

基于集成学习的空间科学卫星工作模式识别*

高立京^{1,2} 陈志敏¹ 郭国航^{1,2} 王春梅¹

1(中国科学院国家空间科学中心 北京 100190)

2(中国科学院大学 北京 100049)

摘要 针对空间科学卫星遥测参数数据量大且特征维度高、需要消耗大量人力资源预先设置海量阈值、预先设置的阈值可能不再适用、现有监测手段可扩展性低等问题,提出了一种基于集成学习的空间科学卫星工作模式识别方法。该方法采用相关系数统计特性和互信息理论对遥测参数数据进行筛选降维,使用数据重采样技术解决数据集中存在的类别不平衡问题,构建集成学习模型,实现空间科学卫星工作模式的识别。借助某型号科学卫星真实遥测参数数据对该方法进行验证,在短时内便可构建完成算法模型,模型对整体类别的识别正确率高达99.67%,可正确识别多数类样本和少数类样本,为地面运控人员判断空间科学卫星工作模式提供了决策依据。

关键词 遥测参数数据, 特征降维, 数据重采样, 集成学习

中图分类号 TP391

Recognition of Working Pattern of Space Science Satellite Based on Ensemble Learning

GAO Lijing^{1,2} CHEN Zhimin¹ GUO Guohang^{1,2} WANG Chunmei¹

1(National Space Science Center, Chinese Academy of Sciences, Beijing 100190)

2(University of Chinese Academy of Sciences, Beijing 100049)

Abstract Aiming at the issues of space science satellite telemetry parameters, such as large amount of data, high dimension, the need of numerous artificial resource consumption for preset massive thresholds, the preset thresholds that may not be applicable, and the current monitoring methods with low scalability, a working pattern recognition method is proposed for scientific satellite based on ensemble learning. Correlation coefficient statistical characteristics and mutual information theory are used to screen and reduce the dimension of telemetry parameter data. Data resampling technology is used to solve the problem of category imbalance for the dataset. An integrated learning model is used to identify the working mode of space science satellite. The method is verified with the real telemetry parameter data of quan-

* 中国科学院空间科学先导专项科学卫星任务运控技术项目资助(XDA15040100)

2022-03-01 收到原稿, 2022-10-11 收到修定稿

E-mail: himeet@163.com

©The Author(s) 2023. This is an open access article under the CC-BY 4.0 License
(<https://creativecommons.org/licenses/by/4.0/>)

tum science satellites. And the algorithm model can be constructed in a short time, and the overall recognition accuracy rate reaches 99.67%, which can correctly identify the majority and minority class samples. The method can provide decision-making basis for ground personnel to judge the working mode of space science satellites.

Key words Telemetry parameter data, Feature dimension reduction, Data resampling, Ensemble learning

0 引言

空间科学卫星在轨工作期间会向地面下传海量遥测参数数据,这些遥测参数数据是地面运控系统判断其在轨工作模式的重要依据^[1]。目前地面运控人员主要通过监视重点遥测参数是否超出事先设定的阈值来识别空间科学卫星当前工作模式,由于空间科学卫星上携带的有效载荷数量众多,需要地面人员事先设定海量的阈值,耗费大量的人力资源。当需要监视其他卫星时,需要重新为新的卫星设置海量的阈值,移植成本高,可扩展性弱。同时,随着卫星在轨运行时间的增加,加之空间环境复杂,辐射因素或温度因素可能引起元器件发生异常,同时机械部件的长期磨损、电子线路的老化、蓄电池因充放电过多导致的性能退化等一系列问题都可能会使得前期预先设定的阈值不再适用,而地面人员又难以根据当前情况适时调整阈值。因此,设计一种可以根据遥测参数自动识别空间科学卫星工作模式的方法是很有必要的。

目前,对于航天器工作模式识别的研究主要集中在航天器故障诊断和遥测参数判读方面。文献[2]开发了 Novelty Detection 组件,利用最大最小值、均值、标准差等统计特征重新表示遥测数据,并结合 LOF(Local Outlier Factor)方法进行异常检测。文献[3]使用混合概率主成分分析(Mixture Probabilistic Principal Component Analysis, MPPCA)与隐半马尔科夫模型(Hidden Semi-Markov Model, HSMM)实现数据维度约简,使用序列蒙特卡罗算法进行参数估计,实现航天器故障诊断。文献[4]针对遥测参数中缓变参数的自动判读问题,提出使用基于历史数据的统计特性自动判读遥测数据参数的方法。文献[5]以极限学习机(Extreme Learning Machine, ELM)预测模型为基础,使用集成学习的方法,针对目标参数的不同类别分别给出判读策略。文献[6]将传感器与

卫星姿态之间的冗余关系作为特征,利用主成分分析(Principal Component Analysis, PCA)进行特征降维,通过监测特征值之间相应比例的变化实现遥测参数的判读。文献[7]使用灰关联分析以及信息熵剔除冗余参数并离散化参数,提出的 Prefix Span 算法能够通过卫星遥测数据检测其异常的模式。

基于上述研究,本文提出了一种基于随机森林算法模型的空间科学卫星工作模式识别方法,将卫星工作模式识别问题归约为机器学习领域中的多标签分类问题。借助皮尔逊相关系数和互信息理论,剔除特征与特征之间相关性高、特征与目标之间相关性低的特征,实现对遥测参数数据的特征降维。使用随机过采样、SMOTE 过采样、SMOTE-Tomek 混合采样等多种数据重采样技术处理数据集,解决数据集中存在的类别不平衡问题。构建以 Bagging 集成学习思想为基础的随机森林算法模型,实现空间科学卫星工作模式的识别。使用某型号科学卫星真实遥测参数数据进行实验验证,比较不同数据重采样技术之间的性能差异,探索了不同基学习器数量下算法模型的性能表现,将随机森林模型与其他常用机器学习算法模型在同一数据集上进行性能对比。实验结果证明,本文方法解决了数据集中存在的类别不平衡问题,同时构建算法模型耗时短,整体类别的识别正确率高,可正确识别多数类样本和少数类样本,能够为地面运控人员判断空间科学卫星工作模式提供决策依据。

1 工作模式识别方法

1.1 问题模型与向量表示

1.1.1 遥测数据形式化描述

定义 1 空间科学卫星的遥测参数集合表示为

$$\mathbf{P} = \{p_i | i = 1, 2, \dots, n\}. \quad (1)$$

其中 p_i 表示第 i 维的遥测参数。

定义 2 以星上时为时间基准,在某一时刻 j 卫星采集到的遥测参数数据向量表示为

$$\mathbf{M}_P^{(j)} = [t^{(j)} \mathbf{P}^{(j)}] = [t^{(j)} p_1^{(j)} p_2^{(j)} \dots p_n^{(j)}]. \quad (2)$$

其中, $t^{(j)}$ 表示某一时刻 j 的星上时,用来在时间维度上标识遥测参数数据向量。 $\mathbf{P}^{(j)} = [p_1^{(j)} p_2^{(j)} \dots p_n^{(j)}]$ 为对应时刻采集到的 n 维遥测参数数据向量。

定义 3 卫星在时间 $1 \sim K$ 的时间段内产生的遥测数据表示为

$$\mathbf{D}_{1,K} = [\mathbf{M}_P^{(1)} \mathbf{M}_P^{(2)} \dots \mathbf{M}_P^{(K)}]^T = \begin{bmatrix} t^{(1)} & p_1^{(1)} & p_2^{(1)} & \dots & p_n^{(1)} \\ t^{(2)} & p_1^{(2)} & p_2^{(2)} & \dots & p_n^{(2)} \\ \vdots & \vdots & \vdots & \dots & \vdots \\ t^{(K)} & p_1^{(K)} & p_2^{(K)} & \dots & p_n^{(K)} \end{bmatrix}. \quad (3)$$

定义 4 某型号空间科学卫星共存在 5 种工作模式,其工作模式集合表示为

$$L = \{l_1, l_2, l_3, l_4, l_5\}. \quad (4)$$

1.1.2 工作模式识别问题模型

空间科学卫星的工作模式指的是卫星在执行特定科学任务时所处的整体状态,可以将空间科学卫星

工作模式识别问题归约为机器学习领域的多标签分类问题。

1.2 算法框架

基于集成学习的空间科学卫星工作模式识别方法的整体算法框架如图 1 所示。首先从数据库中获取以分库分表形式存放的原始遥测参数数据,合并分库分表中的遥测参数数据并将其中的十六进制数据按照配置文件转换为十进制数据。借助星历计划安排获取每个星上时区间内卫星所处的工作模式,根据遥测参数数据的星上时时间戳所处的时间区间为每条数据打上标签,用以标识该条数据产生时卫星所处的工作模式,得到带有标签的原始数据集。之后对原始数据集中的数据进行缺失值填充、异常值剔除、数据标准化等预处理操作。剔除数据集的冗余无关特征,可以减少训练和预测的计算量,因此下一步需对特征进行筛选降维,计算特征与特征之间的皮尔逊相关系数剔除冗余特征,计算特征与目标之间的互信息剔除无关特征。在航天任务中,空间科学卫星工作模式调度的特殊性决定了其处于不同工作模式间的时间占比相差较大,造成遥测参数数据存在严重的类别不平衡问题。为解决该问题,对原始数据集分别进

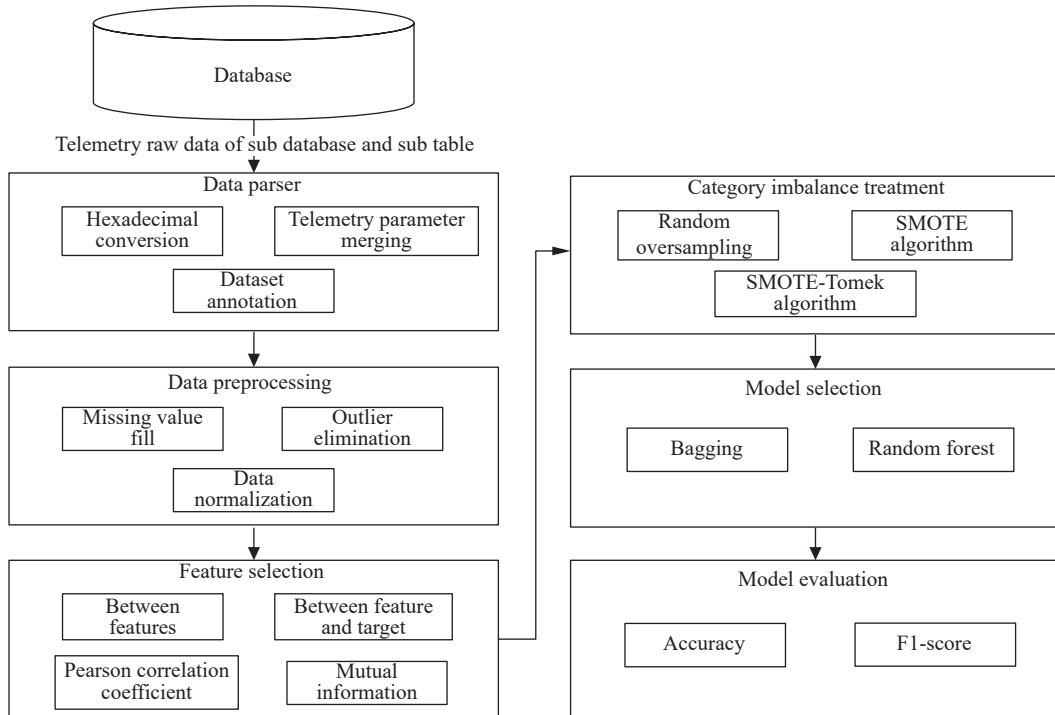


图 1 整体算法框架

Fig. 1 Overall algorithm framework

行随机过采样、SMOTE 过采样、SMOTE-Tomek 混合采样等处理。在模型选择方面,基于 Bagging 集成学习思想,采用随机森林作为识别空间科学卫星工作模式的算法模型。在算法模型的评价指标方面,由于数据集中存在严重的类别不平衡问题,单靠正确率这一评价指标不能完整评价模型的拟合能力和泛化能力,还需要用到 F1-score 这一评价指标。下面以图 1 中的整体算法框架为基础,介绍其每一部分内容。

1.3 数据预处理

数据预处理过程主要进行缺失值填充、异常值剔除和数据归一化的工作。

1.3.1 缺失值填充

遥测参数数据的一个特点是数据变化缓慢,同一个遥测参数的值不会在短期内发生剧烈变化,因此对于某个遥测参数中的缺失值可参考其附近时刻值。这里取该缺失值前后各 30 个采样值的均值作为该缺失值的填充值,若采样值的数量不满 30 则以实际数量为准。

1.3.2 异常值剔除

异常值定义为因设备编解码错误或传输过程出错而产生的严重偏离正常数据范围的值,对于异常值,使用箱型图进行判别。箱型图是一种用作显示一组数据分散情况的统计图,可以用来识别数据分布以及检测异常数据,其主要由 5 个点组成:下边缘 $T_{\min}$,下四分位数 Q_1 ,中位数 Q_2 ,上四分位数 Q_3 ,上边缘 $T_{\max}$ 。若数据位于上下边缘区间之外,则认为是异常值^[8],如图 2 中两个红色点所示,均为异常值。

箱型图的下边缘和上边缘计算公式如下:

$$T_{\min} = Q_1 - 1.5R, \quad (5)$$

$$T_{\max} = Q_3 + 1.5R. \quad (6)$$

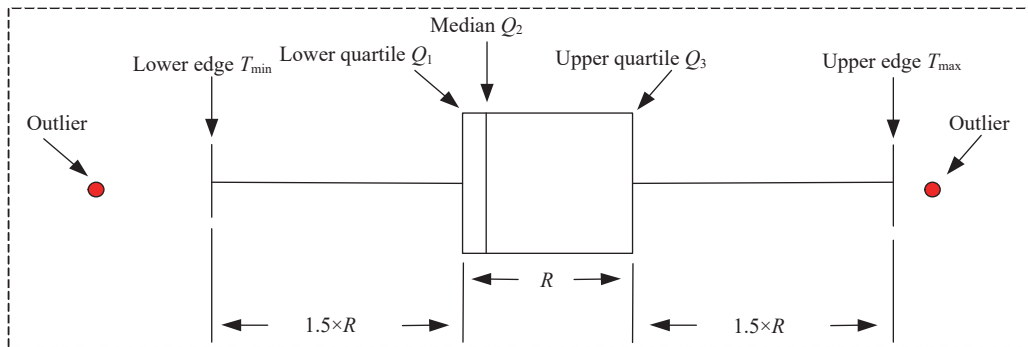


图 2 箱型图检测异常值

Fig. 2 Abnormal value detected by box diagram

其中, $R = Q_3 - Q_1$ 。

对使用箱型图检测出的异常值进行剔除,并使用 1.3.1 节中提到的缺失值填充方法进行填充。

1.3.3 数据归一化

遥测参数数据中不同参数的数值量纲是不同的,特征之间取值范围的悬殊性不利于数值求解,导致模型收敛困难,降低模型的训练效率。针对大部分机器学习算法,在建模之前都需要对数据进行归一化处理,消除参数间数值量纲差异带来的不利影响。本文中使用的数据归一化方式为零均值归一化,该方式可以将数据映射到均值为 0、标准差为 1 的标准正态分布上^[9]。数值 x 的零均值归一化公式为

$$x' = \frac{x - \mu}{\sigma}. \quad (7)$$

其中, μ 表示该特征下的数据均值, σ 表示该特征下的数据标准差。

1.4 特征筛选

对数据集进行特征筛选降维可以删除冗余无关的参数特征并保留重点参数特征,能够在不影响或略微影响模型精度的情况下大大减少计算量,提高训练和预测的效率。本文根据数据集的统计特性判断数据集中的参数特征是否具备冗余参数特征和无关参数特征,借助皮尔逊相关系数和互信息理论分析数据集样本的统计分布特性,剔除特征与特征之间相关性高、特征与目标之间相关性低的特征。

1.4.1 皮尔逊相关系数

皮尔逊相关系数是一种衡量数据统计特性的指标,可以描述遥测参数数据中特征与特征之间的线性相关性。遥测参数特征 X 与遥测参数特征 Y 之间的皮尔逊相关系数为

$$\rho_{X,Y} = \frac{\text{cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{E(XY) - E(X)E(Y)}{\sqrt{E(X^2) - (E(X))^2} \sqrt{E(Y^2) - (E(Y))^2}} = \frac{n \sum_{i=1}^n x_i y_i - \sum_{i=1}^n x_i \sum_{i=1}^n y_i}{\sqrt{n \sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i\right)^2} \sqrt{n \sum_{i=1}^n y_i^2 - \left(\sum_{i=1}^n y_i\right)^2}}. \quad (8)$$

其中, $\text{cov}(X, Y)$ 表示遥测参数特征 X 与遥测参数特征 Y 的协方差, σ_X 和 σ_Y 分别表示遥测参数特征 X 和遥测参数特征 Y 的标准差^[10]。

相关系数 ρ 的取值范围为 $[-1, 1]$, $\rho = -1$ 时表示两个遥测参数特征之间为线性负相关, $\rho = 1$ 表示两个遥测参数特征之间为线性正相关, $\rho = 0$ 时表示两个遥测参数特征之间不存在线性关系。一般地, 使用相关系数的绝对值表示两个遥测参数特征之间的相关程度, 相关系数的绝对值 $|\rho|$ 与相关程度的关系见表 1。在本文中, 对于遥测参数数据中的所有特征, 分别计算两两之间的皮尔逊相关系数, 对于皮尔逊相关系数绝对值满足 $0.9 \leq |\rho| \leq 1$ 的两个遥测参数特征, 只保留其中一个特征, 即剔除了遥测参数数据中特征与特征之间相关程度极高的冗余参数特征。

1.4.2 互信息

互信息用来描述两个随机变量之间共同拥有的信息量, 可以度量遥测参数特征与目标标签之间的线性和非线性依赖程度^[11]。遥测参数特征 X 和目标标签 Y 之间的互信息为

$$I(X, Y) = H(X) - H(X|Y). \quad (9)$$

其中, $H(X)$ 表示遥测参数特征 X 的熵, $H(X|Y)$ 表示在给定目标标签 Y 的条件下遥测参数特征 X 条件概率分布的熵, 二者计算公式分别为

表 1 皮尔逊系数与相关程度的关系
Table 1 Relationship between Pearson coefficient and correlation degree

皮尔逊相关系数绝对值 $ \rho $ 取值	相关程度
$0 \leq \rho < 0.3$	低相关
$0.3 \leq \rho < 0.8$	中相关
$0.8 \leq \rho \leq 1$	高相关

$$H(X) = - \sum_{x \in X} p(x) \text{lb} p(x); \quad (10)$$

$$H(X|Y) = - \sum_{x \in X} \sum_{y \in Y} p(x, y) \text{lb} p(x|y). \quad (11)$$

计算遥测参数数据中每个特征与目标标签之间的互信息, 并对互信息结果进行降序排序, 选取前 50% 的特征进行保留, 即剔除了遥测参数数据中与目标标签相关性较低的特征。

1.5 类别不平衡处理

类别不平衡问题指的是在一个分类问题中某些类的样本数量远多于其他类别的样本数量^[12]。对于分类问题的场景, 如果直接将传统分类器应用于类别不平衡数据集, 由于数据集中各个类别的不平衡比例过大, 分类器极有可能将所有的样本数据均识别为类别数量多的那一类^[13]。在遥测参数数据中类别不平衡问题较为普遍, 这是因为航天任务工作模式调度决定了处于某些工作模式的遥测参数样本占比高, 而处于另一些工作模式的遥测参数样本占比低。因此, 在使用遥测参数识别空间科学卫星工作模式的问题上, 类别不平衡是亟待解决的问题。

处理分类场景下的类别不平衡问题, 通常可以分为数据重采样、改进模型训练算法两大类。其中, 改进模型训练算法主要是代价敏感学习, 但此方法仅改进了模型训练算法, 没有改变数据原始的分布, 存在模型使用的局限性。数据重采样是指从原始数据集入手, 对原始不平衡数据进行调整, 使得不平衡数据在一定程度上达到平衡状态, 从而消除类别不平衡问题。常见的数据重采样方法分为两大类, 即欠采样和过采样。欠采样主要采用某些特定方法移除多类样本中的数据来平衡数据集中的类别分布, 很明显该方法会丢失数据集中有价值的信息, 因此本文使用过采样方法解决数据类别不平衡问题。

常见的过采样技术有随机过采样技术、合成少数类过采样 (Synthetic Minority Oversampling Technique, SMOTE) 技术、合成少数类过采样与 Tomek Link 混合采样 (Synthetic Minority Oversampling Technique-Tomek Links, SMOTE-Tomek) 技术。

1.5.1 随机过采样

随机过采样通过随机性直接复制少数类中的样本达到类别平衡, 该方式虽然不会丢失数据的有价值信息, 但是对少数类中样本的直接复制增加了过拟合的可能性。

1.5.2 SMOTE 过采样

SMOTE 过采样根据预先设定的采样倍率 k , 使用 K 最近邻算法 (K-Nearest Neighbor, KNN) 找到样本点的 k 个近邻样本, 通过在样本点和近邻样本之间人工合成新样本的方式使样本重新达到平衡^[14]。该算法可以有效减缓随机过采样带来的过拟合问题, 但是该方法在进行人工合成样本时, 由于没有考虑到近邻样本可能来自不同类别, 因此会加剧类别间重叠的问题。SMOTE 算法步骤如下。

步骤 1 计算少量类别的样本 X_i 与所有样本点的距离。

步骤 2 根据设定的样本不平衡过采样倍率 k , 找到最近的 k 个样本。

步骤 3 每一个新生成的样本为 $X_{\text{new}j} = X_i + \text{rand}(0, 1) \times (X_j - X_i)$ 。

如图 3 所示, $X_{\text{new}1}$, $X_{\text{new}2}$, $X_{\text{new}3}$ 和 $X_{\text{new}4}$ 即为使用 SMOTE 过采样生成的新样本。

1.5.3 SMOTE-Tomek 混合采样

SMOTE-Tomek 混合采样首先利用 SMOTE 过采样对少数类生成新样本, 对得到的扩充数据集, 找到所有的 Tomek Link 并剔除。其中, Tomek Link

定义为: 假设样本点 X_a 和 X_b 属于不同的类别, $d(X_a, X_b)$ 表示两个样本点之间的距离, 如果不存在第三个样本点 X_c 使得 $d(X_c, X_a) < d(X_a, X_b)$ 和 $d(X_c, X_b) < d(X_a, X_b)$ 成立, 称 (X_a, X_b) 为一个 Tomek Link^[15]。

图 4(a)为原始数据集, 可以看到存在样本类别不平衡和样本类别重叠的问题; 图 4(b)为使用 SMOTE 过采样后得到的数据集, 很明显 SMOTE 过采样后加剧了样本类别重叠问题; 对于得到的扩充数据集找到所有的 Tomek Link, 如图 4(c)中虚线框所示; 剔除所有的 Tomek Link, 可以看到数据集中原本的重叠问题

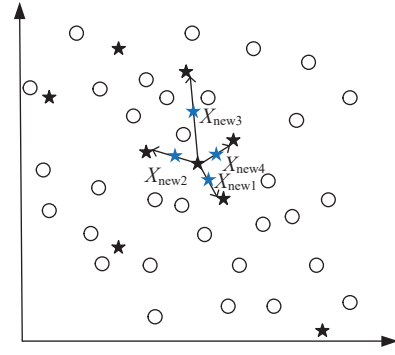


图 3 SMOTE 算法

Fig. 3 SMOTE algorithm

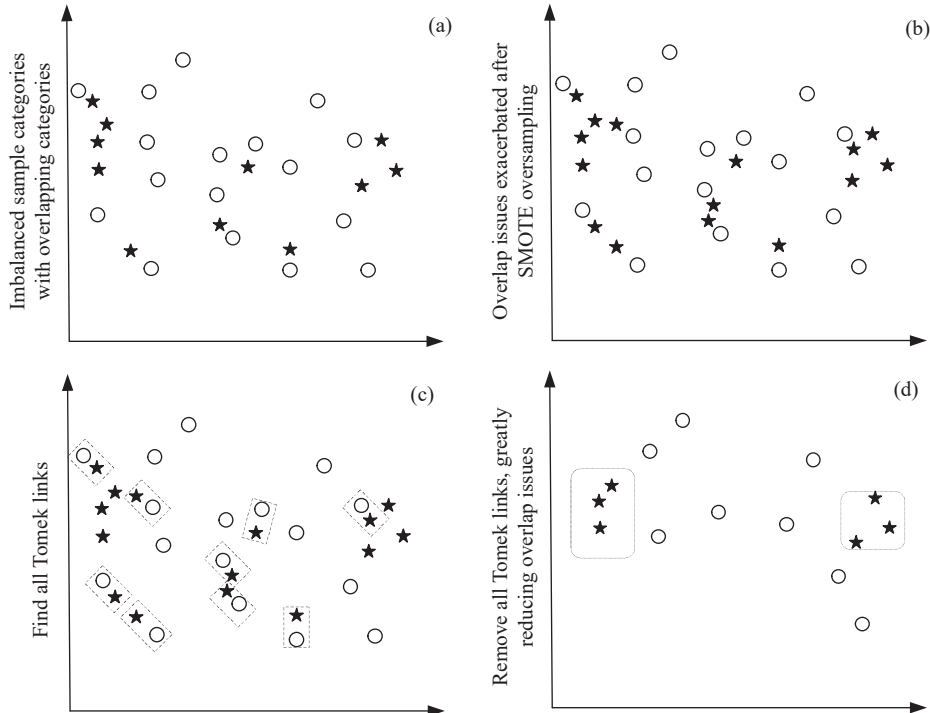


图 4 SMOTE-Tomek 算法

Fig. 4 SMOTE-Tomek algorithm

大大减少,使得分类器更容易对两种类别做出区分,如图4(d)所示。

根据上述三种过采样方式特点可以看出,SMOTE-Tomek 混合采样的性能较为优越,故采取该种过采样方式,同时为了进行比较,也会给出其他过采样方式在数据集上的处理效果。

1.6 随机森林算法原理

随机森林(Random Forest, RF)是一种基于Bagging思想的集成学习算法,由多个弱分类器CART决策树组合成一个强分类器^[16]。相比于传统决策树,随机森林在决策树的训练过程中引入了随机属性选择,具体而言,传统决策树在选择划分属性时是在当前含有 M 个属性结点的属性集合中选择一个最优属性,而在随机森林中,对基决策树的每个结点,先从该结点的属性集合中随机选择一个包含 m 个属性的子集,然后从这个子集中选择一个最优属性用于划分^[17]。构建随机森林并进行预测的示意如图5所示,具体步骤描述如下。

步骤1 对于含有 N 个样本的原始数据集,采取

有放回随机抽样的方式选取 N 个样本。

步骤2 假设每个样本有 M 个属性特征,在决策树的每个结点需要分裂时,随机从这 M 个属性特征中选取 m 个属性特征,满足 $m \ll M$,之后从这 m 个属性中采用Gini系数策略来选择1个属性特征作为该节点的分裂属性。

步骤3 重复步骤2,使得决策树形成过程中每个节点进行分裂,直到不能再分裂为止,整个决策树形成过程中不进行剪枝。

步骤4 重复步骤1~3建立 K 棵基决策树,此时构建完成随机森林。

步骤5 对输入样例进行预测,每棵基决策树都产生一个预测结果,对这 K 个预测结果进行投票,票数最高者作为预测结果输出。

1.7 评估指标

对数据集样本进行分类,结果有4种情况:真阳性(TP),即预测为正,真实标签为正;假阳性(FP),即预测为正,真实标签为负;假阴性(FN),即预测为负,真实标签为正;真阴性(TN),即预测为负,真实标签

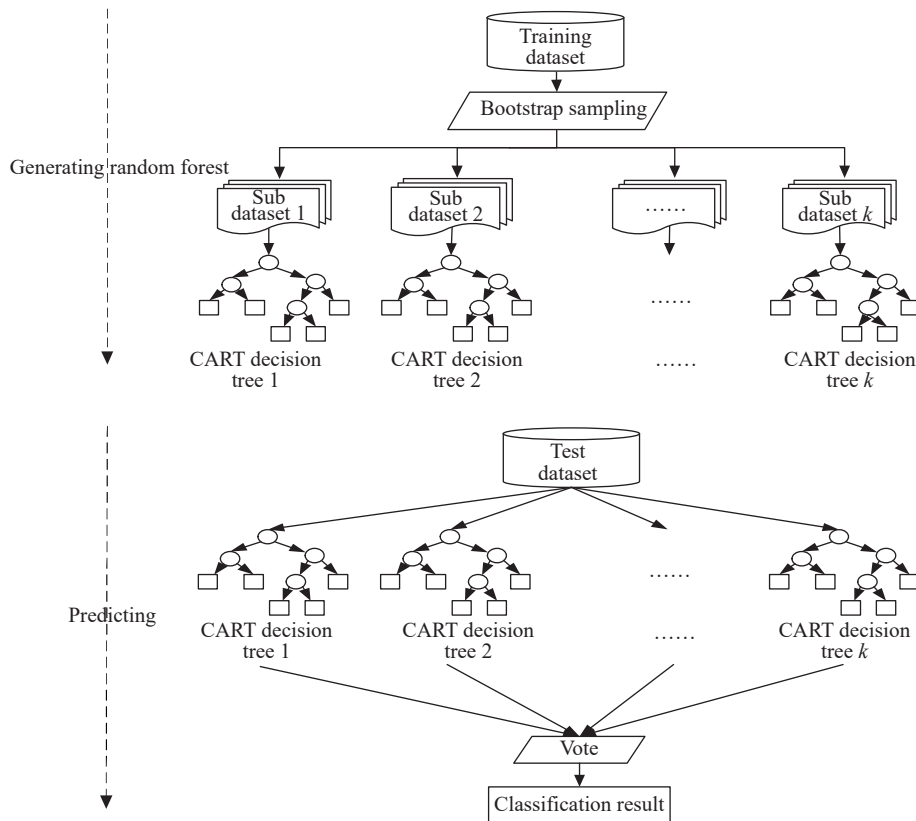


图5 随机森林算法

Fig. 5 Random forest algorithm

为负。本文采用正确率和 F1-score 作为评价指标描述算法模型的分类效果。

1.7.1 正确率

正确率 (Accuracy) 为正确预测样本数占总样本数的百分比, 表征算法模型对所有类别的整体分类表现。对于多标签分类任务, 正确率的计算公式为

$$A = \frac{\sum_{i=1}^K T_{P_i}}{T_{P_1} + F_{P_1} + F_{N_1} + T_{N_1}}. \quad (12)$$

其中, K 表示数据集中的类别数, T_P 表示预测为正且真实标签为正的数, F_P 表示预测为正且真实标签为负的数量, F_N 表示预测为负且真实标签为正的数, T_N 表示预测为负且真实标签为负的数量。

1.7.2 F1-score

F1-score 综合考虑分类结果的精确率 (Precision) 和召回率 (Recall) 的值, 如果精确率和召回率都比较高, 则 F1-score 的值会比较高, 此时分类器对少数类和多数类的识别精度均较高, 适用于类别不平衡数据集场景下的算法模型评估。

某一类别 i 的精确率 (Precision) 为

$$P_i = \frac{T_{P_i}}{T_{P_i} + F_{P_i}}. \quad (13)$$

某一类别 i 的召回率 (Recall) 为

$$R_i = \frac{T_{P_i}}{T_{P_i} + F_{N_i}}. \quad (14)$$

多分类任务中对于全部类别的 F1-score 为

$$F_1 = \frac{1}{K} \sum_{i=1}^K \frac{2P_i R_i}{P_i + R_i}. \quad (15)$$

其中 K 表示数据集中的类别数。

2 仿真实验与结果分析

2.1 数据集准备

仿真实验所用数据集来自 2017 年 1 月 1 日至 2019 年 2 月 27 日期间某型号科学卫星在轨产生的遥测参数数据, 选取其中代号 8362 的源包数据作为原始数据集。原始数据集中共有 87317 条数据, 包含 203 维特征, 某型号科学卫星存在 5 种工作模式, 故数据集的标签标注为工作模式 1、工作模式 2、工作模式 3、工作模式 4 和工作模式 5。原始数据集中数据的总缺失数为 826, 总缺失率约为 0.00466%。使用 1.3 节中介绍的箱型图检测法检测数据集中每一个特征所存在的异常值, 对异常值剔除后使用 1.3 节中介绍的缺失值填充方法进行填充。如图 6 所示, 通过绘制特征 V107 和特征 V141 的箱型图, 可以看出红框内的值判定为异常值, 需要将其剔除并重新填充; 如图 7 所示, 通过绘制特征 V130 和特征 V174 的箱型图, 可以看出这两个特征中不存在异常值, 无须进行特殊处理。之后使用 1.4 节给出的方法对数据集进行特征筛选降维, 降维后的数据集特征维度为 76。

对特征筛选降维后的数据集按照 7 : 3 的比例划分为训练集和测试集, 其中训练集用于训练算法模型, 测试集用于测试算法模型对数据的拟合能力和泛

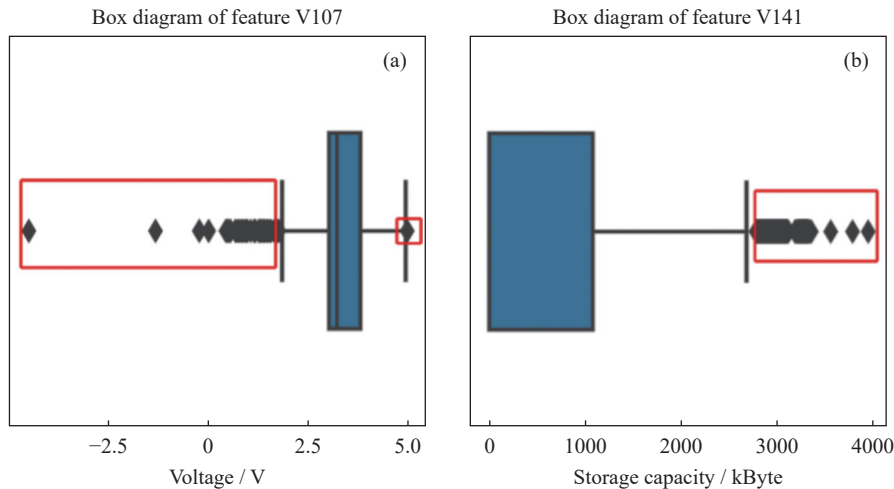


图 6 存在异常值特征的箱型图例

Fig. 6 Examples of box diagram with characteristics of outliers

化能力,此时数据集中的训练集和测试集的样本类别数量分布如图8(a)所示。使用1.5节中介绍的随机过采样、SMOTE过采样、SMOTE-Tomek混合采样三种数据重采样方式对原始数据集中的训练集分别进行处理,处理后得到的扩充数据集的样本类别数量分布分别如图8(b)~(d)所示。

2.2 模型训练与评估

实验环境信息列于表2。使用scikit-learn机器

学习库实现随机森林的算法,在随机森林中,有多个参数需要确定以使模型达到较优的分类效果。其中,参数`max_depth`表示每个基学习器,即CART决策树的最大深度,该值越大则决策树的拟合数据能力越强,该值在超过某个临界值之后,决策树的拟合能力不再发生显著提升;参数`min_samples_split`表示分割节点时考虑特征随机子集的大小,这个值越低,方差减小越多,但偏差的增大也越多。设置`max_depth =`

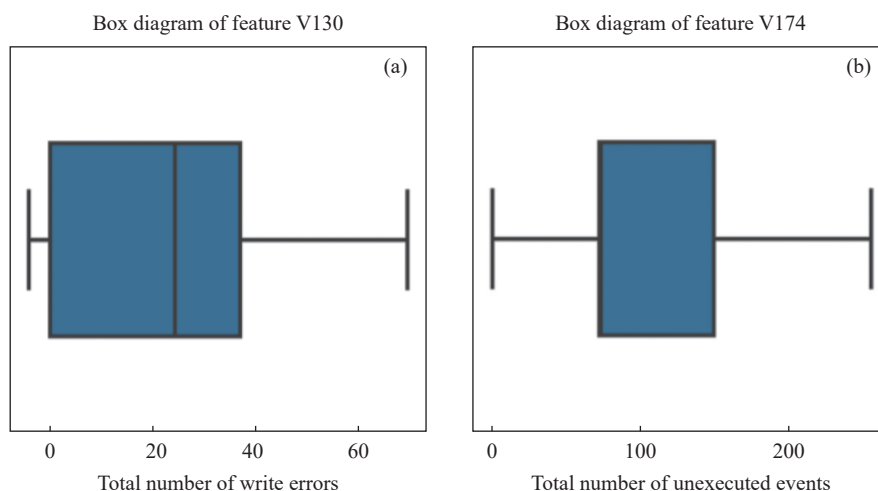


图7 不存在异常值特征的箱型图例

Fig. 7 Examples of box diagram of features without outliers

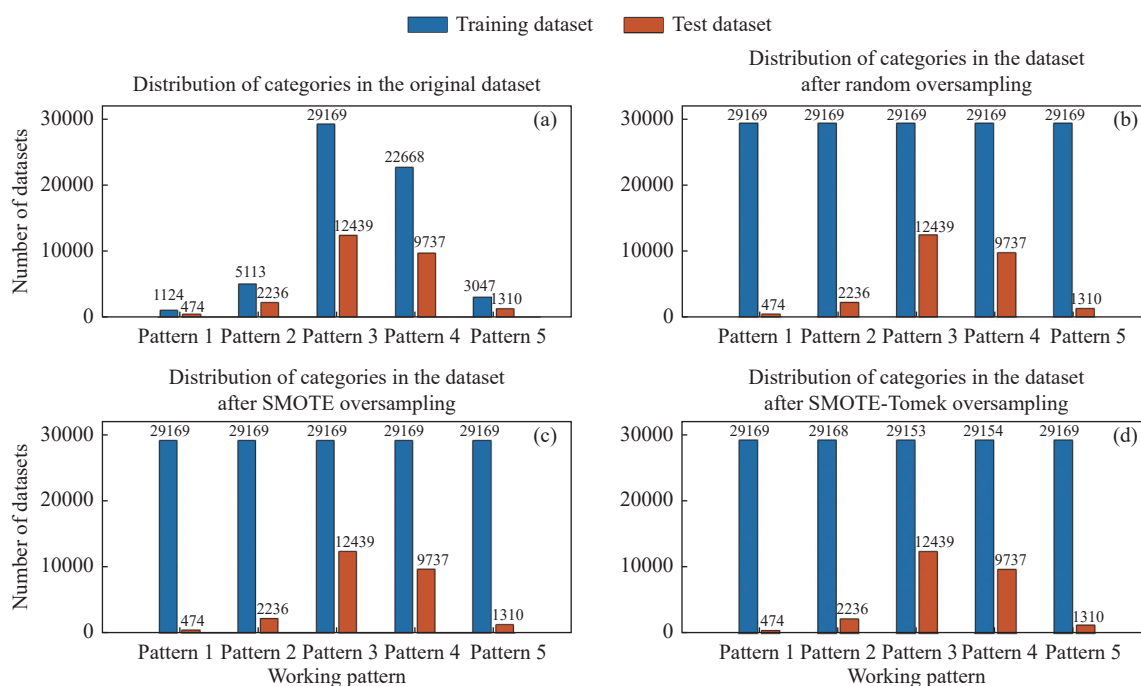


图8 数据集中类别数量分布

Fig. 8 Number distribution of categories in dataset

None结合 $\text{min_samples_spli} = 2$, 为不限制生成一个不修剪的完全树, 可以使得每棵 CART 决策树达到较好的分类性能。

参数 $n_estimators$ 表示随机森林中基学习器即 CART 决策树的个数, 通常数量越大, 算法效果越好, 但是计算时间也会随之增加。当基学习器数量超过一个临界值后, 算法的效果不会显著变好。本文分别在原始训练集、随机过采样处理后的训练集、SMOTE 过采样处理后的训练集和 SMOTE-Tomek 混合采样处理后的训练集上进行训练, 并在测试集上进行测试, 其 F1-score 如图 9 所示, 训练时间如图 10 所示。从图 9 可以看出, 在基学习器数量达到 19 后, 随机森林算法模型的 F1-score 趋于稳定, 不再有显著提升, 同时结合图 10 可以看出, 训练时间随着基学习器数量的增加近乎线性增加, 因此为了平衡训练时间

与算法模型效果, 设定基学习器数为 19, 此时随机森林模型在不同数据集上的性能表现列于表 3。

同时, 为了比较不同机器学习算法模型在空间科学卫星工作模式识别方面的性能差异, 还训练了朴素贝叶斯模型、逻辑回归模型、支持向量机模型和 CART 决策树模型, 每一个算法模型均经过参数调优达到较优的分类性能。几种机器学习算法模型在 SMOTE-Tomek 混合采样处理后数据集上的性能表现列于表 4。

2.3 实验结果分析

观察图 9 可以看出, 随着基学习器数量的增加, 随机森林模型在测试集上的 F1-score 逐渐升高直至趋于稳定, 当基学习器数量 ≥ 19 之后, 分类性能稳定, 不再有显著的变化。这说明基学习器数量的增加确实可以提高分类性能, Bagging 集成学习的思想可以提高弱学习器的分类性能, 体现了群体智慧的优越性。当基学习器数量 ≥ 19 之后, 随机森林模型的性能达到稳定, 此时可以观察到, 随机森林模型在数据重采样后的数据集上的 F1-score 优于原始数据集。这说明随机过采样、SMOTE 过采样、SMOTE-Tomek 混合采样确实能够解决数据集中存在的类别不平衡问题; 并且在三种数据过采样方式中, SMOTE-Tomek 可以达到最高的 F1-score, 这与理论上三者中 SMOTE-Tomek 最优是相符的。

通过观察表 4 可以看出, 在正确率和 F1-score 这

表 2 实验环境信息
Table 2 Experimental environment information

类目	详细信息
CPU	8核i7-1165 G7 2.8 GHz
内存	16 GByte
硬盘	500 GByte
操作系统	Windows 10家庭中文版
编程语言	Python
编程工具	Pycharm Community 2021.3

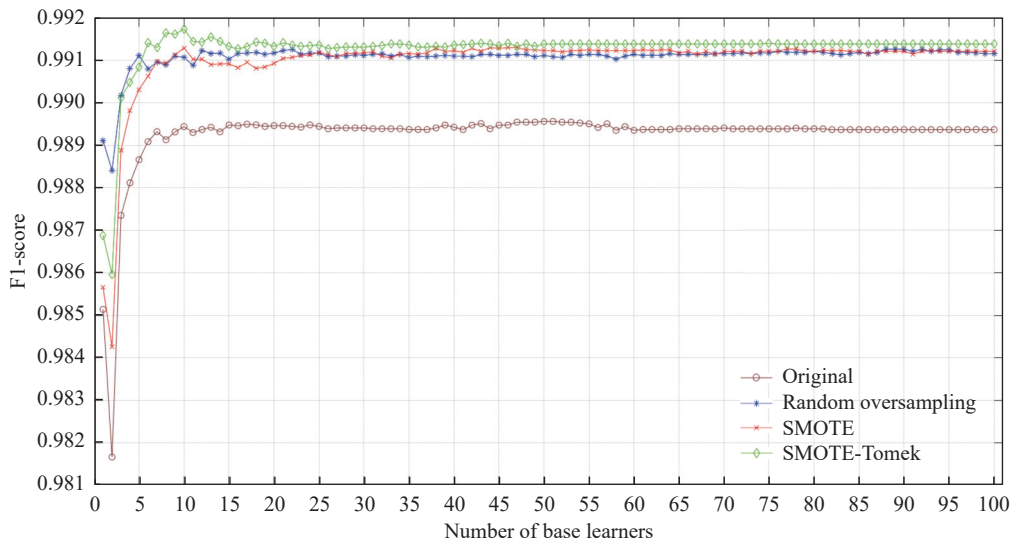


图 9 不同基学习器数量下的 F1-score

Fig. 9 F1-score under different number of basic learners

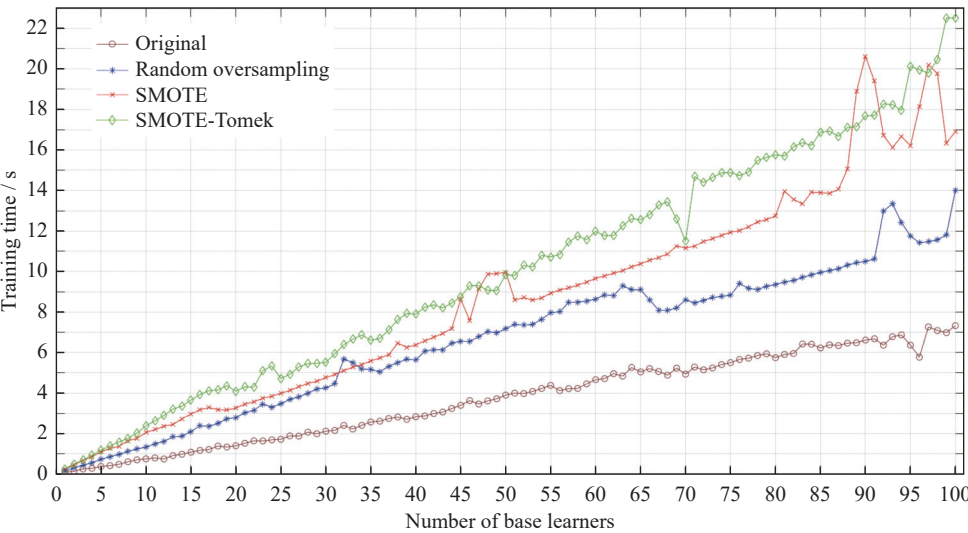


图 10 不同基学习器数量下的训练时间

Fig. 10 Training time under different number of basic learners

表 3 随机森林模型(基学习器数 19)在不同数据集上的性能表现

Table 3 Performance of random forest model (number of base learners 19) on different datasets

数据集类别	正确率	F1-score
原始数据集	0.9966	0.9894
随机过采样后数据集	0.9967	0.9911
SMOTE过采样后数据集	0.9967	0.9908
SMOTE-Tomek过采样后数据集	0.9967	0.9914

注 黑体数字表示该数据在所列中表现最优。

两个评价指标上,相比朴素贝叶斯、逻辑回归、支持向量机和 CART 决策树四种算法模型,决策树模型性能最优,可以达到 99.67% 的正确率和 99.14% 的 F1-score。这说明本文所提方法确实可以学习到遥测参数数据与卫星工作模式之间的内在联系,在空间科学卫星工作模式识别方面可以达到较高的整体预测正确率,对于多数类样本和少数类样本均能正确识别。在训练时间方面,决策树模型虽未达到最短时间,但是 4.36 s 的训练时间也是可以接受的,这说明本文所提方法在构建算法模型方面耗时是比较短的。

3 结语

本文所提出的基于集成学习的空间科学卫星工作模式识别方法,将空间科学卫星工作模式识别问题归约为机器学习领域中的多标签分类问题,借助皮尔

表 4 不同机器学习模型的性能表现

Table 4 Performance of different machine learning models

机器学习模型类别	正确率	F1-score	训练时间/s
朴素贝叶斯	0.7456	0.6176	0.18
逻辑回归	0.9309	0.8664	6.42
支持向量机	0.9169	0.8483	144.36
CART决策树	0.9959	0.9896	2.88
随机森林(基学习器数19)	0.9967	0.9914	4.36

注 黑体数字表示该数据在所列中表现最优。

逊相关系数和互信息理论,实现遥测参数数据筛选降维,使用多种数据重采样技术解决数据集中存在的类别不平衡问题,并构建随机森林集成学习模型,实现空间科学卫星工作模式的识别。本文所提出的方法在工程应用方面具有以下作用。

- (1) 可以作为空间科学卫星在轨运行的辅助监测工具,在耗费少量人力资源的情况下辅助地面运控人员完成对空间科学卫星在轨运行工作模式的监测。
- (2) 为空间科学卫星的指令制定者提供重要的信息反馈,有助于进行指令事后检验。
- (3) 使用遥测参数数据定期训练模型,可以实现模型微调,使其不断适应卫星在轨过程中发生的元器件老化、部件磨损、蓄电池性能退化等问题。
- (4) 当需要更换监视的卫星时,只需要更换新数据集并按照本文方法重新训练模型即可,移植成本低,可扩展性强,耗费人力资源少。

参考文献

- [1] 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 (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)
- [2] MARTÍNEZ-HERAS J A, DONATI A, KIRSCH M G F, *et al.* New Telemetry monitoring paradigm with novelty detection[C]//SpaceOps 2012 Conference. Stockholm, Sweden: AIAA, 2012
- [3] TAGAWA T, YAIRI T, TAKATA N, *et al.* Data monitoring of spacecraft using mixture probabilistic principal component analysis and hidden Semi-Markov models[C]//Proceedings of the 3rd International Conference on Data Mining and Intelligent Information Technology Applications. Macao, China: IEEE, 2011
- [4] 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 (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)
- [5] 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 (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)
- [6] 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 (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)
- [7] XU Yuhang, PI Dechang. Method to mine satellite abnormal patterns[J]. *Journal of Chinese Computer Systems*, 2015, **36**(9): 1988-1992 (徐宇航, 皮德常. 卫星异常模式挖掘方法[J]. *小型微型计算机系统*, 2015, **36**(9): 1988-1992)
- [8] WANG Haotian, LI Xiaorun, ZHAO Liaoying. Cable segmentation method based on box-plot and turning point threshold boundary[J]. *Computer Applications and Software*, 2021, **38**(9): 244-249 (王昊天, 厉小润, 赵辽英. 基于箱型图与折点阈值边界的电缆分割方法[J]. *计算机应用与软件*, 2021, **38**(9): 244-249)
- [9] HAN Xia, LI Xiuxia, SHI Shengnan, *et al.* Research fronts identification based on Z-Score and Sen's Slope method—taking the field of library science as an example[J]. *Information Science*, 2020, **38**(1): 93-97,139 (韩霞, 李秀霞, 史盛楠, 等. 基于Z分数与Sen's斜率的研究前沿识别方法——以图书馆学领域为例[J]. *情报科学*, 2020, **38**(1): 93-97,139)
- [10] JI Deyang, JIN Feng, DONG Lei, *et al.* Data repairing of photovoltaic power plant based on pearson correlation coefficient[J]. *Proceedings of the CSEE*, 2022, **42**(4): 1514-1522 (纪德洋, 金锋, 冬雷, 等. 基于皮尔逊相关系数的光伏电站数据修复[J]. *中国电机工程学报*, 2022, **42**(4): 1514-1522)
- [11] XU Xialing, HU Wei, WANG Chunming, *et al.* Research on power systems key feature selection based on combination effect considering the stability rule[J]. *Proceedings of the CSEE*, 2018, **38**(8): 2232-2238 (徐遐龄, 胡伟, 王春明, 等. 考虑特征组合效应的电网关键稳定特征筛选方法研究[J]. *中国电机工程学报*, 2018, **38**(8): 2232-2238)
- [12] CHAWLA N V, JAPKOWICZ N, KOTCZ A. Editorial: special issue on learning from imbalanced data sets[J]. *ACM SIGKDD Explorations Newsletter*, 2004, **6**(1): 1-3
- [13] VAN DER PUTTEN P, VAN SOMEREN M. A bias-variance analysis of a real world learning problem: the CoLL challenge 2000[J]. *Machine Learning*, 2004, **57**(1/2): 177-195
- [14] CHAWLA N V, BOWYER K W, HALL L O, *et al.* SMOTE: synthetic minority over-sampling technique[J]. *Journal of Artificial Intelligence Research*, 2002, **16**: 321-357
- [15] BATISTA G E A P A, PRATI R C, MONARD M C. A study of the behavior of several methods for balancing machine learning training data[J]. *ACM SIGKDD Explorations Newsletter*, 2004, **6**(1): 20-29
- [16] YANG Sijie. Study on Rapid Recognition of Marine Microplastics Based on Raman Spectroscopy[D]. Harbin: Harbin Institute of Technology, 2021. DOI: 10.27061/d.cnki.ghgdu.2021.004188 (杨思节. 基于拉曼光谱的海水微塑料快速识别方法研究[D]. 哈尔滨: 哈尔滨工业大学, 2021. DOI: 10.27061/d.cnki.ghgdu.2021.004188)
- [17] ZHOU Zhihua. Machine Learning[M]. Beijing: Tsinghua University Press, 2018: 1-415 (周志华. 机器学习[M]. 北京: 清华大学出版社, 2018: 1-415)