文章编号:1671-4598(2021)05-0189-04 DOI:10.16526/j.cnki.11—4762/tp.2021.05.038 中图分类号:TP18 文献标识码:A

面向飞行器遥测数据的关联规则挖掘方法研究

李 智,张丽晔,褚厚斌,蔡斐华,耿 钧
(中国运载火箭技术研究院,北京 100076)

摘要:飞行器遥测数据是飞行器状态的直接体现,对飞行器遥测数据的不断深入分析和研究,可为飞行器的安全性和稳定性提供有效保障;目前复杂飞行器的遥测数据存在试验数据量大、人工判读效率低、数据间关联关系复杂且不易梳理等问题;同时,数据智能化分析程度低,缺少对海量历史试验数据的有效利用;为克服现有技术不足,通过对飞行器遥测数据的关联规则挖掘方法进行研究,提出基于状态转换提取的关联规则挖掘算法,并与FP—Growth算法进行试验挖掘对比分析,实现对飞行器遥测数据参数的关联规则挖掘分析,有效地解决飞行器遥测数据间关联规则的梳理问题,试验结果准确率高,为飞行器工况与参数的关联规则挖掘提供重要参考意义。

关键词:遥测数据;关联规则;数据挖掘;FP—Growth;状态转换

Study on Association Rules Mining Method for Spacecraft Telemetry Data

Li Zhi, Zhang Liye, Chu Houbin, Cai Feihua, Geng Jun
(China Academy of Launch Vehicle Technology, Beijing 100076, China)

Abstract: Spacecraft telemetry data is a direct manifestation of spacecraft status. Continuous in—depth analysis and research on spacecraft telemetry data can provide effective guarantees for the safety and stability of spacecraft. At present, the telemetry data of complex spacecraft has problems such as large amount of test data, low manual interpretation efficiency, complex associations between data, and difficult to sort out. At the same time, the level of intelligent data analysis is low, and the effective use of massive historical test data is lacking. In order to overcome the shortcomings of the existing technology, by researching the association rules mining method of spacecraft telemetry data, the association rule mining algorithm based on state transition extraction is proposed, and conduct experimental mining and comparative analysis on the FP—Growth algorithm, implement the mining and analysis of the association rules of spacecraft telemetry data parameters, and effectively solve the problem of rectifying the association rules among the spacecraft telemetry data. The test results are highly accurate and provide important reference for mining the association rules of spacecraft operating conditions and parameters.

Keywords: telemetry data; association rules; data mining; FP—Growth; state transition

0 引言

近年来数据挖掘技术得到了迅速的发展,如相似模式搜索,模式聚类,事件检测,规则提取等技术和方法已经成功应用于金融、医疗、生物等领域[1]。随着近年来飞行任务的不断增加,大量飞行器的飞行试验积累了海量遥测数据。飞行器遥测数据是地面运管系统判断其运行状态的唯一依据[2],是遥测地面分析系统的最要组成部分[3]。充分利遥测数据进行对比分析,有助于故障分析和掌握飞行器的飞行特征,并不断改进设计以提高飞行器产品的质量[4],对于提高飞行器在轨运行的安全性和可靠性具有重要的意义[5]。而我国在航天领域的大数据挖掘还处于理论研究、探索阶段。

本文针对飞行器电源系统遥测数据的关联规则挖掘算法进行了对比分析,并以基于时间时序的飞行器电源系统遥测参数中,某几个参数之间的关联规则为例,进行试验对比和数据验证,为未来飞行器遥测参数关联规则挖掘提供参考。

1 飞行器遥测数据特点

飞行器在地面试验、发射、在轨飞行等阶段均会产生大量的数据,数据量大是飞行器遥测数据的显著特点。同时由于飞行器本身是个复杂的系统,遥测数据又包含电力、温度、压力、速度等各个方面,因此数据种类多也是飞行器遥测数据的明显特点。在繁多的数据中,数据类型不仅限于数字、文本,图像、视频、音频等数据的频繁使用,也使数据类型的多样性特征凸显出来。

2 关联规则挖掘算法

2.1 基于FP—Growth算法的关联规则挖掘

FP—Growth(Frequent Pattern Growth)算法是一种不产生候选模式,而采用频繁模式增长的方法挖掘频繁模式的算法[6]。它是一种扩展的前缀树(即FP树)结构,存储了关于频繁模式数量的重要信息.树中只包含长度为1

收稿日期:2020-10-13; 修回日期:2020-11-19。
作者简介:李 智(1988-),男,吉林梅河口人,硕士研究生,主要从事虚拟仿真软件开发以及仿真软件平台研发方向的研究。
引用格式:李 智,张丽晔,褚厚斌,等.面向飞行器遥测数据的关联规则挖掘方法研究[J].计算机测量与控制,2021,29(5):189-192,197.

的频繁项作为节点，并且那些频度高的节点更靠近树的根节点，因此，频度高的项比那些频度低的项有更多的机会共享同一个节点[7]。基于这一特性，可以计算出各频繁项间的关联规则。

基于 FP－Growth 算法的关联规则挖掘分为构建 FP 树，利用 FP 树挖掘频繁项集，关联规则挖掘 3 个步骤。以飞行器供电系统中电压与系统指令的遥测数据为例，电压值的变化与发出的指令密切相关，假设由电压 A，B，C 和指令 D，E 构成示例数据集，如表 1 所示。其中第一条数据 {A，B} 表示电压 A 与 B 在时刻 1 时发生变化，而第四条数据 {A，B，C，E} 表示电压 A，B，C 在时刻 4 时发生变化，同时指令 E 发出。

表 1 示例数据集

ID	Item
1	{A,B}
2	{B,C,D,E}
3	{A,C,D}
4	{A,B,C,E}
5	{B,C,D}
6	{A,B}

1）构建 FP 树：

扫描示例数据集（表 1）中全部数据，计算出频繁项集 F1。设定最小支持度为 2，如果任何一个频繁项的频繁度小于等于最小支持度，则将该频繁项从频繁项集中删除。在频繁项集 F1 中，频繁项“E”的频繁度为 2，则从频繁项集 F1 中删除频繁项“E”。然后将频繁项集 F1 按照频繁度降序排序，得出表 2 所示结果。

表 2 频繁项集 F1

Item	Count
B	5
A	4
C	4
D	3

将示例数据集中全部数据按照频繁项集 F1 中的记录重新排序，如表 3 所示。

表 3 排序后的数据

ID	Item
1	{B,A}
2	{B,C,D}
3	{A,C,D}
4	{B,A,C}
5	{B,C,D}
6	{B,A}

建立 FP 树根节点“root”，从根节点出发，第一条数据 {B，A} 各频繁项“B”和“A”按照顺序依次加入 FP 树中，并记录各节点频数 Count 为 1；当第二条数据 {B，C，D} 加入 FP 树时，首个频繁项“B”已经在 FP 树中存在，则只需将“B”节点的频数 Count 加 1，然后再将频繁项“C”作为节点“B”的一个新子节点加入 FP 树中从而得到一个新的分支。如此往复，将表 3 中每条数据依次加入到 FP 树中，得到图 1 所示的 FP 树。

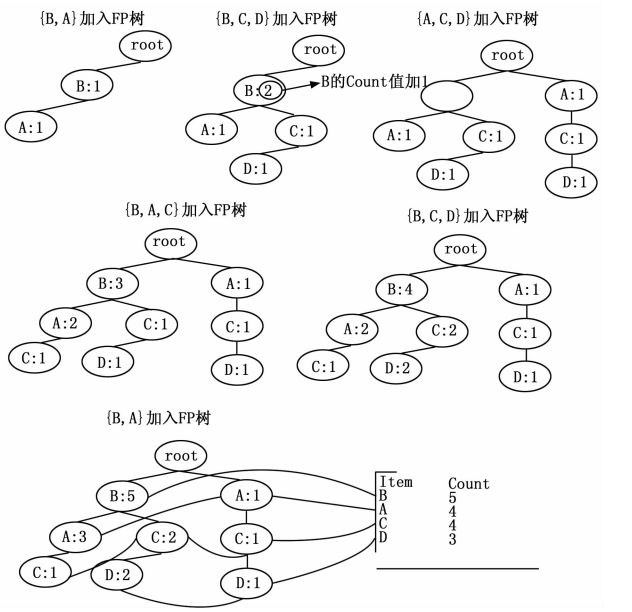

图 1 FP 树构建步骤

2）利用 FP 树挖掘频繁项集：

在得到 FP 树之后，可以进行频繁项集的挖掘。以图 1 中构建完成的 FP 树为例，首先选择一分支末端节点“D”，由“D”向根节点倒推，找出所有包含节点“D”的路径，并找出每个包含“D”的分支：{B，C，D：2}，{A，C，D：1}，其中“2”和“1”分别表示分支 {B，C，D} 和 {A，C，D} 分别出现 2 次和 1 次。分支的“Count”值，由分支后缀节点“D”出现的次数决定。

除去节点“D”，我们得到前缀路径 {B，C：2}，{A，C：1}，根据前缀路径，创建一棵条件 FP 树。然后获取前缀路径的每个节点的前缀路径，并建立条件 FP 树，直到条件 FP 树中只包含一个元素时返回（如图 2 所示）。最后，得到节点“D”的频繁项集为 { {D}，{C，D}}。

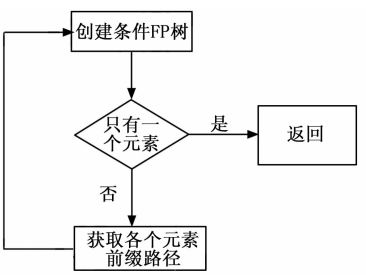

图 2 频繁项集挖掘流程

3）关联规则挖掘：

通过 FP－Growth 算法得到数据的频繁项集后，针对频繁项集中的各个频繁项构建可能的关联规则。例如，对

“D”的频繁项集中的频繁项｛C，D｝而言，可能的关联规则为｛C｝－＞｛D｝，｛D｝－＞｛C｝。

在得到这些可能的关联规则后，按如下公式计算置信度：

置信度（A－＞B）＝（同时包含 A 和 B 的频繁项数量）/（包含 A 的频繁项数量），

在示例中：

置信度（｛C｝－＞｛D｝）＝（包含｛C，D｝的数量：3）/（包含｛C｝的数量：4）＝0.75，

置信度（｛D｝－＞｛C｝）＝（包含｛C，D｝的数量：3）/（包含｛D｝的数量：3）＝1。

在计算得到所有关联规则的置信度后，保留大于置信度最小阈值的关联规则，就可以得到各频繁项之间的关联规则。在例子中，由于｛D｝－＞｛C｝的置信度为 1，可以知道指令 D 发出一定会引起电压 C 的变化，而｛C｝－＞｛D｝的置信度为 0.75，可以得出电压 C 的变化有 75％的可能是由指令 D 引起的。

2.2　基于状态转换提取的关联规则算法

对于连续的参数和离散的指令之间的关联规则挖掘，连续参数的状态转换与指令触发时段之间的联系对关联规则的挖掘有至关重要的影响。基于状态转换提取的关联规则算法，则以此为基础，在连续参数的曲线中，提取参数状态转换位置，再与指令触发的影响时域进行比对，从而得出参数与指令间关联规则的一种算法。

基于状态转换提取的的关联规则算法分为数据预处理，提取数据跳变位置，构建状态矩阵和关联性分析 4 个步骤。

1）数据预处理：

对于试验数据，首先采取预处理措施，除去数据中的非数值和异常跳变数据，以避免异常数据对算法分析的影响。

（1）非数值处理，即将数据中的非数值（NaN）替换成该帧后一帧或前一帧的数据值。

（2）异常跳变数据处理，是在非数值处理后，针对数据中的异常数据进行的异常过滤处理。首先，计算数据标准差，并以±1.5 倍标准差为预估异常范围去筛选异常数据，并得到异常数据集 S1。然后，计算数据的二阶差分，并以二阶差分最大幅度的 1/3 和 10 倍标准差作为标准，对异常数据集 S1 进行二次筛选。最后，对异常数据集 S1 中的每一条数据进行逐一确认，如果数据前后两帧数据均为疑似跳变点，且该数据跳变幅度在前后数据帧跳变幅度的 2 倍以上，则该数据点可以确认为异常数据，并将该数据替换成后一帧或前一帧的数据值。

2）提取状态转换位置：

在数据预处理后，取一次差分前段少量数据，滤除大于 6 倍标准差的数值后，再取其标准差的 n 倍作为最小阈值标准，大于最小阈值的位置，很有可能是状态跳转位置。假设疑似状态跳转位置前后的均值和标准差分别是 μ_1，μ_2，σ_1，σ_2，若其满足下式则当前位置为状态转换位置，并将这个状态转换位置加入到状态转换位置集 $S2$ 中。

$$|\mu_1-\mu_2|>n*\sigma_1 \ or \ |\mu_1-\mu_2|>n*\sigma_2$$

其中：n 是置信度水平（默认为 3），得到的状态转换位置如图 3 所示，虚线框内即为数据状态转换位置。

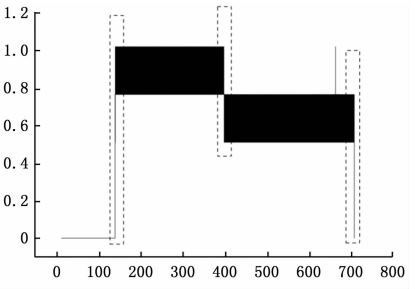

图 3　状态转换位置示意图

3）构建状态转换矩阵：

在得到状态转换位置集 $S2$ 后，对于参数的每一帧数据，按时序分别用 0 和 1 标注该数据是否发生状态转换，即将集合 $S2$ 中的每一个状态转换位置的数据标记为 1。基于考虑参数在指令发出后变化的延迟性，将状态转换位置前后 m 帧的数据范围作为状态转换影响域（转换影响域的范围会影响参数关联性的计算，m 值太小可能会导致关键关联丢失，m 值太大则会造成过多冗余，这里 m 值默认为 3），同时状态标记为 1，其他数据标记为 0。这样可以得到如图 4 所示的参数状态转换矩阵 $\boldsymbol{A}$。

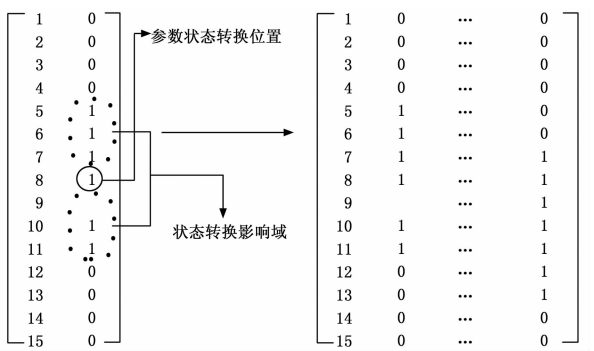

图 4　参数状态转换矩阵

4）关联性计算：

在得到参数状态转换矩阵 $\boldsymbol{A}$ 之后，按照提取矩阵 $\boldsymbol{A}$ 中的各列数据组成 n 个一维数组，如图 5 所示，并将各个数组分别相乘，得到的值便是两个参数间的关联度 C_{ij}，其中 i，j 分别代表两个不同的参数，即 C_{12} 表示参数 1 和参数 2 之间的关联度大小。然后，由计算得到的关联度集合｛C｝组成参数间的关联度表，图 5 中三角区域为集合｛C｝的值，其他部分由 0 填充。

在关联度表的基础上，设置最小关联度 MA（minimum association），筛选关联度表中关联度大于 MA（默认为 2）的值，这些值所在的行列表示的参数，可以视为关联参数对，每一个关联参数对即为两条参数相关联的关联规则。如图 5 中参数 1 与参数 n 可以组成关联参数对，则得到一条参数 1 与参数 n 相关联的关联规则。

3　试验结果与分析

为分析 FP－Growth 算法和基于状态转换提取的关联规

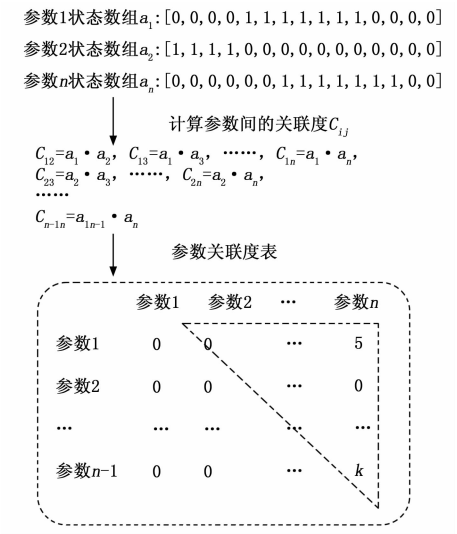

图 5 参数关联度计算

则挖掘算法在遥测参数间关联规则挖掘的表现和可行性，基于 Windows7 操作系统，Python3.8 和 Tensorflow1.15.0 等搭建运行环境，运用对比法设计试验[8]，通过对比两种算法的准确率和冗余率，给出一个更佳的关联规则挖掘方法。

3.1 试验步骤

本次试验以某次飞行器地面试验的供电系统遥测数据中的一段为试验数据，利用系统发出的指令与某电路电流和电压等三十余万条数据按时序构成试验数据集。其中系统指令为离散型数据，电流、电压等参数随时间呈连续变化。试验中验证数据为人工梳理出的参数与系统指令之间已知的关联表，如表 4 所示。

表 4 支路电流 I、电压与 U 指令关联表(部分)

参数	关联指令
I1A1	CM1001,CM1013,CM1056
I1A2	CM1003,CM1013,CM1056
I1B1	CM1001,CM1003,CM1013,CM1056
I1B2	CM1013,CM1056
UM1	CM1001,CM1013,CM1056
UM2	CM1003,CM1013,CM1056
UM3	CM1013

试验数据通过分别执行 FP－Growth 算法和基于状态转换提取的关联规则挖掘算法，对连续变化的参数与离散的系统指令之间的关联规则进行挖掘，得到参数与指令之间的关联规则表。再将试验结果与验证数据进行对比，计算出挖掘结果的准确率和冗余率。

3.2 试验结果分析

试验通过对 15 个不同参数与 5 个指令之间已知的 31 条关联规则进行挖掘，FP－Growth 算法和基于状态转换提取的关联规则挖掘算法分别挖掘到 55 条和 37 条规则，两算法的准确率和冗余率，如表 5 所示。

根据试验结果可见，基于状态转换提取的关联规则挖掘结果明显好于 FP－Growth 算法。产生这种结果的主要原因一方面与飞行器供电系统遥测数据的实际特征有关，在无指令触发情况下，参数值变化较小，几乎无波动。如图 6 中参数随时间变化结果，在指令触发（图中虚线处）前后，参数值几乎无变化，曲线近似为直线。而指令的触发只是改变参数的稳定域，所以参数呈“断崖”式变化。

表 5 挖掘结果准确率、冗余率

算法名称	准确率	冗余率
FP－Growth	0.866 667	0.527 273
基于状态转换提取的关联规则挖掘	0.967 742	0.189 189

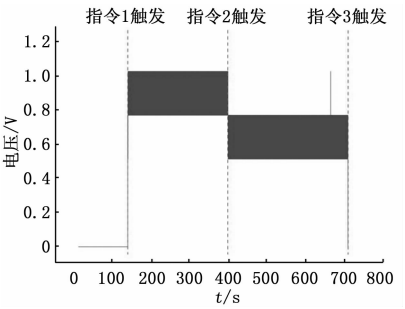

图 6 参数值跳变前后对比图

另一方面，某些指令触发次数少，也是导致 FP－Growth 算法结果准确率低的又一原因，如开机指令（CM1017）和断电指令（CM1056），这两个指令在整个试验中只能触发一次，虽然在 FP－Growth 算法中也设定了指令的作用域，但是这样的触发频率依然很可能被筛选出去，所以无法被挖掘出来。

而基于状态转换提取的关联规则挖掘，则可以很好地利用数据的特点，准确地找到数据跳变的位置即状态发生转换的时刻，再去将这些时刻进行关联计算，从而得出更好的关联结果。如图 7 所示，对于同一参数 UM7 的挖掘，FP－Growth 算法的挖掘结果为 CM1001 和 CM1013 两条指令。而基于状态转换提取的关联规则挖掘算法在 CM1001 和 CM1013 基础上，准确地挖掘出开机指令和断电指令（即图中 CM1017 和 CM1056 两条指令）共四条指令。

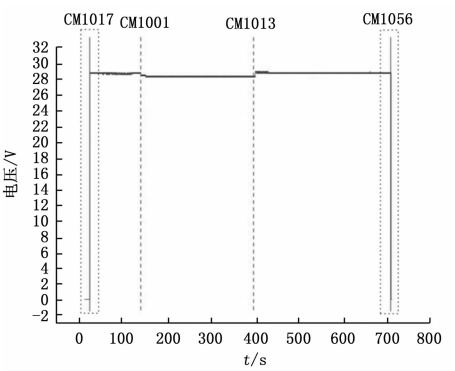

图 7 参数 UM7 随时间变化与指令触发时刻示意图

(下转第 197 页)