Improvement of Gaussian Distribution Based on Apriori Algorithm for Eliminating Differences in Transformer Monitoring Data with High Frequency Access*

WANG Jie[1,2*], SHI Qian[1,2], CHEN Gang[1,2]
(1.Electric Power Research Institute of State Grid Xinjiang Electric Power Co., Ltd., Urumqi Xinjiang 830013, China;
2.Xinjiang Key Laboratory of Extreme Environment Operation and Detection Technology of Power Transmission & Transformation Equipment, Urumqi Xinjiang 830013, China)

Abstract: The quality of transformer monitoring data is closely related to the safety of transformer operation. Therefore, a method based on Apriori algorithm to improve the Gaussian distribution for eliminating differences in high-frequency access transformer monitoring data is studied to improve the quality of transformer monitoring data. A compression matrix is introduced into the Apriori algorithm to improve it. By improving the Apriori algorithm, high-frequency access data within historical transformer monitoring data are mined. By improving the Gaussian process model of Gaussian distribution, the difference data in the high-frequent access transformer monitoring data are detected and eliminated, and the high-frequent access transformer monitoring data difference is eliminated. Experimental results show that this method can effectively mine high-frequency access to transformer monitoring data. This method can effectively detect differential data within high-frequency access transformer monitoring data, complete differential data elimination, and the accuracy of differential data elimination of this method is high.

Key words: Apriori algorithm; improved Gaussian distribution; high frequency access; transformer; monitoring data; difference elimination
EEACC: 7220; 8140 **doi**: 10.3969/j.issn.1005-9490.2026.04.028

基于 Apriori 算法改进高斯分布的高频繁访问变压器监控数据差异消除方法*

王 洁[1,2*], 石 倩[1,2], 陈 刚[1,2]
(1.国网新疆电力有限公司电力科学研究院, 新疆 乌鲁木齐 830013;
2.新疆输变电设备极端环境运行与检测技术重点实验室, 新疆 乌鲁木齐 830013)

摘 要: 变压器监控数据质量与变压器运行安全性息息相关, 为此, 研究基于 Apriori 算法改进高斯分布的高频繁访问变压器监控数据差异消除方法, 提升变压器监控数据质量。在 Apriori 算法内, 引入压缩矩阵, 改进 Apriori 算法; 通过改进 Apriori 算法, 在历史变压器监控数据内, 挖掘高频繁访问变压器监控数据; 通过改进高斯分布的高斯过程模型, 在高频繁访问变压器监控数据内, 检测出差异数据并消除, 完成高频繁访问变压器监控数据差异消除。实验证明: 该方法可有效挖掘高频繁访问变压器监控数据; 可有效检测高频繁访问变压器监控数据内的差异数据, 完成差异数据消除, 且该方法监控数据差异消除精度较高。

关键词: Apriori 算法; 改进高斯分布; 高频繁访问; 变压器; 监控数据; 差异消除
中图分类号: TM411 **文献标识码**: A **文章编号**: 1005-9490(2026)04-0894-06

变压器监控数据质量直接影响数据的使用效果, 尤其是变压器监控数据内的差异数据[1-3], 会直接降低变压器监控精度, 影响变压器运行的安全可靠性。为此需要及时消除差异数据, 提升变压器监控数据质量[4-5], 提高变压器监控准确性, 确保变压器稳定运行。例如, 刘云鹏等[6]通过变分模态分解法, 剔除变压器监控数据时间序列内的趋势项, 利用 3σ 准则对去除趋势项的监控数据进行差异数据识别, 并消除识别到的差异数据。该方法可有效识别变压器监控数据的差异数据, 完成变压器监控数据差异消除。许海林等[7]利用 Shapelet 发现算法, 建立时间序列决策树, 在该决策树内输入变压器监控数据, 输出差异数据, 并消除该差异化数据。该方法对差异化数据识别的灵敏度较高, 可有效消除变压器监测数据的差异数据。但上述方法的泛化能力均较差, 且易受数据维度影响, 在处理大量变压器监控

项目来源: 国网新疆电科院 2022 年第二批成本专项计划项目(新电科研科〔2022〕8 号)
收稿日期: 2024-03-13 **修改日期**: 2024-05-20

数据时,差异数据检测数据量较大,需要处理的差异数据较多,因此消除时间较长。Apriori 算法是以多次扫描数据库的方式,获取频繁项集[8],可在变压器监控数据内,快速挖掘高频繁访问的数据集,加快差异数据消除速度。高斯过程模型属于高斯分布算法的改进,具备较优的适用性与泛化性能[9]。为此,研究基于 Apriori 算法改进高斯分布的高频繁访问变压器监控数据差异消除方法,提升压器监控数据差异消除精度。

1　高频繁访问变压器监控数据差异消除方法

1.1　基于 Apriori 算法的高频繁访问变压器监控数据挖掘

利用 Apriori 算法,在历史变压器监控数据内,挖掘高频繁访问变压器监控数据。令高频繁访问变压器监控数据挖掘项目集合为 $U=\{u_1,u_2,\cdots,u_m\}$;变压器监控事务数据库为 $G=\{g_1,g_2,\cdots,g_n\}$;其中,项目数量是 m;变压器监控事务数据样本数量是 n;各历史变压器监控事务 $g_i(i=1,2,\cdots,n)$ 均属于 U 内一组项目集合,即 $g_i\in U$,其所存在项目个数叫作事务长度。一条高频繁访问变压器监控数据挖掘的关联规则形式为 $X\Rightarrow Y$,其中,变压器监控数据项集 $X\subseteq U$;变压器监控数据项集 $Y\subseteq U$;$X\cap Y=\varnothing$;$\varnothing$ 为空集;$X\Rightarrow Y$ 需符合的条件为:

①存在支持度 $S(X\Rightarrow Y)=\dfrac{N(X\cup Y)}{\|T\|}$,其中,$X\cup Y$ 在时间序列 T 内的出现次数是 $N(X\cup Y)$;历史变压器监控事务数据库记录总数是 $\|G\|$。通过用户设置最小支持度 min$S$。

②存在置信度 $C(X\Rightarrow Y)=\dfrac{N(X\cup Y)}{N(X)}$,其中,$X$ 在 G 内的出现次数是 $N(X)$。通过用户设置最小置信度 min$C$。

同时符合 min$S$ 与 min$C$ 的规则即强关联规则。

Apriori 算法的定理如下:

定理 1　变压器监控数据的频繁 1-项集 L_1 内包含 m 个项集 $\{U_1\},\{U_2\},\cdots,\{U_m\}$,依据支持度计数从小到大排序 $\{U_1\},\{U_2\},\cdots,\{U_m\}$,获取 $\{u_1\},\{u_2\},\cdots,\{u_m\}$,那么排列后的 L_1 会产生很小的候选高频繁访问变压器监控数据 k-项集(k>3)。

定理 2　如果变压器监控数据项集,在高频繁访问变压器监控数据的频繁 k-项集 L_k 内,反复出现的次数低于 k,该项集不属于高频繁访问变压器监控数据的频繁 k+1-项集 L_{k+1} 内的元素。

定理 3　如果某变压器监控事务,存在项目数是 k,那么该事务便存在项数超过 k 的频繁项集。

定理 4　如果某高频繁访问变压器监控数据的频繁 k-项集 L_k 内,项集数量低于 k+1,那么该频繁项集就是最大高频繁访问变压器监控数据的频繁项集[10]。

利用 Apriori 算法,在历史变压器监控数据内,挖掘高频繁访问变压器监控数据的具体步骤如图 1 所示。

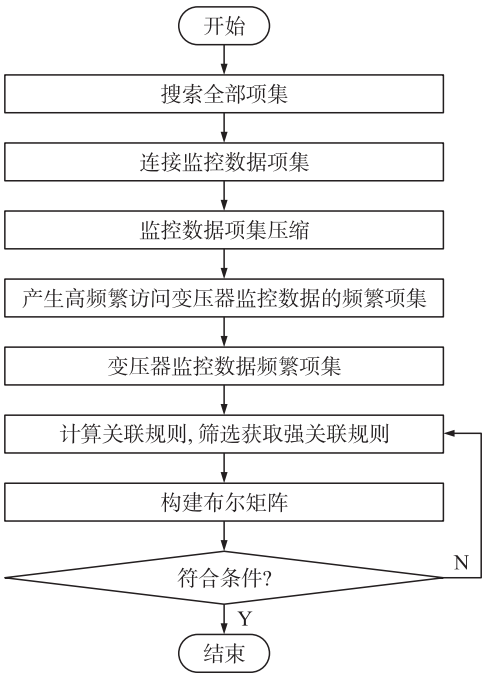

图 1　挖掘高频繁访问变压器监控数据流程图

具体内容如下:

步骤 1　在历史变压器监控数据内,搜索全部 S 大于 min$S$ 的项集,得到变压器监控数据的频繁 1-项集 L_1。

步骤 2　利用 $L_{k-1}(k\geqslant 2)$ 自身连接得到候选高频繁访问变压器监控数据项集 Z_k。

步骤 3　依据频繁项集性质,压缩 Z_k。

步骤 4　产生高频繁访问变压器监控数据的频繁 k-项集 L_k。

步骤 5　反复操作步骤 2 至步骤 4,直到不再生成高频繁访问变压器监控数据的频繁项集。

步骤 6　依据步骤 5 得到的变压器监控数据频繁项集,计算关联规则,并筛选获取强关联规则[11],用于挖掘高频繁访问变压器监控数据。

但 Apriori 算法在压缩候选高频繁访问变压器监控数据项集时,压缩得不够彻底,会浪费存储空间[12],为此,在 Apriori 算法内,引入压缩矩阵,提升候选高频繁访问变压器监控数据项集压缩效果[13]。

U 与 G 的矩阵即布尔矩阵 H,表达形式为:

$$H=\begin{pmatrix} h_{11} & h_{12} & \cdots & h_{1n} \\ h_{21} & h_{22} & \cdots & h_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ h_{m1} & h_{m2} & \cdots & h_{mn} \end{pmatrix} \tag{1}$$

式中：H 内行为变压器监控事务 g_i，列为高频繁访问变压器监控数据挖掘项目 u_j，同时符合 $h_{ij}=\begin{cases}0 & u_j \notin g_i \\ 1 & u_j \in g_i\end{cases}$ 条件。

u_j 的支持数为：

$$\mathrm{S_C}(u_j)=\sum_{i=1}^{m}\sum_{j=1}^{n} w_i h_{ij} \tag{2}$$

式中：权重是 w_i。

H 内，高频繁访问变压器监控数据挖掘项目 u_s 与 u_j 的支持度，就是 u_s 与 u_j 在同一水平列上都出现 1 的次数，u_s 与 u_j 按位展开“与”运算后求和：

$$\mathrm{S_C}(u_s,u_j)=\sum_{i=1}^{m}\sum_{j=1}^{n} w_i(h_{is}\wedge h_{ij}) \tag{3}$$

利用改进 Apriori 算法，在历史变压器监控数据内，挖掘高频繁访问变压器监控数据的具体步骤如下：

步骤 1　扫描变压器监控事务数据库 G，并转换成相应的布尔矩阵 H。

步骤 2　产生变压器监控数据的频繁 1-项集 L_1，先确定 min$S$ 值；再扫描 H，求解各变压器监控数据项集的 S；如果 H 内某变压器监控数据项集的 S 低于 min$S$，那么剔除该项集相应的矩阵列[14]；如果 H 内某行变压器监控数据项集的 S 是 0，那么剔除该变压器监控事务相应的矩阵行[15]。以合计值的方式，从大至小排序 H 内行与列向量，获取 L_1 以及相应的 H_1。

步骤 3　产生变压器监控数据的频繁 2-项集 L_2，以加权向量内积方式，处理 H_1 内列向量和其余列向量，得到不同组合的 S，剔除最小 S 的列，同时从大到小排序各列，获取 H_2，即变压器监控数据的频繁 2-项集 L_2 相应的布尔矩阵。按照定理 1~定理 3 处理 H_2，利于产生更大的变压器监控数据频繁项集。

步骤 4　产生变压器监控数据的频繁 k-项集 L_k，以加权向量内积方式，按照定理 1~定理 3 处理后的 H'_{k-1} 内各列和其余列，剔除 S 低于 min$S$ 的列，同时从大至小排列获取 H_k，即变压器监控数据的频繁 k-项集 L_k 相应的布尔矩阵。

步骤 5　按照定理 1 分析 L_k 是否是最大变压器监控数据频繁项集，如果不是，那么反复操作步骤 4，产生变压器监控数据的频繁 k+1-项集 L_{k+1}，反之，停止搜索。

步骤 6　最终的变压器监控数据频繁项集，就是全部变压器监控数据频繁 k-项集 L_k 的并，并结合最小置信度原则，生成关联规则，挖掘高频繁访问变压器监控数据 $\hat{X}$。

1.2　基于改进高斯分布的高频繁访问变压器监控数据差异消除

高斯过程模型属于高斯分布的改进，可提升监控数据的差异数据检测精度。为此，利用高斯过程模型，在高频访问变压器监控数据 $\hat{X}$ 内，检测差异数据，并消除检测到的差异数据，完成高频繁访问变压器监控数据差异消除。

令高斯核函数的表达公式如下：

$$K(\hat{x}_{i'},\hat{x}_{j'})=\mathrm{e}^{-\frac{\|\hat{x}_{i'}-\hat{x}_{j'}\|^2}{2\delta^2}} \tag{4}$$

式中：函数宽度是 δ；第 i'、j' 个高频繁访问变压器监控数据是 $\hat{x}_{i'}$、$\hat{x}_{j'}$。

高斯过程模型是最小化负对数边缘概率 P 有关超参数 β，获取最佳超参数。P 的计算公式如下：

$$P=\frac{1}{2}\left[\lg|K(\beta)|+\frac{y^{\mathrm{T}}y}{K(\beta)}+\eta\lg(2\pi)\right] \tag{5}$$

式中：高频繁访问变压器监控数据标签是 y；高频繁访问变压器监控数据样本数量是 η；转置符号是 T。

高斯过程模型在解决高频繁访问变压器监控数据的差异数据检测问题时，是通过对高频繁访问变压器监控数据样本 $\hat{X}$ 展开训练的过程，建立用于检测高频繁访问变压器监控数据的差异数据检测模型。

利用高斯过程模型，检测高频繁访问变压器监控数据的差异数据的具体步骤如图 2 所示。

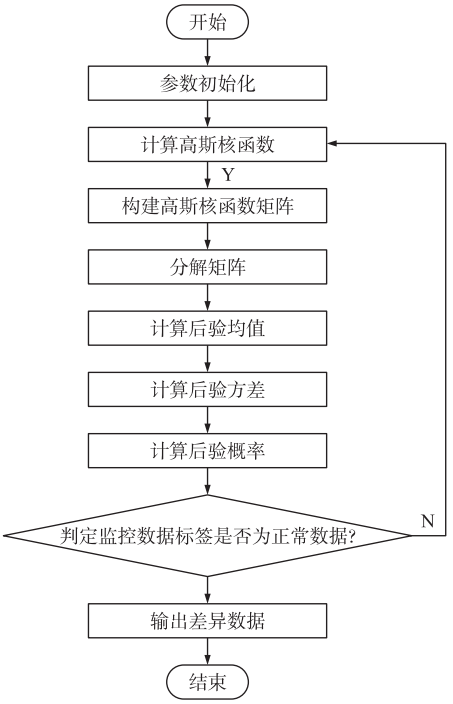

图 2　检测高频繁访问变压器监控数据的差异数据流程图

具体内容如下:

步骤 1　参数初始化。

步骤 2　利用式(4)计算全部高频繁访问变压器监控数据间的 $K(\hat{x}_{i'},\hat{x}_{j'})$,构建高斯核函数矩阵 K。

步骤 3　利用 Cholesky 矩阵分解法,分解 K,令 $P=\text{cholesky}(K+\delta^2 B)$,其中,$B$ 是单位矩阵。

步骤 4　计算 β,公式如下:

$$\beta=\frac{y}{K+\delta^2 B} \tag{6}$$

步骤 5　计算高频繁访问变压器监控数据的差异数据检测的后验均值 μ,公式如下:

$$\mu=\frac{K^{\mathrm{T}}y}{K+\delta^2 B} \tag{7}$$

步骤 6　计算高频繁访问变压器监控数据的差异数据检测的后验方差 θ^2,公式如下:

$$\theta^2=K(\hat{x}_{i'},\hat{x}_{j'})-\frac{K^{\mathrm{T}}K}{K+\delta^2 B} \tag{8}$$

步骤 7　计算高频繁访问变压器监控数据的差异数据检测的后验概率 q,公式如下:

$$q(y=1)=-\frac{1}{2}(y^{\mathrm{T}}\beta-\eta\lg 2\pi)-\sum_{i'=1}^{\eta}\lg P_{i'} \tag{9}$$

式中:第 i'个高频繁访问变压器监控数据的负对数边缘概率是 $P_{i'}$。

步骤 8　设置 μ、θ^2、q 的阈值 ε_μ、ε_{θ^2}、ε_q,对比分析高斯过程模型测试时获取的 μ,与训练时获取的 μ^*,若 $|\mu-\mu^*|<\varepsilon_\mu$,那么判定该高频繁访问变压器监控数据标签为 $y=1$,即正常数据;若 $|\mu-\mu^*|\geqslant\varepsilon_\mu$,那么判定该高频繁访问变压器监控数据标签为 $y=-1$,即差异数据;对比分析高斯过程模型测试时获取的 θ^2,与训练时获取的 θ^{2*},若 $|\theta^2-\theta^{2*}|<\varepsilon_{\theta^2}$,那么判定该高频繁访问变压器监控数据标签为 $y=1$,即正常数据;若 $|\theta^2-\theta^{2*}|\geqslant\varepsilon_{\theta^2}$,那么判定该高频繁访问变压器监控数据标签为 $y=-1$,即差异数据;对比分析高斯过程模型测试时获取的 q,与训练时获取的 q^*,若 $|q-q^*|<\varepsilon_q$,那么判定该高频繁访问变压器监控数据标签为 $y=1$,即正常数据;若 $|q-q^*|\geqslant\varepsilon_q$,那么判定该高频繁访问变压器监控数据标签为 $y=-1$,即差异数据。

步骤 9　整理符合 $|\mu-\mu^*|\geqslant\varepsilon_\mu$、$|\theta^2-\theta^{2*}|\geqslant\varepsilon_{\theta^2}$、$|q-q^*|\geqslant\varepsilon_q$ 条件的高频繁访问变压器监控数据,即输出高频繁访问变压器监控数据的差异数据,并消除标签为-1 的全部高频繁访问变压器监控数据的差异数据。

2　实验分析

2.1　实验参数

以某绝缘三相干式变压器为实验对象,其型号为 SGB10-1000/10H,该绝缘三相干式变压器的技术参数如表 1 所示。

表 1　绝缘三相干式变压器的技术参数

名称	数值
额定容量/kVA	1 000
高压/kV	6
分接/%	±5
低压/kV	0.4
空载损耗/W	1 985
负载损耗/W	13 345
空载电流/%	1
短路阻抗/%	6
器身重量/kg	2 400
长×宽×高(无防护外罩)/mm	1 590×870×1 460
长×宽×高(有防护外罩)	2 090×1 370×1 860
轨距/mm	820×820

2.2　实验结果

2.2.1　频繁项集挖掘数量结果

利用本文方法在历史变压器监测数据内,挖掘高频繁访问变压器监测数据,分析不同支持度,原因是适应不同数据特性和挖掘目标,较高的支持度能够减少频繁项集的数量,这不仅有助于减少噪声数据,聚焦在核心数据上,还能提高算法的效率和挖掘精度。为了确保分析结果的可靠性和稳定性,需要进行多次实验,使用不同的历史数据集来验证算法在不同条件下的性能,以全面评估该方法的优势和应用效果。本文方法改进前后,高频繁访问变压器监测数据挖掘过程中的频繁项集数量越少,高频繁访问变压器监测数据挖掘时重要数据的丢失量越少,高频繁访问变压器监测数据挖掘精度越高,分析结果如图 3 所示。

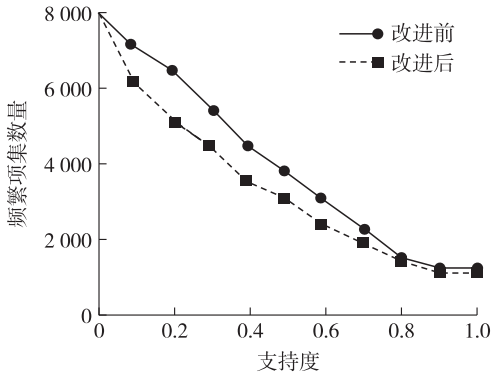

图 3　本文方法改进前后频繁项集挖掘数量对比结果

根据图 1 可知,随着支持度的增长,本文方法改进前后高频繁访问变压器监测数据挖掘的频繁项集数量均呈下降趋势,当支持度达到 0.9 时,本文方法改进前后的频繁项集数量均趋于稳定;在不同支持度时,本文方法改进后的频繁项集数量,均低于本文方法改进前;随着支持度的增长,本文方法改进前后的频繁项集数量差距越来越小,原因是支持度越高,符合条件的候选高频繁访问变压器监测数据项目集越少。实验证明:在不同支持度时,本文方法改进后的频繁项集数量较低,即高频繁访问变压器监测数据挖掘精度较高。

2.2.2 监测数据分布结果

在高频繁访问变压器监测数据内,差异数据的频繁性和分布影响频繁项集的显著性和相关性,使得挖掘结果的精度和实用性降低。在差异数据较多的情况下,本文方法仍然能够展现出较好的适用效果,能够有效地识别并消除差异数据。差异数据增多会增加算法的计算复杂度,通过合理的参数设置,本文共设置 8 个差异数据,能够快速地挖掘出高频繁访问的变压器监测数据中的关键模式。利用本文方法在历史变压器监测数据内,挖掘高频繁访问变压器监测数据,其分布情况如图 4 所示。

根据图 4 可知,本文方法可有效在历史变压器监测数据内,挖掘高频繁访问变压器监测数据,从图中可知,大部分高频繁访问变压器监测数据较为集中,仅有部分高频繁访问变压器监测数据分布在边缘位置。

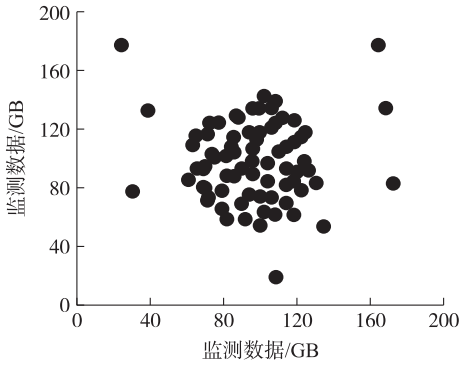

图 4 高频繁访问变压器监测数据分布情况

2.2.3 后验均值、后验方差、后验概率与交叉验证结果

在本文方法挖掘的高频繁访问变压器监测数据内,随机选择 20 个数据样本,利用本文方法计算高频繁访问变压器监控差异数据检测的后验均值,后验方差、后验概率与交叉验证,计算结果如图 5 所示。

根据图 5(a)可知,仅有样本 9 的后验均值低于阈值,说明编号为 9 的高频繁访问变压器监测数据

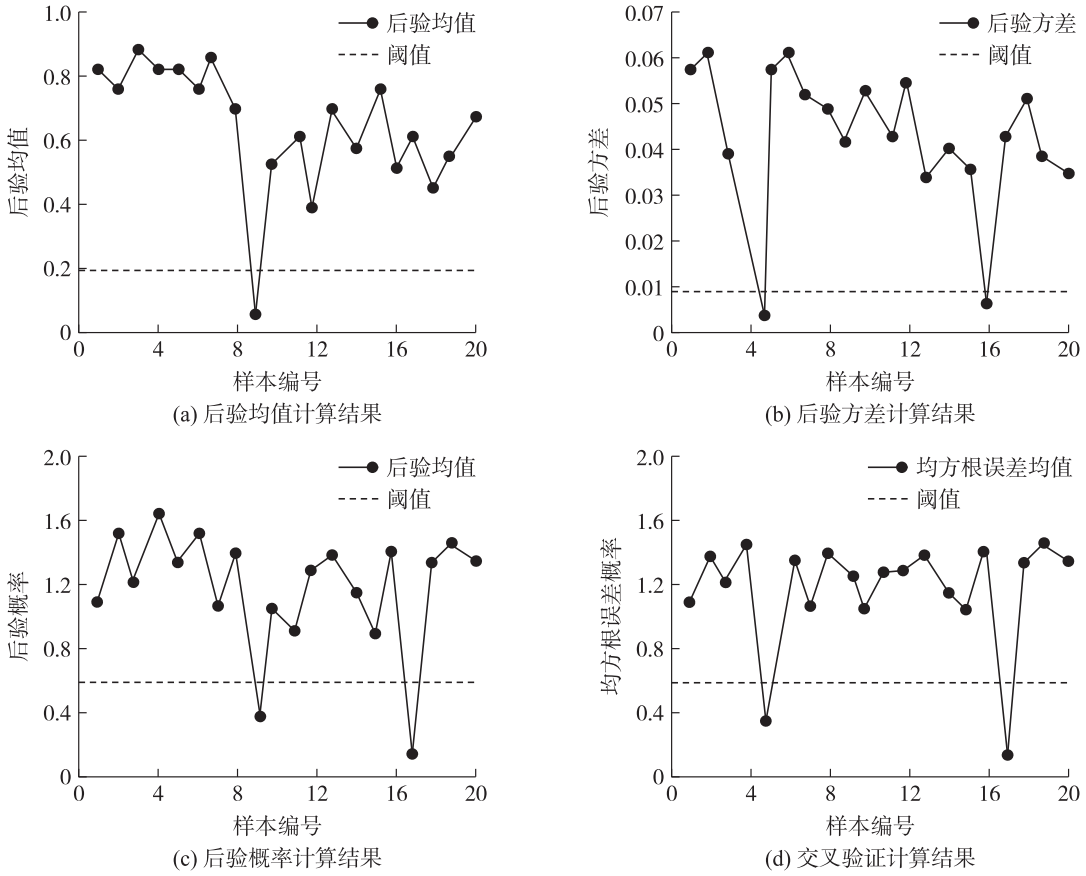

图 5 后验均值、后验方差、后验概率与交叉验证计算结果

为差异数据;根据图 5(b)可知,样本 5 与 16 的后验方差均低于阈值,说明这两个的高频繁访问变压器监测数据为差异数据;根据图 5(c)可知,样本 9 与 17 的后验概率均低于阈值,说明这两个的高频繁访问变压器监测数据为差异数据;根据图 5(d)可知,样本 5 与 17 的均方根误差概率低于阈值,说明这两个的高频繁访问变压器监测数据为差异数据。综合分析可知,样本 5、9、16 与 17 为差异数据。实验证明:本文方法可有效检测高频繁访问变压器监测数据的差异数据。

2.2.4　差异数据消除结果

利用本文方法消除高频繁访问变压器监测数据中的差异数据时,通过统计分析拟合高斯过程模型,并基于该模型确定阈值。根据阈值识别出超出正常范围的数据点作为差异数据,并在图 6 中用方框标出,同时采取基于改进高斯分布的高频繁访问变压器监控数据差异消除方法消除这些差异数据。差异数据消除结果如图 6 所示。

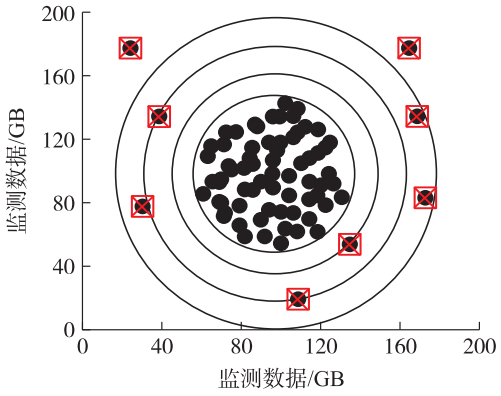

图 6　差异数据消除结果

根据图 6 可知,图中方框表示需要消除的高频繁访问变压器监测数据的差异数据点,图中圆形曲线为高频繁访问变压器监测数据的高斯分布曲线,由图 6 可知,该高频繁访问变压器监测数据内,共包含八个差异数据,与实际情况相符。实验证明:本文方法可精准消除高频繁访问变压器监测数据内的差异数据,可提升变压器监控数据质量。

3　结论

因为变压器监控数据采集类型不同,所以变压器监控数据内会出现差异数据,影响监测数据质量,无法为变压器监控提供精准的数据支持。为此,研究基于 Apriori 算法改进高斯分布的高频繁访问变压器监控数据差异消除方法,提升变压器监控数据差异消除精度,为变压器监控提供高质量的数据。

参考文献:

[1] 陈杰,张浩天,汤奕. 基于改进生成式对抗网络的电网异常数据辨识方法[J]. 电力建设,2021,42(5):9-15.
[2] 孙滢涛,张锋明,陈水标,等. 基于多域特征提取的电力数据异常检测方法[J]. 电力系统及其自动化学报,2022,34(6):105-113.
[3] 严明辉,潘舒宸,吴滇宁,等. 基于非参数核密度估计的电力市场用户电量异常数据辨识与修正方法[J]. 现代电力,2022,39(1):80-87.
[4] 户艳琴,李海明,刘念,等. 面向能源互联网终端用户的异常数据检测方法[J]. 电力建设,2021,42(10):19-27.
[5] 王文森,杨晓西,刘阳,等. 基于层次聚类分析的变压器油中溶解气体在线监测数据异常检测[J]. 高压电器,2023,59(1):142-147.
[6] 殷丽凤,李明状. 基于 Apriori 算法的关联规则分析应用[J]. 电子设计工程,2023,31(15):11-14,19.
[7] 许海林,林春耀,罗颖婷,等. 基于 Shapelet 识别的变压器在线 DGA 异常检测[J]. 高压电器,2021,57(7):175-181,188.
[8] 裘慧奇. 基于向量矩阵的 Apriori 改进算法研究[J]. 上海理工大学学报,2022,44(1):56-61,68.
[9] 刘岩,李文文,周丽霞,等. 基于高斯混合模型的光伏发电出力中高比例异常数据检测方法研究[J]. 电测与仪表,2021,58(9):14-21.
[10] 李文锋,陈俊贤. 基于动态关联规则的 TApriori 算法[J]. 南京邮电大学学报(自然科学版),2021,41(4):67-73.
[11] 樊清川,于飞,宣敏. 基于优化 Bi-LSTM 模型的电力变压器故障诊断[J]. 计算机仿真,2022,39(11):136-140.
[12] 赵天辉,张耀,王建学. 基于空间密度聚类和异常数据域的负荷异常值识别方法[J]. 电力系统自动化,2021,45(10):97-105.
[13] 刘明群,何鑫,覃日升,等. 基于改进 *K*-Means 聚类 k 值选择算法的配网电压数据异常检测[J]. 电力科学与技术学报,2022,37(6):91-99.
[14] 文旭,王浩,黄刚,等. 基于因子分析的母线负荷异常数据辨识方法[J]. 重庆大学学报,2021,44(8):91-102.
[15] 傅世元,高欣,张浩,等. 基于元学习动态选择集成的电力调度数据异常检测方法[J]. 电网技术,2022,46(8):3248-3261.

王　洁(1988—),女,高级工程师,主要研究方向:输变电设备状态检修、电力可靠性规划,chegupand2j@163.com。