基于镁铁-超镁铁岩中单斜辉石主量元素含量的决策树集成算法对比

孙建鹞[1,2],杜雪亮[3],章宝月[1],王 龙[1],金维浚[4],张 旗[4],罗 熊[1,2],朱月琴[2]
SUN Jiankun[1,2], DU Xueliang[3], ZHANG Baoyue[1], WANG Long[1], JIN Weijun[4], ZHANG Qi[4], LUO Xiong[1,2], ZHU Yueqin[2]

1. 北京科技大学计算机与通信工程学院,北京 100083;
2. 自然资源部地质信息技术重点实验室,北京 100037;
3. 兰州大学地质科学与矿产资源学院/甘肃省西部矿产资源重点实验室,甘肃 兰州 730000;
4. 中国科学院地质与地球物理研究所,北京 100029

1. *School of Computer and Communication Engineering, University of Science and Technology Beijing (USTB), Beijing* 100083, *China;*
2. *Key Laboratory of Geological Information Technology, MNR, Beijing* 100037, *China;*
3. *Key Laboratory of Mineral Resources in Western China (Gansu Province)/School of Earth Sciences, Lanzhou University, Lanzhou* 730000, *Gansu, China;*
4. *Institute of Geology and Geophysics, Chinese Academy of Sciences, Beijing* 100029, *China*

摘要:依靠岩浆构造环境的地球化学成分认识岩浆形成过程是岩石地球化学中的重要应用。当前利用岩石地球化学成分判别构造环境的工作还不够深入。用4种基于决策树的机器学习方法对来自全球新生代洋岛玄武岩(OIB)、岛弧玄武岩(IAB)及大洋中脊玄武岩(MORB)等镁铁-超镁铁岩中单斜辉石的13种主量元素构成数据集进行了岩浆构造环境判别和主要特征排序。通过对比4种基于决策树的机器学习方法,验证了树类算法对于地球化学成分识别问题的有效性,并总结出4种方法在处理岩浆构造环境判别问题时的优劣:决策树算法判别过程更易于理解,但是其准确率欠佳;boosting 算法中的 AdaBoost 和 GBDT 对于岩浆构造环境的鉴别准确度较高,但构造过程复杂;bagging 集成算法随机森林在权衡性能和模型可理解性时不失为一个良好的选择。此外,还通过4种算法的特征重要性排序得出 Cr_2O_3,TFeO,TiO_2,FeO 和 Al_2O_3 是进行岩浆构造环境判别的重要成分。

关键词:树算法;bagging 算法;boosting 算法;单斜辉石;岩浆构造环境判别;地球化学特征
中图分类号:P578.594;P628 **文献标志码**:A **文章编号**:1671-2552(2019)12-1981-11

Sun J K, Du X L, Zhang B Y, Wang L, Jin W J, Zhang Q, Luo X, Zhu Y Q. A comparison of tree-based ensemble algorithms on the main element content of monoclinal pyroxene in mafic-ultramafic rocks. *Geological Bulletin of China*, 2019, 38(12):1981-1991

Abstract: Relying on the geochemical composition of the magma tectonic environment to understand the formation process of magma is an important application in rock geochemistry. While the current works to make full use of rock geochemical components

收稿日期:2019-04-16;**修订日期**:2019-07-23
资助项目:国家重点研究开发计划《基于"地质云"平台的深部找矿知识挖掘》(编号:2016YFC0600510)、国家自然科学基金项目《大数据环境下的滑坡危险性评估模型构建方法研究》(批准号:41872253)、国土资源部地质信息技术重点实验室课题《基于知识图谱深度优化技术的地质大数据智能检索服务应用研究》(编号:2017320)、中国地质调查局项目《国家地质大数据汇聚与管理》(编号:DD20190381A)、《资源环境重大问题综合区划与开发保护策略研究》(编号:DD20190463)
作者简介:孙建鹞(1996-),男,在读博士生,从事机器学习和地质信息技术研究。E-mail: sunjk@xs.ustb.edu.cn
通讯作者:罗熊(1976-),男,教授,博士生导师,从事机器学习和地质信息技术研究。E-mail: xluo@ustb.edu.cn

for the tectonic setting discrimination are not enough. In this study, the authors utilized four tree-based machine learning methods to make magma tectonic environment discriminations and feature sorting on the 13 main ingredients of monoclinal pyroxene in mafic-ultramafic rocks from global Cenozoic ocean island (OIB), island arc (IAB), and mid-ocean ridge (MORB). Through the comparison of the four tree-based machine learning methods, the authors proved the validity of the tree-based methods for the identification of geochemical components and derived the advantages and disadvantages of the four methods in dealing with the identification of rock tectonic environments: decision trees gain better comprehensibility but have lower recognition accuracy, boosting algorithms AdaBoost and GBDT have the best recognition accuracy but lower comprehensibility, and random forest is a better choice during trading off and comprehensibility performance. Besides, Cr_2O_3, TFeO, TiO_2, FeO and Al_2O_3 are figured out as the most important ingredients for magma tectonic environment discriminations on this dataset.

Key words: tree algorithm; bagging algorithm; boosting algorithm; magmatic environment discrimination; clinopyroxene; geochemical characteristics

由于区分不同地质环境下的岩浆活动对于岩浆形成过程研究的重要性,岩浆构造环境判别已经成为地球化学领域一个重要的研究方向。在野外,岩石遭受蚀变是常见现象,而蚀变强烈的岩石不利于判别构造环境。但是,组成岩石的某些矿物抗风化和蚀变的能力较强,常可以见到在强烈蚀变的岩石中还有单斜辉石的残晶保留,这些残晶可能保留了原始未受蚀变的岩石的某些信息,因此可利用单斜辉石的残留恢复原岩的成分、岩石成因及其形成的构造环境[1-2]。Nisbet 等收集了329个单斜辉石数据[3],开创了利用单斜辉石矿物成分恢复岩浆岩构造环境的先河,并取得了较好的结果,此后,许多学者也对单斜辉石产出的构造环境进行了研究,结果表明,在组成岩石的造岩矿物和副矿物中,以单斜辉石和尖晶石的效果较好,橄榄石、斜方辉石、斜长石、角闪石及副矿物锆石、榍石、磷灰石、帘石类(绿帘石、褐帘石、斜黝帘石等)等的效果较差[3-5]。其中又以单斜辉石最受学术界的青睐,因为:①单斜辉石是镁铁-超镁铁岩中最常见的造岩矿物;②单斜辉石结晶顺序较晚,保留了岩浆演化时所经历的成分、温度、压力等的变化;③单斜辉石的化学成分往往与构造环境密切相关[3]。因此,在矿物学研究中对单斜辉石的分析较深入。但是,由于研究手段和思路的限制,利用单斜辉石作为构造环境判别指标的研究虽然取得了一些进展,总体效果并不明显。早期的判别图只考虑少量矿物元素[6-7],地球化学研究由于条件限制只能采用抽样或典型的方法进行[8-10], 如图1可见常用的判别图[11-14]。而基于少量特征的判别图方法在面对全体数据时就显得捉襟见肘。笔者先前的研究表明:①目前的玄武岩构造环境判别图都存在问题,尤其是主量元素判别图存在多解性、矛盾性和不准确性。②由于组成成分扩展和大多数早期的识别标志被突破,大洋中脊玄武岩(MORB)和洋岛玄武岩(OIB)不易区分。③岛弧玄武岩(IAB)和其他2种岩石的判别图存在重叠,只有包含 Th, Ta(Nb), Sr 的判别图依然起效[16-18]。与此同时,全部数据的判别图表明各类构造背景中的镁铁-超镁铁岩组成成分变化很大,先前的判别图表现出了它们的局限性[19-20]。

一些用于判别岩浆构造环境和从全部数据中找出最具有区分度元素的方法应时而出,基于全球聚合数据的机器学习方法进一步推动了这一领域的研究[21-22]。在这些机器学习方法中,基于决策树的方法由于较高的可解释性和良好的分类性能被广大研究人员广为采用。Vermeesch[2]和韩帅等[21]分别使用决策树算法和随机森林算法对玄武岩构造环境判别问题进行了分析,朱林奇等[24]和韩启迪等[25]分别将AdaBoost算法和GBDT算法应用于岩性识别问题。基于决策树的方法在地球化学研究中具有良好的表现,但是对决策树类方法(尤其是boosting和bagging集成方法)在地球化学领域的比较分析仍是亟需研究的课题。

本文笔者基于在线 GEOROC 数据库(http://georoc.mpch-mainz.gwdg.de/georoc/)和 PetDB 数据库(https://www.earthchem.org/petdb)[26]中单斜辉石的主量元素,使用决策树、随机森林、Adaboost 和 GBDT 方法对来自3种不同构造背景(OAB, IAB 和 MORB)的镁铁-超镁铁岩数据进行了对比研究,为进一步判别岩浆环境提供技术支撑。本文比较了4种树类算法使用单斜辉石地球化学元素数据在镁铁-超镁铁岩环境判别问题上的性能,为地球化学分类研究的树类机器学习算法选择提供参考依据,同时得出在本数据集中区分不同构造

环境的5种主要元素。

1　树类算法

大量机器学习方法在分类问题上有良好表现,如支持向量机(SVM)、极限学习机(ELM)、人工神经网络(ANN)、决策树及集成算法或深度学习算法[27-29]。但是,大多数方法,如SVM,ANN,ELM和深度学习算法将原始输入映射到更高维的空间,以黑箱模型的方式工作,具有较低的可解释性[30]。而决策树方法的树形结构生成的规则符合人类的决策方式,易于理解,因此决策树方法具有较高的可解释性[31-32],适用于岩浆构造环境判别问题的研究。

随着集成技术的出现,为进一步提高性能,结合决策树和集成技术的判别方法也得以发展。基于决策树的集成算法主要可分为2类:bagging 算法和boosting 算法。bagging 算法,通过结合多棵协作决策分类树进行投票决策,通常能取得更好的泛化性能。随机森林方法是常见的bagging扩展变体。boosting算法是一种串行的集成算法,其在每轮迭代产生弱分类器后,加权求和所有的弱分类器来获得最终分类结果。该算法关注于上一轮错分样本,因此常表现出更好的分类性能。广泛使用的boosting算法是AdaBoost算法和GBDT算法。以下简单描述四种经典的树类算法,即决策树、随机森林、AdaBoost和GBDT。

1.1　决策树

决策树是一种具有强鲁棒性、低计算消耗、高可解释性等优点的机器学习方法,其被广泛应用于分类任务[33]。近年来,许多决策树算法得以快速发展,如ID3,C4.5,CART等。在决策树算法中,决策规则以树的形式组织构造,其中非叶结点表示决策条件,将结点数据分割为更小的结点;叶结点为终止结点,表示决策规则的终点。图2展示了CART算法的一个示例。图2-a为决策树的一种分割模式,其中a_1和a_2表示输入特征,m_1, m_2, m_3, n_1, n_2为分割点。图2-b 表征图2-a相应的判别图,当决策条件得到满足时,左子树工作,否则右子树工作。本次采用的决策树方法为CART算法。

1.2　随机森林

随机森林是一种实用的分类算法,该分类算法的类别确定通过多棵子树投票实现。作为一类基于树的bagging 集成算法,随机森林构建多棵不同

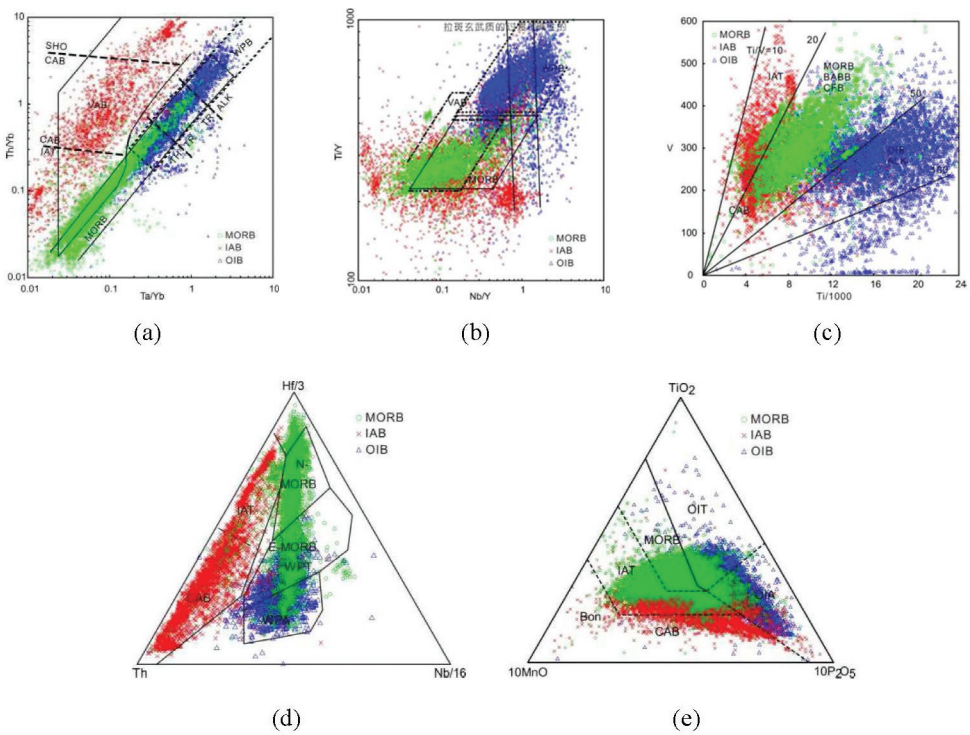

图1　文献中常用经典若干玄武岩构造环境判别图[15]

Fig. 1　Some classical magmatic environment discrimination diagrams

a—Th/Yb-Ta/Yb图解[11];b—Ti/Y-Nb/Y图解[11];c—V-Ti图解[12];d—Hf/3-Th-Nb/16图解[13];e—TiO_2-10MnO-$10P_2O_5$图解[14]

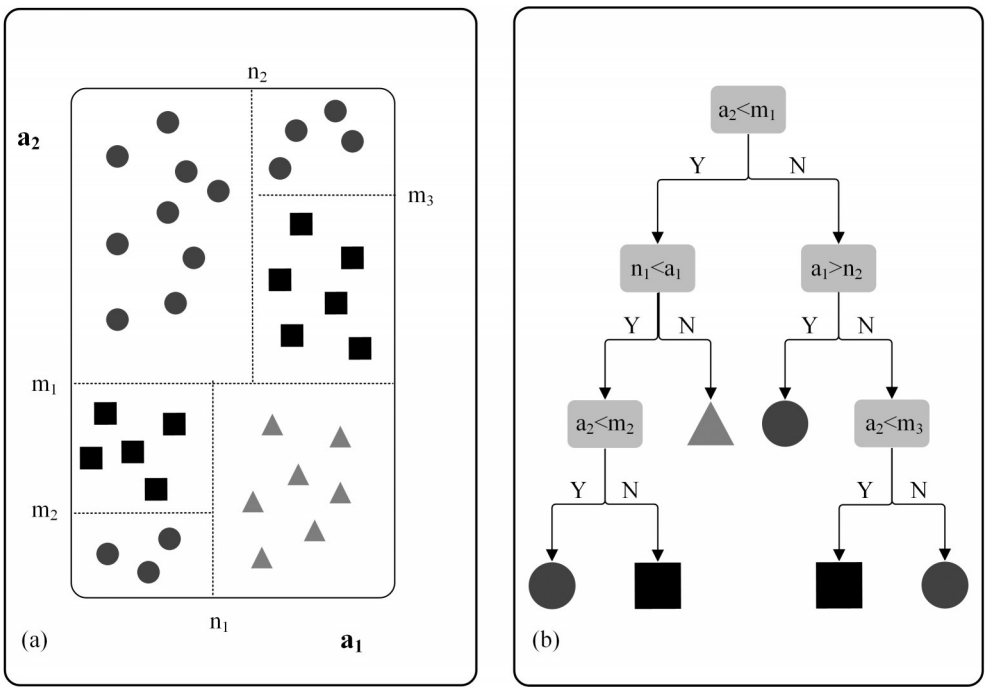

图2 CART示例

Fig. 2 An example of CART

的决策树,其中每一棵子树拥有相同规模的训练数据。为避免过拟合,一方面随机森林采用依赖替换实例的bootstrap方法[34]从原始数据采样获取每棵子树的训练数据,另一方面,随机森林通过随机采样划分每棵子树的训练特征。

1.3 AdaBoost

起初用于特征分类和回归的AdaBoost算法已经成为机器学习领域中最成功的分类算法之一。同bagging 算法相比,AdaBoost算法更关注于错分类的样本,其通过串行的方式来构造学习器[35]。即一个弱学习器先被构造起来,然后下一个分类器的样本权重基于在先前构造中错分类样本来调整。通过重复"构建与调整"过程,一系列侧重于不同样本的学习器被构建起来,而最终的分类结果通过先前的弱分类器加权得到。不同于随机森林,集成AdaBoost算法中的多棵决策树分别拥有不同的权重。

原始的AdaBoost算法只关注于二分类问题,而SAMME 和SAMME.R 算法是2种常见的可用于多分类任务的AdaBoost变种算法,这2种算法的性能都已在实践中得以验证[35]。本文使用性能更好的SAMME.R方法作为AdaBoost的实现进行三分类构造环境判别问题分析。

1.4 GBDT

梯度提升决策树(GBDT)在许多领域都属于一种高泛化性能技术。不同于AdaBoost,GBDT方法主要关注损失函数而不是错分类的样本,最终的强分类器只与前一个学习器有关。需要注意的是,GBDT中的弱学习器必须是决策回归树。

梯度提升决策树由多棵决策树组成,所有树结果的累计输出作为最终决策结论。梯度提升决策树在每次构造新的弱分类器时,关注于之前所有树的结论和残差,通过拟合得到当前残差回归树。与单决策树相比,梯度提升决策树通过弱化训练数据的微小变动对弱分类器的影响来支持最终强分类器的决策结论。

2 实 验

本文对4种树类算法,即决策树、随机森林、AdaBoost和GBDT在岩浆判别环境问题上的性能进行评估。

2.1 岩石数据

本文中使用的实验数据主要来自于2个在线数据库:GEOROC 和 PetDB,它们富含全球新生代镁铁-超镁铁质岩石单斜辉石数据[32]。数据主要包含

来自3种环境的玄武岩,即岛弧玄武岩、洋岛玄武岩和大洋中脊玄武岩,详细数据的地理分布如图3所示;数据的降维可视化可由基于t-SNE的散点图得出,如图4所示。实验数据集包含13种主量元素:SiO_2, TiO_2, Al_2O_3, Cr_2O_3, Fe_2O_3, TFeO, FeO, CaO, MgO, MnO, NiO, K_2O, Na_2O。实验使用python3.6环境在一台Intel(R) Core(TM) i5-4200M, 2.50GHZ, 8.00GB的机器上进行。

在进行分类任务前,需要对原始数据进行有效的清洗[17]。本次实验采用的数据集在初步清洗后仍然存在部分数据缺失,因此需要进行数据填充。由于空缺位的实际元素含量远低于测量仪器范围,通过结合专家知识和常用策略,本文使用0.001填充缺失值[36]。经过初步数据清洗后,共获得大洋中脊玄武岩有效数据795条,洋岛玄武岩有效数据198条,岛弧玄武岩有效数据329条,各项数据统计信息见表1。

由表1可见,数据集中各类玄武岩的样本比例不平衡,为缓解样本不平衡对模型分类性能的影响,本文主要从以下角度考虑问题:①决策树方法通常在不平衡数据上表现良好,而集成算法自身的加权特性使集成方法能更好地训练不平衡数据,本文使用的方法对于处理不平衡数据具有天然的优势;②对训练集的分类进行类加权处理,即每个类的输出概率均嵌入成本误差信息;③在性能度量时,使用多种度量指标,混淆矩阵、F1值等能更好地衡量模型在不平衡数据集上的表现。

2.2　性能测量

准确性(Accuracy)是分类问题性能评估的直观选择。此外,宏平均精度(MaP)、宏平均召回率(MaR)和宏平均F1值(MaF)也是多分类问题常用的指标。MaP表示每一类样本精确度的算术平均值,MaR表示每一类样本召回率的算术平均值,MaF是基于MaP和MaR的调和平均值。本次实验使用以上指标评估4种树类方法的性能,此外,本文使用混淆矩阵和散点图来更直观地展示实验分类效果。

2.3　参数选择

模型参数选择是机器学习算法建模过程中不可避免的过程。考虑到本文中模型参数数量较多,使用贝叶斯优化方法进行参数调整[37]。实验最终确定参数如表2所示。

其中:“max_features”为决策树在进行最优分割时使用的特征占总特征的比例,参数搜索范围为[0.1,1];“max_depth”为树的最大深度,搜索范围为[5,50];“min_samples_split”为分割决策树结点所需最小样本量,参数搜索范围为[2,6];“min_samples_leaf”为叶结点即一条决策规则终止时此结点所需的最小样本量,参数搜索范围为[1,5];“n_estimators”为集成算法中子树的数量,参数搜索

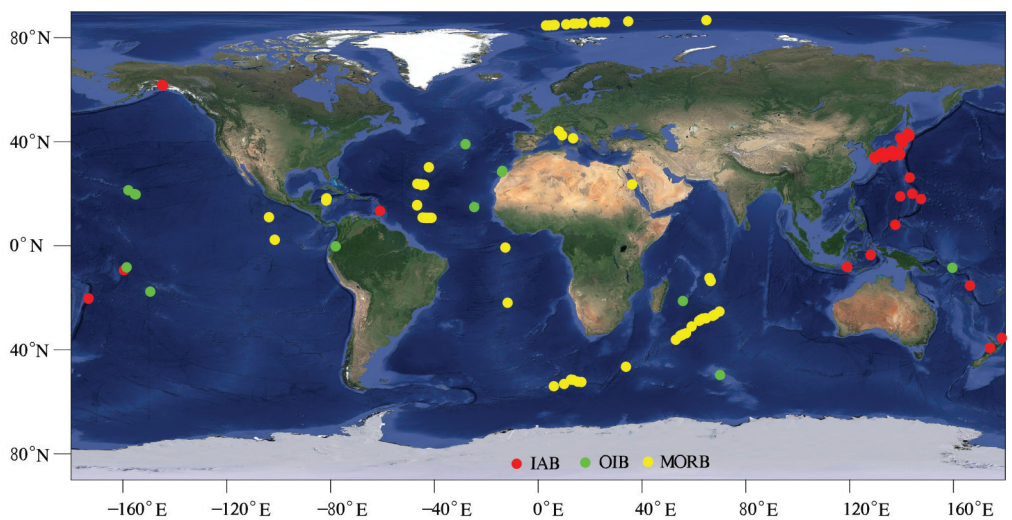

图3　镁铁-超镁铁岩中单斜辉石分布

Fig. 3　Distribution for clinopyroxene of mafic-ultramafic rocks

IAB—岛弧玄武岩;MORB—大洋中脊玄武岩;OIB—洋岛玄武岩

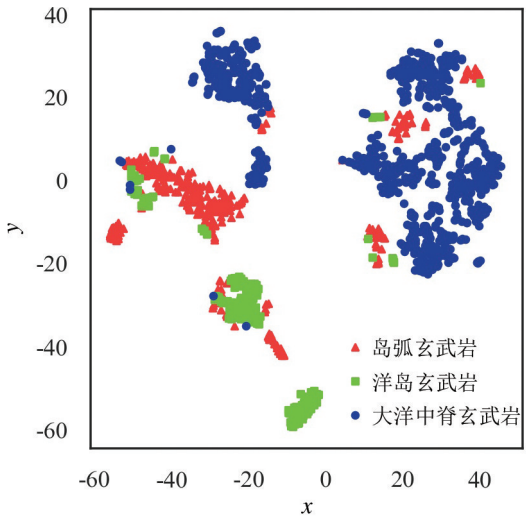

图4 基于t-SNE的散点图

Fig. 4 The t-SNE-based scatter diagram

范围为[20,1000]；“learning_rate”控制boosting算法中每棵树的贡献度，搜索范围为[0.1,1]；“subsample”表征GBDT算法中用于拟合基学习器所使用的样本比例，搜索范围为[0.5,0.8]；“-”代表该算法没有相应的参数需要调整[38]。

3 实验结果

本部分探讨树类模型在单斜辉石主量元素数据集进行10次独立重复实验的对比结果。首先展示基于决策树算法生成的决策图，之后基于随机选择的一次实验结果绘制混淆矩阵，然后列出了4种模型的特征重要性排序，并基于最重要的2个特征绘制散点图以更直观地展示模型的性能，之后采取10次重复实验的平均指标对比了4种模型的分类性能。

3.1 基于决策树的决策图

同其他基于神经网络机器学习的算法相比，决策树算法的一个巨大优势是其优秀的可解释性。决策树图可以通过树形结构清晰地展示其决策过程。考虑到生成决策树的规模，图5只展示了部分决策树的决策图。

图5中的每个结点数据表征当前的统计信息。以根节点为例，TFeO≤4.575表示分类决策条件，这个判别条件将样本分割为“True”和“False”两部分。samples = 100.00%表示当决策到当前结点时所剩余的样本占全体样本的比例。gini = 0.667表示基于当前样本集的Gini指数。values = [0.333, 0.333, 0.333]是当前样本集中每类样本的比例。需要注意，Gini指数为0的结点代表当前结点为叶结点，它们表征决策终点。

3.2 混淆矩阵

混淆矩阵是分类结果的数值化表示，每个类别（IAB / OIB / MORB）的实际样本数量和预测得出的样本数量都易从混淆矩阵中得出，混淆矩阵的判别结果是进行分类性能比对和散点图绘制的基础。图6为4种方法在当前数据集建立模型所得的混淆矩阵，其中每一类的样本数量清晰可见，如图5中对于决策树算法，有7个标签为IAB类的样本被决策树模型错误地预测为OIB类，而有55个标签为IAB类的标签被决策树模型预测正确。

3.3 特征重要性排序和散点图

为进一步探讨影响构造环境判别的主要特征

表1 镁铁-超镁铁岩中单斜辉石主量元素统计信息

Table 1 Major element content of clinopyroxene in mafic-ultramafic rocks in the dataset

主量元素	IAB（岛弧玄武岩）			OIB（洋岛玄武岩）			MORB（大洋中脊玄武岩）		
	数据量	平均数 /%	中位数/%	数据量	平均数 /%	中位数/%	数据量	平均数 /%	中位数/%
SiO_2	329	52.09	52.30	198	48.95	49.92	795	51.87	51.73
TiO_2	324	0.36	0.20	198	1.72	1.24	784	0.16	0.10
Al_2O_3	329	3.63	3.62	198	4.45	3.67	795	4.34	4.55
Cr_2O_3	296	0.68	0.69	135	0.46	0.42	790	1.20	1.23
Fe_2O_3	52	1.27	0.99	1	3.34	3.34	10	1.46	1.65
TFeO	254	3.60	2.81	184	6.13	6.78	225	2.55	2.49
FeO	75	3.28	3.11	14	5.99	5.85	570	2.80	2.69
CaO	329	22.49	22.56	198	22.22	22.33	795	22.16	22.32
MgO	329	16.37	16.52	198	14.38	15.34	795	17.07	17.18
MnO	307	0.10	0.10	192	0.12	0.13	789	0.09	0.09
NiO	171	0.05	0.03	12	0.03	0.01	601	0.05	0.05
K_2O	213	0.01	0.00	70	0.01	0.00	180	0.01	0.01
Na_2O	320	0.43	0.39	198	0.55	0.34	778	0.31	0.18

表 2　镁铁–超镁铁岩中单斜辉石主量元素参数设置

Table 2　Parameter settings of clinopyroxene in mafic–ultramafic rocks

参 数	决策树	随机森林	AdaBoost	GBDT
max_features	0.55	0.36	0.08	0.03
max_depth	21	6	39	27
min_samples_split	2	5	3	4
min_samples_leaf	4	3	2	2
n_estimators	-	90	210	710
learning_rate	-	-	0.123	0.008
subsample	-	-	-	0.52

元素,本文基于排列重要性方法[39]对 4 种方法中的特征重要性进行了排序,排序结果如图 7 所示。根据特征重要性排序的结果,在图 8 中用 2 种最主要成分分别作为横、纵坐标轴构建散点图,展示 4 种算法在岩浆环境判别问题上的性能。

由图 7 可知,决策树方法得出的最重要主量元素,为 TFeO, Al_2O_3, TiO_2, Cr_2O_3, FeO;随机森林方法认为,影响算法进行构造环境判别最重要的主量元素为 Cr_2O_3, TFeO, TiO_2, FeO, Al_2O_3;AdaBoost 方法认为,影响算法判别最重要的主量元素依次为 Cr_2O_3, TiO_2, TFeO, Al_2O_3, FeO;GBDT 方法认为,最重要的主量元素为 TiO_2, TFeO, Al_2O_3, FeO, Cr_2O_3。

由图 8 可见,基于最主要的 2 种主量元素可以区分出大部分测试数据的构造环境,但是仍有小部分的测试数据在二维坐标系中交叠混杂。这种现象一方面说明了 4 种模型在区分岩石构造环境问题上的有效性,另一方面也说明了只使用 2 种特征来区分不同的构造环境会存在部分数据难以区分的现象,使用全数据模式研究地质学构造势在必行。

3.4　性能评估

为了评估树类算法的性能,本文设计了 10 次独立重复实验。表 3 中的宏平均指数表征了决策树类算法的性能。从表 3 的 MaF 和 Accuracy 的结果对比可见,boosting 算法相较于 bagging 算法具有更好的分类性能,单决策树算法的效果劣于集成算法。

4　讨　论

上述实验结果从决策图、混淆矩阵、性能评估等方面展示了树类机器学习方法在单斜辉石构造背景判别问题上的良好表现,验证了树类机器学习方法有效处理地质环境构造等有监督问题的可行性。实验结果表明,相较于传统的判别图方法,依托于全数据模式的机器学习分类模型能更充分地利用地质特征,挖掘数据深层次隐含信息,得出可靠结论。但同时,机器学习为了更大限度利用提供的数据信息,需要综合考虑更多的输入特性,使其模型结构通常较复杂。相对于判别图在地质构造环境研究中的直观性,机器学习模型更不易解释,因此机器学习模型在地质领域的深层次应用,还需要进一步探讨。毫无疑问的是,随着全数据模式在地质领域的兴起,机器学习方法在地质应用上的研究将更加广泛且深入。

本文中决策图、混淆矩阵和性能评估部分的结果也展示了不同树类方法在单斜辉石数据上的分类效果。集成方法基于多棵决策树综合考虑,随机森林通过多棵子树投票确定最终分类结果,boosting 方法基于上一轮分类效果,改进模型以构建基分类

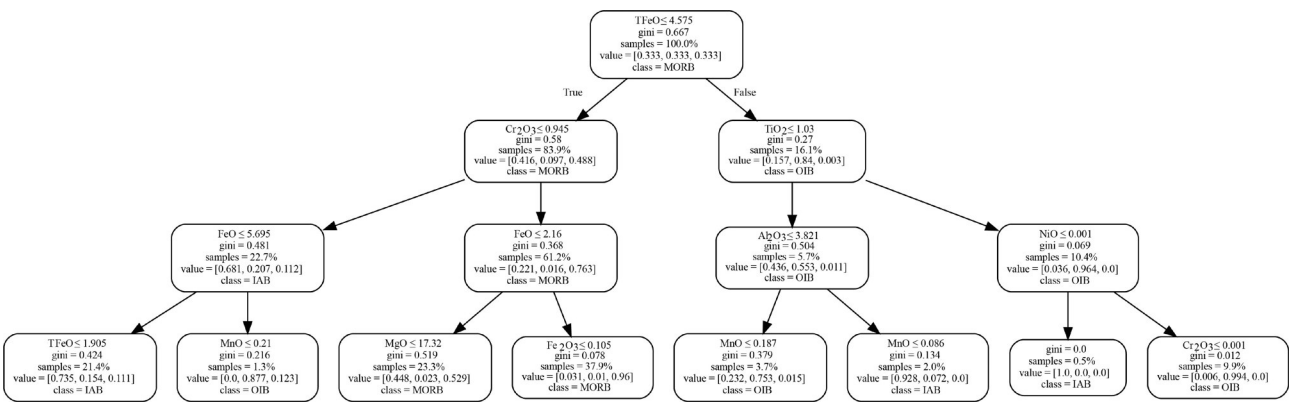

图 5　基于主量元素的决策树(部分)

Fig. 5　Part of decision tree graph

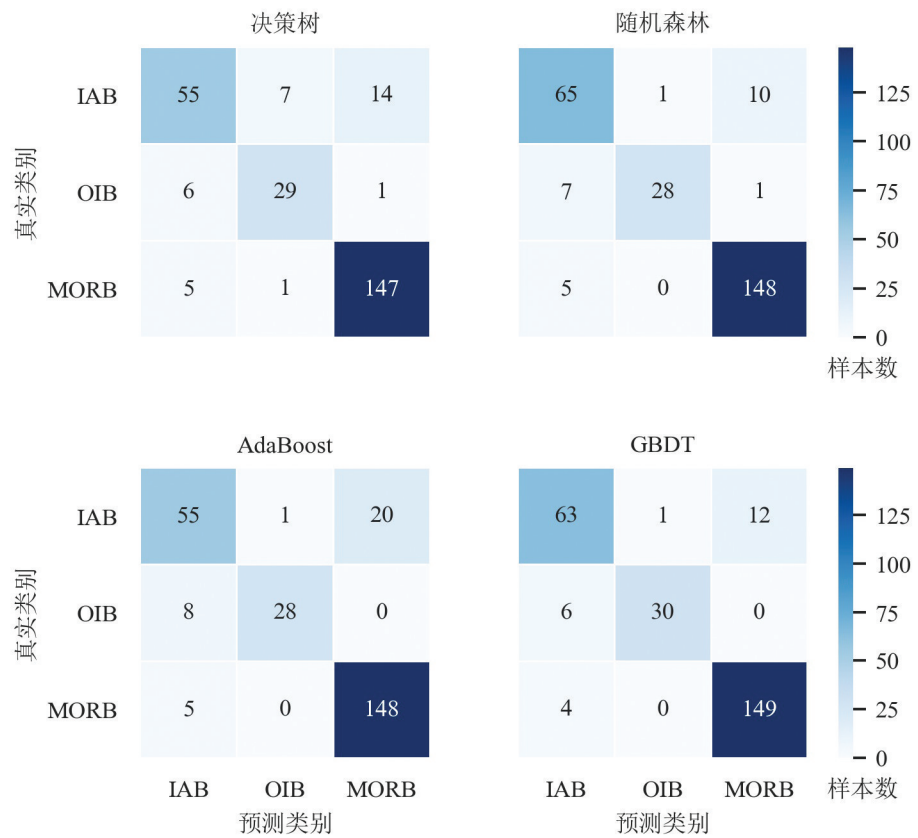

图6 基于主量元素的混淆矩阵(代号同表1)

Fig. 6 Confusion matrix on major element data

器,相对于单决策树模型,随机森林和boosting方法均能明显提高分类准确度。随机森林基于随机划分的数据集和随机自助采样特征对每棵子树进行训练,最终投票选择分类结果,其通过降低模型方差提高模型性能。boosting方法关注上一轮分类器模型的分类结果,以迭代优化基分类器,最终加权多个基分类器获得分类结果,boosting方法通过降低模型偏差提高模型性能。2种方式均能有效提升模型性能,然而如图7所示,本文使用的地质数据中不同的主量元素对判断样本所属类别的影响差异较大,因此相较于boosting分类方法,无权重投票的随机森林方法效果略差。毋庸置疑,多棵子树的集成方法的时间消耗较决策树方法更长,而集成方法中随机森林的多棵子树间并无联系,可以并行执行计算过程,因此随机森林的建模耗时比boosting方法短。

此外,图7展示了4种方法得到的特征重要性排序,因为使用同一数据集,整体上4种方法表现出类似的排序结果,均认为 TFeO,Al_2O_3,TiO_2,Cr_2O_3,FeO 是区分3种构造环境的最主要元素,基于每种方法判定的2种最重要特征的数据分布图展示了特征重要性排序的结果可靠性和4种分类模型的性

表3 基于主量元素的性能指标

Table 3 Performance indexes on major element data

测量指标	决策树	随机森林	AdaBoost	GBDT
MaP	0.8393(+/-0.0357)	0.9120(+/-0.0268)	0.9219(+/-0.0179)	0.9224(+/-0.0292)
MaR	0.8416(+/-0.0347)	0.8904(+/-0.0398)	0.9023(+/-0.0345)	0.9057(+/-0.0302)
MaF	0.8389(+/-0.0294)	0.8997(+/-0.0305)	0.9108(+/-0.0241)	0.9130(+/-0.0243)
Accurcy	0.8715(+/-0.0199)	0.9212(+/-0.0280)	0.9302(+/-0.0193)	0.9315(+/-0.0227)

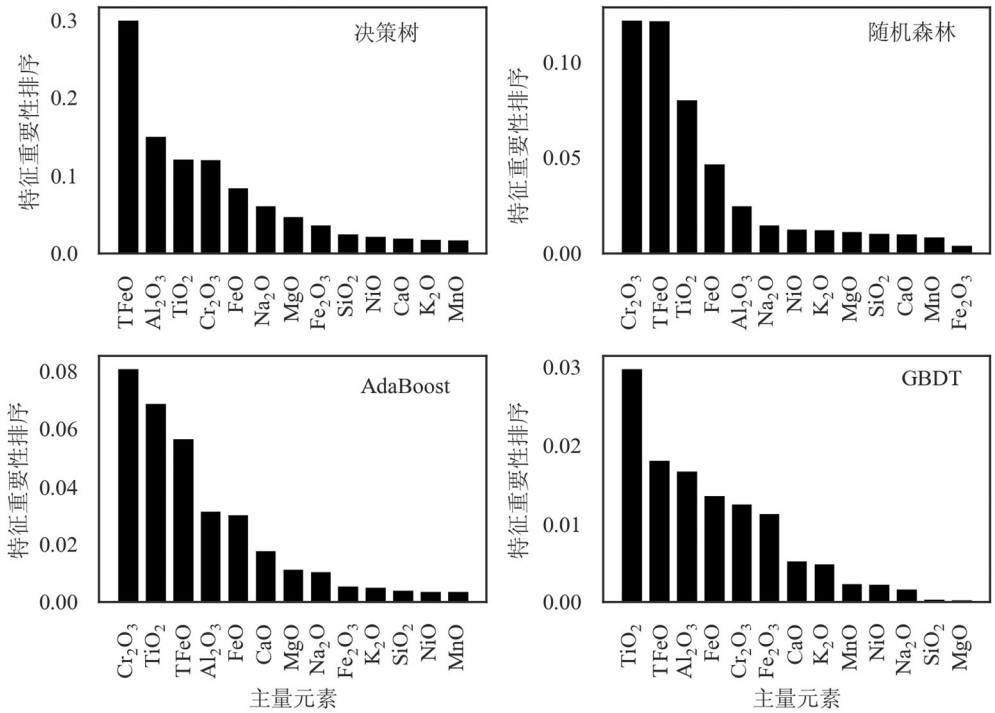

图7　特征重要性排序

Fig. 7　Feature importance on major element data

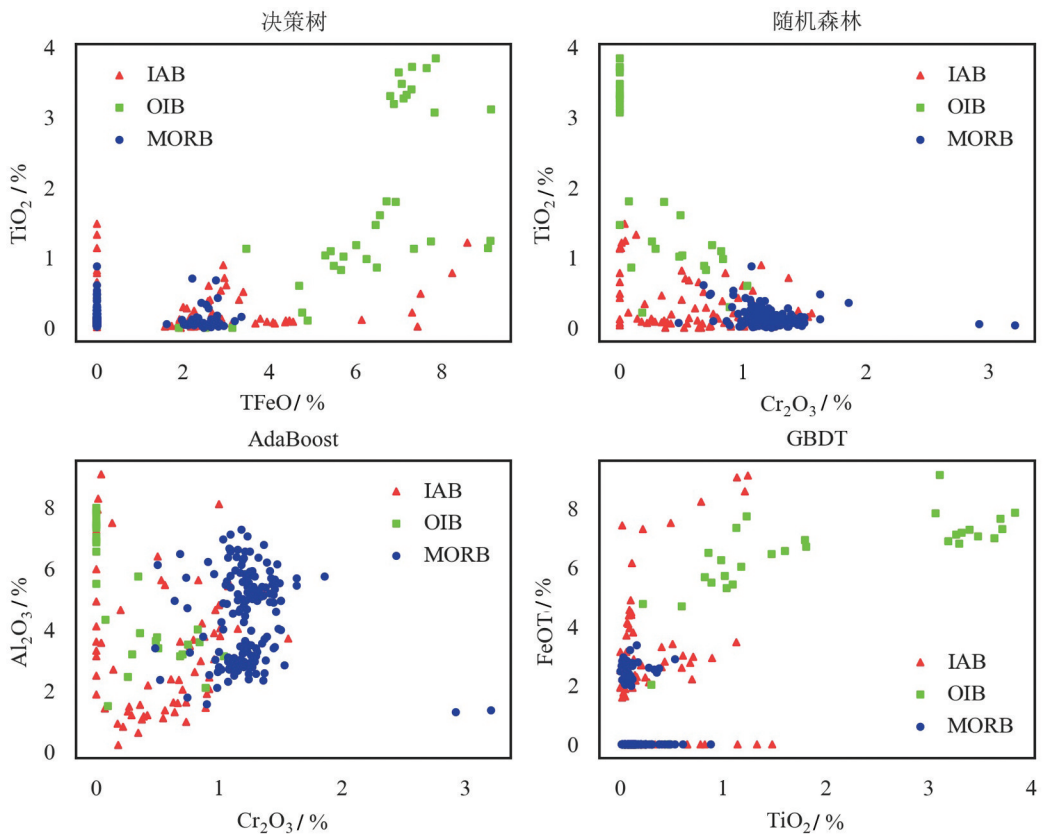

图8　基于主要成分散点图(代号同表1)

Fig. 8　Feature-based scatter diagram on major element data

能。由图8可见,得到的特征可以在一定程度上区分3种构造环境,但是因为只使用了2种主要的特征绘制散点图,不同类别的数据存在重叠部分,可见简单使用2种特征并不能完全划分出不同样本所处的构造环境,全数据模式应当在地质领域发挥更重要的作用。本文的特征重要性排序也展示了机器学习在地质方面的另一个应用,即特征重要性判别,虽然不同的机器学习方法得出的结果不同,但是结合专家经验,机器学习方法能够给地质领域提供更多的辅助功能。

5 结 论

本文基于单斜辉石的矿物元素数据,采用树类算法从可解释性、分类性能、特征重要性排序3个方面对镁铁-超镁铁岩的构造环境进行了分析。考虑到本文所用树类算法均基于决策规则,理论上可以完整绘制4种算法的决策树图,然而受限于集成树的规模和复杂性,本文只展示决策树算法的决策树图。就可解释性和构造过程的易理解性而言,决策树算法更胜一筹。就分类准确性而言,GBDT在当前数据集上表现更好。当对于模型复杂度和准确率均有所要求,随机森林方法对于岩浆环境判别问题更适宜。此外,4种方法对主量元素数据集的特征重要性排序在一定程度上表现出相似性,Cr_2O_3, TFeO, TiO_2, FeO和Al_2O_3可以被认为是识别构造背景的最重要主量成分。

本次研究取得了较好的结果,说明采用单斜辉石来鉴别镁铁-超镁铁岩形成的构造环境是可行的,这就提出了一个判别图研究的新方向。但是,如何进一步将研究结果应用于解决实际地球化学应用中,还需要进一步探讨。

参考文献

[1]Leterrier J, Maury R C, Thonon P, et al. Clinopyroxene composition as a method of identification of the magmatic affinities of paleo- volcanic series[J]. Earth and Planetary Science Letters, 1982, 59(1): 139-154.

[2]Asthana D. Relict clinopyroxenes from within-plate metadolerites of the Petroi Metabasalt, the New England Fold Belt, Australia[J]. Mineralogical Magazine, 1991, 55(381): 549-561.

[3]Nisbet E G, Pearce J A. Clinopyroxene composition in mafic lavas from different tectonic settings[J]. Contributions to Mineralogy and Petrology, 1977, 63(2): 149-160.

[4]Helmy H M, El Mahallawi M M. Gabbro akarem mafic-ultramafic complex, Eastern Desert, Egypt: A Late Precambrian analogue of Alaskan-type complexes[J]. Mineralogy and Petrology, 2003, 77(1): 85-108.

[5]Khedr M Z, Arai S. Chemical variations of mineral inclusions in Neoproterozoic high-Cr chromitites from Egypt: Evidence of fluids during chromitite genesis[J]. Lithos, 2016, 240: 309-326.

[6]Hanson R E, Roberts J M, Dickerson P W, et al. Cryogenian intraplate magmatism along the buried southern Laurentian margin: Evidence from volcanic clasts in Ordovician strata, Marathon uplift, west Texas[J]. Geology, 2016, 44(7): 539-542.

[7]Menand T, Annen C, Blanquat M de S. Rates of magma transfer in the crust: Insights into magma reservoir recharge and pluton growth[J]. Geology, 2015, 43(3): 199-202.

[8]Pearce J A, Cann J R. Tectonic setting of basic volcanic rocks determined using trace element analyses[J]. Earth and Planetary Science Letters, 1973, 19(2): 290-300.

[9]Glassley W. Geochemistry and tectonics of the Crescent volcanic rocks, Olympic Peninsula, Washington[J]. GSA Bulletin, 1974, 85 (5): 785-794.

[10]Pearce J A, Lippard S J, Roberts S. Characteristics and tectonic significance of supra- subduction zone ophiolites[J]. Geological Society, London, Special Publications, 1984, 16(1): 77-94.

[11]Pearce J A. Trace element characteristics of lavas from destructive plate boundaries[J]. Andesites, 1982, 8: 528-548.

[12]Shervais J W. Ti-V plots and the petrogenesis of modern and ophiolitic lavas[J]. Earth and Planetary Science Letters, 1982, 59(1): 101-118.

[13]Wood D A. The application of a Th-Hf-Ta diagram to problems of tectonomagmatic classification and to establishing the nature of crustal contamination of basaltic lavas of the British Tertiary volcanic Province[J]. Earth and Planetary Science Letters, 1980, 50 (1): 11-30.

[14]Mullen E D. $MnO/TiO_2/P_2O_5$: a minor element discriminant for basaltic rocks of oceanic environments and its implications for petrogenesis[J]. Earth and Planetary Science Letters, 1983, 62(1): 53-62.

[15]Zhang Q, Sun W, Zhao Y, et al. New discrimination diagrams for basalts based on big data research[J]. Big Earth Data, 2019, 3(1): 45-55.

[16]王金荣, 陈万峰, 张旗, 等. N-MORB和E-MORB数据挖掘——玄武岩判别图及洋中脊源区地幔性质的讨论[J]. 岩石学报, 2017, 33(3): 993-1005.

[17]王金荣, 潘振杰, 张旗, 等. 大陆板内玄武岩数据挖掘:成分多样性及在判别图中的表现[J]. 岩石学报, 2016, 32(7): 1919-1933.

[18]杨婧,王金荣,张旗,等. 全球岛弧玄武岩数据挖掘——在玄武岩判别图上的表现及初步解释[J]. 地质通报,2016,35(12):1937-1949.

[19]汪云亮, 张成江. 玄武岩类形成的大地构造环境的Th/Hf-Ta/Hf图解判别[J]. 岩石学报, 2001, 17(3): 413-421.

[20]李玉琼, 杜雪亮, 金维浚, 等. 大洋中脊、洋岛、岛弧玄武岩中橄

榄石的对比研究[J]. 地质科学, 2018, 53(4): 1228-1239.

[21]韩帅, 李明超, 任秋兵, 等. 基于大数据方法的玄武岩大地构造环境智能挖掘判别与分析[J]. 岩石学报, 2018, 34(11): 3207-3216.

[22]焦守涛, 周永章, 张旗, 等. 基于GEOROC数据库的全球辉长岩大数据的大地构造环境智能判别研究[J]. 岩石学报, 2018, 34(11): 3189-3194.

[23]Vermeesch P. Tectonic discrimination of basalts with classification trees[J]. Geochimica et Cosmochimica Acta, 2006, 70(7): 1839-1848.

[24]朱林奇, 张冲. 谱聚类-Adaboost集成数据挖掘算法在岩性识别中的应用[J]. 中国科技论文, 2016, 11(5): 545-550.

[25]韩启迪,张小桐,申维.基于梯度提升决策树(GBDT)算法的岩性识别技术[J].矿物岩石地球化学通报,2018,37(6):1173-1180.

[26]Lehnert K, Su Y, Langmuir C H, et al. A global geochemical database structure for rocks[EB/OL][2019-04-10]https://agupubs.onlinelibrary.wiley. com / doi / full / 10. 1029 / 1999GC000026 Geochemistry, Geophysics, Geosystems, 2000.

[27]Phan A V, Nguyen M L, Bui L T. Feature weighting and SVM parameters optimization based on genetic algorithms for classification problems[J]. Applied Intelligence, 2017, 46(2): 455-469.

[28]Luo X, Xu Y, Wang W, et al. Towards enhancing stacked extreme learning machine with sparse autoencoder by correntropy[J]. Journal of the Franklin Institute, 2018, 355(4): 1945-1966.

[29]Luo X, Sun J, Wang L, et al. Short-term wind speed forecasting via stacked extreme learning machine with generalized correntropy[J]. IEEE Transactions on Industrial Informatics, 2018, 14(11): 4963-4971.

[30]Gorissen D, Couckuyt I, Demeester P, et al. A surrogate modeling and adaptive sampling toolbox for computer based design[J]. Journal of Machine Learning Research, 2010, 11(Jul): 2051-2055.

[31]Yan R, Ma Z, Zhao Y, et al. A decision tree based data-driven diagnostic strategy for air handling units[J]. Energy and Buildings, 2016, 133: 37-45.

[32]卢东标. 基于决策树的数据挖掘算法研究与应用[D]. 武汉理工大学硕士学位论文, 2008.

[33]Mantovani R G, Horváth T, Cerri R, et al. Hyper-parameter tuning of a decision tree induction algorithm[C]//2016 5th Brazilian Conference on Intelligent Systems (BRACIS). IEEE, 2016: 37-42.

[34]Rodriguez-Galiano V, Sanchez-Castillo M, Chica-Olmo M, et al. Machine learning predictive models for mineral prospectivity: An evaluation of neural networks, random forest, regression trees and support vector machines[J]. Ore Geology Reviews, 2015, 71: 804-818.

[35]Hastie T, Rosset S, Zhu J, et al. Multi-class AdaBoost[J]. Statistics and Its Interface, 2009, 2(3): 349-360.

[36]杜雪亮, 李玉琼, 金维浚, 等. 镁铁质-超镁铁质岩浆岩中单斜辉石的智能分析研究[J]. 地质科学, 2018, 53(4): 1215-1227.

[37]Snoek J, Larochelle H, Adams R P. Practical bayesian optimization of machine learning algorithms[C]//Advances in neural information processing systems. 2012: 2951-2959.

[38]Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: Machine learning in Python[J]. Journal of Machine Learning research, 2011, 12(Oct): 2825-2830.

[39]Altmann A, Toloşi L, Sander O, et al. Permutation importance: a corrected feature importance measure[J]. Bioinformatics, 2010, 26(10): 1340-1347.