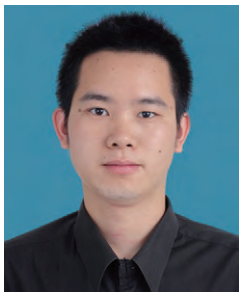


基于Logistic模型的主变压器缺陷概率预测实证研究

主变压器缺陷概率预测对于优化设备运维检修策略具有重要的参考意义。以某省级电网公司35 kV及以上电压等级油浸式主变压器的历史缺陷数据及相关的设备台账数据为基础,利用Logistic模型,实现了基于主变压器个体的缺陷概率预测模型的构建,并采用Kolmogorov-Smirnov检验法和Gini系数方法对预测结果进行评估,证明了该模型的有效性。

吴广财 严宇平 周睿 桂媛/广东电网有限责任公司



吴广财/工程师

设备缺陷概率反映了某台特定的设备在指定条件下发生缺陷的可能性。主变压器是电网中能量转换、传输的核心,其运行状态直接影响整个电力系统的可靠性,主变压器缺陷概率作为提高设备运维效率,提高设备可用性的重要参考数据,是主变压器设备管理的重点关注领域之一。

许多学者已对设备缺陷预测进行了较为深入的研究。文献[1-4]利用数据挖掘中的关联规则、分类、回归和决策树等进行设备缺陷的趋势预测,文献[5]利用支持向量机构建了设备缺陷的预测模型;文献[6-7]基于时间序列法开展了电力系统设备缺陷预测模型的构建及应用。但这些模型都是基于大量个体的宏观信息对特定类型的设备开展的缺陷预测,其模型可应用于特定类型设备的缺陷预测,但不完全适用于设备个体,对于设备个体的巡视工作重点制定缺乏指导意义。为解决该问题,本文尝试提出一种针对主变压器个体开展缺陷概率预测的方法,基于Logistic模型构建了主变

压器缺陷概率预测模型,实现对单台主变压器在未来一段时期内出现缺陷概率的预测,从而提高主变压器缺陷巡视和预防工作的科学性和准确率,促进主变压器运维效率优化。

Logistic模型介绍

Logistic模型是由比利时数学家在Malthus方程的基础上提出的。它是一种概率型非线性回归模型,常用于预测离散因变量与一组解释变量之间的关系。在某些回归问题中,响应变量是分类的,正态线性模型不适用,而Logistic模型的自变量可以是连续变量,也可以是分类变量或哑变量。作为一种广义线性模型,通过logit变换,该模型可转化为线性形式,从而利用线性回归模型对参数进行估计^[8]。相对于其他概率模型,Logistic模型计算更为直接简单,同时,利用最大似然法建立的判别函数所求的估计量比其他模型使用最小二乘法所得的估计量准确性更高。由于其对变量类型、数据的正态性等不做要求,同时具有系数可解释的优点,Logistic模型在医学^[9]、社会学^[10]和经济学^[11]等领域得到了广泛应用。

基于Logistic模型的主变压器缺陷概率预测模型构建

以某省级电网公司35 kV及以上油浸式主变压器缺陷数据为例,利用Logistic模型实现主变压器个体缺陷概率预测模型的构建。

关键词/Keywords

主变压器·
缺陷概率·
缺陷概率预测·
Logistic模型·

1. 模型输入变量选择

依据目前的设备状态评价导则，对主变压器的功能特点、技术特点和运行环境进行分析，并梳理影响主变压器缺陷概率的状态量，结合电网行业专家的经验，初步汇总得到16个模型的输入变量^[12-13]，如表1所示。

表1 主变压器缺陷概率预测模型自变量

模型变量	变量说明
所在站污区等级	所在变电站环境的污染程度。环境污染可能对设备的绝缘性能等造成影响，从而导致缺陷发生
设备厂商	前期数据分析表明，不同厂家主变压器产品缺陷率存在显著差异，是影响主变压器缺陷概率的显著因素
设备电压等级	不同电压等级下主变压器的生产工艺、绝缘性能等均不同，缺陷发生的条件也不相同，对概率预测结果有影响
设备投运时间（月）	不同投运时间设备老化程度存在差异，对设备缺陷发生概率有影响
设备近6个月的缺陷次数	基于历史数据进行预测，一台设备过往历史上的缺陷发生情况是其未来是否发生缺陷的参照
设备近12个月的缺陷次数	同上
设备近24个月的缺陷次数	同上
设备近48个月的缺陷次数	同上
绕组耦合方式	设备工艺的差异导致设备缺陷发生的模式有差异，对缺陷发生概率有影响
绕组外绝缘介质	同上
冷却装置型式	同上
油循环方式	同上
绕组数	同上
调压方式	同上
线圈导线材质	同上
铁心材质	同上

2. 交互式分组

在实际的计算过程中，由于各变量取值较多，为提高预测准确度，可采用交互式分组的方式对输入变量取值进行分组。具体的操作过程如下：应用统计学中的交叉表分析，根据证据权重法（Weight of Evidence，WOE）和信息值法（Information Value，IV）确定模型的特征变量和特征项。WOE值用以衡量自变量取特定的值时对目标变量的影响程度，IV值表征自变量所包含的信

息量大小，用以衡量变量预测能力^[14]。具体步骤如下。

对变量的每个值单独计算WOE，即

$$WOE = \ln \frac{Pct_1}{Pct_0} = \ln(Pct_1) - \ln(Pct_0) \tag{1}$$

式中， Pct_1 为指定条件下发生缺陷的主变压器数量占训练数据中发生缺陷的所有主变压器数量的比例； Pct_0 为指定条件下未发生缺陷的主变压器数量占训练数据中未发生缺陷的所有主变压器数量的比例。

将WOE值相近的项进行合并；根据合并结果以及所述WOE值计算各模型输入变量的IV值，即

$$Pct = Pct_1 - Pct_0 \tag{2}$$

$$IV = \sum (WOE \cdot Pct) \tag{3}$$

对于每个变量，WOE值越大，缺陷概率越高。IV值越大说明变量对缺陷概率的影响越大。选取该省级电网公司2008~2011年主变压器缺陷数据及相关台账数据作为训练数据，计算各变量的WOE值和IV值，结果如表2所示。

3. 预测模型的构建

最常用的Logistic模型是二值型Logistic模型，目标变量只有两个值，即“是”和“否”。在本模型中，把缺陷发生的情况定义为 $Y=1$ ，缺陷未发生的情况定义为 $Y=0$ ，则缺陷概率的值即为取值为0、1的因变量的概率。

假设在自变量 $X_1, X_2, \dots, X_n$ 的作用下，缺陷发生的概率为 p ，则该缺陷不发生的概率为 $1-p$ ， $p/(1-p)$ 为发生概率与不发生概率之比，对其取自然对数，记为 θ_p ，即

$$\theta_p = \text{logit}(p) = \ln \left(\frac{p}{1-p} \right) \tag{4}$$

式中， $\text{logit}(p)$ 是因变量 $Y=1$ 的差异比的自然对数。

很明显， θ_p 以 $\text{logit}(0.5)=0$ 为中心对称，且在 $p=0$ 和 $p=1$ 附近变化幅度较大，当 p 从0变化到1时， θ_p 从 $-\infty$ 变化到 $+\infty$ 。通过自然对数的取值能够克服变量 X_n 非线性和不敏感的问题，得到

$$\theta_p = \ln \left(\frac{p}{1-p} \right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_n X_n \tag{5}$$

表2 各变量WOE值与IV值计算结果（部分）

变量	含义	分组	分组	事件计数	非事件计数	计算的 WOE	信息值
WOE_DIRTINESS_AREA_LEVEL	所在变电站污染等级	A	1	6	116	-0.937 8	0.055
		B	2	50	464	-0.203 8	
		C	3	129	891	0.091 54	
		D, E, 未知或缺失	4	61	391	0.166 24	
WOE_IMP_defect_time24	近2年出现的缺陷次数	IMP_defect_time24<1, 缺失	1	42	1 242	-1.362 7	1.322
		1≤IMP_defect_time24<2	2	66	369	0.302 93	
		2≤IMP_defect_time24<3	3	64	144	1.213 14	
		3≤IMP_defect_time24	4	74	107	1.655 31	
WOE_IMP_defect_time6	近6个月出现的缺陷次数	IMP_defect_time6<1, 缺失	1	139	1 643	-0.445 7	0.556
		1≤IMP_defect_time6	2	107	219	1.307 83	
WOE_MANUFACTURER	设备厂商	设备厂商A ₁ 、设备厂商A ₂ 、设备厂商A ₃ …设备厂商A _n	1	53	113	1.266 98	0.513
		设备厂商B ₁ 、设备厂商B ₂ 、设备厂商B ₃ …设备厂商B _n	2	67	324	0.448 02	
		设备厂商C ₁ 、设备厂商C ₂ 、设备厂商C ₃ …设备厂商C _n	3	85	674	-0.046 5	
		设备厂商D ₁ 、设备厂商D ₂ 、设备厂商D ₃ …设备厂商D _n	4	38	579	-0.699 6	
		设备厂商E ₁ 、设备厂商E ₂ 、设备厂商E ₃ …设备厂商E _n	5	3	172	-2.024 8	
WOE_VOLTAGE_LEVEL	电压等级	220 000, 未知或缺失	1	80	229	0.972 38	0.312
		110 000	2	143	1 270	-0.159 9	
		500 000	3	11	101	-0.193 2	
		35 000	4	12	262	-1.059 4	
WOE_DEV_AGE	设备投运时间/月	AGE<30	1	42	432	-0.306 7	0.194
		30≤AGE<33	2	4	129	-1.449 4	
		33≤AGE<48	3	14	208	-0.674 4	
		48≤AGE<143	4	128	832	0.152 27	
		143≤AGE, 未知或缺失	5	58	261	0.52	

则缺陷发生概率 p 的计算公式为

$$p = \frac{1}{e^{-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_n X_n)}} \tag{6}$$

令 $\alpha = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_n X_n$, 则 $p = \frac{1}{e^{-\alpha}}$ 。

其中 α 与主变压器缺陷概率 p 的单调性一致。 α 与WOE的关系为

$$\alpha = \beta_0 + \beta_1 WOE_1 + \beta_2 WOE_2 + \dots + \beta_n WOE_n \tag{7}$$

式中, WOE_1 、 WOE_2 、 WOE_3 … WOE_n 为各变量的WOE值。

以该省级电网公司2008~2011年的数据作为训练数据, 2012年的数据作为验证数据, 2013年

的数据作为测试数据。利用SAS对数据进行训练, 首先将16个变量全部纳入模型中, 用Wald统计量对模型偏回归系数的显著水平进行检验, 将对因变量解释能力较差的设备近12个月的缺陷次数等10个因素剔除, 最后得到参数估计结果如表3所示。

表3中, Intercept为常数参量, WOE_* 为对应的参量; DF 为自由度; Estimate为最大似然法估计结果; Standare Error为标准误; Wald Chi-Square为卡方检验的值; $Pr>ChiSq$ 为 P 值, 判定假设检验结果的参数。从训练结果可以看出, 设备运行环境、电压等级、设备近2年发生缺陷的次数及近6个月发生缺陷的次数、设备厂商以及设备投运时

表3 参数估计结果

最大似然法估计结果					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr>ChiSq
Intercept	1	2.010 8	0.081 7	605.24	<0.000 1
WOE_DIRTINESS_ AREA_LEVEL	1	0.301 6	0.357 2	0.71	0.398 5
WOE_IMP_defect_ time24	1	0.767 4	-0.079 0	04.49	<0.000 1
WOE_IMP_defect_ time6	1	0.256 3	0.102 6	6.24	0.012 5
WOE_ MANUFACTURER	1	0.441 2	0.121 6	13.17	0.000 3
WOE_VOLTAGE_ LEVEL	1	0.427 1	0.146 2	8.53	0.003 5
WOE_DEV_AGE	1	0.371 9	0.180 0	4.27	0.038 8

间是主变压器缺陷概率预测模型中的重要变量，则

$$\alpha = 2.010\,8 + 0.301\,6\text{WOE_DIRTINESS_AREA_LEVEL} + 0.767\,4\text{WOE_IMP_defect_time24} + 0.256\,3\text{WOE_IMP_defect_time6} + 0.441\,2\text{WOE_MANUFACTURE} + 0.427\,1\text{WOE_VOLTAGE_LEVEL} + 0.371\,9\text{WOE_DEV_AGE} \tag{8}$$

式中，WOE_*为对应的WOE值。

4. 模型的验证

采用Kolmogorov-Smirnov检验法对该模型的预测结果进行评估。Kolmogorov-Smirnov检验法又称单样本K-S检验，是一种拟合优度性检验^[15]。以控制第一类错误概率（误判率）为例，其具体的实施方式如下。

基于将第一类错误概率控制在0.05以下，将预测概率定为0.36，如果预测概率为0.355以上，则认定为该设备发生缺陷。K-S检验法评估结果如表4所示。

表4 K-S检验法评估结果

		实际良好（H0为真）		实际发生缺陷（H0不真）	
		数量	百分比	数量	百分比
预测良好 （接受H0）	数量	2 949	95.0%	287	69.8%
	百分比	91.1%	83.9%	8.9%	8.2%
预测发生缺陷 （拒绝H0）	数量	155	5.0%	124	30.2%
	百分比	55.6%	4.4%	44.4%	3.5%

5.0%为第一类错误概率（误判率）：即设备本来没有缺陷，但是认定设备会发生缺陷的比例；69.8%为第二类错误概率（漏判率）：即设备本来发生了缺陷的，但是认定设备良好的比例；44.4%为准确率，即预测设备发生缺陷，它实际发生缺陷的比例。

对模型的评价准则采用Gini系数^[16]（Gini Index）方法，Gini系数计算方法为

$$\text{Gini}(S) = 1 - \sum_{i=1}^n p_i^2 \tag{9}$$

式中， p_i 为各类变量在S中出现的条件概率； n 为S中所含的类的个数。

利用60%训练集、20%验证集和20%测试集对模型进行评估，得到的评估结果如表5所示。

表5 模型评估结果

模型说明	回归变量筛选模型
训练：ROC 索引	0.832
训练：Gini 系数	0.664
训练：Kolmogorov-Smirnov 统计量	0.516
验证：ROC 索引	0.796
验证：Gini 系数	0.593
验证：Kolmogorov-Smirnov 统计量	0.534
测试：ROC 索引	0.782
测试：Gini 系数	0.564
测试：Kolmogorov-Smirnov 统计量	0.417

分析评估结果，可以得到以下结论：各项指标（ROC、Gini和K-S值）在训练集、验证集和测试集上未发生明显波动，说明主变压器个体缺陷概率预测模型是稳定的；K-S值大于0.3、Gini大于0.4说明预测模型具有显著的区别能力。

5. 模型的应用

利用设备缺陷概率预测模型可辅助设备管理人员优化设备巡视安排、预测技改大修项目规模和预测备品备件采购量等，加强工作的前瞻性，提高相关计划工作的量化水平。比如：可结合模型的误判率和漏判率，将该模型用于确定设备巡视重点。如按照设备缺陷概率进行设备巡视重点排序，当巡视设备数量为50%时，漏判率小于8%，这表明以此为依据安排设备巡视，在节省一半巡视人力物力的情况下，漏检缺陷设备数量小

于8%，结果较为理想，可作为设备巡视重点参考依据。

结束语

以某省级电网公司主变压器缺陷数据及相关的设备台账数据为依据，基于Logistic模型，构建了主变压器缺陷概率预测模型。该模型在训练集、验证集和测试集上，其ROC、Gini和K-S值均未发生明显变动，证明了该预测模型的有效性。

设备缺陷概率预测模型能够辅助生产运行人员把握设备未来的缺陷状况，从而根据设备缺陷概率差异情况有针对性地进行重点巡视和维护，提高工作效率，更有效地发现设备缺陷；另外，基于设备缺陷概率信息，设备管理部门可以优化设备巡视安排、预测技改大修项目规模以及预测备品备件采购量等，辅助管理决策。

参考文献

- [1] 于之虹, 郭志忠. 数据挖掘与电力系统 [J]. 电网技术, 2001, 25 (8): 58-62.
- [2] 张增敏. 数据挖掘技术在变电站设备及缺陷管理系统中的应用 [D]. 西安: 西安理工大学, 2006.
- [3] 张增敏, 谢嘉, 李长河, 等. 数据挖掘技术在变电站设备及缺陷管理系统中的应用 [J]. 山东农业大学学报: 自然科学版, 2006, 37 (4): 642-646.
- [4] 蒋跃强, 傅晨钊, 高凯. LCM视角下的变电站及关键设备缺陷分析 [J]. 华东电力, 2009, 37 (8): 1369-1373.
- [5] 陈义刚, 徐厚东. 基于支持向量机的变电设备缺陷发生率的预测及应用 [J]. 四川电力技术, 2013, 36 (6): 75-77.
- [6] 陈拥军. 电力系统设备缺陷预测 [D]. 杭州: 浙江大学, 2003.
- [7] 陈拥军, 王康元, 黄民翔. 基于决策支持系统平台的设备缺陷预测 [J]. 浙江电力, 2003 (6): 1-4.
- [8] 薛毅, 陈立萍. 统计建模与R软件 [M]. 北京: 清华大学出版社, 2007.
- [9] Kiwuwa Muyingo S, Oja H, Todd J, et al. Dynamic logistic regression model and population attributable fraction to investigate the association between adherence, missed visits and mortality: a study of HIV-infected adults surviving the first year of ART [J]. BMC Infectious Diseases, 2013, 13 (1): 1-14.
- [10] Denham, Bryan E. Measurement of risk perceptions in social research: a comparative analysis of ordinary least squares, ordinal and multinomial logistic regression models [J]. Journal of Risk Research, 2010, 13 (5): 571-589.
- [11] Upadhyay Arun, Bandyopadhyay Gautam, Dutta Avijan. Forecasting stock performance in indian market using multinomial logistic regression [J]. Journal of Business Studies Quarterly, 2012, 3 (3): 16-39.
- [12] 赵文清, 朱永利. 电力变压器状态评估综述 [J]. 变压器, 2007, 44 (11): 9-12, 74.
- [13] 宋新明, 郑益民. 大型变压器整体绝缘下降缺陷原因分析与处理 [J]. 变压器, 1999, 36 (8): 43-46.
- [14] Hope B, Clarkson J. A strategy for using weight-of-evidence methods in ecological risk assessments [J]. Human and Ecological Risk Assessment, 2014, 20 (2): 290-315.
- [15] Wikipedia contributors. Kolmogorov-smirnov test [G/OL]. Wikipedia, The Free Encyclopedia [2014-07-01]. http://en.wikipedia.org/w/index.php?title=Kolmogorov%E2%80%93Smirnov_test&oldid=615035189.
- [16] 龚志民, 刘波. 基尼系数计算的误差估计及其应用 (英文) [J]. 湘潭大学自然科学学报, 2013, 35 (4): 1-8.

(收稿日期: 2014-09-10) EA