葛科宇

基于 Weka 平台知识获取的航空发动机磨损故障诊断专家系统

葛科宇,陈 果

(南京航空航天大学 民航学院,南京 210016)

摘 要: 针对航空发动机磨损故障诊断专家系统知识获取难的问题。提出了一种基于 Weka 平台知识自动获取的航空发动机磨损故障专家系统模型。引入目前国际上著名的数据挖掘软件 Weka,并把它嵌入自己开发的专家系统中作为一个知识自动获取模块。阐述了 C4.5 决策树中连续属性的离散、树的构建、树的修剪及规则的产生等关键技术。采集了一组某型航空发动机实测油样光谱数据,利用基于 Weka 平台的 C4.5 决策树提取了发动机磨损故障知识规则。

关 键 词: Weka; C4.5 决策树; 磨损故障; 知识获取; 专家系统

中图分类号: TP182 **文献标识码:** A **文章编号:** 1003-8728(2011)11-1955-05

Knowledge Acquisition of Aero-engine Wear Fault Diagnosis Expert System Based on Weka Platform

Ge Keyu, Chen Guo

(College of Civil Aviation, Nanjing University of Aeronautics and Astronautics, Nanjing 210016)

Abstract: According to the knowledge acquisition problems of aero-engine wear fault diagnosis expert system, a knowledge acquisition model obtaining automatically the wear fault of aero-engine based on weka platform was proposed in this paper. A world famous data mining software called Weka was introduced and embedded in the expert system as an automatic knowledge acquisition module. The key technologies including continuous attribute discretization, decision tree construction, tree pruning and rule extraction were expounded. This new method was applied to aero-engine wear faults diagnosis. A series of spectral oil analysis samples were acquired from practical aero-engine, the rule extraction from decision tree method was used to extract the diagnosis knowledge rules.

Key words: Weka; C4.5 decision tree; wear fault; knowledge acquisition; expert system

随着航空技术的不断发展,现代化航空发动机的结构日益复杂,其滑油系统中各摩擦副零组件更趋于高载荷、高温、高速及轻质量,因此容易发生各种磨损故障。光谱分析不仅能够建立起元素含量与磨损部位和故障性质之间的关系,而且具有精度高、数据重复性好、分析时间短以及对非铁磁性金属颗粒的灵敏性高等优点。因此,滑油光谱监控是对航空发动机的状态进行监控的有效手段,是开展视情维修的重要保证。

光谱诊断专家系统是利用光谱分析进行航空发动机磨损故障诊断的高级阶段。目前,国内外已有很多滑油检测专家系统。如美国 Mobil 润滑油公司开发的先进快速分析系统 PFALink、美国和加拿大共同研究的润滑油分析专家系统 Lube Analyst 和 Atlas。但这些软件所提供的仅仅是一个框架和管理系统,其核心知识库要用户自己开发,另外还需用户提供所监控对象的磨损元素界限值。国内研究单位针对特定的对象也开发了许多滑油光谱分析专家系统,并取得了许多成绩[1~3]。但是,应该看到,在智能诊断专家系统领域,知识获取能力弱,知识更新困难,知识适应性差等方面的问题仍然没有得到有效的克服。目前的专家系统知识获取基本上是基于经验的机械式的学习方法,知识更新困难、知识规则经常会出现严重

收稿日期: 2010-09-09

基金项目: 国家自然科学基金项目(61179057)资助

作者简介: 葛科宇(1985-),硕士研究生,研究方向为数据挖掘、智能诊断与专家系统,kyge1985@126.com; 陈 果(联系人),教授,博士生导师,cgzyx@263.net

的不一致、冗余、甚至组合爆炸等问题。

数据挖掘可以从大量的数据中挖掘出许多隐含的,有价值的信息[4,5]。其中,决策树技术算法描述简单、分类速度快、具有较好的分类精度。与此同时,自从 Quinlan 于 1986 年提出最早的 ID3 算法后,国内外学者对决策树算法研究一直都没有停止,Cendrowska 根据属性为实例分类提供的有用信息量作为测试标准。DeMantaras 建议利用划分距离的办法选择测试属性。Quinlan 本人也改进了 ID3 算法,提出按信息比值进行估计的方法,即后来的 C4.5 算法[6,7]。杨清等人针对 ID3 学习简单的逻辑表达式能力较差的缺点提出了一种优化算法 MID3。在众多的决策树算法中 C4.5 算法具有能处理连续属性以及克服了用信息增益选择属性时偏向选取多值属性的不足的特点,所以非常适合处理具有连续属性值且要确定界限值的滑油光谱数据。

为了解决传统专家系统知识获取和更新等问题,将 C4.5 决策树理论引入航空发动机磨损故障知识规则提取中。介绍了决策树规则提取过程中树的构建、树的修剪、规则的提取等一些关键技术。通过开源数据挖掘软件 Weka 实现对航空发动机磨损故障知识规则的自动提取,并对其进行解释。

1 Weka 平台介绍

Weka 是新西兰 Waikato 大学开发的全面的数据挖掘系统[8],它不仅提供了多种数据挖掘方法(分类、聚类、关联规则等)的多种常用算法进行知识发现,还提供了适用于任意数据集的数据预处理功能,以及算法性能评估的多种方法。Weka 是由 Java 语言实现的开放性平台,具有非常良好的扩展性和兼容性,用户可以根据具体需要将个性化的算法封装进系统,达到数据处理及算法性能评估的目的,具有良好定义的数据结构和基本的统计接口。图1是利用 Eclipse 开发平台对 Weka 进行了部分汉化后的界面。

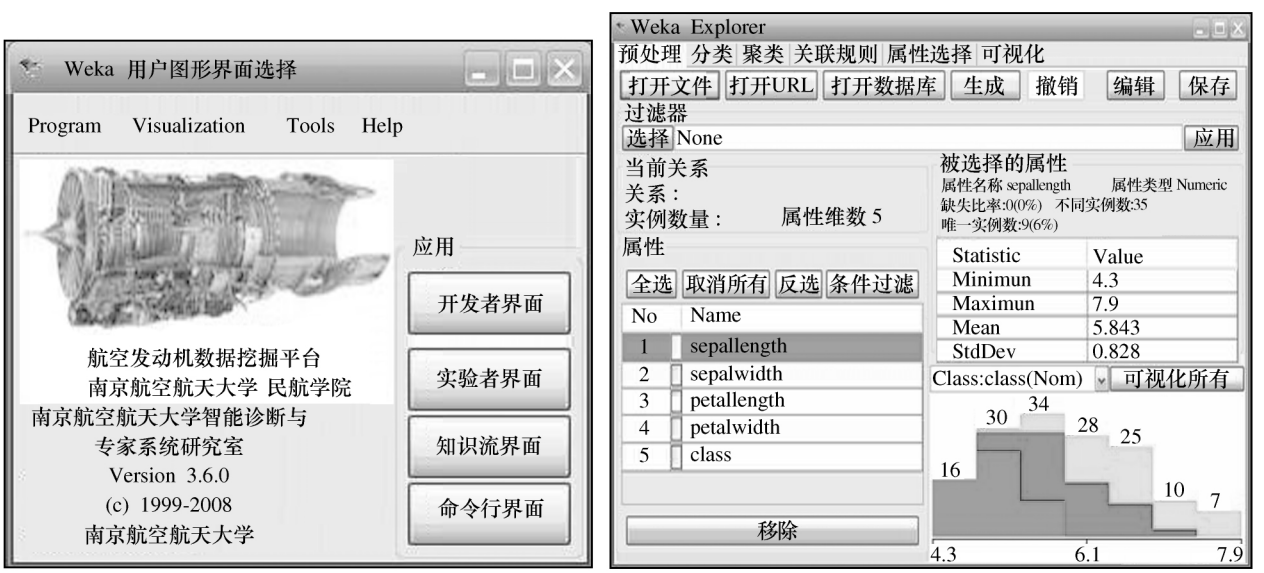

图1 Weka 数据挖掘软件

2 基于 Weka 平台的知识自动获取专家系统

2.1 方法流程

基于知识的专家系统主要由知识库、推理机、人机接口、知识获取子系统、解释子系统、全局数据库组成。传统的基于知识规则的专家系统由于知识获取和更新比较困难,因此,笔者提出了基于 Weka 平台 C4.5 决策树理论的知识自动获取的专家系统,基于 Weka 平台的知识自动获取专家系统模型如图2所示。

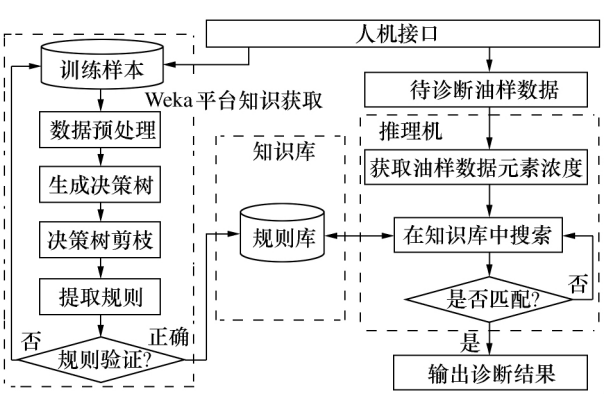

图2 基于 Weka 平台的知识自动获取专家系统模型

在知识库创建和维护阶段,首先确定样本征兆集和故障模式集,建立由训练样本库构成的决策表,然后用 C4.5 决策树理论中数据补齐、数据离散、决策树剪枝及规则提取等方法进行知识规则自动提取,同时用目前比较流行的十折交叉验证法对提取出的规则进行验证。当被验证的规则的准确率和覆盖率都达到指定要求时,则将获取的知识规则存入知识库。对于待诊断样本,将征兆特征值输入到推理机,按一定推理机制经过推理即可得到诊断结果。该模型的主要特点是,利用国际上比较有名数据挖掘平台来作为自己开发的专家系统的知识自动获取模块。这样只要训练样本集具有代表性,提取出的规则的正确性就比较有保证。与此同时,整个知识获取的过程是由软件自动完成,不需要人工干预。

图 3 为专家系统知识自动获取界面。该系统已应用于针对某军用飞机滑油系统磨损故障诊所开发的飞机发动机磨损检测与故障诊断专家系统(图 4 所示)。

图 3　专家系统知识自动获取界面

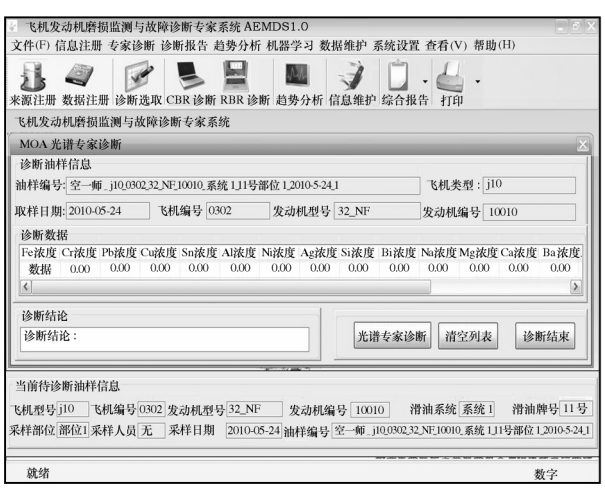

图 4　专家系统推理诊断界面

2.2　基于 Weka 平台 C4.5 决策树知识获取技术

2.2.1　连续属性的离散

针对连续属性,C4.5 算法主要通过下列途径来处理。设在集合 T 中,连续属性 A 的取值为$\{v_1, v_2, \cdots, v_m\}$,则任何在 v_i 和 v_{i+1} 之间的任意值都可以把训练集分成两个部分,即 $T_1 = \{t \mid A \leqslant v_i\}$,$T_2 = \{t \mid A > v_i\}$,因此总共有 $m-1$ 种分割情况。对属性 A 的 $m-1$ 种分割的任意一种情况,作为该属性的两个离散取值,重新构造该属性的离散值,再计算每种分割所对应的信息增益率。然后选取最大增益率的分割作为属性 A 的分支,即 $\text{threshold}(V) = v_k$,其中,$v_k$ 对应的信息增益率为最大。

2.2.2　C4.5 决策树算法

C4.5 算法由 Quinlan 于 1993 年提出,它是一种有指导归纳学习算法,继承了 ID3 算法的全部优点并对其作出改进,在 ID3 算法的基础上扩展了一些 ID3 算法所不能处理的问题,其特点如下:

1) 采用信息增益率来选择属性,克服了用信息增益选择属性时偏向选择取值多的属性的不足;

2) 不仅能处理离散值属性,而且能处理连续值属性;

3) 能对不完整数据集(如个别属性值未知)进行处理;

4) 降低错误修剪率;

5) 提高计算效率等。

C4.5 算法采用信息增益比来描述属性对分类的贡献,用以消除偏向具有大量属性值属性的偏差。设样本集 T 按类别属性 A 的 s 个不同的取值,划分为 $T_1, \cdots, T_s$ 共 s 个子集,则用 A 对 T 进行划分的信息增益为

$$\text{Gain}(A,T) = I(T) - \sum_{i=1}^{s} \frac{|T_i|}{|T|} \times I(T_i) \quad (1)$$

式中: $I(T)$ 表示 T 的信息熵。设 T 中有 m 个类,则

$$I(T) = -\sum_{j=1}^{m} p_j \times \log_2(p_j) \quad (2)$$

式中: p_j 表示 T 中包含类 j 的概率。

用 A 对 T 进行划分的信息增益率为

$$\text{Ratio}(A,T) = \frac{\text{Gain}(A,T)}{\text{SplitInfo}(A,T)} \quad (3)$$

其中

$$\text{SplitInfo}(A,T) = \sum_{i=1}^{n} \left(\frac{|T_i|}{|T_i|}\right) \times \log_2\left(\frac{|T_i|}{|T|}\right) \quad (4)$$

采用此增益率去划分属性得到决策树,其中每个节点取具有最大信息增益率的属性。此方法简单高效、结论可靠,无需很强的相关知识。

具体的算法步骤如下:

输入: 训练样本 samples,候选属性的集合为attribute_list

输出: 由训练数据产生一棵决策树

1) 对训练样本 samples 各项属性数据进行预处理;

2) 创建根结点 root,并确定 attribute_lists 叶结点属性;

3) 计算候选属性 attribute_lists 中每个属性,选取 Gain-Ratio(X) 最大且同时获取的信息增益 Gain(X) 属性又不低于所有属性平均值的属性作为测试属性;

4) 将当前选中的属性赋值给当前结点,将该属性的属性值作为该属性的分叉结点,并且将这些分叉结点插入队列中;

5) 从后选属性 attribute_lists 中将当前使用属性删除;

6) 从队列中取出一个节点,递归进行步骤 3)到步骤 5),直到候选属性 attribute_lists 为空;

7) 为每个叶子节点分配类别属性,对相同的类别属性进行合并,将其进行约减。

基于以上决策算法得到的决策树数据模型,在该模型中之所以选取信息增益率大而信息增益不低于平均值的属性,是因为高信息增益率保证了高分枝属性不会被选取,从而决策树的树形不会因某节点分枝太多而过于松散。过多的分枝会使得决策树过分地依赖某一属性,而信息增益不低于平均值保证了该属性的信息量,使得有利于分类的属性更早地出现。

2.2.3 决策树的剪枝

当得到了完全生长的决策树后,为了消除噪声数据和孤立结点引起的分枝异常,对决策树采取剪枝策略。决策树的剪枝是针对训练数据过分适应问题而提出来的,其修剪方法通常利用统计方法删去最不可靠的分支,以提高分类识别的速度和数据准确分类的能力。其实质是消除训练集中的孤立点和噪声。

C4.5 采用悲观错误修剪法,因为用生成决策树的训练数据集来检验误判率时,实际上对错误的估计过于乐观了,因为决策树是由训练数据集生成的,所以,在多数情况下决策树与训练数据集是符合的。但把决策树用于对训练数据以外的数据进行分类时,很明显这时的错误率将大大增加。基于以上原因,Quinlan 借用二项分布对训练数据中的误判率加以修正,以得到更为符合实际的错误率。显然,与修正前的错误率相比,修正后的错误率增大了不少,因此认为它对错误率的看法是“悲观”的。

算法简化过程如下: 对决策树上所有非叶节点 A 进行计算分析。从树的根节点开始,计算每个分枝节点被剪也即被叶替代后的误判率。采用训练数据集作为测试集,取置信区间的上限作为对误判率的估计。给定一个显著性水平度 α(C4.5 算法中默认 $\alpha=0.25$) ,显然错误的总数服从二项分布,则有

$$P\left[\frac{|p-p^e|}{\sqrt{p^e(1-p^e)/N}}>u_{1-\alpha}\right]=\alpha \tag{5}$$

式中: $p=\frac{E}{N}$是实际观测到得误判率,E 为修剪后出现的错误实例树,N 为被修剪的子树下的实例总数。

令 $z=u_{1-\alpha}$,取置信区间的上限作为这个节点的误判率的估计,则该节点的误判率的计算公式为

$$p^e=\frac{p+\frac{z^2}{2N}+z+\sqrt{\frac{p}{N}-\frac{p^2}{N}+\frac{z^2}{4N^2}}}{1+\frac{z^2}{N}} \tag{6}$$

式中: $p=E/N$ 为观测的误判率; p^e 为估计的误判率。

设定期望误判率的最大值为 C,若剪枝后估计的误判率 p^e 高于 C 时,则保留原来的分枝,否则剪去该分支,用叶片代替。

2.2.4 决策树规则提取

剪枝后生成的决策树,可以直接从决策树中提取相应的决策规则。决策树具有直观性、易理解等特点。分类规则是用 IF-THEN 形式表示,每条规则都是一条从根到叶节点的路径。叶结点表示具体的结论,而叶结点以上的结点及其边表示相应条件的条件取值。从决策树到决策规则见图 5 所示。

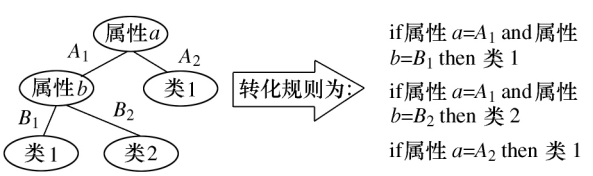

图5 决策树到规则的转换

2.2.5 推理机

推理机作为专家系统的组成控制机构,能通过运用由用户提供的征兆数据,从知识库中选取相关的知识并按照一定的推理策略进行推理,直到得出相应的结论[9~11]。推理机应考虑推理方法、推理方向和推理策略三方面,由于发动机磨损故障诊断的知识不多,知识库结构相对简单,因此,采用精确推理方法,正向推理方向和穷尽式搜索策略。

3　诊断实例

以某军用航空发动机油样分析数据为例，该数据包含了 10 台航空发动机在正常状态下和磨损状态下的 237 个样本，原始数据见文献[10]。Fe、Al、Cu、Cr、Ag、Ti、Mg 这 7 种元素的含量作为样本实例的条件属性分别对应于(A_1 ~ A_7)。磨损状态“F”分为:“1”——正常状态、“2”——轴间轴承磨损、以及“3”——轴间轴承磨损且保持架断裂 3 种形式。磨损状态“F”作为实例的决策属性 D。

表 1 所示为光谱油样分析部分原始数据。由于样本数相对较少，实验中采用目前最流行的 10 折交叉验证准则来比较和评价算法，即将初始样本集划分为 10 个近似相等的数据子集，每个数据子集中属于各分类的样本所占的比例与初始样本中的比例相同，在每次实验中用其中的 9 个数据子集组成训练样本，用剩下的一个子集作为测试集，轮转一遍进行 10 次实验。

表 1　光谱油样分析部分原始数据

Fe(A_1)	Al(A_2)	Cu(A_3)	Cr(A_4)	Ag(A_5)	Ti(A_6)	Mg(A_7)	F(D)
0.50	0.00	0.30	0.00	0.10	0.50	2.00	1.00
1.60	0.00	0.60	0.00	0.10	0.60	2.90	1.00
2.60	0.00,	0.90	0.20	0.20	0.70	3.50	1.00
2.30	0.00	0.60	0.10	0.20	0.50	4.80	1.00
2.60,	0.00	0.60	0.20	0.20	0.60	4.40	1.00
15.60	0.50	2.40	1.40	0.50	1.10	7.20	2.00
3.20	0.00	0.70	0.30	0.20	0.70	5.10	1.00
4.80	0.00	1.50	0.20	0.10	1.00	6.10	1.00
23.90	1.80	9.80	1.10	1.80	1.90	9.30	3.00

表 2 为各属性增益率变化的计算结果，由于信息增益率变化量最大意味着属性的重要性越大，显然，可以看出，属性的重要性依此为: Fe，Ag，Cu，Mg，Ti，Al，Cr。

表 2　属性重要性度量

属性	Fe	Al	Cu	Cr	Ag	Ti	Mg
信息增益率	0.62	0.11	0.26	0.10	0.45	0.17	0.21

图 6 为光谱油样在 C4.5 算法下生成的决策树。表 3 为从决策树中提取的三条规则。由提取的规则可以看出，仅用 3 条规则两个属性就表达了 237 条友油样数据中所蕴含的规律。实现了对数据冗余特征的压缩和信息浓缩。很好地解决了如何用少量元素判断发动机工作状态的问题。其中:

1) 规则 1(Fe≤5.8⇒磨损状态为正常，表示当油样中铁元素含量较低时，发动机的磨损状态处于正常。

2) 规则 2(Fe > 5.8) &(Ag≤0.4) ⇒磨损状态为正常，表示当油样中铁元素含量较高，同时银元素含量较低时，发动机的磨损状态处于正常。

3) 规则 3(Fe > 5.8) & (Ag > 0.4) ⇒ 轴间轴承磨损，表示当油样中铁元素含量较高，同时银元素含量较高时，发动机的轴间轴承磨损严重。

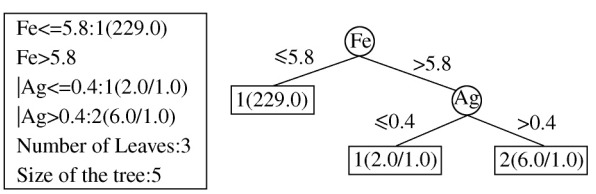

图 6　Weka 平台的 C4.5 决策树算法生成的决策树及其可视化形式

表 3　规则提取结果

规则 1	If Fe≤5.8 Then 磨损状态为正常
规则 2	If Fe > 5.8 And Ag ≤0.4 Then 磨损状态为正常
规则 3	If Fe > 5.8 And Ag > 0.4 Then 轴间轴承磨损

第三类故障，轴间轴承磨损且保持架断裂之所以没有被提取出规则，是因为整个 237 个样本中只有 2 个该类型的故障样本，所以在决策树剪枝过程中被当做噪声数据剪掉。表 4 为用 10 折交叉验证对提取出规则的验证结果。结果表明规则具有很高的精度。

表 4　规则验证的结果

实例数	识别数	误识数	识别率	误识率
237	232	5	97.8%	2.2%

(下转第 1964 页)

对比表1及表2可以看出，带式制动器测点测试结果与计算结果趋势基本一致，说明了带式制动器有限元分析模型较合理。但由于带式制动器有限元分析过程中，是将带式制动器组件作为单一零件来处理，而实际带式制动器组件是由很多零件、多种材料所组成的装配体，这些都与实际有差别，因此测量结果与计算结果存在一定的差别。

5 结论

1）制动上带（紧边）应力值大于制动下带（松边）的应力值，在沉头螺钉孔处均出现应力集中，并且紧靠制动上带的第一个沉头螺钉孔处应力最大；

2）由于截面积较大，在制动上带与制动下带连接处、制动上带固定处及拉杆作用点处的应力值均小于制动带其它位置表面应力；

3）由于制动轮缘的影响，制动带内表面径向位移近似为零，只是沿周向被拉长；制动带最大位移发生在拉杆作用点处；

4）有限元分析过程中，带式制动器组件是作为单一零件来处理的，并且没有考虑焊接等因素，这与实际有差别，建议在以后建立有限元分析模型时，将这些因素都考虑进去，构成更加完善的模型。

参考文献

[1] Yildiz Y，Duzgun M. Streess analysis of ventilated brake discs using the finite element method [J]. **International Journal of Automotive Technology**，2010，11(1)：133～138

[2] Kim D J，Lee Y M，Park J S，Seok C S. Thermal stress analysis for a disk brake of railway vehicles with consideration of the pressure distribution on a frictional surface [J]. **Materials Science and Engineering A**，2008，483-484(15)：456～459

[3] 程相印. 外抱带式制动器的强度设计计算 [J]. 工作研究，2004，(3)：19～23

[4] 冯志强，梁海洋，雷炳怀. 起锚机若干零部件在强度计算中的负荷确定探讨 [J]. 中外船舶科技，2005，(3)：1～4

[5] 李随良. 带式制动器分析及设计 [J]. 设计研究，2004，(3)：14～15

[6] Floquet A，Dubourg M C. Realistic braking operation simulation of ventilated disk brakes [J]. **ASME Journal of Tribology**，1996，118：466～472

[7] 杨莺，王刚. 机车制动盘三维瞬态温度场与应力场仿真 [J]. 机械科学与技术，2005，24(10)：1257～1260

[8] 白金泽等. 应用ANSYS进行复杂结构应力分析 [J]. 机械科学与技术，2003，22(3)：441～446

[9] 吴萌岭. 准高速客车制动盘温度场及应力场的计算与分析(上) [J]. 铁道车辆，1995，33(9)：6～8

[10] 郭乙木编. 有限元法与MSC. Nastran软件的工程应用 [M]. 北京：机械工业出版社，2006

(上接第1959页)

4 结束语

针对航空发动机磨损故障诊断专家系统，知识获取困难的瓶颈问题，提出了一种基于Weka平台的C4.5决策树的知识自动获取方法。建立了航空发动机磨损故障诊断专家系统知识获取模型。阐述了连续属性的离散、决策树的构建、决策树的修剪、及决策规则的提取等关键技术。最后，开发出了飞机发动机磨损检测与故障诊断专家系统（AEMDS1.0），并且采集一组真实航空发动机磨损故障数据对该方法进行了验证，验证结果充分说明了C4.5决策树知识获取方法的有效性。将决策树理论的知识获取方法运用于航空发动机磨损故障系统的知识自动获取，将有效地提升专家系统的智能水平和知识获取能力。

参考文献

[1] 陈果，左洪福. 基于知识规则的发动机磨损故障诊断专家系统 [J]. 航空动力学报，2004，19(1)：23～29

[2] 宋兰琪，汤道宇，陈立波，毛美娟. 航空发动机滑油光谱专家系统知识库建立 [J]. 航空学报，2000，21(5)

[3] 陈果，宋兰琪，陈立波，张占刚. 基于粗糙集理论的航空发动机滑油光谱诊断专家系统知识获取方法研究 [J]. 机械科学与技术，2007，26(7)：898～901

[4] Han J W，Micheline Kamber 著，范明，孟小峰译. 数据挖掘概念与技术 [M]. 北京：机械工业出版，2001

[5] 胡小平，韩泉东，李京浩著. 故障诊断中的数据挖掘 [M]. 长沙：国防科技大学出版社，2009

[6] Quinlan R. **C4.5：Programs for Machine Learning** [M]. Morgan Kaufmann Publishers，San Mateo，CA，1993

[7] Ge G T，William W G. Classification of premalignant pancreatic cancer mass-spectrometry data using decision tree ensembles [J]. **BMC Bioinformatics**，2008，9

[8] Lan H. Witten，Eibe Frank 著，董琳，邱泉，于晓峰，吴韶群，孙立骏译. 数据挖掘实用机器学习技术 [M]. 机械工业出版社，2006

[9] 杨善林，倪志伟. 机器学习与智能决策支持系统 [M]. 北京：科学出版社，2004

[10] 张荣梅. 智能决策支持系统研究开发及应用 [M]. 北京：冶金工业出版社，2003

[11] 文振华. 智能诊断专家系统知识获取方法研究及应用 [D]. 南京：南京航空航天大学，2006