新能源汽车监测平台数据异常识别及质量评价研究*

郝雄博[1,3]　蔡君同[1,3]　谭新治[2]　何山[2]　李昊巍[3]

（1.中汽数据（天津）有限公司，天津 300300；2.深圳供电局有限公司，深圳 518000；

3.中国工业互联网研究院，北京 100000）

【欢迎引用】郝雄博，蔡君同，谭新治，等．新能源汽车监测平台数据异常识别及质量评价研究[J]．汽车文摘，2024(12)：1-7.

【Cite this paper】HAO X B, CAI J T, TIAN X Z, et al. Identification and Quality Evaluation of Abnormal Data in the New Energy Vehicle Monitoring Platform[J]. Automotive Digest (Chinese), 2024(12): 1-7.

【摘要】为了提高新能源汽车监控数据的质量，提出了一种系统性的异常数据识别与质量评估方案。针对监控数据中的多种异常情况，设计了从数据收集到解析后的全面评估流程。该流程涵盖了数据规范、完整性、准确性、一致性和时效性等关键维度，并采用层次分析法与熵权法相结合的方式计算各维度权重。通过模糊综合评价方法，量化数据质量评分，避免了单一主观或客观因素对评估结果的影响。实证分析表明，该方案能够全面识别新能源汽车数据中的异常类型，并提供合理的质量评价结果。

关键词：新能源汽车；数据异常识别；数据评价；熵权法

中图分类号：U469.7；TP274　　**文献标志码：**A　　**DOI:** 10.19822/j.cnki.1671-6329.20230230

Identification and Quality Evaluation of Abnormal Data in the New Energy Vehicle Monitoring Platform

Hao Xiongbo[1,3], Cai Juntong[1,3], Tan Xinzhi[2], He Shan[2], Li Haowei[3]

（1. Automotive Data of China (Tianjin) Co., Ltd., Tianjin 300300; 2. Shenzhen Power Supply Co., Ltd., Shenzhen 518000;

3. China Academy of Industrial Internet, Beijing 100000）

【Abstract】To improve the quality of new energy vehicle monitoring data, a systematic abnormal data identification and quality evaluation scheme is proposed. The scheme addresses various abnormal situations in the monitoring data and designs a comprehensive evaluation process from data collection to data analysis. This process covers key dimensions such as data standardization, completeness, accuracy, consistency, and timeliness. The weights of each dimension are calculated using a combination of the Analytic Hierarchy Process (AHP) and the entropy weight method. Through the fuzzy comprehensive evaluation method, the data quality score is quantified, avoiding the influence of single subjective or objective factors on the evaluation results. Empirical analysis shows that this scheme can comprehensively identify abnormal data types in new energy vehicle monitoring and provide reasonable quality evaluation results.

Key words: New energy vehicles, Identification of abnormal data, Data evaluation, Entropy weight method

0　引言

随着新能源汽车保有量迅速增长，电动汽车的安全问题也日益凸显，行业积极采取监控数据的方法来提高新能源汽车产品的安全水平。数据识别在新能源汽车行业以及其他工程领域发挥着越来越重要的作用，但是由于设备故障和人为因素的干扰，识别的数据中往往会夹杂大量的异常数据。异常数据会影响数据的分析和处理，出现严重情况会引发安全检测系统误报甚至失灵[1]。针对异常数据识别，在车辆交通领域，王英会[2]致力于辨识交通流异常数据，从交通流缺失数据角度出发，结合阈值法和交通流机理，进一步提出了基于混沌理论的错误数据辨识方法。针对异常数据探测，付时瑞等[3]利用某型号动车组振动

*基金项目：规模化电动汽车与电网互动关键技术研究与示范应用（一期）项目（090000KK52210132）。

监控系统捕获的异常振动信息，成功识别出该型号动车组齿轮箱的异常振动故障位置。在环境治理领域，郑涛[4]运用2种统计学方法来判断废气污染源自动监控数据中的异常数据，比较了2种方法的适用性并找出更为适合的方法。在电路元件监控方面，通过传感器采集各类电器件的实时数据，黄雄波等[5]提出一种改进的时序数据流异常值检测算法，有效构建了更高效实用的监控系统。试验表明，在不增加计算成本的同时，该方法检测精度和算法的鲁棒性提升显著。在煤炭安全监控系统中，殷大发等[6]综合应用关联分析、聚类分析和时间序列等方法，通过分析数据异常波动的影响因素，制定了异常识别原则，构建了相应的异常识别模型，从而有效提高了识别效率。

异常数据的识别在各行各业得到越来越多的关注，数据质量评估同样也发挥着越来越重要的作用。数据质量直接决定了数据能否进一步分析和处理应用。目前，国内外其他领域的数据质量评价的方法和技术有很多。例如，数据质量问题是医疗保健数据模型（Clinical Data Management, CDM）发展的主要障碍。Whan[7]基于代表性数据模型CDM创建了高质量数据生成和多中心CDM质量评价方案，并且通过现有CDM质量评估系统的规则，创建了大量高级评估规则并将其纳入系统。最后，通过多家医院的数据质量进行了验证，总体错误率为0.197%。黄国彬等[8]对3种较成熟的国外科学数据质量评估框架进行了比较研究，发现3个较成熟的数据质量评估框架涉及的数据质量维度大致相同，如可信度、准确度、及时性、可访问性等。除了研究和分析国外相关质量评估框架外，国内众多学者在航海航天领域以及能源运输领域也采用了以层析分析法为典型的多种数据质量评价方案。郭昊等[9]率先引入了对船舶自动识别系统（Automatic Identification System, AIS）数据的评价，重点关注了完整性、连续性和时效性这3个关键指标。提出了一种综合质量评分算法，将这3个关键指标结合起来，用于得出AIS数据的综合质量评分。刘承磊等[10]基于改进层次分析法（Improved Analytic Hierarchy Process, IAHP）和模糊综合评价法建立了石油管道的数据质量评估模型并用于划分管道数据质量等级。虞业泺等[11]通过初检数据质量筛选评价对既定卫星装备数据阈值性指标进行首轮质量评价及筛选，在此基础上利用多因素模糊推理下的层次分析法完成了卫星数据质量复检。

数据集质量评价可分为2种方式[12]：一是直接使用属性指标如准确性、完整性、一致性、可用性等进行评价。二是建立不同的分级评价指标体系进行评价，包括基于重复性、准确性、完整性的三维二级评价方法等。全面评估数据质量方法在大数据生产中得到了实际应用，构建了数据质量评价的基础，不仅丰富了数据治理的分析和应用经验，还为数据修正、筛选以及数据价值提取提供了新思路。

目前，行业内针对整车企业按GB/T 32960《电动汽车远程服务与管理系统技术规范》上传的汽车数据尚没有一个成熟的质量评估方案，但针对监测数据的数据评估将有效提升数据后续预警应用的准确性，因此本文提出车辆异常情况识别流程及数据质量评估方案。考虑不同行业中数据出现的主要异常情况（数据缺失、数据超限、数据间关联性错误等），结合监测平台数据特点，将数据异常类型具体化并补充特有异常类型（如难以解析等），力图涵盖车辆数据的全部异常情况。同时，对现有监控数据中的异常类型进行具体化与分类，通过建立科学的评估机制，实现数据质量的量化分析，为数据的整体质量提升和后续预警应用提供基础支撑。

1　监控平台数据质量评价方法介绍

根据GB/T 32960要求，车辆T-box将车辆监测数据传输到平台后，平台需按照国家标准要求，针对上传数据进行解析。因此，该数据质量评价内容从解析前异常数据与解析后异常数据两部分展开。

监测平台收集到车辆T-box上传的数据，并在数据解析前，首先需核查识别的各项内容：判断上传数据是否符合GB/T 32960协议规范；VIN检查合理性；上传数据是否可以解析。

针对解析后的监控数据质量评价，可分为5个一级维度即规范性、完整性、准确性、一致性以及时效性。5个维度又可以分为15个二级维度，如表1所示。

数据规范性是指上传监测平台的数据需满足GB/T 32960的强制要求。不符合数据规范性的异常类型包括：上传数据字段的异常值解析为“FE”；上传数据字段的无效值解析为“FF”。同时区分异常持续时间是否过长，一般根据持续异常时间是否超过3 min判断。

数据完整性指数据长度是否满足要求。不符合数据完整性的异常类型包括：缺失与冗余。缺失具体包括：数据字段的某几帧数值短时缺失及长时缺失。字段冗余包括两种：第一种是采集数据内容相较目标数据内容存在冗余字段，包括各单体电压及温度探针

数的备份字段。第二种为在一时间帧内，某字段内数据重复多次存储在一个数据位。

表1　数据质量评价的维度

一级维度	二级维度
规范性	短时间异常
	短时间无效
	长时间异常
	长时间无效
完整性	短时间缺失
	长时间缺失
	字段冗余
准确性	数据准确性异常
	数据恒值
	数据格式不合规
一致性	关联一致性差
	时变一致性差
时效性	时间规范性差
	及时性差
	时间连续性差

数据准确性是指数据值有效合理并符合数据类型的要求。不符合数据准确性的异常类型包括：字段采集数值不符合预期阈值；数值恒值性；数据格式不合规。

数据一致性用来衡量某数据字段与其他字段或该时间帧前后数据之间的矛盾程度。不符合数据一致性的异常类型包含以下维度：关联一致性，在同一时间帧上，不同字段满足逻辑程度的度量；时变一致性，同一字段随时间变化趋势满足逻辑程度的度量。

数据时效性是指数据采集时间字段的合理程度的度量。不符合数据时效性的异常类型包含以下维度：时间规范性、时间连续性、及时性。

基于上述数据异常情形的定义，针对解析后的监控数据进行异常识别。

2　异常数据识别

2.1　解析前异常数据识别

监测平台收集到车辆T-box上传的数据后，识别解析前异常情况，具体包括：车辆登入登出检测，验证车辆登入登出数据是否符合要求；上传数据是否满足数据包结构；上传数据长度是否与预设的数据单元长度一致。识别当前时间帧数据的VIN是否为平台内部VIN。上传数据是否可以解析。

记录上述异常数据包的上传时间及帧数。同时记录监测数据中标记为补发数据的上传时间及采集时间。

2.2　解析后异常数据识别

解析后，新能源汽车监测平台系统中上传的汽车全生命周期数据已具备物理含义。结合平台内构建的安全预警模型的自身特点，从数据规范性、数据完整性、数据准确性、数据一致性、数据时效性5个维度，开展解析后数据异常识别。

2.2.1　数据规范性

监测平台将上传数据按照GB/T 32960规则进行解析后，识别上传数据中是否存在“FE”及“FF”字段，从而判断上传数据是否存在异常值及无效值。记录异常持续时间、异常帧数。

2.2.2　数据完整性

针对不符合数据完整性的异常类型识别及记录方法包括：记录出现空值的字段及缺失持续时间、缺失帧数。提取上传字段中预设单体电压数及温度探针数，计算实际上传的电池单体电压数及探针温度数，并将其与预设值进行对比，从而识别是否有预留冗余单体电压及探针温度，标记是否存在备份数据字段、备份字段名称。识别字段内上传数据长度、格式是否满足国标要求，从而识别某字段内数据是否重复多次存储在一个数据位。记录数据重复字段、重复对应时间帧的全部数据。

2.2.3　数据准确性

针对数据准确性异常识别，首先筛选字段采集数值是否符合预期阈值。预设的字段阈值范围参考GB/T 32960要求。

数值恒值性检查内容包括：识别相邻2帧电流变化绝对值>20 A时，计算2帧数据间电池单体电压变化值。若存在某单体电压保持不变，则认为数据恒值异常。若识别车速持续大于60 km/h的3 min内，存在累计里程恒定不变的情况，则认为数据恒值异常。

核查数据格式（包括数据类型、数据长度、精度等）是否满足预期要求。如因为存储原因，数值型被误记为字符串。记录各类异常的对应字段及时间。

2.2.4　数据一致性

针对数据一致性异常识别方法包括：关联一致性，静置时车速为0；停车充电时电流为负且车速为0；单体电压最值与各单体电压的关系一致；温度最值与各探针温度的关系一致；单体电池数、温度探针

数与车辆静态上传的数据一致。时变一致性，累计里程跳变是否违背里程随时间的变化规律；充电状态下，相邻时间帧SOC的变化情况。当相邻时间帧递增时，计算相邻两帧的累计里程变化是否≥0；以及累计里程的变化值是否在合理范围内（里程差/时间差≤0.09 km/s）；充电状态下，当相邻时间帧递增时，计算相邻2帧的SOC变化是否≤4%。

2.2.5　数据时效性

针对数据时效性的异常识别方案如下：检查上传两帧数据时间间隔是否>0；针对连续的充电状态，相邻两帧数据时间差是否基本固定，没有丢帧。计算每一帧数据的上传时间与采集时间（若有采集时间）、上传时间与平台接收时间之间的时间差。识别时间差距是否>1 min。

根据识别异常类型的结果情况，得到各自的异常情况的时间帧占比。

综上，数据解析前可能存在3种异常类型、数据解析后可能存在15种异常类型，这些字段的异常类型均会导致数据处理出现错误，甚至会导致数据应用时出现错误，如基于数据的安全预警工作会出现大量误报现象。

3　数据质量综合评价方法

第2章针对监控数据提出了数据质量分析的多个维度，本章则针对多种分析维度提出最终的量化打分方案。利用层次分析法和熵权法，将主观和客观方法相结合确定多维度分析权重。利用模糊综合评价法对数据质量实际情况与理想情况进行了评分比较，实现了解析后数据质量的量化打分。

同时考虑数据存在漏发补发、解析异常等问题，将数据解析前的异常情形也纳入数据质量评估考量范围，在本节最后基于层次分析法，建立解析前数据异常、补发数据、解析后数据异常的得分权重，从而进行整车数据质量情况的综合评价。根据分数区间设置数据质量等级，用于评判数据质量优劣程度。

3.1　基于层次分析法确定主观权重

解析后的监控数据质量评价，可分为5个一级维度及15个二级维度。首先，确定每一维度的各元素的相对重要性，进行两两比较，采用表2所示的1~9标度法进行定量化评价。

针对规范性、完整性、准确性、一致性以及时效性5个维度，分别确定内部细分的二级维度之间的判断矩阵$\boldsymbol{C}_1$,$\boldsymbol{C}_2$,$\boldsymbol{C}_3$,$\boldsymbol{C}_4$,$\boldsymbol{C}_5$。

表2　元素赋值及说明

数值	说明
1	该元素和另一元素具备同样重要性
3	该元素比另一元素具备稍微重要性
5	该元素比另一元素明显重要
7	该元素比另一元素重要的多
9	该元素比另一元素绝对重要
2、4、6、8	表示上述判断的中间值

数据质量规范性维度判断矩阵$\boldsymbol{C}_1$：

$$\boldsymbol{C}_1=\begin{bmatrix}1 & 1 & 1/2 & 1/2\\ 1 & 1 & 1/2 & 1/2\\ 2 & 2 & 1 & 1\\ 2 & 2 & 1 & 1\end{bmatrix} \tag{1}$$

数据质量完整性维度判断矩阵$\boldsymbol{C}_2$：

$$\boldsymbol{C}_2=\begin{bmatrix}1 & 1/2 & 3\\ 2 & 1 & 4\\ 1/3 & 1/4 & 1\end{bmatrix} \tag{2}$$

数据质量准确性维度判断矩阵$\boldsymbol{C}_3$：

$$\boldsymbol{C}_3=\begin{bmatrix}1 & 1/4 & 1/5\\ 4 & 1 & 2\\ 5 & 1/2 & 1\end{bmatrix} \tag{3}$$

数据质量一致性维度判断矩阵$\boldsymbol{C}_4$：

$$\boldsymbol{C}_4=\begin{bmatrix}1 & 1\\ 1 & 1\end{bmatrix} \tag{4}$$

数据质量时效性维度判断矩阵$\boldsymbol{C}_5$：

$$\boldsymbol{C}_5=\begin{bmatrix}1 & 1/4 & 1/5\\ 4 & 1 & 2\\ 5 & 1/2 & 1\end{bmatrix} \tag{5}$$

针对规范性、完整性、准确性、一致性、时效性五个维度的判断矩阵$\boldsymbol{C}_6$：

$$\boldsymbol{C}_6=\begin{bmatrix}1 & 1 & 1/2 & 1/2 & 1/2\\ 1 & 1 & 1/2 & 1/2 & 1/2\\ 2 & 2 & 1 & 1 & 1\\ 2 & 2 & 1 & 1 & 1\\ 2 & 2 & 1 & 1 & 1\end{bmatrix} \tag{6}$$

其中，对应矩阵内第i行第j列数表示为分析维度i比分析维度j的重要程度，如：

$a_{ij}=1$，表示分析维度i与分析维度j同等重要；

$a_{ij}=3$，表示分析维度i比分析维度j略重要；

$a_{ij}=5$，表示分析维度i比分析维度j明显重要；

$a_{ij}=\dfrac{1}{a_{ji}}$，且当$i=j$，$a_{ij}=1$。

使用几何平均法进行权重向量计算，得到相对权重为：

$$\omega_i = \frac{\left(\prod_{j=1}^{n} a_{ij}\right)^{\frac{1}{n}}}{\sum_{i=1}^{n}\left(\left(\prod_{j=1}^{n} a_{ij}\right)^{\frac{1}{n}}\right)}, i = 1,2\cdots,n \tag{7}$$

针对判断矩阵C_1,C_2,C_3,C_4,C_5,C_6计算相关权重为：a_1=(0.166 7, 0.166 7, 0.333 3, 0.333 3), a_2=(0.319 6, 0.558 4, 0.122 0), a_3=(0.098 9, 0.536 8, 0.364 3), a_4=(0.500, 0.500), a_5=(0.098 9, 0.536 8, 0.364 3), a_6=(0.125 0, 0.125 0, 0.250, 0.250, 0.250)。

针对6个判断矩阵进行一致性检验，从而确定构建矩阵的合理性。当指标CR<0.1，认为具备满足需求的一致性。

$$CR = \frac{CI}{RI} \tag{8}$$

式中：CR为一致性比率，判断矩阵是否具有一致性的指标；CI为一致性指标，表示偏离一致性矩阵的程度；RI为随机一致性值，表示基于随机生成的判断矩阵得到的平均一致性指标值，用于平衡CI的大小。

其中，CI满足

$$CI = \frac{\lambda_{max} - n}{n - 1} \tag{9}$$

式中：λ_{max}为特征根最大值，n为待分析的判断矩阵阶数。随机一致性值RI见表3。

表3　随机一致性值标准值

阶数	1	2	3	4	5	6
RI	0	0	0.58	0.9	1.12	1.24

例如对于判断矩阵C_1，λ_{max}为4，计算得CR为0，小于0.1，矩阵具有一致性；对于重要性判断矩阵C_2，λ_{max}为3.018，计算得CR为0.016，小于0.1，矩阵具有一致性；同理可确定判断矩阵C_3、C_4、C_5、C_6均具有一致性。

根据3.1节中的a_1、a_2、a_3、a_4、a_5，确定数据质量评估5个一级维度中二级维度的相对重要程度。基于a_6可确定5个一级维度之间的相对重要程度，基于数据之间关系，可以得到在数据质量评估的目的下，分析的15个二级维度各自的权重，即：

w=[a_1(1,2,3,4)*a_6(1),a_2(1,2,3)*a_6(2)……]=[0.020 833 33, 0.020 833 33, 0.041 666 67, 0.041 666 67, 0.039 952 28, 0.069 803 07, 0.015 244 65, 0.024 720 98, 0.134 206 14, 0.091 072 88, 0.125 0, 0.125 0, 0.024 720 98, 0.134 206 14, 0.091 072 88]。

3.2　基于熵权法确定客观权重

熵值衡量了数据的混乱程度，熵权法就是利用不同维度之间信息量的波动程度从而确定不同维度权重，实现了基于客观可量化因素确定各自维度的客观权重。

为更便于客观评价数据相对熵值，需预设一种最优的数据质量状态。即设定完美状态所有维度对应值均为1。此时对应综合评价结果为100分。各个维度数值范围为0~1，且数值越大，说明此数据异常识别维度的数据质量越好，值的大小与质量优劣为正相关。

因此预设最优数值为：Y_1=(1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1)。

后续计算过程中定义此最优情况为样本一，并定义新能源汽车数据实际异常情况为样本二。设某新能源汽车数据经异常识别的对应的各分值：

$$Y_2 = (0.98, 0.9, 0.95, 0.99, 0.9, 0.9, 0, 0.8, 0.9, 0.9, 0.97, 0.97, 0.99, 0.98, 0.95)$$

为减小数据大小对分析结果的影响，通过密度公式对上述的[Y_1,Y_2]进行标准化。

$$z_{ij} = \frac{y_{ij}}{\sum_{k=1}^{2} y_{kj}} \tag{10}$$

从而得到对应[Y_1,Y_2]的标准矩阵[Z_1,Z_2]为：

Z_1=(0.505, 0.526, 0.513, 0.503, 0.526, 0.526, 0.5, 0.556, 0.526, 0.526, 0.507, 0.508, 0.503, 0.505, 0.513)；Z_2=(0.495,0.474,0.487,0.497,0.473,0.474,0.5,0.444,0.474,0.474,0.492,0.492,0.497,0.495,0.487)。

根据以下公式计算每一个维度对应的熵值E_j：

$$E_j = -k\sum_{i=1}^{2} z_{ij} \ln z_{ij} \tag{11}$$

式中：$k = (\ln 2)^{-1}$，2表示理想最优状态样本与实际样本的数量；E_j在0~1的闭区间范围内，j取值为1~15的整数，表征分析的15个维度。

进一步可计算权重β_j：

$$\beta_j = (1 - E_j) / \sum_{j=1}^{15}(1 - E_j) \tag{12}$$

3.3　基于组合赋权的模糊综合评价

利用层次分析法与熵权法可分别得到数据质量评价中15个维度的2组权重，将其点乘标准化即可得到对应的综合权重值W，即：

$$W = \frac{w\beta}{\sum_i w_i \beta_i} \tag{13}$$

式中：w表示基于层次分析法确定的权重，β表示基于熵权法确定的权重。

在本数据中综合权重为：

$\boldsymbol{W}$=(0.001 46, 0.039 68, 0.018 83, 0.000 72, 0.076 10, 0.132 97, 0, 0.210 21, 0.255 65, 0.173 48, 0.019 92, 0.019 92, 0.000 43, 0.009 41, 0.041 16)。

将各个维度的数据质量分为3个评价等级，即为优秀、可用、较差，相应的评分为100、60、40。模糊评价及相应的评分集是模糊综合评价的基础。

定义各维度需满足的要求，15个维度中设定最低异常占比要求a为0.9，中等要求b为0.95，最高要求c为1。即，认为完全没有数据异常问题为优秀，某种数据异常情况多于数据长度的10%，认为数据较差。本次分析中均为正相关，即数据质量越好，打分越高。

可建立如下隶属函数：

属于“优秀”的隶属函数为：

$$\mu=\begin{cases}1 & x\geqslant c\\ \left(\dfrac{x-b}{c-b}\right)^3 & b<x<c\\ 0 & x\leqslant b\end{cases} \tag{14}$$

属于“可用”的隶属函数为：

$$\mu=\begin{cases}0 & x\leqslant a \text{ or } x\geqslant c\\ \left(\dfrac{x-a}{b-a}\right)^3 & a<x\leqslant b\\ \left(\dfrac{c-x}{c-b}\right)^3 & b<x<c\end{cases} \tag{15}$$

属于“较差”的隶属函数为：

$$\mu=\begin{cases}1 & x\leqslant a\\ \left(\dfrac{b-x}{b-a}\right)^3 & a<x<b\\ 0 & x\geqslant b\end{cases} \tag{16}$$

结合上述3个隶属函数，可得到各个维度分数在3个评价中的比例情况。选用带有立方根的隶属度函数可以放大数据变化程度，增加该方法对数据异常的敏感性。

对应Y_1的评价矩阵$\boldsymbol{R}_1$见表4。

表4　最优状态的评价矩阵

优秀	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
可用	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
较差	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

对应Y_2的评价矩阵$\boldsymbol{R}_2$见表5。

表5　实际状态的评价矩阵

优秀	0.216	0	0	0.512	0	0	0	0	0	0	0.064	0.064	0.512	0.216	0
可用	0.064	0	1	0.008	0	0	0	0	0	0	0.216	0.216	0.008	0.064	1
较差	0	1	0	0	1	1	1	1	1	1	0	0	0	0	0

将综合权重矩阵$\boldsymbol{W}$与评价矩阵$\boldsymbol{R}$相乘可以得到对应的3个评价的占比，即综合模糊评价向量$\boldsymbol{D}$。

对应Y_1的综合模糊评价向量为(1,0,0)。对应Y_2的综合模糊评价向量为(0.005 48,0.069 3,0.888 1)。

根据前期评价集，确定3种评价对应分值为100、60、40。因此将模糊评价向量与评价集合得分相乘得到最终得分：样本1即满分100，样本2得分为40.23。

3.4　数据质量综合评价

本节建立了考虑解析前后异常情况的综合评价放方法。建立解析前异常、补发数据、解析后数据异常3者的判断矩阵，判断矩阵$\boldsymbol{C}_7$：

$$\boldsymbol{C}_7=\begin{bmatrix}1 & 2 & 1/9\\ 1/2 & 1 & 1/13\\ 9 & 13 & 1\end{bmatrix} \tag{20}$$

对于判断矩阵$\boldsymbol{C}_7$，λ_{max}为3.012，计算得CR为0.010，小于0.1，矩阵具有一致性。

使用几何平均法进行权重向量计算，得到判断矩阵$\boldsymbol{C}_7$的相对权重为：$\boldsymbol{a}_7$=(0.103 8,0.057 8,0.838 4)。2.1节中得到了解析前异常的帧数、补发数据帧数。计算异常及补发帧数与上传的数据总量的比值，可得到解析前异常占比为p_1，补发数据占比为p_2。

即最终得分结果R为：

$$R=0.103\,8p_1+0.057\,8p_2+0.838\,4D \tag{21}$$

根据分数区间设置数据质量等级，用于评判数据质量优劣程度。根据相关专家经验，其中数据质量等级与数据得分关系，见表6。

表6　数据质量评价等级

评分等级	分值区间	评分等级	分值区间
Ⅰ（优）	(70,100]	Ⅲ（中）	(30,50]
Ⅱ（良）	(50,70]	Ⅳ（差）	(0,30]

4　实车数据测试

选取多家车企数据进行测试，论证了异常数据类型的全面性及质量评价方法的合理性。

选用A、B、C这3家整车企业，各20台车辆监测数据，单台车辆数据分析时间为1年，数据分析字段为GB/T 32960要求内容，每台车的时间帧数平均约为934 608帧。

针对数据中异常类型进行人为识别，发现数据可能存在的异常类型完全被涵盖在上述提及的解析前后的18种异常情况。通过Python针对车辆数据进行批量处理及异常识别，其中各类异常出现的时间帧数占比如表7所示。

表7　实车测试异常数据占比　　%

异常维度\车企	A	B	C
不符合传输协议	0.001	0.003	0.000 2
VIN不在分析车型内	0	0	0
数据无法解析	0.001	0.003	0.002
短时间异常	0	0.003	0
短时间无效	0	0	0
长时间异常	0	0.068	0
长时间无效	0	0	0
短时间缺失	0.010	0	0.003
长时间缺失	0.082	0	0.005
字段冗余	0.000 5	0	0
数据准确性异常	0.046	0.004	0.007 8
数据恒值	0.000 1	0	0
数据格式不合规	0.000 3	0.000 1	0
关联一致性差	0.000 5	0.000 3	0.000 5
时变一致性差	0.000 2	0.000 3	0.000 3
时间规范性差	0.000 1	0	0
及时性差	0.014	0.007	0.001
时间连续性差	0.000 2	0.000 3	0.000 3

根据表7内容可以直观发现不同整车企业数据的异常占比情况。可通过上述内容确定从硬件层面的数据质量提升策略。

每家车企数据的数据质量评分结果见表8。表8的结果能直观展示车企的数据质量情况。基于以上结果可帮助整车企业建立数据质量优劣的报警机制，并提升数据传输质量。涉及的数据异常识别及数据质量评价方法具有重要的现实意义。

表8　实车测试数据质量评分　　分

评分等级	A	B	C
Ⅰ（优）	11	17	18
Ⅱ（良）	5	2	1
Ⅲ（中）	3	1	1
Ⅳ（差）	1	0	0

5　结束语

本文描述了新能源汽车监控平台数据的异常识别及数据质量评价方法。数据评价的内容包括数据上传平台解析前和解析后两类。解析前识别了异常数据的帧数及补发数据帧数。数据解析后异常情况涵盖数据规范性、数据完整性、数据准确性、数据一致性、数据时效性5个一级维度、15个二级维度。数据异常种类基本涵盖了新能源汽车监控平台数据可能出现的全部异常问题。针对解析后的监控数据，利用层次分析法和熵权法对不同维度的权重进行计算，通过组合赋权确定了主客观两个角度考量的综合权重。基于模糊综合评价方法，针对异常数据识别结果与理想识别结果的对比，确定了最终量化评分，避免了单纯的主客观因素的影响。最终基于判断矩阵将解析前数据情况、补发数据情况、解析后数据情况，得到数据质量最终分析结果。利用多家车企的实际运行状态数据证明数据异常类型的全面性及评价方法的合理性。

参　考　文　献

[1] 沈小军, 付雪姣, 周冲成, 等. 风电机组风速-功率异常运行数据特征及清洗方法[J]. 电工技术学报, 2018, 33(14): 3353-3361.

[2] 王英会. 高速公路交通流异常数据识别及修复方法研究[D]. 北京: 北京交通大学, 2015.

[3] 付时瑞, 卜峰, 吴艳鹏, 等. 动车组齿轮箱异常振动监控数据分析[J]. 城市轨道交通研究, 2022, 25(2): 99-102.

[4] 郑涛, 徐海红. 废气污染源自动监控数据中异常数据的识别方法[J]. 天津科技, 2013, 40(6): 13-16.

[5] 黄雄波, 钟全. 路灯监控系统中时序数据流的异常值检测研究[J]. 微处理机, 2018, 39(6): 47-53.

[6] 殷大发. 煤矿安全监控系统监测点数据异常识别技术研究[J]. 矿山机械, 2013, 41(4): 120-123.

[7] WHAN S O, JEONG S K, SEON Y I, et al. Data Quality Assessment for Observational Medical Outcomes Partnership Common Data Model of Multi-Center[J]. Studies in Health Technology and Informatics, 2023, 302: 322-326.

[8] 黄国彬, 陈丽. 国外科学数据质量评估框架比较研究[J]. 图书与情报, 2021(1): 97-107.

[9] 郭昊, 李海滨, 冯姣, 等. 基于大数据处理的船舶数据质量评价方法研究[J]. 计算机仿真, 2022, 39(2): 298-303.

[10] 刘承磊, 姜晓红, 张翰钊, 等. 基于模糊综合评价的管道内检测数据质量评估[J]. 油气田地面工程, 2023, 42(5): 69-77.

[11] 虞业泺, 施敏华, 邓洛凤, 等. 卫星装备试验鉴定数据质量评价技术及实现[J]. 计算机测量与控制, 2021, 29(8): 233-237.

[12] 盛小平, 焦凤枝. 国内外开放数据评价研究综述[J]. 情报杂志, 2022, 41(8): 131-137.

（责任编辑　**明慧**）