doi: 10.7690/bgzdh.2019.12.010

基于信息熵的遥测数据质量维度量化方法

谷阳阳
（中国人民解放军 92941 部队 45 分队，辽宁 葫芦岛 125000）

摘要：为解决面对遥测数据质量评估依赖专家知识定性分析，人工参与存在主观随意性强、多测站短时间提供处理方案等问题，提出一种基于信息熵的遥测数据质量维度量化方法。在多数据源条件下，通过试验信息构造评估参照标准，利用信息熵指标对数据评价的准确性和完整性维度进行度量，并以某次飞行器的模拟飞行试验数据进行测试分析。测试结果表明：该方法不仅能对遥测数据的准确性和完整性维度给出准确的评估值，而且具有很高的计算效率，易于实现。

关键词：信息熵；多数据源；数据质量；量化评估
中图分类号：TP702　　**文献标志码：**A

Quantization Method of Telemetry Data Quality Dimension Based on Information Entropy

Gu Yangyang
(*No*. 45 *Team*, *No*. 92941 *Unit of PLA*, *Huludao* 125000, *China*)

Abstract: To resolve the problem of the quantization assessment of telemetry data relying on qualitative analysis of expert knowledge, the highly arbitrary artificial participation in subjective assessment, providing solutions in a short time in multi-station environment, this paper proposes a method of telemetry data quality dimensions within context on the basis of information entropy. In multiple-source environment, the method constructs and evaluates the reference standard by test information, measures the accuracy and completeness of data by the information entropy index, and the simulation flight test data of the certain aircraft are used for test analysis. The simulation test shows that the method of the automatic quantization assessment not only evaluates the accuracy and completeness dimension of telemetry data accurately, but also has good adaptability and easy to implement.

Keywords: information entropy; multiple-source information; data quality; quantization assessment

0　引言

随着 QIS 质量管理体系的运用，遥测数据的质量评估是数据处理和数据管理时面临的首要问题，数据质量的管理如同其他质量管理一样贯穿于数据生命周期的各个阶段[1]。现有的遥测数据处理过程中的质量评估只能通过人工参与的方式进行判断，各站位记录数据的准确性、完整性、一致性需依赖专家知识进行定性分析[2-4]。文献[5]提出一种采用 XML 描述数据质量参数的数据质量提高框架，但对有效地评估数据质量没有论述，而遥测数据质量评估需要短时间内对各站位记录的数据情况进行横向比对分析，筛选出记录相对完整、干扰少的使用。基于信息熵的遥测数据质量维度量化方法，是在解决多数据源背景条件下的协作信息系统，在数据预处理阶段从数据质量维度的角度对某条或多个数据源的记录情况进行量化评估，为原始数据的进一步细化处理提供定量依据，避免了主观评估的随意性，实现了数据质量评估的自动化。

1　数据质量维度量化方法

1.1　遥测数据结构

脉冲编码调制（pulse-code modulation，PCM）遥测系统是典型的数字式时分多路传输系统。每路脉冲在采样周期内占据相同的时间间隔，一个采样周期就是一帧。遥测数据由一定记录格式、长度固定的 PCM 数据流子帧构成[6]。飞行器型号记录的帧结构不同，具体格式要根据试验大纲进行设计。帧数据格式中的 BCD 时间码位，记录了当前帧到来时，时码接收器的解调时间；副帧计数字 SFC 是副帧同步字，当副帧锁定时以 0 作为全帧起点，每帧加 1 直至每帧结束，当副帧未锁定时保持 0 不变。子帧同步码组位于数据字区域的末端。具体结构如图 1 所示。

收稿日期：2019-08-22；**修回日期：**2019-09-05
作者简介：谷阳阳（1982—），女，山西人，学士，工程师，从事遥测数据处理及数据工程研究。E-mail: 787743998@qq.com。

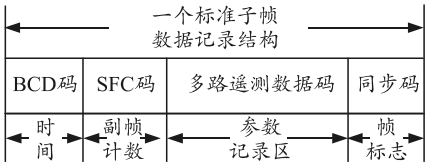


图 1　标准子帧数据结构

1.2　多数据源归一化模型

遥测数据的测量特点是，分布于不同测量站点的各自独立的测量设备，记录同一飞行目标的内环境数据。这些独立的子系统联合构成了一个协作信息系统(collaborative information system，CIS)[7-8]。在实际 CIS 环境中，是由 N 个站位的数据源 $D=\{T_1,T_2,\cdots,T_N\}$提供测量数据，每个站位记录的飞行数据为 $T_i=\{r_1,r_2,\cdots,r_N\}$，所有记录参数的实体集合为 $E=\{e_1,e_2,\cdots,e_N\}$，其中 $r_i(i=1,2,3)$是 e_i的一种记录版本，r_i可能会缺失，可能会是野值等。

在多数据源的环境中，确定好数据质量评估的参照，是实现自动化数据质量评估的关键，而质量评估的任务是评估一条记录 r_i或一个站位的数据源 T_i多大程度逼近于其准确表达 e_i或 E。针对遥测传感器和变换器的记录特点，在实际操作中选择站位数据进行处理时，首要关注时间零点信号 r_{T0} 的记录情况，而时间零点信号 r_{T0} 是一个开关量，所以利用飞行前的验前信息结合实际的飞行零点时刻，根据遥测数据的帧结构，模拟出一个关于时间零点信号 r_{T0}的准确表达，作为数据质量评估的参照 r'_{T0}。当然，在质量评估的过程中也可以选择不同的重要参数进行模拟，作为质量评估的准确表达，但把握的原则是要保证一定的采样率，一般会选择主、副帧参数，实际应用中选择阶跃类型的参数作为“准确表达”参照 r'，模拟得会更准确，操作更为方便、快捷。

1.3　信息熵的概念

C.E.Shannon 于 1948 年提出信息熵，将熵引入了信息论，从而奠定了现代信息论的理论基础[9-11]。对于离散随机变量而言，设随机变量为 X，其成员的分布频率记为 $p(X)$，则 X 的信息熵是：

$$H(X)=H(p_1,p_2,p_3,\cdots,p_n)=-k\sum_{i=0}^{n}p_i\log(p_i)。\quad(1)$$

式中：K 为常数，通常取值为 1，为了计算方便，一般取对数的底为 e，即有：

$$H(X)=H(p_1,p_2,p_3,\cdots,p_n)=-k\sum_{i=0}^{n}p_i\ln(p_i)。\quad(2)$$

这里 $\ln(p_i)$直观地反映了多少信息从信源传递到信宿。

对于给定的 2 个数据分布 $p(X)$和 $q(X)$，其熵差定义为：

$$D(p\|q)=\sum_X p(X)\ln P(X)-\sum_X q(X)\ln q(X)。\quad(3)$$

熵差的概念直观地反映了 2 个信源的信息量差异。基于信息熵和熵差的特点，以此来描述数据准确性和完整性的度量。

1.4　数据准确性度量方法

记录 r_i 的准确性可以用 r_i 和其参照 r_i' 的信息量相似程度来衡量，即

$$\mathrm{acc}(r_i[k])=\frac{\left\|H\left(G(r_i[k])\right)\right|-\left|D\left(G(r_i'[k])\|G(r_i[k])\right)\right\|}{\left|H\left(G(r_i[k])\right)\right|}。\quad(4)$$

其置信度为：

$$\mathrm{conf}\left(\mathrm{acc}(r_i[k])\right)=\frac{\sum_{j=1}^{k}\frac{1}{2^j}}{\sum_{j=1}^{L_{\max}}\frac{1}{2^j}}。\quad(5)$$

这里 k 是记录 r_i 中数据需要进行质量评估的采样个数，而 $L_{\max}$ 则是参照 r_i' 中采样总数。涉及记录中的采样数目越多，得到的准确性就具有越高的可信度。

1.5　数据完整性度量方法

遥测数据的完整性有 2 方面的体现：1) 检查全帧：按照记录格式对帧头、帧长和帧标志的要求，检查是否有丢帧、帧标错或倒时间的情况发生及其程度，这些“问题帧”都不能作为遥测处理的数据；2) 检查记录 r_i 的完整性：记录 r_i 的采样数目越多，其包含的信息量越多；因此，记录 r_i 的完整性可以用它本身传递的信息量和其参照 r_i' 信息量的比值来衡量。

记录 r_i 传递的平均信息量为

$$H(r_i)=\frac{1}{m}\sum_{k=1}^{m}H\left(G(r_i[k])\right)。\quad(6)$$

而其参照 r_i' 传递的平均信息量为

$$H(r_i')=\frac{1}{L_{\max}}\sum_{k=1}^{L_{\max}}H\left(G(r_i'[k])\right)。\quad(7)$$

记录 r_i 的完整性可以度量为

$$\mathrm{comp}(r_i)=\frac{H(r_i)}{H(r_i')}。\quad(8)$$

其置信度与准确性的定义相同。

2 仿真测试分析

利用某次飞行器模拟飞行的试验数据，分别从 4 个站位记录下来的数据流，先按照人工参与的方式对 4 个数据流进行数质量的检查，对每个站位记录的 r_{T0} 进行了预处理，人工统计情况如表 1 所示。

表 1 数据预处理情况统计

类型	总帧数	有效帧数	帧时间错	帧标错误	处理段落/s
A 站	107 071	105 330	0	1 741	6.837～51.776 6
B 站	161 131	161 034	0	97	0.317 0～51.776 8
C 站	162 761	162 735	2	24	0.332 6～51.776 9
D 站	167 653	167 615	2	36	−1.352 7～51.776 9

从初步的人工统计信息来看，A 站位的数据帧标错最多，对应的跟踪处理段落的数据最少，C 站位、D 站位记录的数据有效帧数较多，时间段落较长，数据野值的分布相对集中在首段和尾段，而 A、B 站位野值分布在飞行中段。从人工初步统计信息来看，C 站位、D 站位记录相对较好，可以优先作为下一步处理的数据源进一步分析。

用基于信息熵评测遥测数据准确性、完整性的自动统计方法，验前信息得知 r_{T0} 是一个由 0 跳跃到 3 的阶跃信号，由试验时的零点时刻，能准确地描绘出质量评估的参照 r'_{T0}，自动评测结果如表 2 所示。

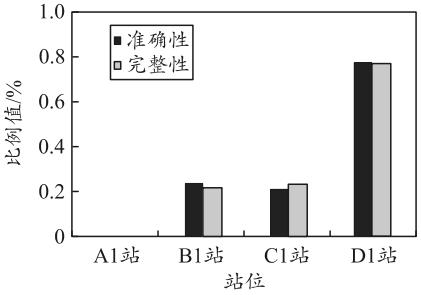
表 2 数据质量评测统计

类型	准确性	完整性	类型	准确性	完整性
A 站	0.785 3	0.809 5	C 站	0.974 2	0.985 3
B 站	0.850 1	0.893 1	D 站	0.980 1	0.991 1

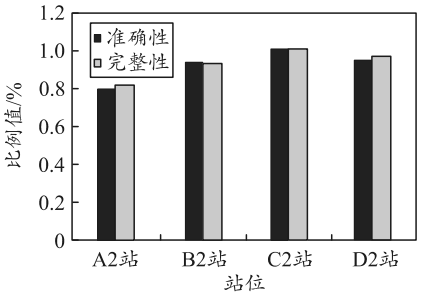
通过自动统计结果，可以很直观地看到 C、D 站位记录的数据完整性和准确性比 A、B 站位要好得多，可以优先作为下一步处理的数据源进一步分析。这与人工统计结果一致，但提高了预处理效率。

由于遥测测量是一个协作信息系统模式，所以在实际处理中经常会根据各站位记录情况，择优选择各测站记录良好的数据段，拼接全程测量数据。根据这一需求，笔者将各测站数据分段进行量化评测，择优选择数据段，再根据多站遥测原始数据精准对接方法[12-13]，进行选段对接，可以实现在帧时间、数据连续性和一致性的前提下，提供高质量的全程测量数据，最大限度地利用各测站数据。

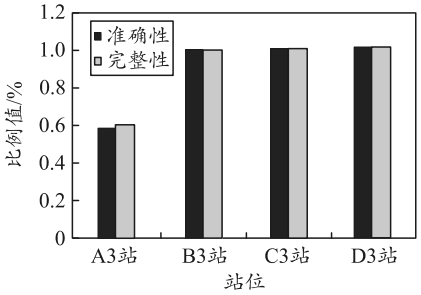
基于上述提供的测试数据，利用验前信息将数据分为−2.000 0～1.000 0 s、1.000 1～10.000 0 s、10.000 1～30.000 0 s、30.000 1～52.000 0 s 4 个段落进行质量分段评估。由于飞行器状态变化时刻，伴随着振动、噪声的发生，测量上野值也会集中发生，所以分段时，根据验前知识应尽量避开飞行器预设的状态变化段落，将分割点尽量选择在稳定飞行段落，数据分段量化评测统计如图 2 所示。



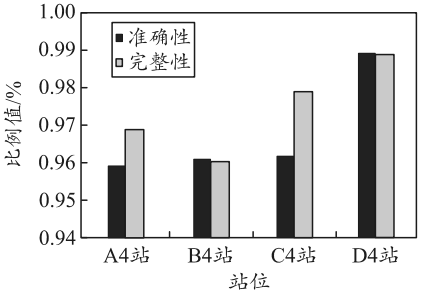
(a) −2.000 0～1.000 0 s 数据段



(b) 1.000 1～10.000 0 s 数据段



(c) 10.000 1～30.000 0 s 数据段



(d) 30.000 1～52.000 0 s 数据段

图 2 数据分段量化评测统计

从统计结果可以很直观地给出数据对接方案：D1+C2+C3+D4 或者 D1+C2+D3+D4，这与人工选段的结果一致。从以上 2 种测试情况可知：基于信息熵的遥测数据质量维度量化方法，在多数据源背景条件下的协作信息系统中，数据的预处理阶段对多个数据源和分段数据均可实现自动量化评估，对各测站数据记录情况提供量化指标，有效避免了人为分析造成的处理方案不确定性，数据分段量化评估更加细化测量段落，提高了数据利用率，同时满足在大数据量条件下，短时间提供数据处理方案的需求。

3　结束语

笔者结合遥测数据帧结构特点，确立多数据源归一化模型，解决了自动化数据质量评估的关键问题，提出基于信息熵建立准确性和完整性的度量，实现了 CIS 环境中遥测数据质量维度的度量统一。仿真测试结果与人工评估结论一致，避免了人工识别的主观随意性，在测站布设多、测量数据量大时，大大提高了预处理效率。作为一种自动化的数据质量评估方法，其准确性和完整性在数据源级别的度量效果满足需求，有效实现了数据管理的质量服务于遥测数据处理。

参考文献：

[1] 龚益鸣. 现代质量管理学[M]. 北京: 清华大学出版社, 2012: 56−68.

[2] 唐家驹, 刘书庆, 程幼明. 质量管理学[M]. 北京: 中国计量出版社, 2004: 101−120.

[3] 李洪敏, 黄晓芳, 张建平. SQL 自动访问攻击框架研究与设计[J]. 兵工自动化, 2015, 34(8): 45−48.

[4] 刑立新, 沈中卿. 某型火箭炮数据采集设备的设计与实现[J]. 兵工自动化, 2014, 33(8): 69−71.

[5] SCANNAPIECO M, VIRGILLITO A, MARCHETETTI C, et a1. The DaQuinCIS architecture：A platform for exchange and improving data quality in cooperative information systems[J]. Information Systems, 2004, 29: 551−582.

[6] 李邦复. 遥测系统[M]. 北京: 中国宇航出版社, 2007: 53−69.

[7] OMAR B, ANISH D S, ALON H, et a1. ULDBs：databases with uncertainty and lineage[C]//Proceedings of the 32nd International Conference on Very large Data Bases. San Francisco：Morgan Kaufmann Publishers, 2006: 953−964.

[8] CHEN H Q, KU W S, WANG H X. Cleansing uncertain databases leveraging aggregate constraints(C)// Proceedings of the 26th International Conference on Data Engineering. Washington, DC: 1EEE Computer Society. 2010: 128−135.

[9] KAO J J, LI P H, HU W S. Optimization models for sitting water quality monitoring stations in a catchment[J]. Environment Monitoring and Assessment, 2012, 184(1): 43−52.

[10] HALL M A. Correlation-based feature selection for discrete and numeric class machine learning[C]//Proc of the 17th Int Conf of Machine Learning(ICML). San Francisco: Morgan Kanfman, 2000: 359−366.

[11] WANG Q, SHEN Y I, ZHANG Y, et al. Fast quantitative correlation analysis and information deviation analysis for evaluating the performances of image fusion techniques[J]. IEEE Trans on Instrumentation and Measurement, 2004, 53(5): 1441−1447.

[12] 朱学锋, 韩宁. 基于互累积量的遥测数据时延估计方法[J]. 遥测遥控, 2009, 30(3): 65−67.

[13] 陈以恩. 遥测数据处理[M]. 北京: 国防工业出版社, 2002: 158−167.