第 50 卷 第 1 期：274-282 高电压技术 Vol.50, No.1: 274-282
2024 年 1 月 31 日 High Voltage Engineering January 31, 2024
DOI: 10.13336/j.1003-6520.hve.20221976

基于组合赋权法和模糊综合评价的电力设备状态数据质量评估

计 蓉[1]，侯慧娟[2]，盛戈皞[2]，张立静[2]，舒 博[2]，江秀臣[2]
（1. 上海交通大学国家电投智慧能源创新学院，上海 200240；2. 上海交通大学电气工程系，上海 200240）

摘 要：随着电力网络的扩大以及工业信息化的迅速发展，在电力领域中采集和待处理的数据量呈现爆发式增长，数据的丢失、冗余、异常、冲突等问题也日益突出，影响数据的质量。数据质量评估作为保证数据质量的关键一环，发挥着重要作用。针对电力设备的状态数据提出一种数据质量评估方法，选择完整性、准确性、唯一性、一致性、及时性 5 个评估指标构建数据质量评估体系，设置了定量计算的评估规则；组合运用模糊层次分析法和熵权法来确定各评估维度的权重，以提高数据质量评估的科学性；再运用模糊综合评价法，依据隶属函数来确定数据质量的所属等级。最后运用上述方法，对某地的油色谱数据质量进行评估，该方法计算得到数据质量的评分为 77.15，处于“中等”等级。评估结果与实际应用场景相符，表明该文提出的方法适用于电力设备状态数据质量评估。
关键词：大数据；电力设备；数据质量评估；模糊层次分析法；熵权法；模糊综合评价

Data Quality Assessment for Power Equipment Condition Based on Combination Weighing Method and Fuzzy Synthetic Evaluation

JI Rong[1], HOU Huijuan[2], SHENG Gehao[2], ZHANG Lijing[2], SHU Bo[2], JIANG Xiuchen[2]
(1. College of Smart Energy, Shanghai Jiao Tong University, Shanghai 200240, China;
2. Department of Electrical Engineering, Shanghai Jiao Tong University, Shanghai 200240, China)

Abstract: With the increasing expansion of power network and the rapid development of industrial informatization, the amount of data collected and to be processed in the field of electric power shows explosive growth. Data loss, redundancy, exception, conflict and other problems are becoming increasingly prominent, affecting the quality of data. As a key part of ensuring data quality, data quality assessment plays an important role. In this paper, a data quality evaluation method is proposed for electrical equipment monitoring data. Five evaluation indexes including completeness, accuracy, uniqueness, consistency and timeliness are selected to construct a quality evaluation system, and evaluation rules for quantitative calculation are set up. The fuzzy analytic hierarchy process and entropy weight method are combined to determine the weight of each dimension, raising the scientificity of the data quality evaluation. Then the method of fuzzy comprehensive evaluation is used to determine the level of data quality based on membership function. Finally, the above method is used to evaluate the quality of oil chromatographic data in a local power grid. The score of data quality calculated by this method is 77.15, rated “medium”. The assessment result is consistent with the actual application situation and verifies that the method proposed in this paper is applicable to the data quality assessment of power equipment condition.
Key words: big data; electrical equipment; data quality assessment; fuzzy analytic hierarchy process; entropy weight method; fuzzy comprehensive evaluation

0 引言

近年来，互联网、物联网、云计算等技术飞速发展且得到广泛应用，各行业中产生的数据呈爆炸式增长，数据规模已从 GB 级、TB 级增长到 EB 级、ZB 级[1]。数据作为信息的载体，能够反映事物内部的联系和规律，是做出决策的基础，具有很高的价值。但在数据采集、运输、储存过程中，主客观因素导致的数据缺失、异常、冗余和内部冲突，会降低数据的质量。国际知名咨询机构 Gartner 的调查显示，全球 1000 强的企业有超过 25%存在数据缺失和数据异常问题[2]。低质量数据中难以提取关键

基金资助项目：国家重点研发计划 (2020YFB1709701)。
Project supported by National Key R&D Program of China (2020YFB1709701).

有效的信息，会影响基于数据做出的判断和决策。因此数据的质量需要保障，数据质量评估作为其中的关键一环具有重要意义。

电力设备状态数据的质量对电力设备运行可靠性、经济性及供电安全性均有影响。目前对电力设备状态数据质量评估的研究还比较少。文献[3-4]设计了适用于电力数据质量评估的大数据框架，但没有给出具体的质量评估案例。文献[5]搭建数据质量评估的模型，设计了具体的评估指标规则，实现定量评估。但指标不全面，没有考虑实时性，且权重的确定过程中存在较强主观性。文献[6]将分类算法应用于数据质量评估模块，并采用基于机器学习的方法完成平台质量评估的模块。文献[7]先利用MapReduce并行化K-means聚类算法实现数据预处理，再利用灰色熵权法判断出数据质量所处等级。文献[8]构建了智能配电网的数据质量评价体系，从准确度、完整度、及时度等5个维度进行评估，但每个指标的评价规则较为复杂，实现难度较大。

模糊综合评价法以模糊数学为基础，利用模糊数学理论将定性的评价转化为定量评价，对受到多因素、多条件影响制约的对象做出综合评价[9]。本文针对电力设备的状态数据，建立了多维度的数据质量模糊综合评估模型，从完整性、准确性、唯一性、一致性、及时性5个维度对数据的质量进行评估，根据隶属度理论建立关于以上评估指标的模糊关系矩阵。在确定评估指标的权重时，组合运用模糊层次分析法(fuzzy analytic hierarchy process, FAHP)和熵权法，提高赋权的科学性和准确性。结合模糊关系矩阵和权重，可以确定数据质量所属的等级，科学客观地反映数据真实情况。最后，以某地电网公司的油色谱在线监测数据为例，采用上述方法对数据质量进行评估，结果表明数据质量处于“中等”的水平，符合实际应用场景，验证了此方法的可行性和有效性。

1　数据质量评估

1.1　数据质量评估概念

数据的质量管理问题涉及多个方面和多个环节，自美国麻省理工学院(Massachusetts Institute of Technology, MIT)提出全面数据质量管理计划(total data quality management, TDQM)，数据质量的知识体系开始建立，目前在数据质量评估、异常数据检测、数据清洗、数据特征提取等方面，已有大量的学术研究和探索。目前对数据质量的研究主要从两方面开展：一方面是对数据的质量进行评估和监测；另一方面是采用多种方法提高数据的质量。数据质量评估是提高数据质量的基础和必要前提，在多个维度上对数据的质量进行评价，定性或定量地评出数据质量的好坏，给数据用户直观的反映，也为后续提高数据质量的操作提供基础[10]。

数据质量在不同领域和背景下的定义有所差别，拥有丰富的内涵，但都可以看作是一个层次分类的概念，数据的质量可分解成具体的数据质量维度[11]。数据质量评估的核心问题是完成对各个维度的评估，并将其综合起来。MIT首先提出信息管理质量评估框架(a methodology for information quality assessment, AIMQ)，国际货币基金组织(international monetary fund, IMF)提出一种通用的统计数据质量评估框架(data quality assessment framework，DQAF)，此外还有十余种框架被提出[12]，应用于不同的专业领域中。

1.2　数据质量评估维度

数据质量的维度是指一个特征或部分信息，被用于分类信息和确定数据需求。Wang 等提出用维度集来衡量数据的质量，且提出了准确性、完整性、一致性、及时性、可靠性的维度[13-14]。文献[15]根据数据用户的需求，经过大量调研后确定了15个常用的质量维度。文献[16]将数据质量评估指标分成4类：内在质量、可访问性质量、上下文质量和表达质量，每一类又分为多种维度，从更广的角度评估数据的质量。

对数据质量的维度有多种选择，在具体的数据质量项目中，根据数据本身的特点和对数据质量的需求来选择合适的维度，并对各维度制定具体的评估规则来进行衡量。

1.3　数据质量评估流程

数据质量的评估流程如图1所示。数据质量评估模型主要包括以下5个步骤：

1）根据实际的需要，确定待进行质量评估的数据集。

2）根据待评估数据集的内容，分析影响数据质量的多种因素，选取合适的维度作为质量评估的指标。

3）根据所选取质量评估的指标，结合数据集本身的内容和特征，设置具体的评估规则，以此定量地计算出各评估指标的情况。

4）采用科学合理的方法计算出各个质量评估指标的权重。

5）根据各评估指标的得分和权重，对数据质量进行综合分析，得到数据质量的评分，确定数据质量的等级，定量地完成质量评估。

2 构建电力设备状态数据质量评估体系

2.1 电力设备状态数据内容及特点

电力设备的状态数据包括运行、运检、环境等方面[17-19]，可分为 3 大类：(1)电网运行和电力设备的检测或监测数据。(2)电力设备的运维、检修等数据。(3)环境的数据。电力设备的数据在统计、收集、整理、分析等环节都面临困难，数据质量急需解决的问题包括数据异常、数据缺失、数据冗余、数据精度和格式不规范、数据不及时等。电力设备的状态信息数据呈现以下特点[20-23]：

1）数据体量巨大、增长迅速：在大规模电力网络建设下，电力设备种类众多、数量大，智能巡检、在线监测、带电检测等设备状态数据积累迅速。

2）数据类型多样：电力设备的状态数据包括结构化数据、非结构化数据、历史/实时数据、GIS 空间数据。

3）数据来源不一：数据来源于各业务应用系统，包括能量管理系统、生产管理系统、地理信息系统等；数据的格式和接口不同，平台间难以共享。

2.2 数据质量评估指标

基于 Wang 的数据质量指标体系，结合电力设备状态数据自身的特征以及数据中存在的质量问题，本文选择以下 5 个指标：完整性(B_1)、准确性(B_2)、唯一性(B_3)、一致性(B_4)、及时性(B_5)，以此描述数据的质量，建立起质量评估体系，如图 2 所示，其中针对各个评估指标设置具体的评估规则。

完整性(B_1)：评估数据的完整程度，数据记录和数据项是否存在缺失和遗漏的现象。

准确性(B_2)：评估数据是否符合客观事实，是否准确描述实际信息。准确性规则包括数据的取值应在合理范围内，格式符合规范，长度和精度满足要求。数据的取值范围根据已知的标准数据确定，不超过理论最大值或理论最小值。

唯一性(B_3)：评估数据记录的唯一程度，是否存在冗余重复，包括数据记录和数据项的冗余。

一致性(B_4)：评估数据对同一实体的描述是否一致，包括数据内部逻辑的一致和函数的一致，是否存在数据冲突的现象。逻辑一致指数据集中某个体的某一指标不对应多个数值，表现为同一行数据内部不存在冲突；函数一致指同类别的数据不存在过大的差别，表现为不同行数据的分布一致。

及时性(B_5)：评估数据的时效性，及时性规则表现为数据的更新时间满足系统的要求。

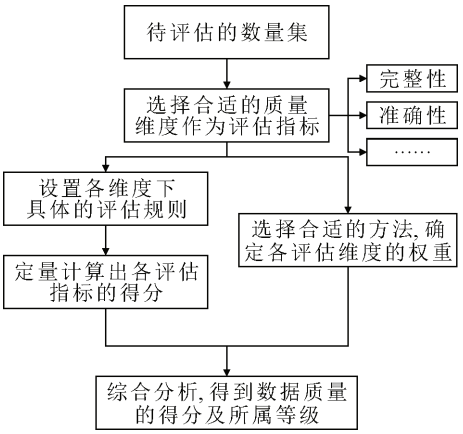


图 1 数据质量评估流程

Fig.1 Data quality assessment process

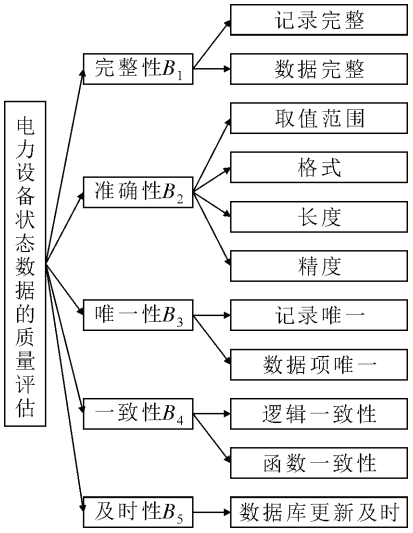


图 2 数据质量评估体系

Fig.2 Data quality assessment system

2.3 评价指标的定量计算

电力设备状态数据质量的定量指标描述数据集满足预先设定的质量标准要求的程度，并提供定量的质量信息。对数据集中的各项指标做如下的定量计算[24]。

1）完整性指标

数据集中实际的记录个数记为 R_a，实际的数据项个数记为 C_a，缺失的记录个数记为 R_s，缺失的数据项个数记为 C_s，空数据的个数记为 D_s。

数据记录的完整率用 S_{11} 表示：

$$S_{11}=(1-\frac{R_s}{R_a+R_s})\times 100\% \tag{1}$$

数据的完整率用 S_{12} 表示：

$$S_{12}=(1-\frac{D_s}{R_a(C_a-C_s)})\times 100\% \tag{2}$$

本文中将数据记录的完整程度和数据的完整程度对数据质量的影响程度视作相同，则完整性指标满足程度用 S_1 表示：

$$S_1=0.5S_{11}+0.5S_{12} \tag{3}$$

2）准确性指标

数据集中违反准确性规则的数据个数记为 D_q，则数据的准确性指标的满足程度 S_2 可表示为：

$$S_2=(1-\frac{D_q}{R_a(C_a-C_s)})\times 100\% \tag{4}$$

3）唯一性指标

数据集中冗余的数据记录个数记为 R_r，非冗余的数据项个数记为 R_{fr}，冗余的数据项的个数记为 C_r。则数据的唯一性指标的满足程度 S_3 可表示为：

$$S_3=(1-\frac{R_rC_a+R_{fr}C_r}{R_aC_a})\times 100\% \tag{5}$$

4）一致性指标

数据集中违反一致性规则的数据记录的个数为 R_c，则数据一致性性指标的满足程度 S_4 可表示为：

$$S_4=(1-\frac{R_c}{R_a})\times 100\% \tag{6}$$

5）及时性指标

数据集中违反及时性规则的数据记录的个数为 T_q，则数据一致性性指标的满足程度 S_5 可表示为：

$$S_5=(1-\frac{T_q}{R_a})\times 100\% \tag{7}$$

由以上计算式可定量地计算出完整性(B_1)、准确性(B_2)、唯一性(B_3)、一致性(B_4)、及时性(B_5)这 5 个数据质量评估指标的满足程度 S_1—S_5，取值范围在 0~1 之间，将其分别乘以 100，可以得到每个评估指标的得分情况 x_1—x_5。根据 1.3 中的数据质量评估流程，接下来需要确定各个评估指标的权重。

2.4　确定评估指标的权重

2.4.1　模糊层次分析法

模糊层次分析法(FAHP)是在层次分析法的基础上引入了模糊的理论，用以改进层次分析法主观性强、科学性弱的不足，解决判断矩阵一致性难以满足的问题。

模糊层次分析法将复杂的评价问题层次化，根据问题的具体情况，确定出目标层、准则层，如图 3 所示。将同一准则层中各因素的重要程度进行两两比较，采用 0.1~0.9 的标度方法[25]，具体的比较原则和标度方法见表 1，根据此建立起两层次之间的模糊判断矩阵 $\boldsymbol{A}=(a_{ij})_{n\times n}(i,j=1,2,\cdots,n)$。$n$ 为判断矩阵所对应的下层的因素数量，数据质量评估的问题中，即为评估指标的个数。

模糊判断矩阵 $\boldsymbol{A}$ 中的元素 a_{ij} 在 0~1 范围内取值，满足 $a_{ij}+a_{ji}=1$，且 $a_{ii}=0.5$，则矩阵 $\boldsymbol{A}$ 为模糊互补矩阵。根据下式计算模糊一致矩阵 $\boldsymbol{F}=(f_{ij})_{n\times n}(i,j=1,2,\cdots,n)$。

$$\begin{cases} f_i=\sum_{k=1}^{n}a_{ik} \\ f_{ij}=\dfrac{f_i-f_j}{2n}+0.5 \end{cases} \tag{8}$$

模糊一致矩阵 $\boldsymbol{F}$ 中的元素满足：

$$f_{ij}=f_{ik}-f_{jk}+0.5 \tag{9}$$

模糊层次分析法计算权重式如下所示：

$$\begin{cases} \beta=\dfrac{n-1}{2} \\ w_{fi}=\dfrac{1}{n\beta}\sum_{j=1}^{n}f_{ij}+\dfrac{1}{n}-\dfrac{1}{2\beta} \end{cases} \tag{10}$$

式中：β 为差异因子，按照式中取值时权重的差异最为合理；w_{fi} 是第 i 个指标所对应的权重，则模糊层次分析法求得的权重向量 $\boldsymbol{w}_f$ 为：

$$\boldsymbol{w}_f=[w_{f1},w_{f2},\cdots,w_{fn}] \tag{11}$$

2.4.2　熵权法

熵权法依据各指标包含信息量的多少来确定指标权重的大小，某指标的熵越小，该指标值的变异程度越大，包含的信息量越多，则该指标的权重

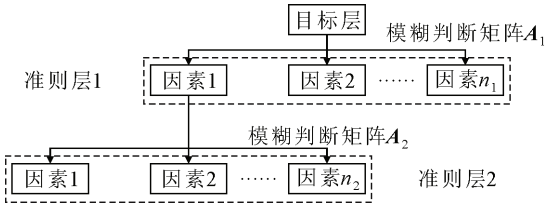


图 3　模糊层次分析法结构

Fig.3　Structure of FAHP

表 1　比较原则及标度方法

Table 1　Principles of comparison and scaling methods

a_{ij}	因素 i 与因素 j 比较	a_{ji}
0.5	同样重要	0.5
0.6	稍微重要	0.4
0.7	明显重要	0.3
0.8	强烈重要	0.2
0.9	极端重要	0.1

也越大，在综合评估中起的作用越大。

先采用德尔菲调查法，收集多位专家对数据质量评估指标的重要性的评分，得到评估矩阵 $\boldsymbol{U}=(u_{ij})_{n\times m}(i=1,2,\cdots,n;\ j=1,2,\cdots,m)$，其中 n 为数据质量评估指标的个数，m 为打分专家的个数。u_{ij} 是第 j 位专家对第 i 个指标的重要性的评估值，采用十标度法。

先按照下式对评估矩阵 $\boldsymbol{U}$ 进行归一化处理：

$$z_{ij}=\frac{u_{ij}-\min(\boldsymbol{U}(*,j))}{\max(\boldsymbol{U}(*,j))-\min(\boldsymbol{U}(*,j))} \tag{12}$$

得到标准化评估矩阵 $\boldsymbol{Z}=(z_{ij})_{n\times m}$，根据信息熵的理论，可按照下式求取信息熵 e_j：

$$\begin{cases} e_j=-\dfrac{1}{\ln n}\displaystyle\sum_{i=1}^{n}p_{ij}\ln p_{ij} \\ p_{ij}=\dfrac{z_{ij}}{\displaystyle\sum_{i=1}^{n}z_{ij}} \end{cases} \tag{13}$$

若 p_{ij} 为 0，则定义 e_j 也为 0。根据信息熵向量可以根据式(14)得到第 j 个专家对应的权重：

$$h_j=\frac{1-e_j}{m-\displaystyle\sum_{j=1}^{m}e_j} \tag{14}$$

专家的权重向量 $\boldsymbol{H}=[h_1,\ h_2,\ldots h_m]$，由评估矩阵 $\boldsymbol{U}$ 和专家的权重向量 $\boldsymbol{H}$ 可得多位专家对各个质量评估指标重要性的综合评分 $\boldsymbol{W}=\boldsymbol{U}\times\boldsymbol{H}$，再对 $\boldsymbol{W}$ 进行归一化处理，即可得到用熵权法计算得出的各评估指标的权重向量 $\boldsymbol{w}_{\mathrm{e}}$：

$$\boldsymbol{w}_{\mathrm{e}}=[w_{\mathrm{e}1},w_{\mathrm{e}2},\cdots,w_{\mathrm{e}n}] \tag{15}$$

2.4.3 求取组合权重

组合应用模糊层次分析法和熵权法来确定权重的过程中，既可以兼顾专家经验，又可以减少主观因素对结果的影响程度，更为客观、合理地确定每个评估指标的权重。

根据上述内容，由模糊层次分析法求得的权重是 $\boldsymbol{w}_{\mathrm{f}}$，由熵权法求得的权重是 $\boldsymbol{w}_{\mathrm{e}}$，采用几何平均的方法来求取组合权重 $\boldsymbol{w}$：

$$\boldsymbol{w}=(w_1,w_2,\ldots,w_n) \tag{16}$$

其中 $\boldsymbol{w}_i(i=1,2,\cdots,n)$的计算式如下：

$$w_i=\frac{\sqrt{w_{\mathrm{f}i}w_{\mathrm{e}i}}}{\displaystyle\sum_{i=1}^{n}\sqrt{w_{\mathrm{f}i}w_{\mathrm{e}i}}} \tag{17}$$

由 2.3 节和 2.4 节可以求出各评估指标的具体得分值及所占权重，但目前对于数据质量优劣程度的判定还没有标准规定，难以用一个准确的数值来评判出具体的等级，本文采用模糊综合评价法来构建数据质量评估的等级标准。

2.5 模糊综合评价

2.5.1 模糊综合评价的因素集和评价集

本文中采用模糊综合评价法进行数据质量的评估，综合评价的因素集即为 $B=\{B_1,B_2,\cdots,B_5\}$，为数据质量评估指标的集合。将电力设备状态数据的质量分为 5 个等级：优秀(v_1)、良好(v_2)、中等(v_3)、较差(v_4)、极差(v_5)，构成评价集 $V=\{v_1,v_2,\ldots,v_5\}$。具体的等级划分规则如下所示：

优秀(v_1)：数据质量评分在[95,100]区间内。

良好(v_2)：数据质量评分在[85,90]区间内。

中等(v_3)：数据质量评分在[75,80]区间内。

较差(v_4)：数据质量评分在[65,70]区间内。

极差(v_5)：数据质量评分在[0,60]区间内。

2.5.2 模糊关系矩阵

针对评价集 $V=\{v_1,\ v_2,\ \cdots,\ v_5\}$中不同等级的阈值，选择梯形隶属度函数，将因素集中 5 个数据质量评估指标的得分 x_1~x_5 代入不同等级的隶属度函数进行计算，可以得到模糊关系矩阵 $\boldsymbol{R}=(r_{ij})_{5\times5}$，隶属度求取规则见附录 A。

2.5.3 实现模糊综合评价

根据上述过程得到模糊关系矩阵 $\boldsymbol{R}$，再结合权重 w，按式(18)得出综合评估向量 $\boldsymbol{D}$：

$$\boldsymbol{D}=\boldsymbol{w}\times\boldsymbol{R}=[d_1,d_2,d_3,d_4,d_5] \tag{18}$$

式中：d_i 对应着评价集 V 中的 5 个等级的隶属度，基于最大隶属度原则，可以确定数据质量的等级，完成对数据质量的评估。

3 算例分析

3.1 某地油色谱状态监测数据

选取某地电网公司的变压器油色谱数据集，共有 994 条数据记录，每条记录有 8 个数据项，包括氢气、甲烷、乙烯等气体的体积分数，部分数据如表 2 所示。

原始的油色谱数据集满足要求，不存在数据空缺、数据冗余、数据冲突的情况，数据值在合理的范围之内。在原始数据集的基础上，按照 GB/T 7252《变压器油中溶解气体分析和判断导则》的规定，根据油中溶解的多种气体的含量，可以判断变压器内部出现的故障，将变压器的状态分为 9 类，并用数字 0—8 来表示，其中 0 表示变压器是正常状态，1—8 分别表示故障类型为：低能放电故障、高能放

电故障、低能放电兼过热故障、中温放热故障、局部放电故障、高温过热故障、低温过热故障、高能放电兼过热故障。

本算例中通过变压器故障诊断的正确率来反映油色谱数据的质量情况。先采用 BP 神经网络对原始的油色谱数据和判定得到的故障结果进行学习和训练，将 8 种气体的体积分数作为神经网络的输入层，输出层为 9 种故障类型，选择前 800 条数据作为训练集，后 194 条数据作为测试集，得到原始油色谱数据集的故障诊断正确率为 91.75%。

3.2　评估指标权重确定

从 B_1—B_5 这 5 个指标对数据的质量进行综合评估，先采用模糊层次分析法，根据专家意见形成模糊判断矩阵如下：

$$\boldsymbol{A}=\begin{bmatrix}0.5 & 0.1 & 0.7 & 0.2 & 0.8\\ 0.9 & 0.5 & 0.9 & 0.9 & 0.9\\ 0.3 & 0.1 & 0.5 & 0.1 & 0.8\\ 0.8 & 0.1 & 0.9 & 0.5 & 0.9\\ 0.2 & 0.1 & 0.2 & 0.1 & 0.5\end{bmatrix}$$

依据模糊层次分析法求取权重的方法，可求得各评估指标的权重为 $\boldsymbol{w}_{\mathrm{f}}$：

$$\boldsymbol{w}_{\mathrm{f}}=[0.19,0.28,0.165,0.235,0.13]$$

再采用熵权法，选取 10 位专家对 5 个评估指标的重要性按照 10 标度的原则打分，得到如下所示评估矩阵：

$$\boldsymbol{U}=\begin{bmatrix}3 & 4 & 4 & 5 & 3 & 3 & 3 & 5 & 3 & 4\\ 8 & 8 & 7 & 9 & 7 & 8 & 8 & 8 & 9 & 7\\ 4 & 3 & 2 & 3 & 4 & 2 & 4 & 3 & 3 & 4\\ 7 & 6 & 5 & 7 & 5 & 7 & 6 & 6 & 6 & 5\\ 3 & 2 & 3 & 2 & 3 & 2 & 3 & 2 & 2 & 3\end{bmatrix}$$

对评估矩阵进行归一化处理再计算信息熵，先得到专家的权重向量 $\boldsymbol{H}$：

$$\boldsymbol{H}=[0.1296,0.0787,0.0718,0.0773,0.1272,\\ 0.1345,0.1309,0.0720,0.1010,0.0771]^{\mathrm{T}}$$

再根据专家评分 $\boldsymbol{U}$ 和专家权重向量 $\boldsymbol{H}$，计算得到各评估指标的权重 $\boldsymbol{w}_{\mathrm{e}}$：

$$\boldsymbol{w}_{\mathrm{e}}=[0.1514,0.3393,0.1399,0.2604,0.1089]$$

将以上两种方法求得的权重求几何平均，得到最终的权重 $\boldsymbol{w}$：

$$\boldsymbol{w}=[0.1703,0.3094,0.1525,0.2483,0.1195]$$

3.3　数据质量评估

原始的油色谱数据集的质量符合要求，在此数据集中增添问题数据，进行以下操作，同时改变数据的完整性、准确性、唯一性、一致性，以改变数据的质量，再用本文提出的数据质量评估方法评估修改后的数据集。

1）改变数据的完整性，删除原始数据中的 200 条记录，并设置 1500 个空数据。

2）改变数据的准确性，在原始数据上设置 2000 个违反准确性规则的数据。

3）改变数据的唯一性，在原始数据上添加 200 个冗余的数据记录和 1 个冗余的数据项。

4）改变数据的一致性，在原始数据上设置 100 条不符合一致性规则的数据记录。

按照前文的定量计算式可得数据质量各个评估指标的得分 x_1—x_5，如表 3 所示。

根据各个评估指标的得分，可求出模糊关系矩阵 $\boldsymbol{R}$ 为：

$$\boldsymbol{R}=\begin{bmatrix}0 & 0.648 & 0.352 & 0 & 0\\ 0 & 0 & 1 & 0 & 0\\ 0 & 0 & 0.2 & 0.8 & 0\\ 0 & 1 & 0 & 0 & 0\\ 1 & 0 & 0 & 0 & 0\end{bmatrix}$$

则综合评估向量 $\boldsymbol{D}$ 为：

$$\boldsymbol{D}=\boldsymbol{w}\times\boldsymbol{R}=[0.1195,0.3857,0.3998,0.1220,0]$$

根据最大隶属度原则，可以从评估向量 $\boldsymbol{D}$ 判断更改后的数据质量等级属于“中等”。也可以计算出数据质量的综合评分 $\boldsymbol{S}$：

表 2　油色谱数据及其故障类型

Table 2　chromatographic data and fault types

故障类型	体积分数/10^{-6}							
	H_2	CH_4	C_2H_2	C_2H_4	C_2H_6	总烃	CO	CO_2
0	17.569	0.496	0	0.189	0	0.685	21.328	114.744
1	60	0.4	0	0	0	0.4	228	0
2	695.9	14.1	10	26.7	3.6	54.4	3172.8	29.2
3	70.43	4.97	0	1.22	1.33	7.52	335.65	104.78
4	42.98	7.54	1.89	2.46	8.35	20.24	648.7	5158.6
5	4166	280	2.53	0.63	35.59	318.75	61.63	405.3
6	440	487	0.9	462	157	1106.9	1278	210
7	176	41.35	0	20.36	71.74	133.45	704.33	4322.29
8	16	29.17	4.96	41.5	4.39	80.02	3	156

表 3　数据质量各指标得分

Table 3　Scores of each index of data quality

指标	得分	
完整性	R_s=200 D_s=1500	x_1=83.24
准确性	D_q=2000	x_2=77.64
唯一性	R_r=200 C_r=1	x_3=71
一致性	R_c=100	x_4=89.94
及时性	T_q=0	x_5=100

$\boldsymbol{S}=[83.24,77.64,71,89.94,100]\cdot$
$[0.1703,0.3094,0.1399,0.2604,0.1089]^{\mathrm{T}}=77.15$

综合评分为 77.15，在 75~80 之间，即数据质量为“中等”。

3.4 验证数据质量评估结果

为直观反映更改后油色谱的数据质量，再次使用 BP 神经网络对其进行学习训练，选择 80%的数据为训练集，神经网络的输入层为添加问题数据后 8 种气体的体积分数，数据集中的空缺值置为 0，输出层仍为根据原始数据集判定得到的 9 种变压器故障类型；将前 800 条数据作为训练集，后 194 条数据作为测试集，进行故障诊断分类。BP 神经网络的训练情况和分类结果见图 4—图 6。

图 4 中横坐标为实际的故障类型，纵坐标为 BP 神经网络输出的结果，由图 4 可知训练集、测试集和总体样本的数据相关性系数均在 0.8 左右；图 5 中给出 BP 神经网络的损失函数；图 6 中以原始数据的故障类型结果为参考，给出添加问题数据后的油色谱数据集得到的故障分类结果，此时的诊断率为 72.68%，比原始数据的故障诊断率低，说明改变后的油色谱数据质量明显下降，可以证明此数据质量评估的结果有效。

4 结论

1）本文针对电力设备的状态数据搭建了数据质量评估模型，从完整性、准确性、唯一性、一致性、及时性 5 个维度设置了质量评估体系和定量计算的评估规则，多角度多方位进行分析。

2）求取各评估指标的权重时，组合运用了模糊层次分析法和熵权法，将主客观方法结合起来，提高科学性和合理性。

3）采用模糊综合评价的方法将数据质量划分为 5 个等级，通过建立模糊关系矩阵，计算模糊综合评估向量来确定数据质量的所属等级和数据质量的综合评分，给用户直观反映。

4）选取某地电网公司的变压器油色谱数据作为算例，通过添加问题数据来降低数据集的质量，再利用本文提出的方法对其进行评估，以油色谱数据的故障诊断率来反映数据的质量，原始数据的故障诊断率为 91.75%，添加问题数据后的故障诊断率为 72.68%。本文所提方法计算出更改后的数据质量处于“中等”水平，得分为 77.15，与实际应用场景相符，证明了本文方法的可行性。

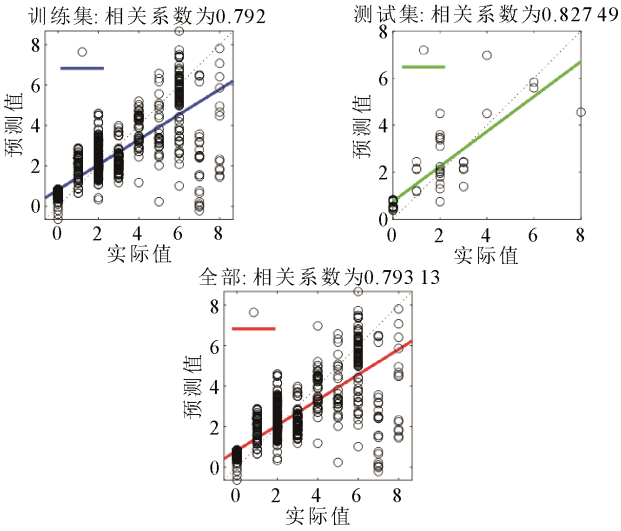


图 4 各样本相关性分析

Fig.4 Correlation analysis of each sample

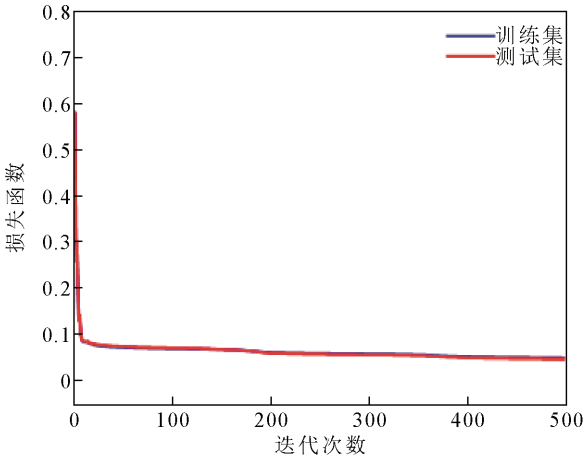


图 5 损失函数曲线

Fig.5 Loss function curves

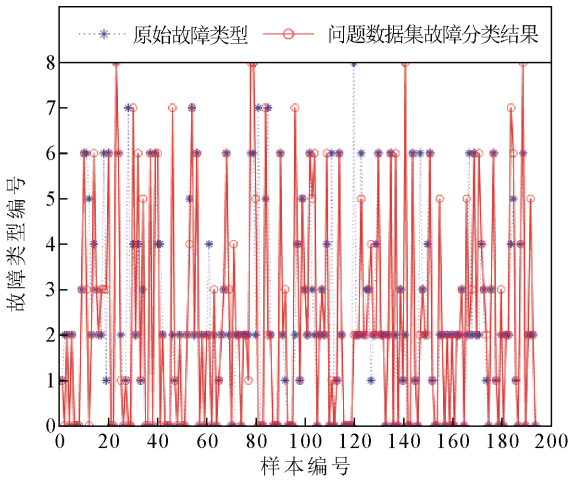


图 6 故障诊断结果

Fig.6 Fault diagnosis results

在后续的研究中将继续细化数据质量的评估指标，做进一步的拓展，设计多层次的评估方法，

同时设置更加合理更加规范的评估规则，扩大质量评估方法的应用范围。

附录见本刊网络版(http://hve.epri.sgcc.com.cn)。

参考文献　References

[1] 李建中，刘显敏. 大数据的一个重要方面：数据可用性[J]. 计算机研究与发展，2013，50(6)：1147-1162.
LI Jianzhong, LIU Xianmin. An important aspect of big data: data usability[J]. Journal of Computer Research and Development, 2013, 50(6): 1147-1162.

[2] SWARTZ N. Gartner warns firms of ‘dirty data’[J]. Information Management Journal, 2007, 41(3): 6.

[3] 高　志，樊锐轶，耿少博，等. 基于大数据技术的电力数据质量评估数据框架研究[J]. 电子器件，2022，45(1)：194-198.
GAO Zhi, FAN Ruiyi, GENG Shaobo, et al. Research on data framework of power data quality assessment based on big data technology[J]. Chinese Journal of Electron Devices, 2022, 45(1): 194-198.

[4] LIU H, HUANG F P, LI H, et al. A big data framework for electric power data quality assessment[C]// 2017 14th Web Information Systems and Applications Conference (WISA). Liuzhou, China: IEEE, 2017: 289-292.

[5] 秦　璇. 电力统计数据的质量评估及其异常检测方法研究[D]. 长沙：长沙理工大学，2013.
QIN Xuan. Data quality evaluation and anomaly detection research based on the statistical data of power system[D]. Changsha, China: Changsha University of Science and Technology, 2013.

[6] 简晓雯. 基于机器学习的传感数据质量评估研究及评估平台构建[D]. 北京：北京邮电大学，2021.
JIAN Xiaowen. Research on data quality assessment and construction of assessment platform based on machine learning[D]. Beijing, China: School of Electronic Engineering, 2021.

[7] 李　刚，焦亚菲，刘福炎，等. 联合采用熵权和灰色系统理论的电力大数据质量综合评估[J]. 电力建设，2016，37(12)：24-31.
LI Gang, JIAO Yafei, LIU Fuyan, et al. Comprehensive evaluation of big data quality in power systems with entropy weight and grey system theory[J]. Electric Power Construction, 2016, 37(12): 24-31.

[8] 潘　旭，王金丽，赵晓龙，等. 智能配电网多维数据质量评价方法[J]. 中国电机工程学报，2018，38(5)：1375-1384.
PAN Xu, WANG Jinli, ZHAO Xiaolong, et al. Multi dimensional data quality evaluation method for intelligent distribution network[J]. Proceedings of the CSEE, 2018, 38(5): 1375-1384.

[9] 李建林，郭兆东，马速良，等. 新型电力系统下“源网荷储”架构与评估体系综述[J]. 高电压技术，2022，48(11)：4330-4341.
LI Jianlin, GUO Zhaodong, MA Suliang, et al. Overview of the “Source-grid-load-storage” architecture and evaluation system under the new power system[J]. High Voltage Engineering, 2022, 48(11): 4330-4341.

[10] 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.
YANG Qingyun, ZHAO Peiying, YANG Dongqing, et al. Research on data quality assessment methodology[J]. Computer Engineering and Applications, 2004, 40(9): 3-4, 15.

[11] LUEBBERS D, GRIMMER U, JARKE M. Systematic development of data mining-based data quality tools[C]// Proceedings of the 29th International Conference on Very Large Data Bases. Berlin, Germany: VLDB Endowment, 2003: 548-559.

[12] BATINI C, CAPPIELLO C, FRANCALANCI C, et al. Methodologies for data quality assessment and improvement[J]. ACM Computing Surveys, 2009, 41(3): 16.

[13] WANG R Y, STOREY V C, FIRTH C P. A framework for analysis of data quality research[J]. IEEE Transactions on Knowledge and Data Engineering, 1995, 7(4): 623-640.

[14] WAND Y, WANG R Y. Anchoring data quality dimensions in ontological foundations[J]. Communications of the ACM, 1996, 39(11): 86-95.

[15] WANG R Y, STRONG D M. Beyond accuracy: what data quality means to data consumers[J]. Journal of Management Information Systems, 1996, 12(4): 5-33.

[16] RAHM E, DO H H. Data cleaning: problems and current approaches[J]. IEEE Data Engineering Bulletin, 2000, 23(4): 3-13.

[17] 姜　骞，刘亚东，严英杰，等. 电力设备多源异构数据空间合成与立体展示方法[J]. 高电压技术，2022，48(1)：66-74.
JIANG Qian, LIU Yadong, YAN Yingjie, et al. Spatial synthesis and stereoscopic display method for multi-source heterogeneous data of power equipment[J]. High Voltage Engineering, 2022, 48(1): 66-74.

[18] 张东霞，苗　新，刘丽平，等. 智能电网大数据技术发展研究[J]. 中国电机工程学报，2015，35(1)：2-12.
ZHANG Dongxia, MIAO Xin, LIU Liping, et al. Research on development strategy for smart grid big data[J]. Proceedings of the CSEE, 2015, 35(1): 2-12.

[19] 刘亚东，陈　思，丛子涵，等. 电力装备行业数字孪生关键技术与应用展望[J]. 高电压技术，2021，47(5)：1539-1554.
LIU Yadong, CHEN Si, CONG Zihan, et al. Key technology and application prospect of digital twin in power equipment industry[J]. High Voltage Engineering, 2021, 47(5): 1539-1554.

[20] 宋亚奇，周国亮，朱永利. 智能电网大数据处理技术现状与挑战[J]. 电网技术，2013，37(4)：927-935.
SONG Yaqi, ZHOU Guoliang, ZHU Yongli. Present status and challenges of big data processing in smart grid[J]. Power System Technology, 2013, 37(4): 927-935.

[21] 江秀臣，许永鹏，李曜丞，等. 新型电力系统背景下的输变电数字化转型[J]. 高电压技术，2022，48(1)：1-10.
JIANG Xiuchen, XU Yongpeng, LI Yaocheng, et al. Digitalization transformation of power transmission and transformation under the back-ground of new power system[J]. High Voltage Engineering, 2022, 48(1): 1-10.

[22] 严英杰，盛戈皞，陈玉峰，等. 基于大数据分析的输变电设备状态数据异常检测方法[J]. 中国电机工程学报，2015，35(1)：52-59.
YAN Yingjie, SHENG Gehao, CHEN Yufeng, et al. An method for anomaly detection of state information of power equipment based on big data analysis[J]. Proceedings of the CSEE, 2015, 35(1): 52-59.

[23] 彭小圣，邓迪元，程时杰，等. 面向智能电网应用的电力大数据关键技术[J]. 中国电机工程学报，2015，35(3)：503-511.
PENG Xiaosheng, DENG Diyuan, CHENG Shijie, et al. Key technologies of electric power big data and its application prospects in smart grid[J]. Proceedings of the CSEE, 2015, 35(3): 503-511.

[24] 焦亚菲. 基于大数据的电力系统信息质量评估[D]. 北京：华北电力大学，2017.
JIAO Yafei. Information quality assessment of power system based on the big data[D]. Beijing, China: North China Electric Power University, 2017.

[25] AZADFALLAH M. AHP rank reversal, the impacts of scale, normalization and aggregation rules[J]. National Journal of System and Information Technology, 2014, 7(2): 113-126.



计　蓉
1999—，女，硕士生
研究方向为电力设备数据质量评估及提升方法
E-mail: jirong123@sjtu.edu.cn

JI Rong



侯慧娟(通信作者)
1984—，女，博士，助理研究员
从事电力设备绝缘在线监测及故障诊断的研究
E-mail: houhuijuan@sjtu.edu.cn

HOU Huijuan
Ph.D.
Associate professor
Corresponding author

收稿日期　2022-12-02　修回日期　2023-03-10　　编辑　卫李静

附录 A

隶属度 r_{ij} 求取规则如下所示：

$$r_{i1}=\begin{cases}1, x_i \geq 95\\ \dfrac{x_i-90}{95-90}, 90 \leq x_i < 95\\ 0, x_i < 90\end{cases} \tag{A1}$$

$$r_{i2}=\begin{cases}0, x_i \geq 95\\ \dfrac{95-x_i}{95-90}, 90 \leq x_i < 95\\ 1, 85 \leq x_i < 90\\ \dfrac{x_i-80}{85-80}, 80 \leq x_i < 85\\ 0, x_2 < 80\end{cases} \tag{A2}$$

$$r_{i3}=\begin{cases}0, x_i \geq 85\\ \dfrac{85-x_i}{85-80}, 80 \leq x_i < 85\\ 1, 75 \leq x_i < 80\\ \dfrac{x_i-70}{75-70}, 70 \leq x_i < 75\\ 0, x_i < 70\end{cases} \tag{A3}$$

$$r_{i4}=\begin{cases}0, x_i \geq 75\\ \dfrac{75-x_i}{75-70}, 70 \leq x_i < 75\\ 1, 65 \leq x_i < 70\\ \dfrac{x_i-60}{65-60}, 60 \leq x_i < 65\\ 0, x_i < 60\end{cases} \tag{A4}$$

$$r_{i5}=\begin{cases}0, x_i \geq 65\\ \dfrac{65-x_i}{65-60}, 60 \leq x_i < 65\\ 1, x_i < 60\end{cases} \tag{A5}$$