基于数据挖掘模型的高压输电线系统故障诊断

廖志伟，孙雅明
（天津大学电气自动化与能源工程学院，天津 300072）

摘要：在大多数故障诊断系统中，由于诊断所依据的实时信息在其形成和传递过程中都有可能产生信息的畸变，从而导致故障诊断结果的错误。文中提出利用基于粗糙集理论的数据挖掘模型来处理实时输入信息的畸变和实现输电线系统的故障诊断，它是依据粗糙集定性分析能力对知识域的数据集进行分析，粗糙集的约简是通过遗传算法求取。还给出了构造测试样本的理论准则，从而使检验故障诊断系统的容错性能具有保证和真正的实用价值。通过仿真测试证明，基于数据挖掘模型的故障诊断与基于神经网络模型的故障诊断相比，具有更高的容错性能。

关键词：输电线系统；故障诊断；容错性能；数据挖掘；粗糙集；遗传算法

中图分类号：TM 732；TP 18

0 引言

高压输电线系统(HTLS)的故障将严重威胁系统的安全运行。准确地进行故障预诊断可防患于未然，对电力系统快速恢复决策具有重要意义。

HTLS 的故障诊断主要依据事故环境下所发生的一系列实时信息序列来识别故障设备及其性能。该领域已有不少研究，如基于逻辑处理、专家系统、人工神经网络(NN)、进化技术等的智能原理故障诊断方法。这些研究方法在容错能力方面都有其局限性，即当诊断所依据的实时信息不充分或信息因畸变而变异、丢失时，形成变异故障模式而导致错误的诊断结论。该问题的实质是由所用诊断原理的容错性能所确定，它也是故障诊断系统推向实用化所要解决的关键问题。因此，如何对信息因畸变而形成的变异故障模式仍能正确识别，涉及实现故障诊断智能原理的自组织和自学习性能。

本文将变异故障模式分为 2 类：一类是变异故障模式与训练故障模式完全不同；另一类是变异故障模式与某个训练故障模式的输入信息组成完全相同。显然，后一类仅从信息的外部形式去研究是不可能的，必须从表征各种故障模式的信息内在知识关联性实现准确诊断。本文的研究仅限于前一类变异故障模式。提出运用基于粗糙集(rough set，缩写为 RS)理论的数据挖掘(data mining，缩写为 DM)技术[1]，提出了对 HTLS 故障诊断模式中存在不确定性信息的处理模型及相应的知识发现求解方法，构造一种综合智能的信息处理模型，在具有不精确、不正确、不完整、近似推理的环境下实现高容错性推理。首先利用 RS 来提取领域知识，其特点是无需预先给定某些特征或属性的数量描述，而是直接从给定问题的描述集合出发，将故障模式集转换成 RS 中的决策表，通过辨识矩阵和辨识函数确定 HTLS 故障模式的近似域，并挖掘出该问题中输入矢量与输出矢量内在的关联性。为了避免对故障模式空间进行盲目的随机搜索或穷举搜索，能快速获得最优知识或规则，本文提出用遗传算法(GA)的全局寻优能力对领域知识进行智能式搜索，求出最佳约简，即从故障样本模式信息中获取其冗余性及内在相关性联系规则。仿真测试结果证明所提出的方法能有效地克服因实时信息畸变而导致的错误诊断，确保 HTLS 故障诊断的高容错性能。本文还提出了构造测试样本模式的理论准则，以确保检验测试的有效性。

收稿日期：2001-04-10；修回日期：2001-05-15。
国家自然科学基金资助项目(59877016)。

1 粗糙集理论概念[2~4]

由 Z. Pawlak 提出的 RS 理论[2~4]是用以处理包含噪声、不确定及不完整信息的数学工具，它为数据挖掘提供了一种新的方法。为能更清晰地表达本文的研究，简略介绍了由 Z. Pawlak 教授提出的 RS 基本理论。本文主要引用波兰数学家 Skowron 提出的决策系统(DS)对故障模式空间进行描述，即将 RS 的方法和模型建立在直观的二维决策表基础上进行研究。

1.1 决策系统

RS 是根据已有的领域知识对给定问题的论域进行划分。称 $S=(U,A,\{V_a\},\alpha)$ 为信息表示系统，

其中 U 为非空有限集合，称论域；A 为非空有限集合，称属性集合；V_α 为属性 $\alpha \in A$ 的值域；$\alpha: U \to V_\alpha$ 为一单值，是论域 U 中的任一元素取属性 α 在 V_α 中的某惟一值。

若 A 由条件属性集合 C 和结论属性集合 D 组成；C，D 满足 $C \cup D = A$；$C \cap D = \varnothing$；则称 S 为决策系统(DS)。在 DS 中，可认为 U 中的每一元素对应一条规则，规则的前件由 C 及其取值所决定，规则的后件由 D 及其取值所决定。可以用 $(U, C \cup D)$ 和 $(U, C \cup \{D\})$ 来表示 DS。

1.2 决策系统的不可辨识关系

对给定 $S = (U, C \cup \{D\})$，$B \subseteq C$ 是条件属性 C 的子集，称二元关系：

$$\mathrm{IND}(B, \{d\}) = \{(x, y) \in U \times U : d(x) = d(y)\} \tag{1}$$

式(1)为 S 的不可辨识关系；x，y 为 U 中的元素。不可辨识关系是一种等价关系，通过它可得到 DS 的划分，称划分后的等价类为不可辨识类，通常用 $[x]_{\mathrm{IND}(B)}$ 表示包含元素 x 的不可辨识类，用 $\mathrm{IND}(B)$ 表示不可辨识关系 $\mathrm{IND}(B, \{d\})$。

1.3 上逼近和下逼近

对于信息表示系统 $S = (U, A)$，设 $B \subseteq A$，$X \subseteq U$，有如下定义：

$$\underline{BX} = \{x \in U \mid [x]_{\mathrm{IND}(B)} \subseteq X\} \tag{2}$$

$$\overline{BX} = \{x \in U \mid [x]_{\mathrm{IND}(B)} \cap X \neq \varnothing\} \tag{3}$$

$$\mathrm{BND}_B(X) = \overline{BX} - \underline{BX} \tag{4}$$

$\underline{BX}$，$\overline{BX}$ 和 $\mathrm{BND}_B(X)$ 分别为 X 的 B 的下逼近、上逼近和边界。$\underline{BX}$ 指根据对象现有知识可确定划归 X 的元素集合；$\overline{BX}$ 为不能确定是否属于 X 的元素集合；若 $\mathrm{BND}_B(X)$ 为空集，称 X 关于 α 是清晰的；若 $\mathrm{BND}_B(X)$ 为非空集，则称 X 为关于 α 的粗糙集。

1.4 删除冗余属性

对于 DS，$S = (U, C \cup \{D\})$，不可辨识关系 $\mathrm{IND}(C)$ 将 U 划分为 t 个不可辨识类 $X_1, X_2, \cdots, X_t$，令 $D(X_i)$ 为结论属性取值为 d 的集合，即

$$D(X_i) = \{v = d(x) : x \in X_i\} \tag{5}$$

若 $D(\mathrm{IND}(C - \{\alpha\})) = D(X_i)$，条件属性 $\alpha \in C$ 称为相对于不可辨识类 X_i 可去除的，即 α 的存在与否不影响 X_i 的结论值的集合。

1.5 约简

对于给定的 DS，$S = (U, C \cup \{D\})$，条件属性集合 C 的约简是指 C 的一个非空子集 C'，它满足：

a. $\mathrm{IND}(C', \{d\}) = \mathrm{IND}(C, \{d\})$；

b. 不存在 $C'' \subset C'$，使得 $\mathrm{IND}(C'', \{d\}) = \mathrm{IND}(C, \{d\})$。

约简可理解为：在不丢失信息的前提下，以最简方式表示 DS 的结论属性对条件属性集合的依赖和关联。所有约简的集合记为 $\mathrm{RED}(C)$。

因此，通过一组相对约简，可获得 DS 的 $S = (U, C \cup \{D\})$ 中最简单的规则集。

1.6 辨识矩阵

对信息系统 $S = (U, C \cup \{D\})$，论域 $U = \{x_1, x_2, \cdots, x_n\}$，$V_\alpha$ 为属性 $x \in U$ 的值域，辨识矩阵可表示为：

$$\boldsymbol{M}_A(x_i, x_j) = \begin{cases} \{\alpha \in A : V_\alpha(x_i) \neq V_\alpha(x_j)\} & D(x_i) \neq D(x_j) \\ 0 & D(x_i) = D(x_j) \\ -1 \quad V_\alpha(x_i) = V_\alpha(x_j), & D(x_i) \neq D(x_j) \end{cases} \tag{6}$$

式中 $i, j = 0, 1, \cdots, n$。

由式(6)可知，辨识矩阵包含了决策表中所有的属性区分信息。

2 HTLS 故障诊断模型

我们在 HTLS 故障诊断的研究过程中，采用了文献[5]中的诊断模型。该诊断模型是具有 12 个输入矢量元素和 17 个输出矢量元素的结构。

矢量所表示的具体物理意义如下。

输入矢量：x_1 为保护启动量达到定值；x_2 为保护发跳闸脉冲；x_3 为保护动作量到达定值；x_4 为故障发生在正方向；x_5 为重合闸过程有记忆；x_6 为开关最终状态为合；x_7 为失灵保护是否动作；x_8 为重合闸发合闸脉冲；x_9 为开关三相位置不一致发保护动作否；x_{10} 为开关是否永跳；x_{11} 为其他保护禁止重合闸；x_{12} 为压力降低闭锁重合闸。

输出矢量：y_1 为保护误启动；y_2 为保护拒动；y_3 为保护反应正确；y_4 为保护范围外故障；y_5 为反方向故障；y_6 为保护误动方向元件故障；y_7 为保护误动，动作元件故障；y_8 为线路瞬时性故障；y_9 为线路永久性故障；y_{10} 为开关拒跳失灵保护动作；y_{11} 为失灵保护拒动；y_{12} 为开关重合后又跳开，重合不成功；y_{13} 为开关拒跳，三相位置不一致保护拒动；y_{14} 为开关永跳；y_{15} 为其他保护禁止重合闸；y_{16} 为压力降低闭锁重合闸；y_{17} 为重合闸装置故障或重合闸退出。

按所有可能故障情况的知识构造了 65 个训练样本模式。

3 构造测试样本模式的理论准则

对任何故障诊断系统，为了进一步证实所建造模型的有效性，都应以一定数量的测试样本来检验。至今尚未见到如何有效地构造测试样本的准则。一般多采用从模型的输入空间获取一定数量的非训练样本作为测试样本模式。这种测试样本的检验法具

有很大的盲目性,不仅给问题的研究增加了难度,因为这样使解空间极为复杂,“清晰解”和“含糊解”交织在一起;甚至可能使检验失去意义,因为某些真正影响诊断错误的且实际应用中可能形成的样本模式恰恰可能未能列入,从而使对诊断系统容错性的检验无实际价值。

本文提出的构造测试样本模式的理论准则是以文献[6]中的可靠性理论分析为基础,以训练样本模式作为范例来衍生全部变异故障模式集,它们是实际环境中可能形成的非训练故障模式作为测试样本模式。本文按在输入信息矢量中仅出现1位畸变来构造衍生变异故障模式集。根据所用硬件和数字处理技术的冗余度及可靠性理论的依据,对所研究系统的65组故障模式的9个输入矢量元素发生畸变,总共为650个故障模式,构成425个衍生的有效变异故障模式(除去前述的第2类变异故障模式160个,它不属本文研究范畴)。

4 决策属性的编码方式

按RS用于DM的概念,应将65个输入和输出矢量的训练故障模式及其衍生变异故障模式集合(共计490个)构成整个信息表,其中全部故障模式为对象的集合,输入矢量集合为条件属性集合,而输出矢量(17个故障属性)集合形成决策(结论)属性集合或对象的类别。随着属性表的增大,计算约简的复杂程度将剧增,已证明求取对象的约简组合属于NP完全问题,难以用穷举法来实现。本文提出用GA作为约简的提炼。GA适用于处理大规模的组合优化问题,并能克服求解的局部极值。利用其对离散解空间充分搜索,快速实现全局最优解的挖掘。

根据RS的概念,若决策属性是一个决策类的特征属性,称单决策属性;若决策属性是多个决策类的特征属性的并集,则称多决策属性。后者是指一个样本模式可能同时属于不同的决策类。在实际诊断系统中多决策属性是普遍存在的,如在HTLS的诊断模型中,对任一故障模式其输出并非都是1个故障属性(或故障症状),可以有多个故障属性,如线路瞬时性故障和开关拒跳失灵保护动作并存,这样就需要用多决策属性约简。

在求取约简之前,必须对所研究故障诊断系统建立相应的信息表,其中条件属性部分由训练样本模式和测试样本模式(共490个)的输入空间确定;而决策属性部分的构造不是惟一的,由决策属性编码决定。决策属性编码可以有以下2种编码方式:

a. 用1个字段表示可能的17个故障属性,即决策属性是通过用1个17位的长整型的编码取值来表示每种故障属性。对单决策属性的样本情况,仅1个信息表,简单直观;但当样本为多决策属性时,所形成决策属性组合的值域空间急剧增大,对17位二进制可有2^{17}种组合。显然,需用大量故障模式才能覆盖决策属性的值域空间,不便于多特征提取。而实际上某些组合是不可能出现的,根据本文构造测试样本空间的机理,变异故障模式集合是极其有限的,更加不适用。

b. 以每一个故障属性表示为一个决策属性去构造信息表,即形成17个子信息表,它们的值域都以{0,1}来表示。这样,虽然信息表数量增加,但可将多决策属性转为单决策属性来处理,采用较成熟的单决策属性的约简法,挖掘各故障属性的特征信息。

本文研究中采用第2种编码方式,并可证明采用此编码方式时对各故障属性之间的关联性是不变的。

5 基于GA的约简形成[7~11]

在1.6节提及的辨识矩阵中隐含了决策表中的所有属性区分信息。设M为决策表的辨识矩阵;$A=\{\alpha_1,\alpha_1,\cdots,\alpha_n\}$是决策表中所有条件属性的集合;$S$是$M$中所有属性组合的集合,且$S$中不包含重复项。令$S$中包含有$s$个属性组合,每个属性组合表示为$B_i$,描述为$B_i\in S;B_j\in S;B_i\neq B_j(i,j=1,2,\cdots,s)$。令$\mathrm{card}(B_i)=m$,则$B_i$中每个条件属性表示为$b_{i,k}\in B_i(k=1,2,\cdots,m)$。这样,从决策表中求取约简的问题转化为在辨识矩阵中求取组合数最小的约简,其步骤如下:

a. 将求取最优约简的解编码成12位二进制染色体。0,1组成的基因代表优化空间的一个解,0表示此条件属性在最优约简中可忽略,1表示此属性与约简相关。随机产生70个染色体,构成初始解群。

b. 选取适应函数。适应函数的选取是GA的关键,它表示在适应函数约束下智能的搜索策略。本文中的适应函数f定义如下:

$$f(R)=(1-\alpha)\frac{\mathrm{card}(B)-\mathrm{card}(R)}{\mathrm{card}(B)}+\alpha\frac{|[\sin S|s\cap R\neq\varnothing]|}{|S|}\tag{7}$$

式中第1项是激励搜索策略朝M中属性组合数最小的约简方向搜索,第2项是确保R为约简;参数α为调节因子。B的子集R是以适应函数作为指导,通过进化搜索而得到的约简。

c. 遗传操作以适应函数为目标函数,通过选择、交叉和变异,实现优化搜索。本文采用的交叉率、变异率分别为0.3和0.05。每个约简在返回约简集

中有一个相应的置信度,用来衡量约简的强度。

d. 遗传操作。以适应函数为目标函数,通过选择、交叉和变异,实现优化搜索。

e. 收敛判据。当遗传操作一定的迭代次数后,适应值没有明显改变,算法结束。

f. 从约简集中提取规则。将所得到约简的每一个对象重新匹配到父决策表中。若该规则对于决策表所有可能取值可完全辨识,则该子规则成立。依次对约简集中的所有约简类推,新规则集即形成,如表1所示。

表1 信息表的17个决策属性的约简及规则
Table 1 Reductions and decision rules of 17 decision attributes in information table

故障属性	计算的最优约简	决策规则数
y_1	x_1	2
y_2	$x_2x_3x_4x_5x_6x_7x_8$	92
y_3	$x_1x_2x_3x_4x_5x_6x_7x_8$	136
y_4	$x_2x_3x_4x_5x_6x_7x_8$	92
y_5	$x_1x_2x_3x_4x_5x_6x_7x_8$	136
y_6	$x_1x_2x_4x_5x_6x_7x_8$	90
y_7	$x_1x_2x_3x_4x_5x_6x_7x_8$	136
y_8	$x_2x_3x_4x_5x_6x_7x_8$	92
y_9	$x_3x_4x_5x_6x_8x_9x_{10}x_{11}x_{12}$	176
y_{10}	$x_2x_5x_6x_7x_8$	23
y_{11}	$x_1x_2x_3x_4x_5x_6x_7x_8$	90
y_{12}	$x_5x_6x_8x_9x_{10}x_{11}x_{12}$	44
y_{13}	$x_5x_6x_8x_9x_{10}x_{11}x_{12}$	44
y_{14}	$x_5x_6x_8x_9x_{10}x_{11}x_{12}$	44
y_{15}	$x_5x_6x_8x_9x_{10}x_{11}x_{12}$	44
y_{16}	$x_5x_6x_8x_9x_{10}x_{11}x_{12}$	44
y_{17}	$x_5x_6x_8x_9x_{10}x_{11}x_{12}$	44

6 DM与基于NN原理的HTLS故障诊断模型的仿真测试及对比

本文对以下2种原理的故障诊断系统做了仿真测试和比较:①基于NN原理的故障诊断系统,其中NN模型具有12个输入矢量元素和17个输出矢量元素,其训练样本为65组样本模式,测试样本模式是425个变异故障模式。②本文提出的基于RS理论的DM模型的故障诊断系统,是用65组样本模式和425个变异故障模式进行数据挖掘,求取决策规则。仿真测试结果与比较如表2所示。

由表2可知,基于NN模型对输入矢量的变异模式具有一定的容错性,但正确率较低,其原因是由于衍生变异模式极其复杂,如有些衍生变异模式与原故障模式之间的汉明距离相等;另外,各原故障模式的变异模式之间又形成相同的变异模式,这给NN的自学习造成难度,故难以正确诊断。而本文提出的算法实现了各种故障模式的信息内在知识联系的挖掘和发现,例如挖掘出决定输出矢量元素与输入矢量元素之间、各输出矢量元素之间的耦合关系,从而确保了判断的高容错性,其诊断结果与NN模型系统相比得到了极大提高。

表2 2种故障模型仿真结果与比较
Table 2 Simulation results and comparison of two diagnosis model

输出矢量元素中不正确的维数	基于NN的诊断系统		基于DM的诊断系统	
	模式数	比率/(%)	模式数	比率/(%)
17维都正确	163	33.3	350	72.4
1维不正确	177	36.2	30	6.1
2维不正确	113	23.1	94	18.6
3维不正确	30	6.1	12	2.2
4维及以上不正确	7	1.3	4	0.7

另外,若基于DM诊断系统输出得不到确切的故障属性时,该系统会给出每个故障属性的置信度,可以提醒运行人员进一步查询判断,利用其结果与其相关的信息进行修正,然后再次诊断。所以,本文提出的诊断模型与基于NN模型的故障诊断相比有极强的容错能力和诊断结论的透明度,使构成的故障诊断系统具有较高的实用价值。

7 结语

由研究证明,DM能从大量、不完全、有噪声、模糊、随机的数据中挖掘出隐藏着的对决策生成有重要参考价值的信息,因此DM和知识发现技术具有广阔的应用前景。

本文将粗糙集理论的定性分析功能同GA全局最优解搜索功能有机相合,构造基于DM的故障诊断模型,具有较强的自学习性能,使系统具有较高的容错性能。

本文提出的构造测试样本的理论准则,对诊断系统的实用化有重要的实用价值。

本文求取约简用的ROSETTA软件包是由波兰华沙大学与挪威科技大学开发的,在此表示感谢。

参考文献

1 廖志伟,孙雅明(Liao Zhiwei, Sun Yaming). 数据挖掘技术及其在电力系统中的应用(Data Mining Technology and Its Application on Power System). 电力系统自动化(Automation of Electric Power Systems), 20001, 25(11): 62～66

2 Pawlak Z. Rough Sets. International Journal of Information and Computer Science, 1982, 11: 341～356

3 Pawlak Z. Rough Sets: Theoretical Aspects of Reasoning About Data. Dordrecht (Netherlands): Kluwer Academic Publishers, 1991

4 Pawlak Z, Grzymala-Busse J, Slowinski R, et al. Rough Sets. Communications of ACM, 1995, 38(11): 89～95

5 孙雅明,宋建文(Sun Yaming, Song Jianwen). 基于NN/ES高压输电线路在线故障综合诊断和分析的智能系统(A NN/ES Based Intelligent System for High Voltage Transmission Line On-Line Faults Synthetic Diagnosis and Performance Analysis). 系统工程理论与实践(Systems Engineering—— Theory & Practice), 1997, 17(2):60～66

6 姜惠兰(Jiang Huilan). 联想记忆神经网络的容错性研究及其在输电线路故障识别中的实现:〔博士学位论文〕(Study on Fault-Tolerance Performance of Associative Memory NN and Its Application in Transmission Line Fault Recognition, Doctoral Dissertation). 天津:天津大学(Tianjin: Tianjin University), 1999

7 Staal A. Vinterbo Predictive Models in Medicine: Some Methods for Construction and Adaptation Norwegian University of Science and Technology, Doctoral Dissertation. 1999

8 ϕhrn A. Discernibility and Rough Sets in Medicine: Tools and Applications NTNU Trondheim Norwegian University of Science and Technology, Doctoral Dissertation. 1999

9 Zhang Q, Han Z X, Wen F S. A New Approach for Fault Diagnosis in Power Systems Based on RS Theory. In: IEE Proceeding of 4th International Conference on Advances in Power System Control, Operation & Management. Hong Kong: 1998. 597～602

10 李勇敏,朱善君,陈湘晖,等(Li Yongmin, Zhu Shanjun, Chen Xianghui, et al). 基于粗糙集理论的数据挖掘模型(Data Mining Model Based on Rough Set Theory). 清华大学学报(自然科学版)(Journal of Tsinghua University (Science and Technology)), 1999, 39(1): 110～113, 117

11 Madan S, Son W K, Bollinger K E. Applications of Data Mining for Power Systems. In: Proceedings of 1997 IEEE Canadian Conference on Electrical and Computer Engineering. 1997. 403～406

廖志伟,男,博士研究生,当前研究方向为电力系统智能控制和故障诊断。

孙雅明,女,教授,博士生导师,长期从事电力系统智能控制和故障识别、变电站和配电网的综合自动化、综合智能型负荷预测等工作。

A NEW DATA MINING APPROACH FOR FAULT DIAGNOSIS OF HIGH VOLTAGE TRANSMISSION LINE BASED ON ROUGH SET THEORY

Liao Zhiwei, Sun Yaming

(Tianjin University, Tianjin 300072, China)

Abstract: In most practical application of fault diagnosis system, misjudgement may be caused by real-time information distorted in the process of generation and transfer. This paper presents rough set (RS) based data mining method to deal with distorted information and to implement the fault diagnosis of HV transmission line system (HVTLS). In this approach, the qualitative analysis ability of RS is used to analyze knowledge region data set and the reductions of RS are solved by a genetic algorithm (GA). At same time, this paper proposes the criterion to build testing samples, in order to get the assurance of fault tolerance performance of tested diagnosis system and have practical application potential of studied system. The higher fault tolerance performance of the proposed approach is confirmed through the comparison with that of NN-model based fault diagnosis system.

This project is supported by National Natural Science Foundation of China (No. 59877016).

Keywords: transmission line system; fault diagnosis; fault tolerance performance; data mining; rough set; genetic algorithm