文章编号:1671-4598(2022)11-0213-06 DOI:10.16526/j.cnki.11—4762/tp.2022.11.031 中图分类号:TP181 文献标识码:A

基于随机森林的卫星快变遥测数据建模

张雪欢，孙剑伟，赵黛岩
（中国电子科技集团公司第十五研究所，北京 100083）

摘要：现代卫星已逐渐成为国家重大基础设施，为了解其在轨运行状态，需要对遥测数据进行分析；其中快变遥测数据包含了大量卫星服务情况信息，对该数据进行基于机器学习算法的分析建模，可以更好利用特征维度高、数据量大的快变遥测数据，为人工智能在卫星数据建模、运维方面提供一种可能方案；提出一种基于随机森林算法对在轨卫星快变遥测数据进行建模的方法，并引入改进的二次网格搜索方法对模型参数进行调优；使用模型对某频点功率测量值进行预测，结果显示 R^2 值达到 0.98 以上，88%以上的预测值误差不超过±2%，且预测值误差均不超过±5%，建立了效果较好的快变遥测数据模型，为实现基于机器学习的快变遥测数据分析提供了一种可能的方案。

关键词：遥测数据；机器学习；网格搜索；随机森林；模型预测

Modeling of Fast Changing Telemetry Data of Satellite Based on Random Forest

ZHANG Xuehuan, SUN Jianwei, ZHAO Daiyan
(The 15th Research Institute, China Electronics Technology Group Corporation, Beijing 100083, China)

Abstract: Modern satellites have gradually become a major national infrastructure. In order to understand the satellite on orbit working status, it is necessary to analyze the telemetry data. The fast changing telemetry data contains a large amount of satellite service information. The analysis and modeling of the data based on machine learning algorithm can make better use of the fast changing telemetry data with high feature dimension and large amount data, and provide a possible scheme for artificial intelligence in satellite modeling, operation and maintenance. A method of modeling the fast changing telemetry data of on orbit satellite based on the random forest algorithm is proposed, and an improved dual grid search method is introduced to optimize the model parameters. The model is used to predict the power measurement value of certain frequency point. The results show that the R^2 value is more than 0.98, 88% of the predicted values reaches an error of less than ±2%, and all the predicted values have an error of less than ±5%. A fast changing telemetry data model with good effect is established, which provides the possible scheme for the analysis of fast changing telemetry data based on machine learning.

Keywords: telemetry data; machine learning; grid search; random forest; model prediction

0 引言

现代卫星功能多、价值大，需要其具备提供高连续性服务的能力[1]。卫星长期运行在距地面数百至数万公里的真空、极温、强辐射太空环境中。为了解其在轨工作状态，及时发现问题，地面技术人员需要对采集的遥测数据进行分析[2]。

卫星快变遥测数据包由该领域专家根据卫星有效载荷提取核心参数组成，数据包参数达到上百个，包含主备钟状态、各频点功率测量值等，是判断卫星工作状态的重要数据。因快变遥测数据复杂、数据量大，现有使用人工分析对快变遥测数据建模的方法存在效率较低的问题，而将机器学习算法引入快变遥测数据建模中，可以提高建模效率，为卫星遥测数据分析和智能运维提供了参考。

目前，许多学者在卫星遥测数据建模方面开展了大量研究。Xu[3]针对遥测数据值不平稳和周期变化的特性，使用小波分析方法建立卫星电压、功率遥测值模型，并利用周期延拓的方法对模型进行完善，结果表明模型预测值和实际值吻合良好。Sazonov[4]使用国际空间站“曙光”功能舱近似遥测数据建立太阳能电池数学模型，可以在 3～4%的误差范围预测发电量。张弓[5]建立基于改进 SumSin 的导航卫星服务舱光学太阳反射镜温度模型，并对温度趋势进行预测，平均误差在 0.01 ℃左右。梅玉航[6]采用动态加权集成学习方法建立遥测数据模型，结合集成学习和多层感知机的算法提高了预测实时性。王旭[7]使用多种机器学习方法，建立星载铷钟遥测参数模型，并使用模型对锁定信号值进行预测，效果较好的模型均方差为 5 左右。但上述研究均只提取了少量遥测数据参数进行模型建立，对于包含大量参数（上百个）的高维遥测数据研究较少。同时，目前尚未有研究将机器学习算法应用到卫星快变遥测数据建模中。本文拟将随机森林算法应用于卫星快变遥测数据回归模型的建立，使用模型对某频点功率测量值进行预测，采用 R^2 值、预测误差率等作为评估标准，结果显示该模型拥有较好的预测效果，为卫星快变遥测数据建模提供了一

收稿日期:2022-07-21； **修回日期**:2022-08-12。
基金项目:国防预研项目(GFZX03010105280203)。
作者简介:张雪欢(1997-),男,河北石家庄人,硕士研究生,主要从事卫星运维方向的研究。
引用格式:张雪欢,孙剑伟,赵黛岩.基于随机森林的卫星快变遥测数据建模[J].计算机测量与控制,2022,30(11):213-218.

种可行方法，为人工智能在卫星运维方面的应用提供思路。

1 随机森林算法

随机森林（random forest，RF）算法由 Leo[8] 在 2001 年提出，它是一种基于决策树的集成学习方法。

决策树是一种经典的机器学习算法，作为一种树模型，其树状结构直观、可解释性强，被广泛应用于数据分析领域[9]。常见的决策树算法包括 ID3（iterative dichotomiser 3）算法、C4.5 算法和 CART（classification and regression tree）算法，三种算法的主要区别在于节点分裂标准。ID3 算法使用信息增益作为节点分裂标准，这种建树方法较为简单，但信息增益标准会偏袒取值较多的属性。C4.5 算法使用信息增益率作为节点分裂标准，这种方法避免了信息增益标准对取值较多属性的偏好，但因其需要对数据集进行多次计算，导致算法效率较低。CART 算法使用基尼系数作为节点分裂标准，这种方法通过建立二叉树的方式简化计算，效率较高。

虽然决策树具有简单直观、可解释性强等优点，但其极易过拟合，为了解决这一问题，随机森林算法应运而生。随机森林算法可以使用多个决策树共同完成学习任务，解决单一学习器训练结果不准确、容易过拟合等问题，提高算法对噪声的容忍度，拥有更好的泛化性能[10-11]。随机森林可以用于解决分类和回归两种问题[12-13]。在解决分类问题时，随机森林方法根据每棵树的分类结果选择多数作为最终结果；在解决回归问题时，随机森林方法则通过计算每棵树预测值的平均值作为结果[14]。

本文主要使用随机森林处理回归问题，随机森林回归算法的基本原理为：首先，通过 bootstrap 抽样在原始数据集中有放回地随机抽取数据组成训练样本集，其中，需要保证训练样本容量与原始样本容量一致，并且重复多次创建不同的训练样本集[15-16]。然后，根据抽取的训练样本集分别构建决策树，得到各决策树的回归结果。最后，对各决策树的回归结果计算均值得到最终结果。随机森林回归算法原理示意图如图 1 所示。

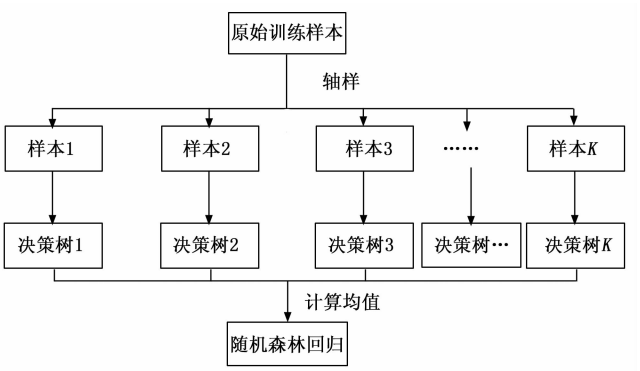

图 1 随机森林回归算法原理示意图

随机森林回归算法的数学推导为：对原始数据集中自变量（输入数据）X 和因变量（需预测输出数据）Y，假设（X，Y）的分布独立，随机在（X，Y）中抽取训练样本集 K，预测结果设为 g（X），则其均方泛化误差为：

$$E_{X,Y}[Y-g(X)]^2 \tag{1}$$

假定有 k 颗决策树，对 k 颗决策树计算其预测值 $\{g(K,X_k)\}$ 的均值得到随机森林回归的预测结果。当 $k\to\infty$ 时，有下式：

$$E_{X,Y}[Y-\bar{g_k}(X,K_k)]^2 \to E_{X,Y}[Y-E_K(X,K_k)]^2 \tag{2}$$

式中，$E_{X,Y}[Y-E_K(X,K_k)]^2$ 表示泛化误差，记为 PE^{**}，当 k 趋近于无穷大时，每颗决策树的泛化误差记为 PE^*，PE^* 满足：

$$PE^* = E_K E_{X,Y}[Y-g(X,K)]^2 \tag{3}$$

其中：K 满足：

$$PE^{**} \leqslant \bar{\rho} PE^* \tag{4}$$

其中：$\bar{\rho}$ 为残差的加权关联系数。最终随机森林的回归函数为：

$$Y = E_k g(X,K) \tag{5}$$

2 数据处理及模型建立

2.1 数据描述

本实验采用某卫星于 2022 年 2 月 21 日 18 时至 2022 年 2 月 21 日 22 时，4 小时内产生的快变遥测数据。其中，数据采样率为 1 条/秒，4 小时内共收集 14 400 条数据，每条数据包含 103 个特征，数据维度和数据量较大。

快变遥测数据以 .csv 的格式存储，为了将数据读入算法中，本文使用 pandas 包中 pandas.read _ csv（）函数。该函数用法简便，只需将原始数据的 .csv 格式文件的绝对路径作为函数参数，便可将快变遥测数据存至 pandas 包中定义的 DataFrame 数据结构中。DataFrame 是一种二维数组，由索引和内容组成，存入 DataFrame 后可以方便的使用 Python 中函数对数据进行分析处理。

需要注意的是，采样得到的快变遥测数据值取自星上发送的原始数据值，部分数据值含有字符，直接进行数据处理会因字符型值无法转换为数值型而出现错误，需要对快变遥测数据进行修正。含有字符的数据值存在三类情况：（1）原始数据值由十六进制数表示导致采样数据值中含有字符，这类情况需要将十六进制数转换为十进制数。（2）原始数据值包含字符用于分隔数据，在这类情况中，字符并无表示数据的实际意义，直接删除即可。（3）快变遥测数据中部分参数为状态参数，使用不同字符代表不同状态，这类情况需要将不同字符转化为离散数值，使用离散数值代表原始数据代表的不同状态。

完成修正后将数据按照 10 000 条和 4 400 条划分为训练集和测试集，准备进行特征预处理。

2.2 特征预处理

2.2.1 野值剔除

在地面接收来自卫星的遥测数据过程中，受天气、磁场等多种环境因素作用，接收到的遥测数据可能与卫星发送的数据产生较大偏差，这种数据被称为野值。对卫星遥测数据进行处理时，其数据准确性会直接影响遥测数据分析建模效果，如果数据中存在野值，容易造成误判，为地面技术人员分析卫星服务状态增加干扰。

常见的野值剔除方法包括 3σ 准则、奈尔准则、53H 准则等。本文使用 53H 准则进行野值剔除，其剔除方法为首先对数据值序列求两次中值得到新的数据值序列。然后将新序列通过下式组合成参考值。

$$y(i) = 0.25 \cdot x_{new}(i-1) + 0.5 \cdot x_{new}(i) + 0.25 \cdot x_{new}(i+1) \quad (6)$$

最后，若有下式成立则当前值为野值，并用参考值替换。

$$|y(i) - x(i)| > t \quad (7)$$

在代码实现上，本文利用 Python 的数据处理功能，循环遍历所有数据值对所需的各类数据进行计算，得到参考序列，以此为标准进行野值剔除。

2.2.2　特征归一化

卫星快变遥测数据维度较高，其中包含多种有效载荷产生的不同类别特征，各个特征量纲不同、物理含义也不同，需要对数据进行归一化，防止部分特征数量级较大导致特征对模型的影响大于其他特征，造成模型偏差变大，影响最终的模型效果。同时，归一化操作还可以使模型收敛速度加快，提高模型构建效率[17]。

归一化方法有 Min-Max 归一化、Sigmod 归一化等。本文使用 Min-Max 归一化方法对快变遥测数据进行处理，其变换函数如式（8）：

$$x^* = \frac{x - x_{min}}{x_{max} - x_{min}} \quad (8)$$

由上述公式可知，Min-Max 归一化将数据中最大值和最小值作为映射标准，对原始数据进行线性变换。由于原始数据均不会超过最大值，因此可以将各个数据值等比例的映射至［0，1］的范围，实现对原始数据的等比缩放。

在代码实现上，本文选择 sklearn.preprocessing 包中 MinMaxScaler 方法。需要注意的是，在对训练集进行归一化后，还需对测试集进行归一化，否则将会因为训练集、测试集特征数量级不一致导致使用测试集得到的预测值大幅度偏离真实值。同时，sklearn 的 MinMaxScaler 方法使用 fit _ transform 函数对训练集进行归一化，使用 transform 函数对测试集进行归一化，保证训练集、测试集的归一化参数一致。如果对测试集也使用 fit _ transform 函数会导致两者归一化参数不同、处理方式不同，从而对预测结果产生影响。

2.2.3　PCA 降维

主成分分析（principal component analysis，PCA）是一种常用特征工程方法，PCA 使用正交变换方法将原始变量转换为不相关的变量，得到的一组新变量为主成分[18]。

维度较高的快变遥测数据直接建立模型可能会造成“维度灾难”，而 PCA 可以将高维向量转换为低维向量来解决问题。

在代码实现上，本文选择 sklearn.decomposition 包中 PCA 方法。参数选择 n _ components＝0.99、svd _ solver＝“full”。其中 n _ components 影响降维后的特征维度，当 n _ components 为正整数 n 时，PCA 方法返回的特征维度为 n；当 n _ components 为［0－1］的浮点数时，PCA 方法返回满足保留 n _ components 指定百分比的信息量的特征维度，并且此时 svd _ solver 需要选择“full”。本文使用 PCA（n _ components＝0.99，svd _ solver＝”full”）函数对归一化后的数据进行降维，处理后特征维度为 18 维，显著降低了数据复杂度。

2.3　模型建立

建立基于随机森林的卫星快变遥测数据回归模型，使用 2.1 节选取的快变遥测数据中某频点功率测量值作为回归模型预测值，快变遥测数据剩余参数作为输入值，实现基于随机森林的卫星快变遥测数据某频点功率测量值回归预测模型，其主要步骤为：

1）卫星快变遥测数据获取。采用某卫星产生的 4 小时快变遥测数据作为原始数据，并根据 2.1 节介绍的原始数据修正方法对数据进行修正。

2）训练集与测试集划分。将步骤 1）中获取的修正后原始数据按照 10 000 条和 4 400 条的比例划分为训练集和测试集。

3）数据预处理。首先将卫星快变遥测数据集进行野值剔除，根据 2.2.1 节介绍的方法对野值进行处理。然后进行特征 Min-Max 归一化处理，根据 2.2.2 节介绍的归一化方法将原始数据等比映射至［0，1］范围内。最后进行 PCA 降维处理，根据 2.2.3 节介绍的 PCA 降维方法降低数据复杂度。

4）模型参数选取。针对随机森林算法，对 4.2 节确定的重要参数 n _ estimators 和 max _ depth 通过改进的二次网格搜索方法循环遍历所有候选参数，并通过 3.1 节介绍的评价指标优选参数。

5）随机森林回归模型构建。根据步骤 4）选取的最优参数，使用训练集数据构建随机森林回归模型。构建模型时采用 sklearn. ensemble 包的 RandomForestRegressor 函数。

6）模型预测结果分析。将测试集数据输入步骤 5）构建的随机森林回归模型，对某频点功率测量值进行预测，使用 3.1 节的评价指标进行模型预测结果分析评价。

根据以上步骤，得到基于随机森林的卫星快变遥测数据回归模型流程图如图 2 所示。

3　实验分析

为验证基于随机森林的卫星快变遥测数据模型效果，使用 Python 语言和 Jupyter Notebook 开发工具进行实验，参照 2.3 节所述流程建立对快变遥测数据中重要参数——某频点功率测量值进行预测的回归模型，再利用运行时间、误差率等指标对预测效果进行评价，从而实现模型效果分析。实验的主要步骤为：

1）数据处理。根据 2.1 节和 2.2 节方法使用 Python 库获取实验所需数据，并依据建模和效果分析需求将数据划分为训练集和测试集。

2）回归模型建立。使用 1）中划分的训练集数据，运行 Python 中 sklearn 库 RandomForestRegressor 函数，依据 2.3 节中构建模型子流程建立训练集数据回归模型。

3）模型预测。使用 2）中建立的回归模型对 1）中划分

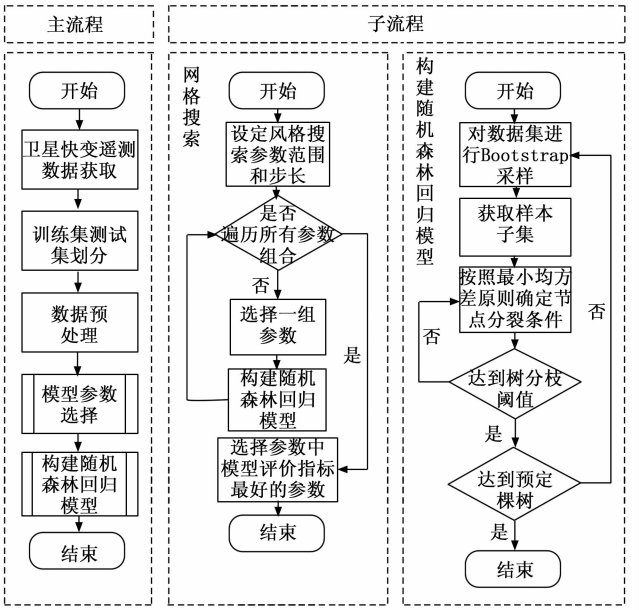

图 2 模型建立流程

的测试集数据进行预测。保存预测值准备进行模型评估。

4）模型评估。按照 3.1 节选取的评价指标，使用 Python 库中 time 函数计算运行时间，可以代表当前模型在数据集上的效率；使用 sklearn 中 score 函数计算 R^2 值，可以表示模型拟合时产生的偏差；使用 sklearn 中 mean _ absolute _ error 函数计算 MAE 值，表示不考虑方向的预测值平均误差程度；使用 Python 库计算误差率及误差率分布，可以直观展示各预测值与其对应的真实值的偏差；绘制模型预测曲线。计算得到各评价指标结果后，根据结果对模型运行效率、模型预测误差进行分析，评估回归模型效果。

5）对比实验及分析。通过对比实验展示随机森林回归模型效果。采用默认参数随机森林、逻辑回归、K 近邻和多层感知机建立回归模型，使用模型进行预测和效果评估，流程参照步骤 2）～4）。模型单独评估后，再根据评价指标对比各个模型预测情况，分析模型效果。

根据以上步骤，得到实验流程图如图 3 所示。

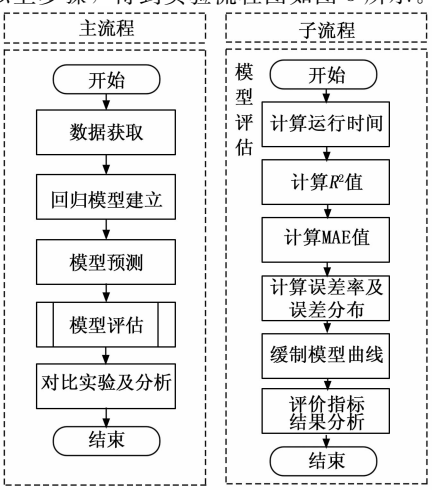

图 3 实验流程图

3.1 评价指标

使用某频点功率测量值作为回归模型预测值，快变遥测数据剩余参数作为输入值，利用 3.2 节选取的使用最优参数的回归模型对某频点功率测量值进行预测，通过预测效果评价模型效果，预测效果指标如下。

3.1.1 运行时间

运行时间为各个模型使用训练集进行拟合和模型使用测试集进行预测的时间。可以代表当前模型在数据集上的效率。针对在轨卫星数据分析这一场景，地面技术人员需要实时了解在轨卫星服务状态，以便对卫星运行时的各类问题做出快速反应。同时，在数据处理方面，卫星快变遥测数据采集间隔短，庞大的数据采集量要求研究人员尽可能提高数据分析处理效率，因此需要选择能够快速生成预测结果的模型。

基于以上要求，本文选取运行时间作为评价指标，通过计算模型在数据集上的运行时间，对模型效率进行表征，运行时间越短，表示模型在数据集上的效率越高。

在代码实现上，选用 Python 中 time 函数，在模型开始拟合前运行 time 函数，并将其记录在 start 变量中，当模型完成预测时再次运行 time 函数，并将其记录在 end 变量中，二者做差便可得到运行时间。

3.1.2 R^2

R^2，亦被称为决定系数、可决系数，表示目标变量在回归中被其他变量（解释变量）拟合时产生的偏差。如果 R^2 小于零，表示模型的预测效果非常差，如果 R^2 大于零，则 R^2 值越大，模型的预测效果越好[19-20]。

计算 R^2 需要样本的残差平方和 RSS（residual sum of squares）以及总平方和 TSS（total sum of squares），其公式如式（9）和（10）：

$$RSS = \sum_{i=1}^{m} (\hat{y}_i - y_i)^2 \tag{9}$$

$$TSS = \sum_{i=1}^{m} (y_i - \bar{y})^2 \tag{10}$$

其中：$\hat{y}_i$ 为预测值，y_i 为真实值，$\bar{y}$ 为真实值的均值。得到 RSS 和 TSS 后，可以由公式（11）计算 R^2 值：

$$R^2 = 1 - \frac{RSS}{TSS} \tag{11}$$

在 sklearn 中，预测模型的 R^2 值可以通过模型的 score 函数得到，例如，针对随机森林模型可以使用 RandomForestRegressor. score（testx1，testy1）。其中，testx1 为测试集输入值，testy1 为测试集真实值。

3.1.3 平均绝对误差

平均绝对误差（mean absolute error，MAE）源于平均误差的度量，是真实值与预测值之差绝对值的和，可以表示不考虑方向的预测值平均误差程度，通常用于评估回归模型。

平均误差的另一种形式是均方根误差（root mean squard error，RMSE），但在大多数情况下，MAE 在测量平均模型精度方面优于 RMSE[21-22]，因此选择 MAE 作为一种评价指标，其公式如（12）所示：

$$MAE = \frac{1}{m}\sum_{i=1}^{m} |(\hat{y}_i - y_i)| \tag{12}$$

其中：$\hat{y}_i$ 为预测值，y_i 为真实值。

在 sklearn 中，预测模型的 MAE 可以通过 metrics 包 mean _ absolute _ error 函数得到。

3.1.4　误差率

为更加直观展示各预测值与其对应的真实值的偏差，除了 3.1.3 节所述平均绝对误差外，本文还引入误差率这一评估指标。通过计算预测值、真实值之差对真实值的比例，可以得到每个预测值的偏差程度，误差率公式如（13）：

$$e = \frac{\hat{y}_i - y_i}{y_i} \tag{13}$$

其中：$\hat{y}_i$ 为预测值，y_i 为真实值，i 可取所有测试集真实值。

3.2　模型参数选择

机器学习算法参数是在开始学习过程之前设置的参数，其对模型效果有较大影响。机器学习算法参数定义了关于模型的更高层次的概念，如复杂性或学习能力。针对随机森林算法，重要的参数包括 n _ estimators 和 max _ depth，分别代表随机森林中基学习器的数量和基学习器的最大深度[23]。

为了建立效果较好的卫星快变遥测数据随机森林回归模型，本文采用改进的二次网格搜索方法对上述两个参数进行调优。二次网格搜索方法设置两次搜索循环，第一次循环时设置较大的参数搜索范围，并设置较大的循环步长，可以在扩大搜索范围的同时防止时间开销过大。第二次循环时，通过第一次搜索得到的较优参数缩小搜索范围，并设置步长为 1，从而得到最优参数组合。二次网格搜索方法相比普通的网格搜索方法，通过一次大范围大步长搜索和一次小范围小步长搜索，显著降低了网格搜索的时间开销。

具体到本文模型，应用二次网格搜索，首先将 n _ estimators 设置为范围 30～300、步长 10，max _ depth 设置为范围 5～100、步长 5，通过嵌套循环搜索每一种参数组合。分析结果，n _ estimators 为 40 和 180、max _ depth 为 5 时均取得 score＝0.984、MAE＝12.27，但 n _ estimators 为 180 时运行时间为 7.87 s，远大于 40 时的 1.9 s，因此将新范围确定为 n _ estimators：30～50、max _ depth：1～10，步长均为 1，并再次进行嵌套循环。对第二次搜索结果进行分析，得到 n _ estimators 为 39、max _ depth 为 3 时有最优结果 score＝0.984、MAE＝12.25、运行时间 1.09 s。因此最终确定基于随机森林的快变遥测模型参数为 n _ estimators＝39、max _ depth＝3。

3.3　模型预测结果分析

使用测试集数据分析模型效果，随机选择 30 对预测值和真实值画出随机森林模型的预测曲线，如图 4 所示。

同时，分析模型的预测误差率，计算得到测试集 4 400 个数据中共有 3 912 个数据误差率小于 2%，并且最大误差率不超过±5%。误差率分布如表 1 所示。

结合 score＝0.984、MAE＝12.25、运行时间 1.09 s 共四个评价指标可知，模型可以较好地预测某频点功率测量值，且模型运行效率高。

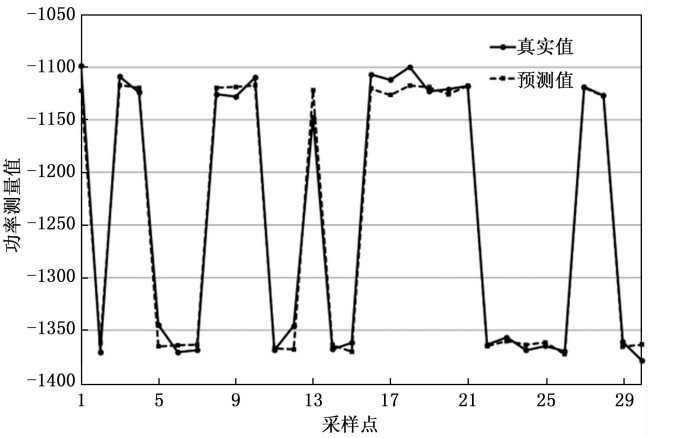

图 4　随机森林模型预测曲线

表 1　随机森林模型预测误差率及分布

误差率	样本数量	所占比例/%
±2%以内	3 912	88.91
−5%至−2%以及 2%至 5%	488	11.09
小于−5%以及大于 5%	0	0

上述模型使用改进二次网格搜索算法对模型进行了参数调优，可以得到模型的最优参数，提高模型预测效果。若不进行参数调优，使用默认参数直接建模其误差率如表 2 所示。

表 2　默认参数随机森林模型预测误差率及分布

误差率	样本数量	所占比例/%
±2%以内	3 821	86.84
−5%至−2%以及 2%至 5%	579	13.16
小于−5%以及大于 5%	0	0

此外，使用默认参数的随机森林模型 score＝0.97、MAE＝12.33、运行时间 39.46 s。由此可知，虽然其在预测误差方面与经过参数调优的随机森林模型差距较小，但默认参数随机森林模型运行时间远远大于参数调优后的模型。通过改进的二次网格搜索得到的最优参数对模型运行效率有显著提升，这对于提高卫星快变遥测数据建模实时性具有一定意义。

除了随机森林算法代表的装袋算法，本文还选择了机器学习中线性算法、非线性算法以及神经网络算法作为对比，具体方法为逻辑回归、K 近邻、多层感知机。三种算法的误差率如表 3～5 所示。

表 3　逻辑回归模型预测误差率及分布

误差率	样本数量	所占比例/%
±2%以内	2 032	46.18
−5%至−2%以及 2%至 5%	205	4.66
小于−5%以及大于 5%	2 163	49.16

表 4 K 近邻模型预测误差率及分布

误差率	样本数量	所占比例/%
±2%以内	711	16.16
−5%至−2%以及 2%至 5%	298	6.77
小于−5%以及大于 5%	3 391	77.07

表 5 多层感知机模型预测误差率及分布

误差率	样本数量	所占比例/%
±2%以内	3 634	82.59
−5%至−2%以及 2%至 5%	761	17.30
小于−5%以及大于 5%	5	0.11

四种算法的 score、MAE、运行时间对比如表 6 所示。

表 6 四种算法的评价指标对比

算法	score	MAE	运行时间/s
随机森林	0.98	12.25	1.09
逻辑回归	0.01	128.69	5.25
K 近邻	0.06	112.37	0.27
多层感知机	0.97	14.60	8.09

对比发现，随机森林算法在±2%以内误差率样本数量、score 和 MAE 三个指标上明显好于逻辑回归和 K 近邻算法。同时，虽然多层感知机在误差率、score 和 MAE 方面较为接近随机森林算法，但多层感知机的训练时间长、效率不高，类似未经参数调优的随机森林模型，多层感知机在遥测数据建模方面实时性较差，具有一定劣势。因此随机森林算法在卫星快变遥测数据建模方面优于其他几种方法。

4 结束语

实现卫星快变遥测数据建模有助于了解卫星服务状态，推动人工智能在卫星运维中的应用。本文使用随机森林算法建立卫星快变遥测数据模型，对某频点功率测量值进行预测，结果显示模型预测效果较好、运行效率高。对比逻辑回归、K 近邻和多层感知机算法，随机森林算法在评价指标上具有明显优势。然而，在实验过程中多层感知机算法也表现出了极大的潜力。作为神经网络的一种基础算法，多层感知机已有较好的效果，在未来的研究中应该重点关注神经网络算法在快变遥测数据建模上的应用，以期获得更好的预测效果。

参考文献：

[1] 李 光，石碧舟，戴永珊，等. 导航卫星载荷自主健康管理研究 [J]. 计算机测量与控制，2021，29 (3)：104－107，113.

[2] 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.

[3] XU J T，LIU P P. Application of Wavelet Analysis in The Prediction of Telemetry Data [J]. International Journal of Advanced Network Monitoring and Controls，2019，4 (2)：28－34.

[4] SAZONOV V V. Determining the Parameters of a Mathematical Model of Spacecraft Solar Batteries from Telemetry Data [J]. Moscow University Computational Mathematics and Cybernetics，2021，45 (3)：120－125.

[5] 张 弓，翟君武，杨海峰. 导航卫星遥测数据趋势预测技术研究 [J]. 航天器工程，2017，26 (3)：70－77.

[6] 梅玉航，贾海艳. 基于动态加权集成学习的遥测数据预测方法 [J]. 计算机测量与控制，2021，29 (10)：144－147.

[7] 王 旭，都晓辉，陈昌麟，等. 机器学习在卫星遥测分析建模中的应用 [J]. 计算机测量与控制，2021，29 (1)：210－214.

[8] BREIMAN L. Random Forests [J]. Machine learning，2001，45 (1)：5－32.

[9] 杨 宁. 决策树模型及其在资本流入急停预测中的应用 [D]. 大连：大连理工大学，2021.

[10] 吴克河，张 英，崔文超，等. 一种基于随机森林算法的 MQTT 异常流量检测方法 [J]. 计算机与现代化，2021 (1)：61－64，119.

[11] 周志华. 机器学习 [M]. 北京：清华大学出版社，2016.

[12] GENUER R，POGGI J M，TULEAU C. Random Forests：some methodological insights [R]. Orsay：Institut National de Recherche en Informatique et en Automatique (INRIA)，2008.

[13] 孙 行. 基于随机森林样本增强的格网 DEM 地形特征点提取方法研究 [D]. 成都：西南交通大学，2021.

[14] QUINLAN J R. Induction of decision trees [J]. Machine Learning，1986，1 (1)：81－106.

[15] BUCHLMANN P，YU B. Analyzing Bagging [J]. The Annals of Statistics，2002，30 (4)：927－961.

[16] 杨 琳，白 钊，寇勇刚. 基于 RFM 模型的随机森林算法对民航客户的流失分析 [J]. 计算机与现代化，2021 (1)：100－104.

[17] 邹建成. 基于非监督算法的卫星异常检测 [D]. 武汉：武汉大学，2020.

[18] 孙宇豪. 基于多变量相关性分析的卫星异常检测技术研究 [D]. 上海：中国科学院大学 (中国科学院微小卫星创新研究院)，2020.

[19] SHAO Y，LI R D，LUO Y J，ZHU M. Research on Running Data Analysis Method Based on Attention-LSTM [C] // Proceedings of 2021 International Conference on Intelligent Transportation，Big Data & Smart City (ICITBS 2021) Part Ⅱ，2021：155－159.

[20] KASUYA E. On the use of r and r squared in correlation and regression [J]. Ecological Research，2019，34 (1)：235－236.

[21] QI J，DU J，SINISCALCHI S M，et al. On Mean Absolute Error for Deep Neural Network Based Vector-to-Vector Regression [J]. IEEE Signal Processing Letters，2020，27 (1)：1485－1489.

[22] 吴 雪. 基于聚类线性回归的预测方法及其在股市预测上的应用 [D]. 绵阳：西南科技大学，2020.

[23] 刘紫亮，居 翔，张永芳，等. 基于改进随机搜索算法的随机森林调参优化 [J]. 网络安全技术与应用，2022 (4)：49－51.