

# 数据挖掘技术在电价执行稽查中的应用研究

利用电力计量自动化、营销系统中的海量用电数据，采用数据挖掘等方法构建了基于自适应模糊神经推理（Adaptive Neural – Fuzzy Inference System, ANFIS）和决策树C5.0算法的电价执行智能稽查模型。该模型首先采用减法聚类方法确定ANFIS的初始结构，并用混合学习算法训练得到ANFIS电价执行诊断模块。其次，利用C5.0决策树算法构建异常类型辨别模块，最后输出电价执行异常的用户及其相应的异常原因。将该智能稽查模型应用于南方某区域电网的电价执行营销稽查中。结果表明，其能快速有效地诊断、识别异常用户，具有良好的实用性和可行性。

彭显刚<sup>1</sup> 林利祥<sup>1</sup> 刘 艺<sup>1</sup> 林慕群<sup>2</sup> 郑伟钦<sup>1</sup> / 1. 广东工业大学自动化学院  
2. 广东电网公司汕头供电局

## 0 引言

电力营销管理是电网企业的重要工作，直接关系到电网企业的经济效益和社会效益<sup>[1]</sup>。然而，受传统管理理念及营销人员业务素质参差不齐等因素影响，目前部分区域电网的营销管理工作仍存在一定问题，特别是在电价执行中，存在部分营销人员的业务失误甚至徇私舞弊、企业违约用电等损害企业利益的现象。因此，有必要研究科学有效的电价执行稽查方法，以确保电价政策正确执行到位，维护正常的电价秩序。

传统的电价执行稽查<sup>[2-3]</sup>主要依靠定期巡检、随机抽样等方法，工作方式被动且稽查目标很不明确，无法及时发现并取缔电价执行异常用户。随着智能电网和电力大数据的发展，国内外电力企业的信息化建设不断加快<sup>[4-6]</sup>。在国内，国家电网公司提出了构建“客户导

向型，机构扁平化，业务集约化，管理专业化，管控实时化，服务协同化”的“一型五化”大营销体系<sup>[7]</sup>；南方电网公司也部署建设包括营销管理等六大系统和决策平台的“6+1”信息集成平台<sup>[8]</sup>。海量电力信息的集成共享与数据挖掘技术的发展为电价执行稽查提供了新的思路，使电价稽查工作由传统大海捞针式的核查转化为在线智能稽查成为可能。通过研究开发基于数据挖掘的电价执行稽查模型，深入挖掘及分析计量营销大数据，可以有效提高电力营销稽查的效率，减少供电企业的损失，是适应电力系统大数据发展的需求和实践。

因此，基于以上需求，提出了一种基于自适应模糊神经推理（Adaptive Neural – Fuzzy Inference System, ANFIS）和决策树C5.0算法的电价执行智能稽查模型（以下简称“智能稽查模型”），该模型具有自动建模和学习特性，能够快速准确地定位电价执行异常用户并诊断其异常原因。模型的输出为电价执行异常用户的嫌疑系数及其异常类型，可为电力稽查人员提供稽查的范围及依据。智能稽查模型首先采用自适应模糊神经推理系统ANFIS构建电价执行诊断模块，用于识别用户是否存在电价执行异常行为。其次，利用C5.0决策树算法构建异常类型辨别模块，用于对异常用户原因的分析和归类。预处理后的数据样本经过ANFIS诊断模块后，系统将输出电价执行异常用户，异常用户数据样本再经过C5.0识别模块，得到电价执行异常的原因。利用该方法对南方



彭显刚 / 副教授

### 关键词/Keywords

数据挖掘 ·  
营销稽查 ·  
电价执行 ·  
自适应模糊神经推理系统 ·  
决策树C5.0 ·

中国南方电网公司科技项目（K-GD2014-0609）。

某区域电网进行电价执行营销稽查。结果表明,其能快速有效地识别、诊断电价执行异常用户,具有良好的实用性和可行性。

## 1 智能稽查模型专家样本库构建

### 1.1 电价执行稽查评价指标体系

#### (1) 大工业客户电价执行异常行为特征

在电力营销中,大工业客户电费数额巨大,是供电企业财务收入的主要来源。确保大客户正确执行规定电价,按时足额缴纳电费是保证供电企业工作业绩及经济效益的客观要求<sup>[9]</sup>。因此,主要以大客户电价稽查为例进行建模与分析。根据供电部门长期的营销稽查经验,大工业客户偷逃电费的主要方式有:①私增容量。按合同容量收取基本电费的客户,采取修改变压器铭牌等方式私自增容用电,以达到少缴基本电费的目的。②超容用电。用户为少缴基本电费而采取措施尽量提高变压器负载率,使其超载运行,造成安全运行隐患。③未按规定执行峰谷分时电价。部分大客户由于经营性质、生产班制等原因,未按规定执行分时电价,以达到少计电度电费的目的。

#### (2) 电价执行评价指标体系设计

用电客户电价异常手段繁多,但其根本目的在于少缴电费,因此电价执行稽查可从分析电费的构成入手。根据我国现行电价政策,大工业用电客户采用两部制电价,由基本电价和电度电价组成,其电费构成为

总电费值 = 基本电费 + 电度电费 +

功率因数调整电费 + 基金及附加电费 (1)

综合分析大用户电费结构和常见的电价异常行为,选取对电价稽查有关联意义的变量构建大客户电价执行评判指标体系。

1) 基本电价类指标。大客户的基本电费有按受电变压器容量或最大需量计算两种计费方式,分别对应两类不同的基本电价,故评判基本电价执行情况的相关指标有:变压器容量和用户最大需量。

2) 分时电价类指标。电度电费除受用电总

量影响外,还受到用电的分布时段影响。评价企业分时电价政策实施情况的主要指标有:峰时电量、平时电量、谷时电量、峰总比、平总比、谷总比和峰谷差等。

3) 设备运行类指标。平均功率因数、平均负载率和超容率。

4) 特定类指标。本文定义“均价比例(Average Price Percentage, APP)”指标表示用户用电均价与其用电类别售电均价的比值,如式(2)所示。APP反映了客户某月份的用电均价 $P_{\text{customer}}$ 与其所属行业平均售电均价 $P_{\text{category}}$ 的大小关系,若PA远小于1,则该用户的异常嫌疑大大增加。此外,特定类指标还包括上月异常嫌疑系数(Suspected Coefficient of Last Month, SCL)。SCL为客户上周期稽查得出的异常嫌疑系数,通过该指标的累计,可增加异常用户的嫌疑系数,提高稽查的准确性。具体为

$$APP = \frac{P_{\text{customer}}}{P_{\text{category}}} \quad (2)$$

将上述各项评价指标按电费、电量、设备运行和特定特征进行归类,可得到如图1所示的电价

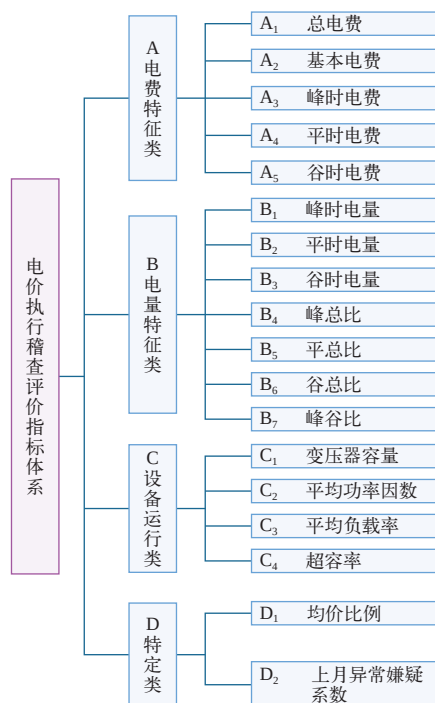


图1 电价执行稽查评价指标体系

执行稽查评价指标体系。

## 1.2 基于因子分析法的指标降维处理

在进行海量数据建模分析时，若模型的输入变量数目较多，将大大增加运算时间；而太少则无法全面覆盖必要的属性，可能遗漏重要的属性关系从而影响模型的准确度<sup>[10]</sup>。因此，为了减少计算量并确保稽查模型的准确度，采用因子分析法对输入变量进行降维处理。因子分析法是数据挖掘中常用的降维方法<sup>[11]</sup>，其在无损或很少损失原有信息量的前提下，将众多初始指标浓缩成少数相互独立的综合指标，达到数据集降维的目的。

## 2 基于ANFIS和决策树C5.0的智能稽查模型

所提出的智能稽查模型主要由ANFIS电价执行诊断模块和决策树C5.0异常类型辨别模块组成，其结构图如图2所示。从计量自动化、营销系统提取所需的数据样本后，经过数据预处理后输入到ANFIS诊断模块中，诊断出电价执行异常的用户；再将异常用户的数据样本输入C5.0异常类型辨别模块，最后得到电价异常嫌疑用户及其异常原因。

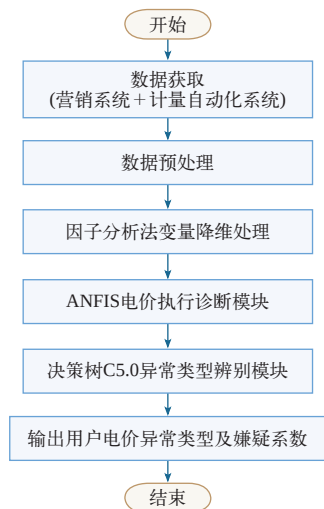


图2 基于ANFIS和决策树C5.0的智能稽查模型结构图

### 2.1 ANFIS电价执行诊断模块

ANFIS是一种基于Takagi-Sugeno模型的自适应模糊神经推理系统<sup>[12]</sup>，其将人工神经网络的自学习功能和模糊推理系统的模糊语言表达能力有

机结合，具有收敛速度快、推广能力强等特点，能以任意准确度逼近非线性函数<sup>[13]</sup>。利用ANFIS构建电价执行诊断模型，其初始结构采用减法聚类法自适应确定，避免了盲目性和随机性；参数则采用融合最小二乘法和梯度下降法的混合算法<sup>[14]</sup>进行训练，以提高优化效率。

#### (1) ANFIS初始结构的构建

ANFIS电价执行诊断模块的输入为1.2节因子分析法处理后的变量。为了避免ANFIS模糊规则数量成指数增长，导致“维数灾难”，采用减法聚类法构建ANFIS模型的初始结构。减法聚类是一种自动快速确定数据组的聚类个数和聚类中心位置的单次算法<sup>[15]</sup>，其计算量与数据点成线性关系，且与变量的维数无关，从而避免了ANFIS建模中可能出现的“维数灾难”问题，其基本原理如下。

1) 假设 $M$ 维空间中的 $k$ 个样本数据点 $x_i (i = 1, 2, \dots, k)$ 均为聚类中心的候选解，并由式(3)计算各数据点 $x_i$ 的密度指标 $D_i$ 。

2) 将密度指标最大的数据点作为第一个聚类中心，记为 $x_{cls\_1}$ ，其密度空间为 $D_{cls\_1}$ 。

3) 将第 $n$ 次取得的聚类中心记为 $x_{cls\_n}$ ，其密度空间为 $D_{cls\_n}$ 。

利用式(4)修正剩余数据点的密度指标

$$D_i = \sum_{j=1}^k \exp \left[ -\frac{\|x_i - x_j\|^2}{(r_a/2)^2} \right] \quad (3)$$

$$D_i = D_i - D_{cls\_n} \exp \left[ -\frac{\|x_i - x_{cls\_n}\|^2}{(r_b/2)^2} \right], \quad i \neq cls\_n$$

式中， $r_a$ 、 $r_b$ 分别为 $x_i$ 和 $x_{cls\_i}$ 的邻域，邻域外的点对聚类中心的密度指标影响较小。

4) 选取密度指标最大的点作为新的聚类中心 $x_{cls_{n+1}}$ 。若 $D_i < 0$ 时，则相应的数据点密度指标显著减小，不会成为下一个聚类中心。

5) 判断 $D_{cls_{n+1}}/D_{cls_i}$ 是否小于阈值 $\varepsilon$ ，若否，返回步骤3)，直至满足约束条件。

通过减法聚类可以高效地确定模糊系统的模糊规则数和输入变量的隶属度函数数目，并确定隶属度函数的位置。

## (2) ANFIS的结构及流程

ANFIS的结构简图如3所示,其模糊规则为

$$R: \text{ if } x_1 = A_1, x_2 = A_2, \dots, x_n = A_n, \\ \text{ Then: } y = c_0 + \sum_{i=1}^n c_i x_i \quad (4)$$

式中,  $x_i$  为输入变量;  $A_i$  为输入模糊语言值;  $c_i$  为结论参数;  $y$  为规则输出。

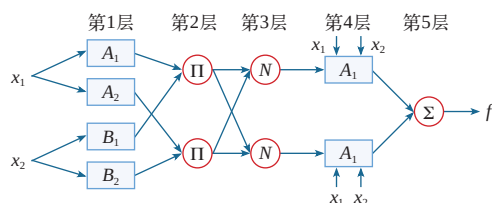


图3 ANFIS结构简图

图3中,模型的输入节点数为 $N$ ,分别对应 $N$ 个电价执行评价指标。ANFIS的第1层为输入变量自适应模糊化层,主要用以对电价执行诊断输入变量的模糊化处理,并计算其隶属度。第2层为模糊推理层,计算模糊规则的激励强度。第3层用于归一化激励强度。第4层计算各规则的输出。第5层是ANFIS网络的总输出。

ANFIS电价执行诊断模块的流程为:首先从计量自动化、营销系统提取所需的历史数据样本,经过缺失值、异常值和因子分析的预处理后,输入ANFIS诊断模型中,输出结果为用户电价执行异常嫌疑系数。嫌疑系数越大,用户存在电价执行异常的可能性越大,经过多次试验设定判定阈值为0.95,则嫌疑系数 $>0.95$ 时,判定为电价执行异常用户。

## 2.2 决策树C5.0异常类型辨别模块

决策树(Decision Tree)算法起源于人工智能机器学习系统,用以探究数据内在规律和对新数据的分类预测中,具有数据分析能力出色、结果显现直观和可解释性强等优点<sup>[16]</sup>。目前,该算法广泛应用于数据挖掘和各类智能分类系统中。C5.0是具有代表性的决策树算法,由C4.5算法改进而来。C5.0可用于建立多叉的分类树,其以信息增益率为标准确定最佳分组变量和分割点,并通过对决策树各叶子的裁剪或合并来提高分类准确度<sup>[17]</sup>。其

基本原理如下。

设 $U$ 为包含 $n$ 个样本的训练集,将样本集 $m$ 个不同的目标属性类 $C = \{c_1, c_2, \dots, c_m\}$ ,且 $c_i$ 在所有样本中出现的频率为 $p_i (i = 1, 2, \dots, m)$ ,则集合 $U$ 包含的信息熵为

$$\text{Ent}(U) = - \sum_{i=1}^m p_i \log_2(p_i) \quad (5)$$

选择属性 $A$ 作为决策树的根节点,则属性 $A$ 对集合 $U$ 的划分熵值为

$$\text{Ent}(U, A) = \sum_{i=1}^k \frac{|U_i|}{|U|} \text{Ent}(U_i) \quad (6)$$

故 $A$ 相对于 $U$ 的信息增益为

$$G(U, A) = \text{Ent}(U) - \text{Ent}(U, A) \quad (7)$$

C5.0采取引入属性的分裂信息对信息增益加以调节,分裂信息定义为

$$\text{SplitE} = - \sum_{i=1}^k \frac{|U_i|}{|U|} \log_2 \frac{|U_i|}{|U|} \quad (8)$$

其信息增益率为

$$\text{GR}(U, A) = \frac{G(U, A)}{\text{SplitE}(U, A)} \quad (9)$$

C5.0算法采用信息增益率作为决策树分枝准则,当样本的属性有较多的分类取值时,即使其信息熵会偏大,但由于信息增益率因分裂信息的增大而降低,从而使分类预测模型可消除属性取值数目带来的影响,保证较高的预测准确度。

基于决策树C5.0的异常类型辨别模型以ANFIS诊断模型中输出的异常用户样本数据为输入,输出为电价异常类型,包括私增容量、超容用电和未执行峰谷分时电价等。根据稽查模型输出的结果,电力稽查人员即可方便快捷地定位电价执行异常嫌疑用户,并根据异常类型报告有针对性地进行现场核查,大大增加营销稽查的智能性和精准性。

## 3 实例分析

### 3.1 样本数据提取

从某供电局计量自动化、营销系统中提取2012年1月~2014年12月3年内记录完整的大工业

用户数据进行建模分析。算例选取电价正常客户200户和已确认异常并开罚单的客户105户的数据进行建模,并将构建的稽查模型应用于2014年12个月的3 527个大客户的用电数据上进行验证。表1中列出了模型构建、模型验证的数据样本信息。

表1 数据样本信息表

| 样本用途 | 时间范围             | 大客户总数   | 样本数目   |
|------|------------------|---------|--------|
| 模型构建 | 2012年1月~2013年12月 | 200+105 | 7 320  |
| 模型验证 | 2014年1月~2014年12月 | 3 527   | 42 324 |

### 3.2 样本数据预处理

从电力应用系统提取的海量数据中常常包含许多噪声或不完整数据,因此必须先对数据进行合理的预处理,以提高稽查模型的准确度。采用均值插补法对缺失值进行插补,采用数学回归法对异常值进行平滑处理,预处理方法的具体原理可参见文献[18]。

### 3.3 输入变量降维

#### (1) 因子分析检验

应用因子分析法的前提是要求原指标变量之间具有较强的相关关系,在数据挖掘中一般采用KMO检验与Bartlett球形检验进行检验。通过计算各电价执行评价指标的相关系数可知,指标 $A_1$ 、 $A_2$ 、 $A_3$ 、 $A_4$ 、 $A_5$ 、 $B_1$ 、 $B_2$ 、 $B_3$ 和 $C_1$ 具有较高的相关性。对上述指标分别作KMO检验与Bartlett球形检验,得到KMO检验值为0.765;Bartlett球形检验的显著性水平为0,拒绝零假设。以上检验结果表明信息有较高重复度,适合作因子分析降维。

#### (2) 因子提取

采用SPSS19.0作因子分析,采用主成分分析法提取,因子提取的结果如表2所示。

表2 因子分析提取结果解释

| 成分 | 初始特征值 |        |        | 提取平方和载入 |        |        |
|----|-------|--------|--------|---------|--------|--------|
|    | 合计    | 方差(%)  | 累计(%)  | 合计      | 方差(%)  | 累计(%)  |
| 1  | 5.871 | 65.233 | 65.233 | 5.871   | 65.233 | 65.233 |
| 2  | 1.803 | 20.037 | 85.271 | 1.803   | 20.037 | 85.271 |
| 3  | 0.583 | 6.483  | 91.753 |         |        |        |
| 4  | 0.401 | 4.460  | 96.214 |         |        |        |
| 5  | 0.175 | 1.945  | 98.159 |         |        |        |

(续)

| 成分 | 初始特征值 |       |         | 提取平方和载入 |       |       |
|----|-------|-------|---------|---------|-------|-------|
|    | 合计    | 方差(%) | 累计(%)   | 合计      | 方差(%) | 累计(%) |
| 6  | 0.074 | 0.823 | 98.983  |         |       |       |
| 7  | 0.063 | 0.698 | 99.681  |         |       |       |
| 8  | 0.024 | 0.263 | 99.945  |         |       |       |
| 9  | 0.005 | 0.055 | 100.000 |         |       |       |

从表2可得,因子分析后的前两个主成分 $PCA_1$ 、 $PCA_2$ 的累积贡献率已在85%以上,表明其已包含绝大多数信息,可用于代替原有的评价指标。由此可得到智能稽查模型的输入指标为: $PCA_1$ 、 $PCA_2$ 、 $B_4$ 、 $B_5$ 、 $B_6$ 、 $B_7$ 、 $C_2$ 、 $C_3$ 、 $C_4$ 、 $D_1$ 和 $D_2$ 。

### 3.4 电价稽查结果及分析

将3.2节中因子分析提取指标的具体数值输入模型算法中进行训练,得到智能稽查模型。将模型验证的数据经预处理后输入到稽查模型中,经过ANFIS诊断模块和决策树C5.0异常类型辨别模型处理后,得到的电价稽查统计结果。为了验证稽查模型的有效性,将其与基于人工神经网络(Artificial Neural Network, ANN)模型的稽查结果进行对比如表3所示。

表3 模型应用结果及对比

| 指标           | 本文稽查模型 | ANN建模 |
|--------------|--------|-------|
| 参与验证用户数      | 3 527  | 3 527 |
| 实际电价异常用户数    | 32     | 32    |
| 模型诊断为异常用户数   | 110    | 413   |
| 其中实际异常匹配数    | 32     | 29    |
| 不匹配用户数       | 78     | 387   |
| 异常类型匹配数      | 29     | 24    |
| 电价诊断准确率(%)   | 97.8   | 89.0  |
| 异常类型辨别准确率(%) | 90.6   | 82.8  |

电价稽查模型的应用效果可通过电价诊断准确率和异常类型辨别准确率来衡量。其中,电价诊断准确率是指电价执行状态诊断正确的用户的比例;异常类型辨别准确率是指电价异常执行状态及异常原因(若为异常用户)均诊断正确的用户的比例。由表3计算可知,提出的稽查模型的电价诊断准确率和异常类型辨别准确率分别为97.8%和90.6%,分别比ANN方法提高了9.89%和

9.42%。对比结果表明了所建立的稽查模型的有效性和实用性。将该智能稽查模型应用于供电企业营销稽查中,能够有效提高电力稽查工作的准确性和智能性,具有很好的实用价值。

## 4 结束语

针对目前供电企业营销稽查主要依靠人工定期巡检等被动方法,无法及时发现并取缔电价执行异常用户的现状,提出了基于自适应模糊神经推理ANFIS和决策树C5.0算法的电价执行智能稽查模型。该模型采用数据挖掘技术方法对电力计量营销的海量数据进行深入挖掘与分析,以提高电力营销稽查的准确性和智能性。实际测试证明,该模型对电价异常客户与异常类型识别具有较高的准确率,能够有效提高电力营销稽查的工作效率,降低用电稽查的差错率,具有很好的实用价值。基于数据挖掘技术的电价执行稽查有利于电力营销工作的信息化建设,也是适应电力系统大数据发展的需求和实践。

### 参考文献

- [1] 康重庆, 庄彦, 胡江溢, 等. 售电市场中售电量的风险分析 [J]. 中国电机工程学报, 2009, 29 (19): 78-84.
- [2] 刘小渊, 庄华松. 电力营销稽查管理中存在的问题及策略探究 [J]. 中国高新技术企业, 2014 (31): 162-163.
- [3] 张婉琳, 王海涛, 张慧琦, 等. 通过营销稽查实现电价执行内控的实践与分析 [J]. 黑龙江科学, 2014, 5 (11): 174.
- [4] Qiu R, Antonik P. Big data and smart grid: a random matrix [M]. John Wiley and Sons, 2014: 33-60.
- [5] Xindong W, Xingquan Z, Gong qing W, et al. Data mining with big data [J]. IEEE Transactions on Knowledge and Data Engineering, 2014, 26 (1): 97-107.
- [6] 王德文, 孙志伟. 电力用户侧大数据分析并行负荷预测 [J]. 中国电机工程学报, 2015, 35 (3): 527-537.
- [7] 许其国. “一型五化”构建大营销 [J]. 中国电力企业管理, 2014 (8): 63-64.
- [8] 建设“6+1”系统打通信息孤岛 [EB/OL]. [http://www.csg.cn/epaper/html/2014-02/28/content\\_53041.htm](http://www.csg.cn/epaper/html/2014-02/28/content_53041.htm).
- [9] 周凯, 栗秋华, 周林, 等. 基于物元分析理论的大电力客户信用评价 [J]. 电网技术, 2009, 33 (16): 75-80.
- [10] 崔巨勇, 于同伟, 黄旭, 等. 智能变电站入侵检测数据降维方法的研究 [J]. 电气应用, 2014, 33 (21): 106-110.
- [11] 赵毅斌. 基于因子分析的变频器故障诊断 [J]. 电气应用, 2011, 30 (11): 72-74.
- [12] 陈继光. 基于ANFIS模型的CPI经济数据预测 [J]. 统计与决策, 2013 (18): 68-69.
- [13] 张钧, 李小鹏, 何正友. 采用自适应神经模糊推理系统的配电网故障分类方法 [J]. 中国电机工程学报, 2010, 30 (25): 87-93.
- [14] 王艳永, 邓方, 孙健. 改进的自适应神经模糊推理系统的角度传感器误差补偿方法 [J]. 控制理论与应用, 2013, 30 (10): 1342-1346.
- [15] 王华强, 袁浩, 杨滁光. 自适应模糊神经网络在EPS中的应用 [J]. 合肥工业大学学报: 自然科学版, 2011, 34 (2): 188-191.
- [16] 陈华丰, 张葛祥. 基于决策树和支持向量机的电能质量扰动识别 [J]. 电网技术, 2013, 37 (5): 1272-1278.
- [17] 张剑清, 段艳. 一种结合分水岭与决策树C5.0的极化SAR分类方法 [J]. 武汉大学学报: 信息科学版, 2014, 39 (8): 891-896.
- [18] 程开明. 统计数据预处理的理论与方法述评 [J]. 统计与信息论坛, 2007, 22 (6): 98-103.

(收稿日期: 2015-04-28) EA