航天工程质量数据体系架构的挖掘

刘国才

中国航天标准化与产品保证研究院,北京 100071

摘要　依据DoDAF体系架构框架理论,针对航天工程中的质量大数据(数据包),采用经验法挖掘出航天工程质量数据体系架构的过程模型。该过程模型与前馈-反馈复合控制系统模型基本吻合,进而识别出航天质量工作恰是遵循了前馈-反馈的复合控制机制。在该机制下,进一步从4个维度选取典型质量数据进行挖掘,得到了关于"热词"、数据基线和质量问题预测等3个方面的重要信息。

关键词　航天工程;质量大数据;体系架构挖掘;质量体系效能

体系架构[1-2]的建模是体系工程理论的重要内容之一,可追溯于美国IBM公司J.A.Zachman提出的信息系统体系架构描述[3],简称为Zachman框架。在其基础上逐渐派生出了美国联邦企业体系架构框架(Federal Enterprise Architecture Framework, FEAF)和美国财政部企业体系架构框架(Treasury Enterprise Architecture Framework, TEAF)。此外,美国在C4ISR框架的基础上发布了DoDAF(U.S. Department of Defense Architecture Framework)框架,分别是DoDAF 1.0、DoDAF 1.5和DoDAF 2.0。当前,美国国防部的最新版本是2020年8月批准发布的DoDAF 2.02版本。

数据挖掘又称为知识发现,主要内容是度量数据之间的距离、相似性等,适用于文本、网络、音频、图像、视频、数值等数据形式,可实现数据的分类、聚类、相关性分析、偏差分析等。当前,深度学习、压缩采样等理论正在成为数据挖掘领域的研究热点[4-6]。

中国航天工程技术专业难度大、全寿命周期链条长、质量管控要求严苛,形成了丰富的质量溯源数据。以某一电子单机产品为例,经过方案阶段、初样阶段、正样阶段和发射在轨服务阶段,形成百份量级的数据包文件。若以分系统为例,则形成千份量级的数据包文件。而对于国家重大专项类的航天项目,最终形成的数据包文件会更多。这些数据包文件是高标准严要求的结果,更是航天事业发展所积累的宝贵大数据资源。深入挖掘和应用这些质量大数据,是提升航天工程高质量治理能力的重要手段。

基于航天工程高质量发展需要、体系架构建模理论和数据挖掘思想方法,本研究的研究思路是以航天工程质量大数据为研究对象,采用数据挖掘的思想方法,对质量大数据自身蕴含的体系架构进行挖掘,称之为航天工程质量数据体系架构挖掘。该思路通过体系架构建模和数据挖掘思想方法的交叉融合,挖掘航天工程质量大数据背后的有用信

收稿日期:2021-11-15;修回日期:2022-02-18

作者简介:刘国才,高级工程师,研究方向为航天工程质量管理,电子信箱:liu_guocai@sohu.com

引用格式:刘国才. 航天工程质量数据体系架构的挖掘[J]. 科技导报, 2022, 40(6): 101-109; doi: 10.3981/j.issn.1000-7857.2022.06.012

息,进而增强或扩展人们对航天工程质量工作规律的认知。

1 体系架构建模与挖掘路径

体系架构的建模包括2部分,其一是建模过程,其二是建模方法。建模过程包括从需求到体系架构的建模、体系架构建模、体系架构到能力的建模,例如文献[7]提出了基于能力需求的武器装备体系架构建模框架(图1[7])。

体系架构的建模方法主要有面向过程的建模方法、面向对象的建模方法、形式化建模方法和多视图建模方法。其中,多视图建模方法是当前DoDAF等体系架构建模的主要方法。本研究思路的主要出发点是基于数据挖掘思想方法的体系架构模型的挖掘,其挖掘路径如图2所示。

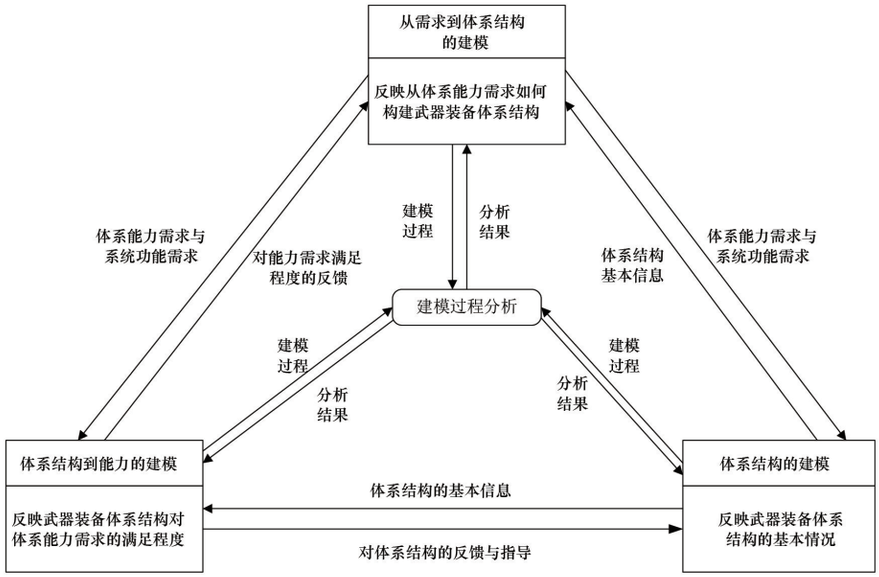

图1 基于能力需求的武器装备体系结构建模框架

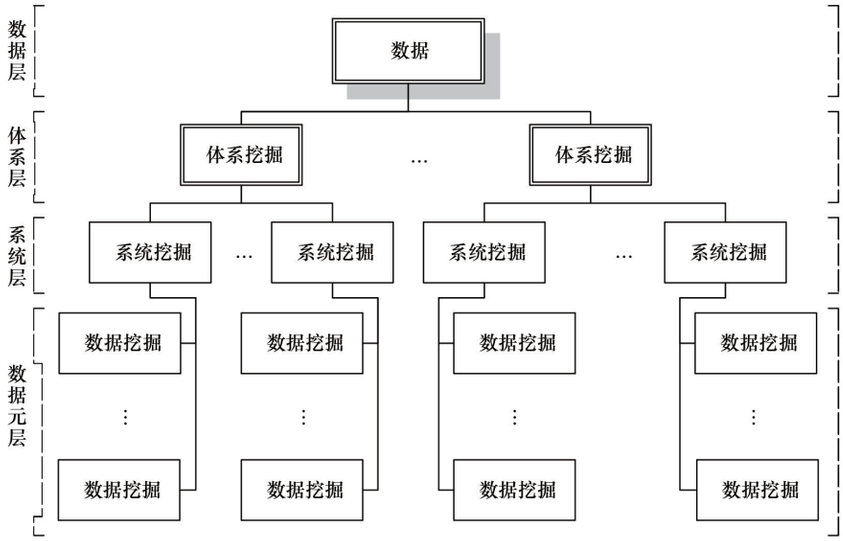

图2 体系架构的挖掘路径

该挖掘路径由4个层次组成,分别是数据层、体系层、系统层和数据元层。数据层即为给定的数据集,例如某个型号的全寿命期数据包、全部型号历年发生的质量问题信息库、质量监督记录库等。这些数据集中可能隐藏着许多体系,例如体系1、体系2、…、体系N,共同构成体系层。每一体系中又包含各式各样的系统,形成系统层。对系统层进一步分解,则形成数据元层。挖掘的先后顺序为:从数据层中挖出体系层,依体系层对数据层进行划分,之后在体系层中挖出系统层,再对某体系内的数据进行划分,之后在系统层中对系统层的组分进行各种挖掘。

航天工程质量数据体系架构模型挖掘需要遵循上述建模过程和挖掘路径。由Zachman架构的相关研究发现:(1)体系架构并没有绝对的、统一的架构形式;(2)这些架构均是"人造产物",而且仁者见仁、智者见智。所以,由于视点的不同,会形成不同的视图分类和视图产品。

在DoDAF 2.0中非常强调数据的重要性,并取代视图产品的概念。其考虑的出发点是,体系架构要为各类人员在不同层次水平提供信息,用于支持决策制定。中国航天工程中的质量数据是基于学科专业、经验积累及几十年的发展逐渐演变而来的,其能否以更精益的方式支持集成、分析和评估体系架构的能力[8],尚是个未知数。究其根源,是中国航天工程的质量数据并不是在体系架构框架下形成和管理的,所以本研究在已知数据的前提下,参照多视图建模方法,以视点给定为出发点,进行逆向溯源体系架构。但这个逆向溯源得到的体系架构与依据架构理论建立的体系架构不完全相同,前者是以"天然"状态存在的体系架构,而后者是"人造"的体系架构。由此可见,逆向溯源得到的体系架构,更能反映事物的本质规律。

2　挖掘方案框架设计

基于上述体系架构建模思路与挖掘实施路径,开展航天工程质量数据体系架构的挖掘方案框架设计,包括目的、工具、算法、步骤和形式。

1)目的。数据挖掘一般是指从大量的数据中自动搜索隐藏于其中的、有着特殊关系性的信息的过程。因此,针对航天工程质量数据开展数据挖掘,其一是挖掘出数据自身蕴含的体系架构信息,其二是挖掘该架构中各组分蕴含的有用信息。

2)工具。当前用于数据挖掘的软件工具非常多样,除了Matlab、Excel等通用型的软件外,还有在某些方面存在更强优势的挖掘软件,例如Rapid Miner、IBM SPSS Modeler、Oracle Data Mining、Teradata、Smartbi Mining、Framed Data、Kaggle、Rattle、KNIME、Python、Orange、SAS Data Mining、Weka等。

3)算法。数据挖掘的算法不仅丰富而且比较成熟,根据挖掘任务的不同,主要有分类、回归分析、聚类、关联规则分析等,部分算法见表1。

4)步骤。数据挖掘的步骤一般包括数据预处理(包括数据选择、类型及格式转换、数据清洗、建立模型假设等),挖掘算法选择,结果的显示、解释、分析与应用等。

5)形式。数据挖掘的实施过程包括非可视化形式和可视化形式。前者需要按照挖掘步骤分步实施,而后者可以基于数据流控件建立从数据源至结果输出的整体流程模型。

表1　数据挖掘算法

挖掘任务	挖掘算法
关联分析	Apriori, FilteredAssociator, FPGrowth, GeneralizedSequentislPatterns, PredictiveApriori, Tertius等
聚类分析	CLOPE, Cobweb, DBSCAN, EM, FarthestFrist, FilteredClusterer, HierarchicalClusterer, MakeDesityBasedClusterer, OPITCS, sIB, SimpleKMeans, XMeans等
分类分析	AnDE, BayesLogisticRegression, BayesNet, ComplementNativeBayes, DMNBtext, HNB, NativeBayes, NativeBayesMultinomial, NativeBayesMultinomialUpdateable, NativeBayesSimple, NativeBayesUpdateable, WAODE, GaussianProcesses, IsotoniRegression, LeastMedSq, LibLINEAR, LibSVM, LineerRegression, Logistic等

3 实证研究

根据上述挖掘方案框架,选取质量数据实施挖掘,主要包括数据源、挖掘方案实现和挖掘结果及其分析。

3.1 数据源简介

航天工程质量数据以数据包的形式进行存储和管理。工程伊始,通过开展数据包的策划,建立适用于本型号、系统、分系统、单机的数据包类型、责任单位及评审方式等。航天工程实行多级管理,分为单机级、分系统级、系统级、工程总体级,每个工程级别关注的质量要点不同,通过全寿命期形成的数据包,可以实现每个级别、环节的质量管控与溯源。本文选取某型号单机级或分系统级的受控数据包为挖掘对象,以质量体系建设、产保大纲、产品测试数据、质量信息汇总等典型数据包为例开展挖掘。

3.2 挖掘方案实现

基于挖掘方案框架,结合已有的质量数据,针对具体目的实现挖掘方案。

1)目的与思路。航天工程中质量工作的难点之一是质量问题防不胜防。即使实施了严格的质量管控,也不能从根本上杜绝质量问题的发生或者保证型号任务的持久成功,而数据挖掘技术恰好可以助力发现质量数据中蕴含的未知有用信息。

基于DoDAF多视图建模框架,首先把质量大数据视为复杂系统[9]或体系,通过数据挖掘识别出数据中自身存在的视点,然后挖掘该视点对应的视图产品,从而建立航天工程质量数据的体系架构模型。考虑到DoDAF多视图建模的复杂性,本研究以过程模型及其组分为例进行挖掘,以期获得新的有用信息。

2)工具。选用R、Minitab、Excel及Weka作为辅助软件。这4款工具各有优势,R可以便利地开展中文文本分词等工作,统计软件Minitab可以便利地开展与质量有关的统计分析,Excel可以便利地绘制散点图、回归分析和检验,Weka不需要用户编程就可以实现关联分析、分类等复杂的数据挖掘算法。

3)算法。算法选择是数据挖掘的核心内容之一,但目前的挖掘算法尚不能较好地适用于体系架构挖掘。受Zachman关于建立体系架构的思想启发,结合工程经验,采用控制系统的思想进行体系架构挖掘,该方法称之为经验法。

对于文本类型质量数据的挖掘,主要采用中文分词、标记等算法实现。对于数值型的产品测试数据主要采用统计分析方法。用于预测的质量数据,根据数据特点,主要采用回归的方法进行分析。为了探寻质量数据间的关系,主要采用数据挖掘中的关联分析方法。前两类质量数据的相关研究比较深入,而后两类质量数据的预测与挖掘研究则较少,本研究提出如下方法。

(1)质量问题预测。将航天质量问题数据用于预测质量问题的研究尚处于起步阶段,其主要研究内容有:一是基于质量问题案例样本的预测,该部分内容在文献[10]中开展了初步理论研究;二是基于质量问题案例样本数量的预测研究,该部分内容是本研究重点。

首先定义自变量:x_1——产品研制过程中质量问题案例样本数量,x_2——总装过程中质量问题案例样本数量,x_3——发射场过程中质量问题案例样本数量。定义因变量y——在轨过程中质量问题案例样本数量。

基于历年的型号质量问题案例样本数据,建立数学模型

$$y = f\left(x_1, x_2, x_3\right) \tag{1}$$

通过式(1),可以实现在轨过程中质量问题数量的预测,同时也对在轨是否发生质量问题进行了预测。

(2)质量体系监督数据挖掘。质量体系监督是质量管控重要手段之一。本研究以质量管理体系的运行监督为例,综合选取产品实现全寿命期的重点监督检查项目,开展关联关系挖掘。

建立如表2所示数据模型。其中,监督项目涵盖产品全寿命期的主要质量工作要求。样本1、样本2、样本3、样本4、样本5分别表示对质量体系开展了5次监督采样,分别形成了检查记录。检查记录中的"0"表示"无不符合项","1"表示"发生不符合项"。本研究以5次质量监督检查记录为例开展分析。

表2　质量监督数据集

监督项目	样本1	样本2	样本3	样本4	样本5
产品有关要求的确定	0	0	0	0	0
与产品有关要求的评审	1	0	0	0	0
顾客沟通	0	0	0	0	0
设计和开发策划	1	1	0	0	0
设计和开发输入	1	1	0	0	0
设计和开发输出	1	1	1	0	0
设计和开发评审	1	1	0	0	0
设计和开发验证	1	0	0	0	0
设计和开发确认	0	0	0	0	0
设计和开发更改的控制	1	1	0	0	0
新产品试制	0	0	0	0	0
试验控制	1	1	0	0	0
采购过程	0	0	1	0	1
采购信息	1	1	0	0	0
采购产品的验证	1	0	0	0	0
采购新设计和开发的产品	0	0	0	0	0
生产和服务提供的控制	1	1	0	0	0
生产和服务提供过程的确认	0	0	0	0	0
标识和可追溯性	1	0	0	0	0
顾客财产	0	0	0	1	0
产品防护	1	0	0	0	0
关键过程	0	0	0	0	0
交付	0	0	0	0	0
交付后的活动	0	0	0	0	0
建设和测量设备的控制	0	1	0	1	0
技术状态管理	1	1	0	0	0
总则	0	0	0	0	0
顾客满意	1	0	0	0	0
内部审核	0	0	0	0	0
过程的监视和测量	0	0	0	0	1
产品的监视和测量	1	0	0	0	0
不合格品控制	0	0	0	0	0
数据分析	0	0	1	0	0
持续改进	0	0	0	0	0
纠正措施	0	0	1	0	0
预防措施	0	0	0	0	0

4）步骤。

经验法主要步骤：参照控制理论，首先明确受控对象和扰动，然后定义系统输入和输出，之后设计反馈通路和控制器。

文本挖掘主要步骤：首先选择内容全面、信息量大并有较大规模的数据包文件作为语料。为具有代表性，本研究选用产保相关文件作为挖掘对象。然后选用R语言编写数据挖掘的软件工具，最后对挖掘结果进行分析。

回归分析主要步骤：首先要满足前提条件（1）ε_i—随机误差，$(\varepsilon_1,\varepsilon_2,\cdots,\varepsilon_n)$服从同一正态分布$N(0,\sigma^2)$，$E(\varepsilon_i)=0$，协方差为零；（2）自变量之间不存在多重共线性；（3）自变量与误差项不相关。然后求解模型参数，最后采用拟合优度等方法进行回归效果检验。

统计分析主要步骤：首先计算数据的均值、方差、中位数等，建立数据基线，之后的测试数据按同样的方法计算均值、方差、中位数等，与数据基线进行比对分析。

关联分析主要步骤：首先建立数据模型，并转换为适宜Weka关联分析算法的格式，然后开展关联分析，得出关联规则。

3.3　挖掘结果分析

3.3.1　体系架构挖掘结果分析

采用经验法得出航天工程质量数据的体系架构过程模型如图3所示。模型主要由4个维度组成，分别是质量调节数据、质量正向数据、质量逆向数据、质量前向数据。质量调节数据起到描述调节受控过程的作用，主要指依据输入数据（Rin）制定的项目研制要求、产品保证要求、技术流程、计划流程、接口数据单等规范性、要求性的文件。质量正向数据起到描述质量正向形成过程的作用，主要指产品全生命周期形成的设计文件、试验文件、测试文件、复核复算等文件。质量逆向数据起到描述逆向监督过程的作用，主要指质量监督检查数据、质量监理数据、质量体系审核数据等。质量前向数据起到描述前向参考过程的作用，主要指质量问题案例集，以及针对项目执行过程中出现新情况而借鉴的以往型号的数据。以上数据组成的过程模型与控制理论中前馈-反馈复合控制系统模型相吻合，这个挖掘结果很好地说明了航天质量工作恰是遵循了前馈-反馈复合控制机理，称之为前馈-反馈复合质量控制系统。前馈-反馈复合质量控制系统与PDCA（plan, do, check, action）在基本思想上高度一致，但关注点比PDCA更精细，主要体现在2

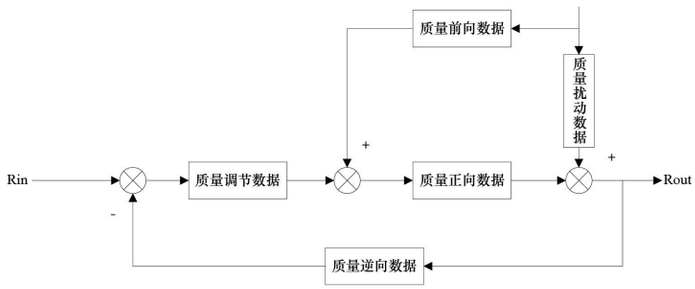

图3 航天工程质量数据的体系架构-过程模型

个方面:一是更加关注质量扰动,二是可以充分借鉴控制理论的研究成果开展质量控制系统的设计与分析。

前馈-反馈复合质量控制系统中的受控对象为质量正向数据和质量扰动数据描述的过程,反馈通道为质量逆向数据描述的过程,前馈通道为质量前向数据描述的过程,质量调节数据为控制器。显而易见,质量扰动数据是客观存在的,但工程中往往很难准确、及时发现。正是由于质量扰动的存在,导致项目执行过程出现问题,例如进度推迟、成本增加、质量问题频发等。

3.3.2 质量调节数据挖掘结果分析

采用文本挖掘[11]得到词频大于100的热词,如表3所示。

表3 频次>100的热词排序

热词	频次	热词	频次
验证	110	控制	161
试验	112	质量	169
管理	126	技术	173
评审	127	工艺	178
软件	136	保证	179
可靠性	138	研制	201
分析	141	设计	271
FPGA	147	要求	330
测试	158	产品	653

可以看出,“产品”出现的次数最多,其次是“要求”,然后分别是“设计”“研制”“保证”等。进一步分析发现,这些热词之间存在一定的逻辑“映射”,例如,针对“产品”提出“要求”,遵照“要求”开展“设计”,“设计”结果用于“研制”,并从“工艺”“技术”“质量”“控制”“测试”“FPGA(field programmable gate array)”“分析”“可靠性”“软件”“评审”“管理”“试验”“验证”等方面开展“保证”工作。这些热词环环相扣,只有吃透“产品”,才能识别出明确的、隐含的“要求”,进而开展全面的“设计”“研制”工作,全方位、有针对性地落实“保证”工作。进一步扩大词频范围,如表4所示。分析发现,表中低频词满足同样的规律,即隶属于“保证”的要素。可见,通过数据挖掘发现的逻辑映射,是蕴含在产品全生命期内的一个有用信息。

表4 频次>40的热词排序

热词	频次	热词	频次
交付	40	文档	86
人员	40	材料	92
焊接	41	关键	93
沿用	45	元器件	94
影响	46	文件	96
组织	47	过程	98
故障	49	状态	99
鉴定	49	验证	110
检查	55	试验	112
清单	55	管理	126
零件	57	评审	127
更改	60	软件	136
选用	61	可靠性	138
实施	62	分析	141
安全性	64	FPGA	147
风险	64	测试	158
措施	69	控制	161
设备	69	质量	169
检验	71	技术	173
计划	75	工艺	178
确认	77	保证	179
问题	77	研制	201
规定	78	设计	271
使用	81	要求	330
验收	83	产品	653

这一信息中与当前中国航天工程采用的“V”模型具有较大区别。因为“V”模型的核心是并行工程,而逻辑映射强调的是类似于数学函数的映射而形成“一一对应”。从映射的观点来看,产品全寿命期各要素间存在复杂的关系。这些关系可能是线性的,也可能是非线性的,有可能是人为产物,也可能是客观存在的物理关系。

3.3.3　质量正向数据挖掘结果分析

利用Minitab软件强大的计算能力和丰富的绘图功能,可以很好地呈现2种形式的挖掘结果,例如数值计算型(表5)。相较于当前的技术状态管控和数据分析方法可以深入挖掘数据中的特征,例如均值、最小值、最大值、中位数、极差、时序图等,进而建立产品的数据基线,实现同类设备历史质量的多维度比对分析。当前航天装备的复杂性决定了数据量非常庞大,从发掘历史数据提质增效的角度看,这些数据尚无有效的管理模式。数据基线的提出及利用现有的计算分析软件恰可以实现数据的技术状态管控。

表5　数值计算型挖掘形式

序号	遥测变量	总计数	均值	最小值	中位数	最大值	极差	众数
1	变量1	8047	2.9794	0	2.98	3.00	3.00	2.98
2	变量2	8047	2.9796	0	2.98	3.00	3.00	2.98
3	变量3	8047	2.9136	0	2.92	2.92	2.92	2.92
4	变量4	8047	2.9796	0	2.98	3.00	3.00	2.98
5	变量5	8047	2.9991	0	3.00	3.00	3.00	3.00
6	变量6	8047	1.4898	0	1.49	1.49	1.49	1.49
7	变量7	8047	37.826	−273.15	39.29	40.53	313.68	40.11
8	变量8	8047	34.700	−273.15	35.76	37.30	310.45	36.91
9	变量9	8047	37.192	−273.15	38.49	39.70	312.85	39.29
10	变量10	8047	35.631	−273.15	36.91	38.49	311.64	37.70
11	变量11	8047	33.804	−273.15	35.00	36.14	309.29	35.76
12	变量12	8047	35.214	−273.15	36.53	38.09	311.24	37.30

3.3.4　质量逆向数据挖掘结果分析

选取经典的关联规则分析算法“Apriori”和“FPGrowth”,运行Weka软件得到最佳关联规则挖掘结果,如表6所示。可以看出,顾客沟通决定产品有关要求和预防措施的确定,而产品有关要求和预防措施的确定是决定后续所有环节的源头。所

表6　质量监督数据关联规则挖掘结果

序号	Apriori算法	FPGrowth算法
1	顾客沟通=0 ==> 产品有关要求的确定=0	顾客沟通=0 ==> 预防措施=0
2	产品有关要求的确定=0 ==> 设计和开发确认=0	预防措施=0 ==> 采购新设计和开发的产品=0
3	产品有关要求的确定=0 ==> 新产品试制=0	预防措施=0 ==> 设计和开发确认=0
4	产品有关要求的确定=0 ==> 采购新设计和开发的产品=0	预防措施=0 ==> 生产和服务提供过程的确认=0
5	产品有关要求的确定=0 ==> 生产和服务提供过程的确认=0	预防措施=0 ==> 新产品试制=0
6	产品有关要求的确定=0 ==> 关键过程=0	预防措施=0 ==> 持续改进=0
7	产品有关要求的确定=0 ==> 交付=0	预防措施=0 ==> 总则=0
8	产品有关要求的确定=0 ==> 交付后的活动=0	预防措施=0 ==> 内部审核=0
9	产品有关要求的确定=0 ==> 总则=0	预防措施=0 ==> 关键过程=0
10	产品有关要求的确定=0 ==> 内部审核=0	预防措施=0 ==> 产品有关要求的确定=0
11	产品有关要求的确定=0 ==> 不合格品控制=0	预防措施=0 ==> 交付后的活动=0
12	产品有关要求的确定=0 ==> 持续改进=0	预防措施=0 ==> 交付=0
13	产品有关要求的确定=0 ==> 预防措施=0	预防措施=0 ==> 不合格品控制=0

以,质量体系建设以及型号质量管理要以顾客沟通、产品有关要求和预防措施的确定为重要抓手。根据型号实际研制情况来看,预防措施是易于疏漏、比较薄弱的环节。

3.3.5 质量前向数据挖掘结果分析

针对自变量x_1、x_2、x_3和因变量y,采用回归建模得到R^2=0.0187,可见,自变量x_1、x_2、x_3和因变量y之间呈弱相关性。直接采用回归的方法不具有可参考意义,所以无法建立$y=f(x_1,x_2,x_3)$进行y值的预测。

为使自变量x_1、x_2、x_3和因变量y有较好的相关性,进一步定义

$$x' = x_1 + x_2 + x_3 \tag{2}$$

$$y' = x_1 + x_2 + x_3 + y \tag{3}$$

同样进行回归建模,建立

$$y' = g(x') \tag{4}$$

针对式(4),分别得到指数回归、一元线性回归和抛物线回归的数学模型

$$y' = 1.2369e^{0.2845x'} \tag{5}$$

$$y' = 0.782x' + 1.235 \tag{6}$$

$$y' = 0.076x'^2 + 0.3868x' + 1.3921 \tag{7}$$

拟合优度分别是R^2=0.5686,R^2=0.6778,R^2=0.7045。

可见,自变量x'与因变量y'呈较强的相关性,其中抛物线回归拟合优度0.7045,明显优于指数回归和一元线性回归。通过x'和y'可进而求得

$$y = y' - x' \tag{8}$$

根据式(8),计算得到预测值如表7所示。可见,预测结果的可读性不太理想,尚不能完全区分质量问题发生的数量及是否发生质量问题。

而采用Logistic方法可以较好的解决这个问题,首先将因变量y分成如下Y类别,对应关系为:

Y=1: $y\leqslant 0$;

Y=2: $y\leqslant 1$;

Y=3: $2\leqslant y\leqslant \infty$;

建立y与x_1、x_2、x_3的Logistic回归模型,计算得到Logistic回归系数,由于x_2、x_3对y的影响不显著,所以如下考虑y与x_1,得到Logistic回归系数,如表8所示。根据表8,得到logit变换,即

$$\text{logit}\left(P\left(y \leqslant 1\right)\right) = 1.2 + 0.695x_1 \tag{9}$$

$$\text{logit}\left(P\left(y \leqslant 2\right)\right) = 1.07565 + 0.695x_1 \tag{10}$$

例如,取x_1=0,得到

$$\text{logit}\left(P\left(y \leqslant 1\right)\right) = -1.2 + 0.695x_1 = -1.2 \tag{11}$$

$$\text{logit}\left(P\left(y \leqslant 2\right)\right) = 1.07565 + 0.695x_1 = 1.07 \tag{12}$$

根据

$$P = \frac{1}{e^{-\text{logit}} + 1} \tag{13}$$

进而得到P_1=0.231475,P_2=0.744597,P_2-P_1=0.513。

可见,当x_1=0时,发生1个质量问题的可能性最大,其概率是0.513。

表7 在轨质量问题预测数量

序号	预测值	序号	预测值	序号	预测值
1	0.8549	12	0.2365	23	1.3921
2	0.4697	13	0.8549	24	1.3921
3	1.3921	14	0.2365	25	0.2365
4	1.3921	15	1.3921	26	0.2261
5	0.4489	16	1.3921	27	0.2365
6	1.3921	17	0.8549	28	0.2365
7	0.8549	18	1.3921	29	1.3921
8	0.8549	19	0.2261	30	0.2261
9	1.3921	20	0.4697	31	0.4697
10	0.2365	21	0.8549	32	0.4697
11	1.3921	22	0.8549	33	…

表8 y与x_1的Logistic回归系数

自变量	系数	系数标准误	Z	P	优势比	下限	上限(95%置信区间)
常量(1)	−1.20048	0.270467	−4.44	0.000	—	—	—
常量(2)	1.07565	0.269528	3.99	0.000	—	—	—
x_1	0.69503	0.197961	3.51	0.000	2.00	1.36	2.95

注:Z表示Z检验的统计量,P表示显著性检验的概率。

4　结论

围绕航天工程质量数据体系架构挖掘的研究思路,剖析其挖掘实施路径、设计挖掘方案框架、研究相应的挖掘算法,得出体系架构、质量调节数据、质量正向数据、质量逆向数据和质量前向数据5个方面的挖掘成果。这些挖掘成果对进一步优化、改进和提升航天工程质量体系效能具有重要参考意义。但尚有如下难点问题需继续开展深入研究。

1）航天工程质量数据体系架构挖掘是体系架构建模理论与数据挖掘思想方法交叉研究的产物,当前的挖掘算法尚不能较好地适用于体系架构挖掘,严重制约航天质量数据中科学规律的挖掘。

2）前馈-反馈复合控制机理适用于线性且干扰已知的受控对象,由于质量扰动的未知性及不确定性,其并不能根本控制质量。只有充分借鉴控制理论,建立传递函数模型或状态空间模型实现可控性、可观性分析,设计更加鲁棒和智能的质量控制系统,才能从控制机理层面提升质量体系效能。

参考文献(References)

[1] 舒振, 刘俊先, 罗爱民, 等. 军事信息系统体系结构设计方法及其应用分析[J]. 科技导报, 2018, 36(20): 48-56.

[2] 张佳, 杨红义, 刘一哲. 美国国防部体系结构框架应用分析[J]. 科技导报, 2019, 37(13): 117-123.

[3] Zachman J A. A Framework for information systems architecture[J]. IBM System Journal, 1987, 26(3): 454-470.

[4] 梁循. 数据挖掘算法与应用[M]. 北京: 北京大学出版社, 2006.

[5] Bao Y Q, Chen Z C, Wei S Y, et al. The state of the art of data science and engineering in structural health monitoring[J]. Engineering, 2019, 5(2): 234-242.

[6] 孙苗, 符昱, 吕憧憬, 等. 深度学习在海洋大数据挖掘中的应用[J]. 科技导报, 2018, 36(17): 83-90.

[7] 赵青松, 杨克巍, 陈英武, 等. 体系工程与体系结构建模方法与技术[M]. 北京: 国防工业出版社, 2013.

[8] Eisner H. Systems architecting: Methods and examples [M]. London: CRC Press Taylor & Francis Group, 2020.

[9] Wang Y X, 靳瑾. 论大数据代数(BDA): 大数据科学与工程的分析方法[J]. 科技导报, 2020, 38(3): 47-67.

[10] 刘国才, 徐欣锋. 航天复杂系统的质量治理——循果度量原理[M]. 北京: 科学出版社, 2020.

[11] 刘国才, 杜庆国, 姚大鹏, 等. 航天标准文献的知识再生与深度应用模式研究[C]//中国航天大会. 北京: 中国宇航出版社, 2019.

The architecture of quality data mining in aerospace engineering

LIU Guocai

China Academy of Aerospace Standardization and Product Assurance, Beijing 100071, China

Abstract　Based on the DoDAF architecture framework theory, for the quality big data (data package) in aerospace engineering, a process model of the aerospace engineering quality data architecture is mined by the empirical method. The process model is consistent with the feedforward and feedback compound control system models, and it is shown that the space quality work follows the compound control mechanism of the feedforward and the feedback. Based on this mechanism, the typical quality data are selected from four dimensions and mined, and the important information about the "hot words", the data baseline and the quality problem predictions is obtained.

Keywords　aerospace engineering; quality big data; architecture mining; quality SOS effectiveness

（责任编辑　王志敏）