基于机器学习的基坑变形预测研究*

杨建新，唐海英
（湖南省核工业地质局 三 0 二大队，湖南 株洲 423000）

摘要： 为确定施工现场深基坑沉降变形值，结合 3 种机器学习算法，建立沉降量与相关因素之间的非映射关系，并以上海某基坑为例对沉降量进行预测。结果表明，相比支持向量机和决策树算法，随机森林算法具有较高的预测精度，其拟合优度 R^2 和均方根误差 $RMSE$ 分别为 0.96、1.13，能很好地预测基坑变形量；内摩擦角对基坑沉降的影响最显著，土层渗透系数的影响较小。

关键词： 公路；基坑；沉降；机器学习
中图分类号：U416.1 **文献标志码：**A **文章编号：**1671－2668(2022)01－0077－04

常见基坑变形研究方法主要包括数值模拟、理论计算及智能算法预测等。在数值模拟方面，Liu Haiming 等利用 FLAC^{3D} 软件，选用 2 种本构模型对地面沉降进行模拟，通过与现场监测数据对比研究了基坑开挖影响范围；刘冰冰采用 ABAQUS 数值软件，对西安地铁四号线基坑工程沉降进行模拟，研究了基坑开挖降水对相邻建筑物的影响。在理论计算方面，Peck R. B.基于大量基坑工程数据，提出了基坑地表沉降计算公式；段绍伟等根据长沙市地铁开挖实测数据，采用回归分析方法对 Peck 沉降计算公式进行了修正。数值模拟及理论计算为现场基坑建设提供了理论指导，但由于基坑变形的复杂性及随机性，现场实际沉降与理论计算存在一定偏差。智能算法能避开基坑变形的内在机理，具有良好的预测能力，目前已成为基坑变形预测的主要技术手段。该文主要利用随机森林、决策树、支持向量机 3 种机器学习算法，结合上海某深基坑实测数据，对基坑变形进行预测，分析基坑沉降的影响因素。

1 机器学习算法

1.1 决策树算法

决策树算法是目前最常见的机器学习算法之一，它以信息熵作为判别标准，将决策树叶节点上的值作为输出样本信息，而非叶节点上的值作为数据样本中某个属性的划分点，样本数据根据该属性上的不同分割点被划分为多个子数据集。建立决策树的核心在于非叶节点上属性的选择，即如何选择适当的属性及属性分割点对样本数据进行划分。

对于回归问题，常采用 CART 决策树算法。对于给定的训练 $T=\{(x_1,y_1),(x_2,y_2),\cdots,(x_n,y_n)\}$，根据训练数据集中的几个或全部特征，按一定方法对样本数据进行分割，从而建立相应决策树，使决策树中叶节点上的值与训练样本中的值相等或接近。建立决策树的核心问题是非叶节点上特征的选择。假如选择训练集 T 中的 j 号特征中的 s 分量作为分割训练集的阈值，原数据集将分为 $R_1=\{x\mid R_j\leqslant s\}$、$R_2=\{x\mid R_j>s\}$ 两部分，分割后模型的输出值与实际 y 值的均方误差可表示为：

$$\min\Big[\min_{x_i\in R_1}\sum_{i=1}^{N}(y_i-f(x_i))^2+\min_{x_i\in R_2}\sum_{i=1}^{N}(y_i-f(x_i))^2\Big]$$

式中：$f(x_i)$ 为模型输出值，其越接近实际值 y，模型精度越高。

1.2 随机森林算法

随机森林的基本思想是通过 Bagging 集成，将多个弱决策树求解结果取平均值，从而获得具有较高精确度和泛化性能的算法。如图 1 所示，通过 Bootstrap重采样技术，从原始训练数据集 D 中有放回地重复随机抽取 k 个样本，生成新的训练数据集，然后基于新生成的 k 个训练集建立 k 颗决策树，将

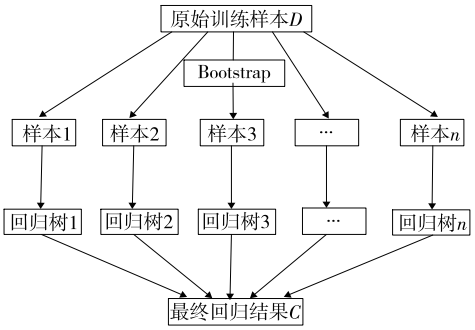

图 1 随机森林算法示意图

k 颗决策树组成随机森林。随机森林的计算结果等于所有决策树计算结果的平均值。

1.3 支持向量机

支持向量机是将实际问题通过非线性变换 $\Phi(x)$ 转换到高维的特征空间，再利用各种优化算法求得最大分类间隔，使样本点能线性可分地转换到所得到的高维空间。在这些样本点中，有一部分位于最大分类间隔的超平面之上，此即支持向量点。

如图 2 所示，设待求解的数据集为 (x_1,y_1)，(x_2,y_2)，…，(x_n,y_n)，$x\in R$，$y\in R$，$i=1,\cdots,N$，x_n 为输入数据，y 为输出数据。通过使所有样本点离超平面的总偏差最小，可建立如下关系式：

$$\begin{cases}\max\left[-\dfrac{1}{2}\sum_{m}^{i=1}\sum_{m}^{i=1}(a_i-a_i^n)(a_j-a_j^n)K(x_i-x_j)-\varepsilon\sum_{m}^{i=1}(a_i+a_i^n)+\sum_{m}^{i=1}(a_i-a_i^n)y_i\right]\\ \sum_{m}^{i=1}(a_i-a_i^n)=0;a_i,a_i^n\in[0,C]\end{cases}$$

式中：a_i、a_i^n 为拉格朗日乘子，系数 $a_i-a_i^n$ 不为零；$K(x_i-x_j)$ 为核函数，常见核函数包括线性核函数、多项式核函数、径向基核函数等；C、ε 为惩罚因子和不敏感损失参数。

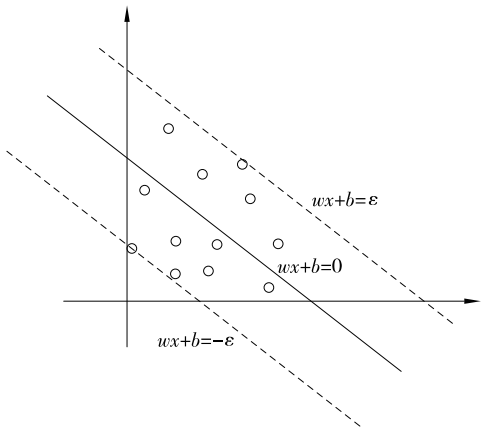

图 2 支持向量机原理示意图

2 基于机器学习的基坑沉降变形预测

2.1 工程分析

基坑开挖对周边地面变形的影响不可忽视。地面变形是多因素共同作用的结果，主要包括施工工况、岩土层参数、支护结构刚度及支撑形式等。各因素的影响程度及方式不同，应用传统的理论计算方法难以考虑多种因素建立准确的基坑沉降预测模型，机器学习方法为此提供了可靠途径。

以上海某基坑工程为例，现场施工过程中记录基坑开挖深度、开挖面以上地层内摩擦角、土体黏聚力、土体重度、地层渗透系数、监测点距离及监测点沉降等。图 3 为选取的输入变量与基坑变形量的 Pearson 相关系数图，用来衡量变量之间的线性相关关系，取值范围为[−1,1]，−1 表示为负相关，1 表示为正相关。数值越接近 1 或−1，表示相关度越强；越接近零，表示相关度越弱。从图 3 来看，输入变量与输出变量之间存在一定的相关性。

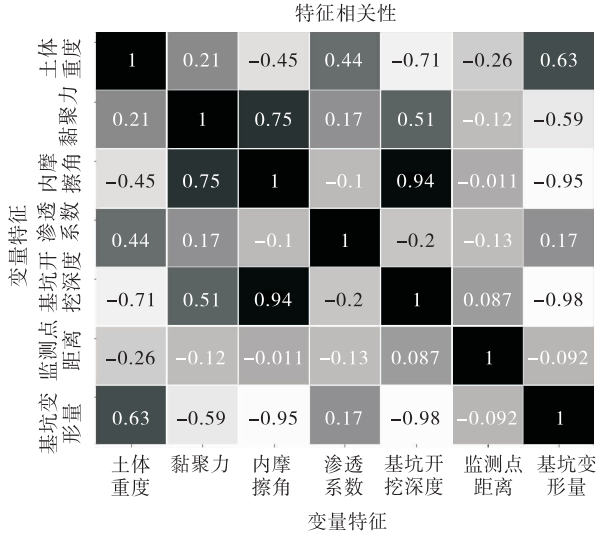

图 3 输入变量与输出变量相关系数图

选取 100 组监测数据（涵盖开挖前、开挖中及基坑施工后全周期）作为训练样本和测试样本建立预测模型，随机抽取 80% 数据作为训练集，剩下 20% 数据作为测试集，分别基于决策树算法、随机森林算法及支持向量机算法进行模型预测。

2.2 机器学习超参数调整及评价指标定义

通过调整模型超参数获得最优化模型，提高机器学习模型的预测准确性。基于网格搜索交叉验证方法进行超参数调整。图 4 为 5 折交叉验证示意图，其原理为通过将超参数数据集分为 n 个子集，以 1 个子集作为验证集，其余 $n-1$ 个子集作为训练

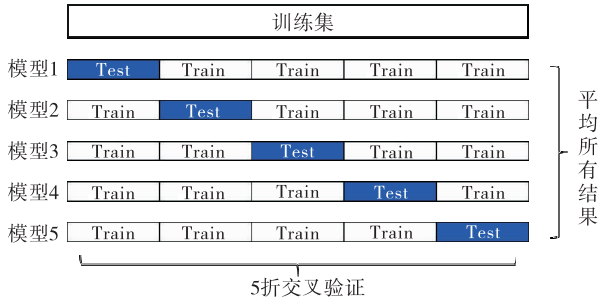

图 4 交叉验证示意图

集，得到模型结果，并通过循环变换验证集。重复上述过程，选取模型表现最优的超参数数据集作为模型的超参数。

采用拟合优度 R^2 和均方根误差 $RMSE$ 统计指标作为机器学习预测模型精度评价指标，公式如下：

$$R^2 = 1 - \frac{\sum (\hat{y}_i - y_i)^2}{\sum (y_i - \bar{y})^2}$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N} (\hat{y}_i - y_i)^2}$$

式中：$\hat{y}_i$ 和 y_i 为训练集中第 i 个样本的预测值、实测值；$\bar{y}$ 为训练样本的平均值；N 为训练样本总数。

R^2 的取值范围为[−1，1]，R^2 越接近 1，拟合越好；$RMSE$ 的取值范围为[0，+∞)，取值越接近零，预测值与实际值拟合越好。

2.3　预测结果及分析

机器学习中，使用网格搜索交叉验证获得的最佳超参数组合进行建模，各模型在测试集上的性能表现见表 1。从表 1 可看出：支持向量机的预测效果较差；随机森林和决策树算法具有较高的预测精度，其拟合优度都超过 0.9，且均方根误差在 2 以下；随机森林算法的预测能力最好，这主要是由于输入数据与输出数据具有高度非线性，集成算法具有较高的表现能力。

表 1　机器学习预测结果对比

预测模型	R^2	$RMSE$
随机森林	0.96	1.13
支持向量机	−35.82	34.34
决策树	0.90	1.51

利用 3 种机器学习模型对整个数据集进行建模分析，结果见图 5。从图 5 可看出：随机森林模型和决策树模型的预测值均较好地分布在理想拟合线附近，最大相对误差为 0.35%，且具有较高的稳定性；支持向量机模型的预测值表现较差，最大相对误差为 10.34%，难以满足工程实际要求。不同机器学习算法，由于其内核计算方法的差别，在同一工程数据的预测应用中表现出精度差异。

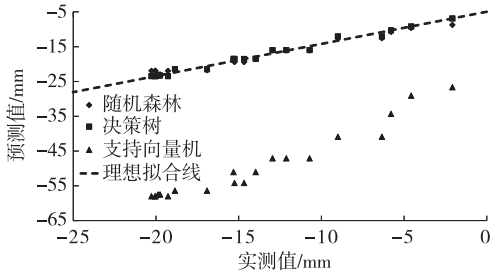

图 5　机器学习模型的预测结果

表 2 为随机森林模型预测值与基坑周边沉降实测值对比。从表 2 可看出：随机森林模型预测结果虽有一定波动，但在可接受范围内，其相对误差为 0.13%～2.01%，平均相对误差为 0.97%，对于基坑变形预测来说其精度满足要求。

表 2　位移实测值与随机森林模型预测值的比较

沉降实测值/mm	沉降预测值/mm	相对误差/%
−2.07	−4.080 08	2.010 08
−4.56	−5.037 48	0.477 48
−5.78	−6.249 55	0.469 55
−6.33	−8.234 09	1.904 09
−8.99	−8.234 09	0.755 91
−10.70	−11.980 00	1.280 00
−12.11	−11.980 00	0.130 00
−12.94	−11.980 00	0.960 00
−13.96	−14.658 80	0.698 80
−15.30	−14.658 80	0.641 20
−14.66	−15.642 10	0.982 10
−15.24	−15.642 10	0.402 10
−16.90	−18.179 40	1.279 40
−18.84	−18.179 40	0.660 60
−19.74	−19.492 30	0.247 70
−19.86	−19.492 30	0.367 70
−20.02	−18.354 90	1.665 10
−20.26	−18.354 90	1.905 10
−20.04	−18.354 90	1.685 10
−19.26	−18.354 90	0.905 10

3　基坑沉降影响因素分析

影响基坑沉降的因素很多，但不同因素的影响程度不一样。在机器学习算法中，函数“feature_importance_”对各影响因素的重要性给出了定量解释，数学过程如下：

(1) 对每一颗决策树，建立决策树前将数据集分为训练集和预测集，选择没有参与建立决策树的预测集数据进行预测，计算预测值与试验值的误差，记为 err_1。

(2) 随机对预测集数据中样本的影响因素(因变量)X 加入噪声干扰(即随机改变样本在特征 X 的值)，再次计算预测值与试验值之间的误差，记为 err_2。

(3) 假设森林中有 N 棵树，则影响因素(因变量)X 的重要性为：

$$ERR_X=\frac{1}{N}\sum_{i=1}^{N}(err_2-err_1)$$

加入随机噪声后，模型的精度会发生变化（即 err_2 改变），err_2 的变化幅度反映输出结果对 X 变量的敏感性。假如 X 变量对结果无影响，则 err_2 与 err_1 相等，即 $ERR_X=0$；ERR_X 越大，X 变量对样本预测结果的影响越大，该特征的重要程度较高。

基于随机森林模型分析各影响因素对基坑沉降的敏感性，结果见图 6，其中所有重要性系数总和为 1。从图 6 可看出：内摩擦角、黏聚力和检测点距离的相对重要性系数分别为 0.245、0.231、0.22，其值在所有影响因素中较高，影响因素的重要性排名为内摩擦角＞黏聚力＞监测点距离＞土体重度＞基坑开挖深度＞土体渗透系数，说明土层本身性质对基坑沉降至关重要。

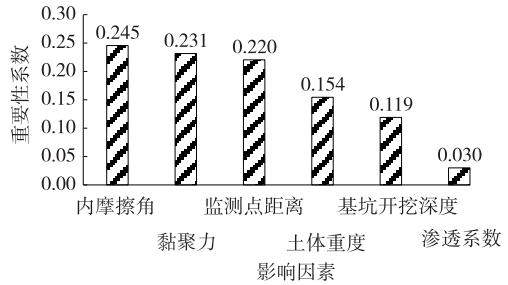

图 6 随机森林模型生成的特征重要性

4 结论

基于机器学习中的决策树、随机森林和支持向量机算法对基坑沉降进行预测，得出如下主要结论：

（1）传统模型一般难以考虑基坑的复杂性。基于基坑实测数据建立基坑沉降预测的机器学习模型，通过与实测数据对比，随机森林预测模型的表现优于其他 2 种模型，其最大相对误差为 2.01%。

（2）内摩擦角对基坑沉降的影响最显著，但土层力学性质等特征的影响较平均，土层渗透系数对基坑沉降的影响较小。

参考文献：

[1] XU Ping，HAN Yuewang，DUAN Honghai，et al.Environmental effects induced by deep subway foundation pit excavation in yellow river alluvial landforms[J].Geotechnical and Geological Engineering，2015，33(6)：1587－1594.

[2] 申琪玉，王建雄，王东.深基坑工程事故常见原因分析及对策[J].建筑技术，2015，46(10)：912－914.

[3] 寇润胜.深基坑周边建筑物沉降预测及安全性评估[D].重庆：重庆大学，2014.

[4] 杨林德，时蓓玲，杨超.基坑变形及其安全性的随机预测[J].同济大学学报(自然科学版)，2002，30(4)：403－408.

[5] 黄燕妮.基于智能算法的深基坑工程变形预测与控制方法的应用研究[D].广州：华南理工大学，2019.

[6] LIU Haiming，CAO Jing，ZHANG Weifeng.Study on effects of ground deformation based on foundation pit excavation[J].Advanced Materials Research，2013(1)：357－362.

[7] 刘冰冰.西安地铁大后区间基坑开挖降水对周围建筑物沉降研究[D].北京：北方工业大学，2015.

[8] PECK R B.Deep excavation and tunneling in soft ground[C]//Proceedings of the 7th lnternational Conference on Soil Mechanics and Foundation Engineering.1969：225－290.

[9] 段绍伟，黄磊，鲍灶成，等.修正的 Peck 公式在长沙地铁隧道施工地表沉降预测中的应用[J].自然灾害学报，2015，24(1)：164－169.

[10] 杨娟丽，徐梅，王福林，等.基于 BP 神经网络的时间序列预测问题研究[J].数学的实践与认识，2013，43(4)：158－164.

[11] 曹聪洁，施冬，韦原原.基于决策树的黄土高原滑坡风险分析[J].农业与技术，2021，41(6)：141－144.

[12] 王志红，王华珍.基于随机森林的基金评级模型选择[J].财务与金融，2009(1)：65－70.

[13] 唐江凌.SVR 对铣削型钢纤维混凝土劈裂抗拉强度的预测研究[J].电脑知识与技术，2020，16(19)：226－227.

[14] 刘沛东，冯江峰，徐文杰，等.肩袖损伤后脂肪浸润基因表达谱及关键通路的生物信息学分析[J].中国组织工程研究，2021，25(11)：1773－1778.

[15] 李帅，李浩，夏玉伟.BP 神经网络在基坑变形预测中的应用研究[J].公路与汽运，2017(5)：97－101.

[16] 朱艳玲.SVM 及信息融合在深基坑施工安全实时监控中的应用[D].西安：西安建筑科技大学，2014.

[17] 邹亮，黄琼，李骜，等.基于随机森林和富集分析的阿尔茨海默症 GWA 研究[J].中国科学(生命科学)，2012，42(8)：639－647.

[18] 史豪杰，李富相，李志勇.公常路下穿改造工程深基坑开挖施工监测与稳定性分析[J].公路与汽运，2021(4)：97－100＋165.

[19] 门彬.超期服役基坑变形关键影响因素模拟分析[J].交通科学与工程，2020，36(3)：50－55.

收稿日期：2021－04－19