第 40 卷　第 3 期(517~531)　　中 国 地 震　　Vol. 40　No. 3
2024 年 9 月　　EARTHQUAKE RESEARCH IN CHINA　　Sep. 2024

王锦红,蒋海昆. 2024. 基于机器学习随机森林算法的地震序列类型判定研究. 中国地震,**40**(3):517~531.

基于机器学习随机森林算法的地震序列类型判定研究

王锦红[1)]　蒋海昆[2)]

1)南方科技大学地球与空间科学系,广东深圳　518055
2)中国地震台网中心,北京　100045

摘要　基于 1970—2021 年中国大陆地震目录、地震序列目录和历史地震震源机制资料,构建地震序列类型判定训练、检验特征样本数据集,将地震序列标签分为多震型、主余型、孤立型三类。采用特征递归消除-随机森林(RFE-RF)机器学习算法建立地震特征参数和地震序列类型之间的非线性映射关系,对震后 3 个不同时间节点的地震序列类型进行早期预测,并对特征重要性进行讨论。结果显示,数据预处理方法对模型分类性能有重要影响,同类样本中位值补齐缺失特征并采用随机重采样方法预处理可达到较高的分类预测效果。对分类结果的交叉检验结果显示,在震后 1 天,三类样本的总体报准率可达 0.93。考察模型最优特征子集随时间的变化可见,在地震刚发生时(即缺乏地震序列资料数据的情况下),相对于传统的历史地震序列类比,主震震源机制相关参数以及主震震源机制 P 轴方位相对于附近区域应力场的偏差等相关参数具有更大的分类贡献率。随着震后时间的延长,序列相关特征成为地震序列类型判定的主要因素。在震后 3 天,在可能已发生最大余震的地震序列数据集中,主震与最大余震震级差成为判定地震序列类型的关键因素。相较于单一的随机森林(RF)模型,RFE-RF 模型的在震后 1 天测试集中报准率提高了 0.41,能够更有效地对地震序列类型加以区分。

关键词:　**地震序列类型判定　中国大陆　随机森林　递归消除**

[**文章编号**] 1001-4683(2024)03-0517-15　　[**中图分类号**] P315　　[**文献标识码**] A

0　引言

地震序列类型判定是震后趋势预测的基础,基于历史地震类比和序列参数计算的序列类型判定是当前广泛采用的余震预测方法(蒋海昆等,2015)。由于单参数序列类型识别正确率并不理想(蒋海昆等,2006),对序列类型的判定较多地依赖于“统计+经验”的方法以及研究者的经验和能力(蒋海昆等,2023)。当前的序列类型判定方法很难识别多震型地震和前震(蒋海昆等,2020),这些类型的地震序列又是可能导致更严重灾害的地震类型,引入人工智能技术或成为解决这一问题的有效途径。

[**收稿日期**] 2023-04-25　[**修定日期**] 2024-06-03
[**项目类别**] 地震动力学国家重点实验室开放基金(LED2022B05)资助
[**作者简介**] 王锦红,女,1999 年生,博士研究生,主要从事人工智能强余震预测研究。E-mail:17866618823@163.com
蒋海昆,通讯作者,男,1964 年生,研究员,主要从事余震统计、余震机理及余震预测研究。
E-mail:jianghaikun@seis.ac.cn

机器学习理论和应用研究始于 1986 年,其主旨是计算机模拟或实现人类学习行为。作为人工智能的重要内容,其目的是从海量、多源、多维度的数据中寻找知识规律,建立学习模型,进而通过已获得的学习模型对其他数据进行分类与预测(杨午阳等,2019)。近年来,机器学习中的随机森林(Random Forest,简称 RF)算法被广泛应用于地震预测(Asim et al,2020; Asencio-Cortés et al,2018)、火山地震事件分类(Hibert et al,2017; Malfante et al,2018)、震后灾害评估(刘金平等,2024)等领域,该算法具有调参少、拟合精度高、泛化能力强、不易产生过拟合的特点,并对异常值和噪声具有较好的容忍度(Breiman,2001)。此外,随机森林模型还能评估指标的重要性,在处理高维数据的分类和回归问题时表现出较强的性能(Ho,1998; Cracknell et al,2013)。尤其在分类问题上,与其他模型相比,随机森林模型提供的预测准确性最高(Fernández-Delgado et al,2014)。Al Banna 等(2020)对 2005—2019 年全球地震预测领域相关文献进行检索分析时也发现,基于机器学习模型的地震预测报准率相对较高。

随着观测站网密度的增加和观测技术的进步,地震序列信息日益丰富。本文的研究重点在于如何应用机器学习算法来进行地震序列的类型判定,并提高判定的准确性。基于 1970—2021 年中国大陆及边邻地区的 $M_S \geq 5.0$ 地震序列数据,从区域和序列地震目录中提取样本特征,建立地震序列样本数据集,在此基础上构建随机森林分类模型,分析不同数据预处理方式对分类结果的影响。为精确过滤冗余特征,提高模型识别准确率,在模型训练之前加入递归特征消除这一特征选择算法。此外,已有研究表明地震序列最大余震往往发生在主震后的早期阶段,例如全球 7 级以上地震最大余震与主震之间时间间隔的平均值小于 50 天,中位值为 3 天(Tahir et al,2012); 针对全球 7 级以上地震的统计结果显示,超过一半的地震序列在主震发生后 3 天内出现最大余震(苏有锦等,2014)。考虑震后早期阶段迫切的实际需求,本文分别在地震刚发生时、震后 1 天、震后 3 天三个时间尺度上进行地震序列早期判定研究。

1 特征数据集

1.1 基础数据及样本标签

根据中国地震台网统一地震目录[①],中国大陆及边邻地区 1970—2021 年共记录 $M_S \geq 5.0$ 地震 1336 次,依据余震破裂范围(Wells et al,1994)及余震活动持续时间(Lolli et al,2003)甄别并删除其中的余震,删除余震之后有 $M_S \geq 5.0$ 地震 902 次,其中 5.0~5.9 级地震 722 次、6.0~6.9 级地震 153 次、7.0~7.9 级地震 25 次、8.0 级以上地震 2 次。根据序列主震和最大余震的震级差 $\Delta M = M_0 - M_1$ 将以上地震序列分为三类(蒋海昆等,2007、2015),分别为多震型(Multiplet Mainshocks Type,简称 MMT):$\Delta M < 0.6$;主-余型(Mainshock-aftershock Type,简称 MAT):$0.6 \leqslant \Delta M \leqslant 2.4$; 孤立型(Isolated Earthquake Type,简称 IET):$\Delta M > 2.4$ 且地震次数较少。

在实际研究中,还有一类地震序列类型在余震预测中使用频率相对低,但对后续地震趋势判定具有一定意义,即主震前发生的显著的前震活动。笼统而言,主震前短时间内震源区

① http://data.earthquake.cn/data/

震级小于主震的地震均可称为前震(蒋海昆等,2020),但为将其与震后趋势研判业务中广泛使用的多震型序列($-0.6<\Delta M<0.6$)相区分,可约定符合 $\Delta M\leqslant -0.6$ 的地震序列为前震序列(蒋海昆等,2023)。由于本文目标地震(序列主震)均为 $M_0\geqslant 5.0$ 地震,因而样本中符合主震 $M_0\geqslant 5.0$ 且震级差 $\Delta M\leqslant -0.6$ 的样本仅 17 例,前震型、多震型均属后续地震危险性较大的序列类型(前震型后续存在发生比主震更大地震的危险,多震型后续存在发生与主震同等大小地震的危险),加之本文前震型样本数量较少,因而按此前做法(蒋海昆等,2023),在模型训练及检验过程中,将前震序列与多震型归为一类,暂且仍称为"多震型"序列。此外对于后续无余震记录从而"无法确认"序列类型的 181 次震例,统一归并为"孤立型"序列,其特点是后续均无显著余震活动。简化后的地震序列基础数据如表 1 所示。

表 1　　1970—2021 年中国大陆及边邻地区 $M_S\geqslant 5.0$ 地震序列基本概况(据蒋海昆等(2023))

震级范围	全部地震	删除余震	1-多震型(含前震型)	2-主余型	3-孤立型(含无法确认类型的地震序列)
M_S5.0 ~5.9	1103	722	88(含前震型 15 例)	320	314(含无法确认类型的地震序列 170 例)
M_S6.0 ~6.9	202	153	20(含前震型 2 例)	90	43(含无法确认类型的地震序列 9 例)
M_S7.0 ~7.9	29	25	4	19	2(含无法确认类型的地震序列 2 例)
$M_S\geqslant 8.0$	2	2		2	
合计	1336	902	112	431	359
所占比例			12.42%	47.78%	39.80%

1.2　样本特征及特征数据完备性

蒋海昆等(2023)提出用于机器学习地震序列类型判定的 44 个备选特征,包括主震相关参数、主震震源机制相关参数及相对于附近区域平均应力场的偏差、主震附近区域历史地震序列类型、指定时段的序列衰减和G-R关系相关参数、归一化能量熵、序列地震震级及频次相关参数。在基于地震序列目录计算的参数中,"震级下限"可依据资料完备情况取多个震级,"指定时段"亦可包含一系列震后不同的统计时段,由此可构建更多的样本特征。例如表 2 特征 20~31 即为历史地震序列类型占比经过变换震级下限 x 后得到的多个特征参数,x 分别取 3.0、4.0、5.0、6.0,对应震后不同时段震中附近区域 $M_S\geqslant x$ 地震样本中多震型、主余型及孤立型所占的比例;同样,特征 66~71 中涉及的震级下限 x 也分别取 M_L3.0 和 M_L3.5。表 2 中特征 46~64 中的时间窗分别设定为震后 1h、2h、3h、6h、12h、18h 以及震后 1 天、3 天等指定时段。最终,通过变换震级下限、统计时段等参数,将上述 8 类 44 项备选特征扩充得到本文使用的 76 个机器学习特征(表 2)。

表 2　　地震序列类型判定机器学习特征(据蒋海昆等(2023))

序号	特征分类	符号表达	特征物理属性描述
1~3	(1)主震相关特征	1La	主震纬度/(°)
		2Lo	主震经度/(°)
		3M_S	主震震级(M_S)

续表 2

序号	特征分类	符号表达	特征物理属性描述
4~11	(2.1)主震震源机制相关特征	4StrA	节面 A 方位角/(°)
		5DipA	节面 A 倾角/(°)
		6 RakeA	节面 A 滑动角/(°)
		7StrB	节面 B 方位角/(°)
		8DipB	节面 B 倾角/(°)
		9 RakeB	节面 B 滑动角/(°)
		10AziP	*P* 轴方位角/(°)
		11DipP	*P* 轴倾角/(°)
12~15	(2.2)主震附近区域平均应力场相关特征	12 MeanAziP	主震附近区域 *P* 轴方位角平均值/(°)
		13 StdAziP	主震附近区域 *P* 轴方位角标准差/(°)
		14 MeanDipP	主震附近区域 *P* 轴倾角平均值/(°)
		15 StdDipP	主震附近区域 *P* 轴倾角标准差/(°)
16~19	(2.3)主震应力场相对于附近区域平均应力场的偏差	16 DiffAziP	主震 *P* 轴方位角与附近区域 *P* 轴平均方位角之差/(°)
		17 DiffAziP/StdAziP	主震 *P* 轴方位角与附近区域 *P* 轴平均方位角之差/*P* 轴方位角标准差/(°)
		18 DiffDipP	主震 *P* 轴倾角与附近区域 *P* 轴平均倾角之差/(°)
		19 DiffDipP/StdDipP	主震 *P* 轴倾角与附近区域平均倾角之差/*P* 轴倾角标准差/(°)
20~31	(3)主震附近区域 $M_S \geqslant x$ 历史地震序列类型占比(注：x=3.0、4.0、5.0、6.0)	20-23M_S Ty1	多震型序列 MMT 占比
		24-27M_S Ty2	主余型序列 MAT 占比
		28-31M_S Ty3	孤立型序列 IET 占比
32~38	(4)指定时段序列衰减相关特征	32pVal	修改的大森公式 *p* 值
		33 AveDeltaN	实际地震频次与修改的大森公式理论频次之差绝对值的平均值
		34 StdDeltaN	实际与理论频次之差绝对值的标准差
		35k_Mcon	折合震级线性衰减速率
		36b_Mcon	折合震级线性衰减纵轴截距(M_L)
		37 AveDeltaMcon	实际折合震级与线性衰减理论折合震级之差绝对值的平均值
		38 StdDeltaMcon	实际与理论折合震级之差绝对值的标准差
39~44	(5)指定时段G-R关系相关特征	39bVal	MLE 方法 *b* 值
		40bSD	MLE 方法 *b* 值标准差
		41aVal	G-R关系 *a* 值
		42aSD	G-R关系 *a* 值标准差
		43M_S Pre	G-R 外推震级(M_S)
		44M_0-M_S Pre	主震与 G-R 外推震级之差
45	(6)指定时段归一化能量熵	45 Entropy	指定时段序列归一化能量熵

续表 2

序号	特征分类	符号表达	特征物理属性描述
46~65	(7)指定时段最大余震震级相关特征（注：时间窗取震后1h、2h、3h、6h、12h、18h）	46-51Mcon	震后指定时段内序列余震的折合震级
		52-57Mmax	震后指定时段内的最大余震震级
		58-63 DeltaM	主震与震后指定时段的最大震级差
		64M_S ASK	最大余震震级(M_S)
		65 DeltaM_S ASK	主震与最大余震震级差
66~76	(8)指定时段 $M_L \leqslant x$ 小震频次及震级相关特征（注：x=3.0、3.5）	66-67CumN_M$_L x$	累积频次
		68-69AveN_M$_L x$	日平均频次
		70-71Nor_N_M$_L x$	相对震后第 1 天的归一化频次
		72-73M_S AveM$_L x$	平均震级(M_L)
		74-75M_S StdM$_L x$	平均震级标准差(M_L)
		76Mcon M_S3.0	$M_L \geqslant$3.0 地震折合震级(M_L)

1.3　特征数据完备性

开展机器学习模型训练前，需要考虑样本中不同特征参数的数据完备性问题。如上所述，地震序列样本的特征提取需计算多个序列参数，其中有些参数要求一定数量的地震样本才能得出准确结果，而部分早期地震序列或发生在监测能力较低地区的地震，以及那些本身就不太活跃的地震序列可能会缺乏足够的样本数量，这导致无法计算出部分特征参数。对 76 个特征参数震后 3 天数据的统计分析结果（图 1）显示，主震参数及主震附近区域历史地震相关参数的数据完备性相对较高，高于 98%；而对于涉及主震震源机制的相关参数，数据完备性相对较低，只有约 70%的样本可以计算相关特征参数。修改的大森公式及G-R关系相关参数的数据完备性最低，仅有约 32%的样本可以计算相关特征参数。

2　模型及算法

2.1　随机森林模型

随机森林分类模型是一种包含多个决策树的集成模型（Breiman，2001），其通过构建多个决策树来进行分类预测。决策树模型结构简单，是基于树结构来进行决策的一种算法，其利用样本特征作为树结构中的节点。节点分为三种，分别为根节点、内部结点和叶节点（图 2）。根节点、内部结点、叶节点是机器学习树模型一些专有名词，根节点是决策树最顶层的节点，其表示具有最初特征的所有样本的集合；内部节点为决策节点，每个内部节点为一个判断条件，该节点包含样本集中满足从根节点到该节点所有条件的样本的集合；叶节点位于决策树的底层，每个叶节点均对应一个样本的分类决策结果。决策树从根节点开始对样本的某一特征进行测试，根据测试结果将样本分配到其子节点，这时每一个子节点对应着该特征的一个取值。如此递归地对样本进行测试并分配，直至达到叶节点，最后将样本分到叶节点的类中，因此，每棵决策树对应着从根节点到叶节点的一组规则。换言之，决策树的构建是一种自上而下的归纳过程。在解决分类问题时，决策树算法具有复杂度低、计算高效的优点，但其不适用于处理特征缺失的数据，因为缺失数据会导致算法无法确定属性取值

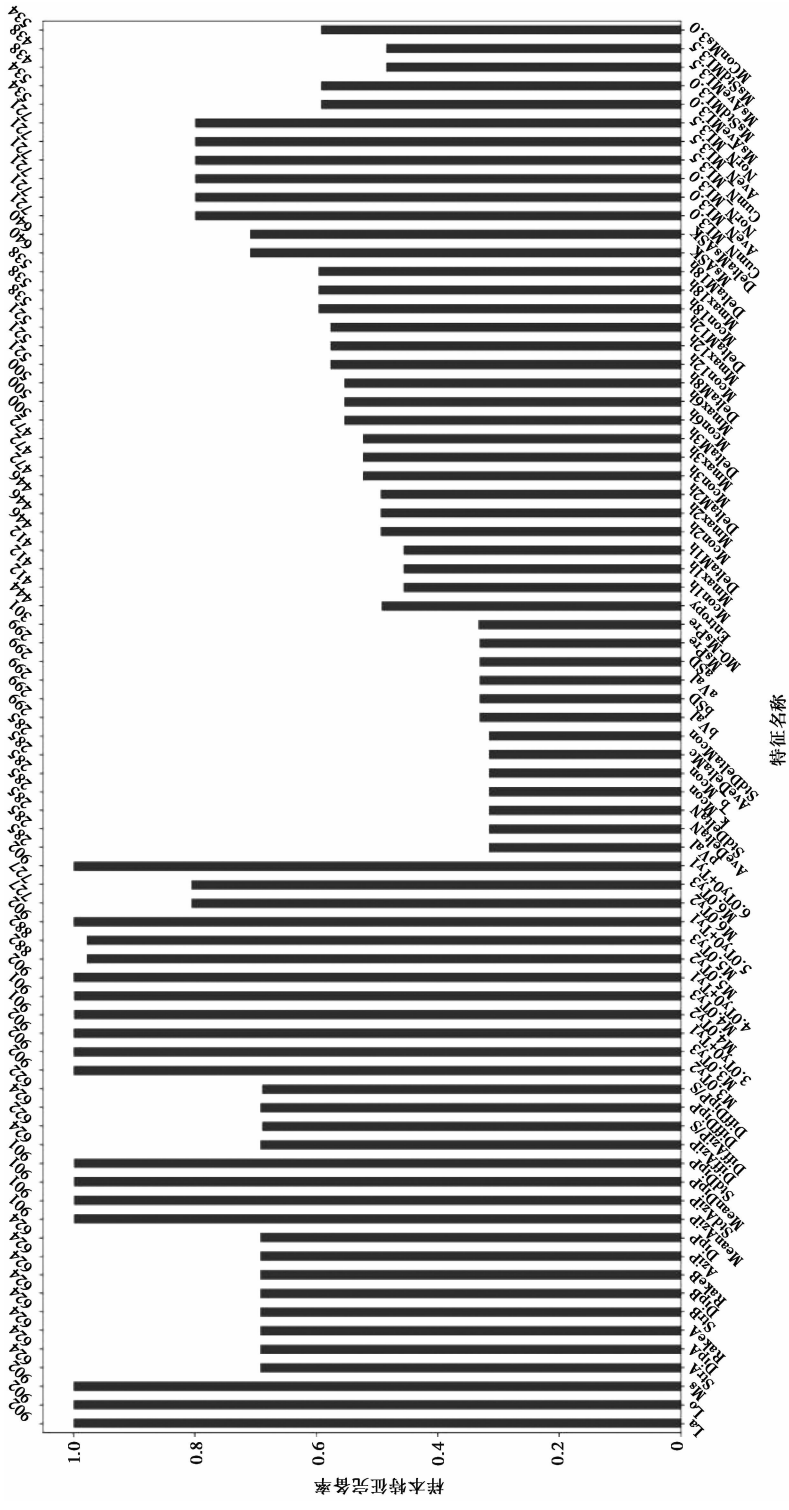
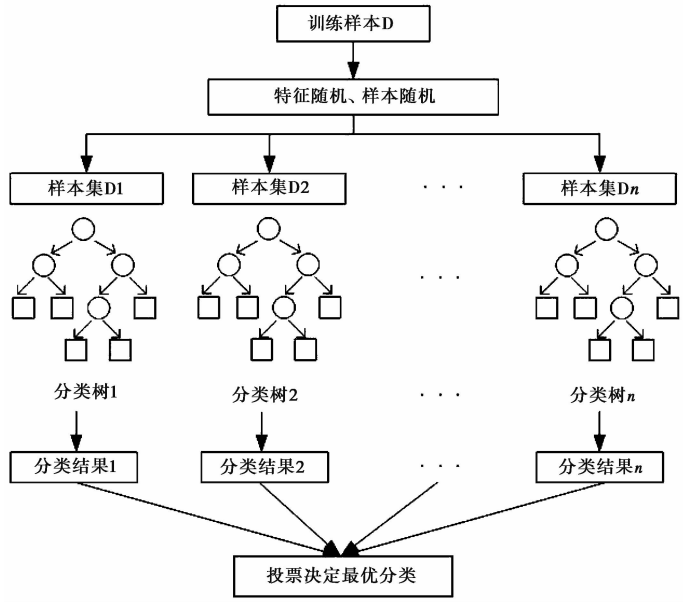


图 1 样本集特征缺失情况统计



注：圆形表示根节点/内部节点，方块表示叶节点。

图 2　随机森林模型示意图（据董红瑶等（2021））

和节点分裂条件，影响分类的准确性，因而需对样本数据进行特征补齐等预处理（Sebastiani，2002）。此外，决策树为追求高的拟合精度，可能会对样本空间过度分割从而导致过拟合问题（Breiman，2001；Domingos，2012）。

将多个决策树弱分类器结合起来进行集成学习，往往可以获得比单一决策树显著优越的泛化性能（周志华，2016）。随机森林分类模型就是多个决策树分类模型的集成，每个决策树分类模型均有一票投票权来选择最优的分类结果。随机森林分类的基本思想是：①在原始样本数据集中随机抽取一定比例的样本，构成 n 个样本不同但样本容量相同的样本数据子集；②对 n 个样本数据子集分别搭建 n 个决策树模型，得到 n 种分类结果；③对得到的 n 种分类结果进行投票，最终分类结果采纳多数投票结果，随机森林分类流程如图 2 所示，包括特征随机化和样本随机化两种重要的随机化步骤。在特征随机化中，每个决策树只使用数据集的一部分特征进行建模，这样可以减少决策树之间的相关性，提高整个随机森林的泛化能力；在样本随机化中，每个决策树只使用数据集的一部分样本进行建模，这样可以减少随机森林对训练集的过度拟合。

2.2　递归随机消除算法

在机器学习分类问题中，去除不相关特征往往会降低学习任务的难度（周志华，2016）。为精确过滤冗余特征，本文在模型训练之前加入递归特征消除这一特征选择算法。递归特征消除（Recursive Feature Elimination，简称 RFE）是使用一个基模型进行多轮训练，每轮训练后，消除若干重要性低的特征，再基于新的特征集进行下一轮训练（Gregorutti et al，2017）。RFE 最初应用于支持向量机模型，通过迭代训练模型对特征进行排序，每次移除一个重要性排序最低的特征，使其能够进行特征选择（Guyon et al，2002）。在本文中，以随机森林 RF 为

基模型,构建递归消除和随机森林相融合的 RFE-RF 模型(Jiang et al,2004; Svetnik et al,2004),RFE-RF 模型能够充分利用特征递归消除和随机森林算法的优势,同时克服单次随机森林的特征选择结果需要反复实验得到特征子集的缺点,使最终得到的特征子集更加可靠(Kuhn M,2011)。

2.3 评价指标

令多震型 MMT、主余型 MAT 和孤立型 IET 三类地震序列分别为地震序列类型 1、类型 2 和类型 3,表 3 给出地震序列类型判定模型的分类结果和实际情况的混淆矩阵。样本总数为 n,矩阵中的每个元素 n_{xy} 表示样本实际情况为 y、预测结果为 x 的样本数量,其中 $x,y=1,2,3$,分别代表地震序列类型 1、序列类型 2 和序列类型 3。例如,n_{23} 表示实际为 IET、但被预测为 MAT 的样本数量。

表 3 **地震序列类型判定模型评估中的混淆矩阵**

混淆矩阵		实际情况		
		1	2	3
预测结果	1	n_{11}	n_{12}	n_{13}
	2	n_{21}	n_{22}	n_{23}
	3	n_{31}	n_{32}	n_{33}

报准率是指地震序列被正确分类的比例,该比例越高越好,但模型有时会以大量的虚报来提高报准率。本文结合地震序列类型判别的实际情况,选用报准率、漏报率和虚报率作为序列分类模型的评估指标。其中,报准率公式为

$$\text{报准率} = \frac{n_{11} + n_{22} + n_{33}}{n} \tag{1}$$

漏报率是危险性高的序列类型被错误分类为危险性低的序列类型所占的比例。例如多震型被错误地划分为主余型或孤立型,或者主余型被错误地划分为孤立型。漏报率越低意味着漏报的强余震越少。漏报率公式为

$$\text{漏报率} = \frac{n_{21} + n_{31} + n_{32}}{n} \tag{2}$$

虚报率是危险性低的序列类型被错误地分类为危险性高的序列类型所占的比例。虚报率越低表征模型误报强余震的可能性越小。虚报率公式为

$$\text{虚报率} = \frac{n_{12} + n_{13} + n_{23}}{n} \tag{3}$$

在实际应用中,序列分类模型的报准率、漏报率和虚报率需要根据具体任务和需求来确定。例如,对强余震预测而言,低的漏报率可能会更加关键,因为漏报强余震会导致严重的不良后果。

3 数据处理及结果分析

数据预处理是机器学习的重要环节,是提高数据可靠性、确保模型训练质量、提高模型分类性能的重要前提。首先探讨多种数据预处理方式对模型分类结果的影响,在此基础上

选择最佳的数据预处理方法对样本数据进行预处理。进而探索特征选择对模型的影响，以缩减特征维度，提高模型计算效率。本文着眼于地震序列类型判别的实际应用，在数据处理阶段为不同震后时段的地震序列判别做准备。

3.1　数据预处理及对结果的影响

根据实际需要，本文训练了适用于震后三个不同时间节点开展序列类型判定的序列分类模型。模型 1 在震后立刻进行序列分类，该模型未使用地震序列资料，所使用的特征来源于震中附近历史地震活动、主震及主震震源机制的相关参数；模型 2、模型 3 在模型 1 的基础上，增加了震后 1 天和震后 3 天的序列资料，即增加了震后不同时段序列的相关参数作为特征。

机器学习模型的训练需要足够的学习样本以确保结果的可靠性，但实际地震序列类型判定样本数据集有三点不足：一是样本数少，仅有 902 个样本；二是如上所述，许多样本存在特征缺失的情况；三是如表 1 所列，三类样本存在明显的样本不均衡，例如主余型样本数约是多震型样本数的 4 倍，这会导致模型学习和训练结果更多地受多样本类型的影响。然而样本数较少的多震型地震序列本身又是余震预测的重点。

由于样本数较少，删除缺失特征样本的做法不可取。针对存在特征缺失的样本，本文采取分类型样本的特征中位值进行填补，这种方法可以更好地保留数据的分布特征，提高模型的鲁棒性（Poulos et al，2018），相对于直接删除含有特征缺失值样本的做法，这种填充方法在保留数据的同时考虑了不同类样本的特性，避免其他地震序列类型对该类型样本的影响，结果更为可靠（Kang et al，2013）。针对样本不均衡问题，对多震型序列样本的特征进行重采样，以构建新的“伪”样本，使其样本数量与其他地震序列类型样本数相当，从而使得数据集中每个类别的样本数量基本相同，这样做的最重要作用是可以避免模型偏向于预测样本数量多的类别。删除缺失特征、缺失特征补齐及不均衡样本处理之后的样本状况如表 4 所列。在此基础上将样本数据集按照 7∶3 的比例划分为训练集和测试集，以进行模型的学习训练和测试。

针对地震序列类型，不同的数据预处理方法可能会对分类器的性能产生显著影响。基于 RFE-RF 模型，探讨表 4 所列三种不同的数据预处理方法对分类结果的影响：①删除含有缺失特征的样本；②利用同类型样本中位值补齐缺失特征；③利用同类型样本中位值补齐缺失特征并重采样少数样本以解决样本不平衡问题。通过比较不同数据预处理方法下 RFE-RF 模型在震后不同时间尺度的报准率、漏报率和虚报率等指标（表 5），寻找最适合地震序列分类任务的数据预处理方法。

对大数据样本而言，直接删除缺失特征样本是一种简单的数据预处理方式，且对结果不会产生大的影响。但对小样本问题、尤其针对地震预测这一类样本数极其有限的情形，直接删除含缺失特征样本并不是一种好的数据预处理方式，即使在震后第 3 天，序列数据已经相对完整，报准率也仅为 0.739，这一结果甚至达不到始终报告“安全”的水平。已往的研究表明，主-余型和孤立型地震约占地震序列的 80%左右（吴开统，1971；蒋海昆等，2006；苏有锦等，2014；薛艳等，2018），在利用分类中位值补齐含缺失特征后，模型报准率有明显提升，在震后 1 天和震后 3 天报准率均提升约 0.20，均达到 0.95 以上，同时漏报率和虚报率均有所降低。在此基础上，针对样本数较少的多震型序列类型，通过重采样算法平衡样本之后，报准率略微降低，震后 1 天和震后 3 天报准率分别为 0.935 和 0.934，漏报率和误报率略微升

表 4 不同预处理情况下的样本数统计

模型	使用特征	序列类型	数据预处理方法		
			删除含缺失特征的样本	缺失特征补齐	缺失特征补齐+样本不均衡处理
模型 1	主震及震源机制相关参数,震中附近历史地震活动相关参数(表 2 中特征 1~31)	MMT	71	112	430
		MAT	286	430	430
		IET	157	360	360
		样本总数	512	902	1220
模型 2	主震及震源机制相关参数,震中附近历史地震活动相关参数,震后 1 天序列资料相关参数(表 2 中特征 1~76)	MMT	28	112	430
		MAT	125	430	430
		IET	7	360	360
		样本总数	160	902	1220
模型 3	主震及震源机制相关参数,震中附近历史地震活动相关参数,震后 3 天序列资料相关参数(表 2 中特征 1~76)	MMT	33	112	430
		MAT	133	430	430
		IET	8	360	360
		样本总数	174	902	1220

表 5 数据预处理对 RFE-RF 模型性能的影响

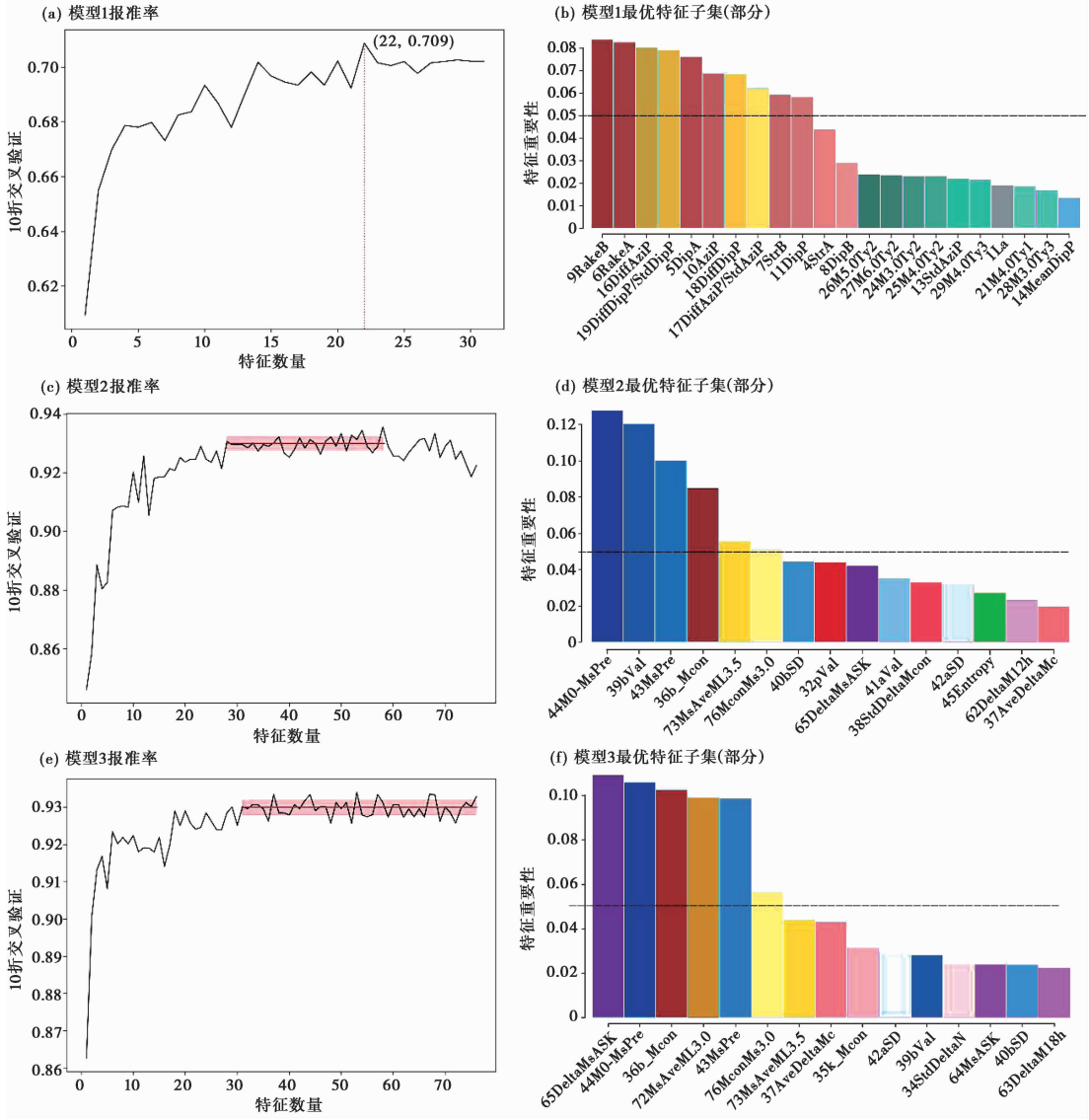
模型 RFE-RF	模型 1			模型 2			模型 3		
	报准率	漏报率	虚报率	报准率	漏报率	虚报率	报准率	漏报率	虚报率
删除含缺失特征样本	0.643	0.206	0.151	0.762	0.115	0.123	0.739	0.137	0.124
缺失特征补齐	0.730	0.165	0.105	0.958	0.019	0.023	0.952	0.023	0.025
缺失特征补齐+样本不均衡处理	0.709	0.176	0.115	0.935	0.052	0.013	0.934	0.032	0.034

高,相较于直接删除含缺失特征样本的处理方法具有明显优势。尽管重采样少数样本能够缓解数据集中的不平衡问题(Kaur et al,2018),但根据本文的结果来看,只采用同类样本中位值补齐缺失特征而不进行数据不平衡处理的方式,其分类准确率似乎更高,这可能是因为数据集中的不平衡程度未达到需要采取数据量不平衡处理方法的标准[②]。同时这也说明,不平衡样本的预处理及对预测效果的影响需要进一步深入研究。换言之,数据预处理对于模型分类效果有着重要的影响。综合考虑报准率、漏报率和虚报率,利用同类样本中位值补齐缺失特征的数据预处理方式适用于地震序列类型判定问题。

3.2 模型检验及特征重要性讨论

采用同类型样本中位值补齐缺失特征并重采样少数样本以解决样本不平衡问题的数据预处理方式,利用 RFE-RF 模型交叉验证,得到模型报准率随特征数量的变化,如图 3(a)、3(c)、3(e)所示。结果显示,随着震后时间的延长,即随着震后序列资料的加入,模型的整体性能逐渐提高,在震后 1 天的报准率达 0.93 以上。此外,随震后时间的延长,不同特征的分类贡献(特征重要性)会发生变化,下面尝试讨论模型选择的最优特征子集随时间的变化

② https://www.cnblogs.com/kamekin/p/9824294.html



注：图(c)、(e)中红色实线表示对应特征数量区间下的报准率平均值，红色阴影为均方差；为便于直观比较，图(d)、(f)中相同特征用相同的颜色表示。

图 3　震后不同时段 RFE-RF 模型分类报准率变化曲线及特征选择

情况。

地震刚发生时，尚无地震序列资料记录，使用历史地震序列类型类比是进行序列类型判定的重要依据(中国地震局，1998；张国民等，2010；蒋海昆等，2006)。从模型 1(震后立刻进行的序列类型判定)模型优选特征子集来看(图 3(b))，在无序列数据加入的前提下，历史地震序列类型相关特征参数的确有助于序列分类。除此之外，模型 1 最优特征重要性排序给出的结果显示，在特征重要性较高的特征中，主震震源机制相关参数、以及主震震源机制 P 轴方位相对于附近区域应力场的偏差等相关参数对序列分类的贡献更大，特征重要性排

序靠前的几个特征分别是表 2 中反映主震破裂形式的滑动角 9 RakeB 和 6 RakeA，以及主震 P 轴方位相对于区域应力场 P 轴方位偏差 16 DiffAziP 及数据波动水平 19 DiffDipP/StdDipP。这是一个全新的认识，一方面意味着主震破裂形式及其与区域应力场的关系有可能影响着余震的发展，另一方面也提示我们在缺乏序列资料的前提下，序列类型判定可能需要更多关注震源机制及应力场的相关参数。

在震后 1 天，由于震后序列资料的加入，模型 2 的报准率得到明显提升，并在特征数量为 30 个左右时分类的准确率趋于稳定，平均报准率约为 0.930，均方差为 0.022，如图 3(c)中红线及红色矩形阴影所示。当使用特征数量达到 60 个左右时，报准率开始下降，表明从报准率的角度，并非特征越多越好。进一步考察特征数量为 30 时模型选择的前 15 个重要特征，整体来看(图 3(d))，震后地震序列资料的加入使得序列相关特征参数特征重要性增加，代替地震当发生时的主震震源机制相关参数，成为地震序列判定的主要依据。在特征重要性高于 0.05 的特征中，主震震级 M_0 与序列 G-R 外推震级 M_S Pre 之差(特征 $44M_0$-M_S Pre)排在首位，这印证了以往序列主震与次大地震之间震级差对序列类型有一定的分类能力的认识(蒋海昆等，2015)。除此之外，G-R关系中的 b 值(特征 39bVal)也成为震后 1 天地震序列类型判定的重要特征，表明在震后早期大小地震的比例关系在某种程度上可能与序列类型有关。折合震级线性衰减纵轴截距(特征 $72M_S$ AveM_L3.5)和 3.5 级以上的平均震级(特征 36b_Mcon)分别从余震序列能量衰减和余震平均活动水平的角度，对地震序列判定做出贡献。

在震后 3 天，模型仍然在特征数量 30 个左右时趋于稳定，平均报准率为 0.930，均方差为 0.020 (图 3(e))。震后 3 天排名前 15 的特征(图 3(f))也可以得到大体相似认识，排名靠前的特征有主震震级 M_0 与序列 G-R 外推震级 M_S Pre 之差(特征 $44M_0-M_S$ Pre)、3.5 级以上的平均震级(特征 36b_Mcon)、折合震级线性衰减纵轴截距(特征 $72M_S$ AveM_L3.5)、G-R 外推震级($43M_S$ Pre)和 $M_L \geqslant 3.0$ 地震折合震级(76Mcon M_S3.0)。对比图 3(d)与图 3(f)可见，尽管前后顺序有变化，但重要性排序靠前的特征与震后 1 天时的特征基本相同。需要注意的是，在震后 3 天，主震与该时段最大余震震级差 65 DeltaM_S ASK 已成为地震序列类型判定中的最关键依据，其原因可能是由于部分地震序列实际最大余震在该时段可能已经发生(苏有锦等，2014；Tahir et al，2012；任雪梅等，2013)，但研究者在当时并不知晓。

3.3 RFE-RF 与 RF 模型预测结果对比

为检验 RFE-RF 模型对地震序列类型判定的能力，基于同类样本中位值补齐缺失特征并重采样少数样本的方式预处理震后 1 天的数据，对比 RFE-RF 与 RF 模型的分类结果，测试集上的实验结果如表 6 所示。结果显示，RFE-RF 模型的各项预测性能均有不同程度的提升。

表 6 **RFE-RF 与 RF 模型预测结果对比**

模型		报准率	漏报率	虚报率
模型 2	RF	0.945	0.049	0.005
	RFE-RF	0.986	0.016	0.003

4　结论和讨论

利用 1970—2021 年中国大陆及边邻地区 5 级以上地震及序列目录数据，建立地震序列类型判定特征数据集，采用递归消除和随机森林融合的分类算法，开展对地震序列类型判定研究，得到以下初步认识：

（1）不同的数据预处理方法对分类模型的性能产生不同的影响。利用同类样本中位值补齐缺失特征并采用随机重采样技术解决样本不平衡问题的数据预处理方法，对地震序列类型判定这一类小样本问题较为适宜，可以达到较好的分类预测效果，同时减少漏报和虚报的风险。对于是否进行不平衡样本处理、如何处理，仍是一个尚待深入探讨的问题。

（2）在无地震序列相关信息输入的情况下，历史地震序列类型占比通常是序列类型判定的重要依据。然而，基于历史地震活动、主震及主震震源机制相关参数模型优选特征结果表明，主震震源机制相关参数以及主震震源机制 P 轴方位相对于附近区域应力场的偏差等相关参数对序列分类的贡献更大。这一全新的认识提示我们，与附近区域历史地震序列类型相比较，主震破裂形式及其与区域应力场的关系可能是影响余震序列发展的重要因素，因此在缺乏序列资料的前提下，除历史地震活动类比外，在对序列类型判定时应同时关注主震破裂形式及应力场相关参数的可能影响。

（3）以震后不同时段的特征样本数据集作为 RFE-RF 序列类型判定模型的输入，结果显示地震序列资料的加入可明显提高地震序列判定模型的报准率。随震后时间的延长，序列相关特征代替主震震源机制和应力场偏差等特征成为地震序列类型判定的主要贡献因素。在震后 3 天，主震与该时段最大余震的震级差（65 DeltaM_S ASK）成为地震序列判别的最重要特征，这或许与部分震例最大余震在该时段可能已发生有关。除此之外，由于序列地震频次相关特征、归一化能量熵等特征也对地震序列类型判定起重要作用，这一时期的分类结果远优于地震刚发生时仅依赖于历史地震及主震相关参数特征的分类结果。

（4）本文机器学习序列类型判定主要基于当前地震序列判定工作中普遍使用的数据类型，诸如主震附近区域历史地震序列类型、地震序列相关参数以及主震及震源机制相关参数。但事实上，还有许多其他数据源可以提供更为详细的地震相关信息。例如地震波形记录可能可以进一步提供关于震源、路径等方面的相关信息，地震地质数据可以提供关于地震发生地区介质及构造背景等方面的特征信息，因此可以考虑综合利用多源数据进行综合分类的做法。但这也存在许多困难，最大的难点在于人们其实并不知道这些多源数据究竟是否与地震序列类型有关。因此，如何有效地融合多源数据以增加有效特征，以及如何进行特征选择，是进一步工作中需要解决的问题。此外，其他机器学习算法和特征选择方法在地震序列类型识别方面的应用也值得探讨，尤其是将基于物理的一些已知经验/信息引入到机器学习模型当中，将机器学习与物理概念相结合，是值得进一步探索的做法。

参考文献

董红瑶，王弈丹，李丽红. 2021. 随机森林优化算法综述. 信息与电脑，**33**（17）：34～37.

蒋海昆，傅征祥，刘杰，等. 2007. 中国大陆地震序列研究. 北京：地震出版社.

蒋海昆，李永莉，曲延军，等. 2006. 中国大陆中强地震序列类型的空间分布特征. 地震学报，**28**（4）：389～398.

蒋海昆，王锦红. 2023. 适用于机器学习的地震序列类型判定特征重要性讨论. 地震研究，**46**（2）：155～172.

蒋海昆,杨马陵,付虹,等. 2015. 震后趋势判定参考指南. 北京: 地震出版社.

蒋海昆,周少辉. 2020. 前震: 预测意义及识别方法. 地震地磁观测与研究,**41**(5):222~225.

刘金平,姜立新,杨天青,等. 2024. 基于随机森林的地震灾害直接经济损失评估研究——以中国西部地区为例. 中国地震, 40(2): 355~367.

任雪梅,谭俊林,马禾育,等. 2013. 中国大陆及边邻地区 6 级以上地震序列的最大余震统计特征. 中国地震, **29**(4): 480~488.

苏有锦,李忠华,赵小艳,等. 2014. 全球 7 级以上地震序列研究. 昆明: 云南大学出版社.

吴开统. 1971. 地震序列的基本类型及其在地震预报中的应用. 地震战线,**7**(11):45~51.

薛艳,刘杰,刘双庆. 2018. 全球浅源巨大地震序列统计特征. 中国地震, **34**(4): 676~694.

杨午阳,魏新建,何欣. 2019. 应用地球物理+AI 的智能化物探技术发展策略. 石油科技论坛,**38**(5):40~47.

张国民,钮凤林,邵志刚,等. 2010. 中国大陆 $M_S \geqslant 7.8$ 大震的余震活动差异性特征及其成因研究. 地震,**30**(4):1~12.

中国地震局. 1998. 地震现场工作大纲和技术指南. 北京: 地震出版社.

周志华. 2016. 机器学习. 北京: 清华大学出版社.

Al Banna M H, Taher K A, Kaiser M S, et al. 2020. Application of artificial intelligence in predicting earthquakes: state-of-the-art and future challenges. IEEE Access, **8**: 192880~192923.

Asencio-Cortés G, Morales-Esteban A, Shang X, et al. 2018. Earthquake prediction in California using regression algorithms and cloud-based big data infrastructure. Comput Geosci, **115**: 198~210.

Asim K M, Moustafa S S R, Niaz I A, et al. 2020. Seismicity analysis and machine learning models for short-term low magnitude seismic activity predictions in Cyprus. Soil Dyn Earthq Eng, **130**: 105932.

Breiman L. 2001. Random forests. Mach Learn, **45**(1): 5~32.

Cracknell M J, Reading A M. 2013. The upside of uncertainty: identification of lithology contact zones from airborne geophysics and satellite data using random forests and support vector machines. Geophysics, 78(3): WB113~WB126.

Domingos P. 2012. A few useful things to know about machine learning. Commun ACM, **55**(10): 78~87.

Fernández-Delgado M, Cernadas E, Barro S, et al. 2014. Do we need hundreds of classifiers to solve real world classification problems? J Mach Learn Res, **15**(1): 3133~3181.

Gregorutti B, Michel B, Saint-Pierre P. 2017. Correlation and variable importance in random forests. Stat Comput, **27**(3): 659~678.

Guyon I, Weston J, Barnhill S, et al. 2002. Gene selection for cancer classification using support vector machines. Mach Learn, **46**(1~3): 389~422.

Hibert C, Provost F, Malet J P, et al. 2017. Automatic identification of rockfalls and volcano-tectonic earthquakes at the Piton de la Fournaise volcano using a Random Forest algorithm. J Volcanol Geotherm Res, **340**: 130~142.

Ho T K. 1998. The random subspace method for constructing decision forests. IEEE Trans Pattern Anal Mach Intell, **20**(8): 832~844.

Jiang H Y, Deng Y P, Chen H S, et al. 2004. Joint analysis of two microarray gene-expression data sets to select lung adenocarcinoma marker genes. BMC Bioinform, **5**(1): 81.

Kang H. 2013. The prevention and handling of the missing data. Korean Journal of Anesthesiology, **64**(5), 402~406.

Kaur J, Gosain J. 2018. A comprehensive review on handling imbalanced data. Artif Intell Rev, **52**(2): 1087~1113.

Kuhn M. 2011. Variable selection using the caret package. International Review of Electrical Engineering, 24.

Lolli B, Gasperini P. 2003. Aftershocks hazard in Italy Part I: estimation of time-magnitude distribution model parameters and computation of probabilities of occurrence. J Seismol, **7**(2): 235~257.

Malfante M, Dalla Mura M, Metaxian J P, et al. 2018. Machine learning for volcano-seismic signals: challenges and perspectives. IEEE Signal Process Mag, **35**(2): 20~30.

Poulos J, Valle R. 2018. Missing data imputation for supervised learning. Appl Artif Intell, **32**(2): 186~196.

Sebastiani F. 2002. Machine learning in automated text categorization. ACM Computing Surveys, **34**(1): 1~47.

Svetnik V, Liaw A, Tong C, et al. 2004. Application of Breiman's random forest to modeling structure-activity relationships of pharmaceutical molecules. In: Proceedings of the 5th International Workshop on Multiple Classifier Systems. Cagliari: Springer, 334~343.

Tahir M, Grasso J R, Amorèse D. 2012. The largest aftershock: how strong, how far away, how delayed? Geophys Res Lett, **39**(4):

L04301.

Wells D L, Coppersmith K J. 1994. New empirical relationships among magnitude, rupture Length, rupture width, rupture area, and surface displacement. Bull Seismol Soc Am, **84**(4): 974 ~ 1002.

Research on Judgements of Earthquake Sequence Types Based on Machine Learning Random Forest Algorithm

Wang Jinhong[1)], Jiang Haikun[2)]

1) Department of Earth and Space Sciences, Southern University of Science and Technology, Shenzhen 518055, Guangdong, China

2) China Earthquake Networks Center, Beijing 100045, China

Abstract Utilizing the earthquake catalog from the Chinese mainland from 1970 to 2021, along with seismic sequence catalogs and historical earthquake source mechanism data, this study constructs a training and testing dataset for determining seismic sequence types. Seismic sequences are categorized into three distinct labels based on prior research: multiplet mainshocks type, mainshock-aftershock type, and isolated earthquake type. The Feature Recursive Elimination-Random Forest (RFE-RF) machine learning algorithm is employed to establish a nonlinear mapping between seismic characteristic parameters and seismic sequence types. This approach enables the early prediction of seismic sequence types at three different time nodes post-earthquake and discusses the significance of various features. The findings underscore the pivotal role of data preprocessing methods in the model's classification performance. Missing features are effectively imputed using the median value of the same sample, and the data is preprocessed using a random resampling method, yielding a high classification prediction effect. Cross-validation of the classification outcomes reveals an overall accuracy rate of 0.93 for the three types of samples one day after the earthquake. The parameters related to the main seismic source mechanism and the deviation of the *P*-axis azimuth from the local stress field are identified as having a greater classification contribution rate than traditional historical seismic sequence analogies at the immediate aftermath of an earthquake, i.e., in the absence of seismic sequence data. As the post-earthquake time progresses, sequence-related features emerge as the primary determinants of the earthquake sequence type. In the seismic sequence dataset three days post-earthquake, where the maximum aftershock has occurred, the magnitude difference between the main shock and the maximum aftershock becomes a critical factor in sequence type determination. Compared to the standalone Random Forest (RF) model, the RFE-RF model demonstrates an enhanced accuracy rate of 0.41 in the test set one day after the earthquake, indicating its superior ability to distinguish between seismic sequence types.

Keywords: Judgement of earthquake sequence types; Chinese mainland; Random forest; Recursive elimination