doi: 10.11720/wtyht.2020.1501
彭刘亚,解惠婷,冯伟栋.基于随机森林算法的砂土液化预测方法[J].物探与化探,2020,44(6):1429-1434.http://doi.org/10.11720/wtyht.2020.1501
Peng L Y,Xie H T,Feng W D.The method of predict sand liquefaction based on random forest algorithm[J].Geophysical and Geochemical Exploration,2020,44(6):1429-1434.http://doi.org/10.11720/wtyht.2020.1501

基于随机森林算法的砂土液化预测方法

彭刘亚,解惠婷,冯伟栋
(安徽省地震局 安徽省地震工程研究院,安徽 合肥 230031)

摘 要: 砂土液化的影响因素较多且复杂。以唐山大地震的 72 个场地的实测液化样本数据为例,在不丢失任何信息的前提下,选取了 8 个砂土液化的判别指标,通过计算样本数据的 Gini 系数,采用 CART 算法的决策树对数据的特征属性进行划分。在此基础之上,通过增加多个决策树构造随机森林的方式,在一定程度上降低了单个决策树学习过度造成的过拟合风险,同时,通过 10 轮交叉验证的方式确定了决策树的最大高度为 5,随机森林中决策树的个数为 20 时,模型的效果达到最佳。研究结果表明,与抗震设计规范中的标贯试验法判别公式相比,决策树模型和随机森林模型的训练结果和预测结果有显著提高,尤其是随机森林模型在训练样本和预测样本上均没有出现误判,稳定性更高。

关键词: 砂土液化;判别指标;决策树;随机森林
中图分类号: P631.4 **文献标识码**: A **文章编号**: 1000-8918(2020)06-1429-06

0 引言

砂土液化通常是指在地震作用下,饱和砂土(或粉土)由于有效应力减小所导致的从固态到液态的变化现象[1]。发生较大震级的地震时,砂土液化容易引起地基承载力降低,造成地面沉陷、滑坡、冒水喷砂、建筑物受损等灾害[2],如 1975 年海城 7.3级地震、1976 年唐山 7.8 级地震、1978 年日本 Miyagiken-okij7.4 级地震、1995 年日本阪神 7.3 级地震、2008 年汶川 8.0 级地震[3-4]中均出现了大面积的砂土液化现象。因此,研究砂土液化的影响因素,建立合理的预测模型,快速判断是否存在砂土液化现象,在一定程度上能够有效地防治砂土液化带来的地震地质灾害。

传统液化判别和危害程度评价方法大多是基于宏观地震灾害现象资料,结合现场试验和室内试验结果,通过总结分析和统计得出的一般规律[5]。如根据剪切波速法、标准贯入法及静力触探法等得出的结果与规范中给出的临界值比较,从而判别是否液化。国内外用于砂土液化的判别方法种类繁多,但由于砂土液化的影响因素多且复杂,因此每种方法都有一定的适用范围和局限性。

砂土液化问题本质上可视为机器学习中的分类问题。近些年来,更多的国内外学者在理论方法和实测数据的基础上,综合多个砂土液化的影响因子,采用不同的分类算法研究砂土液化判别问题。如人工神经网络[6]、支持向量机[7]、距离判别法[8]、Fisher 判别模型[9]等方法都被应用到砂土液化预测中。但由于地震作用的随机性、土层参数的多样性,以及没有足够多的样本数据支撑,使得这些算法均存在一定程度上的局限性,如容易陷入局部极小值,或存在过拟合现象。笔者整理了唐山大地震的砂土液化现场资料,选取了其中的 72 个场地的实际数据,采用机器学习中的随机森林分类算法,并通过数据留出集与交叉验证的测试方式降低模型的过度学习能力,防止出现过拟合现象,一定程度上提高了模型预测的稳定性。

收稿日期: 2019-11-25; **修回日期**: 2020-08-31
基金项目: 中国地震局三结合课题(3JH202002013)
作者简介: 彭刘亚(1990-),男,工程师,主要从事工程地震及地震灾害现场调查等方面的研究工作

1 随机森林算法的基本原理

随机森林算法是机器学习当中比较常用的分类算法,包含多个决策树的分类器,并且其输出的类别是由个别树输出的类别的众数而定[10]。因此,有必要简单地介绍下决策树的分类算法和原理。

1.1 决策树分类原理

决策树(decision tree,DT)是一种常用的分类方法,它通过将大量无规则无次序的数据集进行分类、聚类和预测建模,构造树状结构的分类规则,从而对样本进行分类或预测[11]。图 1 为单个决策树二分类模型的示意,图中最顶端的为根节点,包含了所有样本数据,根据该根节点的某一个属性将数据分成中间层的子节点;以此类推,自上而下,从而划分数据的所属类别,即叶子节点。因此,构造决策树的关键在于在当前状态下选取合适的属性作为划分数据类别的节点,按照一定的目标函数(如信息熵、Gini 系数等)下降最快的方式到达叶子结点,从而对数据类别进行最终判断。

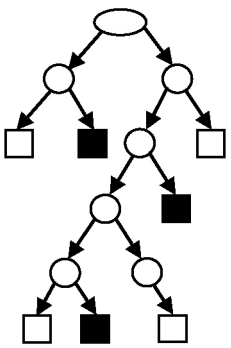

图 1 决策树分类模型示意

Fig.1 Decision tree classification model sketch

常见的决策树算法主要有基于信息熵的 ID3 算法、基于信息增益比的 C4.5 算法和基于 Gini 系数的 CART 算法。其中 C4.5 算法在 ID3 算法基础之上,用信息增益比替代了信息增益,改善了 ID3 算法由于信息增益在可取数值数目较多的属性上存在的倾向性问题。

本文采用的是二叉树 CART 算法,该算法主要以 Gini 系数作为分裂标准,选择具有最小 Gini 系数的属性作为节点,节点处的 Gini 系数值越小,说明该节点数据类别越少,数据集不纯度越低,越有利于划分类别。因此构造 CART 决策树的过程实质上就是层层递归直到某节点 Gini 系数最低,即认为该节点可视为叶子节点,从而对样本数据集进行分类。

假设数据集 $D(X,Y)$ 包含 m 个类别的样本,样本数量为 K,则数据集 D 的 Gini 系数值定义为[10]:

$$\mathrm{Gini}(t) = 1 - \sum_{i=1}^{m} [p(i \mid t)]^2, \qquad (1)$$

式中,$p(i|t)$ 表示节点 t 处当前数据集中类别 i 的概率。Gini 值直观地反映了从数据集 D 中随机抽取两个样本,其类别标记不一致的概率,因此,该值越小,则表示数据集 D 的纯度越高。而 CART 算法将数据集 D 按照某个特征 A 划分为两个子数据集 D_1 和 D_2,则此时在特征 A 条件下,数据集 D 的 Gini 系数值定义如下:

$$\mathrm{Gini}(D,A) = \frac{D_1}{D}\mathrm{Gini}(D_1) + \frac{D_2}{D}\mathrm{Gini}(D_2) \text{。} \qquad (2)$$

CART 算法根据某个特征取值下当前数据集的最小 Gini 系数将数据集划分,生成左右两个子节点,再分别对两个子节点当中的数据集执行相同操作,层层递归,直至叶子节点。不难发现,当决策树的高度无限制地生长时,必然能使得 Gini 系数为零,此时一定可以将训练数据当中的每个样本都能精确地划分类别。因此,不可避免地带来模型的过拟合问题,使得模型的预测能力下降或不稳定。

1.2 随机森林分类原理

随机森林(random forest,RF)是在决策树分类器的基础之上,通过随机有放回采样对数据集当中的样本以及特征进行选取,构造多个决策树,并由各决策树分类结果的众数决定最终的类别划分,从而降低单个决策树的过拟合风险。图 2 为随机森林的分类原理示意,主要包括了以下几个步骤[10]:

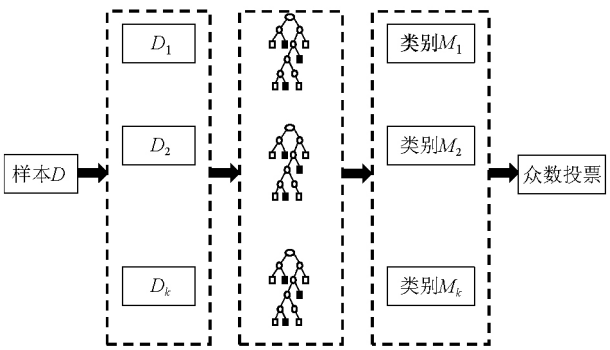

图 2 随机森林分类模型示意

Fig.2 Random forest classification model sketch

1)从包含 M 个特征的总样本数据集 D 中采用有放回采样随机选取 k 个子训练集(D_1,D_2,…,D_k),用于构造 k 个决策树。

2)对每个决策树的每个节点,随机选取 n 个特征(n 应小于 M)计算当前的 Gini 系数作为分裂子节点的优选特征,让决策树完整生长,直至 Gini 系数最小到达叶子节点。

3)遍历所有决策树,得到每个决策树的分类结果,采取众数投票结果作为最终的分类模型,对未知数据进行预测。

2　随机森林砂土液化预测模型

2.1　砂土液化判别指标

影响砂土液化的因素比较多且复杂,主要包括动荷条件、埋藏条件和土性条件三个方面[1]。本次根据文献[4]中唐山大地震的砂土液化现场实测数据选取了地震烈度 $I(x_0)$、震中距 $L(x_1)$、平均粒径 $D_{50}(x_2)$、不均匀系数 $C_u(x_3)$、地下水位 $d_w(x_4)$、砂层埋深 $d_s(x_5)$、标准贯入击数 $N_{63.5}(x_6)$、剪应力与有效上覆应力比 $\tau_d/\sigma'_V(x_7)$ 这 8 个影响因素作为砂土液化的判别指标,并整理出 72 个砂土液化的实测样本数据,其中液化场地 47 个,不液化场地 25 个,并分别用“0”和“1”表示不液化与液化两种现象。

2.2　数据预处理

2.2.1　数据标准化

由于各指标之间的量级差异比较明显,因此需要进行标准化处理以消除量纲的影响。本次采用式(3)所示的 z-score 法进行标准化:

$$x_{i\_z}=\frac{x_i-\mu}{\sigma}, \tag{3}$$

式中:μ 和 σ 分别为样本均值和标准差。标准化之后的数据无量纲,均值为 0,标准差为 1。

2.2.2　数据集划分

为了提高模型的泛化能力,避免由于样本量过少带来的过拟合现象,本次将 72 个样本划分成训练样本集(64 个样本,见表 1,由于篇幅限制,这里仅给出部分数据)和测试样本集(8 个样本,见表 2)。其中测试样本集不参与决策树和随机森林的学习训练过程,仅作为未知样本验证模型的预测能力。

2.2.3　学习过程

对于随机森林算法中的单棵决策树来说,对当前数据集选择某一种属性特征计算 Gini 系数作为分裂子节点的优选特征。图 3 所示为砂土液化预测单棵 6 层决策树模型。首先,选定归一化后的标准贯入击数 $N_{63.5}(x_6)$ 作为根节点,以 $x_6\leqslant -0.054$ 为判定条件计算 Gini 系数为 0.451,此时 64 个样本被划分为两类,分别为 22 个和 42 个。以此为根节点,生长决策树,产生第二层子节点,其中左节点以 $x_4\leqslant 0.129$ 为判定条件,选择地下水位 d_w 为优选特征,样本分别为 5 个和 35 个,而右节点选择以 $x_7\leqslant 1.473$ 为判定条件,选择剪应力与有效上覆应力比为优选特征,样本分别为 17 个和 7 个;两个子节点的 Gini 系数分别为 0.219 和 0.413。相对于根节点的数据集而言,不纯度降低,说明决策树的生长方向是有利

表 1　砂土液化训练样本集

Table 1　Training dataset of sand liquefaction

序号	判别指标								液化情况
	I	L/km	D_{50}/mm	C_u	d_w/m	d_s/m	$N_{63.5}$/击	τ_d/σ'_V	
1	7	68.6	0.410	2.90	1.09	4.15	5	0.1000	1
2	7	83.3	0.187	4.00	1.20	2.45	8	0.0900	1
3	7	83.3	0.111	2.02	0.80	1.35	6	0.0800	1
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
19	7	79.0	0.120	1.55	1.37	3.60	19	0.0940	0
20	7	81.2	0.160	2.67	1.05	4.30	12	0.1050	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
26	8	116.4	0.200	2.70	1.60	8.70	8	0.2120	1
27	8	116.4	0.170	1.91	3.30	5.80	5	0.1600	1
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
43	8	70.9	0.300	2.43	2.30	12.30	13	0.2030	0
44	8	47.0	0.310	2.42	2.00	3.46	8	0.1630	0
45	8	117.0	0.073	7.50	1.53	11.90	26	0.2170	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
50	9	22.0	0.200	1.94	0.43	2.61	10	0.4620	1
51	9	22.0	0.240	2.08	1.15	4.50	22.2	0.4150	1
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
62	9	14.0	0.160	2.25	4.90	9.38	61	0.3180	0
63	9	9.6	0.210	3.15	3.50	8.35	31	0.3470	0
64	9	11.0	0.160	2.76	4.50	4.50	22	0.2480	0

表 2 砂土液化测试样本集

Table 2 Test dataset of sand liquefaction

序号	判别指标								液化情况
	I	L/km	D_{50}/mm	C_u	d_w/m	d_s/m	$N_{63.5}$/击	τ_d/σ'_V	
1	7	76.8	0.166	1.65	0.50	1.70	3	0.1000	1
2	7	60.8	0.360	3.30	1.59	6.65	23	0.1030	0
3	7	70.0	0.145	8.50	0.85	1.80	2	0.0890	1
4	7	49.0	0.140	2.31	1.00	4.80	14	0.1080	0
5	7	81.2	0.140	1.60	1.40	4.35	9	0.1000	1
6	8	116.0	0.265	2.81	3.30	13.80	17	0.1900	0
7	8	117.4	0.134	2.23	3.20	7.20	8	0.1720	1
8	9	17.0	0.185	1.90	0.61	3.80	4	0.4580	1

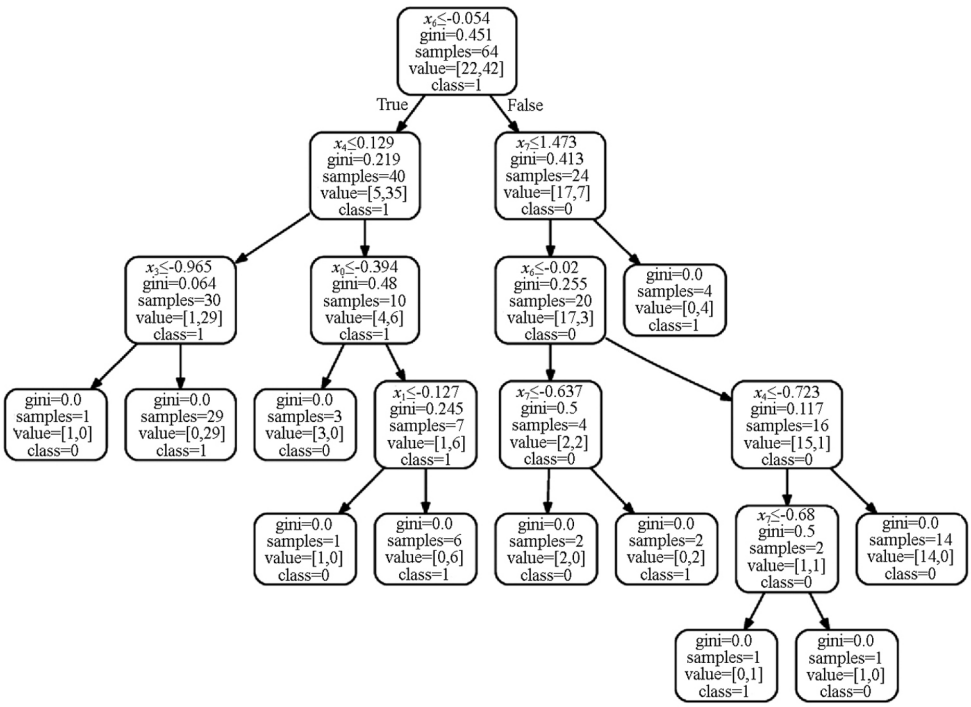

图 3 决策树分类过程示意

Fig.3 Decision tree classification process

于类型划分的。以此类推,直到数据集的 Gini 系数为零,决策树终止生长。不难发现,随着决策树的高度增加,当前数据集的样本量也在减少,因此需要注意的是:当不对决策树的高度和当前节点的最小样本量加以控制时,决策树的规模和计算量会相应地增加,虽然最终能够对每一个样本进行准确地类型划分,但不可避免地增加了过拟合的风险。因此有必要抑制决策树的生长,对模型进行适当的优化。而随机森林算法在决策树的基础上增加了多个分类器模型,避免了单棵决策树由于分类过度带来的过拟合风险。

2.3 模型优化

2.3.1 剪枝处理

虽然随机森林相比于单个决策树分类器来说,通过众数投票的方式在一定程度上能够避免过拟合问题,但如果随机森林当中的每个决策树不加以控制和修剪,必然会带来总体的预测误差及不稳定性。因此,适当地对决策树的生长加以控制,能够提高最终模型的预测稳定性。本次采取预剪枝方法控制决策树的高度和最大叶子节点数[11]来控制决策树的生长,防止出现过拟合现象。

2.3.2 交叉验证

在样本量不足够多的情况下,如果将训练集全部参与学习训练,必然导致学习能力过剩和模型的过拟合。因此,有必要使用留出集的方式从训练集当中随机选取部分数据作为验证集,通过多次交叉验证的方式,让数据的每个子集既是训练集,又是验证集,从而更好地评估模型性能和稳定性。图 4 为五轮交叉验证的示意。

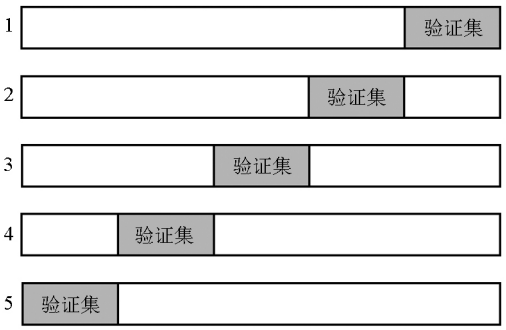

图 4　交叉验证示意

Fig.4　Cross-validation sketch

2.4　预测结果

设置随机森林模型中决策树的个数范围为 10~50,决策树的最大高度范围为 1~10,利用网格搜索和 10 次交叉验证法,获得本次砂土液化预测模型的最优参数,其中,决策树的个数为 20,最大高度为 5。图 5 为本次基于随机森林模型的砂土液化的预测结果,包括了 64 个训练样本和 8 个测试样本的结果。为了更好地说明模型的优越性,本次也加入了《建筑抗震设计规范 GB50011-2010》(2016 年版)中基于标贯试验的判别法[12]进行对比。

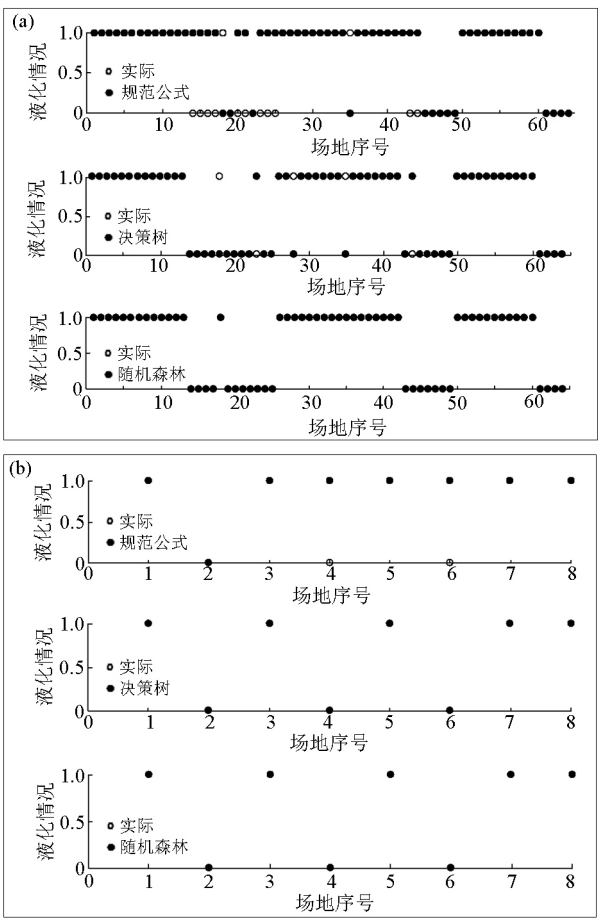

图 5　模型训练(a)及预测结果(b)

Fig.5　Training results(a) and test results(b) of prediction model

图 5 中空心圆表示实测液化结果,实心圆表示不同方法的预测结果。不难发现,抗震规范中的基于标贯试验的计算公式误判率较高,在训练样本上有 13 个样本误判,误判率为 20.3%,在预测样本上有 2 个样本误判,误判率为 25%。而决策树和随机森林模型的训练结果和预测结果明显高于规范公式的计算结果,在预测样本上均没有出现误判。但单个决策树模型在训练样本上有 5 个样本误判,误判率为 7.8%,稳定性明显不如随机森林预测模型。

3　结论

本文选取了 8 个影响砂土液化的判别指标,以唐山大地震中 72 个场点液化情况的实测样本为例,探讨了机器学习中的决策树和随机森林模型在砂土液化预测中的可行性。研究结果表明,与抗震规范中的标贯试验判别公式相比,决策树和随机森林预测模型的成功率有了明显的提高,尤其是随机森林预测模型,在多个决策树分类的基础上降低了样本学习的过拟合风险,提高了模型的预测稳定性,可以在今后砂土液化判别工作中予以推广。

参考文献(References):

[1]　高大钊,袁聚云.土质学与土力学[M]. 北京:人民交通出版社,2001.
Gao D Z,Yuan J Y.Soil science and soil mechanics[M]. Beijing:China Communications Press,2001.

[2]　宫凤强,李嘉维.基于 PCA-DDA 原理的砂土液化预测模型及应用[J].岩土力学,2016,37(s1):448－453.
Gong F Q,Li J W.Discrimination model of sandy soil liquefaction based on PCA-DDA principle and its application[J].Rock and Soil Mechanics,2016,37(s1):448－453.

[3]　刘章军,叶燎原,彭刚.砂土地震液化的模糊概率评判方法[J].岩土力学,2008,29(4):876－880.
Liu Z J,Ye L Y,Peng G.Fuzzy probability comprehensive evaluation method for sand liquefaction during earthquake [J].Rock and Soil Mechanics,2008,29(4):876－880.

[4]　谢君斐.关于修改抗震规范砂土液化判别式的几点意见[J].地震工程与工程振动,1984,4(2):95－109.
Xie J F.Some opinions on the modification of sand liquefaction discriminant of seismic code[J].Earthquake Engineering and Engineering Dynamics,1984,4(2):95－109.

[5]　陈国兴,孔梦云,李小军,等.以标贯试验为依据的砂土液化确定性及概率判别法[J].岩土力学,2015,36(1):9－26.
Chen G X,Kong M Y,Li X J,et al.Deterministic and probabilistic triggering correlations for assessment of seismic soil liquefaction at nuclear power plant[J].Rock and Soil Mechanics,2015,36(1):9－26.

[6] 刘红军,薛新华.砂土地震液化预测的人工神经网络模型[J].岩土力学,2004,25(12):1942 - 1946.
Liu H J,Xue X H.Artificial neural network model for prediction of seismic liquefaction of sand soil[J].Rock and Soil Mechanics,2004,25(12):1942 - 1946.

[7] 刘勇健.基于聚类—二叉树支持向量机的砂土液化预测模型[J].岩土力学,2008,29(10):2764 - 2768.
Liu Y J.Support vector machine model for predicting sand liquefaction based on clustering binary tree algorithm[J].Rock and Soil Mechanics,2008,29(10):2764 - 2768.

[8] 张菊连,沈明荣.基于逐步判别分析的砂土液化预测研究[J].岩土力学,2010,31(s1):298 - 301.
Zhang J L,Shen M R.Sand liquefaction prediction based on stepwise discriminant analysis[J].Rock and Soil Mechanics,2010,31(s1):298 - 301.

[9] 刘年平,王宏图,袁志刚,等.砂土液化预测的 Fisher 判别模型及应用[J].岩土力学,2012,33(2):554 - 557.
Liu N P,Wang H T,Yuan Z G,et al.Fisher discriminant analysis model of sand liquefaction and its application[J].Rock and Soil Mechanics,2012,33(2):554 - 557.

[10] 温博文,董文瀚,解武杰,等.基于改进网格搜索算法的随机森林参数优化[J].计算机工程与应用,2018,54(10):154 - 157.
Wen B W,Dong W H,Xie W J,et al.Parameter optimization method for random forest based on improved grid search algorithm[J].Computer Engineering and Applications,2018,54(10):154 - 157.

[11] 张亮,宁芊.CART 决策树的两种改进及应用[J].计算机工程与设计,2015,36(5):1209 - 1214.
Zhang L,Ning Q.Two improvements on CART decision tree and its application[J].Computer Engineering and Design,2015,36(5):1209 - 1214.

The method of predict sand liquefaction based on random forest algorithm

PENG Liu-Ya, XIE Hui-Ting, FENG Wei-Dong

(*Anhui Earthquake Engineering Institution, Anhui Earthquake Administration, Hefei* 230031, *China*)

Abstract: Among a variety of complicated factors that are related to sand liquefaction, 8 discriminant factors have been picked out of 72 samples in the earthquake event happened in Tangshan without losing any tiny but useful information. By calculating Gini coefficient with CART algorithm, a decision tree has been undertaken to divide the features of original sample dataset. Moreover, in order to reduce overfitting risk of a single decision tree, random forest with multiple trees have been created. Meanwhile, with 10-fold cross validation, best estimators with 5 max-depth and 20 trees can perform with much more stable and reliable results. The research shows that, compared to standard penetration test from Code for seismic design of buildings, both decision tree and random forest have a better predicting precision, especially there have been no false classifications with higher stability using random forest model.

Key words: sand liquefaction; discriminant indicator; decision tree; random forest

(本文编辑:叶佩)