决策树技术在电缆绝缘状态评估中的应用

孙秋野，张化光，张铁岩
（东北大学信息科学与工程学院，沈阳 110004）

［摘要］ 电缆的绝缘状态通常可以分为良好、不好、差和故障等几种，以电缆的日常检修数据、试验数据和在线监测数据为基础，对电缆的状态进行判断是一个非常有意义的课题。采用决策树分类技术来对电缆的绝缘状态进行分类，分别对各种类型数据形成子树，然后通过子树合成技术形成最终的决策树，从而对电缆的绝缘状态进行判断。通过一个实际电缆的各种数据，采用SPSS软件进行实际应用，最终的仿真结果说明决策树技术是一种非常有效的电缆绝缘状态分类技术。

［关键词］ 决策树；分类；数据挖掘；电缆绝缘
［中图分类号］ TP206；TM247 ［文献标识码］ A ［文章编号］ 1009－1742(2010)02－0090－05

1 前言

在线运行的电缆绝缘状态通常与很多因素相关联，对电缆绝缘状态进行评估对于电缆的维护和使用是非常重要的，有利于电缆的寿命管理。更确切的说，也就是要准确地掌握电缆绝缘状态的好坏。

对于日常积累的大量的关于电缆运行和维护的历史数据，使用数据挖掘中的分类技术建立针对电缆绝缘状态的分类决策模型，并形成相应的分类规则，从而可以根据新采集的测量数据判断电缆绝缘的状态以及是否发生了故障。

决策树和决策规则是解决实际应用中分类问题的数据挖掘方法。一般来说，分类是把数据项映射到其中一个事先定义好的类中的过程。由一组输入的属性值向量和相应的类，用基于归纳学习算法得出分类。换句话说，分类是把某个不连续的标识值（类）分配给一个未标识的纪录的过程。分类器是一个在样本的其他属性已知的情况下预测另外一个属性（样本的类）的模型（分类的结果）。

可以采用多种技术作为分类技术，如判定树归纳、贝叶斯分类和神经网络等，笔者使用判定树[1]——决策树技术对电缆的运行数据进行分类，以对电缆的状态进行评估。

决策树是一个类似于流程图的树结构，其中每个内部节点表示一个属性上的测试，每个分支代表一个测试输出，而每个树叶节点代表类或类分布。树的最顶层节点是根节点。为了对未知的样本分类，样本的属性值在判定树上测试，路径由根节点到存放该样本测试的叶节点，判定树容易转换成分类规则。相对于其他数据挖掘方法，决策树的最大特点就是直观明了，易于理解。常用于对离散和连续属性进行预测性建模。

在电力系统、医疗行业以及金融业和零售行业，决策树都有着广泛的应用。特别是对于一些无法量化属性的分类，决策树方法有着较为明显的优势。决策树表示法是应用最广泛的逻辑方法。目前生成决策树方法的算法主要有3种[2~7]：CART算法、CHAID算法、C4.5算法。

不管是在线运行的电缆，还是处于检修状态的电缆，其绝缘状态大致可以分为良好（可以继续运行，不需要检修）、一般（需要引起注意，在监视下运

［收稿日期］ 2008－06－06；修回日期 2008－10－10
［基金项目］ 国家自然科学基金资助项目（60325311，60274017，60572070，60534010）；国家“八六三”重点课题（2006AA04Z183）；辽宁省科学基金（20022030）
［作者简介］ 孙秋野（1971－），男，辽宁沈阳市人，东北大学讲师，研究方向为配电系统分析，故障诊断；E－mail：sunqiuye@mail.neu.edu.cn

行)、不好(需要进行检修和处理)和差(需要更换新的电缆)。

2 训练数据集的建立

能够反映电缆绝缘状态的信息主要来源于电缆的状态监测数据和日常的试验、检修数据,在形成训练数据集的同时假设所采用的这些数据都是处于同一时间范围内的数据,并且假设所采用的数据都是经过预处理的,即不含孤立点和空缺值等数据。对于一根电缆来说,可以用电缆型号,故障类型,故障严重程度,检修效果,服役时间,绝缘状态等属性来描述其日常故障检修数据,其中故障类型可以是各种各样的;而故障的严重程度一般来说可以划分为三档,即严重、一般和轻微;检修效果也可以分为三类,好、一般和差;绝缘状态一般分为不好、一般、良好等,服役时间是指电缆发生故障时的服役时间。

电缆的日常检修数据主要是对电缆的历史状态进行了综合的评估,包括 tanδ、局部放电、交流耐压等,这种评估是结合专家经验来进行的,即绝缘状态是根据专家的经验等得到的。

对于电缆的离线试验来说,由于受外界的干扰较小,所以有确定的状态判定标准。电缆的日常试验数据通常是判断检修效果的标准,所以可获得电缆在线监测数据表格,如表 1 所示。

表 1 在线监测数据

Table 1 Real – time monitering data

序号	tanδ	泄露电流/μA	…	负载电流/kA	环境温度/℃	环境湿度	测量时间	绝缘状态
1	<0.2 %	>40 <20	…	21	12.4	0.2	2000 – 3 – 20	不好
2	>0.2 % <0.5 %	<20	…	12	15	0.4	2000 – 3 – 21	一般
3	>0.5 %	>40	…	23	16.7	0.3	2001 – 4 – 20	一般
4	<0.2 %	<5	…	31	17.8	0.6	2002 – 3 – 20	良好
…	…	…		…	…			…

电缆的在线监测数据没有确定的状态判定标准,要想对电缆的绝缘状态进行判断,只能是结合以往的经验和离线试验数据,虽然离线试验不是经常进行的,但是作为决策分类的训练数据集,应该选择那些能够根据经验或离线试验确定状态的数据集合,而且历史数据量一般来说是较大的,有充分的选择条件。

上述三种数据属性相互之间关联程度非常小,并且电缆的日常检修和试验数据量较少,而在线监测数据的量较大。对于在线监测数据,也是有些属性对电缆绝缘状态的影响较小,并且考虑大量的属性会增加决策树判断的工作量,因此笔者采用分类决策树的建立,然后再将不同的决策树分支进行合成。

3 判断决策树的建立

由于涉及到电缆绝缘状态的数据较多,所以分别针对电缆的日常检修数据和在线监测数据建立相应的决策树,然后将决策树进行合成。

3.1 子树生成

在分别针对日常检修数据和在线监测数据生成子树时,关键的问题是测试属性的选择,当根据最小信息熵的原则来选择测试属性时,有

$$\min \sum_{j=1}^{v} \frac{s_{ij} + \cdots + s_{mj}}{s} I(s_{ij}, \cdots, s_{mj}) \qquad (1)$$

式(1)中,m 为分类的数量,一般为 3 或 4;s 是数据样本总数;$j = 1,2,\cdots,v$ 为被选择属性的可能输出值,将整个样本集合划分为 v 个子集;而 s_{ij} 表示子集 j 中分类 i 的样本数($i = 1,2,\cdots,\mathrm{m}$);$I(s_{ij},\cdots,s_{mj})$ 表示此种划分的信息量,有

$$I(s_{ij},\cdots,s_{mj}) = -\sum_{i=1}^{m} p_{ij}\log_2(p_{ij}) \qquad (2)$$

式(2)中,$p_{ij} = \frac{s_{ij}}{|S_j|}$ 是 S_j 中的样本属于类 i 的概率。当 s,m 一定时,选择不同的测试属性,s_{ij} 是不同的,通常的做法是进行列举,将可能被作为测试属性的分别按照式(1) 计算,然后选择最小的。

3.2 子树合成

由电缆的日常检修数据生成的子树的结构如图 1 所示,而由在线监测数据所形成的决策树如图 2 所示(进行了一定的简化) 。

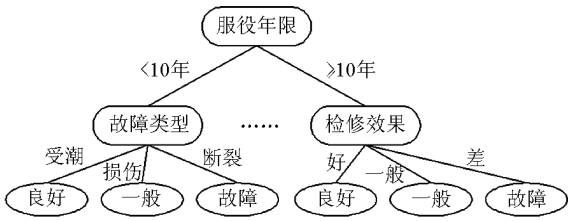

图 1　检修数据决策树

Fig. 1　Decision tree for examine and repair data

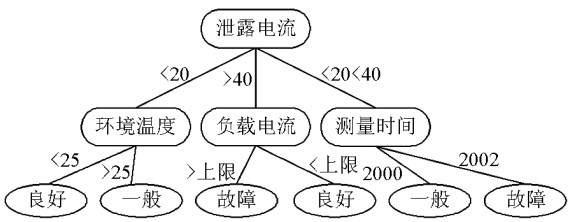

图 2　在线监测数据决策树

Fig. 2　Decision tree for monitoring data

两个子树合成一棵树，必须存在相同或者相类似的属性，否则，两棵子树可以作为单独的树存在，这也是子树合成的基本条件。在电缆的日常检修数据和在线监测数据当中，并没有相同的属性，但是存在相类似的属性，如测量时间与服役年限都属于时间范围内的属性，因此可以将在线监测数据当中的测量时间属性进行变换，即有

服役年限 = 测量时间 – 投入使用的年份　(3)

按照式(3)变换后，测量时间属性完全可以被服役年限属性所代替。将两棵子树在关联属性处相接，则形成一棵完整的树，如图 3 所示。

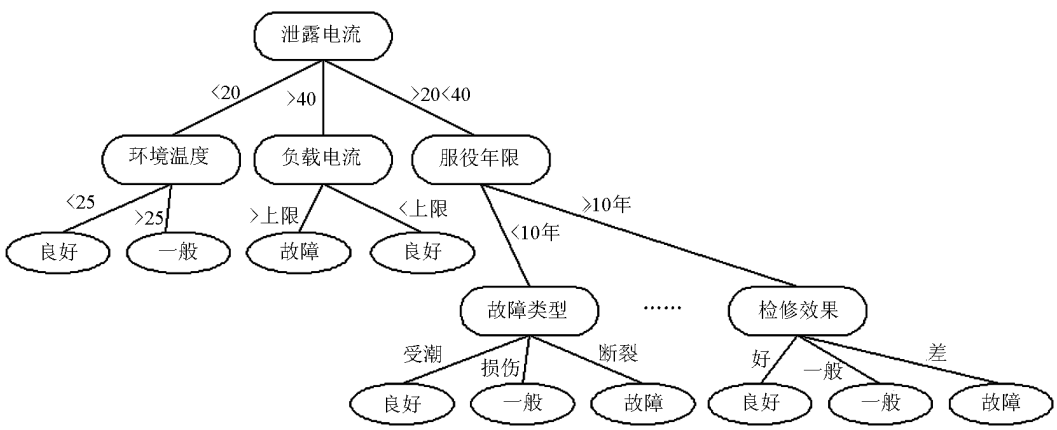

图 3　完整的决策树

Fig. 3　All decision tree

3.3　分类规则的产生

由树的各个分支即可以产生分类规则，分类规则的产生非常像 C 语言中的 If – Then 语句，从决策树中可以一目了然。

分类规则的产生可以作为检修和运行人员判断电缆故障以及状态好坏的依据，也可以提供正确使用电缆、防止事故发生的经验，总之，有利于电缆的寿命管理工作。

4　应用实例

在实际应用过程中，建立电缆的管理数据库是一个十分庞杂的工作，为了检验决策树技术在电缆状态评估中的应用，在文章中只是选取一个实际电缆的在线状态监测数据，通过 SPSS 软件的仿真来生成决策树，选取了比较直观的两个参数：tgδ 与泄漏电流。tgδ 的值往往反映的是普遍性的缺陷，个别集中的缺陷不会引起整根长电缆所测量的 tgδ 的显著变化。根据这些已知的测量数据，利用决策树进行分类，得出这两个主要参数与电缆绝缘状态之间的联系。

由于当前决策树在此领域的运用还处于空白状态，所以采用了另一种方法来证明决策树在这个领域使用的可行性：利用已知的试验标准生成一定的数据加上实测的运行数据，来模拟实际运行的状态，并使用决策树方法对这些数据进行分类，从而最终得出分类标准，最终将标准运用于运行电缆的状态评估。

目前的试验标准都已经非常明确，但对于实际

投入运行的电缆而言,这些标准并不能很好的起到标杆的作用,给实际判断带来了很大的难度。因此在过去经验公式的基础上引入新的数据挖掘的方法,藉此给该领域带来变革。

笔者利用 SPSS 软件提供的决策树分类功能进行分类。以绝缘状态为状态结果,泄漏电流和 tgδ 为独立变量,采用决策树方法进行分类。同时在 output 选项中生成分类规则,以便能够对将来的日常数据进行分类。

由于电力行业的特殊性,对安全性的要求格外高,因此在用决策树分类时就考虑到错误归类的代价必须根据行业特性进行更改。在这里,笔者将绝缘不良误判为绝缘良好的代价设为 3,结果如图 4 所示,基本上使得分类准确率达到了 99 % 以上。

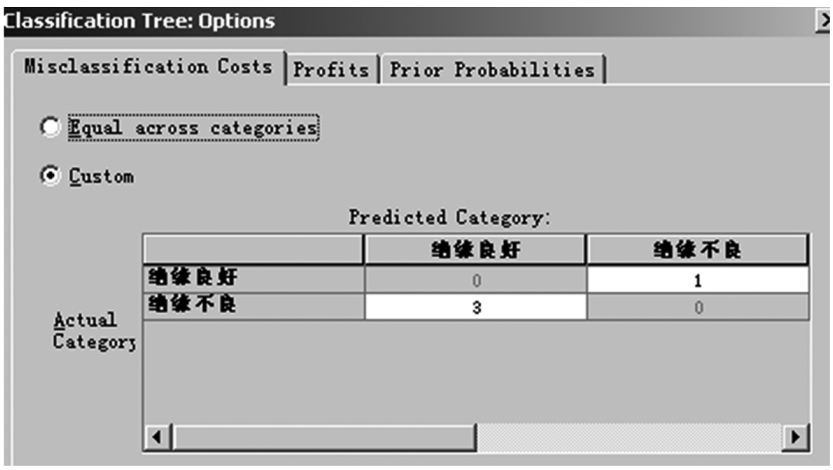

图 4　算法调整

Fig. 4　Algorithm correction

最终采用决策树分类的结果如表 2 所示。

最终有三片树叶 Node 2,Node 3 以及 Node 4,可见决策树方法能够较为准确地从数据中找出规律,泄漏电流以 98.78 mA 为分界,tgδ 以 0.99 % 为分界,基本符合之前的试验标准。计算结果如图 5 和图 6 所示。

表 2　分类结果

Table 2　Sort results

观测结果	预测结果		
	绝缘良好	绝缘不良	准确率
绝缘良好	115	2	98.3 %
绝缘不良	0	283	100.0 %
绝缘占比	28.8 %	71.3%	99.5 %

5　结语

对电缆的绝缘状态以及故障与否进行判断对于电缆的安全运行和寿命管理是非常重要的,笔者将决策树技术应用于电缆绝缘状态的判断中,所得到的结论如下:

1）电缆的在线和离线试验数据之间的关联性不是很大,所以可以单独形成决策子树;

2）对子树的合成必须有相关联的属性,这种

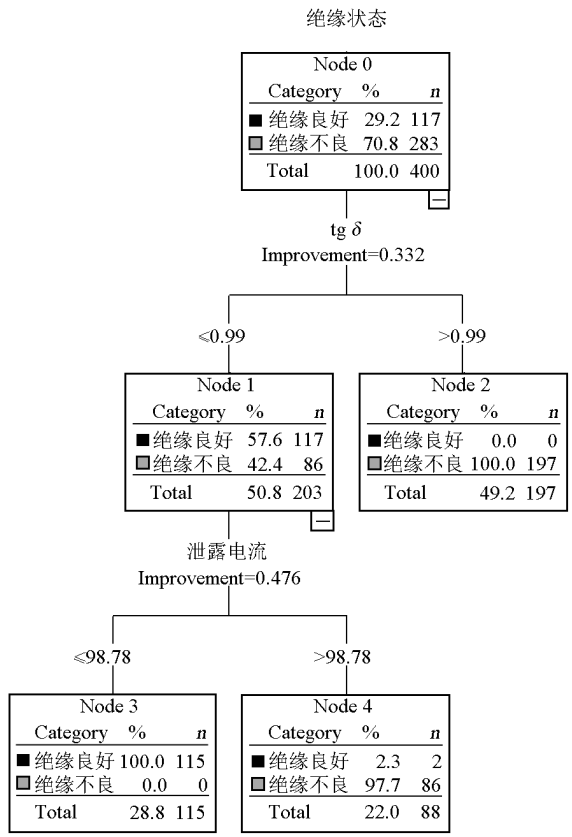

图 5　最终生成的决策树

Fig. 5　The final decision tree

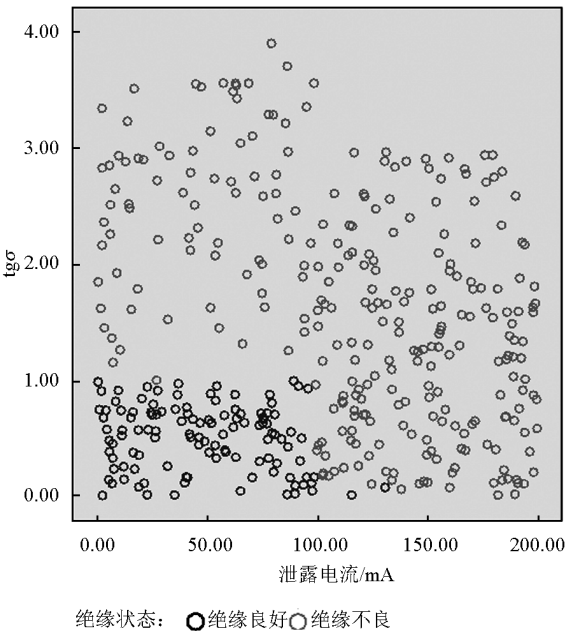

图 6　所有电缆监测数据反应的电缆绝缘状态分布图

Fig. 6　Condition distribution by cable monitering data

相关联的属性可以转化为一个可以结合两棵子树的公共属性；

3）对于电缆的在线监测数据，由于受运行环境的影响，并没有确切的判断电缆绝缘状态的标准，在形成分类规则时必须选取那些能够根据经验或离线试验确定状态的数据集合；

决策树技术作为数据挖掘中的一种典型分类技术，已经被广泛应用于各个领域，但是在电缆绝缘状态评估当中还是很少被应用，笔者将决策树技术应用于这一领域，取得了比较好的结果和值得借鉴的经验。

参考文献

[1] 朱绍文，胡宏银，王泉德，等. 决策树采集技术及发展趋势[J]. 计算机工程，2000，26(10)：61－64

[2] Quinlan J R. Induction of decision trees[J]. MachineLearning，1986，1：81－106

[3] Michalski R S，Larson J B. Selection of the most representative training examples and incremental generation of VL1 hypotheses[R]. Urbana－Champaign：Department of Computer Science，University of Illinois，1978

[4] Cestnik B，Kononenko I，Bratko I. ASSISTANT 86：a knowledge elicitation tool for sophisticated users[J]. Proceedings of EWSL－87，Bled，Yugoslavia，1987. 31－45

[5] Pagallo G，Haussler D. Boolean feature discovery in empirical learning[J]. Machine Learning，1990，5：71－99

[6] Brodley C E，Utgoff P E. Multivariate decision trees[J]. Machine Learning，1995，19：45－77

[7] 刘小虎，李　生. 决策树的优化算法[J]. 软件学报，1998，9(10)：797－800

The application of decision tree to the estimation for cable state

Sun Qiuye，Zhang Huaguang，Zhang Tieyan

(School of Information Science & Engineering，Northeastern University，Shenyang 110004，China)

[**Abstract**] The insulation state of cable can be split into well，bad，worse and fault. The estimation for cable state is a significant topic based on overhaul data，test data and monitor data of the cables. The decision tree is employed to classify the insulation state. The subtrees can be formed by all kinds of data，then the final decision tree is composed of the subtrees，by which the insulation state can be estimated. The application by SPSS with practical data of the cable is carried on. The simulation result for t(he insulation state estimation of cable shows the effectiveness of the approach.

[**Key words**] decision tree；classify；data mining；insulation of cable