memory network,LSTM)建立了工业时序数据质量分析方法。LSTM 具有控制遗忘的结构设计,非常适合处理时序任务[13]。首先基于历史数据对 LSTM 进行训练,然后利用 LSTM 进行时序数据预测,最后使用预测结果与实际数值的差值进行异常区间的判断。其中的关键步骤主要包括:

1)数据预处理。将原始数据按照公式(7)进行归一化,使处理后的数据映射到 0~1 之间。

$$X_{\text{scaled}}=\frac{X-X_{\min}}{X_{\max}-X_{\min}} \tag{7}$$

式中:X_{scaled}为归一化后的数据;X 为待处理数据;$X_{\min}$为样本的最小值;$X_{\max}$为样本的最大值。

2)确定时间步长。时间步长是 LSTM 模型的一个关键参数,会对模型运算速度和预测精度产生影响,可以结合数据量大小和模型表现确定其大小。

3)确定节点数。根据经验公式(8)和模型实际表现确定输入层和隐藏层节点数。

$$m=\sqrt{n+l}+a \tag{8}$$

式中:m 为隐藏层节点数;n 为输入层节点数;l 为输出层节点数;a 为 1~10 之间的常数。

4)确定模型其他参数。结合损失函数和观测函数随训练轮次的表现,选取合适的训练轮次,以避免过拟合与欠拟合。

训练好模型之后,对数据进行预测,并对预测值进行归一化的还原。最后,在专家知识与约束规则的界定分析下,确定异常数据点并修正。

3　数据质量提升效果分析及验证

在对数据质量管理效果进行分析验证的过程中,除了需要比较处理前后的数据质量在评价模型中的得分表现,还需要考量质量提升后的数据是否提高了分析应用的成功率。

本文以某地区的水泵系统数据集为例进行分析。该数据集由 52 组传感器数据和水泵系统状态标签数据组成,每分钟记录一组数据,共计 141 120 组数据。传感器数据记录了水泵系统的压力、温度、流量等信息,水泵系统状态标签包括正常与不正常两种状态。

选取部分传感器数据,依据领域知识和约束规则事先进行异常数据的甄别和标记,然后采用本文基于 LSTM 神经网络的方法对其进行数据质量分析与提升。图 4 所示为部分数据分析过程,图中实线表示实际值,虚线表示预测值,将二者作差并结合专家知识以确定出异常值。

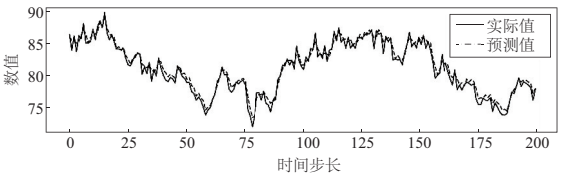

图 4　数据质量分析过程示意图

将分析结果与传统的基于密度和基于统计的方法进行对比,并设置对比项为精度(数据被正确识别的比例)、查准率(识别为正常数据中真正正常数据的比例)以及查全率(数据中正常数据被正确识别的比例)。结合表 3 对于检测结果的定义,给出各对比项的计算公式:

$$A=\frac{T_{\text{P}}+T_{\text{N}}}{T_{\text{P}}+T_{\text{N}}+F_{\text{P}}+F_{\text{N}}} \tag{9}$$

$$P=\frac{T_{\text{P}}}{T_{\text{P}}+F_{\text{P}}} \tag{10}$$

$$R=\frac{T_{\text{P}}}{T_{\text{P}}+F_{\text{N}}} \tag{11}$$

式中:A 为精度;P 为查准率;R 为查全率。

实验结果如表 4 所示。通过对比可以发现,本文的方法在精度、查准率和查全率等方面都具有不错的表现,特别是精度和查全率,相较于传统方法有了较大提升。

表 4　不同方法对数据质量分析结果的对比

单位:%

对比项	基于密度的方法	基于统计的方法	本文方法
精度	61.73	81.98	88.26
查准率	91.53	95.98	93.40
查全率	64.96	84.20	94.05

为了验证数据质量管理的效果,本文基于 Keras 搭建神经网络模型,对该水泵系统进行故障预测。如表 5 所示,经过数据质量分析与提升后,模型对于系统故障的预

(下转第 112 页)

度减小,夹气时刻大体推迟,即推进剂剩余量减小,贮箱出流的流动状态变差,最高流速、输送管口平均流速大幅度提高,出口总压恢复系数减小幅度达7.3%,流动损失增大。

3)圆盘直径对液面塌陷有显著影响。随着圆盘直径从0.5*D*增至1.5*D*后,夹气时刻得到了较大幅度的推迟,即液面塌陷延缓,推进剂剩余量减小,贮箱出口总压恢复系数减小,流动损失增大,输送管内流动差异性增大。

4)大变过载(超重)工况下夹气时刻随*H/D*的变化趋势与常过载(1*g*)工况基本相同,但大变过载有利于延迟液面发生塌陷的时间,减少液氧的剩余量,且大变过载下输送管内的流动差异性减小,输送管内的流动状况有所改善。

参考文献:

[1] 胡平信,刘国球. 液体火箭发动机的技术发展与展望[J]. 导弹与航天运载技术,1998(2):3-12.

[2] 邵业涛,邓新宇,黄兵等. 低温火箭贮箱防漩防塌装置数值模拟研究[C]//第十届全国低温工程大会暨中国航天低温专业信息网2011年度学术交流会, 兰州:[s.n.], 2011:448-452.

[3] 北京宇航系统工程研究所出流装置研制QC小组. 减少火箭不可用推进剂量[J]. 中国质量,2015(11):32-37.

[4] 孙礼杰,褚洪杰,王振剑,等. 液体火箭漩涡与塌陷现象的机理及其抑制措施[J]. 上海航天,2016,33(1):80-84,89.

[5] TAM W,DREY M,JAECKLE D Jr,et al. Design and manufacture of an oxidizer tank assembly [C]//37th Joint Propulsion Conference and Exhibit. Salt Lake City, UT, USA. Reston, Virigina:AIAA,2001.

[6] 杨魏,吴玉林,刘树红. 部分充液贮箱自由液面塌陷的数值研究[J]. 工程热物理学报,2010,31(3):423-426.

[7] 王坤. 液体火箭发动机燃料贮箱出流塌陷夹气现象的研究[D]. 北京:北京理工大学,2015.

[8] 黄晓宁,王磊,毛红威,等. 火箭升空低温推进剂出流特性仿真研究[J]. 制冷学报,2020,41(4):136-143,166.

收稿日期:2021-02-08

(上接第99页)

测准确率由80.18%提升至90.38%,模型损失(二元交叉熵)由0.1982下降至0.0206,从而证明了数据质量管理的有效性。

表5　数据质量管理效果对比验证

对比项	预测准确率/%	损失
数据质量管理前	80.18	0.198 2
数据质量管理后	90.38	0.020 6

4　结语

本文梳理了工业时序数据质量问题的主要表现,引入风险评估机制以完善数据质量评价标准,给出了工业时序数据质量管理方法。提出了一种基于LSTM神经网络的数据质量分析方法,并通过实际数据集进行了验证。后期研究需要将工业时序数据质量管理方法模块化、系统化,提高实用性,使其真正服务于工业大数据。

参考文献:

[1] WANG R Y. A product perspective on total data quality management[J]. Communications of the ACM, 1998, 41(2): 58-65.

[2] JEUSFELD M A, QUIX C, JARKE M. Design and analysis of quality information for data warehouses [C]//Conceptual Modeling - ER '98, 1998. DOI:10.1007/978-3-540-49524-6_28.

[3] BATINI C, SCANNAPIECO M. Data quality: Concepts, methodologies and techniques[M]. [S.I.]:Springer, 2006.

[4] BATINI C,CAPPIELLO C,FRANCALANCI C,et al. Methodologies for data quality assessment and improvement[J]. ACM Computing Surveys,2009,41(3):1-52.

[5] 方幼林,杨冬青,唐世渭,等. 数据仓库中数据质量控制研究[J]. 计算机工程与应用,2003,39(13):1-4.

[6] 杨青云,赵培英,杨冬青,等. 数据质量评估方法研究[J]. 计算机工程与应用,2004,40(9):3-4,15.

[7] 颜宏文,陈鹏. 基于云模型的电网统计数据质量评估方法研究[J]. 计算机应用与软件,2014,31(12):100-103.

[8] 袁满,刘峰,曾超,等. 数据质量维度与框架研究综述[J]. 吉林大学学报(信息科学版),2018,36(4):444-451.

[9] 周艳红. 基于大数据的数据质量评估方法研究[J]. 现代信息科技,2020,4(8):86-89.

[10] 苏佳轩. 面向工业大数据的高维时间序列清洗系统[D]. 哈尔滨:哈尔滨工业大学,2019.

[11] HAN J, KAMBER M, PEI J. Data mining: concepts and techniques, third edition[M]. Waltham, MA, USA: Morgan Kaufmann, 2011.

[12] 丁小欧,王宏志,于晟健. 工业时序大数据质量管理[J]. 大数据,2019,5(6):1-11.

[13] 丁盼,庞晓平,陈进. 基于长短期记忆网络的挖掘机器人视觉跟踪系统设计[J]. 机械制造与自动化,2019,48(4):145-148.

收稿日期:2021-05-23