Advances in Geosciences 地球科学前沿, 2020, 10(12), 1285-1294
Published Online December 2020 in Hans. http://www.hanspub.org/journal/ag
https://doi.org/10.12677/ag.2020.1012125

Hans汉斯

应用随机森林模型识别南宁地铁线路地层结构特征

李浩楠[1]，周念清[1*]，郑茂辉[2]，李晓峰[3]

[1]同济大学，土木工程学院，上海
[2]同济大学，上海防灾救灾研究所，上海
[3]南宁轨道交通集团有限责任公司，广西 南宁
Email: *nq.zhou@tongji.edu.cn

收稿日期：2020年12月2日；录用日期：2020年12月24日；发布日期：2020年12月31日

摘 要

利用传统空间插值方法对钻孔数据进行三维地层建模易受主观因素的影响，为了克服地层建模过程中产生的误差，本文基于随机森林模型提出了一种新型的地层识别方法。选取南宁地铁1号线部分工程勘察资料进行研究，利用网格搜索确定模型参数后，构建随机森林模型，并与支持向量机进行对比分析。通过研究表明，随机森林模型的预测精度达到81.7%，其超参稳定性明显高于支持向量机，且预测精度受其主要参数(树的数量和最小叶子节点数)变化的影响较小；交叉验证评估结果也证实了随机森林的泛化性能更好。当样本数量较少时，随机森林模型不论是在分类精度还是稳定性方面均较好。该方法在三维地质建模中具有良好的应用价值。

关键词

随机森林，地层识别，支持向量机，三维建模，超参稳定性

Research on Recognition of Stratum Structure Using Random Forest Model

Haonan Li[1], Nianqing Zhou[1*], Maohui Zheng[2], Xiaofeng Li[3]

[1]School of Civil Engineering, Tongji University, Shanghai
[2]Shanghai Institute and Disaster Prevention of Relief, Tongji University, Shanghai
[3]Nanning Rail Transit Co., Ltd., Nanning Guangxi
Email: *nq.zhou@tongji.edu.cn

Received: Dec. 2nd, 2020; accepted: Dec. 24th, 2020; published: Dec. 31st, 2020

*通讯作者。

文章引用: 李浩楠, 周念清, 郑茂辉, 李晓峰. 应用随机森林模型识别南宁地铁线路地层结构特征[J]. 地球科学前沿, 2020, 10(12): 1285-1294. DOI: 10.12677/ag.2020.1012125

Abstract

Using traditional spatial interpolation method based on borehole data for 3D stratum modeling is easily affected by subjective factors. In order to overcome the errors in the stratum modeling process, this paper proposes a new stratum recognition method based on the Random Forest model. Part of the engineering investigation data of Nanning Metro Line 1 is selected to carry out the research. After the classifier parameters are determined by the grid parameter search, the Random Forest model is constructed and compared with the Support Vector Machine. Research shows that the prediction accuracy of the random forest model reaches 81.7%, and its hyperparameter stability is significantly higher than that of the support vector machine, and the prediction accuracy is less affected by changes in its main parameters (the number of trees and the minimum number of leaf nodes). The cross-validation results also confirm that the generalization performance of random forest is better. When the number of samples is small, the random forest classifier still has a good performance in terms of classification accuracy and stability. This method has a good application prospect in 3D geological modeling.

Keywords

Random Forest, Lithology Identification, Support Vector Machine, 3D Geological Modeling, Hyperparameter Stability

1. 引言

地层结构是在漫长的地质历史演化过程中形成的，其时空分布具有不均匀、不规则等特点，在宏观上具有一定的统计规律[1]。地层结构的识别有多种方法，其中钻探法是获取地层构造以及岩性分层详细分布信息最为直观、可靠的手段[2]；钻孔数据所记录的信息准确率高，是三维地质建模的重要依据[3]，也是地铁工程选线、设计和施工中必不可少的数据资料。在三维地质建模中连接各钻孔地层时，主要使用线性插值、多项式插值、反距离插值和 kriging 插值等空间插值方法，各种不同的插值方法得到的模拟结果存在一些差异，具有一定的局限性[4]。

本文以南宁地铁 1 号线部分路段作为研究对象，由于线路要穿越人口稠密繁华商业区、市民居住区以及车流量较大的交通要道，勘察施工会给周围环境和市民生活带来影响[5]。南宁地区可溶岩广泛分布，施工时很容易发生岩溶塌陷与突水事故[6]。因此，在南宁地铁建设中，往往由于勘探场地受限，获取钻孔数据成本高，如何利用有限的钻孔数据识别地层岩性及其层序分布是值得关注和探索的问题。

近年来，机器学习算法得到了迅速发展，算法建模既可以用于大型复杂数据集，也可以在较小数据集上建立更准确的模型[7]，此方法在地质学领域得到不断推广和应用；陈玉林等[8]基于 K 近邻算法识别合水地区的岩性，分类准确率达到 89.5%；郭甲腾等[9]基于支持向量机和 BP 神经网络实现钻孔数据自动地层分类，提出需进一步探索超参数的经验值确定方法；Cracknell 等[10]基于地球物理数据对比研究了 5 种机器学习方法的地层分类效果，其中随机森林算法的分类效果最好；马梓程等[11]利用光谱和纹理信息，基于随机森林建立了火成岩分类模型；徐剑波和陈军林[12]应用随机森林算法结合区域的化探数据

来推断地质体的空间分布。因此，在此基础上，本文提出了一种基于随机森林(Random Forest, RF)模型的钻孔数据地层识别方法，并与支持向量机(Support Vector Machine, SVM)模型进行对比分析。

2. 研究方法

2.1. 随机森林算法

随机森林是 Bagging 方法和 Random 子空间的组合[13]，基本构成单元是决策树，通过多棵决策树的组合提高分类的准确性。首先随机生成训练集，利用 bootstrap 方法随机为每棵树生成训练数据，可能重复包含，也可能不重复包含某些数据，并由此构建 K 棵分类决策树，每次未被抽到的样本组成袋外数据(Out-Of-Bag, OOB)。随机选择特征子集：当决策树节点拆分时，随机选择特征子集，该子集的大小 m 通常小于特征总数 M。计算 m 个特征下的基尼系数，选择最佳分割特征。集合每棵决策树的预测结果，且每棵树被采样的机会均等，可以有效地生成随机树，并且将大量随机树组合在一起可以得出准确的模型，通过最终投票对未知类别的样本进行分类。

本文提出的基于随机森林的钻孔数据地层识别流程如图 1 所示。关键步骤包括：1) 对钻孔数据进行预处理，按地层类型划分样本并随机划分训练集和测试集；2) 利用训练集内部交叉验证寻找适合当前问题的随机森林模型超参数组合；3) 根据最优超参数，训练随机森林分类器；4) 由训练好的分类器对测试集进行预测分类，确定其地层类型。

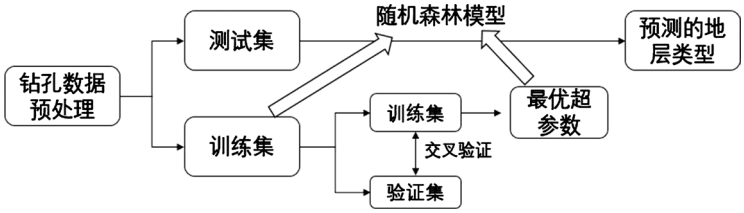


Figure 1. Process of strata recognition for borehole data based on Random Forests
图 1. 基于随机森林模型的钻孔数据地层识别流程

2.2. 地层分类树

本文以 CART 分类树[14]为基学习器构建 RF 地层识别模型。CART 分类树为二叉分类树，由根节点、子节点和叶子节点组成，其中每个从根节点到叶子节点的路径都对应着其依据地层相关属性的分类过程，而叶子节点则对应一种地层类别。使用 CART 决策树进行节点分割时，选择具有最小基尼系数的特征作为最佳分割属性[15]，计算如下：

$$\mathrm{Gini}(t)=1-\sum_{i}^{c}P_i^2 \tag{1}$$

式中：P_i 为地层类型 i 在 t 节点处的概率，Gini(t)为 0 时，表示在 t 节点处的样本数据为同一地层类型；Gini(t)越大，表明在 t 节点处的样本数据越趋于均匀，能获得的分类信息越少。

如图 2 所示，样本集 S_1 的训练过程对应于 CART 分类树的生长过程，即把位于根节点的样本集 S_1 按所给定的属性划分不断进行递归分割。单棵地层分类树的生长训练过程如下：

1) 利用 bagging 方法获取训练集。从具有 N 个样本的总训练集中有放回地随机抽取 n 个组成单棵树的训练集 S_1。

2) 随机选取节点的属性指标。在钻孔数据当中共有 M 个属性，随机地从 M 个指标中选取 m 个作为节点指标。

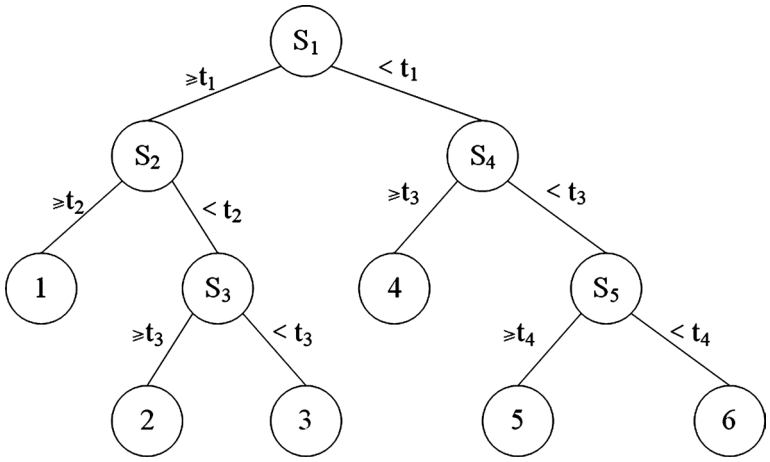


Figure 2. Single strata classification tree
图 2. 单棵地层分类树

3) 节点的递归分割。对于每一个节点的钻孔属性指标都要遍历所有可能的分割方法，选择最小的基尼系数作为此节点的分割标准，对应的属性指标为最优的地层分类指标，然后按最优地层分类指标进行分割。如图 2 中数据集 S_4 根据最优属性指标 t_3 分为两个子数据集 4 和 S_5，其中 4 节点的基尼系数已经很小(通过设定的阈值判断)，可认为该节点所有样本属于同一类别，即地层类别 4，不用继续向下分割；而数据集 S_5 则继续分割。

4) 然后将生成的多棵树组成 RF，用 RF 对新的数据进行分类，分类结果按树分类器投票决定。

2.3. 模型评价

假定 n_{ij} 表示被分类为 j 类的 i 类样本，k 表示地层类别的数量，则分类准确率 A 以正确分类的样本数与总样本数 N 的比值来表示：

$$A = \frac{\sum_{i=1}^{K} n_{ii}}{N},\ K = k \tag{2}$$

该指标是用来衡量分类器对于测试集的总体分类精度，总体分类精度越高说明算法的分类效果越好。除整体分类精度外，各单一地层分类准确率也十分重要，这里选用召回率 R 和精确度的综合指标 F_1 来表达，R 表示被正确分类的地层样本占所有实际为该地层的样本比例，P 表示被正确分类的地层样本占所有预测为该地层的样本比例，F_1 是召回率 R 和查准率 P 的综合指标。随机森林模型的评价指标公式[16]如下：

$$R = \frac{n_{ii}}{\sum_{i=1}^{K} n_{ij}} \tag{3}$$

$$P = \frac{n_{ii}}{\sum_{i=1}^{K} n_{ji}} \tag{4}$$

$$F_1 = \frac{2 * P * R}{P + R} \tag{5}$$

RF 在训练过程中每次的 bootstrap 抽样，N 个地层数据中的每条数据未被抽中的概率为 $(1-1/N)^N$，当 N 足够大时，$(1-1/N)^N \approx 0.368$，即为袋外数据(Out Of Bag, OOB)。RF 利用这部分数据进行内部误差

估计，产生 OOB 误差，为在 RF 中为测试集误差的无偏估计，可利用 python 的机器学习库 Scikit-learn 直接输出。有 2 个主要参数会影响 RF 的效率和性能[17]：树的数量以及叶节点的最小样本数量，可以使用网格参数搜索的办法来确定其最优超参数。

为量化地层分类预测的可靠性，引入余量函数(Margin function) [18]，定义为：

$$mg(X,Y)=av_k\left(h_k(X)=Y\right)-\max_{j\neq Y}av_k I\left(h_k(X)=j\right) \tag{6}$$

式中：$I(*)$为示性函数，余量函数用于度量平均正确分类数超过平均错误分类数的程度，余量值越大，分类预测越可靠。随机森林的泛化误差[18]定义为：

$$PE^{*}=P_{X,Y}\left(mg(X,Y)<0\right) \tag{7}$$

式中：下标 X,Y 表示概率 P 覆盖 X、Y 空间。在 RF 当中，当决策树数量足够多时，PE^{*} 会趋于一个上界，RF 算法不会过拟合(Overfitting) [19]。

3. 应用算例

3.1. 研究区概况和钻孔数据分析

研究区域位于南宁市中心城区，选取地铁 1 号线 170 个钻孔勘察资料进行研究，其钻孔点位主要位于白苍岭站、南宁火车站站、朝阳广场站、新民路站、民族广场站，地铁 5 号线和 7 号线也将穿过本研究区域。朝阳广场站规划成为地铁 1、2、7 号线三线换乘车站，也是南宁市已有规划中唯一一座三线换乘站。钻孔揭露的地层主要为杂填土、粉土、泥岩、粉砂岩、砾石、黏土、粉砂共 7 类，统计的对应地层样本数量分别为 124、75、221、101、131、180、68，总共 900 个样本。

首先将钻孔数据按钻孔编号整理，然后根据地层类型的不同，划分为不同的样本，表 1 为随机选取的 2 个钻孔数据及揭露的部分地层，展示了相关特征属性。钻孔剖面图如图 3 所示，训练中涉及的属性值包括钻孔位置坐标，每个地层分界点的起始深度、终止深度、层厚、钻孔地面标高。

Table 1. Drilling data and feature attributes
表 1. 钻孔数据及特征属性

钻孔编号	钻孔位置横坐标	钻孔位置纵坐标	地层分界点起始深度	地层分界点终止深度	地面标高	地层层厚	地层名称
MAZ3-CY-34	532,528	2,524,505	0	4.6	76.78	4.6	杂填土
			4.6	8	76.78	3.4	黏土
			8	11	76.78	3	黏土
			11	12	76.78	1	黏土
			12	14.4	76.78	2.4	粉砂
			14.4	21	76.78	6.6	砾石
MAZ3-CY-35	532,535.1	2,524,485	10	11.7	76.63	1.7	黏土
			11.7	13	76.63	1.3	黏土
			13	14	76.63	1	粉砂
			14	16.2	76.63	2.2	砾石
			16.2	45.1	76.63	28.9	泥岩

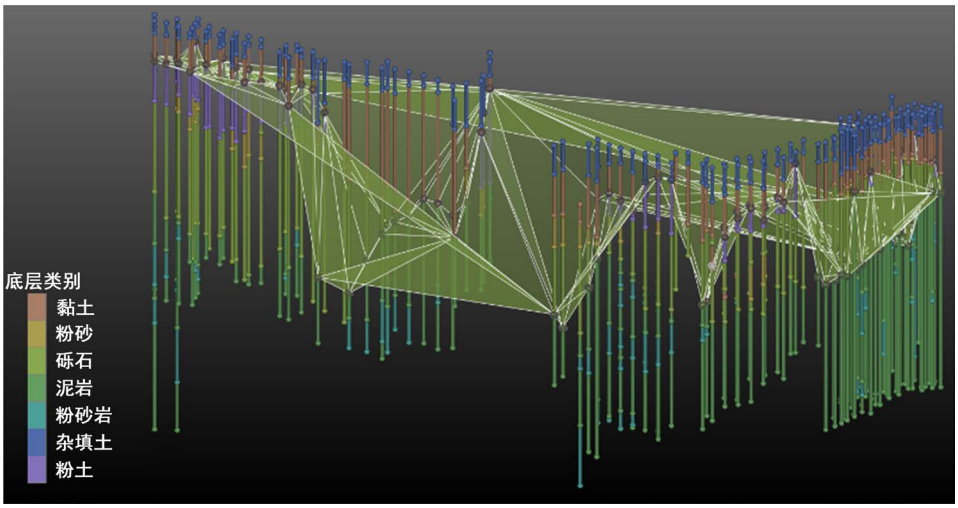


Figure 3. Drilling profile
图 3. 钻孔剖面图

数据样本中钻孔坐标和土层厚度数量级差异很大，为了提高分类器的学习能力，对每个输入特征值进行标准化[20]，将处理后的输入数据标准化为零均值和单位方差，转化函数为

$$X^* = (X - \mu)/\sigma \tag{8}$$

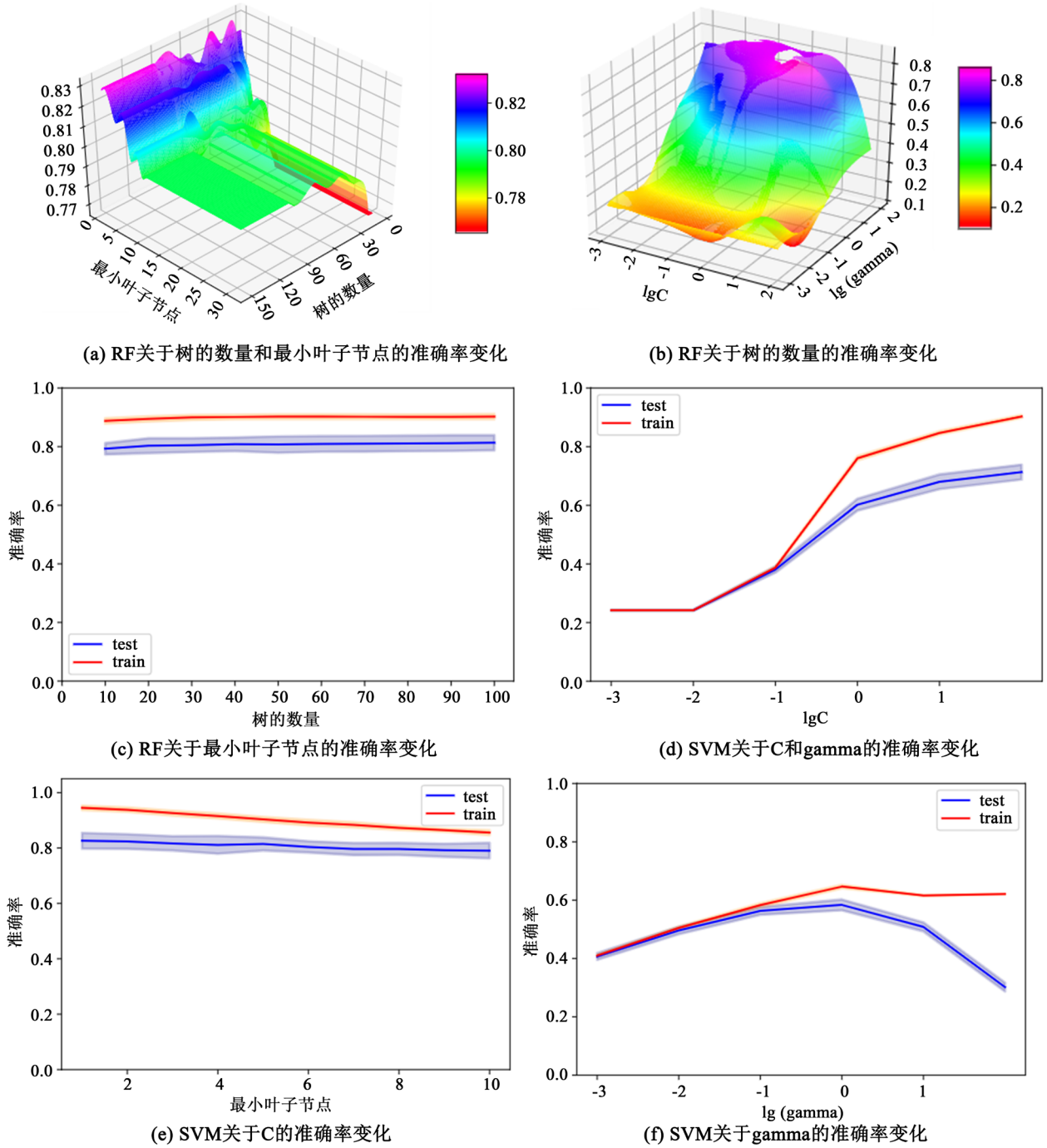
式中： X^* 为标准化后的值；X 为待标准化的值； μ 为样本数据的均值； σ 为样本数据的方差。

3.2. 超参数的敏感性分析

机器学习方法中，对于不同的建模数据，超参数难以确定唯一值，而不同的超参数组合会对建模结果产生很大影响。RF 的性能主要受树的数量以及最小叶子节点数的影响，而 SVM 的性能则受到惩罚因子 C 和 RBF 核函数参数 gamma 这 2 个超参数的影响[21]。因此，需要研究 RF 和 SVM 模型的超参数选取对建模准确率的影响。RF 与 SVM 模型的超参数敏感性分析结果如图 4 所示。在机器学习中，常采用交叉验证分析超参数的敏感性[22]，为方便比较，统一采用 5 折交叉验证和网格搜索方法计算分析不同参数设置下模型分类准确率，结果显示 RF 分类器整体表现更好，且具备较低的超参数敏感性。图 4(a)显示 RF 的两个重要超参数树的数量以及最小叶子节点数变化时，分类准确率的波动很小，图 4(b)和图 4(c)中 RF 的 2 个参数敏感性曲线都很平滑，不存在过拟合(over-fitting)与欠拟合(under-fitting)情况，测试集的波动范围相比于整体准确率而言很小，体现了模型分类的稳定性。图 4(d)、图 4(e)、图 4(f)结果表明，SVM 的参数非常敏感，波动范围很大，甚至会出现训练集和测试集结果相背离的情形，需要重点调整优化关键参数。图 4(e)结果还表明，当 C 值设置不当时，总体精度只有大约 20%。同时，*gamma*值设置对分类性能也有明显影响(见图 4(f))。在实际应用中，使用粗网格搜索可能难于选定最优 SVM 参数，而使用精细网格进行计算，无疑会加大计算工作量。

3.3. 地层识别结果分析

为验证建模准确率，按照 4:1 划分训练集与测试集，即随机选择 180 个地层数据(占全部地层数据的 20%)作为测试集用以评判建模结果，利用 Scikit-learn 机器学习包中的网格搜索进行模型参数调优[23]，混淆矩阵分析方法是评价模型性能好坏最直接有效的方法[24]。图 5 和图 6 以混淆矩阵形式分别给出 RF 和 SVM 的分类预测结果。通过对比发现，2 种模型都具有较好的分类能力，但 RF 模型在不同地层的分



(a) RF关于树的数量和最小叶子节点的准确率变化

(b) RF关于树的数量的准确率变化

(c) RF关于最小叶子节点的准确率变化

(d) SVM关于C和gamma的准确率变化

(e) SVM关于C的准确率变化

(f) SVM关于gamma的准确率变化

Figure 4. Hyperparameter sensitivity map of Random Forest (a)~(c) and Support Vector Machine (d)~(f) under 5-fold cross-validation.
图 4. RF(a)~(c)和 SVM (d)~(f)在 5 折交叉验证下的超参数敏感性图

类结果几乎均强于 SVM 模型。表 2 给出了利用公式(2)、(3)、(4)、(5)得到的模型整体精度、综合指标 F_1 值以及 5 折交叉验证得到的结果，随机森林模型 3 项指标分别为 0.817、0.816 和 0.824，均略高于 SVM，另外，RF 的 OOB 值为 0.824，具备较好的泛化性能。

为应对实际工程中钻孔数据样本有限的问题，此处验证 RF 对样本量的鲁棒性，随机选择不同数量样本，因需要验证在样本数量较少时的模型性能，所以此处按 7:3 划分样本集进行模型训练测试，以减

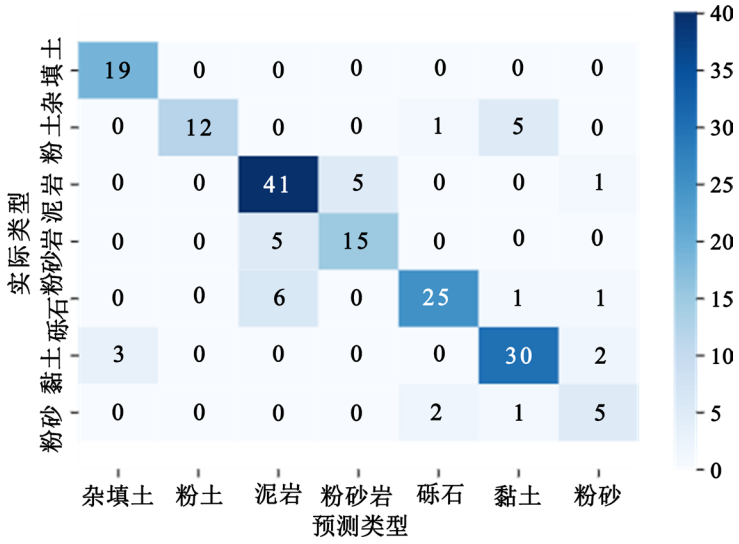


Figure 5. Confusion matrix of the RF classifier
图 5. RF 分类器分类结果的混淆矩阵

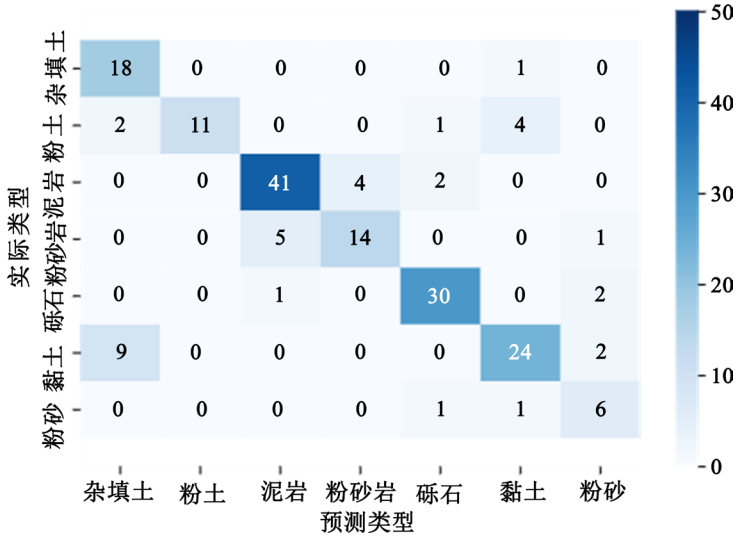
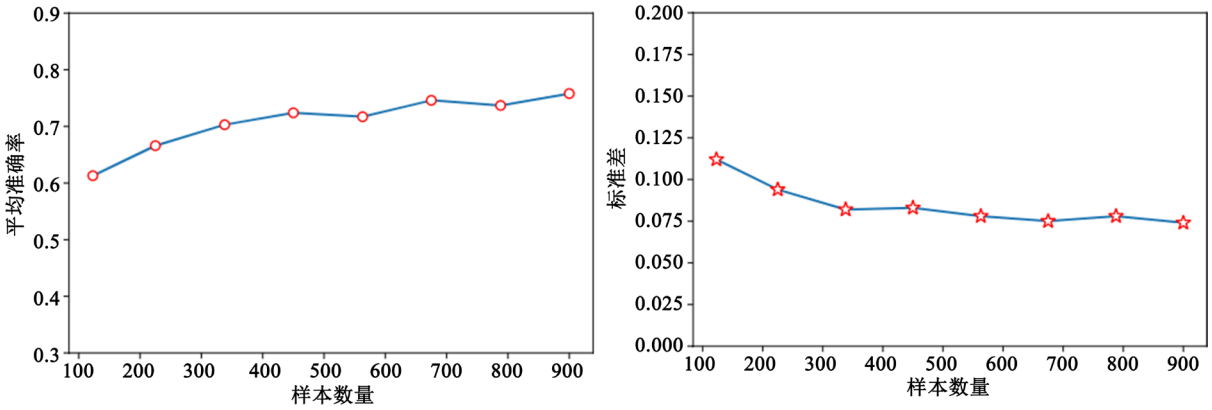


Figure 6. Confusion matrix of the SVM classifier
图 6. SVM 分类器分类结果的混淆矩阵

Table 2. Classification results of RF and SVM classifiers
表 2. RF 与 SVM 模型的分类结果

	Accuracy	F1	CV	模型参数
随机森林	0.817	0.816	0.824	n_estimators = 80, criterion = 'gini', max_depth = 9, min_samples_leaf = 1, min_samples = 5
支持向量机	0.8	0.801	0.817	C = 10, gamma = 1, kernel = 'rbf'

小测试误差。图 7 是不同样本数量下的测试集分类结果。由图 7(a)可知，RF 运行 100 次后获得的平均精度随样本数量的增加而增加，图 7(b)显示标准差则会随之减小。样本数量由 100 增至 900 时，平均精度增加约 20%，标准偏差下降约 0.05。计算结果表明 RF 对于样本集数量的标准差稳定性较好，样本数量为 300 以上时，其准确率就达到 70%以上，基本可以满足实际工程的需求。



(a) 随机森林关于样本数的平均准确率变化　　(b) 随机森林关于样本数的标准差变化

Figure 7. Mean accuracy and its standard deviation (RF runs 1000 times) versus the number of samples
图 7. 随机森林随机划分 1000 次数据集，平均准确度及其标准偏差与样本数的关系

4. 结语

本文以南宁市地铁 1 号线的钻孔勘探资料为研究基础，提出了基于钻孔数据的随机森林地层分类方法，比较分析了随机森林和支持向量机 2 种机器学习算法在地层岩性分类中的应用，得出以下主要结论：

1) 分类模型评价指标为总体准确率和综合指标 F_1 值，随机森林的准确率达到 81.7%，F_1 值为 0.816，不论是整体的分类能力还是各个地层的分类能力随机森林均强于支持向量机，它们的交叉验证结果也在 0.8 以上，保证了其泛化能力，同时随机森林在实际应用当中还可以不用划分测试集，用 OOB 误差精准便捷地评价其泛化能力。

2) 在超参数敏感性方面，与支持向量机比较，随机森林的参数敏感性更低，这在实际应用中会更加便捷、快速。与此同时，随机森林模型对于样本集数量的要求低，在低样本数量时得到的地层预测准确率和标准差良好。

3) 随机森林模型在利用钻孔数据识别地层方面具有明显的优越性，可以有效解决城市区域岩土工程勘探钻孔有限、稀疏的问题，对后续南宁地铁工程建设具有一定指导意义。

基金项目

国家重点研发计划资助项目(2017YFC0803300)，南宁市科技局重点研发项目(02902530072)。

参考文献

[1] 周翠英, 张国豪, 杜子纯, 等. 基于机器学习的地层序列模拟[J]. 工程地质学报, 2019, 27(4): 873-879.

[2] 贺怀建, 白世伟, 赵新华, 等. 三维地层模型中地层划分的探讨[J]. 岩土力学, 2002, 23(5): 637-639.

[3] 刘晓明, 罗周全, 杨彪, 等. 复杂矿区三维地质可视化及数值模型构建[J]. 岩土力学, 2010, 31(12): 4006-4010+4015.

[4] 刘智勇. 曲面插值算法在三维地质建模中的研究[D]: [硕士学位论文]. 成都: 成都理工大学, 2016.

[5] 陈爱侠, 关卫省, 陈宽民. 轨道交通建设对城市生态环境影响分析——以西安市城市轨道交通 2 号线为例[J]. 安全与环境学报, 2007(3): 70-73.

[6] 周念清, 邹朴, 黄钟晖, 等. 多物性参数耦合建模探测地铁站岩溶空间概率分布及风险[J]. 地球科学前沿, 2019, 9(10): 13.

[7] Breiman, L. (2001) Random Forests. *Machine Learning*, **45**, 5-32. https://doi.org/10.1023/A:1010933404324

[8] 陈玉林, 李戈理, 杨智新, 等. 基于 KNN 算法识别合水地区长 7 储层岩性岩相[J]. 测井技术, 2020, 44(2): 182-185.

[9] 郭甲腾, 刘寅贺, 韩英夫, 等. 基于机器学习的钻孔数据隐式三维地质建模方法[J]. 东北大学学报(自然科学版), 2019, 40(9): 1337-1342.

[10] Cracknell, M.J. and Reading, A.M. (2014) Geological Mapping Using Remote Sensing Data: A Comparison of Five Machine Learning Algorithms, Their Response to Variations in the Spatial Distribution of Training Data and the Use of Explicit Spatial Information. *Computers and Geosciences*, **63**, 22-33. https://doi.org/10.1016/j.cageo.2013.10.008

[11] 马梓程, 帅爽, 安志宏, 等. 基于 RF 模型的火成岩提取与分类研究——以吉布提阿里萨比耶地区为例[M]//国家安全地球物理丛书(十五)——丝路环境与地球物理. 西安: 西安地图出版社, 2019: 94-103.

[12] 徐剑波, 陈军林. 利用区域化探数据推断地质体空间分布[J]. 地质与勘探, 2019(5): 1214-1222.

[13] Breiman, L. (2001) Statistical Modeling: The Two Cultures (with Comments and a Rejoinder by the Author). *Statistical Science*, **16**, 199-215. https://doi.org/10.1214/ss/1009213726

[14] Breiman, L., Friedman, J.H., Olshen, R.A., *et al*. (1984) Classification and Regression Trees (CART). *Biometrics*, **40**, 358. https://doi.org/10.2307/2530946

[15] 冯少荣. 决策树算法的研究与改进[J]. 厦门大学学报(自然科学版), 2007, 46(4): 496-500.

[16] 曹正凤. 随机森林算法优化研究[D]: [硕士学位论文]. 北京: 首都经济贸易大学, 2014.

[17] 马骊. 随机森林算法的优化改进研究[D]: [硕士学位论文]. 广州: 暨南大学, 2016.

[18] 姚登举, 杨静, 詹晓娟. 基于随机森林的特征选择算法[J]. 吉林大学学报(工学版), 2014(1): 142-146.

[19] Sun, A. and Lim, E.P. (2001) Hierarchial Text Classification and Evaluation. *IEEE International Conference on Data Mining*, San Jose, 29 November-2 December 2001, 521-528.

[20] 杨根云, 周伟, 方教勇. 基于信息量模型和数据标准化的滑坡易发性评价[J]. 地球信息科学学报, 2018, 20(5): 674-683.

[21] 范昕炜. 支持向量机算法的研究及其应用[D]: [博士学位论文]. 杭州: 浙江大学, 2003.

[22] 郑茂辉, 刘少非, 柳娅楠, 等. 基于粒子群优化极限学习机的排水管结构状况评价[J]. 同济大学学报(自然科学版), 2020, 48(4): 513-516, 551.

[23] Pedregosa, F., Varoquaux, G., Gramfort, A., *et al*. (2011) Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, **12**, 2825-2830.

[24] 张雪蕾, 汪明, 曹寅雪, 等. 3 种典型机器学习方法在灾害敏感性评估中的对比[J]. 中国安全生产科学技术, 2018, 14(7): 80-85.