

融合改进SMOTE与Stacking算法的输电通道树线放电因子风险状态评估

操松元¹, 严波¹, 季坤¹, 夏令志², 方登洲², 程洋², 王康³, 卫鹏³

(1. 国网安徽省电力有限公司, 安徽 合肥 230022; 2. 安徽省电力科学研究院, 安徽 合肥 230601; 3. 杭州数梦工场科技有限公司, 浙江 杭州 310000)

摘要: 当前, 对电力系统输电通道的风险评估手段以规则方法为主, 往往对输电线路结构本身的特征考虑较少。基于KNN、决策树和SVM模型, 提出了一种融合改进SMOTE与Stacking的机器学习算法, 实现了对输电通道中树线放电因子的风险评估预测, 该算法能较好地解决数据集没有明确正常样本数据, 且异常样本类型分布不平衡的问题。通过与Bagging集成算法、融合SMOTE与Stacking算法对比, 验证了所提算法的有效性与先进性, 为后续输电通道其他因子基于机器学习的风险评估奠定了基础。

关键词: 机器学习; SMOTE; Stacking; 输电通道; 风险评估

Assessment of the risk status of tree-line discharge factor in transmission channels using the modified SMOTE and Stacking algorithm

CAO Songyuan¹, YAN Bo¹, JI Kun¹, XIA Lingzhi², FANG Dengzhou², CHENG Yang², WANG Kang³, WEI Peng³

(1. State Grid Anhui Electric Power Co., Ltd., Hefei 230022, China; 2. Anhui Electric Power Research Institute, Hefei 230601, China; 3. Hangzhou Dtdream Technologies Co., Ltd., Hangzhou 310000, China)

Abstract: At present, the rule is the main method of risk assessment for power transmission channels, and the characteristics of transmission line structure are seldom considered. Based on KNN, decision tree and SVM models, a fusing improved SMOTE with the Stacking machine learning algorithm is proposed, it implements the tree line in the transmission channel discharge factor to the risk assessment of forecast, this algorithm can well solve there is no clear normal sample data in data sets and abnormal samples type distribution imbalance issue as well as improve the problem of low classification accuracy of traditional classification algorithms. By comparing with Bagging ensemble algorithm, SMOTE and Stacking algorithm, the effectiveness and advance of the proposed algorithm are verified, which lays a foundation for machine learning based risk assessment of other factors in the subsequent transmission channel.

Key words: machine learning; SMOTE; Stacking; transmission channel; risk assessment

0 引言

架空输电通道是电力系统的重要组成部分, 随着社会的高速发展, 社会用电量也在持续增长, 而用电量的快速增加也对输电线路的

安全运行提出了更高的要求, 同时也对输电线路的检修和诊断带来了更多挑战, 因此对输电通道进行风险评估显得尤为重要。近几年先后出现了基于气象参数^[1]的方法, 即依据传统

的输电线路覆冰增长及风偏角计算模型,实现输电线路风险的评估与预警;基于统计学^[2]的方法,即基于大气质量指数和微气象站信息,然而以上两种评估手段对影响因素的考虑仍较为单一,且对风险状态的划分不够详细;此外,随着机器视觉技术的发展,基于图像、红外视频等^[3-8]的评估手段也都加以应用,但这些评估方式都属于事后评估,即无法通过获取前期对输电通道的杆塔运行状态进而预测高风险的发生。所以目前主流的方法还是采用规则评估,即严重程度、发生频率和后果可能性(LEC)评估方法^[8],但由于输电通道影响因素多且相互之间的影响规律复杂,LEC等规则评估方法在实践环节也出现了以下三点问题:

①L、E、C值的选择主要依靠经验判断,可能存在较大误差;②风险因子相关影响分析不够准确,基于HIRA(Hazard Identification and Risk Assessment)方法对技术要素风险程度评估的区分度不够;③风险要素评估中对输电线路结构本身特性的考虑很少,即没有充分体现不同灾害类型对输电线路各个组成单元的量化影响规律。基于此,本文引入了机器学习算法,解决现有规则评估方法对输电线路结构本身特性缺乏研究的问题。

1 杆塔风险状态评估

在国家电网公司的指导性文件Q/GDW 11450—2015《重要输电通道风险评估导则》中,对线路风险等级的划分为:正常、低风险、中风险和高风险四个等级。其中正常状态为不会引发通道内线路同时高风险,设备处于稳定且良好的运行状态;低风险状态为短期内不会导致通道内线路同时高风险,但应限制其继续发展,加强运行监视,安排治理计划;中风险状态为短期内导致通道内线路同时高风险,需要在近期内安排措施给予管控与治理;高风险状态为极易引发通道内线路同时高风险,需要在近期内安排措施给予管控与治理。本文通过对历史杆塔风险状态的台账数据的积累,设计机器学习算法对未来杆

塔的风险状态进行评估。

2 样本不平衡问题

输电线路中杆塔的运行通常相对稳定,杆塔运行台账数据往往只提供出现异常风险状态时的标记信息,而对于正常状态时的标记信息则没有提供支持。异常数据往往又细分为多个子类异常,而子类异常数据样本经常存在分布不平衡的问题。为减少在使用机器学习进行建模时训练样本分布失衡对模型性能的影响,需要对样本分布进行平衡化处理;同时为了能使模型不仅能够对异常类型作出预测,也能对正常类型作出预测,需要构造相应的带有标记信息的正常样本数据。

常用解决样本分布失衡的方法是随机采样,包括欠采样和过采样,其中以SMOTE算法最为常用,该算法由Chawla Kevin等人提出,能够合成少数类样本^[11]。SMOTE算法为类别中的每个样本找到少数类中最近的 n 个邻居节点,然后这个样本在相邻的节点之间画一条线,并在这条线上产生随机的节点。如图1所示,少数类样本会找到离其最近的5个邻居节点 $x_2, x_3, \dots, x_6$ 。然后给每一个邻居节点画一条线,并分别在每条线上产生红色标注的随机点 $a \sim e$ 作为新得到的合成样本,因此SMOTE是一种过采样算法。除此之外,还有基于SMOTE衍生而来的ADASYN算法。

可以看到,SMOTE算法仅能对已有标记

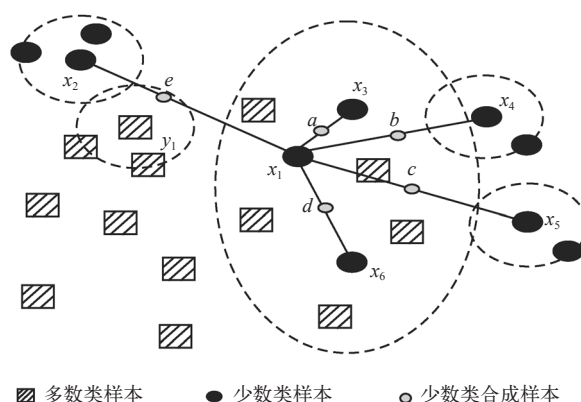


图1 SMOTE算法结构

Fig. 1 SMOTE algorithm structure

信息的样本进行采样,而对未标记信息的样本则不能很好地进行采样。但在实际场景中,对模型的要求是对所有可能出现的情况做出预测。因此,对于树线放电因子,不仅需要要对的数据进行SMOTE采样以达到异常分布的平衡,还需要构造正常样本,保证树线放电因子模型具备对正常状态的预测能力,最终使模型能达到对树线放电因子正常、低风险、中风险和高风险这四种状态事件作出预测。因此设计基于SMOTE的改进算法。

算法1 SMOTE 改进算法

```

Input:      Dataset D
           = {(x1(1), ytype1), (x2(1), ytype1), ...,
             (xm(m), ytype_m), (xn), (xn+1), (xk)}
```

Process:

1. Step1 While l' < m//2:
2. i = randint(1, m)
3. pos_i = (x_{i(-i)}, y_{type_positive})
4. l += l
5. Dl = (pos₁, pos₂, ..., pos_m)
6. end
7. Step2 While l'' < (k-n)//2:
8. i = randint(n, k)
9. pos_i = (x_{i(-i)}, y_{type_positive})
10. l += l
11. D2 = (pos₁, pos₁, ..., pos_m)
12. end
13. Step3 for type = type₁, type₂, ..., type_m:
14. = SMOTE(k, n, (x(·), y_{type}))
15. D3 = (D_{type1}, D_{type2}, ..., D_{typem})
16. end

Output: D' = (D₁, D₂, D₃)

3 融合改进SMOTE与Stacking算法

为防止模型过拟合情况,提高模型的鲁棒性,主流方法是采用基于Bagging或基于Boosting的模型融合手段。Bagging算法更关注于方差的减小^[10],代表是随机森林。Bagging算法通过对数据集进行有放回随机抽样进而构造L组训练样本,即

$$\{z_1^1, z_2^1, \dots, z_B^1\}, \{z_1^2, z_2^2, \dots, z_B^2\}, \dots, \{z_1^L, z_2^L, \dots, z_B^L\} \quad (1)$$

并基于L组训练样本训练多个弱学习器,即

$$w_1, w_2, \dots, w_L \quad (2)$$

最后使用投票或平均方式对所有弱学习器的预测结果进行汇总后输出预测结果,即

$$S_L = \frac{1}{L} \sum_{l=1}^L w_l \quad (3)$$

$$S_L = \arg \max[\text{card}(l | w_l) = k] \quad (4)$$

与Bagging算法同时训练多个弱学习器不同的是,Boosting算法先从初始训练集上训练一个基学习器,再根据此基学习器的表现对原始训练样本分布进行调整,使得先前基学习器做错的训练样本在后续受到更多关注,然后基于调整后的样本分布来训练下一个基学习器,如此反复进行,直到学习L个基学习器,最终将这L个基学习器进行加权输出预测结果

$$S_L = \sum_{l=1}^L c_l w_l \quad (5)$$

$$S_L = S_{L-1} - c_l \times \nabla s_{l-1} E(s_{l-1}) \quad (6)$$

式中, $E(\cdot)$ 为给定模型的拟合误差; c_l 为步长系数。

Boosting算法的代表是GBDT、Adaboost等,更关注于偏差的减小。

然而上述两种模型融合手段都不可避免出现对偏差-方差重视程度失衡的情况。鉴于此,提出一种基于Stacking的融合算法,将训练好的所有基模型对整个训练集进行预测。与Bagging或Boosting不同的是,Stacking并不是由多棵决策树训练得到,而是通过构建多个不同类别的分类或回归模型作为基模型,将每个基模型的预测输出汇总,形成新的训练集,再基于新形成的训练集构造相应元模型,训练参数,所以Stacking模型融合算法通常是异构的。下面的伪代码总结了Stacking算法的实现过程。

算法2 Stacking 算法

```

Input:      Dataset D
           = {(x1, y1), (x2, y2), ..., (xm, ym)}
```

基学习器 $L_1, \dots, L_T$;
元学习器 L .

Process:

1. Step1 for $t = 1, \dots, T$:
 2. $ht = Lt(D)$;
 3. end
 4. Step2 $D = \varphi$;
 5. for $i = 1, \dots, m$:
 6. for $t = 1, \dots, T$:
 7. $z_{it} = h_t(x_i)$;
 8. end
 9. $D = D \cup ((z_{i1}, \dots, z_{iT}), y_i)$
 10. end
 11. Step3 $h = L(D)$;
- Output: $H(x) = h(h_1(x), \dots, h_T(x))$

4 实验结果与分析

本文模型设计流程如图2所示。

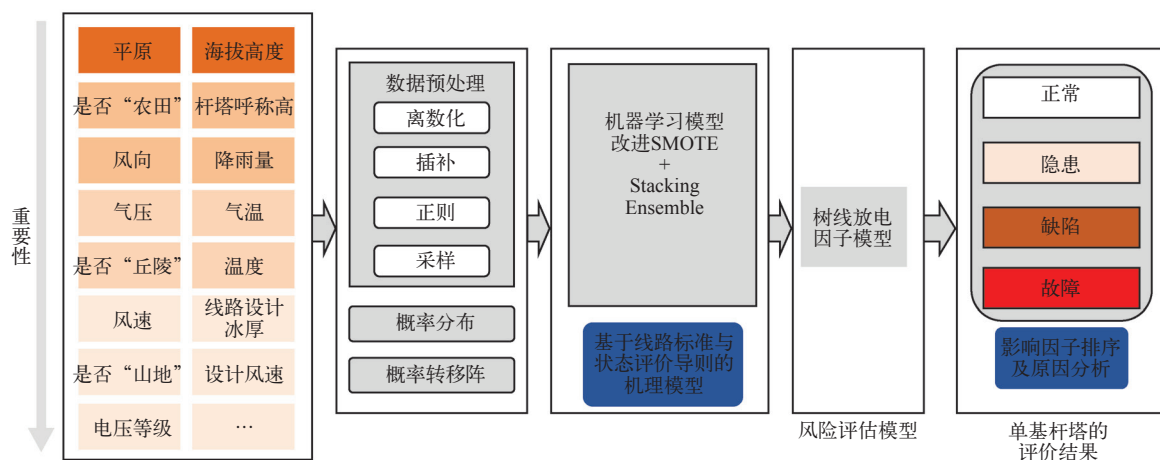


图2 树线放电因子模型训练架构

Fig. 2 Tree line discharge factor model training framework

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (8)$$

式中，TP为真正例；FP为假正例；FN为假负例。

表1 输电通道杆塔台账数据描述

Tab. 1 Description of transfer channel equipment ledger data

数据集	样本数	输入属性数	类别数
杆塔台账数	17 570	20	3

此外，由于F1指标同时衡量了查准率与查全率，所以在分类模型中常使用F1指标作为模型综合效果评估指标。但由于研究的树线放电模型是多分类模型，而F1指标通常是对一个分类效果好坏的度量，所以为了对模型作

4.1 数据集与性能指标

本节主要研究融合改进SMOTE与Stacking算法在输电通道树线放电因子中的应用情况。原始训练数据来自安徽电网输电通道树线放电因子属性数据集（如表1所示），对原始训练数据进行预处理及进行特征工程后的异常数据共有17 570条实例，包含20个输入特征属性以及树线放电低风险、树线放电中风险和树线放电高风险三个类别的标签属性，且三个类别的标签属性数据量分布分别为150:7640:9780。

为了验证本算法的性能，选取以下性能指标

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (7)$$

整体评价，提出一种以每一分类对应样本数量为权重，进行F1指标加权求平均的度量指标对模型整体性能评估，计算公式表示为

$$\text{weight_F} = \frac{\sum_{i=1}^n S_i F1_i}{\sum_{i=1}^n S_i} \quad (9)$$

式中， n 为分类个数； S_i 为第 i 个类别样本数量； $F1_i$ 为第 i 个类别F1指标；weight_F为模型整体评估指标。

4.2 试验设计

由于Stacking算法中用于训练基学习器的

数据同时也会生成元学习器的训练集, 所以容易出现模型过拟合的风险, 因此设计将用于生成元学习器的数据从训练基学习器的数据集 中进行分离, 再使用交叉验证对试验结果进行分析。

具体地, 本文设计采用10折交叉验证, 即原始训练数据集 D 被随机切分成10个几乎相等的部分 $D_1, \dots, D_{10}$, 此外定义 D_j 和 $D_{(-j)} = D/D_j$ 分别为第 j 折的测试集和训练集。在给定 T 学习算法的情况下, 通过输入训练集 $D_{(-j)}$, 完成对 t 个基学习器的训练, 且此时训练好的基学习器记为 $h_t^{(-j)}$ 。对于 D_j 中的每个测试样本 x_i , 即第 j 折的测试集中的样本, 令 z_{it} 表示学习器 $h_t^{(-j)}$ 对测试样本 x_i 的输出。

做完10折交叉验证后, 经过所有基学习器的测试样本的输出 D' 为

$$D' = \{(z_{i1}, \dots, z_{iT}, y_i)\}_{i=1}^m \quad (10)$$

在数据集 D' 的基础上应用元学习算法, 得到的学习器 h' 是关于 $(z_1, \dots, z_T)$ 到输出 y 的模型。在生成新的数据集 D' 之后, 通常通过对整个训练数据的训练来重新生成最终的元学习器模型。

4.3 试验结果

首先使用改进SMOTE算法解决原始训练集缺少正常样本以及异常样本分布不平衡的问题, 得到的类别标签属性信息如表2所示。

表2 改进SMOTE算法修正后的样本分布情况

Tab. 2 Improved sample distribution using the modified

SMOTE

类别标签	原始样本数量	改进SMOTE算法
树线放电正常	0	8 000
树线放电低风险	150	8 150
树线放电中风险	7 640	7 640
树线放电高风险	9 780	9 780

然后使用Stacking算法对处理后的训练集进行模型训练, 这里采用的基学习器分别由KNN、Decision Tree和SVM不同类型的学习

器组成。

为了完整地对本 文所提算法性能作出衡量, 设计三组对照试验, 分别以融合SMOTE与Stacking算法、Stacking算法以及Bagging算法与本文所提出的融合改进SMOTE与Stacking算法作对比, 分别从学习曲线和评估指标两方面分析输电通道树线放电因子风险状态的预测结果。其中, 学习曲线的横坐标表示训练样本的数量, 纵坐标表示模型分类得分weight_F指标。

绘制模型学习曲线观测模型的拟合程度, 如图3~图6所示。其中, 蓝色曲线表示模型在训练集上的学习曲线, 红色曲线表示模型在交叉验证集上的学习曲线。可以观察到, 在平滑性上, 与其他三种算法相比, 使用了融合改进SMOTE与Stacking算法的学习曲线平滑性最好。

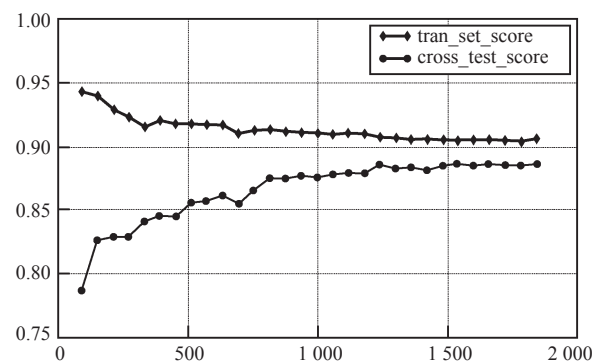


图3 融合改进SMOTE与Stacking算法的学习曲线

Fig. 3 Learning curve of fusing improved SMOTE and Stacking algorithm

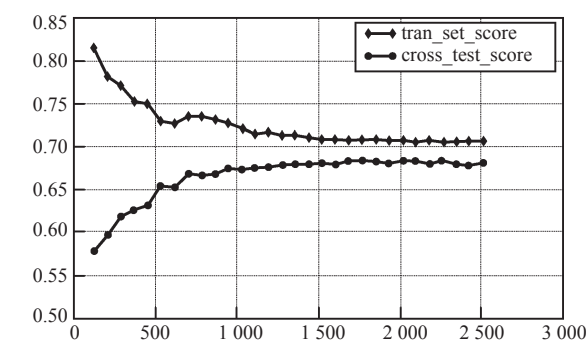


图4 融合SMOTE与Stacking算法的学习曲线

Fig. 4 Learning curves of fusing SMOTE and Stacking algorithm

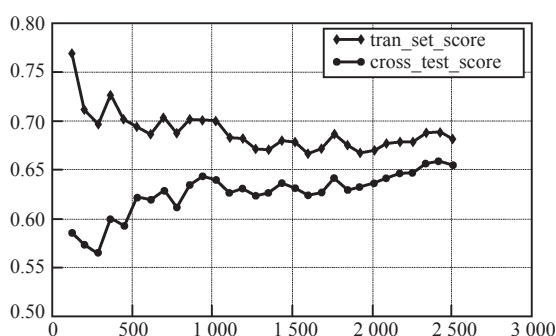


图5 Bagging算法的学习曲线

Fig. 5 Learning curves of Bagging algorithm

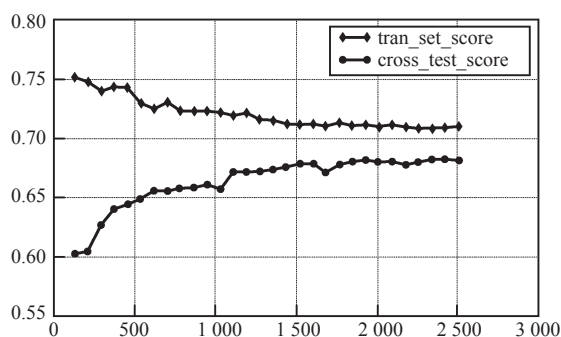


图6 Stacking算法的学习曲线

Fig. 6 Learning curves of Stacking algorithm

在准确率上,融合改进SMOTE与Stacking算法的得分最高,虽然融合SMOTE与Stacking算法的学习曲线整体观测效果也不错,但由于算法缺乏对正常样本的预测能力,导致与本文所提算法在综合准确率方面还是略低一些。

由表3可以看到,本文所提算法不仅优于未考虑正常样本数据的融合SMOTE与Bagging算法和未作异常样本分布不平衡处理的Stacking算法,而且与未考虑正常样本数据的融合SMOTE与Stacking算法相比,在Precision、Recall以及Weight_F这三个指标上都有较大程度提升。

5 结束语

本文以输电线路中树线放电因子为建模对象,使用融合改进SMOTE与Stacking算法对其风险状态进行预测建模,避免了原有评估体系过于依赖经验,实现了基于动态数据分析

表3 基于交叉验证的不同算法性能指标对比

Tab. 3 Comparison of performance indexes of different algorithms

数据集	Precision	Recall	Weight_F
融合改进SMOTE与Stacking算法	0.89	0.86	0.88
融合SMOTE与Stacking算法	0.75	0.83	0.79
融合SMOTE与Bagging算法	0.67	0.67	0.67
Stacking Ensemble算法	0.69	0.72	0.71

的细粒度(时间、空间)输电通道风险评估智能化,同时又保证了模型具有高可解释性。此外,通过与其他主流机器学习算法做对照试验,依次从学习曲线、Precision、Recall以及Weight_F这四个指标说明本文所提算法的先进性。

当然,本文的研究成果还存在很大的改善空间,首先是受制于数据量级,模型在海量样本下的性能表现并没有进行充分评估。其次是模型性能方面,虽然通过绘制学习曲线观察到模型泛化性能较好,但测试集上,模型的Recall指标并不是很高。后续还需要持续地沉淀业务数据,并进行有效的数据治理工作,在特征工程和算法优化上进一步完善和提高,以提高输电通道风险评估的准确性和商业价值。

参考文献

- [1] 王海涛,冯万兴,陶汉涛.基于气象参数的输电线路电气可靠性实时评估与预警系统设计与研发[J].中国电力,2018,51(5):17-23.
- [2] 王忠岳,林济铿,胡世骏.输电线路污闪跳闸概率计算方法[J].中国电力,2016,49(5):53-58.
- [3] 张文峰,彭向阳,陈锐民.基于无人机红外视频的输电线路发热隐患智能诊断技术[J].电网技术,2014,38(5):1334-1338.
- [4] 高军,文善义.架空输电线路跳闸故障智能诊断[J].低碳世界,2017,24(28):35-36.
- [5] 仝伟,仝娜,宋宁宁.智能电网建设中输电线路部分关键技术探讨[J].安徽电力,2012,34(2):

- 78-82.
- [6] 刘晓伟, 涂洁, 张雪峰. 输电线路全状态评价模型与方法研究[J]. 电测与仪表, 2015, 52(12): 113-117.
- [7] 刘家铭. 基于地线取电的高压线路在线监测系统的设计与实现[D]. 哈尔滨: 哈尔滨工业大学, 2018.
- [8] 杨勐峤. 高压输电线路在线监测系统的设计与研究[D]. 北京: 北京交通大学, 2011.
- [9] 吴濡生, 操松元, 季坤. LEC法在重要输电通道风险评估中的应用研究[J]. 电力安全技术, 2016, 28(10): 45-48.
- [10] 周志华. 机器学习[M]. 北京: 清华大学出版社, 2016.

- [11] CHAWLA N V, BOWYER K W, HALL L O, et al. SMOTE: synthetic minority over-sampling technique[J]. Journal of Artificial Intelligence Research, 2011, 16(1): 321-357.

作者简介



操松元

操松元 (1978—), 男, 博士, 高级工程师, 主要研究方向为电气系统及自动化。

(收稿日期: 2019-08-16) **EA**

广告 | AD

全新改版

色彩丰富 凸显主题

2020《电气应用》以鲜明色彩和新颖版式呈现精彩内容, 力求赏心悦目。

栏目设置 独具风格

坚持继承与开拓并举, 以丰富内涵充实经典栏目, 持续深化专家访谈、EA经理人等栏目, 密切关注电气技术的发展和运用。

专家点评 精辟深刻

每期都将围绕一个热点话题, 隆重推出一场专题讨论的盛宴, 共邀知名专家、产品用户和生产厂家汇聚一堂, 针对焦点问题作出评析。

行业案例 生动丰富

关注电力、建筑、石化和冶金等电气行业应用, 以典型深刻的案例分析、方案设计构建技术互动的空间, 促进应用经验交流。

国内邮发代号: 82-341 国外发行代号: M1409

每期定价: 25.00元 全年定价: 300.00元



· 电气应用 盛情呈现 ·

2020