doi:10.11887/j.cn.202106005 http://journal.nudt.edu.cn

遥测参数数据载荷状态判别集成学习方法*

李 虎[1,2],郭国航[1,2],胡 钛[1],杨甲森[3],董振兴[3]
(1. 中国科学院国家空间科学中心 空间科学卫星运控部,北京 100190;2. 中国科学院大学,北京 100049;
3. 中国科学院国家空间科学中心 复杂航天系统电子信息技术重点实验室,北京 100190)

摘 要:针对载荷单机设备遥测参数维度高、数据量大、存在类别不平衡、无法直观判别单机设备运行情况等问题,考虑到航天任务对可解释性的要求,提出一种基于信息增益参数特征选择和集成学习方法的载荷单机状态快速识别方法。采用统计量性质和信息增益子集搜索方法对遥测数据进行特征筛选降维,通过集成学习模型算法实现载荷单机设备状态的自适应识别分类。所提方法将信息增益的参数分类信息量评价准则和集成学习拟合能力强、类别不平衡下准确率高和抗噪能力强等优点相结合,兼顾模型特征和结果的可解释性,提供了重点参数发现功能。采用科学卫星任务真实载荷遥测参数数据对该方法进行了验证,整体识别准确率高于90%,少数样本亦可准确识别,整体效果可达到在轨任务要求,证明了所提方法的有效性和实用性。

关键词:有效载荷;状态判别;集成学习;信息增益;梯度提升决策树;科学卫星
中图分类号:V557.3 **文献标志码**:A **文章编号**:1001-2486(2021)06-033-08

Ensemble learning for state recognition of payload from telemetry data

LI Hu[1,2], *GUO Guohang*[1,2], *HU Tai*[1], *YANG Jiasen*[3], *DONG Zhenxing*[3]
(1. Laboratory of Scientific Satellite Mission Operation, National Space Science Center, Chinese Academy of Sciences, Beijing 100190, China;
2. University of Chinese Academy of Sciences, Beijing 100049, China; 3. Key Laboratory of Electronics and Information Technology for Space Systems, National Space Science Center, Chinese Academy of Sciences, Beijing 100190, China)

Abstract: In order to deal with various complex telemetry problems, such as high dimensionality, huge volume, unbalanced categories and failure to intuitively make sense about states of payload, and considering the requirement of interpretability in space mission, a general method for fast identification of payload based on information gain and integrated learning method was proposed. Sample statistics and information gain was used to select features and reduce the dimension of the telemetry data; meanwhile, the integrated learning algorithm was used to complete the adaptive recognition and classification about payload states. The proposed method combined the advantages of the parameter classification information evaluation criteria of the information gain and strong modeling, high accuracy and strong anti-noise ability under unbalanced category samples. Furthermore, the model had to possess the property of being explanatory and able to find the key parameters. The method was verified by experiments using actual mission data, which was tested using the payload telemetry data on operational scientific satellite mission. Following that, an state-of-art result, of which the overall recognition accuracy is higher than 90 percent and a few samples can also be identified, covered mission requirement in all and proved the effectiveness and practicability.

Keywords: payload; state recognition; ensemble learning; information gain; gradient boosting decision tree; scientific satellite

有效载荷是实现航天任务目标的关键组成部分,直接决定任务的成败。遥测数据是地面运管人员判断有效载荷在轨运行状态最重要的依据[1]。传统地面运管系统主要提供基于门限的常规参数级判读,状态判别则需要专家系统支持。航天任务中有效载荷功能各异,设备参数更多,工作方式更复杂,地面运管系统面临载荷设备遥测参数维度高、数据量大、类别不平衡和无法直观判别设备运行状况等新问题。如何进行高效在轨任务监视、载荷任务调度和参数优化设计等,决定了有效载荷运行的科学性和有效性。

基于遥测数据的航天器统计学习方法,可构建不完全依赖于航天器领域知识[2],由数据驱动

* **收稿日期**:2020-04-29
基金项目:中国科学院空间科学战略性先导专项资助项目(XDA04080201)
作者简介:李虎(1987—),男,山西运城人,工程师,博士研究生,E-mail:lihu@nssc.ac.cn;
胡钛(通信作者),男,研究员,博士,博士生导师,E-mail:hutai@nssc.ac.cn

的分析模型和方法。当前国内外学者主要的研究方向是面向在轨航天器故障异常发现[3]和卫星平台参数判读,其中海量遥测参数数据降维和特征选取方面主要采用主成分分析(Principal Component Analysis, PCA)[4]方法,即主要采用基于时间序列[5]的相似性度量和回归预测。文献[6]采用主成分分析理论对高维遥测数据进行降维处理,从高维数据集中提取低维特征组合,设计了航天器故障定位检测算法。文献[7]针对卫星姿态故障类型和故障源难以辨识问题,利用主成分分析测量卫星姿态与传感器之间遥测数据特征值比例变化进行故障判断。文献[8]对运载火箭飞行过程积累的历史数据进行分析,提出一种基于历史数据统计特性的遥测缓变参数自动判读方法。文献[9]在"天绘一号"01星任务中提出一种基于数据库软件的遥测数据快速处理方法和卫星重点参数监视判读方法。文献[10]以极限学习机(Extreme Learning Machines, ELM)预测模型为基础,采用集成学习方法对目标参数在时间维度上的变化趋势进行预测和判读。文献[11]采用仿真模型对大型充液卫星的在轨模式进行识别。上述文献基于遥测参数处理分析应用研究,围绕航天器通用分系统故障和卫星平台参数判读积累了丰富的经验,所采用的主成分分析方法属于"压缩式"降维,主要存在以下问题:①缺乏对有效载荷设备状态判别的研究;②所使用的方法、模型对类别不平衡支持不够友好;③对面向主题的高维数据特征选择缺少可解释性;④分析结果无法提供影响因素的丰富信息。一方面,航天器任务分析对解释性有较高要求,分析方法和结果要能按遥测量进行准确的人工一致性验证。另一方面,载荷仪器的高精密性、复杂性和任务安排的高灵活性,要求地面运管工作尽可能全面地覆盖载荷领域知识。这些对地面运管系统和运管人员提出了挑战。因此,本文提出一种将信息增益特征筛选方法和集成学习相结合,实现应用于航天任务运行工作的遥测参数数据载荷设备状态判别方法,以支持面向载荷设备任务模式的遥测参数数据自适应学习和判别。

1 问题模型

1.1 遥测参数数据和载荷状态张量表示

定义1 $TM=\{tm_j \mid j=1,2,\cdots,n\}$ 为载荷遥测参数集合,tm_j 为第 j 维遥测参数。

定义2 某一时刻遥测参数数据记录向量 $\boldsymbol{V}_{TM}^{(i)}$:分包遥测中提取到某载荷或某单机设备在星上时某一时刻采集的遥测原始数据。

$$\boldsymbol{V}_{TM}^{(i)}=(c^{(i)}\ \ \boldsymbol{TM}^{(i)})=(c^{(i)}\ \ tm_1^{(i)}\ \ tm_2^{(i)}\ \ \cdots\ \ tm_n^{(i)}) \tag{1}$$

其中,$c^{(i)}$ 为某一时刻星上数据采集对应的星上时,作为数据记录向量对应的时标,$\boldsymbol{TM}^{(i)}=(tm_1^{(i)}\ \ tm_2^{(i)}\ \ \cdots\ \ tm_n^{(i)})$ 为对应时刻的 n 维遥测参数数据向量。

定义3 $P=\{p_k \mid k=1,2,\cdots,l\}$ 为任务有效载荷设备集合,p_k 表示第 k 载荷设备状态。

定义4 载荷设备状态向量 $\boldsymbol{U}^{(i)}$:

$$\boldsymbol{U}^{(i)}=(c^{(i)}\ \ \boldsymbol{P}^{(i)})=(c^{(i)}\ \ p_1^{(i)}\ \ p_2^{(i)}\ \ \cdots\ \ p_l^{(i)}) \tag{2}$$

其中,$\boldsymbol{P}^{(i)}=(p_1^{(i)}\ \ p_2^{(i)}\ \ \cdots\ \ p_l^{(i)})$ 为对应时刻的 l 维载荷状态向量。

1.2 基于遥测参数数据的载荷单机设备状态判别问题模型

载荷单机设备状态是指载荷单机运行所处的各种工作模式。根据遥测参数数据判别载荷设备状态,可归约为遥测参数的多标签分类问题[12],考虑到遥测参数数据与载荷设备状态通过时标进行记录同步,二者等价 $TM\Leftrightarrow V_{TM}$,设 $\boldsymbol{\Omega}_{TM}$ 表示遥测参数数据的 n 维特征空间,考虑一般性和普适性,$\boldsymbol{\Omega}_{P_1}\times\boldsymbol{\Omega}_{P_2}\times\cdots\times\boldsymbol{\Omega}_{P_l}$ 表示为 $\|P\|$ 个不同载荷对应标签类别构成的空间。将其描述为寻求一个目标函数 h,使其满足:

$$h:\boldsymbol{\Omega}_{TM}\rightarrow\boldsymbol{\Omega}_{P_1}\times\boldsymbol{\Omega}_{P_2}\times\cdots\times\boldsymbol{\Omega}_{P_l} \tag{3}$$

给定多标签训练样本集 $\boldsymbol{D}=\{(\boldsymbol{TM}^{(i)},\boldsymbol{P}^{(i)}) \mid 1\leqslant i\leqslant s\}$,对于每条样本记录 $(\boldsymbol{TM}^{(i)},\boldsymbol{P}^{(i)})$,$\boldsymbol{TM}^{(i)}\in\boldsymbol{\Omega}_{TM}$ 为记录的 n 维特征向量,$\boldsymbol{P}^{(i)}\in\boldsymbol{\Omega}_{P_1}\times\boldsymbol{\Omega}_{P_2}\times\cdots\times\boldsymbol{\Omega}_{P_l}$ 为记录 $\boldsymbol{TM}^{(i)}$ 对应的标签。则给定样本记录数据集 $\boldsymbol{D}$ 中学习到多标签分类器为:

$$h(\boldsymbol{TM}^{(i)})=\boldsymbol{P}^{(i)} \tag{4}$$

1.3 标签相关性和遥测参数数据类别不平衡

标签相关性是指多标签问题中,数据集中样本所属的标签类别之间具有的相关性[13],例如互相独立或互斥。基于遥测参数数据的载荷设备状态判别时,由于航天器任务载荷间的协作关系,航天器任务多载荷单机设备状态对应的多标签之间存在相关性,高维度标签和分类数量会影响学习训练的复杂度和运算量,而借助载荷单机设备间的协作相关性,可实现多标签空间的降维,将问题转化为多分类问题。

类别不平衡是指分类问题中出现有些类别的样本量非常少,呈现出不同类别所对应的样本量分布不均匀。类别不平衡会影响以样本量权重为依据的模型分类准确率。遥测参数数据在载荷工作状态中类别不平衡现象较普遍,航天任务工作模式调度决定了处于特定工作状态的载荷遥测参数样本占比比较低,这些状态的判别不能出现漏判或误判。在遥测数据的载荷设备状态判别领域,需要能够准确判别各类状态,载荷设备状态拟合能覆盖到不均匀的样本集。

2　基于遥测参数数据的载荷单机设备状态判别

设计基于遥测参数原始数据进行载荷单机设备状态判别,步骤如下:

步骤 1:依据定义 2 和定义 4 所提遥测参数数据向量和载荷设备状态向量对样本数据建立多标签,按照时标形成问题模型中对应的记录组。

步骤 2:根据任务属性得到的多标签相关性进行降维或问题转换,根据规则对遥测参数进行特征选取,得到训练用记录组($\boldsymbol{V}_{TM}^{(i)}$,$\boldsymbol{P}^{(i)}$)。

步骤 3:根据任务调度时间表对记录组进行采样,分别建立训练集、测试集。

2.1　算法框架

基于问题模型和遥测参数数据分析设计算法框架,见图 1。首先,根据航天器分包遥测得到的海量载荷遥测原始数据集和载荷任务状态文件,进行原始数据数值化、合并和解析处理等;然后采用 3σ 原则一阶数据差分[14]进行野值剔除,并根据遥测数据星上时和载荷状态进行时间对标和分段筛选得到样本特征集;最后基于样本统计的性质、信息增益和任务属性进行特征筛选和多标签特征问题转化。其中,载荷状态数据以可扩展标记语言(eXtensible Markup Language, XML)格式组织。

2.2　算法模型

梯度提升树是集成学习的主要方法之一,其综合了加法模型、回归树模型和梯度提升算法,可更好地拟合训练数据。这种线性组合分类器通过改变样本的权重可以适应类别不平衡问题,并引入 bagging 和正则化项方法应对样本数据中的噪声。梯度提升决策树(Gradient Boosting Decision Tree,GBDT)基于弱学习器,经多次迭代得到特征切分点构成强分类器,并在迭代的每一步构建沿梯度最陡的方向降低损失的学习器来弥补已有模型的不足,每个弱学习器记录损失函数的梯度残差[15]。梯度提升决策树见图 2。

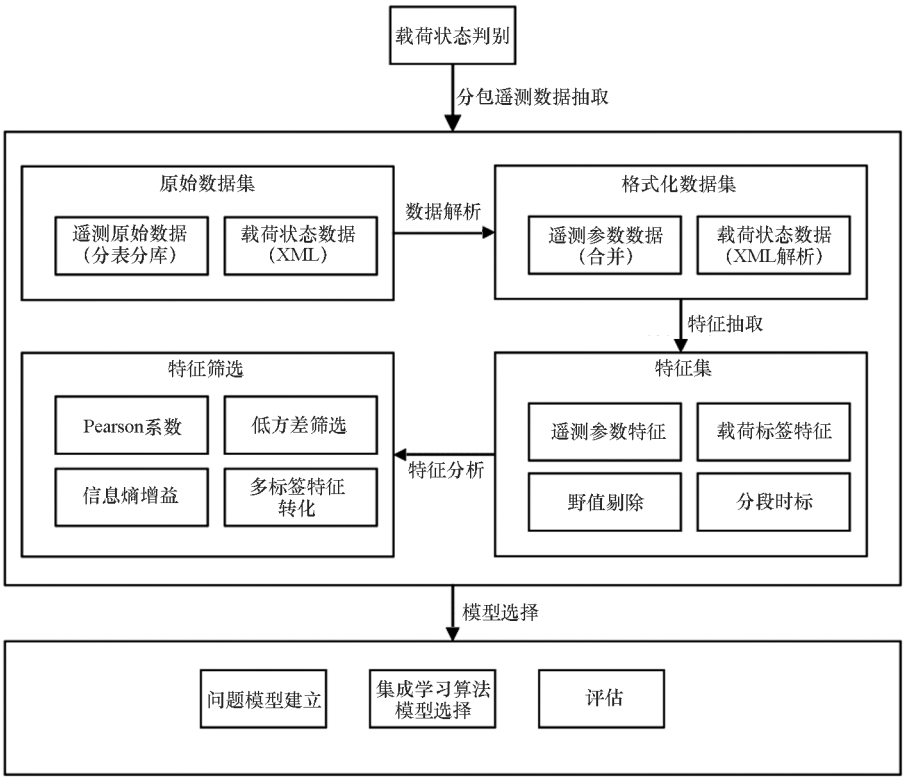

图 1　算法框架

Fig. 1　Algorithm framework

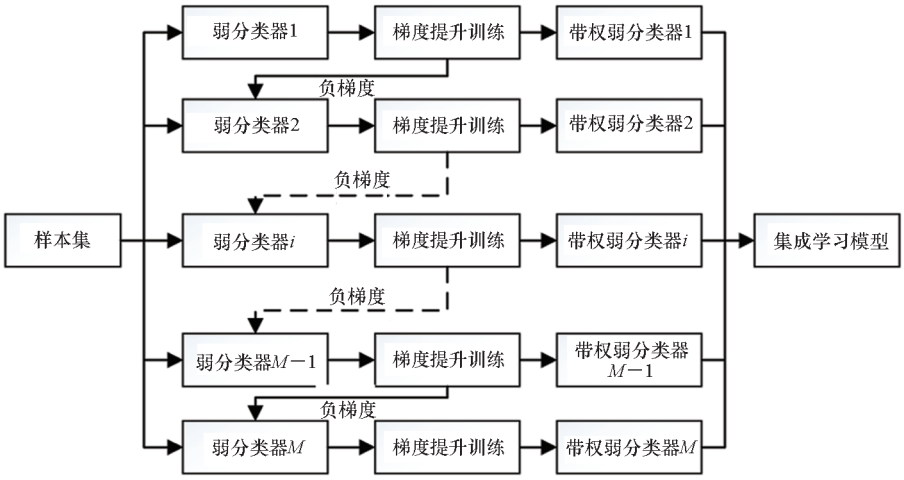

图 2 梯度提升决策树

Fig. 2 Gradient boosting decision tree

文中设计的算法是基于梯度提升的集成学习算法,最终构建遥测参数原始数据与载荷状态的映射关系。将 1.2 节定义的多标签遥测参数数据集 $\boldsymbol{D}$ 作为输入,多标签分类器 $h(\cdot)$ 为输出。梯度提升决策树模型表示为弱分类器的加法模型:

$$\hat{p} = h_M(tm) = \sum_{m=1}^{M} T(tm;\boldsymbol{\Theta}_m) \tag{5}$$

其中,$T(tm;\boldsymbol{\Theta}_m)$ 为弱分类器,$\boldsymbol{\Theta}_m$ 为弱分类器的参数,M 为弱分类器的个数。

式(5)中第 m 个弱分类器在第 i 个样本的梯度残差为:

$$r_{mi} = \left[\frac{\partial L(p,h(tm_i))}{\partial h(tm_i)}\right]_{h(tm_i)=h_{m-1}(tm_i)} \tag{6}$$

其中,L 为损失函数。

2.3 特征选择与降维

模型特征降维的可解释性要求为获得的新特征集能与人工归因一致,模型解释的关键在于特征贡献度,因此需要特征选择方法尽可能保留参数信息并不失可解释性。本文根据样本特征集的统计量性质,借助信息增益分析载荷状态样本特征集分布特性,剔除与目标问题相关性低和参数间相关度高的冗余特征,完成特征筛选和降维,保留重点参数特征以提高载荷单机状态判别模型的训练效率和准确度。方法主要包括皮尔逊相关系数[16]、方差和信息熵增益计算等,与主成分分析、互信息方法相比,效率更高并可以保留载荷参数信息。

1)航天器任务皮尔逊相关系数,即

$$\rho^2(\boldsymbol{a},\boldsymbol{b}) = \frac{E^2(\boldsymbol{a}^{\mathrm{T}}\boldsymbol{b})}{E(\boldsymbol{a}^{\mathrm{T}}\boldsymbol{a})E(\boldsymbol{b}^{\mathrm{T}}\boldsymbol{b})} \tag{7}$$

其中:$\rho^2(\boldsymbol{a},\boldsymbol{b})=1$ 表示两变量相关,$\rho^2(\boldsymbol{a},\boldsymbol{b})=0$ 表示变量不相关;$\rho^2(\boldsymbol{a},\boldsymbol{b})$ 接近 1,表示两变量线性关系密切,$\rho^2(\boldsymbol{a},\boldsymbol{b})$ 值越小表示两变量的线性相关越弱。

对遥测数据特征样本集计算两者间的相关系数,若满足 $|\rho^2(\boldsymbol{a},\boldsymbol{b})-1|\leqslant\varepsilon$,则保留其中之一特征。

2)遥测数据的一个特点是有大量的恒定值或缓变值,这些值给分类模型带来运算量,也会干扰模型的准确率,需根据样本方差性质去除该类遥测数据。由于特征方差小表示该特征中多数样本值接近,分类效果不足;特征方差大表示该特征样本值差别较大,因此设计删除低方差的特征。

3)信息增益。熵可表示随机变量的不确定性,根据随机变量的概率分布将熵定义为

$$H(X) = -\sum_{i=1}^{n} p_i \log_2 p_i \tag{8}$$

其中,p_i 是随机变量的概率。熵只依赖于随机变量的分布,与 P_i 取值无关。

根据随机变量的条件概率可得条件熵

$$H(Y|X) = \sum_{i=1}^{n} p_i H(Y|X=x_i) \tag{9}$$

特征 A 对训练数据集 $\boldsymbol{D}$ 的信息增益定义为集合 $\boldsymbol{D}$ 的经验熵 $H(\boldsymbol{D})$ 与特征 A 给定条件的经验条件熵 $H(\boldsymbol{D}|A)$ 的差,见式(10)。

$$g(\boldsymbol{D},A) = H(\boldsymbol{D}) - H(\boldsymbol{D}|A) \tag{10}$$

根据载荷状态标签特征集和遥测参数特征样本特征集,遍历各参数特征对载荷状态标签的信息增益,获得增益排序 $Rank$,选择信息增益大的特征。

通过皮尔逊相关系数、方差和信息增益计算等处理,可尽可能保留遥测参数的原始信息,并实

现特征维度的降低,同时兼顾可解释性和模型有效性。

2.4　基于遥测数据的载荷状态判别算法流程

根据 1.2 节中问题模型和 2.2 节中算法模型,设计基于遥测数据的载荷状态判别算法如下:

步骤 1:计算遥测参数特征集 $\boldsymbol{TM}^{(i)}=(tm_1^{(i)}\quad tm_2^{(i)}\quad\cdots\quad tm_n^{(i)})$ 各参数数据向量的方差,过滤方差较小的参数特征。

步骤 2:计算遥测参数特征集任意二者间的 Pearson 相关系数,对线性相关度高的特征参数进行处理,保留其中之一特征,得到本步骤筛选后的遥测参数特征集 $\boldsymbol{TM}'$。

步骤 3:对遥测参数特征集 $\boldsymbol{TM}'^{(i)}=(tm_1^{(i)}\quad tm_2^{(i)}\quad\cdots\quad tm_{n'}^{(i)})$ 与 $(\boldsymbol{TM}'^{(i)},\boldsymbol{P}'^{(i)})$,根据式(10)计算遥测参数信息增益 $g(\boldsymbol{TM}'^{(i)},\boldsymbol{P}'^{(i)})$,得到排序结果集 $rank(g(\boldsymbol{TM}'^{(i)},\boldsymbol{P}'^{(i)}))$,并根据排序结果作为训练特征选取优先级顺序,选定特征集 $\boldsymbol{TM}''$。

步骤 4:使用 GBDT 算法进行迭代:

1)初始化弱学习器:

$$h_0(tm)=\underset{c}{\operatorname{argmin}}\sum_{i=1}^{N}L(p_i,c)\tag{11}$$

2)$m=1,2,\cdots,M$,M 即弱学习器数目上限,迭代流程如下:

①对样本特征集 $\boldsymbol{TM}''$,计算负梯度残差

$$r_{mi}=\left[\frac{\partial L(p,h(tm_i))}{\partial h(tm_i)}\right]_{h(tm_i)=h_{m-1}(tm_i)}\tag{12}$$

②将残差作为新样本值得到下棵树的数据集 $\{\boldsymbol{TM}'',r_{mi}\}$,得到新的回归树 $h_m(tm)$ 对应的叶子节点区域 $R_{jm}(j=1,2\cdots,J)$,J 为该回归树的叶子节点数目。

③根据经验风险最小化准则对回归树的叶子区域进行计算最佳拟合

$$c_{jm}=\underset{c}{\operatorname{argmin}}\sum_{tm_i\in R_{jm}}L(p_i,h_{m-1}(tm_i)+c)\tag{13}$$

④更新学习器

$$h_m(tm)=h_{m-1}(tm)+\sum_{j=1}^{J}c_{jm}I(tm\in R_{jm})\tag{14}$$

3)得到最终学习器

$$\hat{h}(tm)=h_M(tm)=h_0(tm)+\sum_{m=1}^{M}\sum_{j=1}^{J}c_{jm}I(tm\in R_{jm})\tag{15}$$

3　实验结果与分析

实验验证在 Python 集成开发环境 Pycharm,采用 Scikit-learn 机器学习库实现算法,以量子科学实验卫星 6 台载荷的在轨运行数据为样本,对基于遥测参数数据的载荷单机设备状态判别算法进行验证。实验中遥测数据特征集是根据任务分包遥测从遥测原始数据中抽取,6 台载荷单机设备的载荷设备状态向量对应 5 种工作模式组合,将这些模式转化成多标签分类问题。经过实验验证,提取特征维度为 $p=6$,弱分类器个数为 150 时,可以得到最优的载荷单机状态识别效果。

模型评价指标采用准确率(A_{cc})和 F1 – Score。A_{cc} 计算正确预测样本占总样本的百分比,代表所有类的整体分类表现;F1 – Score 通过精确率(Precision)和召回率(Recall)对分类器进行整体评价,高 F1 – Score 意味着分类器对少数类别和多数类别均能识别。对于 K 个类别:

$$A_{cc}=\frac{\sum_i TP_i}{TP_1+FP_1+FN_1+TN_1}\tag{16}$$

$$F1=\frac{1}{K}\sum_i\frac{2\times\dfrac{TP_i}{TP_i+FP_i}\cdot\dfrac{TP_i}{TP_i+FN_i}}{\dfrac{TP_i}{TP_i+FP_i}+\dfrac{TP_i}{TP_i+FN_i}}\tag{17}$$

式(16)和式(17)为准确率和 F1 – Score 的计算方法。TP_i、TN_i、FP_i、FN_i 分别代表样本 i 识别为样本 i,非样本 i 识别为非样本 i,非样本 i 识别为样本 i,样本 i 识别为非样本 i。

实验从三个方面进行:①将遥测原始数据按照 2.3 节所述方法处理,计算每维数据相对于标签的信息增益(Information Gain, IG),构建特征样本集,并划分为训练集和测试集;②对比不同特征参数组合下 GBDT 模型实验性能,选择最优特征参数;③和其他算法进行对比实验,验证所提方法的有效性。

3.1　数据准备

选取量子科学实验卫星 2017 年至 2019 年的运行数据来进行算法验证,经过 2.4 节步骤 1 和步骤 2 的预处理后,共获得 579 维特征,76 699 条数据样本。将其中的 70% 作为训练集,剩余 30% 作为测试集。本实验的载荷单机状态识别问题,归约为多标签分类问题之后,采用文献[13]中的方法,将多标签分类问题转化为多分类问题来进行求解。转化成多分类问题后的数据分布如图 3 所示,易见其存在着严重的载荷工作模式类别不平衡,可采用 GBDT 模型,通过集成多个弱分类模型,能很好地拟合该数据分布。

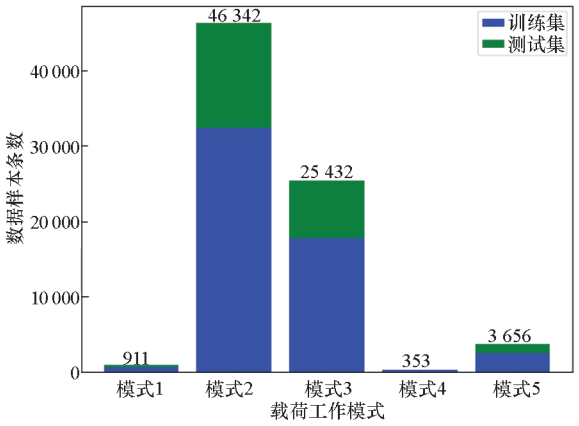

图 3 载荷工作模式数据分布

Fig. 3 Data distribution of payload mode

3.2 IG – GBDT 算法实验结果与分析

载荷单机状态与各组件运行状态息息相关,数据上反映为遥测参数数据与载荷单机状态的相关。因此,采用信息增益作为特征提取的依据。

实现基于 IG – GBDT 算法的载荷单机状态判别,采用 3.1 节中数据完成算法训练和测试,需确定模型中两个参数:应用 IG 算法筛选的特征维度;GBDT 模型中弱分类器集成个数。

首先,选用对数损失函数,固定其他参数,改变特征参数维度,构建 GBDT 分类器,分别计算训练集和测试集数据的损失;之后,选取分类损失最低的特征参数集合作为 IG – GBDT 模型的特征集。本实验中,随特征维度变化,IG – GBDT 模型损失变化如图 4 所示。当特征参数维度 $p=6$ 时,模型损失达到最小值,特征维度增加未能明显降低模型损失,确定该模型特征参数维度为 6。模型在训练集和测试集中的损失都较小,两者的损失曲线差别不大,显示了该模型具有较低的方差和偏差。

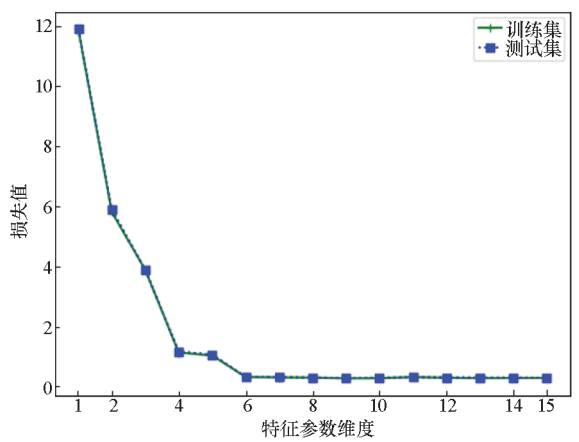

图 4 训练集和测试集损失值与特征参数维度的关系

Fig. 4 Relation between loss of training & testing sets and feature dimension

其次,确定 GBDT 模型规模,即 GBDT 模型中弱分类器的集成个数。参照筛选确定的特征参数集,调整模型中弱分类器个数,观察训练集与测试集损失变化,如图 5 所示。随着弱分类器数量的增加,训练集和测试集的损失值都在下降,起初损失值下降速率很快,当达到一定数目后,损失值变化幅度趋于平缓,继续增加弱分类器会导致计算复杂度的增加。从损失值变化曲线可知,当弱分类器数量达到 150 时,损失值的变化趋于稳定,考虑到模型计算资源消耗,确定弱分类器个数为 150。

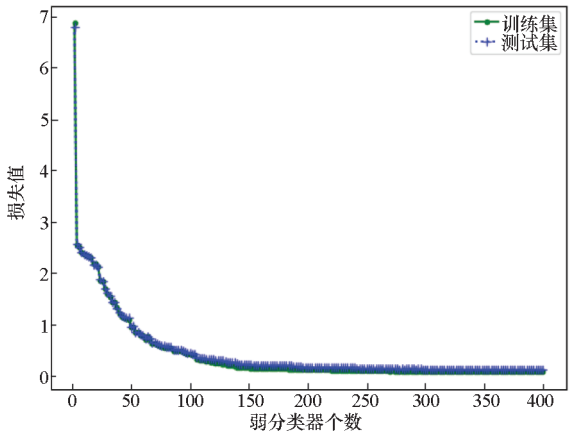

图 5 训练集和测试集损失值与弱分类器个数的关系

Fig. 5 Relation between loss of training & testing sets and number of base classifiers

在确定特征参数维度和 GBDT 规模后,将该 IG – GBDT 算法用于载荷单机状态识别问题,利用训练集中的样本数据训练载荷单机状态判别模型,再利用测试集中的样本数据验证该模型效果。训练所得模型对训练数据和测试数据的分类结果如表 1 和表 2 所示,表示 5 种模式预测结果和真实值之间的关系。其中,训练集准确率为 99.36%,测试集的准确率为 99.27%。由混淆矩阵可知,IG – GBDT 算法对于各个模式都能较准确地进行识别。

表 1 IG – GBDT 算法 – 训练集混淆矩阵

Tab. 1 Confusion matrix of IG – GBDT algorithm-training set

真实	预测					准确率/%
	模式 1	模式 2	模式 3	模式 4	模式 5	
模式 1	563	74	0	0	0	88.38
模式 2	21	32 316	0	10	1	99.90
模式 3	0	0	245	0	0	100.00
模式 4	0	178	0	17 667	1	99.00
模式 5	0	58	0	0	2 555	97.78

表 2　IG-GBDT 算法 – 测试集混淆矩阵

Tab. 2　Confusion matrix of IG-GBDT algorithm-testing set

真实	预测					准确率/%
	模式 1	模式 2	模式 3	模式 4	模式 5	
模式 1	238	37	0	0	0	86.55
模式 2	15	13 976	0	4	0	99.86
模式 3	0	2	106	1	0	97.25
模式 4	0	89	0	7 498	0	98.83
模式 5	0	19	0	0	1 025	98.18

3.3　IG – GBDT 与 PCA – GBDT 算法结果对比

将 IG – GBDT 算法与基于 PCA 特征提取的 GBDT 算法(PCA – GBDT)对比,考虑到数据分布不均衡,采用 F1 值来进行模型精度的衡量,两种算法的 F1 值对比如图 6 所示。

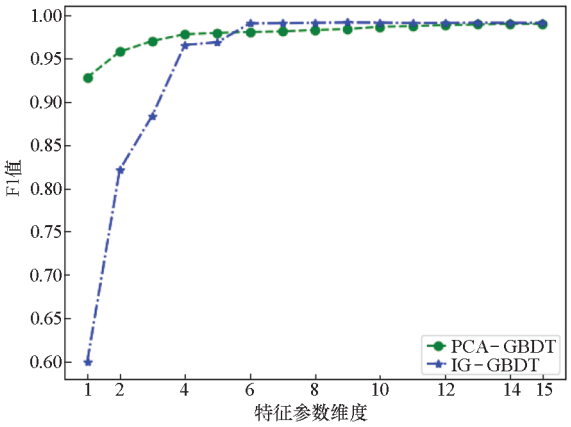

图 6　F1 值与特征参数维度的关系

Fig. 6　Relation between F1 – score and feature dimension

由图 6 可知,随着筛选的特征参数维度增加,两个算法的拟合精度均有提升。当维度 $p=6$ 时,IG – GBDT 算法的精度达到最大值,随着参数的增加,其 F1 值不再显著变化,这与 3.2 节中所得结论一致。对于 PCA – GBDT 算法,当特征参数维度 $p=12$ 时,其精度达到最大值,算法精度不再随 p 值的增加而提升。两条曲线在达到各自的最优值之后,继续增加 p 值,会引入冗余特征,因此曲线不再有上升的趋势。对比两种算法最优情况下的 F1 值,两者的最优 F1 值基本相同,但 IG – GBDT 能够用非常少的特征去表征问题,为了达到同样的效果,PCA – GBDT 则要用 2 倍的参数量。

有效载荷单机状态判别对时效性提出了较高的要求,因此对 IG – GBDT 和 PCA – GBDT 算法执行效率进行了对比。图 7 为 IG – GBDT 和 PCA – GBDT 算法运行时间随特征参数维度 p 的变化情况,特征参数增多时,两种算法的运行时间均在不断增加,PCA – GBDT 相比 IG – GBDT 的运行时间增长较慢。结合图 6 和图 7,当两个算法准确度达到最优时,IG – GBDT 的参数维度为 $p=6$,PCA – GBDT 的参数维度为 $p=12$,此时 IG – GBDT 的运行时间为 56 s,PCA – GBDT 的运行时间为 175 s,可见在相同的准确率下,PCA – GBDT 耗时是 IG – GBDT 的 3 倍。因此,IG – GBDT 算法较 PCA – GBDT 算法具有较高的执行效率。

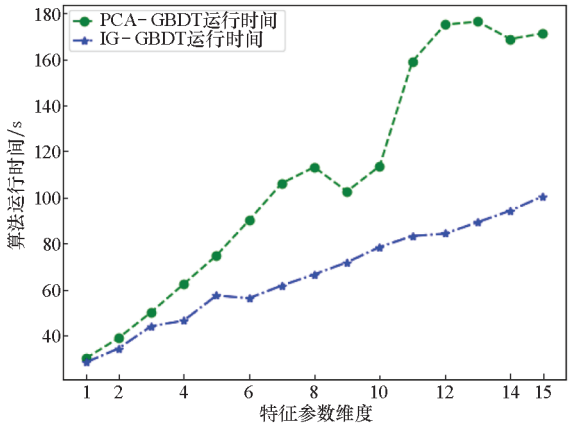

图 7　算法运行时间与特征维度的关系

Fig. 7　Relation between runtime and feature dimension

对比上述 IG – GBDT 和 PCA – GBDT 两种方法。首先,二者均能准确判别载荷单机状态,而 IG – GBDT 筛选出的特征数更少,计算资源消耗少,处理速度快,更能满足有效载荷在轨状态快速识别对时效性的要求;其次,PCA 特征降维会对原始遥测参数进行组合,这样改变了参数的含义,不具可解释性,无法对载荷参数进行人工一致性验证,而 IG 筛选得到的特征参数能确定载荷运行状态判别对应的载荷遥测参数,模型结果具有可解释性;再次,IG – GBDT 所得到的特征参数集,可供地面运控系统重点监视参数选择。

4　结论

本文提出了一种基于载荷遥测参数数据的载荷状态判别方法,将多标签分类集成学习方法应用到载荷设备状态识别问题,并采用真实卫星任务数据进行了应用验证。首先,根据载荷按分包遥测抽取遥测原始数据和任务数据,经合并、解析、数值转换等处理得到数据样本集。其次,在对遥测数据野值剔除的基础上,分别构建遥测参数特征向量和载荷标签特征向量,并以星上时为基准分段对标得到特征集,分析实际问题对多标签特征进行转化,根据特征

样本集统计量性质筛选特征和降维,计算遥测数据特征数据集对标签特征的信息增益并排序,用于构建样本最终的特征向量集。再次,利用各样本的特征向量训练基于 IG - GBDT 集成学习的载荷状态判别模型。通过量子科学实验卫星任务真实数据验证,本文提出的 IG - GBDT 算法具有很高的状态识别准确率。本文提出的载荷状态判别模型和方法能在不依赖于载荷复杂背景知识的情况下适用于载荷遥测数据量大、参数众多、样本分类不平衡等问题,基于 IG 的参数特征降维和集成学习模型将可解释性和拟合效果好的优势相结合,能满足航天任务的高准确度要求,在实际应用验证中表现出良好的性能和适用性。

参考文献(References)

[1] 彭喜元, 庞景月, 彭宇, 等. 航天器遥测数据异常检测综述[J]. 仪器仪表学报, 2016, 37(9): 1929 - 1945.
PENG Xiyuan, PANG Jingyue, PENG Yu, et al. Review on anomaly detection of spacecraft telemetry data [J]. Chinese Journal of Scientific Instrument, 2016, 37(9): 1929 - 1945. (in Chinese)

[2] 周忠玉, 皮德常. 面向卫星遥测数据流的最小稀有模式挖掘方法[J]. 计算机学报, 2019, 42(6): 1351 - 1366.
ZHOU Zhongyu, PI Dechang. Minimal rare pattern mining method for satellite telemetry data streams [J]. Chinese Journal of Computers, 2019, 42(6): 1351 - 1366. (in Chinese)

[3] 刘敏. 数据驱动的卫星姿态控制系统微小故障检测与预测方法研究[D]. 南京: 南京航空航天大学, 2018.
LIU Min. Research on data-driven incipient fault detection and prediction methods for satellite attitude control system[D]. Nanjing: Nanjing University of Aeronautics and Astronautics, 2018. (in Chinese)

[4] NASRI O, GUEDDI I, BENOTHMAN K, et al. Fault diagnosis of spacecraft reaction wheels based on principal component analysis [C]// Proceedings of the International Conference on Systems and Control (ICSC), 2015: 1 - 11.

[5] 余文艳, 肖志刚, 李虎. 时间序列模型在卫星异常检测中的应用研究[J]. 计算机技术与发展, 2018, 28(12): 122 - 126.
YU Wenyan, XIAO Zhigang, LI Hu. Application and research of time series model in satellite anomaly detection[J]. Computer Technology and Development, 2018, 28(12): 122 - 126. (in Chinese)

[6] 柴敏, 杨悦, 徐小辉, 等. 面向故障诊断的航天器遥测数据降维分析技术[J]. 弹箭与制导学报, 2014, 34(1): 150 - 153.
CHAI Min, YANG Yue, XU Xiaohui, et al. The dimension reduction analysis of spacecraft′s telemetry data for fault diagnosis [J]. Journal of Projectiles, Rockets, Missiles and Guidance, 2014, 34(1): 150 - 153. (in Chinese)

[7] 李楠, 张云燕, 李言俊. 一种自旋稳定卫星姿态传感器数据异常的诊断方法[J]. 宇航学报, 2011, 32(6): 1327 - 1332.
LI Nan, ZHANG Yunyan, LI Yanjun. A diagnosis algorithm for abnormal data of spin-stabilized satellite attitude sensors[J]. Journal of Astronautics, 2011, 32(6): 1327 - 1332. (in Chinese)

[8] 李鑫, 高家智, 崔俊峰, 等. 一种遥测缓变参数自动判读的新方法[J]. 宇航学报, 2018, 39(5): 585 - 592.
LI Xin, GAO Jiazhi, CUI Junfeng, et al. A novel method of automatic interpretation for slow-varying telemetry parameters[J]. Journal of Astronautics, 2018, 39(5): 585 - 592. (in Chinese)

[9] 王琨, 李今飞, 张帆. 依据整星遥测数据的卫星状态快速分析方法[J]. 测绘地理信息, 2016, 41(1): 43 - 46.
WANG Kun, LI Jinfei, ZHANG Fan. Rapid analysis method of satellite status based on the entire satellite telemetry data [J]. Journal of Geomatics, 2016, 41(1): 43 - 46. (in Chinese)

[10] 史欣田, 庞景月, 张新, 等. 基于集成极限学习机的卫星大数据分析[J]. 仪器仪表学报, 2018, 39(12): 81 - 91.
SHI Xintian, PANG Jingyue, ZHANG Xin, et al. Satellite big data analysis based on bagging extreme learning machine[J]. Chinese Journal of Scientific Instrument, 2018, 39(12): 81 - 91. (in Chinese)

[11] 孙宝祥, 邹广瑞, 吕振铎, 等. 大型充液卫星在轨模式识别[J]. 控制工程, 2000 (5): 1 - 6.
SUN Baoxiang, ZOU Guangrun, LYU Zhenduo, et al. On-board pattern recognition for large satellites with tanks filled with propellants [J]. Control Engineering of China, 2000(5): 1 - 6. (in Chinese)

[12] ZHANG M L, ZHOU Z H. A review on multi-label learning algorithms [J]. IEEE Transactions on Knowledge and Data Engineering, 2014, 26(8): 1819 - 1837.

[13] 田刚, 何克清, 王健, 等. 面向领域标签辅助的服务聚类方法[J]. 电子学报, 2015, 43(7): 1266 - 1274.
TIAN Gang, HE Keqing, WANG Jian, et al. Domain-oriented and tag-aided web service clustering method [J]. Acta Electronica Sinica, 2015, 43(7): 1266 - 1274. (in Chinese)

[14] 饶云峰, 白燕. 一种基于一阶差分的野值类型判别及处理方法[J]. 时间频率学报, 2015, 38(4): 227 - 234.
RAO Yunfeng, BAI Yan. A method based on first-order difference for type-judging and processing of outliers [J]. Journal of Time and Frequency, 2015, 38(4): 227 - 234. (in Chinese)

[15] FANG W J, ZHOU J, LI X L, et al. Unpack local model interpretation for GBDT[C]//Proceedings of the International Conference on Database Systems for Advanced Applications, 2018: 764 - 775.

[16] BENESTY J, CHEN J, HUANG Y T, et al. Pearson correlation coefficient [M]//Cohen I, Huang Y T, Chen J D, et al. Noise reduction in speech processing. Heidelberg: Springer, 2009: 1 - 4.