

基于SMOTE和决策树算法的电力变压器 状态评估知识获取方法

谢桦¹, 陈俊星¹, 赵宇明², 丁庆², 张沛¹

(1. 北京交通大学电气工程学院, 北京 100044; 2. 深圳供电局有限公司, 广东 深圳 518001)

摘要:提出基于合成少数过采样技术(SMOTE)算法和决策树算法的电力变压器状态评估知识获取方法,首先针对变压器非正常状态样本数量较少的情况,采用SMOTE算法补充非正常状态样本数量,解决了变压器样本集类别不平衡问题。然后将变压器状态评估过程视为分类过程,利用决策树模型为白箱模型的特点,将变压器状态评估知识获取问题转化为构建决策树的问题。最后采用C4.5决策树算法构建决策树,从中提取变压器状态评估知识,得到关键变压器状态量和评估规则。以某地市级供电公司110 kV电压等级油浸式变压器实际数据开展实例分析,结果表明所提出的方法能实现状态评估知识的自动化获取,可以为该地区110 kV油浸式变压器的状态评估工作提供决策支持。

关键词:电力变压器;知识获取;决策树算法;SMOTE;数据挖掘

中图分类号:TM 41

文献标志码:A

DOI:10.16081/j.epae.202002002

0 引言

电力设备的状态检修策略正逐步替代传统的定期检修策略。实现电力设备状态检修的核心工作是确定电力设备的状态^[1]。电力变压器(以下简称变压器)作为电网中的重要变电设备,对其进行准确的状态评估可以及时发现设备异常,降低故障率。目前,变压器状态评估广泛采用的方法是依据Q/GDW 169—2008《油浸式变压器(电抗器)状态评价导则》(以下简称《导则》)进行评分。这种方法根据变压器不同状态量的测量结果,采用扣分方式,然后依据相关的评价标准判断变压器的运行状态。仅依照《导则》的评估标准对变压器进行评估没有考虑在实际条件(如特定地区和特定电压等级条件)下不同型号变压器在评估标准上存在的差异。

变压器在运行过程和运维检修过程中会产生许多类型的数据,如运行数据、试验数据、巡检数据、状态监测数据。随着电力生产管理系统等电力信息系统的应用和推广,变压器的相关数据得到了有效的采集和存储。这些数据中隐含着反映变压器状态劣化过程的知识,但是对这些数据进行人工统计分析提取相关知识的难度较高,因为能够反映变压器状态变化的状态量较多,并且状态变化机理较为复杂。这些客观因素的影响以及人工方式本身存在效率较

低的缺点,导致知识获取工作质量较差,很难对知识进行较为快速、全面的提取^[2]。因此,需要一种能够针对变压器数据进行高效的状态评估知识获取的方法。

近年来,数据挖掘技术由于在处理数据方面具有处理数据量大、处理速度快的优势^[3],已经在变压器状态评估领域得到了应用。文献[4]建立了基于贝叶斯网络的变压器状态评估模型,首先建立变压器健康状态量化的分层模型,然后利用模糊隶属度函数确定分层模型中的状态量阈值,最后采用贝叶斯网络构建变压器状态评估模型。文献[5]建立了基于帝国殖民竞争算法优化支持向量机的非线性多分类模型,结合交叉验证原理对变压器进行了故障诊断。文献[6]应用人工神经网络和Dempster-Shafer证据理论构建多信息融合的变压器状态评估模型。文献[7]提出一种基于AdaBoost-RBF算法与Dezert-Smarandache理论的变压器故障诊断方法。虽然上述数据挖掘方法都可以构建效果较好的评估模型,但是这些模型都属于黑箱模型,难以得到其中的决策规则以及进行知识获取。为了解决此问题,本文提出利用选择决策树算法实现变压器状态评估知识获取。决策树模型是一种白箱模型,能够以决策树图的形式表示模型的决策过程。通过决策树算法分析处理变压器相关数据并构建决策树,将变压器数据中的状态评估知识转化为if-then形式的决策树判断过程,即可实现自动化的知识获取。因此,应用决策树算法可以达到从变压器数据中高效获取状态评估知识的目的。

本文将变压器状态评估看作状态等级分类问题,提出了基于合成少数过采样技术SMOTE(Syn-

收稿日期:2019-07-10;修回日期:2019-12-05

基金项目:国家重点研发计划资助项目(2018YFB0905305);
中国南方电网有限责任公司科技项目(090000KK52190002)

Project supported by the National Key Research and Development Program of China(2018YFB0905305) and the Science and Technology Project of China Southern Power Grid Company(090000KK52190002)

thetic Minority Over-sampling TEchnique)算法和决策树算法的电力变压器状态评估知识获取方法:首先采用SMOTE算法生成状态等级训练样本集,其次构建基于C4.5算法的状态评估知识决策树,然后基于决策树实现状态评估知识的自动获取。最后通过实例对本文方法进行验证和结果分析,并对《导则》中评价规则提出相关建议。

1 理论框架

知识获取是指从大量数据中去除无用信息、提取有用信息的过程^[8]。本文应用决策树算法从变压器相关数据中归纳变压器状态评估知识,达到准确、快速获取知识的目的。变压器状态评估结果使用状态等级进行表示,分为正常状态、注意状态、异常状态、严重状态4个类别,因此变压器状态评估过程属于分类问题^[9]。决策树算法是一种计算量较小、分类精度较高的分类算法^[10],适用于处理变压器状态评估过程。在此基础上,由于决策树模型是白箱模型,可以从中获取决策过程。本文将变压器评估知识获取问题转化为构建决策树的问题。

本文将理论框架分为3个部分:样本集的生成、变压器状态评估知识决策树的构建、基于决策树的知识获取。样本集的生成部分介绍了样本集的构建以及处理过程;变压器状态评估知识决策树的构建部分介绍了决策树算法的选择及构建变压器状态评估知识决策树模型的理论步骤;基于决策树的知识获取部分介绍了从构建的决策树中进行知识获取的过程。

1.1 样本集的生成

样本集的生成是知识获取的前期重要环节。变压器数据主要来源于人工录入与监测装置自动采集,数据质量参差不齐,存在数据缺失、数据冗余等问题。为了提高数据质量,将采集的数据进行数据清理、数据变换、数据集成等处理,形成变压器原始样本集。

变压器属于运行可靠性较高的设备,绝大部分时间处于正常状态,因此在原始样本集中正常状态的样本远多于注意状态、异常状态和严重状态(以下将这3种状态统称为非正常状态)的样本。各状态样本数不平衡会使变压器原始样本集出现类别不平衡问题^[11],即正常状态样本数远多于非正常状态样本数,这可能造成分类算法在进行分类判断时,倾向于忽略非正常状态样本,导致分类性能下降^[12]。

为解决变压器样本集的类别不平衡问题,本文采用SMOTE算法^[13]对原始样本集的非正常状态样本进行过采样处理增加其数量,以减轻样本集的类别不平衡现象,从而提高分类算法性能。SMOTE算法的核心思想是通过对于少数类样本和其近邻样本进

行线性组合来合成新的样本,从而扩充少数类样本的数量^[14]。

SMOTE算法原理如图1所示,生成新样本的步骤如下^[15]。

(1)设少数类样本数量为 N ,过采样倍数为 M ,少数类中样本为 $x_i (i = 1, 2, \dots, N)$,以欧几里得距离为依据搜索 x_i 的 K 个近邻,选取其中 M 个近邻样本 $x_{ij} (j = 1, 2, \dots, M)$ 。

(2)根据样本 x_i 以及选取的近邻样本 x_{ij} ,合成新的少数类样本 y_{ij} 。

$$y_{ij} = x_i + \text{rand}(0, 1) \times (x_i - x_{ij}) \quad (1)$$

其中,函数 $\text{rand}(0, 1)$ 为取0~1间的随机数。

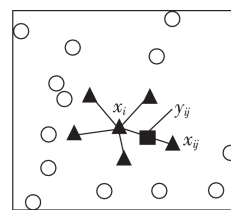


图1 SMOTE算法原理图

Fig.1 Principle diagram of SMOTE algorithm

1.2 变压器状态评估知识决策树的构建

基于经过SMOTE算法处理后的变压器样本集数据,通过使用决策树算法构建变压器状态评估知识决策树,可实现状态评估知识获取。决策树算法中较有代表性的有ID3(Iterative Dichotomiser 3)算法、C4.5算法、CART(Classification And Regression Tree)算法。其中,ID3算法无法处理连续型属性^[16];CART算法只能生成二叉树,无法直接进行多分类^[17];C4.5算法在ID3算法的基础上增加了对连续型属性的处理能力,同时支持决策树的多路划分。由于变压器样本集中含有连续型状态量数据和多分类状态量数据,本文选择C4.5算法构建变压器状态评估知识决策树。基于C4.5算法构建决策树的步骤如下。

(1)计算整个样本集的信息熵。假设整个样本集 S 中第 i 类分类结果对应的样本占比为 $p_i (i = 1, 2, \dots, N)$,则 S 的信息熵为:

$$E(S) = - \sum_{i=1}^N p_i \log_2 p_i \quad (2)$$

其中,样本集 S 为变压器样本集,其分类结果为变压器状态评价等级。

(2)计算每个属性(在本文中属性定义为变压器状态量)对样本集 S 进行划分的信息熵。若属性 a 的数据为离散型数据,有 k 个值,分别为 $a_1, a_2, \dots, a_k$,使用属性 a 来对样本集 S 进行划分,产生 k 个分支节点, S_j 为其中第 j 个分支节点上取值为 a_j 的所有样本。计算以离散属性 a 对样本集 S 进行划分的信

息熵:

$$E(S, a) = \left(\left| S_j \right| E(S_j) \right) / |S| \quad (3)$$

若属性 a 的数据为连续型数据,那么需要对这些数据进行离散化。首先,对这些数据进行递增排列,设每一对数据的中间点作为可能的划分点,划分点左侧的数据集为 S_l ,样本数为 $|S_l|$,右侧的数据集为 S_r ,样本数为 $|S_r|$,总样本集 S 的样本数为 $|S|$,则每种划分方式的信息熵为:

$$E(S, a) = |S_l| E(S_l) / |S| + |S_r| E(S_r) / |S| \quad (4)$$

(3) 计算信息增益。每个属性都有其对应的信息增益,属性 a 的信息增益 I_a 为:

$$I_a = E(S) - E(S, a) \quad (5)$$

(4) 计算属性的分裂信息。每个属性对应一个相应的分裂信息,属性 a 的分裂信息 N_a 为:

$$N_a = - \sum_{j=1}^k \frac{|S_j|}{|S|} \log_2 \frac{|S_j|}{|S|} \quad (6)$$

(5) 计算信息增益率。属性 a 的信息增益率 R_a 为:

$$R_a = I_a / N_a \quad (7)$$

对所有属性进行信息增益率的计算后,选择 R_a 最大的属性为节点,并将其加入决策树。重复步骤 (1) — (5),即可逐步构建决策树。

1.3 基于决策树的知识获取

变压器状态评估知识决策树构建完成后,获取的知识以决策树的形式进行表示,如图 2 所示。图中,根节点 A 和非叶子节点 B 表示变压器状态量;叶子节点 $k_i (i=1, 2, 3)$ 表示变压器状态评估结果;从树的根节点经分支(如图中虚线所示)连接直到某个叶子节点的一条路径表示一条分类规则^[18]。

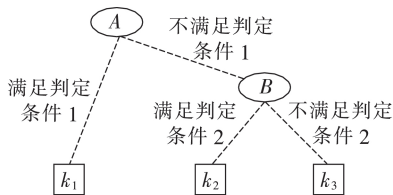


图 2 决策树示意图

Fig.2 Schematic diagram of decision tree

变压器状态评估知识包括经 C4.5 算法筛选后得到的变压器关键状态量 A 、 B , 以及以下形式的变

压器状态评估规则: ①if A 满足判定条件 1, then 状态为 k_1 ; ②if A 不满足判定条件 1 且 B 满足判定条件 2, then 状态为 k_2 ; ③if A 不满足判定条件 1 且 B 不满足判定条件 2, then 状态为 k_3 。

2 实例分析

在某地市级供电公司的实地调研中,从电力生产管理系统中采集 265 台 110 kV 电压等级油浸式变压器相关数据,如表 1 所示。其中变压器型号 25 种,以代号 A—Y 表示,数据的时间跨度为 2015 年底至 2017 年底。

表 1 变压器数据类型及数据量

Table 1 Data types and amount of transformers

数据报告类型	主要包含信息	数据量 / 条
状态评价	待分析状态量、状态评估结果	2462
化学试验	变压器油试验结果	534
电气试验	电气试验结果	365
带电试验	带电试验结果	42
检修决策	上次检修时间	91
台账信息	变压器型号、变压器投运时间	265

2.1 样本集的生成

表 1 所示状态评价报告中的状态量(即待分析状态量)主要有 22 个,如表 2 所示。其中将变压器状态评价报告中的状态量分为 5 个部分,分别为变压器本体部分、套管部分、冷却器系统部分、有载分接开关部分、非电量保护部分,各部分所包含多个待分析状态量。

表 1 中数据量达到了 3759 条,且包含的信息较多,例如状态评价报告包含了多个待分析状态量。如果采用人工分析方法对这些数据进行统计分析以总结提取知识,具有较大难度。而决策树算法具有数据挖掘算法的优点,能够应对数据量大、信息繁杂的情况,从而进行高效的知识获取。

将表 1 中不同类型数据进行数据清理、数据变换、数据集成等处理,形成变压器原始样本集。此样本集总样本数为 769 条,其中正常状态样本 659 条,注意状态样本 59 条,异常状态样本 9 条,严重状态样本 42 条。正常状态样本的数量大约为其他 3 种非正常状态样本的 6 倍,存在类别不平衡现象。

为了平衡正常状态样本数量与非正常状态样本

表 2 变压器状态评价报告中的主要状态量

Table 2 Main state parameters from transformer condition evaluation report

状态量	包含内容
本体	绕组直流电阻各相绕组相互间的差别、绕组绝缘电阻、极化指数、吸收比、铁芯接地电流、渗油、漏油、表面锈蚀、呼吸器状态、本体储油柜油位
套管	电容量偏差、红外测温、主屏、绝缘电阻、末屏绝缘电阻
冷却器系统	漏油、散热片锈蚀、电机运行
有载分接开关	渗漏、油位、呼吸器状态
非电量保护	温度计和分接开关位置等远方与就地指示一致性、温度计

数量,使用SMOTE算法分别对原始样本集中的注意状态样本、异常状态样本、严重状态样本进行过采样。设 $M=5, K=5$,样本集经SMOTE算法过采样处理后样本总数为1319,SMOTE算法处理前后的样本集构成情况如表3所示。由表可见,经SMOTE算法过采样处理后的非正常状态样本数为660,正常状态样本数量与非正常状态样本数量基本达到平衡。

表3 SMOTE算法处理前后的样本集构成情况

Table 3 Sample set composition before and after processed by SMOTE algorithm

样本集类型	状态样本数			
	正常	注意	异常	严重
处理前	659	59	9	42
处理后	659	354	54	252

2.2 变压器状态评估知识决策树构建

以SMOTE算法处理后的样本集为数据基础,使用C4.5决策树算法构建变压器状态评估决策树。以表4所示套管电容量偏差的相应样本统计数据为例解释决策树的构建过程。其中,套管电容量偏差分为“合格”和“不合格”2种类别,属于离散量。

表4 套管电容量偏差状态量的样本数统计

Table 4 Sample number statistics of bushing capacitance deviation

状态	样本数		状态	样本数	
	合格	不合格		合格	不合格
正常	659	0	异常	54	0
注意	354	0	严重	120	132

采用本文方法,可计算得到构建的变压器状态评估知识决策树模型混淆矩阵如下:

$$\begin{array}{c}
 \text{模型评估结果} \\
 \begin{array}{c} \text{正常} \quad \text{注意} \quad \text{异常} \quad \text{严重} \end{array} \\
 \text{实际评估结果} \left\{ \begin{array}{l} \text{正常} \\ \text{注意} \\ \text{异常} \\ \text{严重} \end{array} \right. \begin{bmatrix} 659 & 0 & 0 & 0 \\ 8 & 346 & 0 & 0 \\ 1 & 0 & 53 & 0 \\ 0 & 0 & 0 & 252 \end{bmatrix} \quad (8)
 \end{array}$$

和实际调研结果相比,本文方法的分类正确率为99.318%。其中,正常状态及严重状态正确率为100%,注意状态正确率为97.74%,异常状态正确率为98.15%。上述数据说明本文构建的决策树对变压器各状态的判断准确率均较高。

变压器状态评估知识决策树的Kappa系数^[19]的计算结果为0.9893。依据文献[20]给出的分类精度划分等级(见附录中表A1),本文提出的方法分类精度属于“非常好”等级,验证了本文构建的决策树分类性能出色,可以确保知识获取过程的可靠性。

2.3 变压器状态评估知识表示

以C4.5算法完成自动化的知识获取过程可构

建变压器状态评估知识决策树,如附录中图A1所示,其中各状态量阈值精确到小数点后3位。提取决策树从根节点到叶子节点的if-then形式的决策过程,得到变压器状态的评价结果。以下展示部分提取规则:①若套管电容量偏差不合格,则变压器状态为严重状态;②若套管电容量偏差合格、本体绕组直流电阻各相绕组相互间的差别大于 0.013Ω ,则变压器状态评价结果为严重状态;③若套管电容量偏差合格、本体绕组直流电阻各相绕组相互间的差别大于0且小于等于 0.013Ω ,且本体绕组绝缘电阻大于 $9878.768 \text{ M}\Omega$,则变压器状态评价结果为注意状态。

2.4 获取知识的应用

变压器状态评估决策树中根节点以及非叶子节点上的关键状态量有:本体部件的绕组直流电阻各相绕组相互之间的差别、绕组绝缘电阻、吸收比、铁芯接地电流、渗油、呼吸器状态;套管部件的电容量偏差;分接开关部件的渗漏、油位。在该地区的变压器运检工作中,通过加强对关键状态量的监测可以提前发现状态劣化征兆,及时对变压器采取维护策略,避免变压器状态持续恶化。此外,变压器型号对状态评估过程具有一定的影响,比如:在套管电容量偏差合格、绕组直流电阻偏差大于0小于等于 0.013Ω 时,型号R的变压器对绕组绝缘电阻的变化较为不敏感,而其他24种变压器对绕组绝缘电阻的变化更敏感,因此需要特别注意不同型号变压器的差异性。

除了加强对关键状态量的监测外,利用前述提取的评估规则对《导则》相应规则进行补充优化,形成区域定制化的状态评估规则,可以提高该地区变压器状态评估工作的准确性和针对性。补充优化内容包括在《导则》基础上补充本文导出的if-then形式的判断规则和根据提取的判断规则对《导则》中的状态量阈值进行优化。表5列出了状态量阈值优化结果。

表5 状态量阈值优化结果

Table 5 Optimized results of state parameter threshold

优化条目	阈值	
	优化前	优化后
本体绕组直流电阻各绕组相互间的差别/ Ω	0.02	0.013
本体绕组绝缘电阻/ $\text{M}\Omega$	—	9811.555
本体铁芯接地电流/A	0.3	0.178

注:“—”表示《导则》中未列出。

取该区域12条变压器数据,包含正常状态、注意状态、异常状态、严重状态4种状态的样本各3条,将采用本文提出的区域定制化状态评估规则进行评估结果和采用《导则》给出的方法(以下简称传统方法)得到的结果的正确率进行对比,发现采用传统方

法时评估正确次数是11、评估正确率是91.67%,而采用本文方法时评估正确次数是12、评估正确率达到100%。可以看出,本文所提出方法的正确率高于传统方法,其中传统方法对1条样本的判断出现了错误判断。在算例中样本的变压器型号为C,绕组直流电阻各相绕组相互间的差别为0.013 Ω ,而《导则》中规定该变压器绕组直流电阻各相绕组相互间的差别应小于0.02 Ω (2%),因此此项符合要求,不扣分;样本变压器绕组绝缘电阻为3 265.986 M Ω ,绕组绝缘电阻经查不符合相关规定,扣8分。最终得分为92分,判断此变压器处于正常状态。本文方法可以判断该变压器绕组直流电阻各相绕组相互间的差别不合格,且绕组绝缘电阻不合格,从而判断其处于异常状态,与实际评估结果一致。

2.5 模型性能对比

为了验证本文提出方法的SMOTE算法效果和决策树模型性能,分别以未经SMOTE算法处理的原始样本集与SMOTE算法处理后的样本集为数据基础,以及分别采用C4.5决策树算法与CART决策树算法构建决策树,比较各条件下的Kappa系数,结果如表6所示。

表6 不同样本集和算法的性能对比
Table 6 Performance comparison of different sample sets and algorithms

样本集	Kappa系数	
	C4.5	CART
原始样本集	0.9847	0.9637
SMOTE算法处理后样本集	0.9893	0.9822

对原始样本集使用SMOTE算法进行过采样后,C4.5、CART决策树算法所构建决策树模型的Kappa系数均有提升;C4.5决策树算法所构建决策树模型的Kappa系数在2种样本集条件下均领先于CART决策树算法。结果表明,本文采用SMOTE算法处理样本集可以有效提高决策树模型的性能,且C4.5决策树算法在本文样本集条件下的分类性能优于CART算法。

3 结论

本文提出了基于SMOTE和决策树算法的变压器状态评估知识获取方法。采用SMOTE算法解决了原始样本集非正常状态量数量较少的问题,减轻了变压器相关数据的类别不平衡现象,对提高决策树模型的性能起到了良好效果。基于C4.5算法构建决策树实现了变压器状态评估知识获取,可以帮助运检人员对不同信息系统中获取的数据进行提炼,从而自动获取知识,提取关键状态量和评价规则。对于变压器状态评估工作,在考虑当地设备的差异性的基础上,应该加强对关键状态量的监测,并

对《导则》进行补充优化,形成区域定制化的变压器状态评估规则。在未来变压器采集数据量增加、数据质量提高的情况下,本文方法可以实现知识的自我更新、优化,从而提高变压器状态评估知识获取工作的效益和质量。

附录见本刊网络版(<http://www.epae.cn>)。

参考文献:

- [1] 喇元,王红斌,陈忠东. 变压器状态检修技术的研究及应用[J]. 变压器,2013,50(8):61-65.
LA Yuan, WANG Hongbin, CHEN Zhongdong. Research and application of condition maintenance technology of transformer[J]. Transformer, 2013, 50(8): 61-65.
- [2] 郭小芳,刘爱军,樊景博. 知识获取方法及实现技术[J]. 陕西师范大学学报(自然科学版),2007(增刊2):187-189.
GUO Xiaofang, LIU Aijun, FAN Jingbo. Knowledge acquisition method and its actualization measures[J]. Journal of Shaanxi Normal University (Natural Science Edition), 2007 (Supplement 2): 187-189.
- [3] 于之虹,郭志忠. 数据挖掘与电力系统[J]. 电网技术,2001,25(8):58-62.
YU Zhihong, GUO Zhizhong. Data mining and power system[J]. Power System Technology, 2001, 25(8): 58-62.
- [4] 赵文清,朱永利,姜波,等. 基于贝叶斯网络的电力变压器状态评估[J]. 高电压技术,2008,34(5):1032-1039.
ZHAO Wenqing, ZHU Yongli, JIANG Bo, et al. Condition assessment for power transformers by Bayes networks[J]. High Voltage Engineering, 2008, 34(5): 1032-1039.
- [5] 张懿议,焦健,汪可,等. 基于帝国殖民竞争算法优化支持向量机的电力变压器故障诊断模型[J]. 电力自动化设备,2018,38(1):99-104.
ZHANG Yiyi, JIAO Jian, WANG Ke, et al. Power transformer fault diagnosis model based on support vector machine optimized by imperialist competitive algorithm[J]. Electric Power Automation Equipment, 2018, 38(1): 99-104.
- [6] 阮羚,谢齐家,高胜友,等. 神经网络和信息融合技术在变压器状态评估中的应用[J]. 高电压技术,2014,40(3):822-828.
RUAN Ling, XIE Qijia, GAO Shengyou, et al. Application of artificial neural network and information fusion technology in power transformer condition assessment[J]. High Voltage Engineering, 2014, 40(3): 822-828.
- [7] 刘云鹏,付浩川,许自强,等. 基于AdaBoost-RBF算法与DSmT的变压器故障诊断技术[J]. 电力自动化设备,2019,39(6):166-172.
LIU Yunpeng, FU Haochuan, XU Ziqiang, et al. Transformer fault diagnosis technology based on AdaBoost-RBF algorithm and DSmT[J]. Electric Power Automation Equipment, 2019, 39(6): 166-172.
- [8] 张晶. 基于决策树的知识获取方法研究[J]. 制造业自动化,2011,33(8):154-156.
ZHANG Jing. Based on the decision tree of knowledge acquisition method[J]. Manufacturing Automation, 2011, 33(8): 154-156.
- [9] 贾亚飞,朱永利,高佳程,等. 基于样本加权FCM聚类的未知类别局部放电信号识别[J]. 电力自动化设备,2018,38(12):107-112.
JIA Yafei, ZHU Yongli, GAO Jiacheng, et al. Recognition of unknown partial discharge signals based on sample-weighted FCM clustering[J]. Electric Power Automation Equipment, 2018, 38(12): 107-112.

- [10] 王守相,周凯,苏运. 基于随机森林算法的台区合理线损率估计方法[J]. 电力自动化设备,2017,37(11):39-45.
WANG Shouxiang,ZHOU Kai,SU Yun. Line loss rate estimation method of transformer district based on random forest algorithm[J]. Electric Power Automation Equipment,2017,37(11):39-45.
- [11] 叶志飞,文益民,吕宝粮. 不平衡分类问题研究综述[J]. 智能系统学报,2009,4(2):148-156.
YE Zhifei,WEN Yimin,LÜ Baoliang. A survey of imbalanced pattern classification problems[J]. CAAI Transactions on Intelligent Systems,2009,4(2):148-156.
- [12] 林智勇,郝志峰,杨晓伟. 不平衡数据分类的研究现状[J]. 计算机应用研究,2008,25(2):332-336.
LIN Zhiyong,HAO Zhifeng,YANG Xiaowei. Current state of research on imbalanced data sets classification learning[J]. Application Research of Computers,2008,25(2):332-336.
- [13] CHAWLA N V,BOWYER K W,HALL L O,et al. SMOTE: synthetic minority over-sampling technique[J]. Journal of Artificial Intelligence Research,2002,16(1):321-357.
- [14] BLAGUS R,LUSA L. SMOTE for high-dimensional class-imbalanced data[J]. BMC Bioinformatics,2013,14(1):1-16.
- [15] BLAGUS R,LUSA L. Evaluation of SMOTE for high dimensional class-imbalanced microarray data [C]//Proceedings of 2012 11th International Conference on Machine Learning and Applications. Boca Raton,FL,USA:ICMLA,2012:89-94.
- [16] HSSINA B,MERBOUHA A,EZZIKOURI H,et al. A comparative study of decision trees ID3 and C4.5 [C]//ACM international conference proceeding series. Porto,Portugal:IJACSA,2018:13-19.
- [17] WU X,KUMAR V,QUINLAN J R,et al. Top 10 algorithms in data mining[J]. Knowledge and Information Systems,2008,14(1):1-37.
- [18] 姜欣,徐六通,张雷. C4.5决策树展示算法的设计[J]. 计算机

工程与应用,2003,39(4):93-94.

JIANG Xin,XU Liutong,ZHANG Lei. The design of C4.5 decision tree presentation algorithm[J]. Computer Engineering and Applications,2003,39(4):93-94.

- [19] LANDIS J R,KOCH G G. An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers[J]. Biometrics,1977,33(2):363-374.

- [20] 王军. Kappa系数在一致性评价中的应用研究[D]. 成都:四川大学,2006.

WANG Jun. Application of Kappa coefficient in consistency evaluation[D]. Chengdu:Sichuan University,2006.

作者简介:



谢桦

谢桦(1970—),女,湖北荆州人,副教授,博士,主要研究方向为综合能源系统优化运行、储能规划与控制(E-mail: hxie@bjtu.edu.cn);

陈俊星(1995—),男,湖南衡阳人,硕士研究生,主要研究方向为基于数据挖掘技术的电力设备状态评估(E-mail: 17125975@bjtu.edu.cn);

赵宇明(1978—),男,吉林长春人,教授级高级工程师,博士,主要研究方向为电力系统和直流输电(E-mail: zhaoyim97@sina.com);

丁庆(1982—),男,湖北荆门人,高级工程师,博士,主要研究方向为新能源和电磁环境(E-mail: dingqing91@gmail.com);

张沛(1972—),男,北京人,教授,博士研究生导师,主要研究方向为电力系统规划和运行(E-mail: peizhang166@qq.com)。

(编辑 王欣竹)

Knowledge acquisition method of power transformer condition assessment based on SMOTE and decision tree algorithm

XIE Hua¹,CHEN Junxing¹,ZHAO Yuming²,DING Qing²,ZHANG Pei¹

(1. School of Electrical Engineering, Beijing Jiaotong University, Beijing 100044, China;

2. Shenzhen Power Supply Company, Shenzhen 518001, China)

Abstract: A knowledge acquisition method of power transformer condition assessment based on SMOTE (Synthetic Minority Over-sampling TEchnique) algorithm and decision tree algorithm is proposed. Firstly, for the case that samples in abnormal condition is much fewer than samples in normal condition, SMOTE algorithm is used to supplement the number of abnormal state samples to solve the imbalance problem of the transformer sample set. Then the transformer condition assessment is treated as a classification process, and the knowledge acquisition problem is transformed into a decision tree building problem using the characteristics of decision tree model as a white box model. C4.5 algorithm is applied to construct decision tree, then the transformer condition assessment knowledge and rules are derived from the decision tree. Case study is performed using 110 kV transformer data collected from a distribution company. Study results indicate that the proposed method can automatically acquire knowledge and rules and can provide decision support for future transformer condition assessment work.

Key words: power transformers; knowledge acquisition; decision tree algorithm; SMOTE; data mining

附录

表 A1 Kappa 系数取值范围与分类精度的关系

TableA1Relationship between Kappa statistical value range and its classification

Kappa 系数取值范围	分类精度	Kappa 系数取值范围	分类精度
[0, 0.02)	差	[0.41, 0.60)	中等
[0.02, 0.21)	较差	[0.61, 0.80)	好
[0.21, 0.40)	尚可	[0.81, 1]	非常好

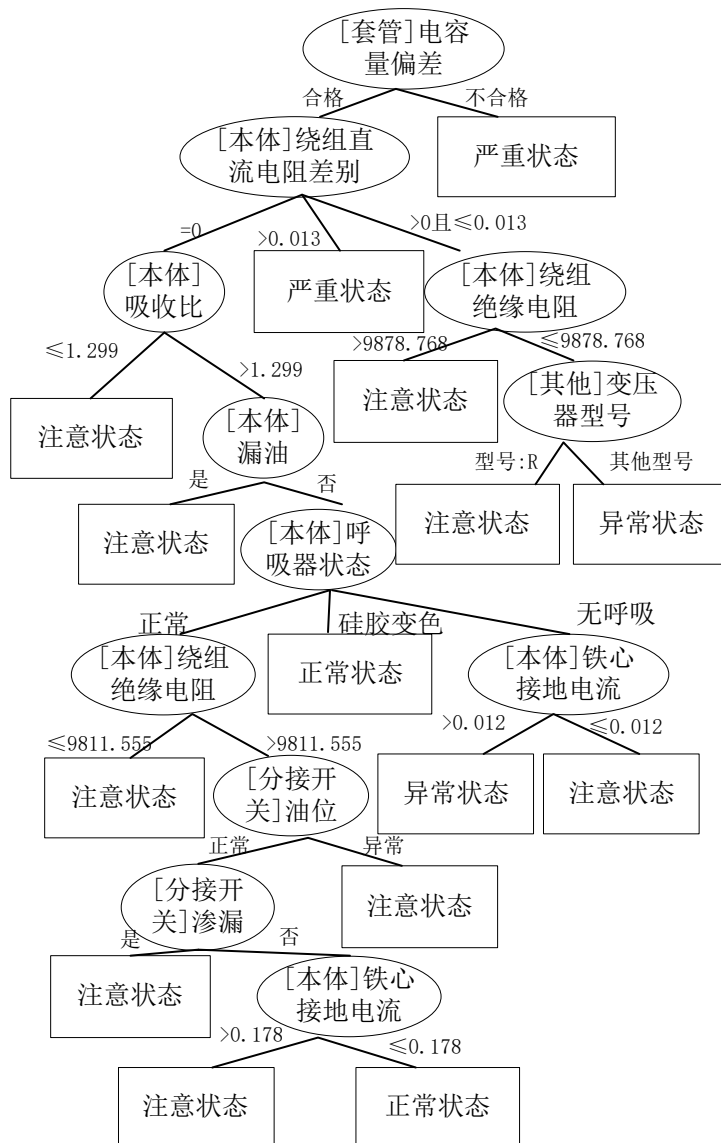


图 A1 变压器状态评估知识决策树

Fig.A1 Decision tree of transformer condition assessment knowledge