

基于数据挖掘的配电网故障风险预警

刘科研¹, 吴心忠², 石琛², 贾东梨¹

(1. 中国电力科学研究院, 北京 100192; 2. 北京交通大学 电气工程学院, 北京 100044)

摘要: 为了提高配电网风险预警的准确性, 提出了基于数据挖掘的配电网故障关联因素分析与风险预警的方法。通过数据清洗、数据变换、数据集成和离群样本剔除, 归纳配电网四大类共 28 个故障特征; 采用改进的 Relief-Wrapper 算法进行故障关联因素分析, 剔除了 6 个冗余特征, 形成了由 22 个故障特征组成的最优故障特征子集; 提出了兼顾故障发生频率和失电负荷比例的配电网故障风险指标和风险等级划分方法, 采用基于径向基函数的支持向量机(SVM)方法和最优故障特征子集进行风险预警。对某市 120 条馈线配电网进行了风险预警算例分析, 结果验证了所提方法的有效性。

关键词: 配电网; 数据挖掘; 故障关联因素; 最优故障特征子集; 风险预警; 支持向量机; 风险指标

中图分类号: TM 732

文献标识码: A

DOI: 10.16081/j.issn.1006-6047.2018.05.022

0 引言

据统计, 用户停电事故中 80% 以上都为配电网发生故障导致的, 因此, 对配电网运行过程中存在的故障风险进行有效的预警, 及时采取风险防控措施, 对保障供电安全性和可靠性显得尤为重要^[1-2]。

停电事故风险预警技术已较多地应用在输电网中, 主要基于潮流、电压、频率等状态监测数据, 并对输电网的运行风险进行评估并预警^[3-4], 但是这些成果很难应用到配电网中。配电网的网络拓扑复杂, 设备种类繁多、分布相对分散, 设备运行状态容易受到外部因素的影响, 且故障发生的随机性强, 因此难以通过分析机理实现风险预警。

文献[5]在配电网实时风险评估中, 考虑了设备实时故障率, 并对故障停电影响后果中的各指标进行量化。文献[6]以用户为中心进行风险评估, 近似采用稳态的失效率计算故障率, 但仅能用于评估短期的运行风险状况。文献[7]通过分析突发事件对设备故障率的影响建立评估模型, 考虑孤岛运行情况, 并给出了风险预警等级和划分方法。文献[8]以最小过热区域为单位计算停运概率, 同时考虑故障恢复重构, 计算不同灾害下的停电指标, 缩短了计算时间。传统的配电网风险评估预警研究主要从故障率和故障影响后果入手, 偏向于建立停电风险指标, 对停电风险关联因素及其风险预警模型的建立涉及较少。

近年来, 数据挖掘算法凭借计算速度快、泛化能

力强等优点广泛应用于电网研究和工程实际。数据挖掘的重要应用之一就是预测性分析, 数据挖掘算法能够从大规模海量数据中挖掘数据间的潜在规则^[9]。文献[10]在电力系统暂态稳定评估中综合利用不同参数训练的支持向量机 SVM (Support Vector Machine), 减少了“误判稳定”样本的个数。文献[11]利用因子分析法提取风险指标中的共有因素, 采用 SVM 构建故障风险模型, 实现了电网的故障风险评估。文献[12]在分析发电机的进相能力时, 建立了基于相关向量机(RVM)的模型, 提高了模型精度。

因此, 针对配电网故障停电风险, 本文提出一种基于数据挖掘的配电网风险预警方法。基于改进 Relief-Wrapper 算法分析配电网故障关联因素, 筛选 22 个与故障强相关的故障特征变量构成最优故障特征子集; 提出一种基于径向基核函数 RBF(Radical Basic Function)的 SVM 方法和最优故障特征子集的配电网风险评估指标和风险等级划分方法; 以某地级市配电网 120 条馈线作为算例进行测试, 验证了所提风险预警模型和方法的有效性、准确性, 对工程实际应用有一定的指导意义。

1 配电网原始故障数据的预处理

本文研究是针对某地级市配电网 120 条馈线展开的。通过对该市配电网信息管理系统进行调研, 从营销业务管理系统、企业资源计划(ERP)系统、配电自动化系统、用电信息采集系统、配电线路在线监测系统、生产管理系统、配电地理信息系统、智能公用配变监测系统这 8 个信息子系统中, 获取了从 2014 年 1 月至 2015 年 6 月 1.5 a 的馈线故障相关数据, 并收集该区域对应时间段的温度、降水等天气数据。

数据预处理包括数据清洗、数据变换、数据集成和离群样本剔除 4 个步骤^[13]。

收稿日期: 2017-07-06; **修回日期:** 2018-03-01

基金项目: 国家电网公司科技项目(支撑精益化管理的配电网大数据分析技术与基础平台开发)(52020116000G)

Project supported by the Science and Technology Program of State Grid Corporation of China (Big Data Analysis Technology and Basic Platform Development in Distribution Network for Supporting Lean Management) (52020116000G)

a. 数据清洗:包含数据空缺值处理、数据异常值处理、数据重复值处理。数据空缺值处理主要是对原始数据中记录缺失和记录中的某个缺失字段缺失进行剔除或补充;数据异常值处理是根据原始数据的特点,制定相应的规则对偏差过大的数据进行剔除或替换;数据重复值处理是根据数据自身的特点,对重复的数据进行剔除。

b. 数据变换:将原有的数据转换为易于分析和应用的形式,主要内容包含特征构造、数据分级及数据量化等,如量化位置信息、投运时间构建特征属性、天气数据的分级分析等。以月降水量为例,据统计该地级市的月降水量主要集中于5—9月,约占全年降水量的57%(其中6月降水量最高,7月次之,5月和8、9月降水量比较相近);3、4月的降水量次之,约占全年降水量的18.9%;10月至次年2月的降水量较少,其降水量和仅约占全年降水量的24.1%。故可将月降水量分成3个等级。同样地,月雷暴日数和月大风日数也分成3个等级。从数据分析的结果可知,馈线的故障情况与温度、降雨、大风和雷击等环境条件密切相关,且在一年四季中是随时间而变化的。

c. 数据集成:进行数据统计,将数据合并到某个统一的数据库中,馈线故障风险预警所需数据来自于不同的配电信息管理系统,因此需要对原始数据进行统计分析与合并。

d. 离群样本剔除:经过前述预处理的原始数据中还可能包含异常的样本,与同一数据集中的绝大部分数据差异很大,这种数据称之为离群样本。可采用基于统计的、基于邻近值或基于聚类的方法,加以识别和剔除。限于篇幅,本文不再展开。

由于各特征变量的影响程度与其取值范围有密切关系,所以所有变量都预先按式(1)进行归一化处理。

$$\bar{x} = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \quad (1)$$

其中, $x_{\max}$ 、 $x_{\min}$ 分别为变量取值的最大值和最小值。

通过数据预处理可将配电网故障特征分成故障因素、外部因素、自身因素和运行因素4类共28个,如表1所示。

2 配电网故障关联因素分析

为了提高配电网风险预警的效率和准确性,必须先从海量数据和众多故障特征中识别和提取与配电网故障密切相关的故障特征,剔除相关性不强和冗余的特征。Relief算法是一种得到广泛应用的特征选择算法,由Kira和Rendell首先提出,该算法的要点是根据特征对近距离样本的区分能力来评价特征的重要性^[14]。其基本思想是:对所有特征各赋予

表1 配电网故障特征变量分类

Table 1 Classification of distribution network fault characteristic variables

类型	变量	故障特征变量	类型	变量	故障特征变量
故障因素	f_1	故障时间	自身因素	f_{15}	变压器数
	f_2	馈线所属变电站		f_{16}	熔断器数
	f_3	$n-1$ 月故障数		f_{17}	分段绝缘导线长度
	f_4	$n-2$ 月故障数		f_{18}	分段绝缘导线数
	f_5	$n-3$ 月故障数		f_{19}	电缆长度
	f_6	失负荷总量		f_{20}	馈线分支线路数
外部因素	f_7	月平均气温	运行因素	f_{21}	馈线月平均负荷
	f_8	月平均高温		f_{22}	馈线月最大负荷
	f_9	月平均低温		f_{23}	熔断器平均投运时间
	f_{10}	极端高温		f_{24}	分段电缆平均投运时间
	f_{11}	极端低温		f_{25}	负荷开关平均投运时间
	f_{12}	月降水量分级		f_{26}	分支线路平均投运时间
	f_{13}	月雷暴日分级		f_{27}	变压器平均投运时间
	f_{14}	月大风日分级		f_{28}	绝缘导线平均投运时间

一个权值,以表征特征与每类的相关程度,然后利用假设间隔的概念对权值进行迭代计算,最后根据权重的大小确定所选择的特征。

Relief算法的步骤如下:对每个特征赋予权值初值 $W_j = 0 (j = 1, 2, \dots, N)$;选择样本 $x_i (i = 1, 2, \dots, t; t$ 为从样本集中抽取的样本数),从同类样本集中找到样本 x_i 的最近邻样本 $H(x_i)$,从异类样本集中找到 x_i 的最近邻样本特征 $M(x_i)$ 。 W_j 的更新方式如下:

$$W_j = W_j - \frac{\text{diff}(j, x_i, H(x_i))}{t} + \frac{\text{diff}(j, x_i, M(x_i))}{t}$$

其中, x_i 为从数据集随机选择的一个样本; W_j 为特征 j 的权值; $H(x_i)$ 为与 x_i 同类的最近邻样本; $M(x_i)$ 为与 x_i 非同类的最近邻样本; $\text{diff}(\cdot)$ 为距离函数。

Relief算法可以给出每个特征的权重系数,但不能识别冗余特征。因此本文提出一种改进的Relief-Wrapper算法,采用Pearson相关系数的绝对值衡量2个特征 f_i 和 f_j 的相关性 ρ_{ij} :

$$\rho_{ij} = \frac{\text{cov}(f_i, f_j)}{\sqrt{\sigma_{f_i} \sigma_{f_j}}} \quad (2)$$

其中, $\text{cov}(f_i, f_j)$ 为 f_i 与 f_j 的协方差; σ_{f_i} 、 σ_{f_j} 分别为 f_i 和 f_j 的方差。

单个特征 f_i 与已选特征集合 s 之间的相关性指标可以用 f_i 与 s 中任一特征之间相关系数绝对值的最大值描述:

$$\rho_{f_i, s} = \max_{f_j \in s} (|\rho_{ij}|) \quad (3)$$

综合考虑特征的重要性与冗余性,定义如下的特征评价指标:

$$J_{f_i} = W_{f_i} - R\rho_{f_i, s} \quad (4)$$

其中, $W_{f_i} \in [0, 1]$ 为归一化后的单个特征 f_i 的Relief权重,该值越大,表示该特征与配电网故障之间的关

系越密切; $\rho_{f_i, s} \in [0, 1]$ 为特征 f_i 与已选特征集合 s 之间的相关性指标, $\rho_{f_i, s}$ 的值越大, 表示特征 f_i 与已选特征集 s 中的特征越冗余; R 为相关性指标所占的惩罚因子, 当 $R=0$ 时, 式(4)退化为不考虑冗余因素的原始 Relief 权重计算公式, 当 R 的值较大时, 相当于加重了由相关系数描述的冗余性的惩罚力度, 使冗余度越小的特征综合权重越大, 越容易被选中。

3 基于 RBF-SVM 的配电网风险预警

SVM 是数据挖掘中用于分类识别的常用技术, 本文采用的是基于 RBF 的 SVM, 其非线性映射能力能将低维空间的非线性问题映射到高维空间, 增强识别对象的线性可分性。基于 RBF 的 SVM 算法的详细步骤参见文献[15], 本文不再赘述。

配电网的风险不但与停电的频繁程度有关, 还与每次停电事故中的失电范围有关^[16]。频繁程度可以用停电次数表示, 它是一个基于统计的概率量; 而失电范围可以用每次停电的失负荷总量表示, 它表征了此次事故的严重程度。实际工作中每个馈线分区核定的报装负荷总量不一样, 给本文的故障风险评估和风险等级划分带来一定的困难。为此本文提出一种配电网风险评估指标和风险等级划分方法。以每一个馈线分区每个月的运行数据和状态为统计分析对象和风险预警对象, 记 S_{iN} 为第 i 个馈线分区的核装容量, S_{ij} 为统计对象月中第 i 个馈线分区第 j 次停电事故中的失负荷总量, n_t 为当月停电事故总数, 则该馈线分区当月的失电负荷比例累积值 C_i 为:

$$C_i = \sum_{j=1}^{n_t} \frac{S_{ij}}{S_{iN}} \quad (5)$$

根据以上指标将配电网馈线风险划分成 3 个等级, 如表 2 所示。综合考虑 n_t 和 C_i 这 2 个指标, 取其中最严重的风险等级, 如在 1 次停电中失电负荷比例超过 50%, 则当月的风险等级就应该定为 2 级。

表 2 配电网风险等级划分

Table 2 Classification of distribution network risk level

风险等级	状态	n_t	C_i
1	正常	≤ 2 次	$< 50\%$
2	较高	3~6 次	50%~100%
3	严重	≥ 7 次	$> 100\%$

本文采用预警准确率及 Kappa 统计指标作为馈线故障预测模型评估指标。Kappa 统计指标一般用于评判预测模型的分类预测结果与随机分类的差异度, 通过误差矩阵的值计算得到。误差矩阵是一个 $n_s \times n_s$ 阶矩阵 (n_s 为分类数), 矩阵的行代表分类点,

列代表预测点, 对角线部分指某类测试样本与预测类别完全一致的数量。文献[17]给出了 Kappa 统计指标计算方法:

$$K = \frac{N_s \sum_{i=1}^r x_{ii} - \sum_{i=1}^r x_{i+} x_{+i}}{N^2 - \sum_{i=1}^r x_{i+} x_{+i}} \quad (6)$$

其中, K 为 Kappa 统计指标; r 为误差矩阵的行数; x_{ii} 为第 i 行第 i 列(主对角线)上的值; x_{i+} 和 x_{+i} 分别为第 i 行的和与第 i 列的和; N_s 为测试样本总数。 $K \in [0, 1]$, K 值越大表明准确率越高。

本文的研究思路和风险预警流程如图 1 所示。

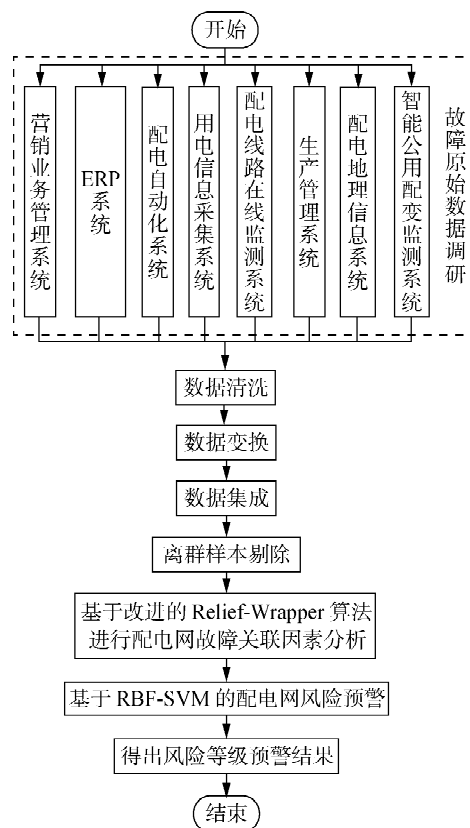


图 1 风险预警流程图

Fig.1 Flowchart of risk early warning

4 算例分析

4.1 故障关联因素分析

以某地级市配电网 120 条馈线 1.5 a 的运行数据为基础, 以从 8 个配电网管理信息子系统中采集的数据为分析对象, 以每一条馈线每个月的运行数据为一个统计单位, 采用前述的方法进行数据清洗、数据变换、数据集成和离群样本剔除, 归纳得到如表 1 所示的 4 类 28 个故障特征; 采用前述改进的 Relief-Wrapper 算法进行故障关联因素分析, 形成由 22 个故障特征组成的最优故障特征子集, 剔除 6 个冗余特征, 如表 3 所示。

表3 故障特征关联因素分析

Table 3 Correlation factor analysis of fault characteristics

类型	最优故障特征子集	剔除故障特征	类型	最优故障特征子集	剔除故障特征
故障因素	f_1-f_6	—	自身因素	$f_{15}-f_{17}$	$f_{18}-f_{20}$
外部因素	f_7-f_{14}	—	运行因素	$f_{21}-f_{25}$	$f_{26}-f_{28}$

采用 Relief 算法可以计算得到所有特征值的权重大小,如图2所示。本文作为特征筛选阈值的设定遵循统计学计算的规律,即将这些权重值归一化为 $[0,1]$ 的值。基于统计的置信区间的概念,通常95%的置信水平是大概率事件,因此,可以假设特征的权重值大小表示该特征所包含整个特征集中信息量的多少,那么,如果信息的累积比例达到95%,则可以认为这些特征包含了全体特征集合的大部分信息量,同时,这些特征可以被认为是最重要的。根据这一假设,门槛值 τ 的设定规则如下:归一化权重值已经按照降序进行排序并累加计算,当计算停止时,该累积值达到95%;上述假设中所提到的通过 Relief 算法计算的权重值表示特征的重要程度,因此,当累加计算停止时,所对应的归一化的权重值就可以设定为特征筛选阈值 τ 。图2中的虚线即为通过计算得到的阈值 τ ,因此,这里被筛选剔除的关联度较低的特征为 $\{f_{18}, f_{19}, f_{20}, f_{26}, f_{27}, f_{28}\}$ 。

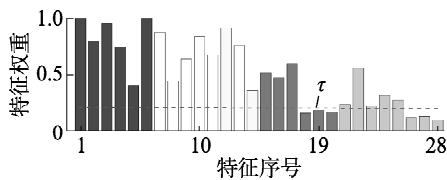


图2 特征权重值

Fig.2 Weighted values of characteristics

从特征选择的结果来看,与馈线故障直接相关的故障特征全部得以保留,它们代表了故障的时间特征、地域特征、故障自相关因素等,它们与故障强相关且相互独立。而在自身和运行因素2类故障特征中均存在一些冗余因素得以有效的剔除,说明了本文所提方法的选择性和有效性。从关联因素分析结果还可看出,馈线的故障情况不仅与馈线长度和负荷情况有关,还与温度、降雨、大风和雷击等环境因素密切相关。所以本文旨在提出一种基于大数据的配电网故障关联因素分析和故障预警的模型与方法,针对不同地区的配电网必须从当地配电系统中获取运行与故障数据,所得到的故障关联因素也可能是不一样的。

4.2 风险预警

由于不同馈线的故障风险是随时间变化的,考虑到这种时变因素,本文选取120条馈线的2014年4月至2015年3月的一整年数据作为训练样本数据,2015年4月的数据作为测试样本数据,2015年

5、6月的样本作为预测样本。其中,训练样本用于训练获取模型,测试样本用于模型的优化,预测样本用于验证模型的实际预警效果。

各风险等级的馈线样本数量如表4所示。

表4 各风险等级样本数量

Table 4 Sample quantity of each risk level

数据类型	样本数量			总数
	风险等级1	风险等级2	风险等级3	
训练样本	1 078	290	72	1 440
测试样本	101	15	4	120
预测样本	198	36	6	240

采用基于 RBF 的 SVM 方法和22个最优故障特征子集,对预测样本中120条馈线在2015年5、6月的月风险等级进行预警,具体的馈线预警结果如表5所示,表中月预测准确率为当月准确预测的样本数与当月馈线样本总数的比值。

表5 配电网故障风险等级预警结果

Table 5 Early warning results of fault risk level for distribution network

预测月	馈线实际故障等级	馈线样本数			等级预测准确率/%	月预测准确率/%
		预测为故障等级1	预测为故障等级2	预测为故障等级3		
5月	1	96	4	0	96	93.33
	2	3	14	0	82.35	
	3	0	1	2	66.70	
6月	1	94	4	0	95.92	94.17
	2	2	15	0	88.24	
	3	0	1	4	80	

由表5可见,每个月的预警准确率都在90%以上,2个月的平均预警准确率为93.75%,说明本文提取的故障特征子集和预警方法具有较高的准确性和实用性;预警失误的样本最多只偏差一个等级,不会跨越2个等级,说明本文方法的鲁棒性较强;5月份等级2和3的样本预测精度较低,是因为这2个等级的原始样本太少。在原始样本中,风险等级1的样本占74%,等级2和3的样本分别只占20%和6%,这也符合配电网故障的实际情况。若能在训练样本中增加高风险等级的样本数量,还能进一步提高预警的准确率。由于各特征变量的时变性,导致配电网的风险也是随时间变化的,而本文提出的方法可以很好地适应这种时变特征,为风险防控提供及时的技术支持。

为了验证本文采用的基于 RBF 的 SVM 进行风险预警的优越性,还与人工神经网络(ANN)、C4.5 决策树和基于线性核函数的 SVM 这3种常用的预测方法进行了对比,其准确率和 Kappa 指标如表6所示。

由表6可见,本文采用的基于 RBF 的 SVM 预警方法,不论是在预警准确率还是 Kappa 统计指标上,都优于其他3种方法。

表6 预测方法结果对比

Table 6 Results comparison among four forecasting methods

方法	预警准确率/%	Kappa 指标	方法	预警准确率/%	Kappa 指标
ANN	83.69	0.413 8	线性核函数 SVM	90.14	0.810 2
C4.5 决策树	88.91	0.765 1	RBF-SVM	93.75	0.863 1

5 结论

a. 本文基于数据挖掘的方法以某地级市配电网120条馈线1.5a的数据为研究对象,对配电网故障关联因素分析和风险预警的模型和方法进行了研究;提出了基于改进Relief-Wrapper算法的配电网故障关联因素分析方法,提取了22个与故障强相关的故障特征构成最优故障特征子集,提高了风险预警的效率和准确性。

b. 提出了兼顾故障发生频率和失电负荷比例的配电网故障风险指标和风险等级划分方法,采用基于RBF的SVM方法和最优故障特征子集进行风险预警。对某地级市120条馈线进行了风险预警算例分析,验证了本文所提风险预警模型和方法的有效性和正确性,为后续配电网在线风险防控提供了理论依据。

c. 建议加强配电网信息管理系统的互联与融合,加强配电网海量数据的挖掘和应用力度,进一步提高配电网的精益化管理水平,降低故障停电风险。

d. 本文旨在提出一种基于大数据的配电网故障因素关联分析和风险预警的模型和方法,由于配电网故障具有地域性、环境敏感性和时变性,所以针对不同地区的配电网,必须从当地的配电系统中获取第一手的运行与故障数据,识别的故障关联因素也可能不尽相同。

参考文献:

- [1] BILLITON R, WANG P. Reliability-network-equivalent approach to distribution-system-reliability evaluation[J]. IEE Proceedings - Generation, Transmission and Distribution, 1998, 145(2): 149-153.
- [2] 李蕊,李跃,苏剑,等. 配电网重要电力用户停电损失及应急策略[J]. 电网技术, 2011, 35(10): 170-176.
LI Rui, LI Yue, SU Jian, et al. Power supply in interruption cost of important power consumers in distribution network and its emergency management[J]. Power System Technology, 2011, 35(10): 170-176.
- [3] 文云峰,崔建磊,张金江,等. 面向调度运行的电网安全风险管控系统(一)概念及架构与功能设计[J]. 电力系统自动化, 2013, 37(9): 66-73.
WEN Yunfeng, CUI Jianlei, ZHANG Jinjiang, et al. Design of a security risk management system for power system dispatching and operation part one concepts and design of architecture and function[J]. Automation of Electric Power Systems, 2013, 37(9): 66-73.
- [4] 李碧君,方勇杰,徐泰山. 关于电网运行安全风险在线评估的评述[J]. 电力系统自动化, 2012, 36(18): 171-177.
LI Bijun, FANG Yongjie, XU Taishan. Review on online operational security risk assessment of power system[J]. Automation of Electric Power Systems, 2012, 36(18): 171-177.
- [5] 赵会茹,李娜娜,郭森,等. 配电网设备故障停电风险实时评估[J]. 电力自动化设备, 2014, 34(11): 89-94.
ZHAO Huiru, LI Nana, GUO Sen, et al. Real-time risk assessment on equipment failure outage of distribution network[J]. Electric Power Automation Equipment, 2014, 34(11): 89-94.
- [6] 刘健,韩磊,张志华. 面向用户并考虑紧迫性的配电网运行风险评估[J]. 电力自动化设备, 2015, 35(2): 97-102.
LIU Jian, HAN Lei, ZHANG Zhihua. Customer-oriented distribution network operational risk assessment considering urgency[J]. Electric Power Automation Equipment, 2015, 35(2): 97-102.
- [7] 李锐,陈颖,梅生伟,等. 基于停电风险评估的城市配电网应急预警方法[J]. 电力系统自动化, 2010, 34(16): 19-23.
LI Rui, CHEN Ying, MEI Shengwei, et al. An early warning method for emergency response based on power failure risk analysis of distribution systems[J]. Automation of Electric Power Systems, 2010, 34(16): 19-23.
- [8] 俞隼亚,王增平,田红雨,等. 基于最小过热区域的城市配电网灾害风险评估[J]. 电力系统保护与控制, 2015, 43(4): 115-120.
YU Junya, WANG Zengping, TIAN Hongyu, et al. Risk assessment of urban distribution network under disaster based on minimum over-heated region[J]. Power System Protection and Control, 2015, 43(4): 115-120.
- [9] 刘道伟,张东霞,孙华东,等. 时空大数据环境下的大电网稳定态势量化评估与自适应防控体系构建[J]. 中国电机工程学报, 2015, 35(2): 268-276.
LIU Daowei, ZHANG Dongxia, SUN Huadong, et al. Construction of stability situation quantitative assessment and adaptive control system for large-scale power grid in the spatio-temporal big data environment[J]. Proceedings of the CSEE, 2015, 35(2): 268-276.
- [10] 赵庆周,李勇,田世明,等. 基于智能配电网大数据分析的状态监测与故障处理方法[J]. 电网技术, 2016, 40(3): 774-780.
ZHAO Qingzhou, LI Yong, TIAN Shiming, et al. Analysis of the data of state monitoring and fault processing method based on intelligent distribution system[J]. Power System Technology, 2016, 40(3): 774-780.
- [11] 汤昶烽,卫志农,李志杰,等. 基于因子分析和支持向量机的电网故障风险评估[J]. 电网技术, 2013, 37(4): 1039-1044.
TANG Changfeng, WEI Zhinong, LI Zhijie, et al. Risk assessment of power grid failure based on factor analysis and support vector machine[J]. Power System Technology, 2013, 37(4): 1039-1044.
- [12] 翟学锋,卫志农,范立新,等. 基于相关向量机的发电机进相能力建模[J]. 电力自动化设备, 2015, 35(3): 146-151.
ZHAI Xuefeng, WEI Zhinong, FAN Lixin, et al. Generator leading phase capability model based on relevance vector machine[J]. Electric Power Automation Equipment, 2015, 35(3): 146-151.
- [13] 张友强,寇凌峰,盛万兴,等. 配电变压器运行状态评估的大数据分析方法[J]. 电网技术, 2016, 40(3): 768-773.
ZHANG Youqiang, KOU Lingfeng, SHENG Wanxing, et al. Big data analytical method for operating state assessment of distribution transformer[J]. Power System Technology, 2016, 40(3): 768-773.
- [14] KIRA K, RENDELL L. A practical approach to feature selection [C]//Proceedings of the 9th International Workshop on Machine Learning. San Francisco, CA, USA: ACM, 1992: 249-256.
- [15] 王东,史晓霞,尹交英. 不同核函数的支持向量机用于空调负荷预测的对比研究[J]. 电工技术学报, 2015, 30(增刊1): 531-535.

WANG Dong, SHI Xiaoxia, YIN Jiaoying. Prediction on hourly load of air conditioning by RBF support vector machine[J]. Transactions of China Electrotechnical Society, 2015, 30 (Supplement 1): 531-535.

- [16] 周淦, 廖婧舒, 廖瑞金, 等. 含分布式电源的配电网停电风险快速评估[J]. 电网技术, 2014, 38(4): 882-887.

ZHOU Quan, LIAO Jingshu, LIAO Ruijin, et al. Rapid assessment of power system blackout risk with distributed generation[J]. Power System Technology, 2014, 38(4): 882-887.

- [17] 叶圣永. 基于机器学习的电力系统暂态稳定评估研究[D]. 成都: 西南交通大学, 2010.

YE Shengyong. Study on power systems transient stability assessment based on machine learning method [D]. Chengdu: Southwest Jiaotong University, 2010.

作者简介:



刘科研

刘科研(1978—),男,河南郑州人,高级工程师,博士,研究方向为配电网建模仿真与智能分析(**E-mail**: liukeyan@epri.sgcc.com.cn);

吴心忠(1990—),男,山东济宁人,硕士研究生,研究方向为配电网可靠性评估、优化运行(**E-mail**: 15121489@bjtu.edu.cn);

石琛(1994—),男,山西朔州人,硕士研究生,研究方向为配电网优化运行(**E-mail**: 13291142@bjtu.edu.cn)。

Fault risk early warning of distribution network based on data mining

LIU Keyan¹, WU Xinzhong², SHI Chen², JIA Dongli¹

(1. China Electric Power Research Institute, Beijing 100192, China;

2. School of Electrical Engineering, Beijing Jiaotong University, Beijing 100044, China)

Abstract: The fault correlation factor analysis and risk early warning method of distribution network are proposed based on the data mining to improve the accuracy of the warning. Four kinds of general categories including 28 fault features of distribution network are summed up through data cleaning, data transformation, data integration and outlier sample elimination. The fault correlation factors are analyzed by the improved Relief-Wrapper algorithm, which eliminates 6 redundant features to form the optimal fault feature subset composed of the rest 22 fault features. The fault risk indexes and risk level classification method for distribution network are proposed considering the frequency of failure and the proportion of electricity load loss, and the risks are warned by the radial basis function based SVM(Support Vector Machine) and the optimal fault feature subset. The risk early warning of a distribution network with 120 feeders is analyzed, and the results verify the effectiveness of the proposed method.

Key words: distribution network; data mining; fault correlation factors; optimal fault feature subset; risk early warning; support vector machines; risk index