

基于 PCA 和粗糙集构建决策树的变电站故障诊断

张延松, 赵英凯

(南京工业大学自动化与电气工程学院, 江苏 南京 210009)

摘要: 提出一种基于主元分析 (PCA) 和粗糙集理论结合继而构建决策树的故障诊断方法。该方法利用 PCA 对原始故障决策表的条件属性集进行降维处理, 得到由主元变量构成的故障决策表, 采用等频分割方法对这一决策表的数据离散化, 进而采用基于主元属性重要度的粗糙集属性约简算法得到离散后的决策表的最小约简, 以约简数据集为样本基于核属性采用一种改进的决策树算法训练学习, 构建故障决策树进行诊断决策。测试实例证明了该方法能简化故障诊断系统, 提取容错性较强的诊断规则, 提高了故障的识别率。

关键词: 主元分析; 粗糙集; 决策树; 变电站; 故障诊断

Fault diagnosis of substation by the constructed decision tree based on principal component analysis(PCA) and rough set

ZHANG Yan-song, ZHAO Ying-kai

(Automation and Electrical Engineering College, Nanjing University of Technology, Nanjing 210009, China)

Abstract: A method for substation fault diagnosis based on principal component analysis(PCA) and rough set theory, then to constructed decision tree, is proposed. By this method, PCA is used to decrease the dimension of all the condition attributes of the original fault decision table and get fault decision table that consists of principle component variables. Then, an equal-frequency deviation method is used to discrete the value of the above decision table. The next is that a rough attribute reduction based on the significance of principle component variables is applied to obtain a minimum reduction of the discrete decision table. Finally, based on the core attributes, an improved decision tree algorithm is used to train the reduced data set and construct a decision tree for the diagnosis. Test proved that this method can simplify the fault diagnosis system, extract the diagnosis rules of better fault tolerance and increase the fault recognition rate.

Key words: PCA; rough set; decision tree; substation; fault diagnosis

中图分类号: TM711 文献标识码: A 文章编号: 1674-3415(2010)14-0104-06

0 引言

变电站是电力输配电系统中非常重要的一环, 涉及到许多安全运行、可靠供电方面的问题。变电站故障诊断就是从变电站的某些检测量中得到故障征兆信息, 再通过对这些故障征兆信息的分析与处理, 判断出故障源的位置。其中检测量由厂站监控系统 and 故障录波器提供, 故障征兆包括保护开关动作、断路器跳闸等, 而需判断的故障源通常有母线、变压器、无功补偿设备等^[1]。

多元统计理论中的主元分析法可以提取样本集的主元, 降低样本的维数, 甚至可以实现样本的最优压缩^[2]。粗糙集理论^[3]是由 z. Pawlak 于 1982 年提出的一种处理不精确和不确定知识的数学工具, 其主要思想就是在保持分类能力不变的前提下, 通过知识约简, 将决策系统简化, 提高系统潜在知

识的清晰度。决策树算法是分类发现算法中最常见的一种方法^[4], 在故障诊断的决策信息规则提取中有着广泛的应用。

主元分析能将高维的故障样本空间投影到相对独立的低维空间, 以降低分析难度, 但不适合解决非线性^[5-6]、时变动态^[7]、故障隔离^[8]等问题, 单独使用时需要对其进行改进。粗糙集理论被用于电力系统故障诊断时, 使用不同的属性约简^[9-11]和值约简方法可能会得到不尽一致的故障诊断规则, 势必对故障诊断结果的正确性造成影响; 目前属性约简方法很多, 但值约简方法研究较少^[12-14], 这些也给约简方法的选取和使用带来困难, 所以仅依靠粗糙集理论的故障诊断方法优势不是很明显。决策树具有对无次序的实例进行快速分类并记忆的能力, 其产生的规则是树状结构, 规则间关系清晰, 使得推理的结果便于解释, 但难点在于分支节点的选择

上, 选择的好坏关系到计算的复杂度和决策规则的质量, 文献[15-16]分别以信息熵降大小和信息增益比高低为依据确定分支节点, 但存在计算复杂、效率不高的问题。本文吸收了 PCA 在特征提取和数据降维的优点, 以及粗糙集理论能进行知识约简的优势, 在尽可能少丢失信息的前提下浓缩故障样本, 构造出易操作、复杂度小、直观性强的决策树进行故障诊断。

1 基于PCA的故障特征提取

根据PCA原理, 基于PCA从故障决策表中提取故障信号特征的算法(算法1)如下:

(1) 设故障信号的采样点为 r 个, 也就是决策表的条件属性个数为 r 个, 将其排列成一维行向量, 可以表示成: $x = (a_1, a_2, \dots, a_r)$ 。故障信号样本总数为 N , 每个故障样本均为 r 维, 分别对应行向量 x_1, x_2, \dots, x_N , 构成 N 行 r 列的故障数据矩阵 X 。

(2) 计算全体故障样本的平均值向量

$$m = \frac{1}{N} \sum_{i=1}^N x_i$$

(3) 计算协方差矩阵

$$C_x = \frac{1}{N} \sum_{i=1}^N (x_i - m)^T (x_i - m)$$

(4) 求协方差矩阵 C_x 的特征值 λ_i 及相应的正交归一化特征向量 u_i , 其中 $i = 1, 2, \dots, r$;

(5) 将特征值按由大到小顺序排列, 并按照下式计算前 p 个主元的累积贡献率:

$$\eta(p) = \sum_{i=1}^p \lambda_i / \sum_{i=1}^r \lambda_i, p < r$$

累积贡献率用于衡量新生成分量对原始数据的信息保存程度, 通常要求其大于85%即可;

(6) 取前 p 个较大特征值对应的特征向量构成变换矩阵 T : $T = (u_1, u_2, \dots, u_p)$, $p < r$;

(7) 通过 $Y = XT$ 计算前 p 个主成分, 达到降低维数的目的, 将每种故障信号样本的矢量向步骤

(6)中的特征矢量 T 张成的子空间投影, 就可提取故障信号的 p 维特征矢量。

2 基于粗糙集的故障决策表约简

主元分析方法将原始故障样本决策表浓缩成由主元变量 Y 构成的新故障样本表, 表中的数据信息是经过投影后的连续型数据, 在用粗糙集理论解决连续系统的问题时, 要求信息表必须是离散值, 所以必须通过离散化方法, 将连续型数据转化成离

散化数据。离散化方法很多, 本文采用等频率离散化方法。设主元变量故障决策表为: $DT = (U, A = C \cup D, V, f)$, 其中: U 为样本集序号, C 为条件属性集, $C = Y = \{y_1, y_2, \dots, y_p\}$, D 为决策属性集, $V = \{a_1, a_2, \dots, a_p\}$ 为条件属性值集合, f 为相应的决策属性值。等频法思想如下^[17]:

在 y_i 上有 N 个离散值, 给定一个参数 k , 把这 N 个离散值分成 $k+1$ 段, 每段有 $N/(k+1)$ 个离散值, 则断点 c_j^i 的确定方法如下: 对于条件属性 y_i , 以值升序集合表示它的值域, 即:

$$\begin{cases} y_i(U) = \{y_1^i, \dots, y_m^i, y_{m+1}^i, \dots, y_N^i\} \text{ 设断点 } c_j^i \text{ 落入值} \\ y_1^i = c_0^i, \dots, y_N^i = c_{k+1}^i, i = 1, \dots, p \end{cases}$$

y_m^i, y_{m+1}^i 之间, 则断点 $c_j^i = \frac{(y_m^i + y_{m+1}^i)}{2}$, 这样可求出 c_j^i 中的

k 个断点 $(c_1^i, c_2^i, \dots, c_k^i)$ 。于是 y_i 就被划分为 $k+1$ 个区

间, 而每个区间可以用一个离散值来代替即完成了离散化过程, 得到了离散化的故障决策表。然后从粗糙集的相对正域出发, 采用基于属性重要度的属性约简算法约简离散的故障决策表^[18]。该算法(算法2)如下:

输入: 决策表 $DT = (U, A = C \cup D, V, f)$

输出: DT 的一个相对约简 B

(1) 令 C 相对于 D 的核 $CORE_C(D) = \emptyset$;

(2) $\forall c_i \in C$, 如果 $POS_{C-c_i}(D) \neq POS_C(D)$, 则令 $CORE_C(D) = CORE_C(D) \cup \{c_i\}$, 直到遍历条件属性全集中的每一个属性;

(3) 设 DT 的一个相对约简为 B , 令 $B = CORE_C(D)$;

(4) 如果 $POS_B(D) = POS_C(D)$, 则 DT 的一个约简为 $B = CORE_C(D) \in RED_C(D)$, 属性约简完成转到第(6)步, 否则转到第(5)步;

(5) $\forall c_i \in C \setminus B$, 计算属性重要度 $sig(c_i, B; D) = |POS_{B \cup \{c_i\}}(D)| - |POS_B(D)|$, 求得 $c_m = \arg \max_{C_i \in C \setminus B} sig(c_i, B; D)$, 这里的 \arg 表示取使重

要度达到最大的属性 c_i , 而非最大的重要度值(若同时存在多个属性满足最大值, 则从中选取一个与 B 的属性值组合数最少的属性作为 c_m), 令 $B = B \cup \{c_i\}$, 转到第(4)步;

(6) 输出 B , 算法结束。

至此, 在不影响分类的前提下我们得到了复杂

4.1 故障决策表的主元特征提取

首先, 使用本文中的算法 1 对表 1 进行主元分析, 提取出故障样本的主元变量, 得到主元 Y1, Y2, Y3, Y4, Y5, Y6, Y7, Y8, Y9, Y10。前 9 个主元的贡献率如图 2 所示。根据图 2 贡献率的数值, 取累计方差贡献率 $\eta_k \geq 85\%$ 的主元。前 8 个主元 Y1~Y8 的累积贡献率为 91.56%, 大于给定的 85% 的取值, 用这 8 个主元对样本进行分析不会损失太多的信息。X 为表 1 中各条件属性对应的数据矩阵, 维数为 13×15。T 为前 8 个主元表达式对应的系数矩阵, 其维数为 15×8。令 $Y=[Y1, Y2, Y3, Y4, Y5, Y6, Y7, Y8]$, 根据算法 1 可知 $Y=XT$, 得到主元变量故障决策表的数据矩阵, 由此建立由主元作为条件属性的故障样本表。

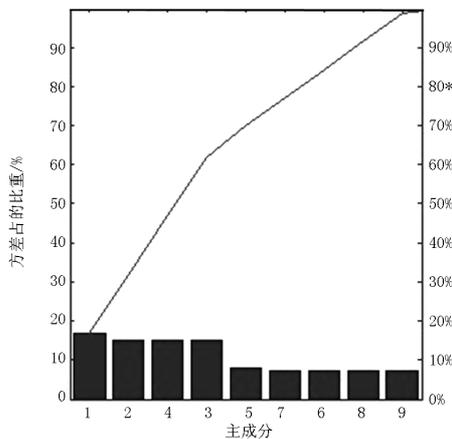


图2 主元贡献率

Fig.2 Percentage of principal components

4.2 基于粗糙集的主元故障决策表的离散化及其约简

设参数 $k=3$, 使用本文中的等频率离散化方法对主元样本表数据进行离散化, 每个主元下均可得到 3 个数据大小范围不同的离散区间, 区间由小到大分别以离散数 0、1、2 代替, 然后删除有重复的离散化的样本, 离散化结果如表 2 所示。

在不影响分类决策的前提下, 为了约简冗余的主元变量, 我们使用本文中的算法 2 对表 2 进行主元条件属性的约简, 进一步浓缩样本空间。经过运算, 得到表 2 的核条件属性 $core=[Y4, Y5]$, 一个最小属性约简 $redu=[Y2, Y3, Y4, Y5]$ 。以 $redu$ 为条件属性, 决策属性不变, 从表 2 构建约简故障样本, 并将重复样本删除, 我们就可以得到粗糙集约简后的故障决策表, 如表 3 所示。

表2 离散化决策表

Tab.2 Discreted decision table

U	Y1	Y2	Y3	Y4	Y5	Y6	Y7	Y8	FA
1	0	1	0	1	1	0	1	1	R1
2	0	1	0	1	1	2	1	1	R1
3	0	0	2	0	1	1	0	2	R2
4	0	0	2	0	1	1	2	0	R2
5	2	1	1	1	2	1	1	1	R1
6	2	1	1	1	2	1	1	1	R2
7	2	1	1	1	0	1	1	1	R1
8	2	1	1	1	0	1	1	1	R2
9	0	2	2	2	1	1	0	0	R4
10	0	2	2	2	1	1	2	2	R4
11	0	0	2	2	1	1	2	0	R5
12	0	0	2	2	1	1	0	2	R5
13	2	1	1	1	1	1	1	1	NO

表3 约简后的决策表

Tab.3 Rduced decision table

U	Y2	Y3	Y4	Y5	FA
1	1	0	1	1	R1
2	0	2	0	1	R2
3	1	1	1	2	R1/R2
4	1	1	1	0	R1/R2
5	2	2	2	1	R4
6	0	2	2	1	R5
7	1	1	1	1	NO

4.3 决策树的生成

由表 3 知, 约简后的故障决策表的核条件属性 $core=[Y4, Y5]$, $redu=[Y2, Y3, Y4, Y5]$ 为故障决策表的一个约简。根据上述决策树构建方法, 选择 $core$ 为根节点测试属性, 该节点有 5 种数据组合, 产生了 5 个分支, 其中 (0, 1)、(1, 2)、(1, 0) 组合只能产生叶节点, 不再分裂, 即可形成规则。(1, 1) 组合的下层候选节点为 $[Y2, Y3]$, 根据表 3 可知 $Y3$ 的属性值的类别多于 $Y2$, 形成决策的重要度大于 $Y2$, 选择 $Y3$ 作为该组合下的下层分支节点。其他节点据此递推, 最终得到的故障分类决策树如图 3 所示。

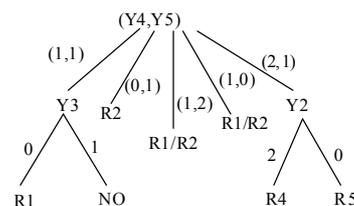


图3 故障决策树

Fig.3 Decision tree of faults

4.4 故障诊断测试与结果分析

我们从变电站监控系统中取得图1所对应的5条实际故障数据如表4所示,表中每个记录均包括了某些保护开关拒动或误动的故障数据,即为故障测试样本S。我们根据上文思路,首先利用算法1得到

表4 测试样本

Tab.4 Test sample

U	CB1	TP1	OP1	CB2	TP2	OP2	CB3	TP3	OP3	CB4	TP4	OP4	CB5	TP5	OP5
1	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0
2	1	1	0	0	0	1	0	0	0	0	0	0	0	0	0
3	0	1	0	1	1	0	0	0	0	0	0	0	0	0	0
3	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0
5	0	0	1	0	0	1	0	0	1	0	0	0	1	0	1

表5 表4的离散化

Tab.5 Discretization of Tab.4

U	Y1	Y2	Y3	Y4	Y5	Y6	Y7	Y8
1	0	1	0	1	1	1	0	2
2	0	1	0	1	1	0	2	0
3	0	0	1	0	1	0	0	2
4	2	2	2	2	1	1	1	1
5	0	0	1	1	2	1	0	0

图3的决策树有7个分支,相应地我们可得到7条诊断规则。我们可以根据决策树的各个节点的取值选择决策树的一条支路得到相应的故障区域。按照图3结合表5进行故障诊断,逐行比照,得到相应的故障分别为:FA={R1, R1, R2, R4, R5},而这与实际的故障是一致的。该方法对这5个测试样本故障识别率达100%。

综上所述,构建的决策树不仅能大部分反映故障的信号特征,而且诊断规则简单、易懂、直观,诊断起来速度更快。经过PCA和粗糙集的处理,故障样本的浓缩率为53.85%,大大减少了数据样本,计算量小;特征约简率为26.67%,能使最主要的信号特征凸显出来,使工程人员在生产过程中更容易发现故障并及时诊断。诊断实例证明了该方法是行之有效的。

5 结论

本文在研究PCA特征提取、粗糙集属性约简的基础上,结合变电站的故障诊断特点,提出了一种改进的决策树构建方法,并用于变电站的故障诊断,是个有益的尝试和探索。使用主元分析处理故障样本,使得原本规模大、数据变量多的数据信息维数大

S的主元变量数据矩阵 $D=S*T$,从而建立实际故障样本的主元决策表,然后依照上述的离散化准则对其进行离散化,我们得到离散后的主元故障决策表如表5所示。

大降低,同时也降低了决策诊断信息系统的分析难度;采用基于粗糙集核属性和属性重要度的方法构建的决策树简单、直观、规则易于提取,而且复杂度小,具有很好的容错性和识别效果。这一故障诊断方法还可以推广到其他对象,能被用于解决其他领域的故障诊断、模式识别等问题上。

参考文献

- [1] 刘青,等. 基于多知识库与粗糙集理论的变电站故障诊断方法[J]. 华东电力, 2006, 34(12): 27-28.
LIU Qing, et al. Substation fault diagnosis based on multi-knowledge base and rough set theory[J]. East China Electric Power, 2006, 34(12): 27-28.
- [2] Johnson R A, Wichern D W. 实用多元统计分析[M]. 陆璇,译. 北京: 清华大学出版社, 2001.
Johnson R A, Wichern D W. Applied multivariate statistical analysis[M]. LU Xuan, trans. Beijing: Tsinghua University Press, 2001.
- [3] Pawlak Z. Rough sets[J]. International Journal of Computer and Information Science, 1982, 11: 341-356.
- [4] Quinlan J R. Induction of decision trees[J]. Machine Learning, 1986 (1): 81-106.
- [5] David Antory, Uwe Kruger, Irwin G W, et al. Industrial process monitoring using nonlinear principal component models[C].//Second IEEE International Conference on Intelligent Systems. Kine: 2004: 293-298.
- [6] Choi Sang Wook, Lee Changkyu, Lee Jong-Min, et al. Fault detection and identification of nonlinear processes based on kernel PCA[J]. Chemometrics and Intelligent Laboratory Systems, 2005, 75(1): 55-67.
- [7] Choi Sang Wook, Lee In- Beum. Nonlinear dynamic process monitoring based on dynamic kernel PCA[J]. Chemical Engineering Science, 2004,59(24): 5897-5908.
- [8] Thaddeus T Shannon, David Abercrombie, et al. Improved

- process monitoring with independent components[C]. //IEEE/SEMI Advanced Semiconductor Manufacturing Conference. 2004: 170-175.
- [9] 王加阳, 谢颖. 基于量子粒子群优化的最小属性约简算法[J]. 计算机工程, 2009, 35(12): 148-150.
WANG Jia-yang, XIE Ying. Minimal attribute reduction algorithm based on quantum particle swarm optimization[J]. Computer Engineering, 2009, 35(12): 148-150.
- [10] WANG Xiang-yang, YANG Jie, TENG Xiao-long. Feature selection based on rough sets and particle swarm optimization[J]. Pattern Recognition Letters, 2007, 28(4): 459-471.
- [11] 程京, 等. 一个基于差别矩阵的属性约简改进算法[J]. 湖南大学学报: 自然科学版, 2009, 36(4): 85-88.
CHENG Jing, et al. An updated algorithm for attribute reduction based on discernibility matrix[J]. Journal of Hunan University: Natural Science, 2009, 36(4): 85-88.
- [12] 张利, 等. 基于粗糙集的启发式值约简的改进算法[J]. 仪器仪表学报, 2009(1): 82-85.
ZHANG Li, et al. Improved heuristic algorithm used in attribute value reduction of rough set[J]. Chinese Journal of Scientific Instrument, 2009(1): 82-85.
- [13] ZHANG L, Lu X Y, WU H Y. An improved heuristic algorithm used in attribute reduction of rough set[C]. // Proceedings of the First International Symposium on Data, Privacy and E-Commerce(ISDPE). 2007: 44-46.
- [14] 徐凤生. 一种属性与值约简及规则提取算法[J]. 计算机工程与科学, 2008(2): 61-63.
XU Feng-sheng. An attribute and value reduction and rule extraction algorithm[J]. Computer Engineering & Science, 2008(2): 61-63.
- [15] 王英英. 基于粗糙集与决策树的配电网故障诊断方法[J]. 高电压技术, 2008, 34(4): 794-798.
WANG Ying-ying, et al. Fault diagnosis method for distribution networks based on the rough sets and decision tree theory[J]. High Voltage Engineering, 2008, 34(4): 794-798.
- [16] 孙卫祥, 等. 基于 PCA 与决策树的转子故障诊断[J]. 振动与冲击, 2007, 26(3): 72-74.
SUN Wei-xiang, et al. Rotor fault diagnosis based on PCA and decision tree[J]. Journal of Vibration and Shock, 2007, 26(3): 72-74.
- [17] 高贇, 等. 连续量信息表决策值的离散化方法[J]. 西安科技大学学报, 2004, 24(4): 486-488.
GAO Bin, et al. Discretization methods on decision values of successive quantity information table[J]. Journal of Xi'an University of Science and Technology, 2004, 24(4): 486-488.
- [18] 王国胤. Rough 集理论与知识获取[M]. 西安: 西安交通大学出版社, 2001.
WANG Guo-yin. Rough set theory and knowledge acquisition[M]. Xi'an: Xi'an Jiaotong University Press, 2001.
- [19] Quinlan J R. Introduction of decision trees[J]. Machine Learning, 1986 (1): 84-100.
- [20] 梁道雷, 等. 一种多变量决策树方法研究[J]. 计算机科学, 2008, 35 (1): 211-212.
LIANG Dao-lei, et al. A new multivariate decision tree algorithm[J]. Computer Science, 2008, 35 (1): 211-212.

收稿日期: 2009-08-04; 修回日期: 2009-09-30

作者简介:

张延松 (1983-), 男, 硕士研究生, 从事智能控制算法研究; E-mail: favourite200811@163.com

赵英凯 (1943-), 男, 教授, 博导, 从事智能控制、机器人控制研究。

(上接第 54 页 continued from page 54)

- [16] 王联国, 洪毅. 随机交叉粒子群优化算法[J]. 计算机工程与应用, 2009, 45 (16): 69-71.
WANG Lian-guo, HONG Yi. Stochastic crossover particle swarm optimization[J]. Computer Engineering and Applications, 2009, 45 (16): 69-71.

作者简介:

董岳昕 (1985-), 女, 硕士研究生, 研究方向为电力系统无功优化、电能质量; E-mail: dyxscu@126.com

杨洪耕 (1949-), 男, 教授, 博士生导师, 从事电能质量分析与控制、区域电压无功控制等方面的研究与教学工作。

收稿日期: 2009-09-02; 修回日期: 2009-11-16