惯性仪器故障诊断模型设计与实现

李笔锋,李富荣,于建立,秦 浩
(海军航空工程学院青岛分院,山东 青岛 266041)

摘 要:为了挖掘隐藏在惯性仪器测试数据背后的信息知识,解决数据丰富而知识贫乏的问题,运用数据挖掘技术筛选出典型的故障测试数据,借鉴CRISP-DM行业标准并以Clementine12.0为平台进行惯性仪器故障诊断模型的设计与实现。提出一种基于两阶段聚类的C5.0算法,即在两步聚类和k-means聚类的基础上使用C5.0算法,与传统C5.0算法相比,提高了预测精度和普适能力。结果表明,基于两阶段聚类的C5.0模型具有较好的分类能力和较强可解释性,为建立基于数据挖掘技术的惯性仪器故障诊断系统提供了研究基础。

关键词:惯性仪器;两步聚类;k-means聚类;C5.0
中图分类号:TP39 **文献标识码**:A **文章编号**:1673-629X(2012)01-0143-04

Design and Implementation of Inertial Apparatus Fault Diagnosis Model

LI Bi-feng, LI Fu-rong, YU Jian-li, QIN Hao
(Qingdao Branch of Naval Aeronautical and Astronautical University, Qingdao 266041, China)

Abstract: In order to tap the information and knowledge hidden behind the test data of inertial apparatus, solving the problem that data is rich but information is poor, apply data mining technology to select typical fault test data, referring to CRISP-DM and taking Clementine12.0 as the platform to design and implement the model of inertial instrument fault diagnosis. Propose C5.0 algorithm based on two-stage clustering that is using C5.0 based on two-step and k-means clustering, compared with traditional C5.0 algorithm, the prediction accuracy and universal capacity has been improved. The results prove that the model of C5.0 based on two-stage clustering has good classification capacity and strong interpretability, which provides a research base for fault diagnosis system of inertial instrument based on data mining technology.

Key words: inertial apparatus; two-step clustering; k-means clustering; C5.0

0 引 言

惯性仪器是陀螺仪、加速度表等惯性仪表和陀螺稳定平台以及捷联惯性测量组合等惯性测量装置的总称,是飞行器制导与控制系统中的核心部件[1]。因此,无论部队还是研究部门对惯性仪器的性能测试都十分重视。针对惯性仪器的测试特点,部队对其测试内容主要包括功能测试和精度测试,具体测试项目包括:电源电压测试、电流测试、电阻阻值测试、位置性能测试、回路前放测试、陀螺惯性时间测试、陀螺漂移测试、陀螺修正速度测试、加速度表启动电流测试、程序功能检查、温控系统和温度状态监测等。

惯性仪器的测试是部队日常训练、执行任务及定期维护保养所必须进行的内容,日积月累就积攒了大量惯性仪器LRU(外场可更换单元)的测试数据,其中不乏典型的故障数据。面对这些海量的高维数据,其中又有许多空缺的、冗余的、含噪声的、不一致数据,使得传统的故障诊断方法无法直接利用这些数据信息。另外,由于故障的随机性、模糊性和不确定性,一个故障的形成往往是众多因素造成的结果,且各因素之间的联系又非常复杂,这使得传统的故障诊断方法已不能满足现代设备的要求。而与此同时,数据挖掘技术的迅猛发展正好可以提供一个故障诊断的应用平台,其在特诊提取、状态识别、诊断决策支持等方面具有独特优势,在对数据处理过程中自动生成知识规则、自动学习故障诊断知识模型,克服了专家系统知识获取瓶颈以及智能故障诊断方法所带来的诊断推理过程解释困难等问题[2,3],为我们提供了一种如何从大量的数据中提取有效数据从而帮助我们进行正确决策分析的方法,而基于数据挖掘技术的故障诊断也是目前一个重要研究方向。

收稿日期:2011-06-05;**修回日期**:2011-09-16
基金项目:海军资助项目(TJ2007_0011)
作者简介:李笔锋(1983-),男,湖北枣阳人,硕士研究生,研究方向为惯性测试技术与仪器;李富荣,副教授,硕士生导师,研究方向为惯性测试技术与仪器。

1 基于数据挖掘的故障诊断分类模型

数据挖掘工具选用 SPSS 公司的 Clementine12.0 作为平台，借鉴 CRISP-DM (Cross-industry Standard Process for Data Mining)行业标准，逐步以专业理解、数据理解、数据准备、建立模型、模型评估的步骤实施惯性仪器的故障诊断过程。

1.1 专业理解

本阶段主要是结合专业背景知识明确数据挖掘的目标，而挖掘目标的定义要求非常明确，任何不明确的定义都会严重影响模型建立的准确性和应用时的效果。这就要求对专业中存在的诸多问题进行深入调查和了解，并对这些疑问转化为数据挖掘问题的可行性进行分析。针对惯性仪器 LRU 测试数据的特点，筛选出典型故障数据，以建立一个基于数据挖掘的故障诊断分类模型作为目标。定性归纳故障字段(输入)和故障类别(输出)，定量分析并建立分类模型，而通过样本学习不断完善模型，为建立基于数据挖掘技术的惯性仪器故障诊断系统提供研究基础，同时也为测试人员进行数据分析提供一种研究方法。

1.2 数据理解

数据来源于两个航空部队同类型某轰炸机惯性仪器 LRU 的测试数据，测试系统为某军事院校自主研发的仪电综合检测与诊断设备，测试时间为 2009 年 3 月至 2011 年 3 月，存储格式为 Microsoft Office Access 数据库，数据类型既有数值型也有分类型。每个惯性 LRU 都包括很多测试字段，对应相应的名称和编号，鉴于名称较长，处理时均以编号表示，下文中仅对涉及到的编号予以字段说明。

1.3 数据准备

此阶段是数据挖掘中的一个重要环节，通常也称作数据预处理，约占数据挖掘过程 70% 的工作量，包括数据选择、清理、重构、整合及归约等[4]。观察分析发现，并非每条测试记录都是完整的，其中由于采集时间短或测试突然中断等原因不乏缺失值、异常值和乱码符号，处理时对于记录中缺失字段较多的予以删除，个别空缺或乱码则采用相关算法进行填充，Data Audit 节点可完成此操作。而对于文字性描述的字段需要将其数字化，例如测试通过用“0”替换，测试不合格用“1”替换，Reclassify 节点可实现此操作。

根据故障数据的分布特点，选择 8 个惯性仪器 LRU 作为故障诊断类别，用 1 ~ 8 表示，分别为组合陀螺故障、航向位置指示器故障、角速度信号器故障、垂直陀螺故障、角位移传感器故障、垂直陀螺仪故障、角速度陀螺仪组故障和航向联系盒故障，用数字 9 表示两个以上惯性 LRU 故障，用数字 0 表示全部 LRU 工作正常。只有 LRU 的每个测试字段均合格，该 LRU 工作正常，否则即判为故障。造成每个 LRU 故障的测试字段较多，不可能均用来参与建模，对于个别测试字段出现的故障记录比例甚少可认为是测试设备不稳定因素造成，重测正常，俗称“RTOK”问题，这样的字段应剔除。这里采用软件中的特征选择节点 Feature Selection 进行维归约，以 Pearson 卡方检验对字段的重要性进行判断，阈值取 0.65，大于阈值的为重要字段，经过筛选，每个 LRU 的典型故障字段为 1 ~ 5 个，均用编号表示，例如角速度信号器仅有的一个典型故障字段 240005 表示 A 相工作电流的起动时间测量。整理合并后形成 25 个典型故障字段作为输入，前面确定的故障类别设为输出，每个故障类别筛选出 100 条测试记录数据，其中故障类别 9 为 200 条，故障类别 0 为 1000 条，共计 2000 条有效测试记录数据用于建模。鉴于数据预处理阶段每个步骤都是反复进行，数据节点流程复杂，下面仅给出数据预处理阶段针对组合陀螺的节点流程图，如图 1 所示。数据源数据依次经过过滤(Filter)、实例化(Type)、纵向合并(Append)、字段选择(Select)、属性值替换(Reclassify)、排序(Sort)、故障字段选择(Select)、横向合并(Merge)、空缺值补充(Data Audit)、特征选择节点(FS)后将组合陀螺的典型故障数据输出到 Excel 表格，用于后续数据再处理。

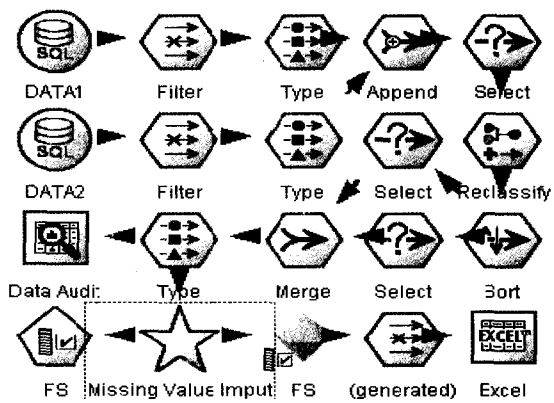

图 1 数据预处理节点流程图-组合陀螺

1.4 建立模型

本质上，故障诊断是一个模式分类与识别问题，即把系统或设备的状态分为正常和异常两类，并判别异常的样本究竟又属于哪种故障[5,6]。因此，此次建模的目的是对惯性仪器 LRU 故障样本做出正确地分类与识别，识别的过程即是诊断的过程。传统的分类方法有很多，如统计算法中的 Logistic 回归、计算机算法中的神经网络、机器学习算法中的决策树等。鉴于 Clementine12.0 模块限制，选用 C5.0 算法作为分类方法。C5.0 是决策树算法之一，它能产生决策树或规则集，在决策树的产生过程中能自动根据最大信息增益进行样本拆分，一直到样本子集不能再拆分为止[7]。而规则集则是规则的集合，它用一种更加简炼的方式陈述

决策树中的信息,结合文中的研究,规则集更加有利于对挖掘信息地解释和利用,因此,选择使用 C5.0 算法产生规则集。

由于惯性测试中存在很多不稳定因素,诸如电源变化、环境条件、基座运动、软件漏洞、人为操作等必然导致测试异常结果的出现,这对于分类预测意义不大,同时也会影响分类精度,但需要确定这样的数据,便于测试和研究人员根据奇异值特点进行故障字段调整或 LRU 部件维修。于是提出一种基于两阶段聚类的 C5.0 分类算法,在两步聚类和 k-means 聚类的基础上使用 C5.0 算法[8]。

k-means 是使用最为广泛和经典的划分方法,它以 k 为参数,把 N 个对象分为 k 个类,使类内具有较高的相似度,而类间的相似度较低[9]。其复杂度接近线性,比较适合对大规模数据进行挖掘,伸缩性好且简单易行,具有高效性。其缺点也同样突出,一是必须指定 k 值,且对初始中心点的选择比较敏感,如果初始值选择不当,将会收敛成为一个局部最小的准则函数;二是对噪声和异常数据敏感,少量的该类数据能够对平均值产生极大的影响。而利用两步聚类算法得到的初始类可以弥补 k-means 方法对初始类选择的需求,同时通过层次聚类发现的异常值,可以通过除去异常值再将剩下的数据集作为 k-means 聚类的基础数据,有效减缓了噪声和异常数据对平均值的负面影响[10,11]。

建立基于两阶段聚类的 C5.0 模型如图 2 所示。首先经过两阶段聚类(Two Step)得到初始类个数,即 k-means 初始聚类值 k,但为了挖掘出异常值,需要增加 k 值,通常取 1,作为异常值的类别值,通过 k-means 聚类即可得出异常的测试值,筛选出这些异常值作为维修人员和研究人员进一步研究相关 LRU 的参考数据,剩下的测试正常数据和典型故障数据经分割节点(Partition)划分为训练数据和检验数据,最后由 C5.0 模型进行分类并产生规则。每个模型节点所生成的结果都加入到数据流中,其中 C5.0 模型结果节点中包含着分类规则,经过完善即可作为 LRU 故障诊断的分类模型,只要将数据按照事先规定的格式存入 Excel,经过此节点即可知道数据状况以及所属故障类别。整个数据挖掘过程浑然一体,凸显出 Clementine 软件可视化的优势。

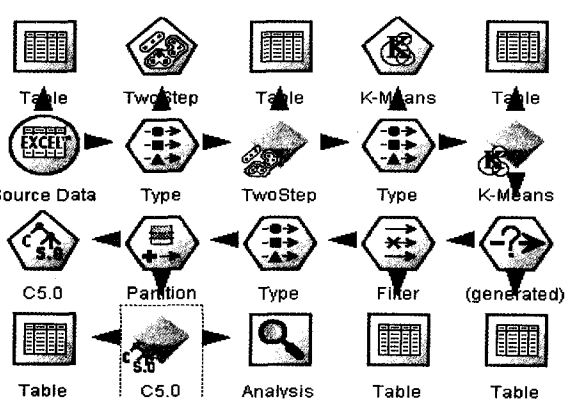

图 2　基于两阶段聚类的 C5.0 模型

1.5　结果分析及模型评价

分阶段执行模型结果,经过两步聚类自动确定出两个类别,样本记录分别为 1313 和 687,简单描述为正常数据和故障数据,添加一类作为离群类的存取,即设定 k-means 聚类初始 k 为 3。k-means 聚类结果显示为三类,样本记录分别为 1837、130 和 33,可理解为正常数据和故障数据、疑似异常数据和异常数据,筛选 33 条异常测试记录,通过表格分析或散点图观察,造成异常的主要字段是 344230,字段名称为俯仰对倾斜交叉影响,是角速度陀螺仪组其中的一个字段,用于输出性能测试的描述,充分说明了角速度陀螺仪出现异常的频率较高,需要重点关注。

在执行 C5.0 算法时,产生规则集,取 70% 样本作为训练样本,30% 样本作为检验样本,经过反复训练学习,当修剪纯度设为 50,子分支最少记录数设为 1 时分类效果最佳。利用 Clementine 的 Analysis 节点生成分类预测结果的分析值如图 3 所示。对训练样本的分类正确率达到 98.83%,对检验样本的分类正确率达到 97.67%。

⊟-Results for output field 故障类别

⊟-Comparing $C-故障类别 with 故障类别

'Partition'	1_Training		2_Testing	
Correct	1,436	98.83%	502	97.67%
Wrong	17	1.17%	12	2.33%
Total	1,453		514	

图 3　C5.0 分类预测结果

同时,建模过程可以将产生的规则集显示出来,直观明了,生成的规则集部分如图 4 所示。以规则 3 为例,当字段 240005 大于 0.3 并且 260025 小于等于 0.163 时,判断故障类别为 3,即角速度信号器故障。

利用预测准确率、查准率和查全率评估模型对惯性 LRU 故障数据的分类能力,并与传统 C5.0 模型做比较。预测查准率是预测为故障记录中实际故障记录的比例,体现了模型对故障记录的预测是否精确。预测查全率是实际故障记录中预测为故障记录的比例,体现了模型预测结果的覆盖程度[12]。这三个指标越大,表明模型的预测效果越好。根据 Analysis 节点的 Coincidence matrices for $-N 故障类别值综合计算这三个指标生成表 1,比较可知基于两阶段聚类的 C5.0 算法在分类预测效果上明显优于传统 C5.0 算法,可以应用于惯性仪器故障诊断之中。

```
Rules for 1 - contains 4 rule(s)
Rules for 2 - contains 3 rule(s)
Rules for 3 - contains 1 rule(s)
  Rule 1 for 3.0
    if   240005.000000 > 0.300
    and  260025.000000 <= 0.163
    then 3.000
Rules for 4 - contains 3 rule(s)
  Rule 1 for 4.0
    if   180121.000000 > 130.362
    and  240005.000000 <= 0.288
    and  260025.000000 > 0.044
    then 4.000
  Rule 2 for 4.0
    if   260015.000000 > 58.445
    and  333010.000000 <= 6.440
    then 4.000
  Rule 3 for 4.0
    if   240005.000000 > 0.192
    then 4.000
Rules for 5 - contains 2 rule(s)
  Rule 1 for 5.0
    if   320420.000000 > 0.032
    then 5.000
  Rule 2 for 5.0
    if   260025.000000 <= -0.195
    and  320420.000000 <= -0.011
    then 5.000
Rules for 6 - contains 2 rule(s)
```

图 4 C5.0 生成的规则集

表 1 传统 C5.0 与改进 C5.0 算法预测结果指标值

算法	准确率(%)	查准率(%)	查全率(%)
传统 C5.0	91.65	92.00	90.77
改进 C5.0	98.53	98.65	98.34

2 结束语

针对传统故障诊断方法在数据预处理以及知识获取瓶颈方面存在的问题，文中尝试将数据挖掘技术应用于惯性仪器 LRU 故障诊断之中，在对大量不乏典型故障信息测试数据定性分析的基础上，应用基于两阶段聚类的 C5.0 分类算法，剔除异常值后利用生成的规则集实现对惯性仪器的故障诊断，并与传统的 C5.0 算法做比较，提高了预测精度和泛化能力。同时验证了基于数据挖掘的惯性仪器故障诊断的理论可行性，对于数据挖掘技术应用于惯性仪器测试领域具有一定的推广价值。但是，如何统筹解决硬件系统和软件平台问题，真正建立一个基于数据挖掘技术的惯性仪器故障诊断系统，有待进一步开展实践性研究。

参考文献：

[1] 余 红. 惯性仪器通用测试技术应用研究[J]. 工业控制计算机，2005，18(11)：19-20.

[2] 刘洁瑜，钱培贤. 惯性仪器测试标定数据库的数据挖掘技术[J]. 计算机工程，2004，4(2)：11-12.

[3] 高毅龙. 数据挖掘及其在工程故障诊断中的应用[D]. 西安：西安交通大学，2000.

[4] 廖 芹，郝志峰，陈志宏. 数据挖掘与数学建模[M]. 北京：国防工业出版社，2010.

[5] Yeh Ruey-Ling, Liu Ching, Shia Ben-Chang, et al. Imputing manufacturing material in data mining[J]. Journal of Intelligent Manufacturing, 2008, 19(1): 110-113.

[6] Fong A C M, Hui S C. An intelligent online machine fault diagnosis system[J]. Computing & Control Engineering Journal, 2001, 25(10): 217-220.

[7] 邱 涛，李 雯. 决策树算法在智能导学系统中的应用[J]. 计算机技术与发展，2009，19(12)：191-192.

[8] 吴玉霞，牟援朝. 基于两阶段聚类的洗钱行为识别[J]. 计算机工程，2010，36(15)：60-62.

[9] 朱云贺，张春海，张 博. 基于数据分段的 K-means 的优化研究[J]. 计算机技术与发展，2010，20(11)：131-132.

[10] Hosseini S M S, Maleki A, Gholamian M R. Cluster analysis using data mining approach to develop CRM methodology to assess the customer loyalty[J]. Expert Systems with Applications, 2010, 37(7): 5260-5262.

[11] 李 睿，肖维民. 基于孤立点挖掘的异常检测研究[J]. 计算机技术与发展，2009，19(6)：169-170.

[12] 周生宝，郭俊芳. 客户流失预测模型设计与实现[J]. 计算机系统应用，2009，46(5)：170-172.

(上接第 132 页)

控制算法[J]. 计算机研究与发展，2008，45(7)：1099-1105.

[3] 解文斌，李 佳，鲜 明，等. 基于拓扑特性的分布式虚拟骨干网算法[J]. 软件学报，2010，21(6)：1416-1425.

[4] 齐迎迎，禹继国. 无线传感器网络的节能分布式分簇算法[J]. 计算机工程，2011，37(3)：83-86.

[5] 李建波，黄刘生，徐宏力. 一种密集部署传感器网络的分簇算法[J]. 计算机研究与发展，2008，45(7)：1106-1114.

[6] 刘林峰，刘 业. 基于热点区域场景的传感器网络拓扑控制算法[J]. 计算机技术与发展，2010，20(10)：8-12.

[7] 郭晓莲，林志伟，黄榕宁. 自组网分簇算法仿真设计[J]. 计算机技术与发展，2007，17(9)：92-95.

[8] Ling Qing, Tian Zhi. Impact of mobility on topology control of wireless sensor networks[C]//WiCom'07. [s. l.]: [s. n.], 2007: 2483-2486.

[9] Li Jianbo, Huang Liusheng, Xiao Mingjun. Energy Efficient Topology Control Algorithms for Variant Rate Mobile Sensor Networks[C]//The 4th International Conference on Mobile Ad-hoc and Sensor Networks. [s. l.]: [s. n.], 2008: 23-30.

[10] 冯雪玲，于 炯，马 俊，等. 基于簇结构的移动 Ad Hoc 网络的认证协议[J]. 计算机技术与发展，2008，18(9)：127-130.

[11] Bao Lichun, Garcia-Luna-Aceves J J. Stable energy-aware topology management in ad hoc networks[J]. Ad Hoc Networks, 2010, 8(3): 313-327.

[12] Zorzi M, Rao R R. Geographic Random Forwarding (GeRaF) for Ad Hoc and Sensor Networks: Energy and Latency Performance[J]. IEEE Transactions on Mobile Computing, 2003, 2(4): 349-365.