第 40 卷　第 2 期(355~367)　　中 国 地 震　　Vol. 40　No. 2
2024 年 6 月　　EARTHQUAKE RESEARCH IN CHINA　　Jun. 2024

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.

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

刘金平[1)]　姜立新[2)]　杨天青[2)]　刘钦[2)]

1)中国地震局地震预测研究所,北京　100036
2)中国地震台网中心,北京　100045

摘要　为快速评估地震直接经济损失,针对我国西部地区,尝试采用随机森林机器学习回归算法,以 1993—2017 年震害数据为基础,结合各年份经济数据与抗震设计数据,经特征选择与参数优化后,进行模型的训练与测试。实验结果表明,在减少模型输入特征的情况下,优化后的随机森林模型可得到更优的评估结果。通过删除含有缺失特征样本的数据预处理方法,评估模型的决定系数 R^2 达到 0.86,优于中值补齐缺失特征数据预处理下的评估模型,更适用于地震直接经济损失的评估。实例验证表明该模型评估结果与实际经济损失有较好的一致性,可为抗震救灾提供决策支持。

关键词:　地震直接经济损失　随机森林　特征选择　超参数优化

[文章编号] 1001-4683(2024)02-0355-13　　[中图分类号] P315　　[文献标识码] A

0　引言

地震具有突发性、成灾时间短、破坏波及范围广与经济损失大等特点(张培震等,2013)。地震发生后,迅速掌握灾情是重中之重,对地震灾害直接经济损失的估计是其中一个重要环节。

国外首先对地震灾害损失评估开展研究,相关工作起源于美国,由 Freeman(1932)首次开展了针对地震直接经济损失的区域损失评估研究。目前关于地震灾害直接经济损失评估已有易损性分类清单法(Algermissen et al,1984;徐国栋等,2008),该方法利用灾区详细的建筑物资料或各建筑物类型比例进行地震直接经济损失的评估;还有利用社会经济数据作为评估基础的 GDP 的经济指标方法(陈棋福等,1999;刘双庆等,2010);以及利用基于遥感影像与 GIS 的方法(Kim et al,2016;丁香等,2019;Zhang et al,2021),开发出基于地理信息系统的中国地震灾害损失评估系统等。但上述方法普遍存在工作量大且繁琐、易受环境因素影响等缺点。

[收稿日期] 2023-04-30　[修定日期] 2023-06-12
[项目类别] 国家重点研发计划(2018YFC1504506)资助
[作者简介] 刘金平,女,1997 年生,硕士研究生,研究方向为地震灾害风险评估。E-mail:ljp9709@163.com
姜立新,通讯作者,男,1966 年生,研究员,主要从事震害预测、地震应急技术等研究。E-mail:jlx@seis.ac.cn

机器学习具有在动态、大容量和复杂的数据环境中处理各种数据格式等优点。随着人工智能的发展,机器学习的各种方法更广泛地应用于地震研究中。随机森林(Random Forest,RF)作为一种主流的集成学习算法,其稳定性和准确率上均优于 AdaBoost 和 CART 算法(Jia et al,2019)。在不同样本训练量、不同验证数据量下,集成算法中随机森林的识别效果和算法稳健性高于决策树(庞聪等,2020)。

因此,本研究利用 1993—2017 年中国西部地区的震害数据,在随机森林模型的基础上,利用相关性分析得到最优特征子集,减小数据冗余,再通过交叉验证的格网搜索与随机搜索结合调整超参数,以降低学习难度,加快学习速度,提高地震直接经济损失评估精度。

1 模型研究

集成学习(Ensemble Learning)通过构建并结合多个学习器来完成学习任务(周志华,2016)。相比于单一学习器,将多个学习器结合能获得显著优越的泛化性。本研究采用由单一学习器决策树构建的集成学习模型随机森林进行地震灾害直接经济损失评估研究。

1.1 随机森林模型

1.1.1 决策树

决策树(Decision Tree,DT)是归纳学习和数据挖掘的重要方法,通常可以用来建立预测模型(杨学兵等,2007)。其目标是创建一个模型,通过在原始数据特性中的学习,总结决策规则,达到预测目标变量的目的(栾丽华等,2004)。决策树计算复杂度低,输出结果易于理解,对缺失值不敏感,但容易产生过拟合的问题,对此引入树的集成学习模型随机森林,继承决策树算法优点的同时减轻了过拟合现象,能有效提高模型的性能。

1.1.2 随机森林

随机森林算法在 2001 年被正式提出,其具备准确率高、不易过度拟合、对噪声及异常值容忍度高等优点,相对于其他方法具有较高的精度(Breiman,2001)。随机森林是 Bagging 的一个扩展变体(曹正凤等,2014),在以决策树为基学习器构建 Bagging 集成的基础上,进一步在决策树的训练过程中引入了随机属性的选择(刘永垚等,2018)。

随机森林回归概念如下:假设训练集是从随机向量 X 和 Y 分布中独立提取出来的,令 $h_i(x)$ 表示其中一个决策树的回归预测值,然后对决策树的回归预测值取平均得到随机森林回归的预测值(董红瑶等,2021),如下式所示

$$M(X)=\frac{1}{n}\sum_{i=1}^{n}h_i(x) \tag{1}$$

其中,n 为决策树个数。

1.2 Spearman 相关系数

在机器学习算法中,输入的特征数量会影响模型预测的精度(周志华,2016)。特征选择可以从原始特征集合中选择使评价准则最大化的最小特征子集,缩短原始数据获取时间,减小数据存储空间,提高模型的可解释性和模型性能(Bolón-Canedo et al,2016)。

过滤式特征选择算法中的 Spearman 相关系数通用性强,算法复杂性低,可以快速去除大量不相关的特征相关系数,该相关系数对原始数据的选取、相关形式及分布类型均无要求(Headrick,2016)。Spearman 相关系数大于 0 为正相关,小于 0 则为负相关,越接近 1 和-1

相关性越强(张文耀,2016)。其计算方法如下

$$r_s = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n}(n^2 - 1) \tag{2}$$

其中，n 为样本的数量，d_i 代表数据之间的等级差。

利用 Spearman 相关性分析方法对与地震相关的影响因素进行相关性排序，将相关性小于 0.3 的特征剔除后得到的最优特征子集输入随机森林模型，以提高评估速度，降低模型复杂度。

1.3　模型性能优化

1.3.1　交叉验证

交叉验证(Cross Validation，CV)是重复使用数据矫正模型性能的过程(Geisser，1975)，其原理为将原始数据平均分为 N 份，每次取出 1 份作为测试集，剩下 N-1 份数据作为训练集进行模型训练，最后将 N 次的平均结果作为预测误差的估计，以此衡量模型的泛化能力。本研究利用五折交叉验证方法选择参数最优模型，如图 1 所示。

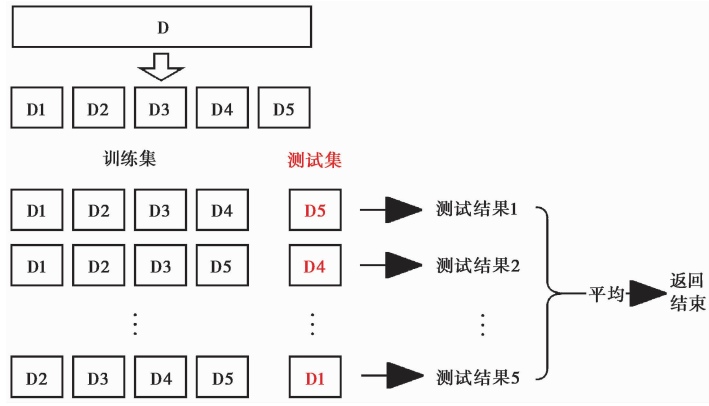

图 1　五折交叉验证

1.3.2　格网搜索

随机森林超参数众多，且默认超参数数值不一定为模型预测带来最大效益，为使模型预测达到更好的效果，需要进行超参数调优。取决定系数最大时的超参数组合作为模型输入最终结果，最大程度地提高模型的泛化能力。

需要调整的超参数如表 1 所示，其余参数选择默认值。其中，n_estimators 主要影响模型学习情况，通常数量越大，效果越好，模型越稳定，但是计算时间也会随之增加；max_depth 越小，模型越简单，但树深过大容易产生过拟合现象；max_features 限制分支时考虑的特征个数，超过限制个数的特征均会被舍弃；min_samples_split、min_samples_leaf 越小，模型越简单。上述超参数对模型的预测能力产生不同程度的影响。

交叉验证的网格搜索算法(Grid Search CV)是一种通过遍历给定的参数组合来优化模型表现的方法(王雪洁等,2022)。在指定的参数范围内，按步长依次调整参数，利用调整的参数训练学习器，从所有参数中找到在验证集上精度最高的参数组合。由于格网搜索需要

表 1 模型超参数

超参数	描述	默认值
n_estimators	模型中树木的数量	100
max_depth	模型中树的最大深度	None
max_feature	最大特征数	auto
min_samples_split	拆分内部节点所需的最少样本数	2
min_samples_leaf	在叶节点处需要的最小样本数	1

遍历所有可能的参数组合,非常耗时。因此,首先采用交叉验证的随机搜索(Randomized Search CV)进行参数粗选择,再利用网格搜索在最优解附近范围进行参数的遍历,最终得到模型参数最优解(温博文等,2018)。

2 研究过程与分析

2.1 数据分析

2.1.1 震害数据

本研究使用 1993—2017 年中国西部地区震害数据,共 251 例数据,剔除边境地震数据以及震害直接经济损失极小的数据,经过筛选最后使用的数据为 236 例。西部地区包括重庆市、四川省、贵州省、云南省、西藏自治区、陕西省、甘肃省、青海省、宁夏回族自治区、新疆维吾尔族自治区,其震中分布如图 2 所示,数据分布统计结果见表 2。

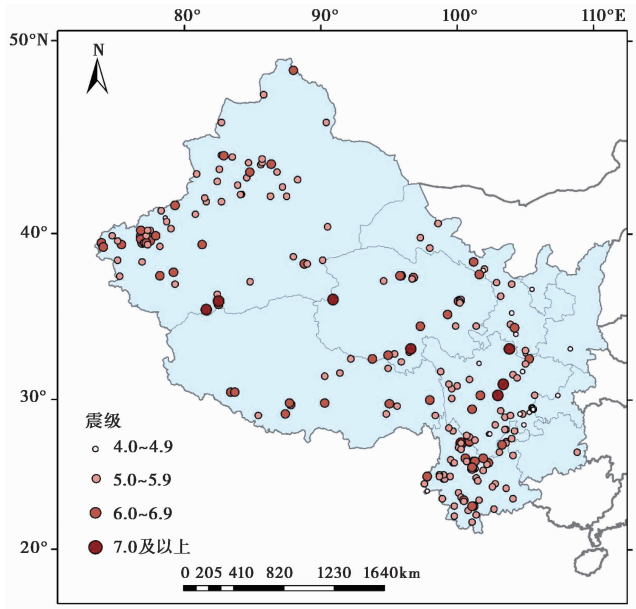

图 2 1993—2017 年西部地区震中分布

2.1.2 特征参数

地震灾害发生后,地震产生的经济损失受多方面因素的影响,按地震本身产生的影响、经济方面的影响、区域因素的影响三方面因素进行统计。研究选取的影响因素共 43 个,如

表 2　　**西部地区震级样本分布**

震级	样本数量	震级	样本数量	震级	样本数量	总样本数
$M<4.0$	0	$5.0\leqslant M<6.0$	152	$7.0\leqslant M<8.0$	6	236
$4.0\leqslant M<5.0$	20	$6.0\leqslant M<7.0$	56	$8.0\leqslant M<9.0$	2	

表 3 所示。其中地震影响因素，如发震时间及地点、地震震级、震中烈度、震源深度、建筑物破坏、受灾人口、受灾范围等数据主要来自于《中国大陆地震灾害损失评估报告汇编》(中国地震局监测预报司，2001；中国地震局震灾应急救援司，2010、2015)；区域抗震能力中设防烈度、地震分组等指标取自各版本 GB 50011—2010《建筑抗震设计规范》(中华人民共和国住房和城乡建设部等，2010)；经济、财政指标等社会发展指标来自国家统计局网站[①]以及各地方统计年鉴。

表 3　　**地震直接经济损失的影响因素**

特征类型	影响因素
地震	时间、地点(经度、纬度(°))、震级、极震区烈度、震源深度(km)、受伤人数、死亡人数、房屋破坏情况(毁坏、严重及以上、中等及以上、轻微及以上(m^2))、各烈度区面积(Ⅵ及以上、Ⅶ及以上、Ⅷ及以上、Ⅸ(km^2))、受灾人口(万人)
经济	人口出生率(‰)、死亡率(‰)、人口自然增长率(‰)、地区总人口(万人)、人口密度(人/km^2)、年 GDP(亿元)、第一产业(亿元)、第二产业(亿元)、第三产业(亿元)、居民消费水平(亿元)、城镇居民消费水平(亿元)、农村居民消费水平(亿元)、居民消费价格指数(%，以上一年为 100)、城镇居民消费价格指数(%，以上一年为 100)、农村居民消费价格指数(%，以上一年为 100)、地方财政收入(亿元)、地方财政税收收入(亿元)、地方财政支出(亿元)、城镇居民可支配收入(亿元)、农村居民可支配收入(亿元)、城镇居民人均消费性支出(亿元)、农村居民人均消费性支出(亿元)、社会消费品零售总额(亿元)
区域	抗震设防烈度、设计基本地震加速度、乡镇数

2.2　研究过程

本研究共分四个步骤，第一步为数据预处理，分别利用中值补齐缺失特征与删除含有缺失特征样本两种方法，解决数据存在缺失值和数据不平衡问题；第二步，特征选择及超参数调优，利用 Spearman 相关性分析计算各个特征与地震直接经济损失的相关性，得到最优特征子集，按照 8∶2 将数据集划分为训练集和测试集，利用交叉验证的格网搜索和随机搜索相结合的方式得到最优超参数组合；第三步，模型训练，利用最优特征子集及最优超参数组合进行模型训练，得到地震直接经济损失评估模型；第四步，精度评定，利用回归分析评价指标评价模型精度。

2.2.1　评价指标

选取平均绝对误差(MAE)、均方根误差(RMSE)和决定系数(R^2)来衡量模型的精度，如表 4 所示。R^2 越接近 1，同时 MAE、RMSE 越小，表明模型评估精度越高。

2.2.2　数据预处理

由于各年份震害数据撰写年份不同，数据收集不完整，导致数据中存在缺失值，对西部

① https://www.stats.gov.cn/

表 4 **评价指标**

评价指标	定义	公式
MAE	平均绝对误差	$\mathrm{MAE}=\frac{1}{N}\sum_{i=1}^{N}\lvert y_i-\hat{y}_i\rvert$
RMSE	均方根误差	$\mathrm{RMSE}=\sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i-y_i)^2}$
R^2	决定系数	$R^2=1-\frac{\sum_{i=1}^{N}(y_i-\hat{y}_i)^2}{\sum_{i=1}^{N}(y_i-\bar{y})^2}$

注：y_i为真实值，$\hat{y}$ 为评估值。

地区震害数据缺失值进行统计，结果如图 3 所示。其中数据完备程度的范围为 0～1，数据越完备，数值越接近 1。

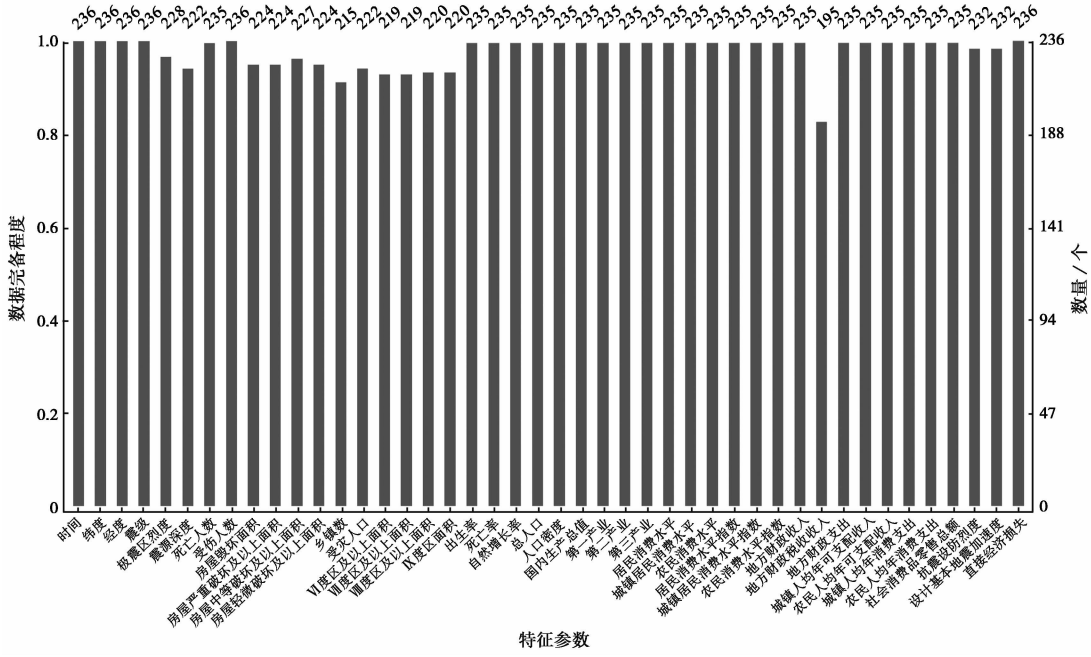

图 3 缺失值统计

图 3 中地方财政税收收入缺失值利用前后两年的平均增长率计算得到。对于其他缺失值，使用两种方法进行处理。方法一：利用中值填补缺失特征样本，将数据分为 $M_S<5.0$、$5.0\leqslant M_S<6.0$、$6.0\leqslant M_S<7.0$、$7.0\leqslant M_S$四类，分别利用中值补齐缺失特征；方法二：删除含有缺失特征样本。将两种方法处理后的数据分别带入模型进行训练。

由于时间跨度较大，货币价值发生变化，导致直接经济损失没有可比性，使用居民消费水平指数（CPI）进行换算（朱达逦等，2021），以 1978 年（CPI 为 100）作为标准，计算得到后续年份的 CPI 值，见表 5。本研究通过使用国家统计局统计公布的 CPI 对经济损失进行折

算，其公式为

$$L_A = L_B \times \frac{\mathrm{CPI}_A}{\mathrm{CPI}_B} \tag{3}$$

其中，L 为地震损失，下角标 A、B 为年份。将 1993—2017 年的直接经济损失按公式换算为 2017 年水平，直接经济损失经过换算后，提高了数据可对比性，增加了科学客观性。

表 5　**我国历年 CPI**

年份	CPI	年份	CPI	年份	CPI	年份	CPI
1978	100.0	2000	434.0	2008	522.7	2016	627.5
1993	273.1	2001	437.0	2009	519.0	2017	637.5
1994	339.0	2002	433.5	2010	536.1	2018	650.9
1995	396.9	2003	438.7	2011	565.0	2019	669.8
1996	429.9	2004	455.8	2012	579.7	2020	686.5
1997	441.9	2005	464.0	2013	594.8	2021	692.7
1998	438.4	2006	471.0	2014	606.7	2022	706.6
1999	432.2	2007	493.6	2015	615.2	2023	721.4

2.2.3　特征选择

在已有研究中，大多凭借经验法选择特征参数，建立地震直接经济损失评估模型（李云飞等，2021；赵士达等，2016），缺少客观性。本研究利用 Spearman 相关性分析对 43 种影响因素进行相关性排序，将相关性小于 0.3 的特征（图 4 中橙色部分）剔除后得到的最优特征子集输入模型，以提高评估速度，降低模型复杂度。两种不同数据预处理方法得到的 spearman 相关性排序结果如图 4 所示，由图中结果可以看出两种方法结果相似，排名靠前的分别为房屋破坏情况、各烈度区面积以及人员伤亡情况。

2.2.4　超参数调优

根据前文提到模型参数优化方法，设置参数取值范围，利用本文方法对最优特征子集进行参数寻优，具体情况如表 6 所示。

2.3　结果分析

本研究在两种不同数据预处理方法下进行特征选择和超参数调优，得到模型最优参数组合，最终得到关于中国西部地区地震直接经济损失评估的两种训练模型，由中值填充缺失特征的数据预处理方法训练得到的模型称为模型Ⅰ，由删除缺失特征样本的数据预处理方法训练得到的模型称为模型Ⅱ。对两种模型的特征重要性和模型精度两方面进行结果分析。

2.3.1　特征重要性排序

经特征选择与超参数调优后，西部地区的特征重要性排序结果如图 5 所示，由图可见两种模型得到的特征重要性排序结果基本一致。整体看来，房屋受损情况、受灾人口、各烈度区面积以及极震区烈度排名靠前，相对于地震本身产生的影响，经济方面的因素重要性排序比较靠后。房屋中等破坏及以上面积占据影响直接经济损失的主导地位。西部地区高烈度区面积对地震直接经济损失影响大于低烈度区面积，但将Ⅸ度区面积判定为不相关特征或相关性较小特征，其原因是西部地区较少发生破坏性较强的地震，产生较大烈度区的比例低，特征数据量小将影响结果的准确性。

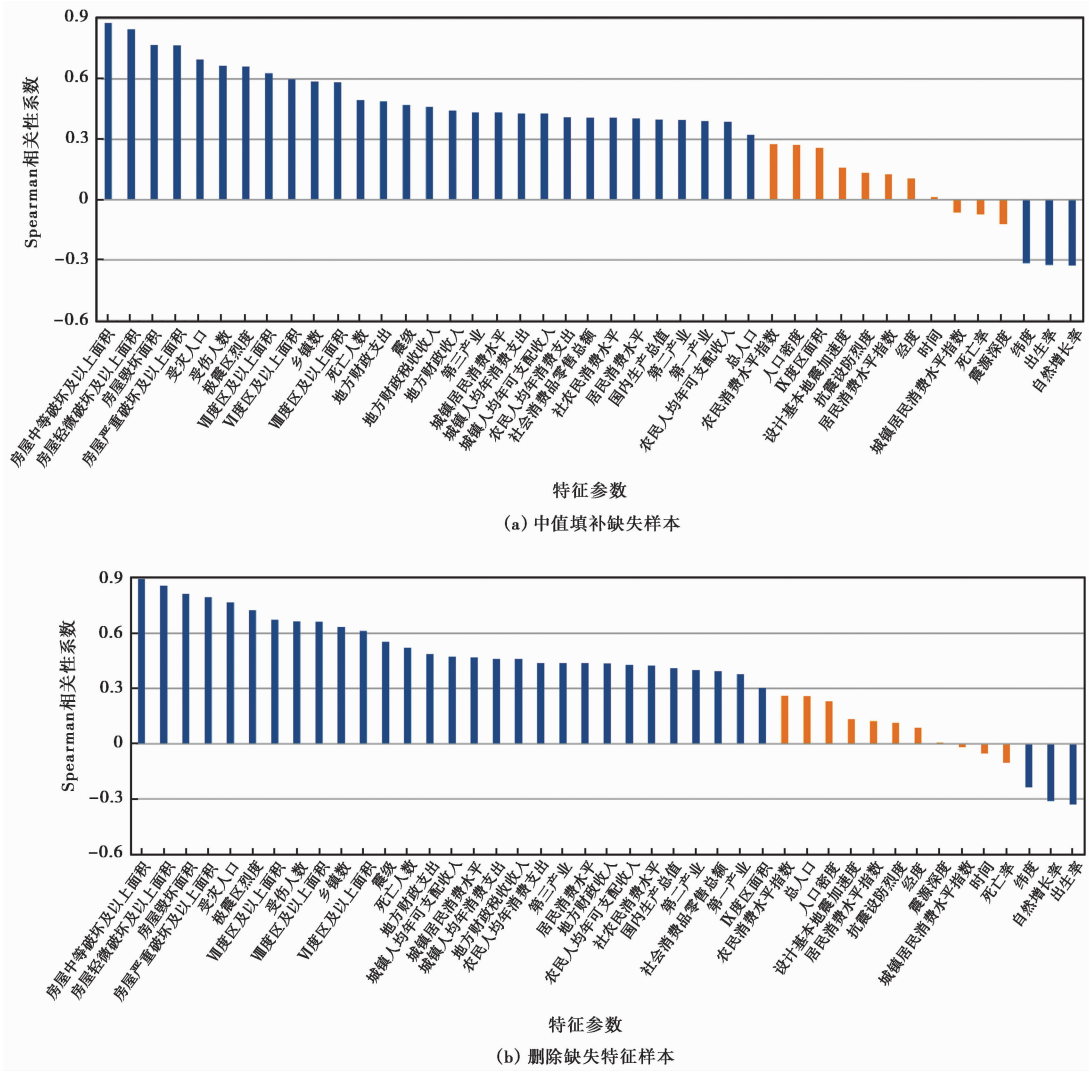

图 4 西部地区 spearman 相关性排序

表 6 **参数设置**

超参数	取值范围	中值填补缺失样本	删除缺失特征样本
n_estimators	(10,500,num=60)	356	86
max_depth	(1,100,num=50)	90	68
max_feature	['sqrt','auto']	sqrt	sqrt
min_samples_split	[2,5,10]	2	2
min_samples_leaf	[1,2,4,8]	2	1

2.3.2 精度评价

根据本研究所提出的地震直接经济损失评估模型优化方法,得到两种训练模型,其各项评价指标情况如图 6 所示。优化后的随机森林在西部地区模型评估精度有所提高,反映出特征选择和超参数优化对提高随机森林模型评估精度有积极的影响。在特征数减少三分之

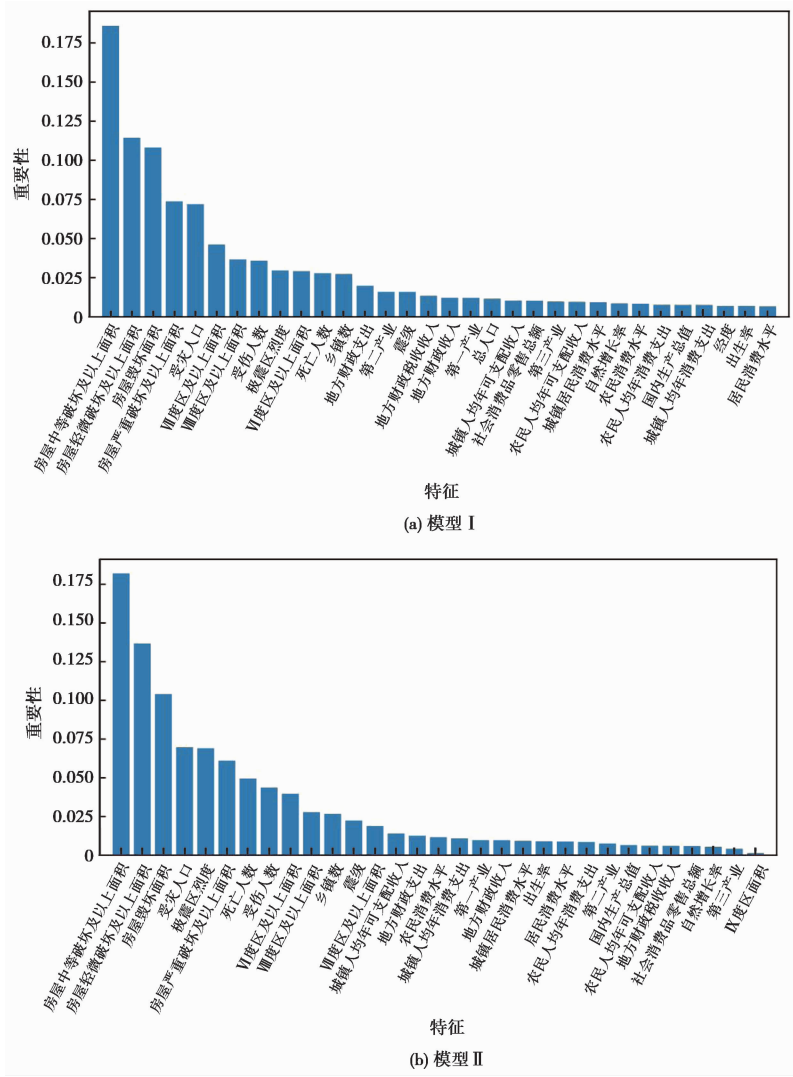

图 5　西部地区特征重要性排序

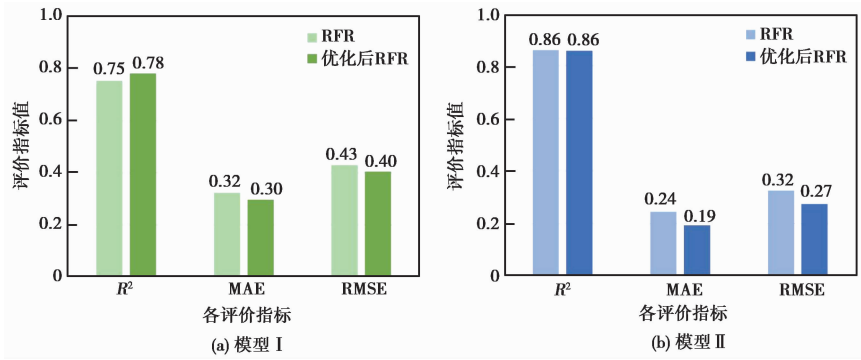

图 6　模型评估效果对比

一的情况下,依然得到较高精度,表明相关性较低的特征对模型训练基本没有影响,将其删除后反而可以提升评估精度与效率。另外,模型Ⅰ中经本文方法优化后的模型(优化后RFR)的决定系数 R^2 为 0.78,相较于普通随机森林回归模型(RFR)提升了 4%,MAE 和 RMSE 分别下降了 7.9%和 13.1%;模型Ⅱ中优化后 RFR 的决定系数 R^2 为 0.86,MAE 和 RMSE 分别下降 20.8%和 15.6%。模型Ⅰ精度较低,考虑其原因为特征缺失数据占比较小,中值填充后反而增大了误差。

两种预处理方法得到的西部地区评估值与真实值的拟合情况及误差分布情况,如图 7 所示,其中对于地震直接经济损失值取对数处理。可以看出,两种数据预处理方法下的模型均达到了较好的评估效果,误差分布情况符合正态分布规律。由误差分布情况可见,两种方法均存在评估误差较大的数据,但相较于模型Ⅰ,模型Ⅱ有更小的模型误差与评估精度。

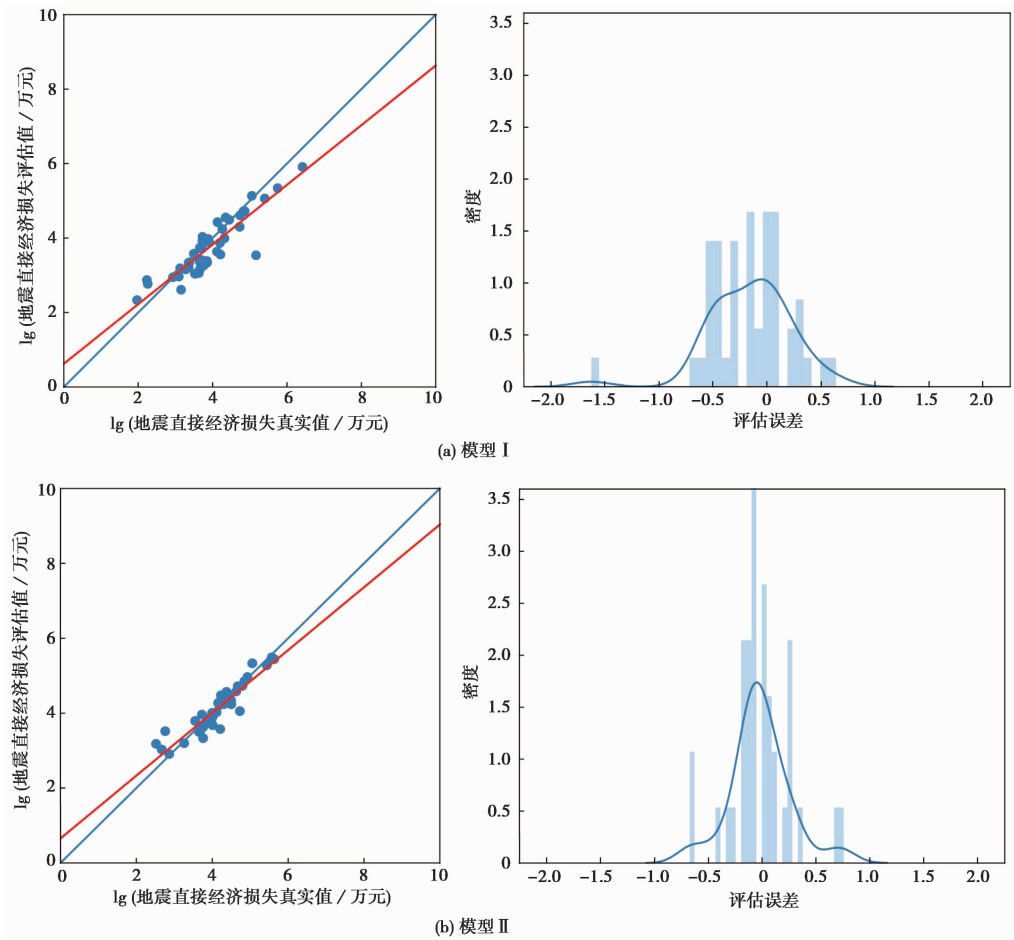

图 7 评估值分布与误差分布

2.4 实例应用

针对中国西部地区地震直接经济损失评估模型,选取 2020 年 1 月 16 日新疆库车市 5.6 级、2020 年 1 月 19 日新疆伽师 6.4 级、2022 年 1 月 8 日青海门源 6.9 级和 2022 年 6 月 1 日

四川芦山 6.1 级地震进行实例验证，将各个特征参数输入模型，利用模型Ⅱ进行地震灾害直接经济损失评估，模型评估结果如表 7 所示。

表 7　　模型Ⅱ评估结果对比

震例	lg(真实值/万元)	lg(评估值/万元)	MAE
新疆库车 5.6 级地震	2.82	3.46	0.34
新疆伽师 6.4 级地震	5.15	5.22	
青海门源 6.9 级地震	5.46	4.83	
四川芦山 6.1 级地震	5.50	5.49	

注：lg 表示地震直接经济损失取对数处理。

从表 7 可以看出，运用优化后的随机森林模型对实际地震进行直接经济损失评估，能达到良好的评估精度，评估值与实际值较为吻合，表明模型Ⅱ可以用于地震灾害直接经济损失的评估。

3　结果与讨论

本研究基于随机森林算法，在两种数据预处理方法下，利用 Spearman 相关性分析方法对中国西部地区震害数据进行特征选择，采用格网搜索与随机搜索相结合的方法进行参数调优，对随机森林算法进行了优化，减少了特征数量，使得对地震灾害的直接经济损失的评估精度和效率均有所提高，得到了特征的重要性排序，初步结论如下：

(1)对于相同原始数据，利用不同数据预处理方法会影响模型训练结果，中值填充缺失特征后模型精度低于删除缺失特征数据，考虑其原因为数据量较小，缺失数据较多，导致中值填充后数据质量降低，影响模型评估精度。实例应用结果表明，在删除缺失特征数据预处理方法下的西部地区随机森林模型具有一定的准确性，适用于对地震灾害经济损失进行评估研究。

(2)改进的随机森林算法利用更少的特征得到了更高的评估精度与评估效率，表明特征冗余对模型评估精度有一定影响，参数调整也会进一步影响模型训练，体现出特征选择与超参数调优的必要性。

(3)特征选择及其评估结果表明，西部地区房屋受损情况、受灾人口、各烈度区面积以及极震区烈度均对震害直接经济损失有较大影响，其中房屋的损失情况影响最大，其原因可能为西部地区房屋抗震性能较差。特征重要性排序结果可为实际房屋建设提供指导，相关部门在房屋抗震结构的设计上应给予重视，加强抗震措施，提高抗震意识。

参考文献

曹正凤，谢邦昌，纪宏. 2014. 一种随机森林的混合算法. 统计与决策，**30**(4)：7~9.

陈棋福，陈凌. 1999. 地震损失预测评估中的易损性分析. 中国地震，**15**(2)：97~105.

丁香，王晓青，窦爱霞，等. 2019. 基于格网的全国尺度地震灾害损失预测系统设计与实现. 中国地震，**35**(2)：238~247.

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.

李云飞，许才顺，池招招，等. 2021. 基于 Softmax 回归模型的地震灾害损失预测评估研究. 合肥工业大学学报(自然科学版)，**44**(12)：1676~1681.

刘双庆,邱虎,王晓青. 2010. 一种基于宏观经济指标的地震灾害快速评估方法及实现. 灾害学,**25**(3):16~19,31.

刘永垚,第宝锋,詹宇,等. 2018. 基于随机森林模型的泥石流易发性评价——以汶川地震重灾区为例. 山地学报,**36**(5):765~773.

栾丽华,吉根林. 2004. 决策树分类技术研究. 计算机工程,**30**(9):94~96,105.

庞聪. 江勇. 廖成旺,等. 2020. 基于机器学习的强震动监测环境抗干扰方法对比研究. 内陆地震,**34**(2):119~124.

王雪洁,施国萍,周子钦,等. 2022. 基于随机森林算法对 ERA5 太阳辐射产品的订正. 自然资源遥感,**34**(2):105~111.

温博文,董文瀚,解武杰,等. 2018. 基于改进网格搜索算法的随机森林参数优化. 计算机工程与应用,**54**(10):154~157.

徐国栋,方伟华,史培军,等. 2008. 汶川地震损失快速评估. 地震工程与工程振动,**28**(6):74~83.

杨学兵,张俊. 2007. 决策树算法及其核心技术. 计算机技术与发展,**17**(1):43~45.

张培震,邓起东,张竹琪,等. 2013. 中国大陆的活动断裂、地震灾害及其动力过程. 中国科学:地球科学,**43**(10):1607~1620.

张文耀. 2016. 用斯皮尔曼系数衡量网络的度相关. 合肥:中国科学技术大学.

赵士达,张楠,张斯文,等. 2016. 基于 LM-BP 神经网络的地震直接经济损失快速评估方法研究. 地震研究,**39**(03):500~506,528.

中国地震局监测预报司. 2001. (1996~2000)中国大陆地震灾害损失评估汇编. 北京:地震出版社.

中国地震局震灾应急救援司. 2010. 2001—2005 年中国大陆地震灾害损失评估汇编. 北京:地震出版社.

中国地震局震灾应急救援司. 2015. 2006—2010 年中国大陆地震灾害损失评估汇编. 北京:地震出版社.

中华人民共和国住房和城乡建设部,中华人民共和国国家质量监督检验检疫总局. 2010. GB 50011-2010 建筑抗震设计规范. 北京:中国建筑工业出版社.

周志华. 2016. 机器学习. 中国民商,**3**(21):93.

朱达邈,王东明. 2021. 基于贝叶斯网络的地震直接经济损失预估模型在灾评推演训练中的应用. 自然灾害学报,**30**(5):51~63.

Algermissen S T,Steinbrugge K V. 1984. Seismic hazard and risk assessment:some case studies. Geneva Pap Risk Insur Issues Pract,**9**(1):8~26.

Bolón-Canedo V,Sánchez-Maroño N,Alonso-Betanzos A. 2016. Feature selection for high-dimensional data. Prog Artif Intell,**5**(2):65~75.

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

Freeman J R. 1932. Earthquake Damage and Earthquake Insurance:Studies of a Rational Basis for Earthquake Insurance; Also Studies of Engineering Data for Earthquake-Resisting Construction. New York:McGraw-hill.

Geisser S. 1975. The predictive sample reuse method with applications. J Am Stat Assoc,**70**(350):320~328.

Headrick T C. 2016. A note on the relationship between the Pearson product-moment and the Spearman rank-based coefficients of correlation. Open J Stat,**6**(6):1025~1027.

Jia H X,Lin J Q,Liu J L. 2019. An earthquake fatalities assessment method based on feature importance with deep learning and random forest models. Sustainability,**11**(10):2727.

Kim H S,Chung C K. 2016. Integrated system for site-specific earthquake hazard assessment with geotechnical spatial grid information based on GIS. Nat Hazards,**82**(2):981~1007.

Zhang Y X,Zheng S S,Sun L F,et al. 2021. Developing GIS-based earthquake loss model:a case study of Baqiao District,China. Bull Earthquake Eng,**19**(5):2045~2079.

Research on Assessment of Direct Economic Losses of Earthquake Disasters Based on Random Forest—A Case Study of the Western Region

Liu Jinping[1)], Jiang Lixin[2)], Yang Tianqing[2)], Liu Qin[2)]

1) Institute of Earthquake Forecasting, CEA, Beijing 100036, China

2) China Earthquake Networks Center, Beijing 100045, China

Abstract In this study, we aim to expedite the assessment of direct economic losses induced by earthquakes, focusing on China's western region. We employ a random forest machine learning regression algorithm for this purpose. Leveraging earthquake damage data spanning from 1993 to 2017, in conjunction with economic and seismic design data from various years, we train and test the model following feature selection and parameter optimization steps. The findings reveal that the optimized random forest model yields superior evaluation outcomes while reducing the model's input features. Specifically, the evaluation model achieves an R^2 value of 0.86 under the data preprocessing method involving the deletion of missing feature samples, surpassing the evaluation model's performance under the median filling missing feature data preprocessing approach. This optimized model proves more suitable for assessing direct economic losses attributable to earthquakes. Validation using real world examples demonstrates that the evaluation results derived from this model align closely with actual economic losses, underscoring its utility in providing decision support for earthquake relief efforts.

Keywords: **Earthquake direct economic loss**; **Random forest**; **Feature selection**; **Hyper parameter optimization**