

基于 GBDT 的电力计量设备故障预测

刘金硕¹ 刘必为² 张密³ 刘卿⁴

(武汉大学国家网络安全学院 武汉 430070)¹ (武汉大学计算机学院 武汉 430070)²

(中国电力科学研究院 北京 100089)³ (天津电力科学研究院 天津 300041)⁴

摘要 电力计量设备的故障风险预测可以减少国家电网因为故障风险带来的损失。文中首先进行了数据的预处理和特征选取;其次,设计了基于 GBDT 的故障大类、故障小类以及设备寿命周期的预测;最后,对设计的模型进行了有效性和先进性的验证。实验在中国电力科研院提供的数据上进行。由实验结果可知,所提算法对 6 种故障类型的预测准确率为 90.56%,查全率为 92.95%,F1 值为 91.71%。相比回归、BP 神经网络、Adaboost、决策树算法,梯度提升决策树算法在参数调优条件下的性能最优。

关键词 GBDT, 计量风险预测, 数据清洗

中图分类号 TP206+.3 **文献标识码** A

Fault Prediction of Power Metering Equipment Based on GBDT

LIU Jin-shuo¹ LIU Bi-wei² ZHANG Mi³ LIU Qing⁴

(School of Cyber Science and Engineering, Wuhan University, Wuhan 430070, China)¹

(School of Computer Science, Wuhan University, Wuhan 430070, China)²

(China Electric Power Research Institute, Beijing 100089, China)³

(Electric Power Science & Research Institute of Tianjin Electric Power Company, Tianjin 300041, China)⁴

Abstract The fault risk prediction of power metering equipment can reduce the loss caused by the fault risk of the national grid. Firstly, the data preprocessing and feature selection are carried out. Secondly, the GBDT-based fault categories, fault subclasses and equipment life cycle prediction are designed. Finally, the validity and advancement of the designed model are verified. Data used in the experiment are provided by China Electric Power Research Institute. The experimental results show that the prediction accuracy of the six fault types by using the proposed algorithm is 90.56%, the recall rate is 92.95%, and the F1 value is 91.71%. Compared with regression, BP neural network, Adaboost and decision tree algorithm, the gradient lifting decision tree algorithm has the best performance under parameter tuning conditions.

Keywords GBDT, Measurement risk prediction, Data cleaning

国家电网公司于 2010 年开始启动用电信息采集系统的建设。目前,采集覆盖率在全国电网系统达到 92.2%,已安装智能电表 3.98 亿只。在 7 年左右的运行中,累积了计量设备的大量数据。利用营销信息化系统的数据资源,多维度挖掘计量装置的故障分布规律及影响因素,按照时间、供电区域、供应商、设备类别等维度统计公司每个生产环节的故障情况、故障率,探索故障高发的原因,对故障之间的关联关系、派生关系和因果关系进行分析,挖掘故障之间的潜在关联,对计量设备潜在的运行风险类别进行有效预测。

另外,营销计量数据具有典型的数据体量巨大、类型繁多、价值密度低和复杂的特点。从业务的角度,对数据进行数据清洗和规约转化等,可以提高数据精度,进而提高数据挖掘的分析速度。对营销信息化系统数据进行预处理,包括数据重采样、变量选择、数据替换、过滤异常、变量变换、数据分割等。

目前,国内外对计量数据尤其是负荷数据做了大量的分

析工作^[1-4]。Zakaria 等^[5]采用了 FCM 算法对负荷曲线进行聚类,但是该算法存在一些不足,如对初始聚类中心敏感、聚类结果容易受到孤立点的影响等。为了提高聚类结果的准确性,一些学者针对常用算法提出了相关改进。周开乐、杨善林^[6]提出了基于模拟退火遗传算法的模糊 C 均值聚类,该算法提高了负荷分类的有效性和精确性。刘永光等^[7]提出了一种改进的模糊 C 均值聚类算法对日负荷曲线进行分类,该算法减小了聚类中人为输入聚类数和初始聚类中心带来的影响。董瑞、黄民翔^[8]为了避免 FCM 法陷入局部最优解,采用了减法聚类找到初始聚类中心,然后将这种改进的 FCM 算法应用在了电力负荷分类上。从上述这些改进的算法中可知,聚类算法的改进主要从聚类初始中心选择和聚类数确定这两个方面进行。在负荷曲线聚类中需要解决两个问题:1)聚类算法的选择;2)聚类数的确定。针对聚类算法选择问题,Bidoki 等^[9]采用了 K-means 算法、WAF (Weighted Fuzzy Average) K-means 算法、层次聚类算法和 SOM 算法对负荷曲

线进行聚类,并通过聚类有效性评价指标 CID 和 MIA 对其进行评价,结果表明 k-means 算法比其他算法好。

针对上述问题,本文对电力计量设备运行数据进行了预处理和特征选取,并构建了 GBDT^[10-11]的故障风险预测模型,该模型可以对计量设备的故障大类、故障小类以及设备寿命周期进行预测,进而减少国家电网因为故障风险带来的损失。最后,本文对设计的模型进行了有效性和先进性的验证。

1 相关理论

梯度提升(Gradient Boosting)是一种用于回归、分类和排序任务的机器学习技术,属于 Boosting 算法族的一部分。Boosting 是一族可将弱学习器提升为强学习器的算法,属于集成学习(Ensemble Learning)的范畴。Boosting 方法基于这样一种思想:对于一个复杂任务来说,将多个专家的判断进行适当的综合所得出的判断,比其中任何一个专家单独的判断要好。通俗地说,就是“三个臭皮匠顶个诸葛亮”的道理。梯度提升同其他 boosting 方法一样,通过集成(Ensemble)多个弱学习器(通常是决策树),来构建最终的预测模型。

梯度提升决策树(Gradient Boosting Decision Tree,GBDT)是基于梯度提升技巧的一种集成学习算法,与随机森林算法类似,但属于不同的决策树模型组合方式。GBDT 由多棵决策树构成,输出为每棵决策树输出结果的累加,利用梯度提升和回归决策树的组合方式,每次建立新的决策树模型都是在之前模型损失函数的梯度的下降方向,使得决策模型不断改进。

2 故障预测模型数据构建

2.1 数据预处理

因为人员、设备等一些外部原因,从现场计量设备实际应用场景中提取的历史数据,有时会存在一个或多个字段产生的数据项的缺失,有时是信息量更大的字段出错,需要对数据进行增强处理的过程。

数据预处理系统是由原始数据、数据采样、变量选择、过滤异常、数据替换、变量变换、数据分割和可用数据集等组成。数据预处理的流程如图 1 所示。

1)数据采样:原始数据的数据量过大,不利于后续的处理训练,需要提取部分数据并保留原始数据特征,同时后续在数据量相对较小的数据集上进行数据挖掘可以极大地提高数据挖掘的效率。

2)变量选择:通过丢弃与挖掘任务无关的变量,或者丢弃包含大量丢值的变量,来减小数据挖掘算法输入变量的数目,降低数据空间的维数,加快数据挖掘算法的执行速度。

3)数据替换:通过用区间变量的均值、中值、区间中值等替换区间变量的丢值,用最常出现的分类变量值替换分类变量的丢值,能够保留数据记录中其他有意义的变量,解决由于变量丢值所造成的数据量不足的问题。

4)过滤异常:过滤掉数据采集时未被发现的局外样本,消除数据噪声对数据挖掘模型的干扰,可以产生精度更高的数据挖掘模型。

5)变量替换:现有数据集中的有些变量信息无法被模型直接使用或无法很好地满足要求,需要进行变换或添加新变量来生成新数据集。

6)数据分割:为了构造高精度的数据挖掘模型,将原始数据集分割为训练集、验证集和测试集。训练集用于数据挖掘模型的构造;验证集用于数据挖掘模型评估,也可用于数据挖掘模型修正;测试集用于测试得到的数据挖掘模型的精度。

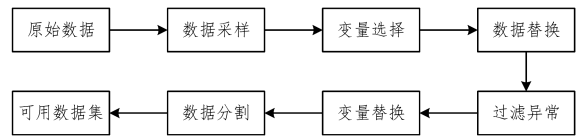


图 1 数据预处理流程图

2.2 特征选取

2.2.1 原始数据抽取特征

为了提高预测模型的运算速度和优化建模方法,本文采用了如下的特征变换方法:

闲置时间 = 安装时间 - 到货时间

寿命 = 拆除时间 - 安装时间

将故障小类及产权单位标签化

最终,经过上述数据预处理过程后,选取的特征如下(见表 2):厂商编号、到货招标批次号、故障大类、故障小类、闲置时间、寿命、产权单位。

表 1 提取后特征数据示例

厂商编号	到货批次	故障类别	闲置时间	寿命	产权单位
22047	26150423481497	485 接口损坏	278	727	5963942
22047	26150423481497	参数设置不正确	279	727	5963944
22047	26150423481497	时钟错误	279	727	5963944
22047	26150423481497	火灾	215	794	5963942
22047	26150423481497	运行中外力损毁	170	775	5963949
22047	26150423481497	过负荷	212	733	5963942
22047	26150423481497	运行中外力损毁	273	636	59639
22047	26150423481497	计量性能	316	602	596394944

2.2.2 气候特征

考虑到电表内部电子元器件的使用寿命会受到天气的影响,例如在多雨潮湿的气候条件下电子元器件的寿命会缩短。根据设备所属产权单位所属的地理位置,在天气网中爬取了详细的天气信息。例如:月气温走势图、风力、每日气温的最值、一年中的气温最值、天气情况以及风力形成新的特征项。同时,考虑到提取所有设备的气候信息会出现很多重复的信息,以及同时在一个市的范围内气候差异极小,因此爬取气候信息的最小范围为市。

3 预测模型构建

在计量设备故障风险预测的学习任务中,预测目标有 3 种:故障大类、故障小类以及设备寿命。将整个模型分为 3 个子模型,利用梯度提升决策树分步预测 3 个目标值,同时由于 3 个目标值之间分别具有相关性,因此每个步骤完成后均将得到的目标值作为下一步骤的输入值。本文采用先预测故障大类,然后将故障大类作为中间结果来进一步预测故障原因小类的方案,通过对闲置时间、产权单位、故障大类进行建模分析,找出当前招标批次电表的闲置时间、产权单位与故障大类之间的潜在关系。具体模型的流程图如图 2 所示。

GBDT 故障大类预测模型、故障小类预测模型、寿命预测模型的参数性能如表 2 所列,其中 44.57 天代表预测值与实际值的平均误差。

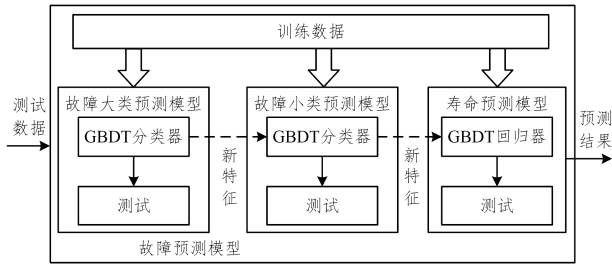


图2 GBDT预测模型

表2 模型参数性能

名称	所选模型	n_estimators	learning_rate	max_depth	准确度(误差)
故障原因大类预测模型	GBDT分类器	200	0.1	4	95.03/%
故障原因小类预测模型	GBDT分类器	250	0.2	3	97.13/%
寿命预测模型	GBDT回归器	150	0.25	4	44.57天

首先以电表的闲置时间、产权单位为输入,利用故障大类预测模型得到概率最大的两个故障大类,然后将两个故障大类分别加入输入特征,利用故障小类预测模型得到概率最大的两个故障小类,将大类与小类的概率相乘即可得到4个概率最大的故障大类与小类的组合,最后利用寿命预测模型分别预测各组合的电表寿命,即可得到模型的最终输出。以上述编号为2的样本为例,真实值为故障大类4,故障小类41103,实际寿命794,而组合模型的输出如表3所列。

表3 模型输出结果

输出编号	故障大类	故障小类	概率/%	寿命/天
1	4	485 接口损坏	25.35	796.84
2	4	显示单元故障	10.90	731.44
3	4	时钟错误	25.74	794.65
4	4	误差超差	2.41	790.81

受模型准确度的限制,必须同时输出4种结果才能保证模型的可用性和对实际工作的高指导性,此时将4种输出结果均视为当前计量设备的可能故障原因及寿命。假设模型预测寿命服从正态分布,则利用寿命预测模型的测试误差及电表安装日期即可得到以下累计分布曲线。

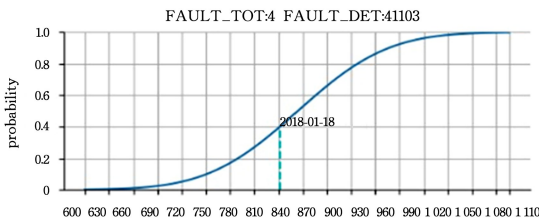


图3 时钟错误累计分布曲线图

图3中,曲线代表当电表使用日期到达某一天时,电表故障的发生概率开始随时间不断增加,直至为1;虚线代表当电表发生故障的概率到达40%时,即发出警告提示运维人员,方便其提前做好更换设备及人员安排,能够有效地为国网运维工作提供技术支撑。

4 实验与分析

4.1 数据说明

为了验证模型的可靠性,本文利用了中国电力科学研究院提供的全国智能电表历史故障数据,选择了招标批次为

20110101 3个招标批次(BID_BATCH_NO)的故障点表数据,其中总样本数量各为50000,过滤掉拆除时间在2016以前的和闲置时间180天以后的数据,样本数量还有140000条,其中应标厂商总共有4家,分别为宁波三星、哈尔滨汇鑫、杭州西力、威胜集团,各厂商故障电表数量为39368,35141,44881,49098。

由于各厂商各招标批次的故障电表数量存在差异,且样本过少的情况会影响模型的精确度,各厂商的占比情况为:0.28,0.24,0.32,0.36,选择宁波三星、哈尔滨汇鑫、杭州西力等厂商作为实验数据,其中威胜集团厂商的到货批次有2013000506885490,2616082131164320,2011000144242720,数量为1000,3000,300。因为2011000144242720批次故障样本数量过少,影响实验精度,故舍弃。

最终实验数据以招标批次为20100104,厂商为宁波三星、哈尔滨汇鑫、杭州西力,省市为宁波、哈尔滨、杭州,到货批次为2013000472287010,2614082530500330,2616082931175790,共140550条样本作为实验数据。取其中66.7%作为训练集,33.3%作为测试集。实验环境为window10系统,使用python中的sklearn作为实验工具。

4.1.1 数据采样

在得到的原始故障数据中,由于各厂商各招标批次损坏电表的数量都不相同,而数据量过小的样本集无法很好地代表当前招标批次的故障特点,因此将其作为训练集时无法得到很好的模型效果。

例如:本文选择故障电表占总电表数量的比例大于10%时的招标批次。表4列出了编号为22047的厂商部分招标批次的故障表数量及所占比例。

表4 部分招标批次故障表数量及所占比例

到货招标批次号	故障表数量	招标批次电表总数量	比例/%
26150423481497	7776	47819	16.26
700000000425	4708	84033	5.60
3385115980	2975	112907	2.63
3385116070	23308	64063	3.60
DHPH201407181158	1954	27578	7.09
26150615508678	1730	1849	93.56
261310115171	1346	19406	6.93

如表4所列,仅有招标批次号为26150423481497的故障表所占比例为16.26%,超过10%,具有一定分析价值,因此本方案中数据采样最终选择了厂商编号为22047、到货招标批次号为26150423481497的故障表数据作为处理数据集。

4.1.2 变量选择

经过分析,现有数据项中不少特征无用或具有强相关性,整理后的数据具有如下特点:1)条形码、首检合格时间属于无用信息;2)同一供货招标批次必属于同一招标批次;3)同一供货招标批次必属于同一省公司;4)同一供货招标批次必拥有同一设备规格;5)电能表运行状态属性的趋同性较高,待分拣、待分流、已报废、待报废都表示没有正在运行;6)实际数据中电能表状态主要集中在状态14(已报废)上,剩余其他状态也没有利用价值,基本都表达的是电表故障之后采取的措施,对预测故障没有帮助;7)故障数据中真故障电表的拆除原因基本一样,均为03(故障更换),不具有参考意义,偶尔会存在02(功能性更换)或04(客户用电变更),但是这不属于电表的自身原因,属于异常值,需在过滤异常时同步过滤掉;8)不同

厂商有不同特点,同一厂商的不同招标批次也可能具有不同特点,但是厂商和招标批次数量属于不定项,是随着时间不断增加的,因此需要对不同厂商的不同招标批次分开建模;9)产权单位代表电表的安装地区,电表故障与地区具有高相关性。分析提取后,原始数据中的有用变量为招标批次号、厂商编号、到货招标批次号、故障大类、故障小类、到货时间、安装时间、拆除时间、产权单位。其中部分电表状态如表 5 所列。

表 5 电表部分状态

编码	具体状态
001	新购
002	待耐压
003	待校验
004	待走字
005	返修在库
006	合格在库
007	待修理
008	待返厂
009	待报废
010	丢失
011	领出待装
012	运行
013	已返厂
014	已报废
015	拆回待退
016	预配待领
017	预领待装

4.1.3 数据替换

在数据丢失的情况下,数据记录是不完整的。含有丢值的数据记录是不能被轻易删除的,因为这可能造成数据量的不足;另一方面,数据记录中的其他属性可能含有非常有意义的信息,所以进行如下替换:

1)若安装时间和拆除时间均缺失,则安装时间用检定时间或者其他时间代替,拆除时间采用故障相关前后的时间代替。

2)若只有一种缺失,则根据当前故障记录的所属故障找到其他所有不含缺失值相同故障的故障记录,并计算它们的寿命(拆除时间-安装时间)均值,根据寿命均值来计算安装时间或拆除时间的替换值。

4.1.4 过滤异常

原始数据中,由于以往人员、设备等原因,采集到的数据无法直接作为模型输入,异常数据占了很大比例,无法作为有效的数据集被使用。在研究实际数据的特征后,选择了拆除时间在 2016 年之后且闲置时间在 180 天以下的故障数据。经过分析处理后,数据中还存在以下异常:

- 1)拆除时间在安装时间之前;
- 2)拆除时间距离安装时间过短,属于离群点;
- 3)拆除原因属于功能性更换的一类,实质上不属于故障,需要全部过滤掉。

针对异常 1)和异常 2),二者都属于离群点,在整体数据集中只占很少的一部分,因此采用四分位距法过滤极端值。

其中,分位数是将总体的全部数据按大小顺序排列后,处于各等分位置的变量值。如果将全部数据分成相等的两部分,它就是中位数;如果分成四等分,它就是四分位数;如果分成八等分,它就是八分位数等。4 分位数也称为四分位点,它是将全部数据分成相等的 4 部分,其中每部分包括 25% 的数据,处在各分位点的数值就是四分位数。四分位数有 3 个:第

一个四分位数就是通常所说的四分位数,称为下四分位数;第二个四分位数就是中位数;第三个四分位数称为上四分位数,分别用 Q_1, Q_2, Q_3 表示。

第一四分位数(Q_1),又称“较小四分位数”,等于该样本中所有数值由小到大排列后第 25% 的数字。

第二四分位数(Q_2),又称“中位数”,等于该样本中所有数值由小到大排列后第 50% 的数字。

第三四分位数(Q_3),又称“较大四分位数”,等于该样本中所有数值由小到大排列后第 75% 的数字。

第三四分位数与第一四分位数的差距又称四分位距(Inter Quartile Range, IQR)。

四分位距法中极端值定义为: $\min = Q_1 - 3 * IQR$, $\max = Q_3 + 3 * IQR$ 。

用箱型图对电表使用时间的分布情况进行整体的统计汇总,如图 4 所示。

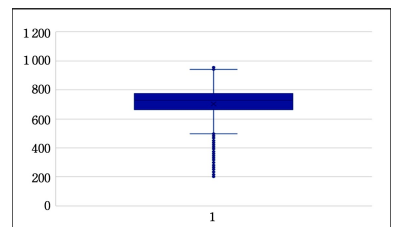


图 4 电表使用时间分布箱型图

由图 4 可知,大部分电表的使用时间集中在 600—800 天;异常值主要分布在底部,少数分布在主箱体之上。以招标批次号为 26150423481497 的数据集为例,本方案利用四分位距法过滤掉设备寿命中的极端异常值,其中四分位距法的各参数如表 6 所列。

表 6 四分位距法参数展示

名称	参数值/天
Q_1	639
Q_3	771
min	243
max	1167

根据表 6 中的参数,结合电力行业的规律,过滤掉数据集中寿命小于 150 天和大于 3000 天的异常数据,将原始数据由 7776 条最终减少到 7244 条,经验证,采用四分位距法过滤后极大地提高了模型的准确度。

4.2 有效性实验

训练完成后可得到故障预测模型,仿真实验的结果如表 7 所列。

表 7 模型性能评价指标表

故障类型	查准率	查全率	F1 分数
040101	90.34	96.75	93.44
040201	93.52	91.34	92.42
...
040902	89.74	90.83	90.28

查准率、查全率和 F1 分数都是评价分类模型性能的常用指标。查准率也称精确率,是指在一定实验条件下的多个测定值中,满足限定条件的测定值所占的比例,查全率也称召回率,是指预测出的真正值和实际值所有的正例的比率。而 F1 分数可以看作是模型的查准率和查全率的一种加权平均。由表 7 可知,模型在测试集上的查准率、查全率和 F1 分数的

宏平均值基本达到了 90%，体现了其分类精度高、泛化能力强的优点。

4.3 先进性实验

为了验证 GBDT 在该学习任务上的先进性,本方案首先选择几种常用的分类器:逻辑回归、神经网络、adaboost、决策树,在默认参数的条件下与 GBDT 进行比较。将预测后的结果与原始结果进行比对,并以智能电表的故障原因大类的命中率即作为该类的准确率。同样进行故障小类预测模型和寿命预测模型比较,结果如表 8 所列。

从实验结果可以看出,只有决策树以及 GBDT 的表现较优。于 BDT 的思想是将多个决策树组合在一起不断逼近真实值,在调参后模型精确度仍有上升空间,由此可见本方案选择的 GBDT 作为故障原因大类的预测分类器具有更好的表现性能。

表 8 故障小类预测模型比较结果

模型名称	预测准确度/%	生命周期 (均方根误差)/天
逻辑回归	44.32	66.72
BP 神经网络	39.51	141.59
Adaboost	53.84	88.40
决策树	74.14	51.78
梯度提升决策树	74.20	48.38

表 8 中,故障小类预测和寿命预测采用的对比模型存在差别,故障小类预测是分类任务,因此比较模型都以分类器评价指标为准确度,而寿命预测是回归任务,比较模型使用的是回归器,评价指标为均方根误差。如表 8 中的数据所示,GBDT 在本学习任务上具有优异的表现性能,在进行多源融合后,组合模型的性能也会更加优异。

结束语 本文主要设计了基于 GBDT 分类器和 GBDT 回归器的故障风险预测。该方法首先预测计量设备在发生故障的前提下具体的故障原因大类和故障原因小类,在得到 4 种可能后,分别以其作为条件预测设备的预期寿命,其中寿命的误差服从正态分布,因此可以得到设备故障日期的累计分布曲线,并以此提出在故障发生概率超过 40%后报警的预警手段。研究提出的基于 GBDT 多模型融合方法,可以发掘出故障电表中的隐含关系,对设备故障及寿命的预测准确度可

以达到 92.2%,对现场运维工作具有极大的价值。同时,由于本方案只利用了电表的故障数据,对于电表的运行数据还没有充分的利用,若能考虑到电表当前的运行状态,最终的故障预测结果将会更加准确。

参考文献

- [1] LIU J S. Analyzing Electricity Consumption via Data Mining [J]. Journal of Wuhan University, 2015, 12(10): 7-8.
- [2] NIKOVSKI D N, WANG Z, ESENTER A, et al. Smart meter data analysis for power theft detection [M] // Machine Learning and Data Mining in Pattern Recognition. Springer Berlin Heidelberg, 2013.
- [3] SAHOO S, NIKOVSKI D, MUSO T, et al. Electricity theft detection using smart meter data [C] // IEEE Innovative Smart Grid Technologies Conference. 2015: 1-5.
- [4] DEPURU S, et al. Support Vector Machine Based Data Classification for Detection of Electricity Theft [J]. Power Systems Conference & Exposition, 2011: 1-8.
- [5] ZAKARIA Z, LO K L. Two-stage fuzzy clustering approach for load profiling [C] // 2009 Proceedings of the 44th International Universities Power Engineering Conference (UPEC). IEEE, 2009.
- [6] 周开乐, 沈超, 丁帅. 基于遗传算法得微电网负荷优化分配 [J]. 中国管理科学, 2014, 22(3): 68-73.
- [7] 刘永光, 孙超亮, 牛贞贞, 等. 改进型模糊 C 均值聚类算法的电力负荷特性分类技术研究 [J]. 电测与仪表, 2014, 51(18): 5-9.
- [8] 董瑞, 黄民翔. 基于减法聚类的 FCM 算法在电力负荷分类中的应用 [J]. 华东电力, 2014, 42(5): 917-921.
- [9] BIDOKI S, MAHMOUDI-KOHAN N, GERAMI S. Comparison of several clustering methods in the case of electrical load curves classification [C] // IEEE Electrical Power Distribution Networks. 2011: 1-7.
- [10] MONEDERO I, et al. Detection of Frauds and Other Non-technical Losses in A Power Utility Using Pearson Networks and Decision Trees [J]. International Journal & Energy Systems, 2012, 34(1): 90-98.
- [11] 王立平, 邓芳明. 基于小波包和 GBDT 的瓦斯传感器故障诊断 [J]. 测控技术, 2016, 35(12): 30-33.
- [12] http://www.cs.ubc.ca/~lowe/keypoints/.
- [13] http://blog.sina.com.cn/s/blog_49cd44fb010005qb.html.
- [14] http://image.fengniao.com/360/3600114_all.html #3601563
- [15] TRALIC D, ZUPANCIC I, GRGIC S. CoMoFoD — New database for copy-move forgery detection [C] // 55th International Symposium ELMAR-2013. IEEE, 2013: 49-54.
- [4] ALCANTARILLA P F, BARTOLI A, DAVISON A J. KAZE Features [M] // European Conference on Computer Vision. Berlin; Springer, 2012: 214-227.
- [5] HUANG H, GUO W, ZHANG Y. Detection of Copy-Move Forgery in Digital Images Using SIFT Algorithm [C] // Pacific-Asia Workshop on Computational Intelligence and Industrial Application, 2008(PACIA'08). IEEE, 2009: 272-276.
- [6] ARDIZZONE E, BRUNO A, MAZZOLA G. Detecting multiple copies in tampered images [C] // IEEE International Conference on Image Processing. IEEE, 2010: 2117-2120.
- [7] 池静, 张婷, 杨振宇. 基于 SIFT 和区域生长的图像复制粘贴盲检测算法 [J]. 科学技术与工程, 2013, 13(4): 917-920.
- [8] AMERINI I, BALLAN L, CALDELLI R, et al. Copy-move forgery detection and localization by means of robust clustering with J-Linkage [J]. Signal Processing Image Communication, 2013, 28(6): 659-669.
- [9] 李云, 陈帆, 和红杰. SIFT 复制粘贴图像取证算法的安全性分析及改进 [C] // 全国信息隐藏暨多媒体信息安全学术大会. 2015.
- [10] 巩家昌. 基于局部颜色不变性特征的图像复制粘贴篡改取证研究 [D]. 天津: 天津大学, 2015.
- [11] FISCHLER M A, BOLLES R C. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography [J]. Comm. of the ACM, 1981, 24: 381-395.

(上接第 384 页)