文章编号：1672-8785(2021)08-0033-05

基于随机森林回归分析的脉管制冷机性能预测模型

赵　鹏[1,2]　陆　志[1]　蒋珍华[1]　曲晓萍[1]　吴亦农[1*]

(1. 中国科学院上海技术物理研究所，上海　200083；
2. 中国科学院大学，北京 100049)

摘　要：为了探索星载脉管制冷机相关参数对制冷性能的影响和提高制冷性能的一致性，建立了基于机器学习的随机森林回归(Random Forest Regression，RFR)模型，然后对制冷性能与各个自变量进行了回归预测。制冷性能预测的平均相对误差为 5.62%，平均确定性系数为 0.805。按照特征重要度从高到低排序，前两位分别为丝网填充率和磁感应强度，与实际的实验结果相符(丝网填充率和磁感应强度的实际输入功的变化值分别为 6.11 Wac 和 3.52 Wac，远大于其他 4 个自变量)。研究结果表明，RFR 具有较高的精确度和鲁棒性，为提高星载脉管制冷机性能的一致性提供了新的思路。

关键词：脉管制冷机；随机森林回归；特征重要度

中图分类号：TK123　　**文献标志码：**A　　**DOI：**10.3969/j.issn.1672-8785.2021.08.005

Cooling Performance Prediction Model of Pulse Tube Cryocooler Based on Random Forest Regression Analysis

ZHAO Peng[1,2], LU Zhi[1], JIANG Zhen-hua[1], QU Xiao-ping[1], WU Yi-nong[1*]

(*1. Shanghai Institute of Technical Physics, Chinese Academy of Sciences, Shanghai 200083, China;*
2. University of Chinese Academy of Sciences, Beijing 100049, China)

Abstract: In order to explore the influence of relevant parameters on the cooling performance of space-borne pulse tube cryocooler and improve the consistency of cooling performance, a random forest regression model based on machine learning is established to make regression prediction of the cooling performance and various independent variables. The average relative error of cooling performance prediction is 5.62%, and the average certainty coefficient is 0.805. In terms of the influence degree of the variables, the first and second feature are mesh filling rate and magnetic induction intensity, which are consistent with the actual experimental results (the actual input power changes of mesh filling rate and magnetic induction intensity are 6.11 Wac and 3.52 Wac, which are much larger than the other four independent variables). The results show that RFR has the high accuracy and robustness, which provides a new idea for the consistency improvement of the cooling performance of space-borne pulse tube cryocooler.

收稿日期：2021-05-10

基金项目：国家自然科学基金项目(51806231)

作者简介：赵鹏(1984-)，男，江西抚州人，博士，主要从事空间低温制冷机研究。

*通讯作者：E-mail：wyn@mail.sitp.ac.cn

Key words: pulse tube cryocooler; random forest regression; feature importance

0 引言

脉管制冷机(简称制冷机)是星载红外焦平面探测器组件的主要组成部分之一，其制冷性能直接影响红外相机的成像质量。随着载荷的增加，卫星平台电力资源愈发紧缺，用户对制冷机性能一致性的要求也越来越高。因此，通过回归方法从众多影响制冷性能的参数中识别出重要参数，对于提高制冷性能的一致性具有重要意义。

国内外学者从制冷机理的角度对制冷机性能的优化做了大量的研究工作[1-2]，推动了制冷机性能的逐步提升。过去大部分研究中参数变量的变化范围相对较大。制冷机各参数变量在工艺控制范围内小幅变化的性能一致性研究，暂时未见公开报道。制冷机样本数据的积累为制冷性能的回归预测提供了数据支撑。RFR 具有运行效率高、抗干扰能力强以及预测精确度高等优点。因此，本文采用 RFR 分析制冷性能，并将其与实验结果进行对比分析。

本文首先对用于 RFR 的自变量和因变量(制冷性能)的选取准则进行了说明，并给出用于回归的数据；然后对 RFR 的原理及流程进行介绍；最后分析 RFR 结果的精确度和特征重要度，并将其与制冷机的实验结果进行对比。

1 回归自变量和因变量

在工艺控制范围内，各自变量参数的选取原则如下：理论上会对制冷性能产生一定影响，或波动百分比((最大值－最小值)/平均值)大于 1%。它们分别为密封间隙 x_1、磁感应强度 x_2、电机阻值 x_3、丝网填充率 x_4、丝网丝径 x_5 和丝网厚度 x_6。因变量 y 为 12 W@85 K 制冷量的实际输入功(制冷性能)。

1.1 各自变量的实验结果

通过实验研究得到了在工艺控制范围内的自变量对制冷性能的实际影响结果(见表 1)。其中，丝网填充率 x_4 对制冷性能的影响最大，磁感应强度 x_2 次之，其余四个自变量的影响相对较小。

1.2 回归分析的原始样本数据

表 2 列出了用于回归分析的编号为 1～18 的制冷机原始样本数据。在进行回归分析时，随机选择其中的 14 组数据作为回归模型的实验样本，并将另外 4 组作为模型的验证集。

2 随机森林算法的基本原理

随机森林算法是由 Breiman L 和 Cutler A[3]在 2001 年提出的一种基于决策树的集成算法。它结合了 Breiman L 提出的 Bagging 算法和 Cutler A 提出的随机子空间算法，被广泛应用于医学、金融以及人工智能等领域。

2.1 原理

2.1.1 Bagging 算法取样[4]

在原始样本中随机有放回抽取 N 个训练样本，并进行 n_{tree} 次循环。每个循环样本相互独立，并用于构建 n_{tree} 个回归决策树。最后取各个决策树的平均值作为最终的预测结果。在

表 1 基于实验的自变量与制冷性能的关系

自变量	工艺控制范围	验证实验结果
密封间隙	9～12.25 μm	在工艺控制范围内，制冷性能与密封间隙无显著相关性
磁感应强度	986～1032 mT	在工艺控制范围内，实际输入功变化为 3.52 Wac
电机阻值	1715～1865 mΩ	在工艺控制范围内，实际输入功变化为 0.62 Wac
丝网填充率	24.8%～26.37%	在工艺控制范围内，实际输入功变化为 6.11 Wac
丝网丝径	22～23 μm	在工艺控制范围内，实际输入功变化为 1.04 Wac
丝网厚度	45～54 μm	在工艺控制范围内，理论上对制冷性能无影响

表 2 用于回归分析的因变量和自变量

制冷机编号	密封间隙 x_1 (10^{-2}μm)	磁感应强度 x_2/mT	电机阻值 x_3/mΩ	丝网填充率 x_4/%	丝网丝径 x_5/μm	丝网厚度 x_6/μm	制冷性能 y/Wac
1	1075	1009	1795	26.06	23	54	190
2	1225	1012	1845	26.37	23	54	187
3	1100	1012	1715	26.31	23	49	202
4	975	1012	1840	25.86	22	46	203
5	1050	1009	1785	26.18	23	46	202
6	950	1017	1800	26.18	23	46	207
7	1225	1008	1865	26.2	22	47	185
8	1050	1017	1815	26.2	22	47	195
9	1150	1032	1810	26.2	22	47	219
10	1050	998	1845	26.2	22	47	194
11	900	1003	1825	25.93	22	47	189
12	1050	1014	1830	25.92	22	47	194
13	1125	995	1830	25.55	22	45	185
14	1200	997	1860	25.57	23	48	193
15	1125	1009	1820	24.81	22	46	192
16	1050	1019	1800	24.80	22	46	171
17	1050	986	1840	24.80	22	46	177
18	1025	1019	1810	24.90	23	48	176

生成训练子集时，理论上每次各训练样本被抽中的概率是相同的。经过多次抽取，每个样本被抽中的概率为$(1-\frac{1}{N})^N$。当 $N\to\infty$时，概率约等于 0.368，即每次约 36.8%的样本未被抽取。这些样本称为袋外数据(Out-Of-Bag，OOB)。

2.1.2 随机子空间算法思想[5]

在构建回归决策树时，每个分裂节点的特征子空间都从总特征空间中随机抽取，然后在其中选取最优特征进行分裂，以保证树与树之间的随机性和独立性，从而降低过拟合程度。

定义 1[3] RFR 算法基于一组决策树组合的模型，最终取各决策树的均值作为回归预测的值，其关系式为

$$\bar{h}(x)=\frac{1}{T}\sum_{t=1}^{T}\{h(x,\theta_t)\} \tag{1}$$

式中，θ_t 为服从独立分布的随机变量；x 为自变量；T 表示决策树的棵数；$h(x,\theta_t)$为各决策树基于 x 和 θ_t 的输出。

2.2 泛化误差

定义 2 泛化误差反映了对训练集外数据的预测能力，是判断模型好坏的重要指标。泛化误差越小，分类结果越好；泛化误差越大，分类结果越差。均方泛化误差的定义为 $E_{X,Y}(Y-h(X))^2$。

定理 1[3] 假定对于所有的随机变量 θ_t，回归决策树均是无偏估计，则每一棵决策树的泛化误差 PE_c^* 应满足以下关系：

$$PE_c^*\leqslant\bar{\rho}PE_b^* \tag{2}$$

式中，$\bar{\rho}$ 为残差 $Y-h(X,\theta)$与 $Y-h(X,\theta')$的相关系数，θ 与 θ'相互独立，E_b^* 为 RFR 的泛化误差。

3 随机森林算法的构建

随机森林算法的流程如图 1 所示，主要步骤如下：

(1)利用 Bagging 思想建立样本子集。

(2)每棵决策树选取 m 个特征($m\leqslant M$，M 为总特征数)，进行最后节点分裂并构建单棵决策树。

(3)重复步骤(1)和(2)来构建其余的决策树。

(4)所有决策树组成决策森林。对所有决策树的预测值取平均值，从而得到最终预测结果。

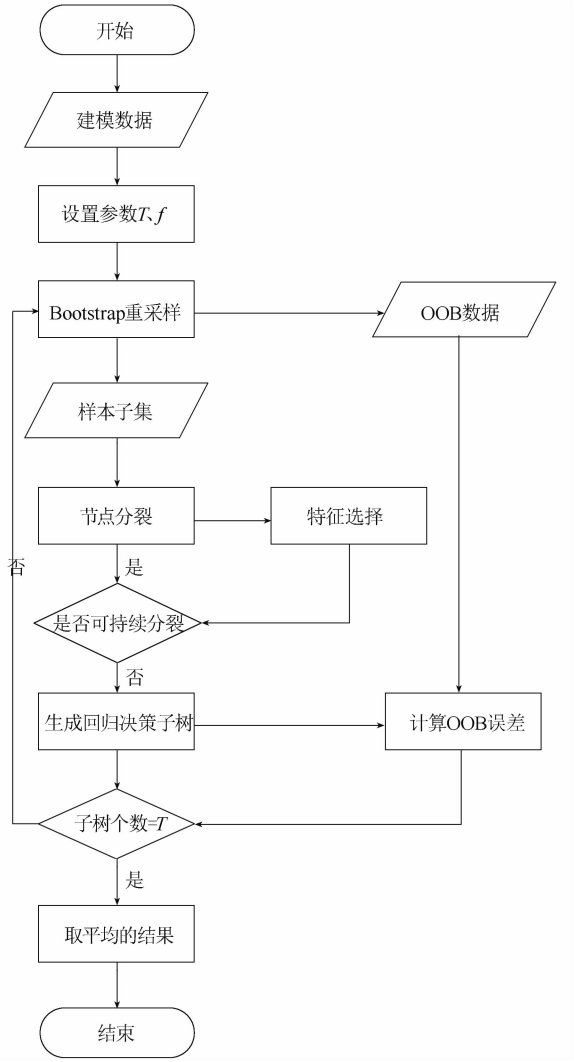

图 1 RFR 的流程图

4 随机森林算法的实验结果分析

采用模型精确度作为评测指标，对随机森林模型的实验结果进行分析。精确度主要包括平均相对误差(Mean Relative Error，MRE)和确定性系数 R^2 两个指标：

$$MRE=\frac{1}{N}\sum_{i=1}^{N}\left|\frac{\hat{Y}_i-Y_i}{Y_i}\right| \tag{3}$$

$$R^2=1-\frac{\sum_{i=1}^{N}(\hat{Y}_i-Y_i)^2}{\sum_{i=1}^{N}(Y_i-\bar{Y}_c)^2} \tag{4}$$

式中，Y_i 为真实的 Y 值，$\hat{Y}_i$ 为预测的 Y 值，$\bar{Y}_c$为真实 Y 值的平均值，N 为样本数。

本文在 Python 环境下进行实验，对其中 14 组数据进行建模分析，并随机选出编号为 2、9、15、18 的 4 组制冷性能数据作为验证集。决策树的棵数分别为 10、15 和 20。此时得到 3 组不同的制冷性能预测模型，然后利用该模型对 4 组验证集进行预测并取平均值。表 3 列出了根据式(3)和式(4)算得的模型精确度。

表 3 RFR 模型的平均性能分析

决策树的数量	MRE	R^2
10	5.70%	0.782
15	5.56%	0.828
20	5.61%	0.804
平均值	5.62%	0.805

从表 3 中可以看出，决策树的数量取 10、15 和 20 时对预测结果的精确度有小幅影响，但差异不大：平均相对误差分别为 5.70%、5.56%和 5.61%；确定性系数分别为 0.782、0.828 和 0.804。对三组不同数量的决策树模型的 MRE 和 R^2 取平均值。它们分别为 5.62%和 0.805，说明预测模型具有较好的预测精确度。

决策树的棵树不同时，自变量的特征重要度如图 2、图 3 和图 4 所示。可以看出，改变决策树的棵数对特征重要度的排序没有影响。重要程度从高到低依次为丝网填充率 x_4、磁感应强度 x_2、丝网厚度 x_6、电机阻值 x_3、密封间隙 x_1 和 丝网丝径 x_5。但是当改变决策树的棵数时，各特征重要度的值会有小幅变化，这与随机森林模型的目标函数有关。通过与表 1 中的结果进行对比发现，丝网填充率 x_4 对性能的影响特征重要度最高，磁感应强度 x_2 次之，其他 4 个自变量的特征重要度较低，与制冷性能的实验研究结果相符。当决策树增加到 20 棵时，丝网厚度 x_6 的特征重要度反而提升了，与实际情况不符。根据制冷机一维热力学模型，丝网的两大重要参数分别为填充率和丝径，理论上丝网厚度不会对制冷性能产生显著影响。此外，当决策树增加到 20 棵时，确定性系数 R^2 反而下降了，说明此时模型存在过

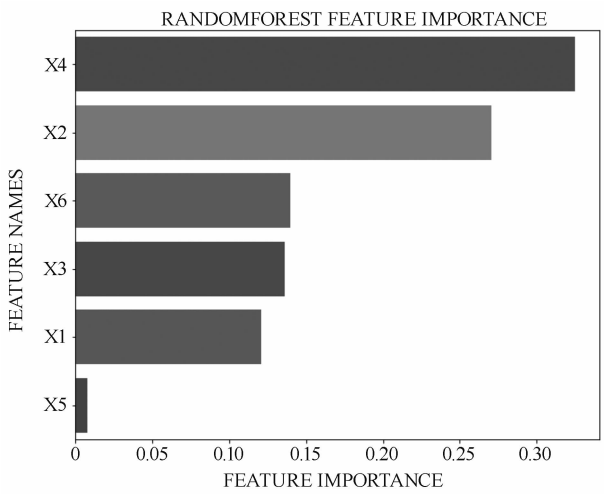

图 2　10 棵决策树时的特征重要度

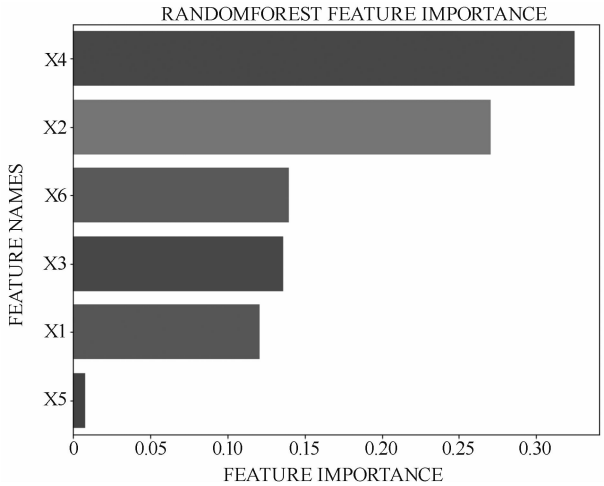

图 3　15 棵决策树时的特征重要度

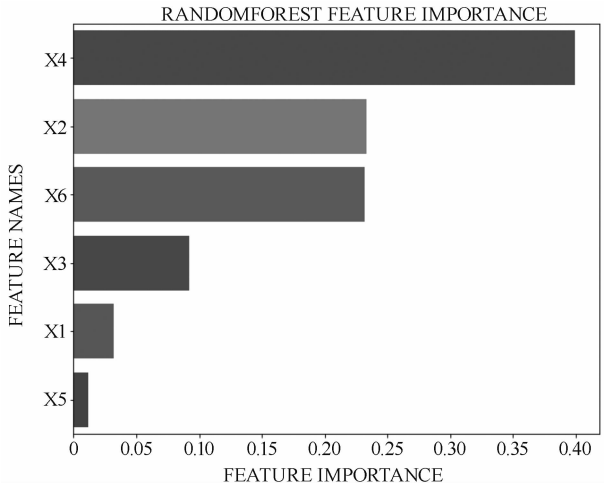

图 4　20 棵决策树时的特征重要度

拟合现象。

5　结论

本文利用基于机器学习语言的 RFR 对制冷性能与各自变量进行了回归预测分析。制冷性能的 RFR 具有较高的精确度，预测的平均相对误差为 5.62%，平均确定性系数为 0.805，且自变量的特征重要度与实际的实验结果基本相符。基于机器学习语言对数据进行回归分析研究，为提高星载脉管制冷机的一致性提供了新的思路。该工作可与传统的制冷机性能机理研究进行相互验证和补充。

此外，由 RFR 得出的前两大特征重要度变量分别为丝网填充率和磁感应强度，与实验结果相符。但其他变量回归分析的特征重要度与实验结果略有差异。这跟样本数量偏少导致的过拟合有关。随着以后制冷机样本数据的积累，还需进一步优化回归模型，以得到更优的预测结果。

参考文献

[1] Wilson K B, Fralick C C, Gedeon D R, et al. Sunpower's CPT60 pulse tube cryocooler [J]. *Cryocoolers*, 2007, **14**: 123–132.

[2] Liu S S, Wu Y N, Zhang H, et al. Investigation on the inertance tube of pulse tube refrigerator operating at high temperature [J]. *Energy*, 2017, **123**(3): 378–385.

[3] Breiman L, Cutler A. Random forests [J]. *Machine Learning*, 2001, **45**(1): 5–32.

[4] Breiman L. Bagging predictors [J]. *Machine Learning*, 1996, **24**(2): 123–140.

[5] Ho T K. The random subspace method for constructing decision forests [J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1998, **20**(8): 832–844.