评语：帮助运维人员快速定位高风险台区，优化运维策略和技术改造计划。在迎峰度夏、春节保供电等业务中发挥作用，经济社会效益较为显著，推广应用价值大。

基于关联分析与机器学习的配网台区重过载预测方法

张国宾，王晓蓉，邓春宇
中国电力科学研究院，北京 100192

摘要
针对配电网运行中长期存在的台区重过载问题，提出基于关联规则挖掘的重过载影响因素分析方法，从设备和用户属性、自然环境、短期负荷特性中挖掘针对各类重过载事件的关联规则。从关联项中提取重过载影响因素，并基于机器学习模型，建立重过载事件预测模型，实现对重过载事件的短期预测。最后利用业务系统实际数据，对所提方法进行了效果验证。算例结果表明，新方法能够更为系统、全面地刻画重过载事件，提出的重过载预测模型在命中率和准确率方面表现良好。

关键词
台区重过载；关联分析；机器学习

中图分类号：TP393　　文献标识码：A　　doi: 10.11959/j.issn.2096-0271.2018011

Prediction method in distribution transformer heavy and overload supply areas with relevance analyze and machine learning

ZHANG Guobin, WANG Xiaorong, DENG Chunyu
China Electric Power Research Institute, Beijing 100192, China

Abstract
Aiming at the heavy and over load issue in operation of distribution network, a heavy overload forecasting model to achieve short-term forecast of heavy and overload events was established. Using the actual data of the business system, the proposed method was validated. The results show that the proposed method can describe the heavy and overload events more systematically and comprehensively, and the heavy overload forecasting model based on association rules performs well in the hit rate and accuracy rate. This method provides a new technical means that has certain practical value to enhance the distribution network management which could be seen as important experience of attempts in grid big data as well.

Key words
distribution transformer heavy and overload, relevance analyze, machine learning

1 引言

配变台区作为面向低压用户的最末一级供电单位，台区供电设备的运行状态直接影响台区内的供电质量。设备的重过载运行是引起故障停电的主要原因之一，而重过载现象通常也伴随着三相不平衡、电压偏移等其他问题，严重影响用户安全可靠用电。此外，设备长时间处于重过载状态会加快元件的非正常损耗，降低设备使用寿命，给电网带来故障隐患和运行风险。因此，台区重过载治理一直是配网运维检修工作的重要内容。

目前，对于配变台区的重过载治理通常分为事中监测和事后处理两个阶段，即实时监测配变运行情况，在发现重过载事件后，向运维人员发出告警信息，然后由现场专工根据具体情况采取相应措施进行处理。在大部分情况下，考虑到实际工况中允许设备短时间重过载运行，为保证持续供电，对一般的重过载事件只进行监测，对频繁出现重过载的配变进行报备，并列入周期性技术改造大修计划。而对于可能或已经造成停电的重过载事件，可以通过切改用户线路或临时替换大容量变压器消除重过载现象，但临时停电依然不可避免。

在现有设备水平和电网运行技术条件下，较为理想的重过载治理方式仍是通过合理地安排技术改造大修计划，及时进行调整用户线路和配变增容，减少非计划停电。因此开展重过载影响因素分析并实现事前预测，对于及时发现重过载隐患、优化设备升级改造计划具有重要作用[1]。

在理论研究方面，参考文献[2,3]分析了气象指标、用电类别、行业类别与配变重过载发生概率的变化关系，并就可能的重过载成因进行了初步探讨，最后利用基于随机森林理论改进的决策树模型对配变重过载状态进行预测。参考文献[4]面向负荷增长较快的高速发展区域，从用户、气象和历史负荷出发，提出基于逻辑回归的重过载中长期预测方法，但未给出逻辑回归模型中各项参数的选择过程，方法中对重过载的定义也与参考文献[2]中不同。参考文献[5]针对春节期间的重过载现象，提出基于BP神经网络与灰色模型的重过载预测方法。将春节前后配电变压器的负荷变化作为模型输入对负载率进行预测，进而判断设备重过载情况。通过该方法获取的重过载预测模型不具备泛化能力，也不能适应大规模配电网的快速分析。参考文献[6]以一周为周期，对6 000个居民用户台区每周同一天共7类负荷曲线进行了研究，并加入气象和经济数据，采用不同的机器学习方法对负荷曲线进行了预测。但该方法忽略了负荷的季度年度规律，且对异常值较为敏感，预测效果受到了影响。参考文献[7]基于线性回归模型讨论了在聚合和分解两种模型处理下预测效果的差异，通过算例说明了通过对训练节点进行分解能够有效提高负荷预测精度，但算例只采用了天气和时间数据用于模型训练，对配变负荷影响因素分析不足。参考文献[8]提出了一种趋势分析与指数加权模型的超短期负荷预测方法，该方法能够提高对负荷峰谷的预测精度，且对负荷缺失值和异常值有较好的适应性。参考文献[9]在基本的气象指标之外，研究了风寒指数、空气污染指数、人体舒适指数等气象指数与配变重过载的关系，在此基础上运用小波神经网络（wavelet neural network，WNN）模型对配变负荷进行预测。

在应用实践方面，国网黑龙江省电力有限公司哈尔滨供电公司[10]依托运监中心数据开展重过载监测，同时利用配电变压器型号、设备容量、运行年限等设备台账信

息排查重过载隐患。国网福建省电力有限公司福州供电公司[11]针对城网公变开展重过载预测研究，从风险角度对重过载配电变压器进行多方面综合评估，能够在一定程度上指导运维工作，但模型效果仍有提升空间。

由上可见，目前的研究大多数仍然以配变负荷预测为切入点，包括利用历史负荷数据外推负荷的典型变化规律，加入气象条件、经济指标等外部间接因素分析与负荷的关联程度，尝试采用各类机器学习方法构建负荷预测模型，最后基于负荷预测结果对重过载进行判断。但目前各类负荷预测方法在预测精度上的局限性导致基于负荷预测结果的重过载预测不能适应业务实际情况。对于短期负荷预测，考虑到目前消除重过载的手段以实施周期较长的线路切改和设备改造为主，短期的预测结果不足以提供足够的时间裕度消除重过载隐患；对于中长期负荷预测，考虑到台区负荷的突变性和波动性，负荷预测结果本身不能满足对重过载进行判断的基本精度。

本文直接以重过载事件为目标，从历史运行数据中抽取重过载记录，融合设备台账数据、台区用户数据以及供电环境属性，利用多维关联分析方法挖掘可能与重过载相关的影响因素组合，利用历史数据训练机器学习模型，建立配电台区重过载预测模型。最后，利用某地市供电公司实际数据，对方法分析过程和模型效果进行了验证。

2 整体研究思路

配电台区重过载预测研究大致分为以下3个阶段。

- 数据收集与预处理：从各个业务系统中获得原始数据后，对重过载相关数据项进行初步筛选和清洗，量化分析目标和样本规模。
- 重过载影响因素分析：从原始数据中抽取或设计特征变量，通过多变量多层关联找到单个变量或变量组合与重过载事件之间的强弱关联。
- 重过载预测模型训练：结合上一阶段中获得的相关特征构造机器学习模型，通过历史数据样本对重过载模型参数进行训练及效果评估。

上述研究思路可以进一步分解为以下步骤，如**图1**所示。

- 明确业务定义与模型目标。长期以来重过载管理水平不高导致业务上对于重过载的判定过于简单模糊，没有统一的标准。因此，必须对重过载定义做出清晰、可执行的限定。此外，需要结合业务需求明确模型的技术目标，包括模型的预测精度、预测效率、预测时间跨度等，模型的目标一定程度上决定了特征变量和模型架构的选择。
- 数据获取与预处理。随着电力系统信息化程度的不断加深，目前在电力企业的各业务环节中都具备信息化系统支撑，系统之间的数据贯通和融合不在本文讨论范围内，但数据的获取是开展任何大数据应用研究的前提，本文涉及的业务系统及数据规模会在下文中说明。数据质量决定了模型效果的上限，因此从业务系统中获取原始数据后需要对数据质量进行评估，必要时通过清洗、补齐等技术手段对原始数据进行预处理。

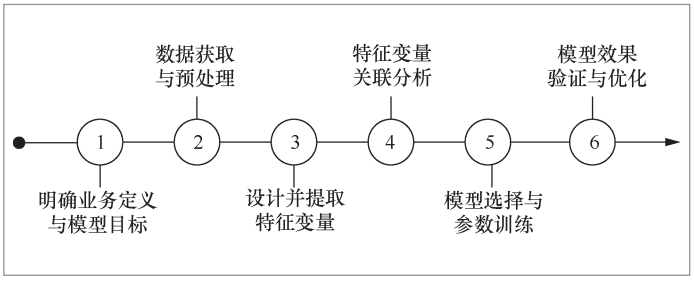


图1 整体研究思路

● 设计并提取特征变量。特征变量是进行关联分析的直接输入，特征变量可以来自原始数据字段，也可以通过对原始数据字段进行加工获得。特征变量的设计可以依靠业务经验，也可以依据某些标准原则。特征变量的提取是指在从原始数据到特征变量数据的实现过程中，通过提取形成用于关联分析的特征变量数据集。

● 特征变量关联分析。通过关联分析方法对特征变量数据集进行操作，得到与重过载相关的关联规则，关联规则中的特征变量集合代表了重过载影响因素。关联规则中的特征变量大多不唯一。关联分析方法的选择需要根据数据类型和分析目标综合确定。

● 模型选择与参数训练。重过载预测属于数据挖掘中的预测问题范畴。从传统的回归方法到ANN、支持向量机（support vector machine，SVM）等人工智能方法，各类预测模型在具体问题上的表现都有所差异。而从实际应用情况来看，机器学习在非线性拟合问题上具有天然优势，同时对于输入特征的自学习性也是本文选择机器学习的重要原因。需要注意的是，成为影响因素的特征变量一般需要进行一定的数据变换才能成为预测模型的输入变量。

● 模型效果验证与优化。通过一部分历史数据对模型效果进行验证，除了准确率、命中率等模型性能指标外，业务层面对模型的认可程度也是评价模型效果的重要依据。

3 方法与技术路线

3.1 业务定义与模型目标

目前实际业务中使用较多的配变重过载的定义通常为：每天00:00配变功率曲线中，连续3个点负载率为80%~100%，视为重载；连续3个点负载率超过100%，视为过载。但从实际数据和业务需求来看，上述定义存在3方面问题：

● 配变重过载中，负载率经常频繁波动，存在大量重载与过载混杂存在的情况，无法按照现有定义进行识别；

● 相当一部分重过载现象在持续时间上不足2 h，这类重过载对设备和电网的影响理论上虽然小于长时间重过载，但对于分析重过载成因具有很高的价值；

● 大部分重过载发生于单相，还有部分重过载会同时发生在多相，如果能够对相别进行识别，则对于重过载治理措施的选择至关重要。

针对上述3方面问题，本文对重过载定义做出调整：每天00:00配变功率曲线中，连续两个或以上点三相负载率中的最大值超过80%，视为重过载。连续3个点负载率超过100%，视为过载。针对上述3方面问题，本文对重过载定义做出调整：每天00:00配变功率曲线中，连续两个或以上点三相负载率中的最大值超过80%，视为重载；连续3个点负载率超过100%，视为过载。

重过载定义优化如**图2**所示。

在模型目标方面，模型结果必须能够与业务环节形成衔接，才能实现模型的应用价值。这就要求模型预测结果必须与目前配变重过载治理方法和周期匹配。针对重过载现象，目前电网侧的治理手段包括以下两方面：

● 对用户进行临时线路切改，这种调整方式相对便捷，处理周期短，但由于台区用户情况复杂，部分重过载台区可能不具备调相条件；

● 通过设备升级改造，永久性解决重过载问题，但该途径周期较长，不适用于突发情况。

结合数据支撑情况和业务流程，对重过载预测的目标确定为：预测精度以天为单位，预测时间跨度不低于3天。

3.2 数据选择与获取

重过载现象本质上反映了相对静态的电源容量不能满足某些条件下的用户用电需求。台区内的用户类型、用户数量、用电行为等不同程度上决定了用户的用电负荷特征。进一步扩展来看，用户方面的各项变化又受到天气、时节、群体活动、经济波动等众多自然环境与社会环境的影响。此外，设备自身的缺陷以及台区内拓扑的不合理同样会导致重过载的发生。恶劣的运行环境引起的设备非正常损耗也会增加重过载风险。

通过数据调研，在充分考虑数据获取途径和难度后，确定用于本项研究的数据规模，如图3所示。内部数据包括某地市公司所属的所有配变设备台账、台区用户的部分档案信息以及变更记录、配变负荷曲线、配变改造记录，覆盖近3万个配电台区。同时受数据粒度所限，外部数据选择了时间数据、节假日信息、各类气象指标等数据。数据时间跨度为2014—2015年底。总数据量超过3亿条。

数据的获取通过系统间接口完成。通过大数据平台与业务数据中心进行数据集成，不直接与各源业务系统（调度、运检、营销等业务信息系统，如EMS/配网自动化/PMS/营销/95598/用电信息采集系统等）对接。业务数据中心直接将相关数据库表提供给大数据平台及应用直接访问，以供大数据平台及应用接入数据，创建中间库，用户在中间库下创建接口视图，源数据根据中间库抽取数据到大数据平台基础库（Hive），或从Hive推送数据到应用业务库（MySQL）。

3.3 特征工程与关联分析

重过载现象受到台区下用户数量和用电行为的直接影响。但用户数量和用电行为会随着外界环境动态变化。而各项环境因素之间常常存在内部横向关联。因此在设计特征变量时，要保证特征变量之间的独立

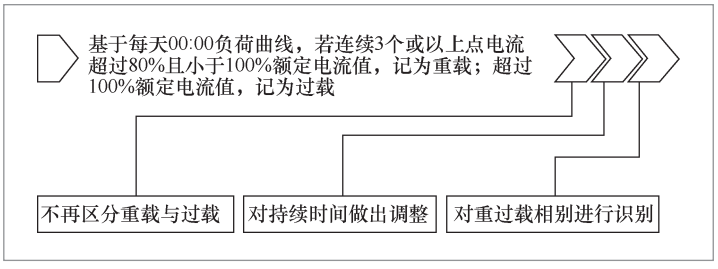


图2 重过载定义优化

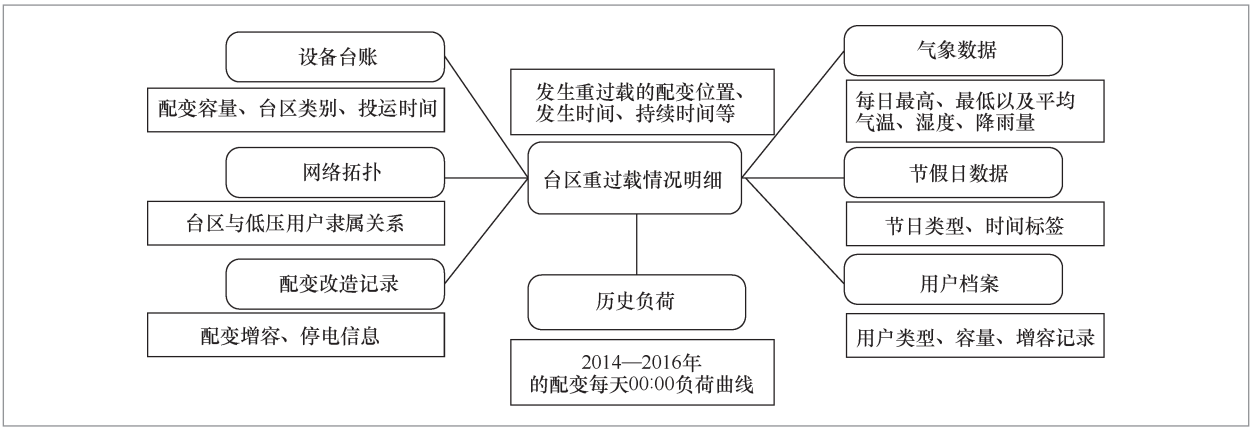


图3 数据分析范围与规模

性。基于数据现状和特征独立性，选择部分原始数据字段设计以下三大类特征。

- 静态特征：额定容量、设备型号、台区类型、用户构成比例。主要用于分析以上各维度下的重过载配变分布规律。
- 时序特征：重要节假日标志、温度、湿度、空气质量指数等气象指标。主要用于分析期间各项指标下的重过载台区随时间变化的趋势。
- 衍生特征：从负荷曲线中提取的短期用电特征，用于分析短期用电特征与重过载的相关性。

具体的特征定义见**表1**。

按照上述特征定义，每一次重过载事件都可以形成一条关联特征向量。关联分析的操作对象是由关联特征向量构成的关联样本集。由于关联特征中同时存在离散型和连续型变量，为了尽量降低数据泛化对分析结果的影响，选择HotSpot算法对关联样本集进行频繁模式挖掘。

频繁模式p形式如下：

$$p:\{a_1\in R_1,a_2\in R_2,\cdots,a_i\in A_i\}\Rightarrow b[\text{surpport}] \quad (1)$$

其中，$a_i\in R_i$为谓词，表示特征项a_i的取值属于某一取值空间R_i；b表示某重过载事件类型标号。surpport表示支持度，定义如下：

$$\text{surpport}(p)=\frac{\text{count}(p)}{\text{count}(\text{Re})}\times 100\% \quad (2)$$

即包含频繁模式p的样本数目在整个关联样本集中的百分比。频繁模式要满足最小支持度min:surpport的要求。

频繁模式体现了某些关联特征在特定的取值范围下出现在重过载事件中的频率。但这些特征与重过载事件的关联关系需要进一步通过计算置信度和相关度确定。

置信度为：

$$\text{confidence}(A\Rightarrow B)=\frac{\text{surpport}(I)}{\text{surpport}(A)}\times 100\% \quad (3)$$

相关度为：

$$\text{co}(A\Rightarrow B)=\frac{\text{surpport}(I)}{\sqrt{\text{surpport}(A)\times\text{surpport}(B)}}\times 100\% \quad (4)$$

在满足支持度阈值的前提下，同时满足最小置信度min:confidence和最小相关度min:co的频繁模式称为关联规则。

3.4 模型训练

通过关联规则确定影响重过载发生的主要因素后，对重过载事件的预测就可以看作数据挖掘中的预测问题，即通过若干条件综

表1 关联特征含义

序号	大类	特征名称	特征类型	特征描述
1	静态特征	额定容量	因子型	配电变压器额定容量
2		设备型号	名义型	配电变压器设备型号
3		用户构成比例	数值型	台区内居民、商业及工业等三类用户的容量占比
4		台区类型	因子型	台区的供电区域类型
5	外部时序特征	温度	数值	重过载当日的平均温度
6		温度	数值	重过载当日的平均温度
7		空气质量指数	数值	重过载当日的空气质量指数
8		节假日标签	名义型	重过载当日的法定节假日代码
9	负荷特征	近5日平均负载率	数值	重过载发生前5日内每天00:00负载率的平均值
10		近5日小时负载率区间次数	数值	重过载发生前5日内每天00:00负载率超过30%、45%、70%的次数
11		近5日负载率变化趋势	数值	重过载发生前5日内每日最大负载率连续上升天数

合判断某一类型重过载事件是否发生。为了降低关联规则中各特征之间可能存在的共线性对模型预测效果的影响，同时防止对数据的过度拟合，选用深度逻辑网络（deep logical network，DLN）模型。模型结构如下：

$$F=f(y) \tag{5}$$

$$y=k_0+k_1C_1+k_2C_2+\cdots+k_iC_i \tag{6}$$

其中，$C_i=\begin{cases}0, & a_i \notin A_i\\ 1, & a_i \in A_i\end{cases}$，$A_i$为该逻辑项的判断域。重过载事件样本$a_i$若满足判断条件，则$C_i$生效。判断域的范围由关联规则中的谓词确定。

$k_1,k_2,\cdots,k_i$各项系数反映了该项特征对预测目标的影响程度，通过样本数据对模型进行训练自动获得。$f(y)$为对数函数，取值范围为(0,1)，表示该样本发生重过载的概率。

对于DLN模型，最大化似然函数和最小化损失函数实际上是等价的。因此在参数训练方法上，为了进一步降低逻辑项之间可能存在的共线性影响，采用了基于不完全数据的混合最大似然估计（maximum likelihood estimation，MLE）方法。

在模型验证方面，采用柯尔莫可洛夫—斯米洛夫（Kolmogorov-Smirnov，K-S）校验方法，比较样本数据的累计频数分布与理论分布，若两者间的差距很小，则推论该样本取自某该类重过载样本集。

方法整体流程如**图4**所示。

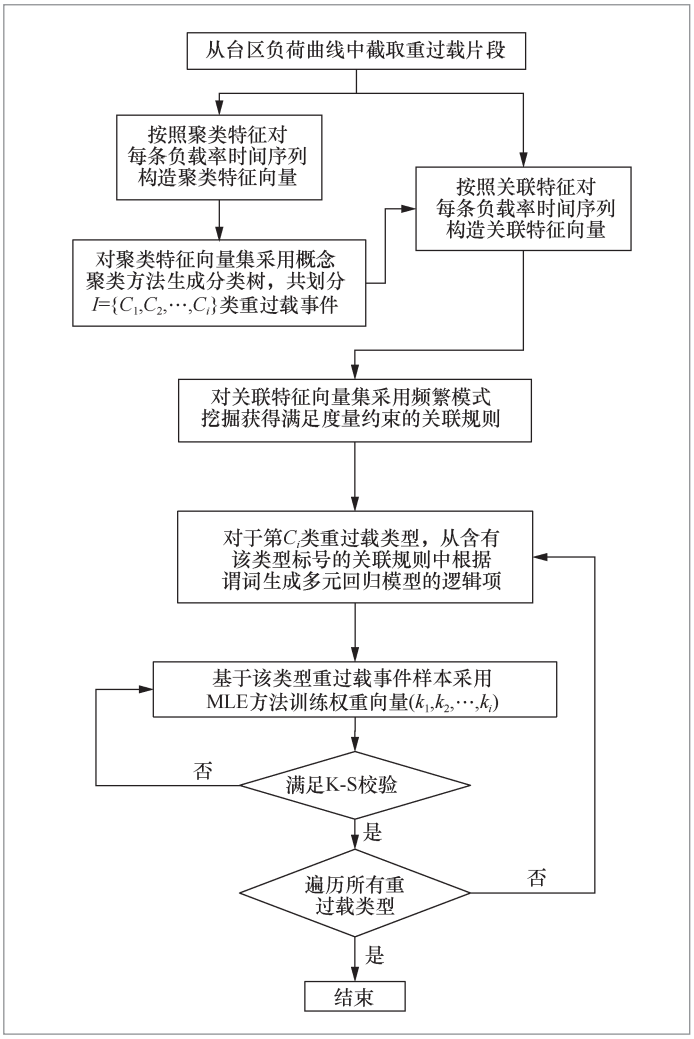


图4 重过载预测方法整体流程

4 算例分析

以某地市供电公司实际数据为例，对本文提出的关联分析以及预测方法进行验证。

4.1 数据质量诊断与预处理

本文运用的数据覆盖近14 000台配变近3年每天00:00的负荷曲线数据、相应的配变设备台账（型号、容量、运行年限、升级改造记录）、台区类型以及用户数量、每日最高气温、每日最低气温、每日湿度、每日降雨量、当日星期类型、节假日标签等。

4.1.1 台账数据

设备台账方面，共收集配变设备台账记录14 904条，数据情况见**表2**。

- 设备型号字段中存在包含中文字符的无效记录共246条，其中222条记录可通过自动修正处理为有效数据；另外，在共计1 342种型号中，保留其中166种主要型号，

对应设备数量占总数的99.5%。

● 额定容量字段中有464条记录不符合标准容量，占设备总数量的3%；生产厂商字段存在552条无效记录，另外保留其中239个主要厂商，对应设备数量占总数的91.3%；区域特征字段存在306条空值记录。以上各项字段中的空值记录和无效记录不纳入本次分析范围。

● 其他字段数据情况良好可用。

4.1.2 负荷数据

历史负荷数据方面，获得2014—2016年共3年的每天00:00的负荷曲线数据，数据情况见**表3**。

从负荷曲线中对重过载事件进行识别，需要尽量确保数据的连续性，按照以下规则对负荷曲线数据进行插补。

表 2 台账数据情况

字段名称	总记录数/条	空值记录数/条	无效记录数/条
配变标识	14 904	0	0
额定容量	14 904	0	464
设备型号	14 904	0	246
生产厂商	14 904	0	552
投运日期	14 904	0	0
区域特征	14 904	306	0

● 数据连续缺失在两个点（包括两个点）以内的按照插值法进行补齐。

● 在数据连续缺失2个点以上、6个点以下的情况，若缺失区间两端时刻的负载率为重过载，则节假日在该端采用移动平均法进行补齐，时期数N=3；非节假日，按照前5日（相同类型）同一时刻的负载率均值计算。对于仅有一个端点负载率在重过载区间（节假日）的情况，在该端采用移动平均法进行补齐，时期数N=3。

● 其他情况一律不做补齐处理。

4.1.3 用户信息数据

用户信息数据方面，共获得低压用户档案记录4 947 519条，数据情况见**表4**。

● 13 697条用户档案记录无法与相应台区进行匹配，属于无效数据。

● 户均容量方面，由于存在多台配变器划归同一台区的情况，因此户均容量记录条数少于配变设备数量。台区户均容量为空的记录共6 453条，其中存在无法与相应用户进行匹配的台区3个，其他空值可以根据现有台区域用户匹配关系进行补充。

● 原始数据中包含410个行业类别，按照《2015年行业用电分类标准》进行泛化处理后，共保留47类。

表 3 负荷曲线数据情况

有功功率	年份	总记录数/条	全空值记录/条	占比
A相	2014年	21 202 803	1 673 224	7.89%
	2015年	117 227 510	5 344 980	4.56%
	2016年	110 545 711	638 220	0.58%
B相	2014年	21 181 283	1 769 469	8.35%
	2015年	116 988 021	5 697 187	4.87%
	2016年	110 445 020	758 437	0.69%
C相	2014年	21 183 270	1 736 968	8.20%
	2015年	116 986 380	5 577 581	4.77%
	2016年	110 443 841	756 313	0.68%

4.1.4 气象指标数据

对气象类信息进行排查，选用了温度、湿度、降水量和风速信息，数据粒度为小时。在进行后续分析之前，需要对坏数据进行处理，剔除无效数据。

4.2 重过载关联因素

为了保证能够获得足够多的频繁模式，同时基于频繁模式产生的关联规则能够反映样本整体规律，一方面可以尝试多组支持度、置信度和相关度阈值，另一方面，也可以对部分关联特征进行适当的预处理。根据特征数值类型，相应的预处理方式见**表5**。

共155 768个重过载事件生成重过载事件样本。采用R语言中的Hotspot包对样本集进行分类。支持度support、置信度confidence和相关度co阈值分别设置为0.3、0.7、0.4。关联分析结果见**表6**。

表6中每条关联规则的谓词都表示了一种影响重过载事件最主要的条件组合，分别如下：

- 台区属于城网类型，当日平均温度低于6℃，近5日内每天00:00负载率超过80%的次数大于20次且小于45次；
- 当日平均温度大于26℃，平均湿度小于60%；
- 近5日内00:00负载率平均值大于45%；
- 近5日内每天00:00负载率平均值大于30%且小于45%，近5日内每天最大负载率连续上升天数为0，日期为周一或周二；
- 近5日内每天最大负载率连续上升天数大于1天且小于4天，近5日内每天00:00负载率超过80%的次数大于45次。

整体上看，外部气象和配变的短期负载特征与重过载事件具有较强的相关性。从指标上看，关联规则2、关联规则3具有较高的确定性，其中关联规则2中的关联因素与重过载事件相关度最高。而用户特征和设备台账特征对重过载事件无明显影响。

表4 用户信息数据情况

字段名称	总记录数/条	空值记录/条	无效记录/条
用户总数	4 947 519	0	13 697
户均容量	12 754	6 453	0
负荷性质	4 947 519	1	0
用电类别	4 947 519	0	0
行业类别	4 947 519	0	0

表5 数据预处理方式

特征名	类型	处理方式
容量、用户类型构成	离散数值型	分箱
负荷特征、温度、湿度、空气质量指数	连续数值型	标准化
台区类型、节假日标签	名义型	编码

表6 关联规则

编号	谓词	support	confidence	co
1	{temp<6℃, Dist=2, 20<Der2<45}	0.62	0.9	0.69
2	{temp>26℃, Hum<60%}	0.87	0.92	0.74
3	{Der1>45%}	0.7	0.88	0.6
4	{30%<Der1<45%, Der3=0, ***T***=Mon|Tur}	0.44	0.73	0.53
5	{Der2>45, 1<Der3<4}	0.39	0.72	0.67

4.3 重过载预测模型训练

根据第4.2节获得的关联规则，模型涉及当日平均温度、湿度、台区类型、日期类型以及短期负载特征等共7项指标，采用11项关联规则中的谓词作为模型参数，其结构如式（7）所示。

$$T=\text{int}(1\ 000/(1+\exp(-y))) \quad (7)$$

其中，$y=k_0+k_1\times$ldrate_aver_d5_1$+k_2\times$ldrate_aver_d5_2$+k_3\times$cnt80_d5_1$+k_4\times$cnt80_d5_2$+k_5\times$tf_loadrate_1$+k_6\times$tf_loadrate_3$+k_7\times$mean_humidity_1$+k_8\times$min_tem_1$+k_9\times$min_tem_3$+k_{10}\times$trans_type$+k_{11}\times$weektype_1。

模型中各项参数意义见**表7**。

每组训练集的样本总量为2 000，按照正负3:1的比例共分割训练集103组，对参数权值进行多次训练，最终结果见**表8**。

模型的最终输出需要给出对重过载事件是否发生的逻辑判断。在确定模型参数和权重取值后，按照K-S准则对y值进行区间划分，结果如**图5**所示。在y值为3%时，K-S值达到最大值95%，以此作为判断重过载事件是否发生的分界点。

在剩余的1 268个重过载样本中抽取20%样本（256条重过载样本），并加入150条非重过载样本作为检测集，对模型效果进行验证，结果见**表9**。

根据**表9**可以得到该模型准确率（ac_rate）和命中率（tar_rate）分别为：

$$\text{ac\_rate}=\frac{RS_3}{RS_3+RS_4}\times 100\%=85\% \quad (8)$$

$$\text{tar\_rate}=\frac{RS_3}{RS_1}\times 100\%=90.6\% \quad (9)$$

预测结果中实际发生重过载的台区数量决定了准确率，预测结果中没有包含但实际发生重过载的台区数量决定了命中率。实际业务中考虑到运维和抢修资源配置，当两项指标不能达到业务要求时，需要对以上两项指标进行取舍。在模型训练过程中，通常采取“优先保证准确率，尽量提高命中率”的策略保证预测结果。

对于业务人员感兴趣的其他类型过载事件，同样可以按照上述过程建立相应的重过载预测模型。所有的重过载模型构成完整的重过载模型库，能够实现对重过载事件的综合判断。

特别地，对于同一个台区预测出现多类重过载事件的情况，可以通过对各预测模型给出的y值进行统一标准化后，取其中

表7 参数描述

参数名	描述
ldrate_aver_d5_1	近5日内负载率均值45%以上
ldrate_aver_d5_2	近5日内负载率均值大于30%小于45%
cnt80_d5_1	近5日内每天00:00负载率超过70%的次数大于45次
cnt80_d5_2	近5日内每天00:00负载率超过70%的次数大于20次且小于45次
tf_loadrate_1	近5日内每日最大负载率无上升趋势
tf_loadrate_3	近5日内每日最大负载率连续2天或者3天连续上升
mean_humidity_1	平均湿度在60%或以下
min_tem_1	平均气温在12℃以下
min_tem_3	平均气温在26℃以下
trans_type	属于城网台区
weektype_1	工作日周一或周二

最大值代表的重过载类型作为对该台区的最终预测结果。

5 结束语

本文针对配网重过载管理中存在的难题和业务空白，结合大数据中的数据挖掘技术提出了重过载分类、重过载关联分析以及重过载预测方法，基于实际数据对方法进行验证，效果达到了预期目标。

随着智能配电网的深化推进，未来电网的大数据特征必将更加明显，大数据技术在处理海量数据、揭示事物间隐藏关联等方面的天然优势，能够成为解决电网业务问题、提升电网运营水平的重要技术手段。

表8 参数权重

权重	取值	权重	取值	权重	取值
k_0	−6.512	k_4	3.182	k_8	0.846
k_1	1.942	k_5	3.474	k_9	2.745
k_2	−2.545	k_6	0.501	k_{10}	1.520
k_3	2.729	k_7	0.451	k_{11}	0.712

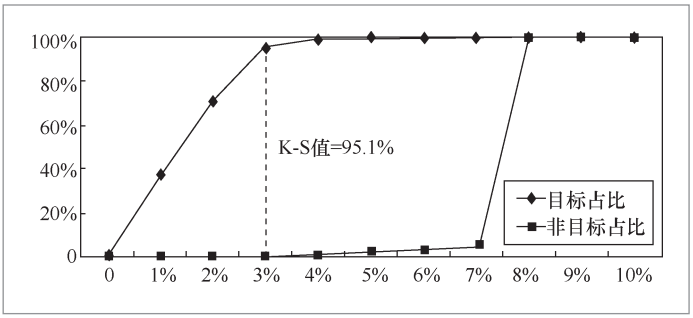


图5 K–S检验

表9 模型效果验证结果

样本集合	集合类型	数量/条
RS_1	实际发生	256
RS_2	实际未发生	150
RS_3	预测发生，实际发生	232
RS_4	预测发生，实际未发生	41

参考文献:

[1] 王继业. 智能电网大数据[M]. 北京: 中国电力出版社, 2017.
WANG J Y. Big data in smart grid[M]. Beijing: China Electric Power Press, 2017.

[2] 贺建章, 王海波, 季知祥, 等. 面向智能电网的配电变压器重过载影响因素分析[J]. 电网技术, 2017, 41(1): 279−284.
HE J Z, WANG H B, JI Z X, et al. Analysis of factors affecting distribution transformer overload in smart grid[J]. Power System Technology, 2017, 41(1): 279−284.

[3] 贺建章, 王海波, 季知祥, 等. 基于随机森林理论的配电变压器重过载预测[J]. 电网技术, 2017, 41(8): 2594−2597.
HE J Z, WANG H B, JI Z X, et al. Heavy overload forecasting of distribution transformers based on random forest theory[J]. Power System Technology, 2017, 41(8): 2594−2597.

[4] LI M, ZHOU Q. Distribution transformer mid−term heavy and over load pre−warning based on logistic regression[C]//2015 IEEE Eindhoven Powertech, June 29−July 2, 2015, Eindhoven, Netherlands. [S.l.:s.n.], 2015: 1−5.

[5] 史常凯, 闫文棋, 张筱慧, 等. 基于BP网络和灰色模型的春节配变重过载预测[J]. 电力科学与技术学报, 2016, 31(1): 140−145.
SHI C K, YAN W Q, ZHANG X H, et al. Heavy overload forecasting of distribution transformer during the spring festival based on BP network and grey model[J]. Journal of Electric Power Science and Technology, 2016, 31(1): 140−145.

[6] PADMANABH K, SINGH M J. Load forecasting at distribution transformer using IoT based smart meter data[C]//2016 IEEE International Conference on Contemporary Computing and Informatics, December 14−17, 2016,

Noida, India. New York: IEEE Press, 2016: 758-763.

[7] CAMPEZIDOU S I, GRIJALVA S. Dsitribution transformer short-term load forecasting models[J]. IEEE Transactions on Power System, 2016(19): 267-273.

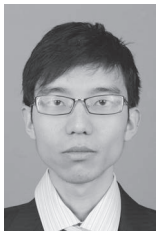
[8] NGO V C, WU W C, ZHANG B M. Ultra-short-term load forecasting using robust exponentially weighted method in distribution networks[J]. Journal of Information, Control and Management Systems, 2015(9): 301-308.

[9] SUN X R, LUH P B, CHEUNG K W. Efficient approach to short-term load forecasting[J]. IEEE Transactions on Power Systems, 2016(3): 301-307.

[10] 赵宇, 董莉丽. 基于运监大数据挖掘的公变台区重过载监测分析[J]. 企业管理, 2016 S(2): 276-277.
ZHAO Y, DONG L L. Heavy and over load analysis on distribution transformer using operative and monitoring data[J]. Enterprise Management, 2016(S2): 276-277.

[11] 王山林, 林正祥. 国网福建电力应用大数据预测重过载台区[N]. 国家电网报, 2016-08-19.
WANG S L, LIN Z X. Practise of big data on heavy and over load forecasting[N]. State Grid News, 2016-08-19.

作者简介



张国宾(1988-),男,中国电力科学研究院工程师,主要从事智能电网大数据、配电网规划方面的研究工作。

王晓蓉(1973-),女,博士,中国电力科学研究院教授级高级工程师,主要从事新能源发电规划设计和功率预测、智能电网大数据等方面的研究工作。

邓春宇(1983-),男,博士,中国电力科学研究院高级工程师,主要从事电力信息化建设及电力大数据应用工作。

收稿日期:2017-12-17

评语：北京宜信致诚信用管理有限公司阿福平台通过接入各种金融数据并采用先进的数据处理技术，提升了信用判断的准确性，有效防范了多头负债等过去难以规避的风险。

金融科技在互联网金融行业性风险防范领域的应用

胡鹏飞
北京宜信致诚信用管理有限公司，北京 100022

摘要
互联网金融领域出现的越来越多的欺诈风险、多头负债等现象，已经严重影响了行业的健康发展。详细分析了互联网金融行业面临的挑战，采用分布式架构设计及API调用技术，设计了一个行业机构间的共享生态系统，并以知识图谱为底层架构，设计了更适用于金融场景的社交网络图谱，构建了反欺诈的风控体系，实现了互联网金融行业数据的互联与互通，有效防范了行业多头负债的发生，提升了行业反欺诈能力。

关键词
反欺诈；共享平台；知识图谱

中图分类号：TP393　　文献标识码：A　　doi: 10.11959/j.issn.2096-0271.2018012

Application of FinTech in internet financial industry risk prevention

HU Pengfei
Zhicheng Credit Service Co., Ltd., Beijing 100022, China

Abstract
In view of the increasing fraud risk and long-term liabilities appearing in the field of internet finance, the healthy development of the industry has been severely affected. The challenges facing the internet financial industry were analyzed in detail. Based on distributed architecture design and API calling technology, a shared ecosystem between industry organizations was designed. A social network knowledge graph for a financial scene was proposed. The anti-fraud risk control system was designed. The industry's anti fraud ability was promoted.

Key words
anti-fraud, share-platform, knowledge graph