基于大数据平台的
XX 应用场景
分析模型设计说明书

二〇一五年三月

一、 分析模型需求

（一）分析挖掘目标

由于近年来经济的持续稳定发展，城区用电负荷出现了相应的较快增长，每年均有配变发生重过载的现象，供电单位满足用户用电需求的难度也在不断提高。同时由于配网资金有限，不能完全满足改造需求。因此，科学地开展配变重过载的中期预警工作，可为来年配网的升级、改造规划提供参考，提高配网对迎峰度夏期间峰荷的应急能力。还能够为独立判断、分析各地市重过载程度、合理分配资源提供客观依据。

（二）分析挖掘成功标准

短期预警模型预警结果的好坏关系到基于其结果做出决策的损失大小，好的、稳定的、与实际拟合度高的模型能为决策层提供可靠的参考依据，因此对模型预警能力的评价至关重要，这就要求有一系列科学合理的评价指标从不同的需求考虑对模型预警能力做出量化的衡量，由于不同算法建立的模型输出结果不同，因此针对不同模型也有不同的评价标准，下面介绍几种评价标准并给出它们适用的算法。

1. 查全率、查准率

表 4.1.5 混淆矩阵

实际 / 预测	不发生	发生	查准率
不发生	202	3	
发生	2	5	71.43%
查全率		62.50%	97.64%

混淆矩阵（confusion matrix）是可视化工具，每一列代表了实际测得信息，每一行代表了预测的分类信息，查全率和查准率正是基于混淆矩阵上计算得出的。

查全率是在实际发生的范畴中讨论预测的准确性，它是衡量预测模型正确预测实际发生的能力，其计算公式为：（预测发生且实际发生的台变数/实际发生的总台变数）*100%。如表 4.1.5，查全率=5/(3+5)*100%=62.50%。

查准率是在预测发生的范畴中讨论预测的准确性，它是对预测模型预测发生中预测正确的肯定，其计算公式为：（预测发生且实际发生的台变数/预测发生的总台变数）*100%。如表 4.1.5，查准率=5/(2+5)*100%=71.43%。

由于查全率和查准率是根据混淆矩阵中分类信息计算出的指标，在三种模型中只有 Logistic 模型能输出混淆矩阵中的分类信息，故查全率和查准率只适合 Logistic 模型的评价。

2. 准确率

准确率又称"精度"、"正确率"，用来衡量预测值与实际值的拟合程度，其计算公式为：（|预测值-实际值|/实际值）*100%。

由于准确率的计算公式适用的是连续变量而不是分类变量，而只有多元回归模型和时间序列模型的输出结果是连续变量，故该评价指标适用多元回归模型和时间序列模型。

（三）模型构建策略

本次配变重过载短期预警模型的建立主要分为以下五个步骤：

第一、数据的准备。根据模型构建的整体框架，收集建模所需的相关变量的原始数据，再导入数据，通过分析平台对原始数据进行初步的观察，制定数据清洗规则，根据准则编写相应的程序，然后再对原始数据进行初步的清洗，以确保进入模型的数据质量。

第二、数据的观察与探索。再次对数据进行观察，设计并计算特征变量，分析单个自变量与因变量之间的显著性关系，对变量进行初步筛选，同时分析各个自变量间的相关关系，为排除自变量间的相关关系提供参考依据，为下一步模型的建立做好准备。

第三、模型的建立。通过对不同模型进行优劣分析，选择合适的模型算法，初步建立预警模型。

第四、模型的验证及优化。模型的建立需要不断的优化，通过比较取舍，确定最优的预警模型。首先，通过所建立的模型的相关验证指标，如拟合优度、系数显著性检验等，对变量进行优化筛选，建立可行的预警模型；然后，通过改变参数计算新的特征量数值，或根据观察模型误判结果，设计新的特征量，建立新的预警模型；最后，通过对比不同可行模型拟合优度，预警的准确性，综合各方面的因素，确定最优的预警模型。

第五，结果的输出及应用。将确定的最优模型对配变短期的重过载现象进行预警，得出配变重过载的概率，对配变的升级改造等实际应用提供科学的参考依据。

二、 影响因素梳理

1. 数据需求

根据建模需要，需要收集四个维度的数据信息，包括负荷信息、用户信息、天气信息和设备信息。具体数据需求清单，见表 4.1.1。

表 4.1.1 数据需求清单表目录（详见附录 1：数据需求表）

输入维度	数据需求表清单	数据来源	系统
负荷信息	负荷数据需求表	计量中心	用电采集系统
用户信息	用户数据需求表	营销部	SG186 营销系统
天气信息	天气数据需求表	防灾减灾部	防灾减灾天气系统
设备信息	设备数据需求表	运检部	GPMIS

2. 影响因素选择

考虑所有相关的影响因子进行分析：

附表 1 负荷数据需求表

数据项	备注	数据示例
采集记录编号	采集记录的编号	125
采集点编号	采集点的编号	0001636045
台区编号	对应台区的编号	30617030
台区名称	对应台区的名称	闽侯县白沙镇梧桐下村一片公用变
数据日期	负荷曲线的日期	2011-1-1

额定容量	配变容量	80
配变用途	配变用途	城区/农网
采集时间	数据采集的日期	2011/1/2
供电单位	9 个地市信息	福州
二级供电单位	二级供电单位	鼓楼区
P1~P96	负荷曲线信息	15.5...

取数说明:

1、统计范围：福福建省 9 地市及其下属区县二级供电单位（城网）

2、统计时间：2011-1-1 至 2013-12-31

附表 2 用户信息

数据项	数据示例 1
用户编号	0019939869
用户名称	XXX
用电地址	XX
用户分类	XX
用电类别	城市居民生活用电
用户状态	正常
供电单位	泉州
二级供电单位	南安
三级供电单位	XX
行业分类	XX
供电电压	220V
城农网标识	农网
合同容量	30
台区编号	XX
台区名称	XX
配变编号	XX
配变名称	XX

附表 3 抄表记录

用户编号	0019939869
月份	2011 年 1 月
用电量	100

取数说明:

1、统计范围：福建省 9 地市及其下属区县二级供电单位（城网）

2、统计时间：用户信息：最新； 每月抄表记录：2011 年 1 月~2013 年 12 月

附表 4 温度信息

数据项	备注	数据示例
地区/区域名称	福州/闽侯/宁德/莆田/漳州	闽侯
日期	气温采集日期	2011/11/1
最高温度	当日最高温度	23
时点温度 P1	当日时点温度	12
……	当日时点温度	……
时点温度 P24	当日时点温度	23

附表 5 体感信息

数据项	备注	数据示例
地区/区域名称	福州/闽侯/宁德/莆田/漳州	闽侯
日期	气温采集日期	2011/11/1
体感温度/舒适度指数	当日体感指数	5

附表 6 天气信息

数据项	备注	数据示例
地区/区域名称	福州/闽侯/宁德/莆田/漳州	闽侯
日期	气温采集日期	2011/11/1
天气	当日天气状态	雷雨/晴/多云

取数说明:

1、统计范围：福建省 9 地市及其下属区县（城网）

2、统计时间：2011-1-1 至 2013-12-30

3、统计内容：气温信息

附表 7 台区及设备信息

数据项	备注	数据示例
台区编号	对应负荷信息	13248114
台区名称	对应负荷信息	红星三村箱变 1#变
额定容量	对应负荷信息	630
投运日期		20120524
配变编号		
配变使用性质		公变
配变用途		城区配变
配变类型		油式
安装地址		

取数说明：

1、统计范围：福建省 9 地市及其下属区县二级供电单位（城网）

2、统计时间：最新

通过业务分析、影响因子相关相关性分析、数据完整性分析等方面进行综合性考虑设计并编写了 10 个特征变量，多角度地描述配变特征，如表 4.1.2。

表 4.1.2 影响因素汇总表

序号	特征量	自变量名称	自变量	自变量的意义
1	负荷特征量	加权日均负荷	LD_AVG	代表变压器整体的负荷水平
2		是否发生重载统计	LD_HLT	统计是否发生重载
3		是否发生过载统计	LD_OLT	统计是否发生过载
4		负荷突变统计	LD_SdChg	代表了每日平均负荷的突变情况
5		谷峰比	LD_VPR	每周负荷曲线的平均谷峰比
6		跨周负荷变动率	LD_ROL	上一周的平均负荷到下一周的的平均负荷变动程度和方向
7	用户特征量	居民用电占比	UR_Y2（UR_Y2SP）	描述居民用户用电量占总用电量的百分比例
8		居民用电增长指数	UR_GR (UR_GRSP)	描述在特定变压器下的居民用电占比变化
9	天气特征量	天气负荷回归系数	WX_Y12_Coef	描述不同气温下负荷的变化情况
10		天气负荷敏感度	WX_Y12_Cor	描述最高负荷和最高气温的相关关系

三、 分析方法选择

短期预警模型的建立对电力系统近期输变电建设、运行和计划都非常重要。短期负荷除具有明显的周期性外，还受到各种环境因素的影响，如天气因素、季节变换、电力市场、重大事件等，使得负荷的时间序列变化呈现出非平稳的随机过程。由于短期预警的不准确性和条件性，因此要对负荷在各种情况下可能的发展状况进行预警有一定难度。短期预警的建模算法有多种，它们也都有各自的特点和适用条件。在分析短期负荷特点和影响因素的基础上，综合考虑短期预警算法的特点，分析归纳出三种比较适合可行的算法，其分别是 Logistic 回归模型、多元回归模型和时间序列模型，下面分别对这三种方法的理论进行介绍。

（一） 算法方法介绍

1. Logistic

传统线性回归模型在实际定量分析中，受到多种条件的限制。例如，当因变量是一个分类变量而不是一个连续变量，线性回归模型不再适用。在分析分类变量时，通常采用的统计方法是对数线性模型（Log-linear-model）。在本期研究的课题是对配变是否发生重过载现象进行短期预警。配变是否发生过重过载，可分为“发生”和“不发生”两类，也被称为二分变量。当对数模型中的一个二分变量被当作因变量时，对数线性模型即成为 Logistic 回归模型。Logistic 回归分析为解决这种因变量为定性变量的问题提供了有效的分析工具。

设Y为一个随机变量，且服从两点分布。当事件发生时，Y的取值为 1，当事件没有发生时，Y的取值为 0。

假设影响实验结果的自变量为$X=(X_1,X_2,\cdots,X_p)$，则在给定X

的条件下，$Y=1$的概率为：

$$P(Y=1|X)=\frac{e^{\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_px_p}}{1+e^{\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_px_p}}$$ （式 4.1.1）

公式 4.1.1 就是 Logistic 函数，它具有 S 分布，如图 4.1.2：

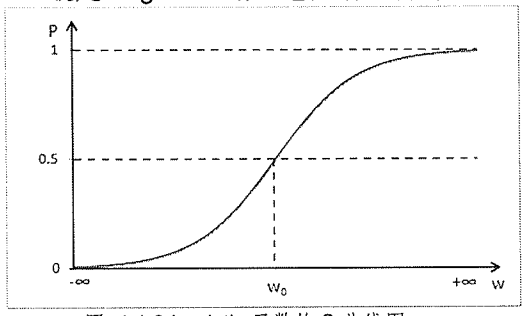

图 4.1.2 Logistic 函数的 S 曲线图

令 $w=\beta_0+\sum_{i=1}^{p}\beta_ix_i$，可以看出，当 w 趋于负无穷大时，$P(Y=1|X)=\lim_{w\to-\infty}\frac{e^w}{1+e^w}=0$，当 w 趋于正无穷大时，$P(Y=1|X)=\lim_{w\to+\infty}\frac{e^w}{1+e^w}=1$。正如图 2.1.1 所示，无论 w 取任何值，Logistic 函数的取值范围均在 0-1 之间变动。Logistic 函数的这一性质保障了由 Logistic 模型估计的概率决不会大于 1 或小于 0。同时这个函数的形状对于研究概率也很适合，当 w 从负无穷开始向右移动时，函数值先是很缓慢地增加，在接近 w_0 时开始迅速增加，之后增加的速度又开始逐渐减缓，最后当 w 趋于正无穷大时，函数值趋于 1。Logistic 函数的 S 曲线表明，w 的作用对于案例发生某一事件的可能性是变化的，在 w 值很小时其作用也很小，然后在中间阶段对应的可能性增加很快，但是在 w 值增加到一定程度以后，可能性就保持在几乎不变的水平，这说明，w 在 $P(Y=1|X)$[1]接近于 0 或 1 时的作用要小于当 $P(Y=1|X)$ 处于中间阶段时的作用。

由 Logistic 回归的定义可知，事件发生的概率为：

$$\pi=P(Y=1|X)=\frac{e^{\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_px_p}}{1+e^{\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_px_p}}$$ （式 4.1.2）

所以，事件不发生的概率为：

$$1-\pi=1-P(Y=1|X)=1-\frac{e^{\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_px_p}}{1+e^{\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_px_p}}=\frac{1}{1+e^{\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_px_p}}$$ （式 4.1.3）

因此，由公式 4.1.2 和公式 4.1.3 可以得出：

$$\frac{\pi}{1-\pi}=e^{\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_px_p}$$ （式 4.1.4）

我们称比率 $\frac{\pi}{1-\pi}$ 为事件的优势比，对上式进行对数变换，则有：

$$g(x_1,x_2,\cdots,x_n)=\log(\frac{\pi}{1-\pi})=\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_px_p$$ （式 4.1.5）

优势比的对数称为 Logit。Logit 变换产生了参数为 $\beta_0,\beta_1,\beta_2,\cdots,\beta_p$ 的一个线性函数，拟合 Logistic 回归模型的参数问题转换为拟合线性模型的参数。通常采用极大似然法（后文简称 MLE）来估计参数。下面介绍 Logistic 回归模型的假设前提、适用条件及验证方法。

Logistic 回归模型也有其假设前提适用条件及验证方法。Logistic 回归模型估计的假设条件与 OLS 线性回归分析的假设条件有相似之处，也有所区别。相似的地方包括：首先，数据必须是总体数据或者来自随机样本。第二、因变量 y_i 被假设为 K 个自变量 $x_{ki}\ (k=1,2,\cdots,K)$ 的函数。第三，Logistic 回归也对多元共线性敏感，自变量之间存在的多元共线性会导致标准误的膨胀。

Logistic 回归模型还有一些与 OLS 回归不同的假设。第一，Logistic 回归的因变量 y_i 是二分变量，这个变量只能取值 0 或 1。研究的兴趣在于事件发生的条件概率，即 $P(y_i=1|x_{ki})$。第二，Logistic 回归中因变量和各自变量之间的关系是非线性的。第三，在 OLS 回归中要假设相同分布性或称方差不变，类似的假设在 Logistic 回归中却不需要。最后，Logistic 回归也没有关于自变量分布的假设条件。各自变量可以是连续变量，也可以是离散变量，还可以是虚拟变量。并且，也不需要假设它们之间存在多元正态分布。

[1] $P(Y=1|X)$ 指事件发生，及 Y=1 时的概率大小

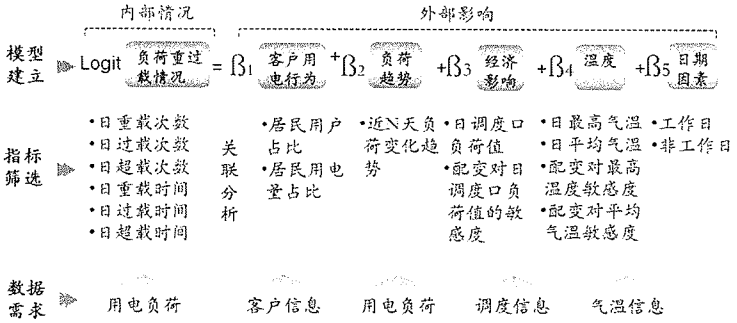

图 4.1.3 短期的 Logistic 模型

在短期预警中应用此模型的建模框架如图 4.1.3 所示,通过输入客户、用电负荷、调度和气温等维度的特征信息建立 Logistic 模型，然后用模型对配变未来发生重过载的概率进行预警。

2. 多元回归

多元回归分析预测法，是指通过对两上或两个以上的自变量与一个因变量的相关分析，建立预测模型进行预测的方法。其模型的数学表达式如下:

设 y 是一个可观测的随机变量，它受到 p 个非随机因素 x_1, $x_2,\ldots,x_n$ 和随机因素 ε 的影响，若 y 与 $x_1,x_2,\ldots,x_p$ 有如下线性关系:

$$y=\beta_0+\beta_1x_1+\cdots+\beta_px_p+\varepsilon \quad (式\ 4.1.6)$$

其中 β_0，$,\ldots,\beta_p$ 是 $p+1$ 个未知参数，ε 是不可测的随机误差，且通常假定 $\varepsilon\sim N(0,\sigma^2)$.我们称式（4.1.6）为多元线性回归模型.称 y 为被解释变量（因变量），$x_i(i=1,2,\cdots,p)$ 为解释变量（自变量）。

对式 4.1.6 取期望得

$$E(y)=\beta_0+\beta_1x_1+\cdots+\beta_px_p \quad (式\ 4.1.7)$$

式 4.1.7 为理论回归方程。

对于一个实际问题，要建立多元回归方程，首先要估计出未知参数 β_0,β_1，$\ldots,\beta_p$，为此我们要进行 n 次独立观测，得到 n 组样本数据 $(x_{i1},x_{i2},\cdots,x_{ip};y_i)$，$i=1,2,\cdots,n$。然后对未知参数进行估计，与一元线性回归时的一样，多元线性回归方程中的未知参数 $\beta_0,\beta_1,\cdots,\beta_p$ 仍然可用最小二乘法来估计。

给定因变量 y 与 $x_1,x_2,\ldots,x_p$ 的 n 组观测值，利用前述方法确定线性回归方程是否有意义，还有待于显著性检验。主要包括回归方程显著性的 F 检验、回归系数的 t 检验以及衡量回归拟合程度的拟合优度检验。

以上便是多元回归模型算法，下面具体举例来说明多元回归模型的具体结构。

$$y_t=\beta_0+\beta_1\times Temp+\beta_2\times Temp^2+\beta_3\times Load+\sum_{i=1}^{7}\sum_{j=6}^{9}s(i,j,t)\cdot c_{ij}+d\cdot t \quad (式4.1.8)$$

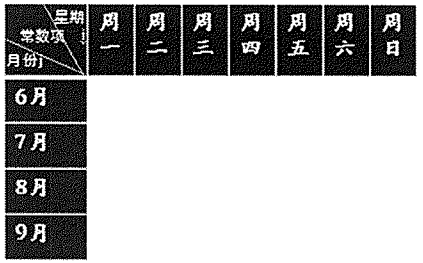

图 4.1.4

在短期总应用多元回归建模大体结构如式 4.1.8，式中，t 表示日期，y 表示日电量，Temp 表示平均温度，β_0表示回归模型的中常数项，Temp 表示温度信息，Load 表示符合信息，$d\cdot t$ 表示电量随时间的线性增长趋势。符号函数$s(i,j,t)$来表征不同星期类型、不同月份的影响，如图 4.1.4 所示，当且仅当第 t 日的星期恰好等于 i，月份恰好等于 j 时，$s(i,j,t)$才取值为 1，否则函数值一律为 0。c_ij表示星期类型为 i,月份为 j 时的回归常数项。

3. 时间序列

时间序列是指以时间顺序形态出现的一连串观测值的集合，或者更确切的说，是对某动态系统随时间连续观察所产生的有顺序的观测值的集合。在统计研究中，常用按时间序列排列的一组随机变量（…，$X_1,X_2,\cdots,X_t,\cdots$）来表示一个随机事件的时间序列，简记为$\{X_t,\ t\in T\}$或$\{X_t\}$。用 $x_1,x_2,\cdots,x_n$ 表示该随机序列的 n 个有序观察值，称之为序列长度为 n 的观察值序列。一个观察值序列属于随机过程的一次样本实现。随机时间序列的一个最基本特征就是相邻两个数据之间有相互依赖性，即：两个随机数据呈现一定的相关。时间序列分析就是依据不同时刻变量的相关关系进行分析，生成随机动态模型来揭示其相关结构并进行预测。模型的数学表达式如下:

若有平稳零均值随机序列$\{X_t\}$及白噪声序列$\{\alpha_t\}$[2]满足

$$X_t-\phi X_{t-1}-\cdots-\phi_p X_{t-p}=\alpha_t-\theta_1\alpha_{t-1}-\cdots-\theta_t\alpha_{t-q}$$ （式 4.1.9）

引入延迟算子 B,记

$$\Phi(B)=1-\phi_1 B-\phi_2 B^2-\cdots-\phi_p B^p$$

$$\Theta(B)=1-\theta_1 B-\theta_2 B^2-\cdots-\theta_q B^q$$ （式 4.1.10）

则式(4.1.9) 又可写为

$$X_t\Phi(B)=\alpha_t\Theta(B)$$ （式 4.1.11）

若$\Phi(B)=0$与$\Theta(B)=0$的根都在单位圆外,则上面的模型即为 ARMA 模型[5]。它是时间序列法的一般形式,可视为一个单入单出的线性系统。当将 ARMA 模型用于预报时,$\{\alpha_t\}$就是残差序列。

[2]唐鸿龄, 等. 应用概率[M] . 南京: 南京工学院出版社, 1988.

表 4.1.3 ARMA（p,q）模型的自相关系数和偏自相关系数特征

模型	相关系数	偏自相关系数
AR（p）	拖尾	P 阶截尾
MA（q）	q 阶截尾	拖尾
ARMA（p,q）	拖尾	拖尾

如表 4.1.3，作为 ARMA 的特例,当 q=0 时称为 AR(p)(p 阶自回归)模型。其特点为:偏自相关函数具有截尾性,自相关函数具有拖尾性;当 p=0 时称为 MA(q)(q 阶滑动平均)模型,其自相关函数具有截尾性,偏相关函数具有拖尾性;对非平稳序列作 d 阶差分后再拟合 ARMA 模型则称为 ARIMA(p,d,q)模型。

（二）算法优选

表 4.1.4 模型对比分析

	Logistic	多元回归	时间序列
输入	各个特征量信息	各个特征量信息	历史负荷信息
输出	每台配变未来一周重过载概率清单	每台配变每天负荷值	每台配变每个小时负荷值
可解释性	高	中	低
预测精度	中	中	高
建模个数	1	建模个数等于配变个数	建模个数等于配变个数
实现难度	中	中	高

根据表 4.1.4 中，综合评估 3 种算法在输入、输出、可解释性、预测精度、建模个数和实现难度上的优劣。在本次建模中，采用 Logistic 算法作为建模算法的首选方案。多元回归和时间序列作为备选方案。