粗糙集与决策树理论在输电线路巡检中的应用①

段其昌，程有富

(重庆大学 自动化学院，重庆 400030)

摘 要：针对输电线路巡检需求，为了提高巡检的效率，充分考虑了影响输电线路运行状态的各种因素，采用基于粗糙集理论与决策树的数据挖掘方法，建立了输电线路运行状态预测模型，将从输电线路历史异常数据中提取的规则应用于输电线路异常预测，根据预测结果可以制定出科学合理的巡检计划，经过测试样本的验证，该模型有较高的准确率。

关键词：输电线路巡检；数据挖掘；决策树；粗糙集

Application of Rough Set and Decision Tree Theory to Inspection of Power Transmission Line

DUAN Qi-Chang, CHENG You-Fu

(College of Automation, Chongqing University, Chongqing 400044, China)

Abstract: As the inspection demands of the transmission, it is necessary to enhance the efficiency of power inspection. Taking all factors which have significant effects on the running status of the transmission lines into consideration, this system establishes a state prediction model of transmission line based on rough set theory and decision tree. The rules which extract from the model apply in the abnormity prediction of the transmission lines. The manager can work out a scientific and reasonable inspection plan in accordance with these rules. This model is proven and have a higher accuracy.

Keywords: transmission lines inspection; data mining; decision tree; rough set

1 引言

输电线路巡检是为确保电力线路正常运行而采取的检查措施，主要是通过制定和执行固定和临时的巡检计划来实现。

其中，固定的定期巡检的不足之处在于巡检计划一旦制定，将严格执行，这样就忽略了特殊突发事件对输电线路的影响，不能及时处理和防范突发自然灾害对输电线路造成的损害。因此电力巡检部门在定期巡检计划进行的同时，常常辅助开展临时巡检工作。但是这种巡检机制并不科学，其适用范围只是在巡检线路出现重大险情后才开展，因此也不能有效的预先检查出输电线路的故障。

当前很多电力公司都采取了信息化的管理方式，在电力巡检管理系统中积累了大量的设备运行状态数据，所以，我们可以充分利用这些历史运行数据，采用数据挖掘技术实现对输电线路未来一段时间可能出现的异常运行情况进行预测，以此来制定科学的巡检计划。

针对这个问题，参考文献一提出了一种基于巡检知识的线路巡检智能决策方案[1],该方案的基本原理是在一定推理策略的控制下，利用知识库中的规则对事实数据进行匹配并获得结论的过程，匹配过程是通过将一个巡检事实与知识化的历史巡检记录逐个匹配的方法来得到对应的巡检事实属性值，然后再与上层知识库进行匹配，最后与第三层知识库进行匹配，直至找出最终的巡检决策方案。从该方案的基本原理来看，由于匹配过程较为复杂，所以匹配效率还有一定的提升空间。

① 收稿时间:2010-09-14;收到修改稿时间:2010-11-03

本文提出一种基于粗糙集—决策树（RS-DT）的输电线路异常预测方法[2]，将可能引发输电线路设备异常的各种因素作为条件属性，线路设备的运行状态（正常/异常）作为决策属性构成决策表，然后采用粗糙集理论进行属性约简，降低向量空间维数，减少特征数，从而提高分类速度，然后采用 C5.0 决策树算法构建决策树，提取规则，用来对输电线路设备的运行情况进行预测。

2　粗糙集理论与决策树简介

2.1 粗糙集

粗糙集理论是研究不完整、不确定知识和数据的表达、学习、归纳的有效方法,特别是在处理大数据量、消除冗余信息等方面具有一定的优势, 因此广泛应用于数据挖掘的数据预处理、 属性约简等方面。

粗糙集属性约简指在保持知识库分类能力不变的条件下，删除其中不相关或不重要的属性，从而简化原有的系统。一般来说，描述对象特征的属性集是较大的，但是对于信息系统分类的知识发现来说，不同属性的重要程度是不同的。有些是绝对不必要的，去掉这种属性并不影响分类的知识发现；还有一些属性是相对必要的，去掉这种属性必然会影响分类的知识发现。属性约简就是要在属性集中去掉所有不必要的属性，得到一个最小的属性集，它能完全确定知识发现，也即由这个最小的属性集确定的分类知识与由全体属性集确定的分类知识是相同的[3]。

2.2 决策树基本概念

决策树是应用最广的分类算法之一，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墒的测试，每个分支代表一个测试输出，每个树叶节点代表类或者类分布，决策树通过把实例从根节点排列到某个叶子节点来实现对实例的分类，树上的每个节点说明了对实例的某个属性的测试，节点的后继分支对应于该属性的一个可能值，要构造决策树，需要一个叫做训练样本集合的数据集作为输入，训练集由一组数据记录组成，每条记录由若干特征属性和一个用作决策的类别属性组成，一个样本可以表示为 $(v_{1,}v_{2,}...v_{n,}c)$，其中 v_i 表示特征属性，c 表示决策属性。决策树的生成是一个从根节点开始、从上到下的递归过程，一般根据分而治之的思想，通过不断的将训练样本分割成子集来构造决策树[4]。

3　线路运行状态预测模型的构建

构建预测模型要经过以下步骤：构建原始样本数据集；对输电线路设备运行状态样本集进行数据预处理，包括数据清理、数据转换和离散化等步骤，形成初步的数据集合；然后使用粗糙集相关算法进行属性约简，约去不必要的条件属性，形成新的数据集；最后把形成的数据集作为决策树构建算法的输入，形成决策树，最后提取决策规则，用来对将来的设备运行情况进行预测。方案如图 1 所示：

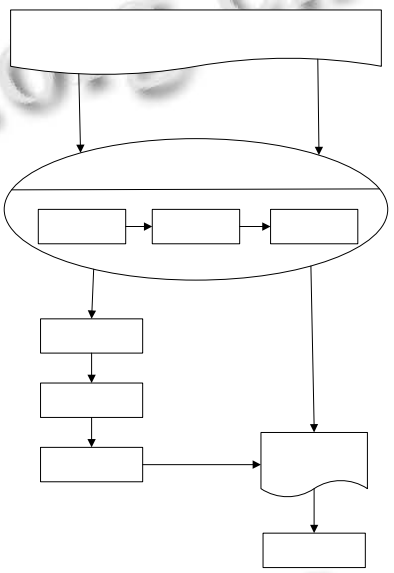

图 1　线路运行状态预测模型图

3.1 构建原始样本数据集

为构造原始样本集合，首先要取得实际的线路运行数据，为此从四川省和重庆市的几个地方电业局获得了各单位管理的部分输电线路的相关信息，包括近几年历史运行异常信息和输电线路属性信息，用来构建决策树模型并用作验证数据。

3.2 线路运行状态决策表

构建输电线路运行状态样本集不仅需要考虑具体线路本身的因素，包括线路等级、线路位置、线路材料等，还要考虑线路所处的周围环境，包括自然因素（天气状况，地质状况）、人为因素（人为破坏）的影响。不同时期、不同地点、不同位置的线路所面临的突发事件各不相同，因此，要综合考虑各种因素对线路运行的影响[1,5]。

经过对原始数据中各输电线路设备运行情况及运行环境的分析，最终选取了对输电线路设备的运行情

况影响较为突出的几项因素，包括：巡检线路所在地区天气因素和历史气候因素（包括平均温度、平均湿度和极端天气发生频率）、巡检线路所在地区地质因素、巡检线路材料、巡检线路上次巡检距现在的巡检天数与该巡检线路的巡检周期的比值、巡检线路役龄与线路运行年限的比值、线路设备历史异常情况、线路当前用电负荷情况和历史用电负荷情况，十一项巡检因素综合考虑，对于可能影响输电线路设备运行状态的人类活动因素纳入地质条件的范围来考虑，对于这点将在下面说明。

由于获得的原始数据中关于输电线路运行状态的记录都是线路异常记录，为满足数据挖掘需要，必须要有足够的输电线路正常运行记录，而实际中输电线路在大部分的运行时间内都处于正常运行状态，所以，我们只需要选取各输电线路运行历史中的几个时间点，将各时间点内正常运行的输电线路信息记录下来，作为数据挖掘原始数据中的一部分，对于时间点的选取，为了突出与运行异常线路的对比，将每一个线路发生运行异常的时间作为时间点，选取该线路发生异常时同地区仍然正常运行的输电线路作为数据集中的正常记录。

根据粗糙集对信息系统的定义，将所有的线路运行状态数据的集合作为论域，每条状态记录作为论域的对象，对线路运行有影响的因素集合作为条件属性集，线路设备运行是否正常作为决策属性，这样形成的决策表构成了决策树的训练集合，表中每一行代表一条线路设备运行状态记录，每列代表一个条件属性，用 C_1 到 C_{11} 表示，D 代表决策属性，也就是设备运行状态。示例如表 1 所示：

表 1 初始决策

C_1	C_2	C_3	C_4	…	C_{11}	D
晴	20	84%	5	…	11	正常
大风	18.3	70%	6	…	16	正常
…	…	…	…	…	…	…
覆冰	20.6	81%	1		31	异常

C_1：巡检线路所在地区天气情况；

C_2：巡检线路所在地区近几年平均温度，单位摄氏度；

C_3：巡检线路所在地区近几年平均湿度；

C_4：巡检线路所在地区近几年极端天气发生频率，用发生次数表示；

C_5：巡检线路所在地区地质因素，将地质条件分为复杂、简单、中等三级；

C_6：巡检线路使用材料等级，分为高级与一般两类；

C_7：巡检线路上次巡检距现在的巡检天数与该巡检线路的巡检周期的比值；

C_8：巡检线路役龄与线路运行年限的比值；

…

C_{11}：历史高峰负荷持续时间，单位为天；

D：决策属性，表征线路设备运行状态（正常 or 异常）。

最终的决策表中共有 372 条记录，其中异常记录有 189 条，占总记录数的 50.81%，正常记录为 183 条，占总记录数的 49.19%。

3.3 决策表数据预处理

数据预处理是数据挖掘过程中的一个重要步骤，尤其是对包含有噪声、不完整，甚至是不一致数据进行数据挖掘时，更需要进行数据预处理，以提高数据挖掘对象的质量，并最终达到提高数据挖掘所获模式的知识质量的目的。

对于本系统来说，不同指标（变量因素）一般都有不同的量纲，并且有不同的数量级单位，为了不同量纲、不同数量级的数据能放在一起比较，通常需对数据进行变换处理。

① 巡检线路设备所处地区天气因素的处理

本文以国家气象中心于 2007 年 9 月 10 日发布的《灾害性天气个列入库标准》作为气象等级划分依据，结合西南地区重庆和四川的气候特征，对影响输电线路的主要多发气候进行量化处理。最低量化结果为 1，最高量化结果为 10。天气因素量化结果如表 2 所示，并且将表中量化等级在“大风”以上(包括“大风”)的列为“极端天气”。

表 2 天气因素量化表

晴/阴天	小雨/雪	中雨/雪	大风	大雨/雪	暴雨/雪	覆冰	特大雨/雪	雷电
1	2	3	4	5	8	9	10	10

② 巡检线路所在地区地质因素

巡检线路所在地区地质因素的影响对巡检线路的影响也是十分重要的，表 3 为地质环境条件复杂程度分类表，表中复杂项每项评分为 10，中等项每项评分

为 5，简单项每项评分为 1，总分即为各项评分之和（总分处于 5 到 50 之间），从而实现对地质因素的量化。巡检计划制定者可以方便的对照表 3，对不同电力设备所处地质环境评分，录入数据库，方便数据的读取与使用。对于每一类的最后一项：破坏地质环境的人类工程活动，因为其不可预知性与突发性，可以随时更改保证数据的实时性。

表 3　地质环境条件复杂程度分类表

复杂	中等	简单
地质灾害发育强烈	地质灾害发育中等	地质灾害一般不发育
地形与地貌类型复杂	地形较简单，地貌类型单一	地形简单，地貌类型单一
地质构造复杂，岩性岩相变化大，岩土体工程地质性质差	地质构造较复杂，岩性岩相不稳定，岩土体工程地质性质较差	地质、构造简单，岩性单一，岩土体工程地质性质良好
工程地质、水文地质条件差	工程地质、水文地质条件较差	工程地质、水文地质条件良好
破坏地质环境的人类工程活动强烈	破坏地质环境的人类工程活动较强烈	破坏地质环境的人类工程活动　一般

③ 其他属性的处理

对于类似平均温度、平均湿度以及其他用小数表示的属性值，首先要映射到区间内，并四舍五入整数方便处理。线路所用材料可分为高级与一般两种，分别用数字 1、2 表示。线路当前是否用电高峰用数字 0、1 表示，线路设备历史异常情况用该设备每年发生异常的次数来表征。

3.4 决策表粗糙集属性约简

在构建了属性模型与决策表以后，需要对条件属性集进行约简，本文使用了基于 Pawlak 属性重要度的属性约简算法对决策表进行条件属性的约简[3,6]。

根据粗糙集理论，在前文提出的决策表中不同属性的重要度并不相同，如果在决策表中删除某个属性后，决策表的分类能力变化越大，说明该属性就越重要，反过来，变化越小重要性就越小。属性重要性的定义如下：

定义 1(属性的重要度). 设给定一个信息系统 $IS=(U,C,V,f),\forall B\subseteq C$，以及 $\forall\alpha\in C-B$，定义 $sig(\alpha,B;C)=$

$$\frac{card(U/IND(B\bigcup\{\alpha\}))-card(U/IND(B))}{card(U)}$$

α 为属性对属性集 B 的重要度。式中 $card(U)$ 运算表示有限集合 U 中的元素个数。

对决策表进行属性约简处理后发现历史平均温度与平均湿度对系统决策结果基本没有影响，属性重要度最小，可以约去，由剩下的属性构成新的决策表。

3.5 构建决策树

常用的决策树算法有 ID3、CHAID、CART、 Quest 和 C4.5 等，我们使用 SPSS Clementine12.0 建立挖掘模型，所以使用 SPSS Clementine 中的决策树模型 C5.0。 C5.0 算法是 C4.5 算法应用于大数据集上的分类算法，主要在执行效率和内存使用方面进行了改进。C5.0 模型根据能够带来最大信息增益（information gain）的字段拆分样本。第一次拆分确定的样本子集随后再次拆分，通常是根据另一个字段进行拆分，这一过程重复进行直到样本子集不能再被拆分为止。最后，重新检验最低层次的拆分，那些对模型值没有显著贡献的样本子集被剔除或者修剪。

我们首先将决策数据分为两部分，从中随机抽取 70%作为训练集，剩下的 30%用做预测验证，然后开始建模构建决策树。我们使用 SPSS Clementine 建立了如图 2 所示模型：生成的决策树如图 3 所示。

决策树的分析结果表明，天气因素是影响电力设备工作状态的最主要因素，这与实际情况也是相符的，从决策树中导出的决策规则可以直接应用于今后的输电线路设备的运行状态预测。

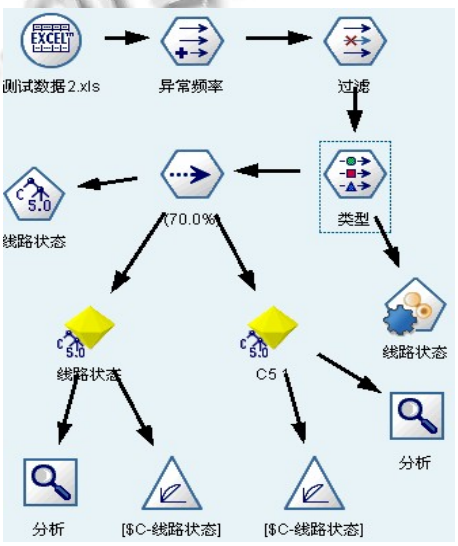

图 2　输电线路运行状态预测模型图

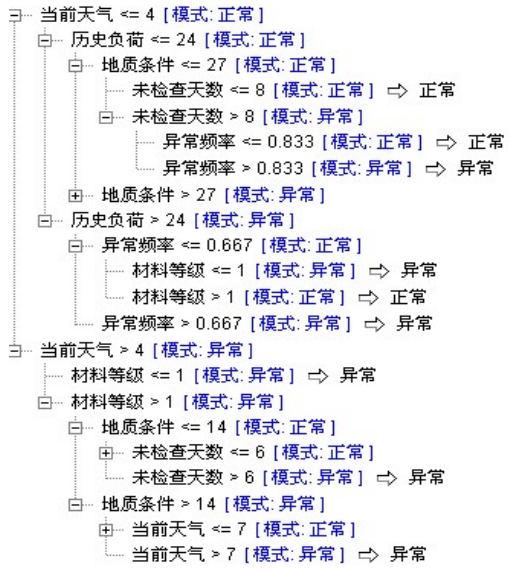

图 3 生成的部分决策树

3 模型验证与评价

最后要对生成的模型进行验证，由于电力系统设备在实际运行中出现问题的几率并不是很大，所以本系统在实际中短时间的应用也无法获得大量的验证数据，所以我们用原数据中剩下的 30%作检验集来模拟实际的验证效果。

将剩下的 30%数据输入决策树模型，然后加入分析节点与评估节点进行验证，得到的结果如图 4 所示：

比较 $C-线路状态 与 线路状态

正确	87	85.29%
错误	15	14.71%
总计	102	

$C-线路状态 的重合矩阵（行表示实际值）

	异常	正常
异常	49	4
正常	11	38

图 4 预测结果统计图

图中$C-线路状态指的是模型预测值，如图 8 所示，通过与实际值的比较，在占总记录数 30%的 102 条记录中，有 87 条记录预测准确，预测准确率有 85.29%，其中，53 条异常记录中有 49 条预测正确，准确率为 92.45%；49 条正常记录中，有 38 条预测正确，准确率为 77.55%。

图 5 是模型累积增益评估图，累积增益被定义为在每个分位点上的成功总数的一定百分比。它是用公式“(百分点位上的成功数/总的成功数)*100%”来计算的，对于一个好的模型来说，增益图表将陡峭地升高到 100%然后变得平缓，如图 5 中的$BEST-线路状态所示。从图中可以看出，系统有较好的预测性能，基本可以满足预测要求。

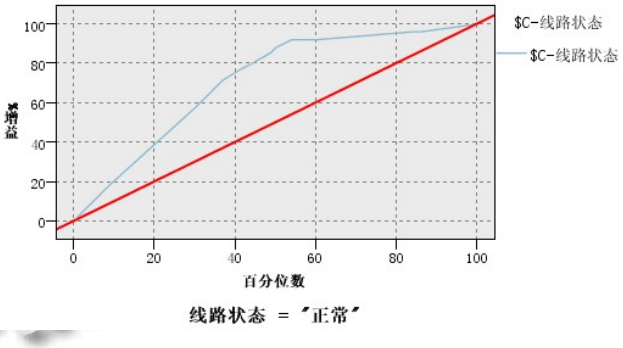

图 5 模型累积增益评估图

4 模型改进方法研究

在初始模型的预测结果(图 4)的基础上，我们将线路异常被误判成线路正常的错误称为 A 类错误，线路正常被误判为线路异常的错误称为 B 类错误，对于我们的系统来说，A 类错误要比 B 类错误严重的多，所以，在保持较高的总体预测准确率的前提下，应该尽量降低 A 类错误发生的几率。

误判成本值是C5.0中对于某一种误判所产生后果的严重性的反映，值越高，说明误判的后果越严重，这种错误在模型的构建过程中应该得到重视以减少该类误判。

C5.0 允许为每一种误判设定成本值，然后根据这些成本值以最小化期望误判成本总和为目标生成决策树，通过这种方法生成的决策树称为 Cost-sensitive tree。

我们将 A 类错误的成本值称为 COST(A)，B 类错误的成本值称为 COST(B)，采取多次试验比较结果的方法确定最佳的成本值，即在模型其他参数不变的情况下，不断提高 COST(A)的值，进行多次试验，建立不同的决策树模型，经过比较预测结果来挑选 COST(A)的最佳值。挑选的标准如下：

1) 对于测试样本，总错误率不能明显高于 COST(A)取其他值时的模型，并且总体错误率越低越好；2) 在满足第一条标准的前提下，A 类错误越低越好。

根据以上标准，我们经过试验，得到了 COST(A)

取不同值时的分类错误率如表 4 所示：

表 4　COST（A）取不同值时模型预测准确率

COST（A）	1.0	1.2	1.4	1.6	1.8
总错误率（%）	14.71	15.69	17.64	20.58	21.56
A 类错误率（%）	7.54	5.66	3.77	3.77	1.89
B 类错误率（%）	22.44	26.53	32.65	38.78	42.86

从表 4 可以看出，随着 COST(A)的不断增大，模型的预测总错误率是不断上升的，而 A 类错误率则呈下降趋势。接下来我们通过分析比较，找出最佳的 COST(A)：

1) 当 COST（A）>1.4 时，总错误率超过了 20%，这样的错误率太高了，显然不符合我们的选择标准。

2) 通过比较 COST(A)分别为 1.2 和 1.4 的模型，发现当 COST(A)等于 1.4 时，不仅总错误率相对较小，而且 A 类错误率也是最小的，因此，选择 COST(A)为 1.4 是相对比较合适的。

参数调整后决策树的预测结果如表 5 所示：

表 5　改进后模型预测结果

		预测类别		合计	错误率(%)	正确率(%)
		正常	异常			
实际类别	正常	33	16	49	32.65	67.35
	异常	2	51	53	3.77	96.23

改进后的模型与初始模型预测结果错误率对比表如表 6 所示：

表 6　改进模型与初始模型预测错误率对比

	初始模型	改进模型	差额
总错误率（%）	14.71	17.64	-2.93
A 类错误率(%)	7.55	3.77	3.78
B 类错误率(%)	22.45	32.65	-10.2

表 6 可以看出，模型改进后，在总错误率稍有增加的情况下降低了 A 类错误率，是异常预测准确率达到 96.23%，这种结果与我们的建模目标是吻合的。

5　结语

输电线路巡检工作需要有一个科学可靠的智能决策系统来辅助巡检工作的实施，本文尝试将粗糙集理论与 C5.0 决策树算法应用在输电线路工作状态预测中，初步取得了较为满意的效果，有较大的研究价值和实用价值，对于输电线路巡检计划的制定有较强的指导意义。本文采用决策树算法形成匹配规则，相对于参考文献一中的知识匹配方法，匹配效率有了较大的提升。

参考文献

1 徐家明,张树友,万昌江,傅顺强.基于知识的线路巡检计划层次推理技术研究.计算机工程,2003,29(19):45–47.

2 吴成东,许可.基于粗糙集与决策树的数据挖掘方法.东北大学学报,2006,(5):481–484.

3 Pawlak Z. Rough sets. International Journal of Computer and Information Science, 1982,11(5):341–356.Ernesto Vazquez.

4 Quinlan JR. C4.5: Programs for Machine Learning. San Mateo, CA: Morgan-Kaufmann Publishers, 1993:34–106.

5 An online expert system for fault section diagnosis in power systems. IEEE Trans. on Power Systems, 1997,12(1): 357–362.

6 Yao YY. On generalizing pawlak approximation operators. Lecture Notes in Artificial Intelligence 1998,1424:1998, 298–307.

(上接第 64 页)

2001,8(5):390–393.

3 陈卫东,杨绍全.利用累量不变量对MPSK信号分类.西安电子科技大学学报,2002,29(2):229–232.

4 一种基于高阶累积量的数字调相信号识别方法.系统工程与电子技术,2008,30(9):1611–1615.

5 张贤达.现代信号处理.第 2 版.北京:清华大学出版社,2002.

6 包锡锐,吴瑛,周欣.基于高阶累积量的数字调制信号识别算法.信息工程大学学报,2007,8(4):463–467.

7 陈筱倩,王宏远.基于联合特征向量的自动数字调制识别算法.计算机应用研究,2009,26(7):2478–2480.

8 郭双冰.基于小波和分形理论的调制信号特征提取方法研究.信号处理,2005,21(3):316–318.

9 吕铁军,郭双冰,肖先赐.基于复杂度特征的调制信号识别.通信学报,2002,23(1):111–115.