工程管理
Journal of Project Management
第 4 卷◆第 8 期◆版本 1.0◆2023 年
文章类型：论文| 刊号（ISSN）：2737-4580(P) / 2737-4599(O)

基于随机森林的地基土未采样点 CPT 空间参数预测方法

谭景文
武汉轻工大学土木工程与建筑学院
DOI：10.12238/jpm.v4i8.6169

[摘　要] 机器学习技术已经作为一种辅助方法协助人们在多个领域中应用。随机森林是一种集成机器学习技术，与其它算法相比，具有处理高维数据方式灵活，较好表现性能的特点。本文针对土体性质评估难度高的问题，采用一种基于随机森林算法构建的模型来对未采样点的空间土性参数进行预测和评估，并选取了一个静力触探试验(CPT)的案例用于验证模型的正确性，最终利用已有的 CPT 数据在该模型中预测未取样点桩位的锥尖阻力(Q_c)和侧摩阻力(F_s)。结果表明基于随机森林算法建立的机器学习模型具有预测准确度高，适用性强的特点，在工程中具备良好的应用前景。
[关键词] 机器学习；随机森林；静力触探试验
[中图分类号] TU451　　[文献标识码] A

Prediction of Geotechnical Parameters of Unsampled Points by Random Forest Algorithm Based on CPT Data

(1. College of Civil Engineering and Construction,Wuhan Polytechnic University Wuhan 430023, China;)
[Abstract] Machine learning technology has been used as an auxiliary method to assist people in many fields. Random forest is an integrated machine learning technology. Compared with other algorithms, it has the characteristics of flexible processing of high-dimensional data and better performance. Therefore, aiming at the problem of high difficulty in the evaluation of geotechnical properties, a model based on random forest algorithm is used to predict and evaluate the geotechnical properties of unsampled points, and a case of static cone penetration test (CPT) is selected to verify the correctness of the model. Finally, the existing CPT data are used to predict the cone tip resistance (q_c) and side friction resistance (f_s) of the pile position at the unsampled point in the model. The results show that the machine learning model based on random forest algorithm has high prediction accuracy and strong applicability, and has highly application prospects in engineering.
[Key words] machine learning; random forest; CPT

1 引言

路基是公路、铁路的重要结构型式，路基的沉降控制是保证交通安全运行的前提，路基的变形与沉降计算则依靠所在地基物理力学参数的精准获取。由于地层结构是经历了漫长且复杂的地质作用形成的，其土体性能在三维空间表现出极大的复杂性[1]。工程地质勘察能够有效反映建设工程的地质条件，其中原位测试是详细勘察阶段常用的手段。由于勘察人员通常只能对场地内有限数量的勘测点进行测量，而未测区域的土体性质只能凭经验估测[2]。因此导致地基处理常常趋于保守，大大增加工程建设成本。

近年来，国内外学者就如何根据有限的勘测数据准确预测未采样处的地层参数进行了一定的研究。M. Lloret-Cabot 等[3]在 2012 年提出了一种处理空间变异性引起土体性能不确定性的便捷方法，即利用野外测量中由位置约束产生的随机场。Li 等[4]提出了一种将 3D Kriging 与现有随机场相结合的方法。这是 Kriging 首次应用于 3D 问题。但由于各方向的半变异函数难以表征，使得 Kriging 插值方法难以应用于各向异性三维问题[5]。静力触探试验(CPT)是一种高效、准确的原位试验方法。它可以检测软土层的分布情况，区分砂土液化情况，并可评估土体的软化度、密实度、黏聚力、内摩擦角和压缩程度。关于 CPT 试验方面已开展了大量的研究，但主要集中在静力触探数据与土体物理力学参数的关联性研究，对未知地层参数预测较少。Ching 等[6]通过对台中市五峰区 CPT 测深的分析，充分说明了处理统计数据的不确定性以及现存点评估和降趋势法的局限性。李镜培等[7]采用空间递推平均法和改进相关函数法分析部分静力触探比贯入阻力实测数据，为确定土性指标自相关特征参数提供了方法，但计算结果还存在一定的局限性。

本研究以 CPT 勘测数据为基础，利用决策树派生的随机森林算法，建立机器学习模型，解决三维土体中未采样点的岩土参数确定问题。所建模型的特点是能够将有限点的三维数据应用到机器学习中，适用于大数据集且不需要降维，具有较好的精度。基于选取算例的 CPT 测试数据，对未采样点进行分析预测，探讨该方法在工程实践中的适用性。

2 随机森林算法简介

随机森林是基于多个决策树的集成学习算法。集成学习类似于集思广益，即训练出多个决策树评估器，然后将多个评估

Copyright © This work is licensed under a Commons Attibution-Non Commercial 4.0 International License.

文章类型：论文|刊号（ISSN）：2737-4580(P) / 2737-4599(O)

进行组合得到最终的模型，从而提高模型的预测精度。决策树是一种应用广泛的机器学习模型。本质上，决策树模型通过学习"If-Else"问题的层次结构建模并做出决策[8]。常用的决策树算法有 C4.5、ID3 和 CART。决策树从原始数据开始，通过"If-Else"进行扩展。

2.1 集成学习

决策树的主要缺点是即使有了预剪树枝，模型也容易出现过度拟合，从未降低了模型的泛化能力。因此，通常采用集成的方法来代替单一的决策树。集成方法将多个机器学习模型结合起来，建立一个更强大更准确更稳健的模型。集成学习常用的处理方法是 Bagging 和 Boosting[9]。二者的区别阐述如下。从结构上比较，Bagging 是并行处理，Boosting 是串行处理。从训练的角度，Bagging 的子模型训练是独立的，Boosting 的子模型训练是存在依赖关系，即后面的子模型依赖前面的子模型。从作用上讲，Bagging 减小的是样本的方差，而 Boosting 减小的是样本的偏差。

前文提及决策树的主要缺点是模型训练过拟合，随机森林为解决该问题提出了一种解决方案。随机森林的思想是，每棵树都可能做比较好的预测工作，但可能是过拟合模型。但是如果构建了多棵树，所有这些树都会很好地工作，并且以不同的方式进行修剪，过拟合的次数可以通过结果的平均化来减少，随机森林是 Bagging 的一个扩展变体。

随机森林中的分类问题通常是通过对每个模型的预测结果进行投票得到的，而回归问题通常是通过加权平均得到的。

2.2 随机森林优点

为了防止过拟合或欠拟合，每个决策树模型必须在最大程度上存在差异，使得集成前每个模型的预测效果各具特色，最终达到较好的预测效果。为了达到上述目的，随机森林采用了两种方法。一是训练数据集的差异，通过引导采样使每棵决策树建立的数据集不同；二是特征的差异，通过对各自节点中的特征进行选择，将每棵决策树的节点划分为不同的特征集。高特征分类意味着随机森林中的树会具有更高的相似度，模型复杂度也较高，而低特征分类则相反。因此，可以通过调整特征集的大小来实现提高模型精度的目标。

对于模型验证方法，除了使用传统的交叉验证外，其本身自带验证属性。在生成单棵树时，算法对训练集进行抽样过程中，约有 1/3 的样本没有被选择。这些未选样本数据(也被称为 out-of-bag data)可以用以验证决策树的预测精度。这部分未选取的数据可以用来估计随机森林单个决策树的分类强度和决策树之间的相关性，进而可以得到随机森林泛化误差。L. Breiman 在 2001 年证明了 out-of-bag data 评估误差是一种可以代替测试集的评估误差方法。这种方法被称为 out-of-bag 误差法，这也是随机森林本身的独特优势。

随机森林具有较好的计算精度，既不需要像 SVM 那样对参数进行过多的调整，也不需要对样本进行缩放。同时，它适用于高维问题，不需要降维，也可以弥补数据的不足。从本质上讲，随机森林共享了决策树的所有优点，弥补了它们的一些不足。虽然在大数据集上构建随机森林可能有点费时，但在一台计算机上很容易跨多个 CPU 核并行化，即使在非常大的数据集上，随机森林通常也能很好地工作。

3 机器学习模型构建

3.1 建立样本集

本研究以实测 CPT 试验数据为样本，采用随机森林算法建立场地内的未采样点 CPT 空间预测的机器学习模型，解决了三维土质条件下通过有限数据预测土体整体特性范围的问题，为后续岩土工程应用提供可靠的参数。图 1 为 CPT 试验钻孔位置示意图。

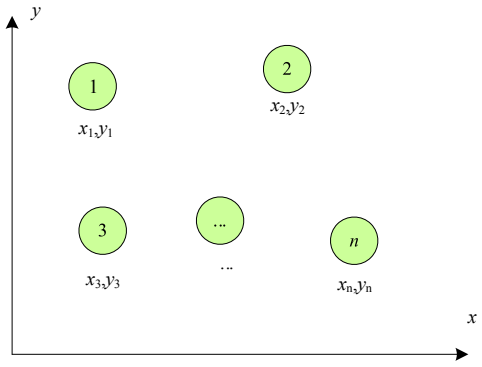


图 1　CPT 钻孔位置示意图

现场采集的 CPT 数据分为特征组和标签组。采用 CPT 采样点的三维坐标(x, y, z)作为特征组，对应采样点的岩土参数锥尖阻力 Q_c 和侧摩阻力 F_s 作为标签组。对于 n 个 CPT 采样点，表 1 给出了实测参数的样本表。

表 1　机器学习样本集

采样点	特征			标签	
	X 坐标	Y 坐标	Z 坐标	F_s	Q_c
1	x_1	y_1	z_{11}	f_{s11}	q_{c11}
	x_1	y_1	z_{12}	f_{s12}	q_{c12}
	...	...	...	...	...
	x_1	y_1	z_{1m}	f_{s1m}	q_{c1m}
2	x_2	y_2	z_{21}	f_{s21}	q_{c21}
	x_2	y_2	z_{22}	f_{s22}	q_{c22}
	...	...	...	...	...
	x_2	y_2	z_{2m}	f_{s2m}	q_{c2m}
...	...	...	...	...	..
n	x_n	y_n	z_{n1}	f_{sn1}	q_{cn1}
	x_n	y_n	z_{n2}	f_{sn2}	q_{cn2}
	...	...	...	...	...
	x_n	y_n	z_{nm}	f_{snm}	q_{cnm}

传统的模型测试采用留一法，即将样本集随机分为训练集和测试集，通过训练集训练机器学习模型，通过测试集测试模型的准确性。交叉验证是目前比较流行的一种测试方法。它在评价泛化性能方面比使用分割成训练集和测试集后再输入模型更稳定。

在交叉验证中，不是将数据集划分为训练集和测试集，而是反复对数据进行拆分，训练多个模型。交叉验证的主要缺点是增加了计算所需时间，因为交叉训练生成多个模型而不是单个模型，比传统的保留方法速度慢很多倍。对于模型的验证，本文采用的交叉验证法。

3.2 实施步骤

本研究的目的是通过随机森林算法构建未采样点的 CPT 空间预测机器学习模型，主要通过以下步骤实现：

(1)原始数据准备

原位数据通过 CPT 获得，以三维坐标为特征(x, y, z)，以目标岩土参数锥尖阻力 Qc 和侧摩阻力 Fs 为标签(Qc, Fs)。

(2)设定参数

引入随机森林算法，初步设定 RF 框架的主要参数和控制单棵树的参数。比如设置树的棵为 100，指定单棵树的参数准则，最大特征数，最大分支深度等。

(3)模型训练

利用分配好的训练集，通过特征组和标签组重复输入随机森林算法模型进行训练，并对每个子决策者的回归结果采用简

Copyright © This work is licensed under a Commons Attibution-Non Commercial 4.0 International License.

单平均法进行算术平均，初步形成针对这一岩土问题的独特随机森林算法模型。

(4)模型验证

采用交叉验证的测试方法评估模型的测试精度。

(5)参数调整

根据测试精度适当调整合适的参数，然后重复进行训练和测试，以达到预测效果最佳的模型。

(6)未采样点预测

利用最终的机器学习模型，将未采样点的特征(x, y, z)输入到模型中即可得到精度保证的预测标签(Qc, Fs)，即此时的岩土参数为锥尖阻力 Qc 和侧摩阻力 Fs。

(7)可视化

利用图形处理软件对预测结果进行可视化，如 Matplotlib、Mayavi 和 Paraview，并将预测结果以二维平面或三维立体的形式表达。

(8)其他潜在应用

利用预测的岩土参数，根据实际工程需要，应用如土层分层确定、承载力分析、液化评估等。

4 案例研究

4.1 研究场地

本研究采用的 CPT 数据采集地点在南澳大利亚阿德莱德，该场地包括 222 个垂直 CPT，深度为 5 米，在 50 * 50 米范围内进行方格网布置。场地土层包括硬粘土、超固结粘土。M. Jaksa 等[10]对原位测试场地进行了较为详细的描述。

在实际工程中，原位测试的样本点是非常有限的，所以本研究选取 6 个钻孔位置的数据作为建模数据，如图 2 中所示。

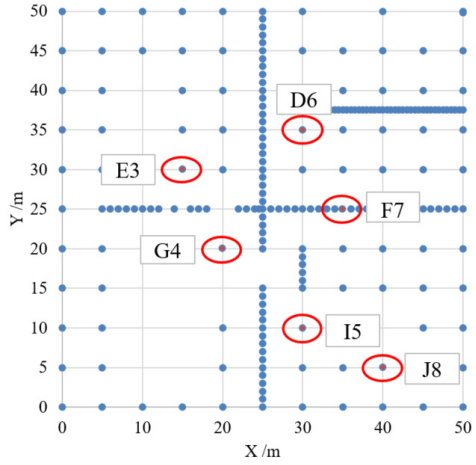
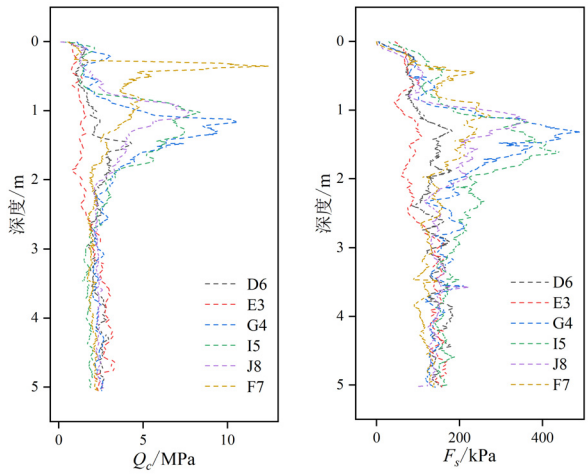


图 2　钻孔位置示意图

4.2 原始数据

本研究根据已有的相关 CPT 资料，其可视化下图所示。



(a) 尖端阻力(Qc)　　(b) 侧摩阻力(Fs)

图 3　目标岩土参数随深度的变化

4.3 建模与评估

本研究采用 Python 编程语言，利用 SCIKIT-LEARN 机器学习库进行建模和评估。

首先对现场数据进行处理，然后利用随机森林算法对目标岩土参数进行预测。以三维坐标为特征(x, y, z)，以锥尖阻力 Q_c 和侧摩阻力 F_s 的岩土参数为标签。每组特征对应一个对应的标签。已划分的数据组约有 5000 个。验证方法采用 5 折交叉验证，即将 5000 份数据单独分为 5 份，每份 1000 个样本，4 分作为训练样本，1 份作为测试样本。训练样本又划分为训练集和测试集，每份数据单独进行训练和测试。

然后引入随机森林算法，随机森林回归参数采用默认值，如表 2 所示(SCIKIT-LEARN 中随机森林算法的参数)。利用分配好的训练集对模型进行反复训练，结合交叉验证，得到最终的预测模型。试验结果如图 4 和图 5 所示。将 6 个不同坐标的 CPT 数据的一半划分为训练集，生成机器学习模型，然后对 6 个不同坐标的钻孔进行预测，并利用测试集对模型的精度进行评价。结果表明，锥尖阻力 Q_c 和侧摩阻力 F_s 的预测值与实测值吻合较好，模型建立正确，同时还进一步验证了该随机森林算法预测准确度高，处理高纬度数据能力强的特点。

表 2　随机森林回归算法参数默认值

参数	默认值	参数	默认值
n_estimators	100	min_impurity_split	None
criterion	mse	bootstrap	True
max_depth	None	oob_score	False
min_samples_split	2	n_jobs	None
min_samples_leaf	1	random_state	None
min_weight_fraction_leaf	0.0	verbose	0
max_features	auto	warm_start	False
max_leaf_nodes	None	ccp_alpha	0.0
min_impurity_decrease	0.0	max_samples	None

Copyright © This work is licensed under a Commons Attibution-Non Commercial 4.0 International License.

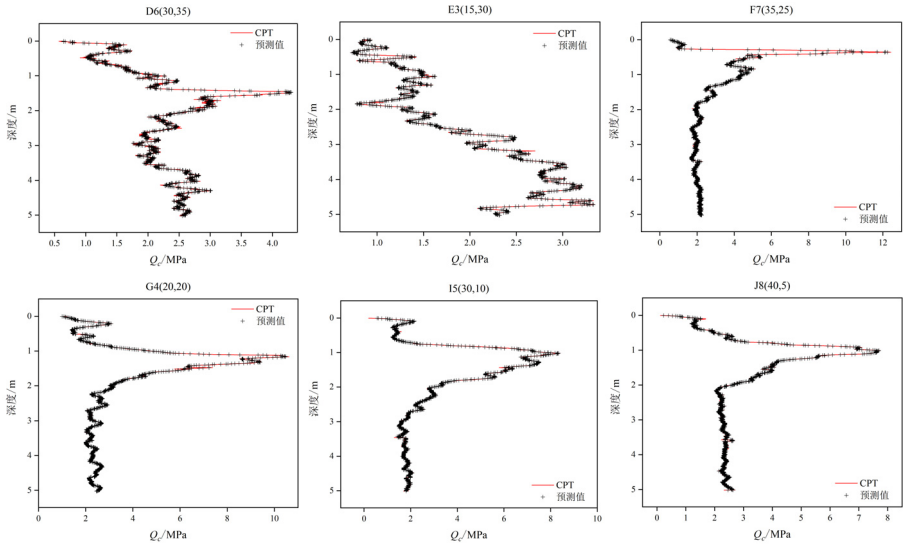


图 4　CPT 钻孔位置锥尖阻力 Q_c 实测值与预测值的比较

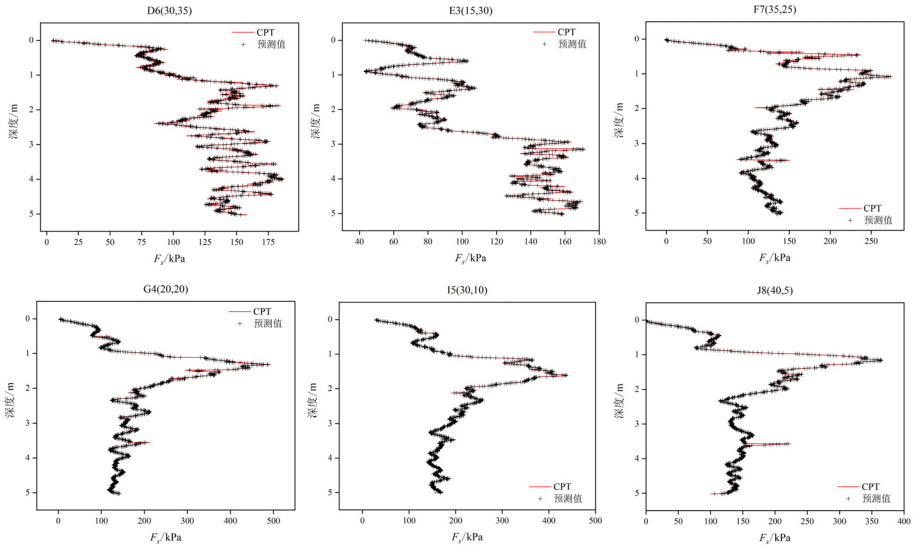


图 5　CPT 钻孔位置侧摩阻力 Fs 实测值与预测值的比较

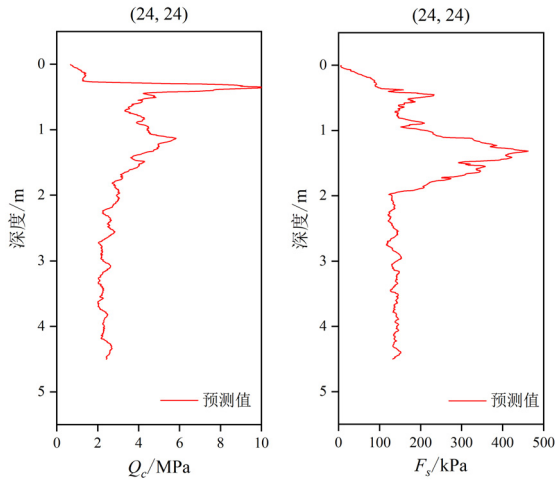


图 6　CPT(24，24)钻孔位置 Q_c，Fs 预测结果

利用上述模型对任意场地范围内的 Q_c 和 F_s 进行预测，图 6 即为场地坐标(24，24)和深度 4.5m 范围内的 Q_c 和 F_s 预测结果。从图中可以看出，Q_c 和 F_s 在深度大于 2 时，预测结果趋于稳定，此时二者预测值分别为 2.2 MPa 和 150 kPa。这也说明了，基于随机森林建立的机器学习模型可以输出较好的预测结果，模型的泛化能力强，适用性较为广泛，采用该方法对场地内任意点位处的 CPT 参数进行预测非常便捷。

5 结论

机器学习在各个领域发挥了重要的作用，它可以作为一种辅助方法来帮助研究者进行性能的预测与分析。本文采用随机森林机器学习算法构建模型，对未采样点的岩土特性进行预测和评估。试验中选取了静力触探试验的案例，利用已有的 CPT 数据预测未取样点桩位的锥尖阻力和侧摩阻力。预测结果与实测值吻合较好。利用随机森林建立的模型取得了良好的预测效果，未来可以考虑将其应用于实际工程中，在减少工程勘察工作量的同时，提升对地层三维空间参数变化的判断水平。

Copyright © This work is licensed under a Commons Attibution-Non Commercial 4.0 International License.

[参考文献]

[1]田密，张帆，李丽华. 间接测量数据条件下岩土参数空间变异性定量分析方法对比研究[J]. 岩土力学，2018，39(12):4673-4680.

[2]胡越，王宇. 静力触探识别场地土层分布的贝叶斯学习方法研究[J]. 工程地质学报，2020，28(05):966-972.

[3]Lloret-Cabot M. Hicks M A, van den Eijnden, A P. Investigation of the reduction in uncertainty due to soil variability when conditioning a random field using Kriging. Géotechnique Letters, 2012, 2(3), 123-127.

[4]Li Y J, Hicks M A, Vardon P J. Uncertainty reduction and sampling efficiency in slope designs using 3D conditional random fields. Computers and Geotechnics, 2016, 79, 159-172.

[5]Cai Y M, Li J H, Li X Y, Li D Q, Zhang L M. Estimating soil resistance at unsampled locations based on limited CPT data. Bulletin of engineering geology and the environment. 2019, 78 (5), 3637-3648.

[6]CHING J Y, PHOON K K, WU S H. Impact of statistical uncertainty on geotechnical reliability estimation[J]. Journal of Engineering Mechanics, 2016, 142(6): 04016027.

[7]李镜培，舒翔，丁士君. 土性指标的自相关特征参数及其确定原则[J]. 同济大学学报(自然科学版)，2003，(03):287-290.

[8]L. Breiman. Random forests. Machine Learning. 2001. 45 (1), 5-32.

[9]Bauer E, Kohavi R. An empirical comparison of voting classification algorithms: bagging, boosting, and variants. Machine Learning. 1999, 36, 105-139.

[10]Jaksa M B, Kaggwa W S, Brooker P I. Experimental evaluation of the scale of fluctuation of a stiff clay. ICASP8, Int. Conf. Applications of Statistics and Probability in Civil Engineering. Sydney, 1999, 1, 415-422.

Copyright © This work is licensed under a Commons Attibution-Non Commercial 4.0 International License.