DOI：10.19666/j.rlfd.202403044

考虑激励特性的汽轮机做功模型辨识数据优选方法

郝晓光[1]，王　辉[1]，金　飞[1]，王腾辉[2]
（1.国网河北能源技术服务有限公司，河北　石家庄　050021；
2.华北电力大学控制与计算机工程学院，北京　102206）

［摘　　要］针对历史运行数据中难以选择合适样本辨识汽轮机做功模型问题，提出一种考虑激励特性的辨识数据优选方法。首先，采用费歇尔（Fisher）信息矩阵条件数提取历史运行数据的激励特性，与数据的趋势特性和参数间相关性共同构成特征变量集。其次，以特征变量作为输入，基于标准汽轮机做功模型生成的标识结果作为输出，采用随机森林分类算法生成辨识数据分类规则模型，实现辨识数据的在线选择。最后，对模型分类结果的准确性与所选数据的辨识效果进行验证。结果表明，分类规则模型的准确度为 97.561%，可准确选出历史运行数据中含有充分激励的样本段，其汽轮机做功模型辨识结果与标准模型具有较高的一致性。

［关　键　词］辨识数据；数据激励特性；汽轮机；随机森林

［引用本文格式］郝晓光，王辉，金飞，等. 考虑激励特性的汽轮机做功模型辨识数据优选方法[J]. 热力发电，2024，53(11)：130-138.　HAO Xiaoguang, WANG Hui, JIN Fei, et al. Selection method for identification data of steam turbine work model considering excitation characteristics[J]. Thermal Power Generation, 2024, 53(11): 130-138.

Selection method for identification data of steam turbine work model considering excitation characteristics

HAO Xiaoguang[1], WANG Hui[1], JIN Fei[1], WANG Tenghui[2]
(1.State Grid Hebei Energy Technology Service Co., Ltd., Shijiazhuang 050021, China;
2.School of Control and Computer Engineering, North China Electric Power University, Beijing 102206, China)

Abstract: A method of identifying data by considering the excitation characteristics is proposed to solve the problem that it is difficult to select suitable samples from the historical operation data to identify the turbine work model. Firstly, Fisher's information matrix condition number is applied to extract the excitation characteristics of the historical operating data, which together with the trend characteristics and the correlation between parameters constitute the set of feature variables. Secondly, by using the feature variables as inputs and the identification results generated based on the standard turbine work model as outputs, the Random Forest classification algorithm is used to generate a classification rule model for the identification data to realize the online selection of identification data. Finally, the accuracy of the model classification results and the identification effect of the selected data are verified. The result proves that the accuracy of the classification rule model is 97.561%, which can accurately select the sample segments containing sufficient incentives in the historical operation data, and the identification results of the turbine work model are in high consistency with that of the the standard model.
Key words: data identification; data excitation properties; steam turbine; random forest

在大量新能源电力接入电网的背景下，电力行业的能源结构不断发生调整，但火力发电仍占据主要部分[1-2]。汽轮机做功系统作为电源侧重要的调频系统，其辨识研究对掌握机组一次调频能力具有重

收 稿 日 期：2024-03-26　网络首发日期：2024-05-20
基 金 项 目：国网河北省电力有限公司科技项目（TSS2023-03）
Supported by：Technology Project of State Grid Hebei Electric Power Co. Ltd. (TSS2023-03)
第一作者简介：郝晓光（1980），男，本科，正高级工程师，主要研究方向为综合能源控制技术的研究及应用，dyy_haoxg@he.sgcc.com.cn。

要意义[3]。目前，系统辨识建模过程往往需要高质量的输入输出数据[4]，然而，在生产过程中加入长时间激励信号获得试验数据的方法不利于正常的机组运行。相比辨识试验得到的数据，基于海量历史运行数据的辨识方法具有对机组运行影响小、经济成本低等优点，但是，在机组运行较平稳，负荷变动不大时，忽略了输入输出辨识数据的激励特性，辨识结果的输出和实际运行输出相差较大，因此，亟需开展相关研究，筛选历史运行数据中含有充分激励特性的辨识数据。

针对获取模型辨识数据的研究主要集中在模型特性试验与基于历史运行数据两方面。文献[5]直接采用现场阶跃响应试验，建立机组全工况数学模型，反映协调控制系统被控对象的特性，但是现场试验步骤复杂，耗时长，且影响正常生产运行。文献[6]提出对机组历史运行数据进行分析和挖掘，解决了现场试验繁琐、耗时长的问题，但该方法未充分考虑现场数据的可辨识性。文献[7]基于辨识理论，分析了火电机组现场数据用于辨识的基本条件，当机组负荷波动较大时，含有噪声的运行数据能够满足充分激励，且火电机组的对象大多具有滞后性，能够保证参数辨识的准确性。文献[8]认为历史数据变化具有较大的时间常数，满足可辨识条件，并利用历史数据辨识了除氧器水位模型。文献[9]通过对 SCR 烟气脱硝系统模型构造费歇尔（Fisher）信息矩阵，以矩阵最大、最小特征值之比作为评价历史数据是否可辨识的指标，但该方法仍需要人为设置筛选阈值。

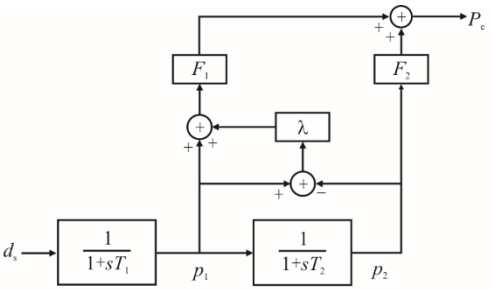
考虑到当前辨识数据筛选工作多为人工选取，且数据变量间的相关性、数据变化趋势与辨识数据激励是否充分等诸多因素对辨识结果均有影响，同时对人工选取数据也造成了很大的困难。随着机器学习与智能分类算法的深入发展，文献[10]提出一种对故障信号提出关键特征分量重构信号的方法，通过使用智能分类算法构建分类模型，完成故障分类的辨识。文献[11]采用随机森林智能算法，建立智能分类模型，将划分粉煤灰活性问题转化为快速判断是否能够作为辅助凝胶材料的分类问题。文献[12]从电压数据中提取特征构造样本，对随机森林分类器进行样本训练，实现根据电压数据的特征量判断模块工作状态。因此，可以考虑将随机森林分类算法用于辨识数据选择，从实际运行历史数据中提取能够影响辨识结果的数据特征，重构历史数据，建立分类规则模型，将某段数据是否适宜辨识转化成分类问题，避免人工选取的盲目性、复杂性等问题。

鉴于此，在上述方法的基础上，本文提出一种考虑激励特性的辨识数据优选方法。首先，从汽轮机历史运行数据提取影响辨识结果的数据趋势性特征、变量相关性特征和数据激励性特征，从而构造成辨识数据特征变量集，在包含原始机组动态信息的同时纵向降低了数据维度；其次，为了准确生成辨识数据的标识，对辨识响应输出拟合度设置阈值，当拟合度大于阈值时，认为该段数据适用于辨识，反之则标识为该段数据不适用于辨识，转化为二分类问题；最后，采用随机森林分类算法，基于构建的特征变量和标识结果生成分类规则模型，实现辨识数据的选择。

1 汽轮机做功模型及辨识机理

1.1 汽轮机做功模型

作为典型火电机组一次调频模型的主要组成部分，汽轮机做功模型的调频动作响应能力与汽轮机做功密切相关[13]。图 1 为汽轮机做功模型结构，该模型中，蒸汽流经高压缸、再热器、中低压合缸，在汽轮机膨胀做功输出机械功率。



d_s—主蒸汽流量，t/h；p_1—调节阀后压力，MPa；p_2—再热器压力，MPa；P_e—机组负荷，MW；T_1—高压缸前汽室容积时间常数，s；T_2—再热器容积时间常数，s；F_1—高压缸功率分配系数；F_2—中低压合缸功率分配系数；λ—高压缸功率自然过调系数。

图 1 汽轮机做功模型结构

Fig.1 Structural diagram of the turbine work model

该模型以主蒸汽流量 d_s 为输入，以机组负荷 P_e 为输出[14]。传递函数 $G(s)$可以整理为：

$$G(s)=\frac{P_e(s)}{d_s(s)}=\frac{b_1 s+b_0}{a_2 s^2+a_1 s+a_0} \tag{1}$$

其中：

$$a_0=1 \tag{2}$$

$$a_1=T_1+T_2 \tag{3}$$

$$a_2=T_1T_2 \tag{4}$$

$$b_0=F_1+F_2=1 \tag{5}$$

$$b_1 = T_2 F_1 (1+\lambda) \tag{6}$$

由式(2)—式(6)可知，汽轮机做功模型传递函数 $G(s)$中，a_0 和 b_0 为已知参数，需要辨识的参数分别为 a_1、a_2 和 b_1，共 3 组。

1.2 辨识数据优选方法框架

基于随机森林分类算法的汽轮机做功模型辨识数据优选方法框架如图 2 所示。该框架主要分为 2 个模块：生成分类规则模型模块与辨识数据筛选模块。在生成分类规则模型模块中，主要考虑利用历史运行数据作为辨识数据：首先，利用 Fisher 信息矩阵提取历史数据的激励特性，与数据趋势特性、参数间相关特性共同构成特征变量集；然后，基于标准汽轮机做功模型的响应输出计算辨识数据响应输出拟合度，对应生成“1”或“0”的标识结果；最后，基于随机森林分类算法，将特征变量集作为模型输入，标识结果作为模型输出，生成分类规则模型，实现辨识数据的快速筛选。在辨识数据筛选模块中，将待分类的特征变量集输入到训练好的分类规则模型中进行分类预测。如果分类结果为“0”，则可以判定为该段历史数据不适用于汽轮机做功模型进行参数辨识；若分类为“1”，则判定该特征集合对应的历史运行数据适用于模型参数辨识。

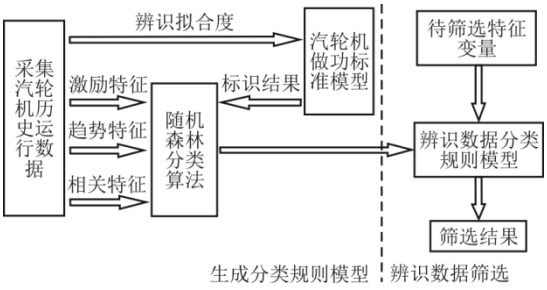


图 2 辨识数据优选方法框架

Fig.2 Block-chart of the identification data selection method

2 辨识特征选择

2.1 基于 Fisher 信息矩阵条件数的激励特性提取

考虑到历史运行数据中存在测量噪声，并且参数向量的维度小于辨识的数据量，可认为历史数据中存在含有持续充分激励的辨识数据段。鉴于此，考虑使用 Fisher 信息矩阵条件数作为判别用于辨识数据是否含有持续且充分的激励[15-16]。

假设辨识过程输入和输出存在如下关系：

$$\overline{y_t} = f(u_t, \theta) \tag{7}$$

式中：f 为任意函数；$\boldsymbol{\theta}=(\bar{\theta}_1, \bar{\theta}_2, \cdots, \bar{\theta}_r)$ 为 r 维的参数向量。则 f 对应的雅可比矩阵 $\boldsymbol{J}$ 为：

$$\begin{aligned}\boldsymbol{J} &= \frac{\partial f(u_t,\theta)}{n}\Big|_{u=u_t} \\ &= \left[\frac{\partial f(u_t,\bar{\theta})}{\partial \theta_1}, \frac{\partial f(u_t,\bar{\theta})}{\partial \theta_2}, \cdots, \frac{\partial f(u_t,\bar{\theta})}{\partial \theta r}\right]\end{aligned} \tag{8}$$

根据上式求得输入变量的雅可比矩阵，进而构造成矩阵 $\boldsymbol{M}$：

$$\boldsymbol{M} = \begin{pmatrix} \dfrac{\partial f(u_1,\bar{\theta})}{\partial \theta_1} & \cdots & \dfrac{\partial f(u_t,\bar{\theta})}{\partial \theta_t} \\ \vdots & \ddots & \vdots \\ \dfrac{\partial f(u_t,\bar{\theta})}{\partial \theta_1} & \cdots & \dfrac{\partial f(u_t,\bar{\theta})}{\partial \theta_t} \end{pmatrix} \tag{9}$$

$$\boldsymbol{M}\bar{\theta} = \bar{y} \tag{10}$$

$$\boldsymbol{M}^{\mathrm{T}}\boldsymbol{M}\bar{\theta} = \boldsymbol{M}^{\mathrm{T}}\bar{y} \tag{11}$$

$$\boldsymbol{F}\bar{\theta} = \boldsymbol{M}^{\mathrm{T}}\bar{y} \tag{12}$$

$$\eta = \frac{\max(|\lambda(\boldsymbol{F})|)}{\min(|\lambda(\boldsymbol{F})|)} \tag{13}$$

式中：$\lambda(\boldsymbol{F})$为 Fisher 信息矩阵 $\boldsymbol{F}$ 的特征值；η为 Fisher 信息矩阵的条件数。当矩阵最大特征值 $\max(|\lambda(\boldsymbol{F})|)$与最小特征值 $\min(|\lambda(\boldsymbol{F})|)$的比值越大时，辨识系统对数据中的扰动更敏感，从而导致辨识参数的准确性降低，即认为该段数据缺乏持续激励而缺少系统的动态信息，不适用于辨识；相反，当条件数 η越小，辨识系统对噪声的鲁棒性更高，参数辨识的收敛性与稳定性越好，可认为该段辨识数据满足充分持续激励条件。

2.2 辨识数据优选过程特征变量集构建

构建清晰、准确、简明的数据特征对表征汽轮机做功模型动态信息十分重要。数据趋势描述了汽轮机做功随时间或其他相关因素的动态变化方向或模式，相关系数度量了系统输入输出变量间的相关程度，因此本文考虑通过数据激励特性和数据趋势特性、参数间相关性对特征变量集进行描述。表 1 给出了 3 类特征的 12 种特征变量。

趋势分析作为时间序列分析中的重要组成部分，能够寻找数据长期变化过程中可能存在的规律。因此，采用均值、标准差、变异系数、众数带等来表征数据趋势特征。

$$c_i = \frac{\sigma_i}{\mu_i} \tag{14}$$

$$M_i = \max_i - \min_i \tag{15}$$

$$\Delta M_i = \frac{n_{M_i}}{n_i} \tag{16}$$

式中：c_i 为变异系数；σ_i 为标准差；μ_i 为均差；

M_i 为极差；$\max_i$ 和 $\min_i$ 分别为数据序列中的最大值和最小值；ΔM_i 为众数带；n_{Mi} 为数据序列中位于区间$[\mu_i-\varepsilon M_i, \mu_i+\varepsilon M_i]$的个数；$\varepsilon$通常取 0.5。

表 1 特征变量
Tab.1 Feature variables

特征	符号	名称
数据趋势特征	μ_1	主蒸汽流量的均值
	μ_2	机组负荷的均值
	σ_1	主蒸汽流量的标准差
	σ_2	机组负荷的标准差
	c_1	主蒸汽流量的变异系数
	c_2	机组负荷的变异系数
	M_1	主蒸汽流量的极差
	M_2	机组负荷的极差
	ΔM_1	主蒸汽流量的众数带
	ΔM_2	机组负荷的众数带
参数间相关特征	r	主蒸汽流量与机组负荷的相关系数
数据激励特征	η	主蒸汽流量与机组负荷的 Fisher 信息矩阵条件数

采用输入输出变量间的相关系数 r 作为参数间相关特征，其相关系数为：

$$r=\frac{\mathrm{cov}(d_s,P_e)}{\sqrt{\mathrm{var}(d_s)\,\mathrm{var}(P_e)}} \tag{17}$$

式中：$\mathrm{cov}(d_s,P_e)$为输入变量和输出变量的协方差；var(.)为对应变量序列的方差。

根据式(7)—式(13)计算数据的 Fisher 信息矩阵条件数，以此表征辨识数据的激励特征。

通过计算以上数据特征，将其组合成历史数据的特征变量集样本，不仅包含了汽轮机做功模型的动态信息和激励特性，还降低了数据的纵向维度。将采集到的汽轮机做功模型的历史运行数据按照数据长度 n 进行顺序切分，获得 m 份数据段，对每段数据分别计算表 1 中的 12 个特征变量，第 k 个特征向量 $\boldsymbol{x}$ 如式(18)表示。然后将 m 个向量 $\boldsymbol{x}$ 合并为 1 个 $m\times12$ 维的特征变量矩阵 $\boldsymbol{X}$。

$$\boldsymbol{x}_k=(\mu_{1k},\mu_{2k},\sigma_{1k},\sigma_{2k},\cdots,r_k,\eta_k) \tag{18}$$

2.3 基于标准模型的数据标识

根据汽轮机做功标准模型，对切分后的各段历史数据生成标识。将每段数据样本进行汽轮机做功模型传递函数 $G(s)$的参数辨识，对辨识得到的模型和标准模型输入单位阶跃信号，依据辨识模型响应输出拟合度 Fits，对应特征变量集生成标识结果。

$$\mathrm{Fits}=1-\frac{\sqrt{\sum_{z=1}^{n}(P_z-P_z)^2}}{\sqrt{\sum_{z=1}^{n}(P_z-\overline{P_z})^2}}\times100 \tag{19}$$

针对主要辨识对象汽轮机做功模型：P_z 是标准模型响应输出，P_z 是辨识模型响应输出；$\overline{P_z}$ 是标准模型响应输出的均值。若某段辨识数据的输出拟合度大于设置阈值，标识为“1”，即代表该段数据适用于辨识；反之，标识为“0”，即该段数据不适用于辨识。可获得如下标识结果集：

$$\boldsymbol{Y}=\{(y_1,y_2,\cdots,y_k),y_k\in\{0,1\},k=1,2,\cdots,m\} \tag{20}$$

根据以上步骤，基于汽轮机做功模型历史运行数据，将构造的特征变量集与生成的标识结果组合为以下数据集合：

$$T=\{(x_k,y_k),x_k\in\boldsymbol{X},y_k\in\boldsymbol{Y},k=1,2,\cdots,m\} \tag{21}$$

3 基于随机森林分类的辨识数据选择

3.1 随机森林分类算法原理

将筛选辨识数据问题转化为基于随机森林分类的历史运行数据是否适宜辨识的二分类问题。随机森林分类算法的本质是一种基于决策树的集成学习方法，其基本原理是将多个基础学习器组合，最终通过投票表决的方式决定最终的分类预测结果[17]。随机森林分类算法示意如图 3 所示，该算法主要包括构建训练样本集合、构建决策树和投票最终分类结果 3 个步骤。

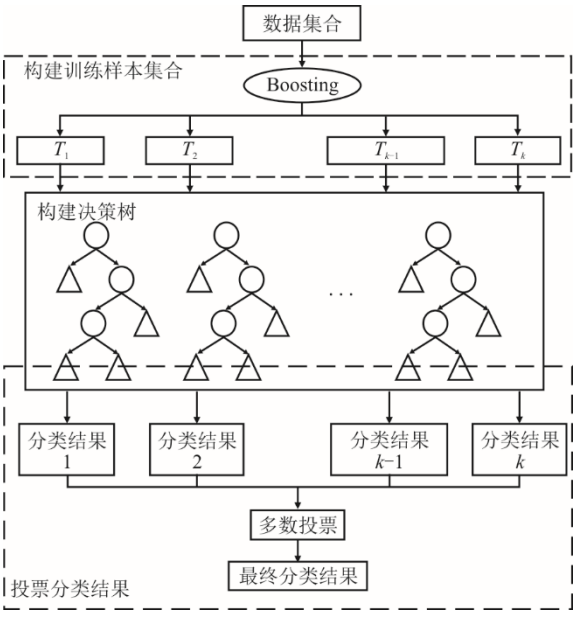


图 3 随机森林分类算法示意
Fig.3 Schematic diagram of the random forest classification algorithm

首先，在训练单个基础学习器决策树时引入随机属性，采用 Boosting 的方法构建训练样本集合，根据式(18)可知，原样本集合的容量为 m，其中的

元素为 12 维的样本向量，有放回地对样本集合抽样 m 次，可以得到 1 个样本数量为 m 的新训练集合，重复抽样 K 次，获得 K 组训练集合样本。

其次，根据每组训练样本集合构成决策树，每棵决策树代表根据随机选取的特征集合训练获得的分类规则，根据式(22)计算出的最小 Gini 系数决定决策树分裂节点[18]。

$$\mathrm{Gini}(T)=1-\sum_{i=1}^{12}p_i^2 \tag{22}$$

式中：p_i 为第 i(i=1,2,⋯,12)个数据特征在样本集合 T 中出现的概率。

最后，依据多数投票法对 K 棵决策树的分类结果进行选择，当 K 棵决策树的分类标识结果中分类标识“1”的结果超过半数时，即表示该段数据特征对应的历史数据适用于辨识，反之，则不适用于辨识。

3.2 优选辨识数据分类规则模型

将 2.2 节中提取的特征变量集作为分类规则模型的输入，基于 3.1 节的算法原理，可以将辨识数据的筛选转化为分类问题。构建基于随机森林分类算法的辨识数据优选分类规则模型的具体步骤如图 4 所示，分为以下几步。

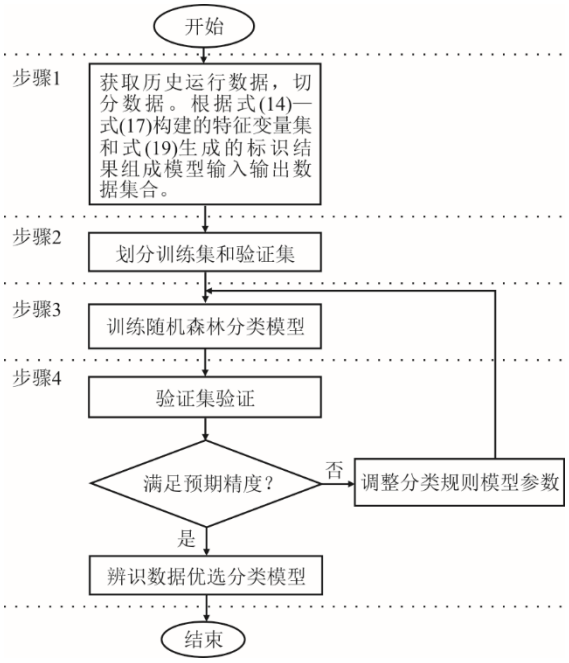


图 4 基于随机森林分类的辨识优选模型构造方法
Fig.4 Identify data selection methods for random forest classification algorithm

步骤 1　获取汽轮机历史运行数据，设置辨识数据段长度 n，对历史运行数据进行切分，获得 m 份长度为 n 的历史数据段。分别提取每段数据的辨识数据特征，构建 1 个由 12 维向量组成的特征集合。根据辨识模型响应输出拟合度，设置阈值为 70，比较生成标识结果，标识数据适用于辨识与不适用于辨识，分别用“1”和“0”标识。从而获取模型的输入输出数据集。

步骤 2　将步骤 1 中的原始数据集划分训练集 D_{train} 与验证集 D_{test}。将训练集中的样本按照 Boosting 方法有放回地抽取样本得到占 D_{train} 样本数量 90%的实际训练数据集和 10%的袋外验证数据。

步骤 3　设置分类模型初始化参数，基础学习器数量 N，子树最大特征数 M_{f}。由于辨识数据筛选是标准的二分类问题，因此依据式(22)计算系数，选择当前分支值最小的特征作为分裂点。以袋外分类精度（Accuracy）作为模型的评价指标[19]，训练分类规则模型。

步骤 4　将验证集 D_{test} 导入分类规则模型进行验证，对比模型分类结果与实际分类标识，若验证精度较低，则返回步骤 3，调整模型参数，重新训练分类规则模型；当验证精度满足预期时，固定最终模型。

4 仿真分析

4.1 辨识数据的 Fisher 信息矩阵条件数分析

某超临界 350 MW 燃煤机组，汽轮机为哈尔滨汽轮机厂有限责任公司生产的一次中间再热、三缸两排汽、再热凝汽式汽轮机。采集不同工况下汽轮机做功模型中主蒸汽流量与机组负荷的历史运行数据（表 2），采样时间为 1 s，每份样本数据量为 10 000。根据式(7)—式(13)分别计算在 48%、57%和 82%负荷工况下汽轮机历史数据的 Fisher 信息矩阵条件数（图 5），评估历史数据的激励特性，计算窗口数据长度设置为 500 个采样点。

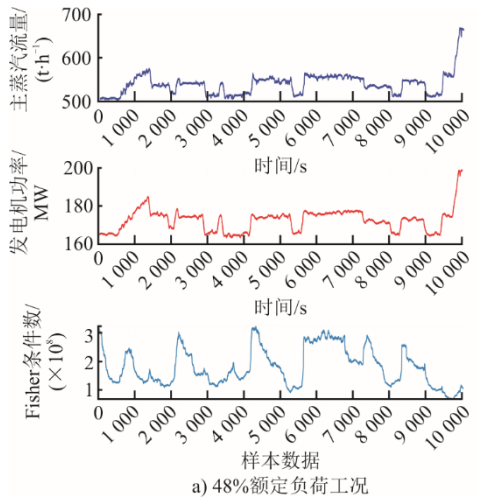
由图 5 可以看出：在图 5a)历史运行数据中的 5 700~6 700 数据段、图 5b)中的 3 100~5 000 数据段与图 5c)中的 8 000~9 000 数据段，火电机组运行较平稳，主蒸汽流量作为输入信号的幅值较小，发电机功率输出变化平稳，含有动态信息较少；图 5a)中 0~1 500 数据段、图 5b)中 1 500~3 100、6 100~6 300、7 500~7 800 数据段和图 5c)中 1 000~3 000 数据段中，主蒸汽流量有明显大幅调节动作，输出随之升降，符合汽轮机实际做功规律，且输入信号有充分的激励。

由图 5 中不同工况下的 Fisher 条件数曲线可

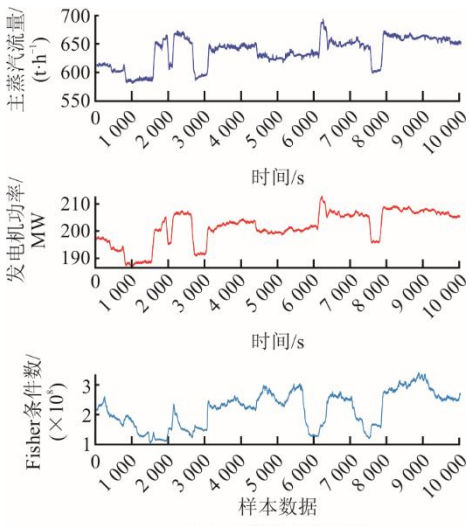
知，欠激励数据段的条件数明显大于充分激励数据段。特别是 48%负荷工况下第 4 200 时刻、57%负荷工况下第 3 100 时刻和 82%负荷工况下第 7 900 时刻，滑动窗口进入欠激励数据段，其相应的条件数曲线迅速上升。从 48%负荷工况下第 10 时刻、57%负荷工况下第 5 600 时刻开始，计算条件数的窗口逐渐开始包含充分激励数据段数据，其对应的条件数由局部最大值逐渐减少。当滑动窗口内的数据全部来自充分激励段，48%负荷工况下第 500 时刻、57%负荷工况下第 6 100 时刻和 82%负荷工况下第 2 650 时刻对应条件数曲线值达到局部极小值。

表 2 样本数据
Tab.2 The sample data

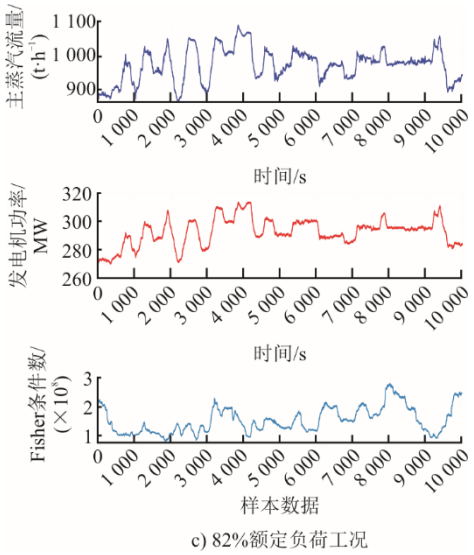
编号	数据起止时间	样本工况	样本数据量
a	2023 年 8 月 2 日 06:19:57—09:06:37	48%	10 000
b	2023 年 7 月 14 日 07:26:37—10:13:17	57%	10 000
c	2023 年 7 月 31 日 15:46:57—18:33:37	82%	10 000



a) 48%额定负荷工况



b) 57%额定负荷工况



c) 82%额定负荷工况

图 5 辨识数据 Fisher 信息矩阵条件数曲线
Fig.5 Fisher information matrix conditional number curves for identified data

根据不同工况下的条件数分析可知，充分激励段数据对应的 Fisher 信息矩阵条件数比欠激励数据段对应的条件数相对较小，但是由于汽轮机做功的复杂性，无法通过设置明确的阈值来划分辨识数据，因此可以考虑将 Fisher 信息矩阵条件数作为表征辨识数据激励是否充分的特征数据，用于生成辨识数据分类规则模型。

同时，由于随机森林具有计算每个特征变量重要性的特点[20]，因此对表 1 中的 12 个特征变量进行重要性分数的计算，结果如图 6 所示。由图 6 可知，Fisher 信息矩阵条件数η的重要性分值为 2.219 2，远高于其他数据特征，进一步证明了提取激励特征对辨识数据筛选具有一定的参考意义。

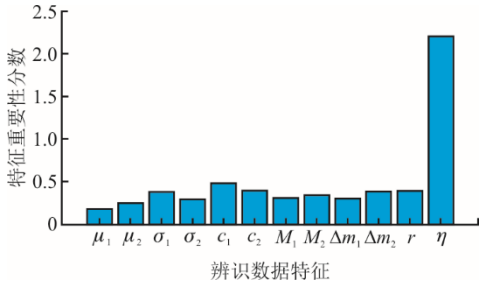


图 6 数据特征重要性
Fig.6 Significant scores for data characterization

4.2 分类规则模型结果对比

对上述机组 2023 年 7 月 13 日 12:00:00 至 8 月 4 日 23:09:57 间的运行历史数据（图 7）进行采样，抽取共 1 005 000 组样本数据，采样时间 1 s。将数

据顺序划分为每段 5 000 个数据点，预处理后得到 199 组原数据集，取其中 158 组样本数据用于生成分类规则模型，41 组样本数据作为验证集合，用于检验模型分类效果。

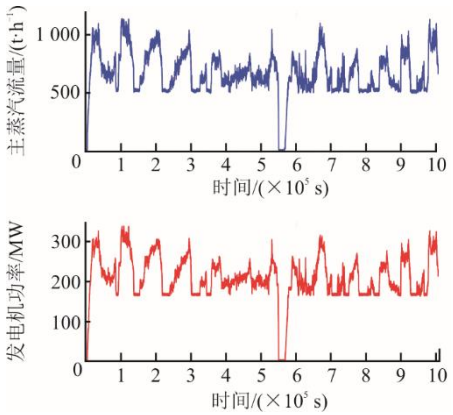


图 7 某机组历史运行数据
Fig.7 Historical operation data of a unit

数据适宜辨识分类为“1”，不适宜辨识分类为“0”，对于这种二分类问题采用召回率（Recall）、准确率（Precision）和 F 值（F-score）来评价预测分类结果的有效性[21]。混淆矩阵见表 3。

表 3 混淆矩阵
Tab.3 Confusion matrix

分类标识	1	0
实际 1	T_P	F_N
实际 0	F_P	T_N

$$\text{Accuracy}=\frac{T_P+T_N}{T_P+T_N+F_P+F_N} \tag{23}$$

式中：T_P 和 T_N 为分类标识“1”和“0”预测正确的样本数量；F_P 和 F_N 为分类标识“1”和“0”预测错误的样本数量。

$$\text{Recall}=\frac{T_P}{T_P+F_N} \tag{24}$$

$$\text{Precision}=\frac{T_P}{T_P+F_P} \tag{25}$$

$$F\text{-score}=\frac{(1+\gamma^2)\text{Recall}\times\text{Precision}}{\gamma^2\times\text{Precision}+\text{Recall}},\gamma\in(0,1] \tag{26}$$

将训练集与验证集分别进行模型的训练与验证。设置模型基础学习器数量 N=200，最大特征树 M_f=10，模型学习训练迭代曲线如图 8 所示。当基础学习器数量 N=17 时，其分类精确度得分最高为 0.905。

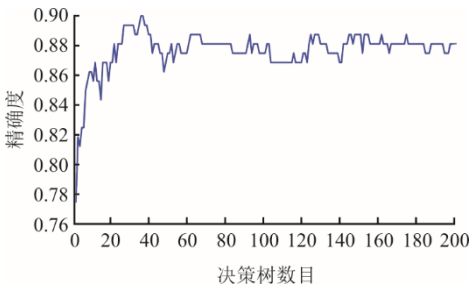


图 8 模型精确度曲线
Fig.8 Model Accuracy curve

分类规则模型在验证集上的验证结果如图 9 所示。通过实际验证样本分类标识与预测验证样本分类的重合情况来表征模型的准确性，其中实际为适宜辨识的数据段有 19 组，非适宜辨识的数据段有 22 组，仅在适宜辨识数据段有 1 个验证集样本分类错误。

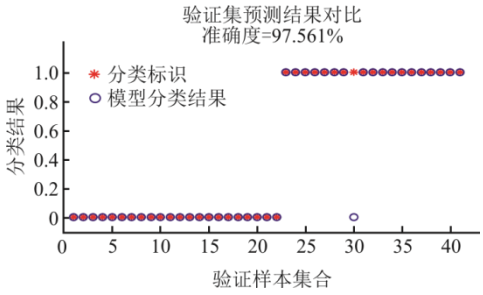


图 9 验证集数据分类结果
Fig.9 Data classification results for verification set

为了进一步验证所选分类器进行辨识数据筛选的有效性与准确性，更好地评判该分类规则模型的优劣程度，将随机森林（random forest，RF）算法生成的分类规则模型与支持向量机（supportive vector machine，SVM）算法、决策树（decision tree，DT）和 K-近邻（K-nearest neighbor，KNN）算法生成的模型对汽轮机历史运行数据进行辨识数据筛选的分类结果进行比较，结果见表 4。

表 4 各分类模型综合评价指标对比
Tab.4 Comprehensive performance comparison of classification models

评价指标	适宜辨识数据（标识“1”）			
	DT	SVM	KNN	RF
Recall	0.769 23	0.769 23	0.428 57	1.000 00
Precision	0.769 23	0.833 00	0.230 76	0.947 37
F-score	0.769 23	0.799 84	0.299 99	0.972 97
分类模型运行时间/s	0.641	0.215	0.822	0.156

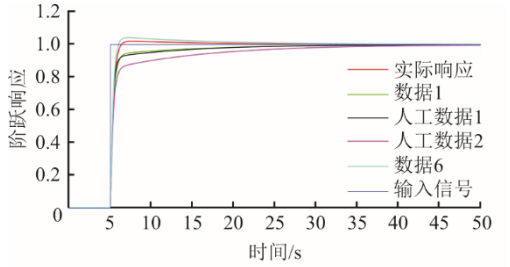
从表 4 分析可知，RF 算法生成的分类规则模型的 Recall、Precision 和 F-score 值以及整体性能均优于其他分类模型，在运行时间上也优于其他分类模型，从而验证了基于随机森林分类算法的分类规则模型

筛选汽轮机做功模型辨识数据的优越性与适用性。

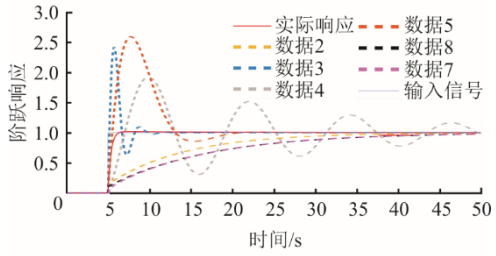
4.3 辨识结果分析

为进一步验证所选出数据的可辨识性，选取采样时间为 1 s、40 000 个连续采样点的历史运行数据，按照 5 000 数据长度顺序划分为 8 组样本。将构建好的辨识数据特征集合输入训练好的分类规则模型，进行辨识数据优选。结果显示，2 组数据样本划分为适宜辨识的数据，6 组数据样本划分为不适宜辨识的数据。

同时，根据人工挑选辨识数据经验选取 2 组数据样本，对模型的数据样本和人工数据样本进行参数辨识，并以单位阶跃信号作为输入信号，输出各组辨识结果的响应曲线，与汽轮机做功标准模型的响应曲线进行对比，结果如图 10 所示。



a) 标识“1”的数据、人工挑选数据辨识响应曲线



b) 标识“0”的数据辨识响应曲线

图 10 参数辨识结果的响应曲线对比

Fig.10 Comparison of response curves for parameter identification using different data samples

由图 10 可知：人工挑选的 2 组数据样本的拟合度分别为 89.99%和 81.50%；分类结果为适宜辨识的样本数据的辨识模型单位阶跃响应曲线与汽轮机做功标准模型响应曲线有较好的一致性，且拟合度分别为 96.41%和 90.8%，同时优于人工挑选数据的辨识结果，不仅能够较好地反应实际汽轮机动作响应，还能替代人工挑选出较优的辨识数据；相比下，分类结果为不适宜辨识的数据样本的辨识模型的单位阶跃响应曲线与标准模型响应曲线相差较大，数据样本 3、4、5 均出现超调响应，且该 6 组数据样本的拟合度均低于 30%。进一步验证了本文所提辨识数据优选方法的有效性。

5 结　论

1）基于历史运行数据在汽轮机做功模型系统辨识时的激励特性，提出了使用 Fisher 信息矩阵条件数提取数据激励特征，并与数据趋势特征、参数间相关特征构造数据特征集合，实现从历史运行数据中提取含有激励特征的数据特征以构造分类规则模型的样本集。

2）采用 RF 分类算法，提出基于 RF 分类算法的优选辨识数据分类规则模型。通过构建训练样本集合、构建决策树和投票分类结果 3 个关键步骤生成模型，通过数据特征量判断是否适用于系统辨识，实现从历史运行数据中快速筛选出充分激励的辨识数据，降低人为挑选数据的盲目性。

3）通过所提出的辨识数据优选方法从分类准确性、数据可辨识 2 个方面进行验证。该方法整体分类的准确性为 97.561%，适宜辨识数据的 Recall、Precision、F-score 分别为 1.000 00、0.947 37、0.972 97，100%的不适宜辨识数据被预测出来，仅在适宜辨识数据段有 1 个验证集样本分类错误；且分类结果为适宜辨识的样本数据的辨识结果能够较好地反应汽轮机做功模型的实际动作情况。仿真结果表明，本文提出的方法实现了在大量历史运行数据中较为便捷且准确地优选出辨识效果较好的数据段，大大提高了生产效率，节约人工成本，可为系统辨识中辨识数据筛选提供一定的参考。

[参 考 文 献]

[1] 王月明, 牟春华, 姚明宇, 等. 二次再热技术发展与应用现状[J]. 热力发电, 2017, 46(8): 1-10.
WANG Yueming, MOU Chunhua, YAO Mingyu, et al. Review of the development and application of double-reheat power generation technology[J]. Thermal Power Generation, 2017, 46(8): 1-10.

[2] 蒋敏华, 黄斌. 燃煤发电技术发展展望[J]. 中国电机工程学报, 2012, 32(29): 1-8.
JIANG Minhua, HUANG Bin. Prospects on coal-fired power generation technology development[J]. Proceedings of the CSEE, 2012, 32(29): 1-8.

[3] 谢昌亚, 朱龙飞, 胡娱欧, 等. 660 MW 超临界火电机组汽轮机及其调速系统精细化模型研究和应用[J]. 热能动力工程, 2023, 38(6): 58-67.
XIE Changya, ZHU Longfei, HU Yuou, et al. Research and application of refined model for 660 MW supercritical thermal power unit steam turbine and its governing system[J]. Journal of Engineering for Thermal Energy and Power, 2023, 38(6): 58-67.

[4] 李树明, 刘青松, 朱小东, 等. 350 MW 超临界热电联产机组灵活性改造分析[J]. 发电技术, 2018, 39(5): 449-454.

LI Shuming, LIU Qingsong, ZHU Xiaodong et al. Flexibility transformation analysis of 350 MW supercritical cogeneration unit[J]. Power Generation Technology, 2018, 39(5): 449-454.

[5] 于国强, 胡尊民, 张天海. 全工况下的 1 000 MW 超超临界机组协调控制系统多模型广义预测控制方法及其工程应用[J]. 热能动力工程, 2020, 35(5): 9-16.
YU Guoqiang, HU Zunmin, ZHANG Tianhai. Engineering application of multi-model generalized predictive control method to coordinated control system of 1 000 MW ultra supercritical power units under all working conditions[J]. Journal of Engineering for Thermal Energy and Power, 2020, 35(5): 9-16.

[6] 王竹, 吴鹏, 张锐锋, 等. 基于历史数据的汽轮机调节阀流量特性优化[J]. 热力发电, 2019, 48(2): 39-44.
WANG Zhu, WU Peng, ZHANG Ruifeng, et al. Research on flow characteristics of steam turbine regulating valve based on historical data[J]. Thermal Power Generation, 2019, 48(2): 39-44.

[7] 张小桃, 倪维斗, 李政, 等. 基于现场数据热工对象建模的可辨识性[J]. 清华大学学报(自然科学版), 2004(11): 1544-1547.
ZHANG Xiaotao, NI Weidou, LI Zheng, et al. Identifiability of building thermal system models using on-line data[J]. Journal of Tsinghua University (Science and Technology), 2004(11): 1544-1547.

[8] 尹琦, 潘蕾, 沈炯, 等. 基于闭环辨识的除氧器水位最优前馈-反馈控制器设计[J]. 工程热物理学报, 2020, 41(10): 2380-2385.
YIN Qi, PAN Lei, SHEN Jiong, et al. Design of optimal feedforward-feedback controller based on closed-loop identification model of deaerator water level system[J]. Journal of Engineering Thermophysics, 2020, 41(10): 2380-2385.

[9] 石饶桥. 基于历史数据的燃煤电厂 SCR 脱硝系统辨识与控制研究[D]. 南京: 东南大学, 2017: 1.
SHI Raoqiao. Identification and control of coal-fired power plant SCR de-NO_x system based on historical operation data[D]. Nanjing: Southeast University, 2017: 1.

[10] 陈尚年, 李录平, 张世海, 等. 基于EEMD-LSTM的汽轮机转子碰磨故障诊断模型及其工程应用[J]. 热能动力工程, 2023, 38(8): 159-168.
CHEN Shangnian, LI Luping, ZHANG Shihai et al. EEMD-LSTM-based turbine rotor rub-impact fault diagnosis model and its engineering application[J]. Journal of Engineering for Thermal Energy and Power, 2023, 38(8): 159-168.

[11] 胡涛, 武梦婷, 胡巍, 等. 基于机器学习的粉煤灰活性分类预测[J]. 中南大学学报(自然科学版), 2023, 54(10): 3829-3839.
HU Tao, WU Mengting, HU Wei, et al. Reactivity classification prediction of coal fly ash based on machine learning[J]. Journal of Central South University (Science and Technology), 2023, 54(10): 3829-3839.

[12] 杨贺雅, 邢纹硕, 陈聪, 等. 基于随机森林二分类器的模块化多电平换流器子模块开路故障检测方法[J]. 中国电机工程学报, 2023, 43(10): 3916-3928.
YANG Heya, XING Wenshuo, CHEN Cong, et al. A fault detection and location strategy for sub-module open-circuit fault in modular multilevel converters based on random forest binary classifier[J]. Proceedings of the CSEE, 2023, 43(10): 3916-3928.

[13] 张小科, 王子杰, 夏大伟, 等. 基于长短时记忆神经网络的深度调峰火电机组一次调频能力在线估计[J]. 热力发电, 2023, 52(8): 172-178.
ZHANG Xiaoke, WANG Zijie, XIA Dawei, et al. On-line estimation of primary frequency regulation capability of deep peak regulation thermal power unit based on LSTM neural network[J]. Thermal Power Generation, 2023, 52(8): 172-178.

[14] 张小科, 王子杰, 夏大伟, 等. 一种面向深度调峰运行火电机组的一次调频能力建模新方法[J]. 电网技术, 2022, 46(12): 4947-4954.
ZHANG Xiaoke, WANG Zijie, XIA Dawei, et al. New modeling for primary frequency regulation capability of thermal power units under deep peak regulation[J]. Power System Technology, 2022, 46(12): 4947-4954.

[15] SHARDT Y A W, HUANG B. Data quality assessment of routine operating data for process identification[J]. Computers & Chemical Engineering, 2013, 55: 19-27.

[16] SHARDT Y A W, HUANG B. Closed-loop identification with routine operating data: effect of time delay and sampling time[J]. Journal of Process Control, 2011, 21(7): 997-1010.

[17] 吕红燕, 冯倩. 随机森林算法研究综述[J]. 河北省科学院学报, 2019, 36(3): 37-41.
LYU Hongyan, FENG Qian. A review of random forest algorithm[J]. Journal of Hebei Academy of Sciences, 2019, 36(3): 37-41.

[18] CUTLER D R, EDWARDS JR T C, BEARD K H, et al. Random forests for classification in ecology[J]. Ecology, 2007, 88(11): 2783-2792.

[19] 马骊. 随机森林算法的优化改进研究[D]. 珠海: 暨南大学, 2017: 1.
MA Li. Research on optimization and improvement of random forest algorithm[D]. Zhuhai: Jinan University, 2017: 1.

[20] 王芳, 马素霞, 王河. 基于随机森林变量选择的飞灰含碳量预测模型[J]. 热力发电, 2018, 47(11): 89-95.
WANG Fang, MA Suxia, WANG He. Prediction model of carbon content in fly ash using random forest variable selection method[J]. Thermal Power Generation, 2018, 47(11): 89-95.

[21] 肖黎, 罗嘉, 欧阳春明. 基于半监督学习方法的磨煤机故障预警[J]. 热力发电, 2019, 48(4): 121-127.
XIAO Li, LUO Jia, OUYANG Chunming. Research on coal mill fault prediction based on semi-supervised learning method[J]. Thermal Power Generation, 2019, 48(4): 121-127.

（责任编辑 李园）