基于智能电表运行故障数据的纵向分析模型

刘紫熠[1] 刘 卿[1] 王 崇[1] 王季孟[1] 王 玥[1] 刘金硕[2] 殷泽豪[2]
(国网天津市电力公司电力科学研究院 天津 300000)[1] (武汉大学国家网络安全学院 武汉 430076)[2]

摘 要 智能电能表作为当前电能计量和经济结算的主要工具,它的故障率直接关系到国计民生。文中设计了基于智能电表运行故障数据分析的纵向分析模型。该模型可以对已经安装的不同厂家、不同批次的智能电表的故障率随时间的变化进行分析。首先清洗不可用数据,然后对基本的数据条目进行线性回归分析,得到每个批次的智能电表的故障率和故障率的变化率,进而再对批次故障率的变化率进行聚类分析,评估各厂家的智能电表质量的稳定性,为智能电表使用单位提供参考。该模型对智能电表的批次质量进行了评估,对厂家的智能电表质量评判起到重要的借鉴作用。

关键词 智能电表,回归分析,聚类,多层模型
中图法分类号 TM934 **文献标识码** A

Vertical Analysis Based on Fault Data of Running Smart Meter

LIU Zi-yi[1] LIU Qing[1] WANG Chong[1] WANG Ji-meng[1] WANG Yue[1] LIU Jin-shuo[2] YIN Ze-hao[2]
(State Grid Tianjin Electric Power Company Electric Power Research Institute, Tianjin 300000, China)[1]
(School of Cyber Science and Engineering, Wuhan University, Wuhan 430076, China)[2]

Abstract As the main tool of electricity measurement and economic settlement, the failure rate of smart meter is directly related to the national economy and livelihood of the masses. This paper devised a vertical analysis model of fault data of running smart meter. The model can analyze the operation failure rate data of smart meters from different manufacturers and batches. The model firstly cleans the useless data, then carries out linear regression analysis on the basic data items, and gets the fault data and changing rate of the failure rate of each batch, which are utilized to do the cluster to evaluate the stability of the factory quality. The method and the result of the model can assess the quality of the batch of the smart meter, and can be beneficial to estimate the quality of factory.

Keywords Smart meter, Regression analysis, Cluster, Multilayer model

1 引言

智能电网的发展已成为世界各国电力工业应对未来挑战的共同选择,是21世纪电力系统的发展方向。智能电表作为智能电网建设的重要基础装备,加快其发展对电网实现信息化、自动化、互动化具有重要支撑作用[1]。与此同时,随着电力市场的逐步发展,电能表需要承担的功能越来越多,从而对电能表的故障率提出了更高的要求。目前国内各大电能表生产厂家生产的此类电能表在使用过程中,可能出现各种原因导致的故障,以致无法继续使用[2]。

传统的可靠性预计和可靠性实验评估都采用的是横向分析,并且是在电表装出前进行加速实验得到的。据统计,预测值与真实值的相对误差高达百分之几十到几千不等[3]。影响智能电表可靠性和可信性的因素有:功能、复杂性、设计、制造过程、失效判据、工作条件、安装维护等[3]。同时,由于技术革新以及不同厂家的产品的质量水平,智能电表的不同生产批次的故障率不同[4]。

顾伟等[5]在《基于故障统计模型的可修系统维修周期预测法》中,提出了一种可修系统故障统计模型和维修周期的顶测方法,它的基本原理是系统局部故障变化率的单样本参数估计,被应用于设备维修管理。

该方法在可维修系统的维修周期预测上有其实用性,但在智能电表这类相比维修更倾向于轮换的设备而言,并不能直接套用,因此本文在对智能电表相关的大量数据进行分析的基础上,提出了针对智能电表寿命周期预测的新算法模型。中华人民共和国国家标准 GB17215.911—200X/IEC/TR 62059—11:2002[6]阐述了电测量设备的可信性的基本概念。国际标准 13/1437/FDIS[7]阐述了电能计量装置的可信性的"温度和湿度"的加速可靠性实验。但是文献[8]的国际标准仅仅对温度和湿度信息影响因子进行了可靠性实验,没有考虑影响智能电表使用寿命的其他因素,而且这种实验是在智能电表安装使用前进行的。目前,由于智能电表的运行故障数据的采集与存储技术得到提高,可以有大量的运行故障数据用于分析。本文利用了天津电力公司电力研究院提供的智能电表"运行中"故障信息数据。设计一种多层次纵向分析模型,在考虑了智能电表故障率与时间的相关度的基础上,进一

本文受国家电网公司科技项目(用电信息采集系统运行维护及现场移动作业关键技术研究)资助。
刘紫熠(1987—),女,工程师,主要研究方向为电能计量;**刘 卿**(1980—),男,高级工程师,主要研究方向为电能计量;**刘金硕**(1974—),女,博士,副教授,硕士生导师,主要研究方向为数据挖掘、高性能计算,E-mail:896290784@qq.com(通信作者)。

步考虑了生产批次以及智能电表的生产厂家，更为全面地反映了不同厂家在智能电表质量上的表现。在评估单个购入批次的智能电表质量的基础上，进一步比较了不同厂家生产的智能电表的质量，为智能电表使用单位提供了一定的参考[9]。

本文第 2 节介绍了相关的理论背景；第 3 节详细阐述了算法模型的设计；第 4 节给出了实验的设计与结果分析；最后总结全文。

2　背景

纵向研究主要用来分析一段时间或某几个时间点总体的平均增长趋势和个体之间的差异。也就是说，对于纵向研究，目前主要集中在两方面：一个是描述总体的平均增长趋势，另一个是用来描述不同个体之间增长趋势的差异[10]。

聚类是一个无监督的分类，它没有任何先验知识可用，它是将数据划分成有意义或有用的组。聚类的形式描述如下：令 $U=\{p_1,p_2,\cdots,p_n\}$ 表示一个模式（实体）集合，p_i 表示第 i 个模式 $i=\{1,2,\cdots,n\}$，$C_t\subseteq U$，$t=\{1,2,\cdots,k\}$，$C_t=\{p_{t1},p_{t2},\cdots,p_{tw}\}$，$proximity(p_{ms},p_{ir})$。其中，第 1 个下标表示模式所属的类，第 2 个下标表示某类中的某一模式，函数 $proximity$ 用来刻画模式的相似性距离。若 C_t 为聚类的结果，则 C_t 需满足如下条件：

1）$U_{t=1}^{k}C_t=U$；

2）对于 $\forall C_m,C_r\subseteq U,C_m\neq C_r$，有 $C_m\cap C_r=\varnothing$（仅限于刚性聚类）；

$$\underset{\forall P_{mu}\in C_m,\forall P_{rv}\in C_r,\forall C_m,C_r\subseteq U\,\&\,C_m\neq C_r}{\mathrm{MIN}}(proximity(P_{mu},P_{rv}))>\underset{\forall P_{mx},P_{my}\in C_m,\forall C_m\subseteq U}{\mathrm{MAX}}(proximity(P_mx,P_{my})) \quad (1)$$

典型的聚类过程主要包括数据（或称为样本或模式）准备、特征选择和特征提取、接近度计算、聚类（或分组）、对聚类结果进行有效性评估等步骤[11]。

3　数据分析模型的建立

智能电表的运行故障数据最能反映真实的产品故障率水平，它们可以看作是一系列随时间发展变化的数据，将其按时间顺序排列起来，便能组成一组时间序列。本文借鉴了外推预测方法[8]，设计了多层次的基于时间序列的纵向分析法，对天津电力公司电力研究院提供的故障率数据进行了分析，模型综合运用了回归分析法和聚类法。

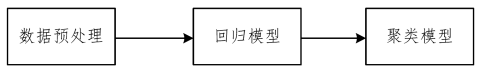

图 1　模型流程图

3.1　模型第一层线性——回归模型

读取数据表中的每个条目，使用线性回归模型对每一个批次的故障率随时间的变化做回归分析，从而得到批次的故障率关于时间的图线，图中，横轴为使用时长，纵轴为故障率，设第一个点为初始点，将初始点的纵坐标值作为截距 intercept，此后，计算以后读入的每个点与初始点之间的直线的斜率，并取其和前一个点与初始点之间直线的斜率的平均值 slope，直到读入所有点为止，将最终的斜率作为该批次的故障率的增长率，即：

$$intercept=errorRate_0 \quad (2)$$

$$slope=\frac{\dfrac{errorRate_{next}-errorRate_0}{workingDays_{next}-workingDays_0}+\dfrac{errorRate_{current}-errorRate_0}{workingDays_{current}-workingDays_0}}{2} \quad (3)$$

其中，$errorRate_0$ 为该批次最早发生故障的一个时间段内的故障率，将其作为图线的截距；$slope$ 则为图线的斜率；其中 $errorRate_{current}$ 表示非 $errorRate_0$ 的当前时间段内的故障率；$errorRate_{next}$ 表示下一时间段内的故障率；$workingDays_{current}$ 表示当前的工作时长；$workingDays_0$ 表示该批次最早发生故障时的设备工作时长；$workingDays_{next}$ 表示下一个时间段的设备工作时长。

3.2　模型第二层——聚类模型

使用聚类模型将批次号与其厂家关联，去掉每个厂家的批次集中的离群点，即与其他批次的表现差距较远的批次，通过综合分析批次的故障表现，来反映厂家的产品质量。以所有厂家每个批次的故障率为横轴，它们的故障率增长率为纵轴，对这些数据点进行聚类之后，可以直观地看出不同厂家之间、不同批次之间的质量差异，即横坐标越大表示故障率越高，纵坐标越大表示故障率增长率越大。

3.3　基于第一层第二层模型进行厂家分析

在计算出同一厂家的所有批次的中心点后，去除离群点，将最终的中心点作为该厂家的质量表征。中心点的定义为：与所有其他点之间的欧几里得度量之和最小的点。例如，对于点 (x_0,y_0) 以及其点集内的其他点 $\{(x_1,y_1),(x_2,y_2),\cdots,(x_n,y_n)\}$，该点与其他点的欧几里得距离之和为：

$$\sum_{i=1}^{n}\sqrt[2]{(x_0-x_i)^2+(y_0-y_i)^2} \quad (4)$$

离群点的定义为：与中心点的欧几里得距离大于平均欧几里得距离的 1.2 倍的点，该参数可以进一步调整以使分析结果更为准确。另外，每个厂家的中心点上有一个紧致度，由处理后满足条件的点集中点的数量除以原点集的点的数量确定，用于表征同一厂家的不同批次产品质量的稳定性。

3.4　模型评价

3.4.1　类内紧致度

本文给出了一种计算每个厂家各个批次智能电表故障率的紧致度 D_w 的方法。D_w 为在去掉离群点的点集中，距中心点单位长度的圆内的点的数目 n 与处理后的点集中点的总数 N 的比值，即：

$$D_w=\frac{n}{N} \quad (5)$$

紧致度反映了同一厂家各批次电表故障率随时间的增长率之间差别的大小，体现了同一厂家生产的不同批次智能电表产品质量的差异程度。

3.4.2　类间离散度

用不同厂家中心点之间的距离来评估不同厂家生产电表的质量的差别。定义 D_b 为两个厂家各批次电表中心点之间的距离，即：

$$D_b=\sqrt{(B_i-B_j)^2+(K_i-K_j)^2} \quad (6)$$

其中，$(B_i,K_i),(B_j,K_j)$ 分别为第 i 个和第 j 个厂家所有批次电表的中心点，计算公式为：

$$B_i=\sum_{s=1}^{M}b_s \quad (7)$$

$$K_i = \sum_{s=1}^{M} k_s \tag{8}$$

批次的总数 b 为一批次电表初始故障率的均值，k 为一批次故障率增长率的均值。

4 实验

4.1 清理数据

由于不是所有的数据都可用于分析，因此在处理前要清理一些不可用数据，如使用时长很短的批次、装出数量很少的批次、以及故障数量很少的批次。智能电表使用单位有硬性规定，对于每个购入的智能电表批次，其最大故障率不得超过2%，因此，对于 errorNum/installNum，大于 2% 的批次可以直接被归类为不合格批次。综合考虑，清理的原则为：批次故障总数量少于 7 的数据需要被清除；由固定的最大故障率可知，装出数量少于 350 的数据也不能被使用。分析每个批次的故障率的增长率是为了评估其所在厂商的产品质量，因此，若同一个厂的批次号的数量少于 5，由于样本容量过小，也不予采集。

3.2 处理数据

输入的数据格式如表 1 所列。表 1 中，从左至右的列分别代表厂家、安装日期、故障日期、购入批次号、装出数量、故障数量以及使用时长，算法将按行读入。

表 1 数据处理的输入格式

Factory (String)	installDate (Date)	errorDate (Date)	Batch (String)	installNum (Integer)	errorNum (Integer)	workingDays (Integer)
A 厂	2012-06-29	2012-07-02	DHPH006364	260	7	3
A 厂	2012-06-11	2012-06-28	DHPH006364	260	7	17

4.3 结果分析

算法最后产生的点集为厂家的中心点的集合，这是为了直观地比较厂家之间的产品质量的差异。在算法处理过程中，为了方便处理，本文将工作时长少于 30 天的智能电表按工作了 30 天处理，将工作时长少于 60 天的智能电表按 60 天处理，其他工作时长类似。

在模型中，可以综合紧致度 D_w 和类间离散度 D_b 来评价不同厂家生产的智能电表的质量。利用提出的模型对天津电力公司电力研究院提供的智能电表故障数据进行批次层次上和厂家层次上的分析。

4.3.1 批次层次上的结果分析

以北京富根厂家为例，对其进行批次层次上的分析，实验结果如表 2 所列。

月平均增长率都是相对最小的，批次智能电表的质量相对很好。批次 0000001793 的初始故障率为 0，其中，所有批次的故障率都是乘以 10000 来表示的，44 个月内的故障率为 34.95，使用时间内的故障率月平均增长率为 0.79，虽然其故障率相对较高，但是 44 个月内故障率的月平均增长率很低，因此批次 0000001793 智能电表的质量也很好。

表 2 批次层次上的分析结果

厂家	批次号	到货批次数量	装出数量	批次故障数量	初始故障率	工作时长/月	故障率*10000	故障率月平均增长率
北京富根	0000001653	1990	1990	7	0	53	35.18	0.66
北京富根	0000001683	2300	2300	8	0	44	34.78	0.78
北京富根	0000001793	2003	2003	7	0	44	34.95	0.79
北京富根	0000001797	2004	2004	9	0	49	44.91	0.91
北京富根	0000001808	2002	2002	9	0	46	44.96	0.98
北京富根	0000001842	1994	1994	12	0	57	60.18	1.06
北京富根	0000001881	1769	1769	12	0	60	67.83	1.13
北京富根	0000001858	5000	5000	13	0	19	26.00	1.36
北京富根	0000001912	2002	2002	14	0	49	69.93	1.44
北京富根	0000003227	5000	5000	14	0	19	28.00	1.45

4.3.2 厂家层次上的结果分析

厂家层次上的结果分析如表 3 所列。其中，中心点的横坐标 b 为故障率，纵坐标 k 为故障率的增长率；AvgDi 为类内平均距离，即一个厂家的所有批次与其中心点距离的平均值，代表了这个厂家平均质量的好坏；n 为这个厂家中距中心点 AvgDi 的圆内的批次的数目。

表 3 厂家层次上的分析结果

厂家	批次数目 N	AvgDi	中心点(b,k)	n	Dw/%
哈尔滨汇鑫	31	0.02	(0.21,0.006)	24	77.41
杭州百富	16	0.69	(2.48,0.19)	15	93.75
河南许继	8	1.03	(17.58,0.72)	7	87.50
北京富根	13	0.23	(4.75,0.19)	6	46.15
宁夏隆基	14	7.66	(3.0,0.67)	11	78.57

从表 2 中可以看出，北京富根各批次智能电表之间的故障率以及故障率增长率的差异。批次 0000001793 的初始故障率为 0，在使用的 44 个月内的故障率为 34.95，故障率的月平均增长率为 0.79，批次 0000001793 的初始故障率为 0，在使用的 44 个月内的故障率为 34.78，故障率的月平均增长率为 0.78，这两个批次差异得到了体现，但是在北京富根这一厂家的所有批次中，各批次的故障率增长率差距较大，表 3 中北京富根的类内紧致度为 46.15%，紧致度越小，厂家生产的批次电表质量的差异越大，稳定性也越差，相反，紧致度越大，该厂家生产的电表的质量差异越小，稳定性也越好。

并且，通过比较厂家之间的中心点，哈尔滨汇鑫的中心点为(0.21,0.006)，表示该厂家的平均故障率为 0.21，平均故障率增长率为 0.006，杭州百富的中心点为(2.48,0.19)，表示该厂家的平均故障率为 2.18，平均故障率增长率为 0.19，通过对比可以看出，中心点越靠近原点(0,0)的厂家，该厂家的平均质量越好。通过以上分析初步可以得出，哈尔滨汇鑫和杭州百富两个厂家生产的智能电表质量较好并且性能较为稳定。

另外，类间离散度 D_b 表示两个厂家之间中心点的距离，

(下转第 456 页)