文章编号:1671-6833(2003)01-0109-04

粗糙集与决策树结合诊断故障的数据挖掘方法

石金彦,黄士涛,雷文平

(郑州大学机械工程学院,河南 郑州 450002)

摘 要:根据数据挖掘技术用于故障诊断的基本思想,利用粗糙集进行数据归纳,过滤大量故障数据中的冗余属性,得到精简故障数据集.而后,用决策树分类方法处理精简数据集,产生分类所需的规则以进行分类,并结合实例说明了该方法的工作步骤.由实例可知该方法用于故障诊断的可行性,最后指出实际应用过程中的一些技术难题.

关键词:数据挖掘;故障诊断;关联规则;粗糙集;决策树

中图分类号:TP 391 **文献标识码**:A

0 引言

随着数据库技术的迅速发展及大型关键设备自动化程度的提高,各工厂对重要设备都实施了实时监控并由传感器不间断地传回反映机组运行状态的各种数据及参数,形成大型的数据库或数据仓库.这些数据和参数中包含了机组运行状态的各种特征,而数据和参数本身往往是杂乱无章的,其特征并不明显、不直观.而数据挖掘中的知识发现是从数据中识别出有效的、新颖的、潜在有用的以及最终可理解的模式的高级应用过程[1].利用数据挖掘进行故障诊断,就是根据该机组的历史运行记录,对其可能的运行状态进行分类并对其趋势进行预测.故障诊断的本质是模式识别,对机器故障进行诊断的过程,其实也就是模式获取及模式匹配的过程.而知识获取是智能诊断乃至人工智能发展的瓶颈,考虑数据挖掘技术近几年来的迅猛发展及它在知识获取方面的独特优势,将数据挖掘中粗糙集技术与决策树技术结合应用于故障诊断也不失为一种方法.在各种机械设备中,旋转机械比较典型,本文主要根据旋转转子的一些故障特征进行分析.

1 故障诊断的数据挖掘方法策略

对机械故障诊断而言,首先要获取关于本机组的大量运行参数,既要有机器平稳运行、正常工作时的数据,更要有机器出现故障时的数据,并且应已获知故障的类别.这样,由已知故障类别、故障发生时的各运行参数、历史记录组成的数据库或数据仓库便构成了数据挖掘的训练/学习样本库.数据挖掘的任务就是从这些海量的杂乱无章的样本库中找出隐藏在其中的内在规律,提取出不同故障的各自特征.在数据挖掘处理分类问题时,采用一种分类方法未必会取得良好的效果.对同一问题可根据需要选用不同的分类方法,依照不同的判决规则完成分类工作.本文采用发展较为成熟的粗糙集与决策树理论结合来处理实际问题,即利用粗糙集理论进行数据规约、过滤冗余属性,然后利用决策树方法来产生分类所用到的规则,即达到分类的目的.依据这些规则,对新来数据进行判别并对故障数据进行归类,识别出故障的种类,依此找到故障的原因并消除故障,如图 1 所示.

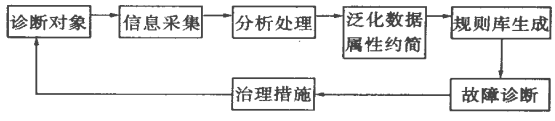

图 1 基于数据挖掘方法的故障诊断示图

Fig. 1 The sketch map of fault diagnosis based on data mining method

2 基本原理

旋转转子的特征属性一般由其时域特征、频域特征(低频与高频成分)及一些敏感参数如相位、轴心轨迹、振动方向、进动方向、临界转速等组

收稿日期:2002-09-03;修订日期:2002-10-30

作者简介:石金彦(1979-),河南省新乡市人,郑州大学硕士研究生.

成.在大数据量背景下,导致某一故障发生时的特征属性值及属性之间的关系可能会有某种规律存在,这与数据挖掘技术中的规则相吻合.数据挖掘技术中的关联规则反映一个事件和其它事件之间依赖或关联的知识,如果两项或多项属性之间存在关联,那么其中一项的属性值就可以依据其它属性值进行预测,为 $A \Rightarrow B$ 的形式[2].用支持度、置信度、最小支持度、最小置信度四个参数可确定一条规则.利用数据挖掘方法,找到这样的规则,根据规则将很容易地将未知故障数据归类.

在大量故障数据的收集中,很难知道哪些属性是重要的,那些属性是不重要的,所有的属性都被认为是有用的,并全部存入数据库,大大增加了信息存储量和处理量.实际对于特定的知识发现,可能只有某一特征子集有用,需要对特征属性进行约简.粗糙集能够在缺少数据先验知识的情况下,以对观测数据的分类能力为基础,解决模糊或不确定性数据的分析与处理[3].设有信息系统 $S=\{U,Q,V,f\}$,$Q=C\cup D$,其中 $U=\{x_1,x_2,\cdots,x_n\}$ 为全部有限、非空的对象集,为属性集,为所有属性值域的并集,f 为信息函数,C 为条件属性集,D 为决策属性集.本文所涉及到的粗糙集方法就是通过属性归约算法,达到删除冗余属性的目的.当存在 $1\leqslant j\leqslant i\leqslant n$,$m_{ij}=\{c\}$ 时,有核心 $\mathrm{CORE}=\{c\in C: m_{ij}=\{c\}\}$,其中 m_{ij} 为 S 关于 C 的分辨矩阵元素.所谓属性归约算法,即以核心 $\mathrm{CORE}(C,D)$ 为计算起点,若用户未指定任何属性,则算法找出最佳归约集,包括那些具有最大重要性属性.若用户指定某些属性,则算法产生用户定义最小属性集,而不丢失信息.步骤如下:

(1) 计算决策表的可辨识矩阵 $\boldsymbol{C}_D$;

(2) 对于可辨识矩阵中的所有取值为非空集合的元素 C_{ij}($C_{ij}\neq 0, C_{ij}\neq\phi$),建立相应的析取逻辑表达式 $L_{ij}=\bigvee\limits_{a_i\in C_{ij}} a_i$;

(3) 将所有的析取逻辑表达式 L_{ij} 进行合取运算,得一个合取范式 $L=\bigwedge\limits_{C_{ij}\neq 0, C_{ij}\neq\phi} L_{ij}$;

(4) 将合取范式 L 转换为析取范式的形式,得 $L'=\bigvee\limits_i L_i$;

(5) 输出属性约简结果.

析取范式中的每个合取项就对应一个属性约简的结果,每个合取项中所包含的属性组成约简后的条件属性集合.

传统的数据挖掘方法有一定的局限性,例如只重视从数据库中提取规则,而忽视库中数据的变化,其人为干预数据少,所以灵活性较差.而决策树却有一定的优势,即使训练库中数据发生变化,通过遍历树也会容易调整树的结构[4].使用决策树不仅可以达到分类的目的,而且当故障数据样本增加时,扩充样本库内容时其灵活性也得到充分的体现[3].决策树中的每一节点与特征属性相关联,该属性包含有重要的特征信息.

Entropy(熵值,平均信息量)用于表示节点信息的重要程度.公式如下:

$$AE=\sum_b \frac{n_b}{n_i}\times\Big[\sum_c -\Big(\frac{n_{bc}}{n_b}\Big)\Big]\log_2\Big(\frac{n_{bc}}{n_b}\Big)\Big] \quad (1)$$

式中:n_i 为总样本数;n_b 为属性值为 b 的样本数;n_{bc} 为当属性值为 b 时导致某种故障类 c 出现的样本数.从树根开始遍历整个树,每一路径就形成一条决策规则,通过对树的修剪可得到精练的规则集.

3 应用实例

建立该机组的运行记录样本库.令 F 表示整个故障集,则 $F=(F_1,F_2,\cdots,F_J)$,其中 J 为故障类型的个数.对每一类型的数据,选择一定数量的文件,抽取对分类影响最大的特征,组成属性集.若令 A 表示整个属性集,则 $A=(A_1,A_2,\cdots A_I)$,I 为属性(特征)的个数.故障集和属性集组成用于训练/学习的故障诊断样本库,训练的方式是利用基于粗糙集理论和基于决策树的数据挖掘方法进行学习.以下训练/学习库给出了转子实验台信号的时域特征、频域特征(低频与高频成分)及一些敏感参数.这里我们所选取的识别参数属性集 $A=\{(0\sim0.4)f,(0.41\sim0.5)f,(0.51\sim0.99)f,1f,2f,(3\sim5)f,(>5)f$,相位,轴心轨迹,振动方向,进动方向,临界转速$\}$,选取的样本数据如表1所示,共30组.

3.1 精简属性

首先泛化样本数据表,基于粗糙集的属性归约算法确定归约集,指定条件属性集为{0,1,2,3,4,5,6,7,8,9,10,11},决策属性集为{12}.运行规约算法,产生的规约集如下:{0,4},{1,4},{2,4},{4,6},{0,7},{1,7},{2,7},{6,7},{0,9},{1,9},{2,9},{6,9},所有规约集的并集为{0,1,2,4,5,6,7,9},说明其中根据属性$\{1f,(3\sim5)f$,轴心轨迹,进动方向,临界转速$\}$来判断故障类型不太合理,故不考虑这些属性.从表中可以明显看出,假如以上实例中进动方向属性的值全是"正进动",根据进动方向也就不可能去判断故障属于哪一类

型(当数据量大的时候就不会出现一个属性只有单一值的情形).因此,只考虑对结果有影响的一些属性,精简后样本数据集如表 2 所示.

表 1　训练样本集

Tab.1　The training sample set

编号	(0~0.4)f	(0.41~0.5)f	(0.51~0.99)f	$1f$	$2f$	(3~5)f	(>5)f	相位	轴心轨迹	振动方向	进动	临界转速	故障
1	0.0000	0.0035	0.0000	1.0000	0.0890	0.0039	0.0010	稳定	规则	轴向	正进动	不变	不平衡
2	0.0012	0.0000	0.0000	1.0000	0.0890	0.0139	0.0000	稳定	规则	轴向	正进动	不变	不平衡
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
11	0.0000	0.0000	0.0000	1.0000	1.4140	0.1950	0.0000	半稳定	规则	混合	正进动	不变	不对中
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
30	0.6580	2.0740	0.5390	1.0000	0.4090	0.0240	0.0900	稳定	规则	轴向	正进动	不变	油膜涡动

表 2　精简后的训练样本集

Tab.2　The reduced training sample set

编号	(0~0.4)f	(0.41~0.5)f	(0.51~0.99)f	$2f$	(>5)f	相位	振动方向	故障类型
1	0.0000	0.0035	0.0000	0.0890	0.0010	稳定	轴向	不平衡
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
30	0.6580	2.0740	0.5390	0.4090	0.0900	稳定	轴向	油膜涡动

3.2　构造决策树并产生规则库

对样本数据集的每个属性执行平均熵值(Average Entropy)计算,简称 AE,依次把具有最小 AE 的属性作为决策树子树的根,见图 2 及表 3.

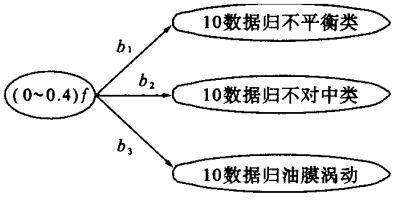

图 2　样本分布

Fig.2　Distribution of samples

表 3　(0~0.4)f 属性取值

Tab.3　Value of (0~0.4)f attribute

区间名	取值区间
b_1	[0.0000,0.0895]
b_2	[0.0895,0.6580]
b_3	[0.6580,1]

计算属性(0~0.4)f 的 AE,由公式(1)可得

$$AE=\sum_b\left(\frac{n_b}{n_i}\right)\times\left[\sum_c-\left(\frac{n_{bc}}{n_b}\right)\log_2\left(\frac{n_{bc}}{n_b}\right)\right]=$$

$$\frac{20}{30}\times\left[1-\left(\frac{10}{20}\right)\log_2\left(\frac{10}{20}\right)-\left(\frac{10}{20}\right)\log_2\left(\frac{10}{20}\right)\right]=0.5.$$

以此类推其它属性的 AE,如表 4 所示.

表 4　各属性的 AE

Tab.4　AE of each attribute

属性	AE	属性	AE
(0.41~0.5)f	0.5	(>5)f	0.5
(0.51~0.9)f	0.5	相位	0.5
$2f$	0	振动方向	0.5

由表 4 明显看出,属性 $2f$ 的 AE 值最小,把 $2f$ 作为决策树的根.对 $2f<0.2295$ 的子集中各属性进行下一步的搜索.计算 $2f<0.2295$ 子集中其它属性的 AE 值.可知各属性 AE 均为 0,图 3 所示最左边待定节点为叶节点.依次计算中间节点和右节点,可知均为叶节点.停止搜索.产生的决策树结构如图 4,其结构简单的原因是选取的训练集数据过少,不能代表全部故障数据集.

根据上面的决策树,见图 4,产生的规则如下:

Rule 1: $2f<0.2295 \Rightarrow$ class =1

(Sup:33.333% Conf:100.000% 10 10 Cov:33.333%)

Rule 2: $2f=[0.2295, 0.7715] \Rightarrow$ class =3

(Sup:33.333% Conf:100.000% 10 10 Cov:33.333%)

Rule 3: $2f>=0.7715 \Rightarrow$ class =2

(Sup:33.333% Conf:100.000% 10 10 Cov:33.333%)

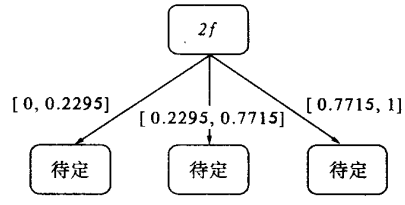

图 3　第一步搜索后形成的树

Fig.3　Tree generated after the first search

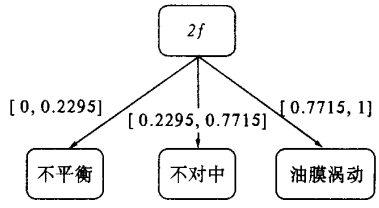

图 4　决策树

Fig.4　The decision tree

3.3 根据规则库进行故障类型匹配

设新来数据样本如表5所示.根据规则1,第一条样本数据中的 $\frac{2}{1}f = 0.082 < 0.2295$ 应归于不平衡故障;根据规则2,第二条样本数据应归于不对中故障;最后一样本数据由规则3可知应为油膜涡动故障数据.

表5 测试数据集

Tab.5 The test data set

编号	(0~0.4)f	(0.41~0.5)f	(0.51~0.99)f	$\frac{1}{1}f$	$\frac{2}{1}f$	(3~5)f	(>5)f	相位	轴心轨迹	振动方向	进动	临界转速
1	0.0000	0.0000	0.0000	1.0000	0.0820	0.0030	0.0000	稳定	规则	轴向	正进动	不变
2	0.0000	0.0000	0.0000	1.0000	1.1710	0.2920	0.0000	半稳定	规则	轴向	正进动	不变
3	0.7680	2.4030	0.7040	1.0000	0.4350	0.0270	0.0170	稳定	规则	轴向	正进动	不变

4 结束语

用粗糙集与决策树相结合的数据挖掘方法能较好地构造故障样本库,生成规则库,完成故障的识别工作,为故障分类提供决策依据,并达到预期的目的.然而,一些技术难题仍不可避免.数据挖掘技术需要大量甚至是海量数据的支持,否则精度会大打折扣.表1所示的样本仅仅是作者按照经验为了验证数据挖掘技术可用性而构造的一些数据,在实际应用中没有太多的参考价值.数据样本的预处理,包括各种变换、消噪、数据浓缩技术.本文所用粗糙集理论仅能处理离散的数据,在处理具有连续属性的数据样本时,泛化样本数据成为另一技术难题,即数据的预处理中连续属性值区间步长的划分比较关键.很可能把包含有重要信息的属性当作冗余属性处理,造成结果的偏离或错误.可能存在这样的数据,它既满足规则库中的一个规则,而同时又满足另外的规则,这两个规则所推导的结论并不一致并因而出现两种或两种以上的故障.而在现实中这样的数据也是存在的,它同时具有两个或多个故障的特征.

参考文献:

[1] 高毅龙. 数据挖掘技术及其在工程诊断中的应用[D]. 西安:西安交通大学, 2000.

[2] 韩家炜,坎 伯.数据挖掘概念及技术[M]. 范 明,孟小锋,译.北京:机械工业出版社,2001.149~183.

[3] 王国胤. Rough 集理论与知识获取[M]. 西安:西安交通大学出版社, 2001.23~51.

[4] QUINLAN J R. Simplifying decision trees[J]. International Journal of Man—Machine Studies, 1987,27:221~248.

[5] QUINLAN J R. Improved use of continuous attributes in C4.5[J]. Journal of Artificial Intelligence Research, 1996,(4):77~90.

The Method Combining Rough Set and Decision Tree in Fault Diagnosis

SHI Jin—yan, HUANG Shi—tao, LEI Wen—ping

(College of Mechanical Engineering, Zhengzhou University, Zhengzhou 450002, China)

Abstract: The principle and the basic idea of data mining method used in fault diagnosis are presented. Reduced fault data set is obtained after filtered redundant attributes by rough set theory from a mass of fault data. Then the decision tree wiss constructed on the basis of the data set, rules for classification are generated and the steps of getting rules are shown through one instance. It can be concluded that combining rough set theory and decision tree can be practicably used in fault diagnosis from the instance. Finally, some difficulties of this technology in practical application are mentioned.

Key words: data mining; fault diagnosis; association rule; rough set; decision tree