# 液体火箭发动机故障的数值型关联规则挖掘

李京浩，胡小平，韩泉东
（国防科技大学 航天与材料工程学院，湖南 长沙 410073）

**摘 要**：数值型关联规则的算法大多是将多值属性关联规则挖掘问题转化为布尔型关联规则挖掘问题，而连续属性的离散化是数值型关联规则的核心问题。本文基于数值型关联规则的理论，用一种数理统计的方法进行连续属性的离散化。将该方法应用于某大型液体火箭发动机稳态段的热试车数据，然后利用 FP-Growth 算法对其进行测试，挖掘出了故障数据，进而验证了其可行性。

**关键词**：液体推进剂火箭发动机；故障诊断；数值型关联规则；FP-Growth 算法
**中图分类号**：V434.1 **文献标识码**：A **文章编号**：(2007) 02-0007-06

# Mining quantitative association rules of liquid propellant rocket engine

Li Jinghao, Hu Xiaoping, Han Quandong
(College of Aerospace and Material Engineering, NUDT, Changsha 410073, China)

**Abstract**: Most quantitative association rules transform mining association rules of numeric property into boolean property, and the kernel problem is to divide the numeric data into intervals. Based on the theory of quantitative association rules, the numeric data is divided into intervals with statistical method. This method is applied to the measured steady data of a large -scale liquid propellant rocket engine and tested with FP-Growth arithmetic. The fault data is mined and the feasibility of the method is verified.

**Key words**: liquid propellant rocket engine; fault diagnosis; quantitative association rules; FP-Growth arithmetic

**收稿日期**：2006-11-14；**修回日期**：2007-01-11。**基金项目**：国家自然科学基金资助项目(50376073)。

**作者简介**：李京浩（1980—），男，硕士研究生，研究领域为数据挖掘、火箭发动机故障诊断。

# 1 引言

液体火箭发动机故障诊断需要检测和分析大量的数据。数据挖掘是从大量的数据中提取隐含在其中的人们事先未知的、但潜在有用的信息和知识的过程。提取的知识表示为概念、规则、规律和模式等形式。

本文介绍了一种数值型关联规则的方法，并利用 FP-Growth 算法对划分区间后的某大型液体火箭发动机热试车数据进行了检测，挖掘出了隐藏着的对决策有重要参考价值的信息。

# 2 基本原理

## 2.1 数值型关联规则

关联规则的挖掘是数据挖掘领域中的一个重要研究课题，它用来挖掘发现大量数据中项集之间有趣的关联或相关联系，由 Agrawal.R 最早提出。它一般表示为 $X \Rightarrow Y$ (sup，conf) 的形式，其中，$X$，$Y$ 为项目集合，简称项集；sup 为规则的支持度；conf 为规则的置信度。

根据规则中处理的变量的类别，关联规则可以分为布尔型和数值型[2]。布尔关联规则挖掘技术只能挖掘出事务数据库中的布尔关联规则。关联规则挖掘技术要得到广泛的应用必须能够处理各种类型的数据集，这就要用到数值型关联规则的挖掘技术。在存在连续属性和多值属性的数据库中，如果能把连续属性和多值属性映射为相应的布尔属性，就可以用布尔关联规则挖掘算法来解决数值型关联规则的挖掘问题。对数值型关联规则挖掘算法主要研究把连续属性和多值属性映射到布尔属性的方法、数值型关联规则挖掘问题的定义及分解、连续属性的离散化、规则的兴趣度等。对数值型关联规则挖掘问题的研究，拓展了关联规则挖掘的范围，使得关联规则挖掘可以在更多领域的数据库中进行。

数值型关联规则挖掘技术是在布尔关联规则挖掘技术的基础上发展起来的。数值型关联规则挖掘技术的中心问题是连续属性的离散化，在完成了连续属性的离散化之后，数值型关联规则挖掘问题就可以映射为布尔关联规则挖掘问题。数值属性和类别属性到布尔属性的映射可通过下面的方法完成：

(1) 对类别属性或取值较少的数值属性，将每个属性值映射为一个布尔属性；

(2) 对取值较多的数值属性，先将其属性值划分为多个子区间，然后把每个子区间映射为一个布尔属性。

给定数据库 DB 和用户指定的最小支持度和最小置信度，数值型关联规则的挖掘可通过以下各步骤来完成：

(1) 确定对各数值属性的划分方法；

(2) 实现连续属性和多值属性到布尔属性的映射；

(3) 找出所有的频繁项目集；

(4) 由频繁项目集导出满足最小置信度的关联规则 (强规则)。

本文采用 FP-Growth 算法寻找频繁项集和挖掘关联规则。

## 2.2 FP-Growth 算法[4]

1999 年 Jiawei Han 等人提出了不产生备选项目集的关联规则挖掘方法，这种方法被称之为频繁模式增长 (Frequent-Pattern Growth)，或简称为 FP-Growth。它采取如下的分治策略：将提供频繁项集的数据库压缩成一棵频繁模式树 (或称为 FP-树)，但仍保留项集关联信息；然后，将这种压缩后的数据库分成一组条件数据库 (一种特殊类型的投影数据库)，每个关联一个频繁项，并分别挖掘每个数据库。算法的实现步骤如下。

按以下步骤构造 FP-树：

(1) 扫描事务数据库 DB 一次。收集频繁项的集合 F 和它们的支持度。对 F 按支持度降序排序，结果为频繁项表 L；

(2) 创建 FP-树的根节点，以“null”标记它。对于 DB 中每个事务 Trans，执行：选择 Trans 中的频繁项，并按 L 中的降序排序。设排序后的频繁项表为 [p|P]，其中，p 是第一个元素；P 是剩余元素的表。调用函数 insert_tree ([p|P] ,T)。该过程执行情况如下：如果 T 有子女 N

使得式 N.item-name=p.item-name 成立，则 N 的计数加 1；否则创建一个新节点 N，将其计数设置为 1，接到它的父节点 T，并且通过节点链接结构将其链接到具有相同 item-name 的节点。如果 P 非空，递归地调用函数 insert_tree (P,N)。

FP-树的挖掘通过调用 FP-growth (FP_tree, null) 实现。该过程实现如下：

procedure FP_growth (Tree, a)

(1) if Tree 含单个路径 P，then；

(2) for 路径 P 中节点的每个组合 (记作$\beta$)；

(3) 产生模式$\beta \cup \alpha$，其支持度 support 等于$\beta$中节点的最小支持度；

(4) else for each $a_i$ 在 Tree 的头部 {；

(5) 产生一模式$\beta=a_i \cup \alpha$，其支持度 support=$a_i$.support；

(6) 构造$\beta$的条件模式基，然后构造$\beta$的条件 FP-树 $Tree_\beta$；

(7) if $Trees_\beta \neq \varnothing$ then；

(8) 调用 FP_growth ($Tree_\beta$, $\beta$) ;}。

和 Apriori 算法相比，FP-Growth 算法有两大优点：

(1) 不需要产生大量候选项集。大型的液体火箭发动机一般都包含几十个属性，发生故障时很多参数会发生变化。如果使用 Apriori 算法，它的候选项集会非常大，给计算带来困难。

(2) 不需要重复地扫描数据库，对于挖掘长的和短的频繁模式，都是有效的和可伸缩的。本文使用的试车数据是以 0.1 秒左右为步长取一组数据，每组数据基本都包含了几千个样本，数据量非常庞大。利用 FP-Growth 算法不用重复地扫描数据，其速度大约比 Apriori 算法快一个数量级。

FP-Growth 算法具体实现过程可参考文献[4]。

## 3 液体火箭发动机数据挖掘实例

本方法主要适用于挖掘液体火箭发动机稳态工作段发生的故障。因此，本文采用某大型液体火箭发动机稳态段的试车数据进行试验。

### 3.1 数据预处理

数据预处理是数据挖掘的一个极其重要的环节。液体火箭发动机试车测得的参数种类很多，数据量非常大，数据采样的步长也可能不一致，这些给数据的选取带来了极大的困难。这里通过选取多次试车数据中共有的且参数值相对稳定的数据，从 80 多个测量参数中初步提取出了 30 个参数的数据。然后采用属性选择的方法（主要选取与易发生故障的组件所对应的参数），将上述 30 个参数进一步压缩到 19 个，用英文字母表示，如表 1 所示。

表 1 筛选后的发动机参数

Tab.1 Selected LRE parameters

| 参数名称 | 符号 | 参数名称 | 符号 |
|---|---|---|---|
| 氢泵转速 | A | 氢泵出口温度 | K |
| 氧泵转速 | B | 冷却套出口压力 | L |
| 氢泵流量 | C | 发生器氢喷前压力 | M |
| 氧泵流量 | D | 发生器氧喷前压力 | N |
| 氧泵入口压力 | E | 燃气发生器压力 | O |
| 氧泵出口压力 | F | 氢涡轮入口压力 | P |
| 燃烧室氧喷前压力 | G | 氧涡轮入口压力 | Q |
| 燃烧室压力 | H | 氢涡轮出口压力 | R |
| 氢泵入口压力 | I | 氧贮箱压力 | S |
| 氢泵出口压力 | J | - | - |

数据处理步骤：

(1) 试车实测数据可能变化较大，为了减小阈值的搜索范围以提高数据挖掘的效率，本文采用了 $(X_{ij}-\overline{X}_j)/\overline{X}_j$ 形式的函数把每个数据转换成与正常数据的统计均值的偏差值，式中 $X_{ij}$ 表示第 $i$ 个样本中第 $j$ 个参数的测量值，$\overline{X}_j$ 表示该型号发动机多次正常试车稳态段第 $j$ 个参数的统计均值。

(2) 连续属性的离散化是数值型关联规则的核心问题。本文采用统计方法对数据进行离散化并划分区间。通过对所有偏差值数据的统计分析可以看出，它大致服从以参数值 $X$=0 为中心的正态分布，如图 1 所示。

因此，可以用假设检验的方法，把数据划分为三个区间：偏小的数据；正常范围内的数据；

3) 偏大的数据。即假定正常数据落在区间[ $X_1$, $X_2$] 内。这时候，原点两侧数据落在正常范围内的置信度 α 的取值就比较关键。经过多次取值和试验分析，这里取 α=0.998，（1−α/2）=0.001。可求出左右两侧的分界点大约为 $X_1$=−0.164，$X_2$=0.325，因为中心点两侧的数据分布不一样，因此分界点也不对称。求出分界点以后，可以划分子区间，区间 1：[$X_{min}$，−0.164]；区间 2：[−0.164，0.325]；区间 3：[0.325，$X_{max}$]，每个子区间包含一个布尔属性，$X_{min}$ 和 $X_{max}$ 分别表示数据的最小值和最大值。如：氢泵转速属性可划分为 A1：[$X_{min}$，−0.164]，A2：[−0.164，0.325]，A3：[0.325，$X_{max}$]；氧泵转速可划分为 B1：[$X_{min}$，−0.164]，B2：[−0.164，0.325]，B3：[0.325，$X_{max}$]，如表 2 和表 3 所示。可以看出：落在区间 1 上的数据是比统计均值偏小的数据，其物理意义相当于参数值的降低；落在区间 3 上的数据是比 $\overline{X_j}$ 偏大的数据，其物理意义相当于参数值的增加。

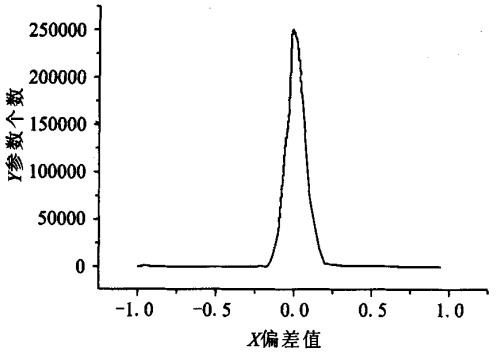


图 1　数据分部图

Fig.1　Distribution of the data

表 2　偏差值数据

Tab.2　Windage value LRE data

| 属性 / 参数项 | 氢泵转速 | 氧泵转速 | … |
|---|---|---|---|
| 1 | −0.11657 | −0.1273 | … |
| 2 | −0.13473 | −0.14043 | … |
| 3 | −0.1512 | −0.15481 | … |
| 4 | −0.16087 | −0.16224 | … |
| 5 | −0.18411 | −0.17866 | … |
| … | … | … | … |

在得出划分属性以后的一组布尔型数据之后，就可以利用 FP-Growth 算法进行挖掘。

表 3　离散化后的数据

Tab.3　The boolean data

| 属性 / 参数项 | A1 | A2 | A3 | B1 | B2 | B3 | … |
|---|---|---|---|---|---|---|---|
| 1 | 0 | 0 | 0 | 0 | 1 | 0 | … |
| 2 | 0 | 1 | 0 | 0 | 1 | 0 | … |
| 3 | 0 | 1 | 0 | 0 | 1 | 0 | … |
| 4 | 0 | 1 | 0 | 0 | 1 | 0 | … |
| 5 | 1 | 0 | 0 | 1 | 0 | 0 | … |
| … | … | … | … | … | … | … | … |

### 3.2　利用 FP-Growth 算法挖掘频繁项集

前面已假定落在区间 2 上的数据为正常数据。而本文需要挖掘的是火箭发动机各参数的异常数据之间的关联规则，因此，在利用 FP-Growth 算法进行挖掘时只考虑区间 1 和区间 3 的数据即可。

本文所用来挖掘的数据是某大型氢氧发动机的 23 次热试车稳态段的数据（即 3 次故障试车数据和 20 次正常试车数据，共 76005 个样本）。

在用 FP-Growth 算法进行挖掘时最小支持度计数的取值也是一个关键问题。如果支持度设定得过高，挖掘的频繁项集太少，甚至不能产生频繁项集，这样挖掘的工作就毫无意义；如果设置得过低，就会产生大量的频繁项集，过多的所谓的信息与数据垃圾是一样的，无法从数据中得到真正有用的决策信息，这对我们分析数据也是毫无用处的。

经过几次试验分析和对比，这里将最小支持度 min_sup 设定为 2 是比较合理的。火箭发动机是一个整体系统，当故障出现时，某一部件的损坏会影响到其他系统，引起其他参数的变化。因此，小的频繁项集对火箭发动机故障分析是没有帮助的，所以本试验在挖掘频繁项集的时候过滤掉了参数个数小于 3 的频繁项集。

通过对这 23 组数据进行挖掘频繁项集，其结果如下表 4 所示。

表 4 对试车数据的挖掘结果

Tab.4 Results of data mining

| 数据代号 | 频繁项集数目 | 所需时间/s | 数据代号 | 频繁项集数目 | 所需时间/s | 数据代号 | 频繁项集数目 | 所需时间/s |
|---|---|---|---|---|---|---|---|---|
| T10 | 968 | 0.078 | T155 | 16 | 0.047 | T1921 | 1 | 0.047 |
| T145 | 0 | 0.031 | T183 | 0 | 0.063 | T1922 | 0 | 0.031 |
| T146 | 0 | 0.031 | T184 | 0 | 0.078 | T1931 | 0 | 0.047 |
| T147 | 0 | 0.048 | T1851 | 0 | 0.063 | T1932 | 0 | 0.015 |
| T1511 | 0 | 0.047 | T1852 | 0 | 0.016 | T20 | 10 | 0.062 |
| T1512 | 0 | 0.016 | T1861 | 0 | 0.047 | T232 | 0 | 0.057 |
| T152 | 0 | 0.063 | T1911 | 1 | 0.047 | T233 | 0 | 0.078 |
| T153 | 0 | 0.062 | T1912 | 0 | 0.016 | | | |

表 5 挖掘出的频繁项集

Tab.5 The mined frequent items

| 数据代号 | T10-1 | T155 | T20 |
|---|---|---|---|
| 频繁项集 | J3 N1 O1 (7) | A1 C1 K1 (3) | E1 J3 K1 (90) |
| | N1 O1 P1 (7) | C1 K1 M1 (2) | D3 J3 K1 (12) |
| | J3 O1 P1 (7) | A1 K1 M1 (2) | D3 E1 K1 (12) |
| | J3 N1 P1 (7) | A1 C1 M1 (2) | D3 E1 J3 (12) |
| | J3 N1 O1 P1 (7) | A1 C1 K1 M1 (2) | D3 E1 J3 K1 (12) |
| | … | C1 K1 N1 (2) | C1 J3 K1 (7) |
| | A1 B1 C1 D1 E3 F1 I1 J3 K1 N1 O1 P1 Q1 (2) | A1 K1 N1 (2) | A3 C1 K1 (3) |
| | A1 B1 C1 D1 E3 F1 I1 J3 K1 M1 N1 O1 P1 Q1 (2) | A1 C1 N1 (2) | A3 C1 J3 (3) |
| | A1 C1 D1 E3 F1 H1 I1 J3 K1 N1 O1 P1 Q1 (2) | A1 C1 K1 N1 (2) | A3 J3 K1 (3) |
| | A1 C1 D1 E3 F1 H1 I1 J3 K1 M1 N1 O1 P1 Q1 (2) | A1 M1 N1 (2) | A3 C1 J3 K1 (3) |
| | A1 B1 C1 D1 E3 F1 H1 I1 J3 K1 N1 O1 P1 Q1 (2) | K1 M1 N1 (2) | |
| | A1 B1 C1 D1 E3 F1 H1 I1 J3 K1 M1 N1 O1 P1 Q1 (2) | A1 K1 M1 N1 (2) | |
| | | C1 M1 N1 (2) | |
| | | A1 C1 M1 N1 (2) | |
| | | C1 K1 M1 N1 (2) | |
| | | A1 C1 K1 M1 N1 (2) | |

用来计算的计算机是普通的 PC 机，配置是：Pentium43.0 CPU 和 1G 物理内存。可以看出 FP-Growth 算法的速度很快，效率非常高。表 5 分析了 T10-1、T155 和 T20 三组数据中挖掘的频繁项集。实际上，T10-1 试车、T155 试车和 T20 是三组稳态故障数据，故障类型分别为氧副文氏管出现多余物、氢涡轮破坏以及氢泵次同步振动。

通过对频繁项集的分析可以看出每个频繁项集的支持度都不大，这是因为故障发生以后很快引起了一系列连锁反应，使得测试系统中止试车的缘故。通过所挖掘出的频繁项集可以很直观地看出 LRE 故障是从系统的某一部件损坏开始，迅速引起其他系统参数的变化。比如 T155 试车数据。（下转第 58 页）

容器生产中。

**参考文献：**

[1] 姜焕中. 焊接方法及设备 [M]. 北京：机械工业出版社，1986.

[2] 周振丰. 金属熔焊原理及工艺(下册)[M]. 北京：机械工业出版社，1986.

[3] 陈裕川. 低合金结构钢的焊接 [M]. 北京：机械工业出版社，1992.

[4] 单黎波. 波纹板加层结构高温钎焊焊缝 X 射线影像分析 [J]. 火箭推进. 2006, 32(2).

(编辑:马 杰)

(上接第 11 页)

其中一个最大频繁项集 "A1 C1 K1 M1 N1 (2)" 的物理意义是氢泵转速 ($A$) 的下降，引起了氢泵流量 ($C$)，氢泵出口温度 ($K$)，发生器氢喷前压力 ($M$)，发生器氧喷前压力 ($N$) 的下降，这与实际试车中出现的氢涡轮破坏的故障现象相吻合。

通过以上测试可以看出 FP-Growth 算法很好地挖掘了液体火箭发动机试车数据中的故障数据，与文献[5]中采用神经网络方法及文献[6]中采用的支持向量机方法进行对比，所得结果也保持一致，故该方法是可信的。

## 4 结束语

将数据挖掘应用于液体火箭发动机的故障检测和诊断中是一种新的思路。本文从数据挖掘的角度，对传统的布尔型关联规则进行了改进，利用数值型关联规则挖掘火箭发动机试车数据中的频繁项集。结果发现，该方法能够准确地检测发动机中发生的故障，对于理论分析和工程应用都具有重要意义。但是关联规则主要用于事后分析试验得出的数据，在故障预报和实时诊断方面的性能不是很好，这些是将来继续有待深入研究的课题。

**参考文献：**

[1] 张育林，吴建军. 液体火箭发动机健康监控技术[M]. 长沙：国防科技大学出版社，1998.

[2] 张会容. 关联规则挖掘的研究及其应用[D]. 上海：华东理工大学，2004.

[3] 邓景毅，张小康. 事务间数值型关联规则的挖掘[J]. 计算机应用，2004, 24 (4) .

[4] 韩家炜，等. 数据挖掘：概念与技术 [M]. 北京：机械工业出版社，2001.

[5] 黄敏超. 液体火箭发动机故障的神经网络诊断研究[D]. 长沙：国防科技大学研究生院，1998.

[6] 韩泉东，胡小平，李舟军. 决策树和支持向量机方法在液体火箭发动机故障诊断中的应用[C].

(编辑：马 杰)