海洋地质与第四纪地质
MARINE GEOLOGY & QUATERNARY GEOLOGY

基于机器学习的大洋玄武岩构造环境判别研究

徐　，关馨儿，吕豪哲，赵　霄，热则耶·如孜，陈艳虹

Tectonic discrimination of oceanic basalt by machine learning

XU Kun, GUAN Xiner, LV Haozhe, ZHAO Xiao, Rezeye RUZI, and CHEN Yanhong

在线阅读 View online: https://doi.org/10.16562/j.cnki.0256-1492.2023041101

ISSN 0256-1492
CN 37-1117/P

海洋地质与第四纪地质
MARINE GEOLOGY & QUATERNARY GEOLOGY

第 44 卷第 4 期
Vol.44, No.4

徐堃, 关馨儿, 吕豪哲, 等. 基于机器学习的大洋玄武岩构造环境判别研究 [J]. 海洋地质与第四纪地质, 2024, 44(4): 190-199.
XU Kun, GUAN Xiner, LV Haozhe, et al. Tectonic discrimination of oceanic basalt by machine learning[J]. Marine Geology & Quaternary Geology, 2024, 44(4): 190-199.

基于机器学习的大洋玄武岩构造环境判别研究

徐堃[1], 关馨儿[1], 吕豪哲[1], 赵霄[1], 热则耶·如孜[1], 陈艳虹[1,2]

1. 中国地质大学(北京)海洋学院, 北京 100083
2. 中国地质大学(北京)海洋矿产与极地地质教育部重点实验室, 北京 100083

摘要: 玄武岩的地球化学成分与其产出构造环境密切相关, 是研究地球深部物质组成与动力学过程的重要岩石。为了判别玄武岩形成的构造环境, 前人根据玄武岩的地球化学特征建立了一系列构造判别图, 然而这些判别图仅限于二维或三维判别。随着全球玄武岩样品地球化学数据的爆发性增长, 这些构造判别图逐渐暴露出其局限性强、准确率较低的缺点。在地学与大数据结合发展的背景下, 利用机器学习方法有利于更全面和深入分析数据, 建立高准确率和高效率的构造环境判别模型。因此, 本文利用 GEOROC 和 PetDB 数据库, 经过一系列数据下载、处理等步骤, 建立了全球现代大洋玄武岩数据集。通过支持向量机(SVM)和随机森林(RF)机器学习算法, 训练出高准确率的高维判别模型。本文分析了不同机器学习算法和不同地球化学成分数据集对现代大洋玄武岩构造环境判别的影响, 并将各个判别模型应用于蛇绿岩数据当中, 探讨机器学习模型在判别古老大洋岩石圈(蛇绿岩)形成构造环境下的应用前景。这项工作为大洋玄武岩形成的构造环境判别提供了更高维度的判别手段, 是大数据时代下机器学习如何在地球科学领域应用的一次有益尝试。

关键词: 地球化学; 构造环境; 机器学习; 大洋玄武岩

中图分类号: P736.3　　**文献标识码:** A　　**DOI:** 10.16562/j.cnki.0256-1492.2023041101

Tectonic discrimination of oceanic basalt by machine learning

XU Kun[1], GUAN Xiner[1], LV Haozhe[1], ZHAO Xiao[1], Rezeye RUZI[1], CHEN Yanhong[1,2]

1. School of Ocean Sciences, China University of Geosciences, Beijing 100083, China
2. Key Laboratory of Marine Mineral Resources and Polar Geology, Ministry of Education, China University of Geosciences, Beijing 100083, China

Abstract: The geochemical composition of basalt is closely related to the tectonic setting of the formation, thus basalt is an important window for viewing the deep Earth and the composition and geodynamic processes. To discriminate the tectonic setting of basalt formation, although a series of tectonic discrimination diagrams have been established based on the geochemical characteristics of basalt, those discrimination diagrams are limited to two-dimensional or three-dimensional data. With the explosive growth of global geochemical data of basalt, these discrimination diagrams show gradually the shortcomings of being local and inaccurate. Therefore, using machine learning methods is beneficial to analyze data multi-dimensionally and comprehensively, and to establish accurate and efficient discriminant models. A global modern oceanic basalt dataset was established by using GEOROC and PetDB databases through a series of steps from data downloading, training, and analyzing. The dataset was trained by the support vector machine (SVM) and random forest (RF) machine learning algorithms and a high-accuracy and high-dimensional discrimination model was built. In addition, the accuracies of different machine-learning algorithms training were analyzed against different geochemical composition datasets of modern oceanic basalt, and the discrimination models were applied to ophiolitic basalt to explore the application of machine learning models for ancient oceanic basalt. This work provided a higher-dimensional approach to discriminate oceanic basalt, and a successful attempt of using machine learning in earth science in the era of the big data.

Key words: geochemistry; tectonic setting; machine learning; oceanic basalt

玄武岩广泛分布于各种构造环境中, 为地幔部分熔融的直接产物, 其成分记录了岩浆源区、岩浆形成温度、压力及氧逸度等地幔深部过程的重要信息, 是研究地球深部物质组成与动力学过程的重要

资助项目: 中国地质大学(北京)创新创业训练计划项目(S202211415138); 中央高校基本科研业务费专项资金(2652021007)
作者简介: 徐堃(2001—), 女, 本科, 海洋科学专业, E-mail: 1011201201@email.cugb.edu.cn
通讯作者: 陈艳虹(1990—), 女, 博士, 主要从事大洋岩石圈的岩石学、地球化学和全球构造研究, E-mail: chenyh@cugb.edu.cn
收稿日期: 2023-04-11; **改回日期:** 2023-06-01.　张现荣编辑

岩石[1-2]。根据玄武岩的产出位置，可分为大洋中脊玄武岩、岛弧玄武岩、弧后盆地玄武岩、大陆弧玄武岩、洋岛玄武岩、洋底高原玄武岩、大陆板内玄武岩等。不同构造位置产出的玄武岩由于地幔源区及其经历的岩浆作用（部分熔融、结晶分异、同化混染）的不同，导致其成分具有不同的特征，所以利用玄武岩的这种成分差异可推导其形成的构造环境。前人根据玄武岩地球化学特征和构造环境的这种关系建立了一系列构造环境判别图，如 Th/Yb-Ta/Yb 图解[3]、Nb/Yb-Th/Yb 图解[4]、Th-Hf/3-Ta 图解[5]、Ti/1 000-V 图解[6]、Y-Cr 图解[7]。但随着岩石地球化学分析测试技术的完善及推广，早先的部分构造判别图也暴露出使用元素种类较少、岩石数据较少、岩石数据区域性较强的缺点，这导致判别图受区域地质背景的影响较大而不一定适用于全球范围玄武岩判别[8-9]。正是这些限制，使得不同构造判别图解具有不同的应用范围和条件，不当的使用极易造成误判。同时，常规全岩地球化学分析可生成高维地球化学数据，但玄武岩的传统判别图仅能应用其中二至三维数据特征，造成大量有效信息被浪费[10]。

随着地质大数据的发展，可以对大量数据进行处理的机器学习成为地质大数据的研究热点[11]。与传统的研究方法相比，机器学习在数据分析中更加深入和全面[12]。机器学习是指利用计算机指定的算法，通过学习相关的经验数据，生成应用者所需要的模型，在面对新数据时，计算机可以根据模型对新数据做出判别[13]。在机器学习应用于地球化学领域的过程中，通过选择优胜算法，改进现有算法，计算机可以高效率地学习高维度地球化学数据从而获得极高的判别分类能力[2,14-17]。GEOROC 数据库（https://georoc.mpch-mainz.gwdg.de/georoc）和 PetDB 数据库（https://search.earthchem.org）是目前使用较为广泛的岩石地球化学数据库，这两个数据库整合了各类文献中来自不同构造背景的火成岩岩石地球化学信息，数据查询方式多样，可支持大量下载，为机器学习的开展提供了可能[18-19]。

在各种玄武岩中，大洋玄武岩的形成经历的地壳混染作用比较少，经历的岩浆过程相对简单，同时其形成构造环境的判别对于恢复蛇绿岩（古老大洋岩石圈）形成环境具有重要意义。因此本文利用 GEOROC 和 PetDB 数据库，结合玄武岩在现代大洋中的产出构造背景，建立现代大洋中脊玄武岩（MORB）、岛弧玄武岩（IAB）、弧后盆地玄武岩（BABB）、洋底高原玄武岩（OPB）和洋岛玄武岩（OIB）数据集。结合已有机器学习方法，训练出高准确率的高维度大洋玄武岩构造环境判别模型，并尝试将这一模型应用于蛇绿岩中的玄武岩中，探讨机器学习方法在大洋玄武岩构造环境判别方向的应用可能。

1　数据与方法

本文使用的样品数据全部来源于 GEOROC 和 PetDB 数据库，这两个数据库根据样品的岩性、成分、产出构造环境等，对数据库中样品数据进行了大致的分类。数据一般包括样品名称、岩石类型、采样位置、岩石成分（主量元素、微量元素、同位素）等信息。本文数据搜集、处理及建模流程如下（图 1）：首先下载所需玄武岩数据，对玄武岩数据进行初步的构造环境判别，筛选出属于现代大洋构造环境下的玄武岩数据并根据构造环境类别添加标签，然后对添加标签的数据进行质量筛查，剔除不符合要求的样品，确保机器学习的准确性，之后利用这些数据通过机器学习建立判别模型，如果模型准确，判别的准确率高，则将模型代入蛇绿岩数据中进行应用，以下详细描述各个步骤。

1.1　数据下载与初筛

本文挑选了相关构造背景下的火成岩数据并下载（下文将该数据称为原始火成岩数据）。其中，GEOROC 原始火成岩数据共 368 339 条，PetDB 原始火成岩数据共 85 639 条。在原始火成岩数据中，根据岩性描述并结合 SiO_2 含量为 45%～52% 进行筛选，得到原始玄武岩数据，其中 GEOROC 共 80 660 条，PetDB 共 45 589 条（图 2a）。

1.2　数据预处理——构造环境判别并加标签

根据构造环境，大洋玄武岩可分为大洋中脊玄武岩、洋岛玄武岩、岛弧玄武岩、弧后盆地玄武岩和洋底高原玄武岩 5 个类别，每一个类别相当于一个标签。计算机只有对同一标签下的玄武岩地球化学数据进行学习后才能得出该类玄武岩的地球化学成分特征与构造环境之间的关系，并通过训练得出判别模型。本文利用 ArcGIS 软件将样品采集位置投影到全球地形图上，通过样品位置与大地构造位置的关系确定其产出构造环境，为数据添加标签。为确保生成数据集的准确性，本文仅考虑具有明确产出构造背景的现代大洋玄武岩样品。筛选后得到 GEOROC 数据共 39 045 条，PetDB 数据共

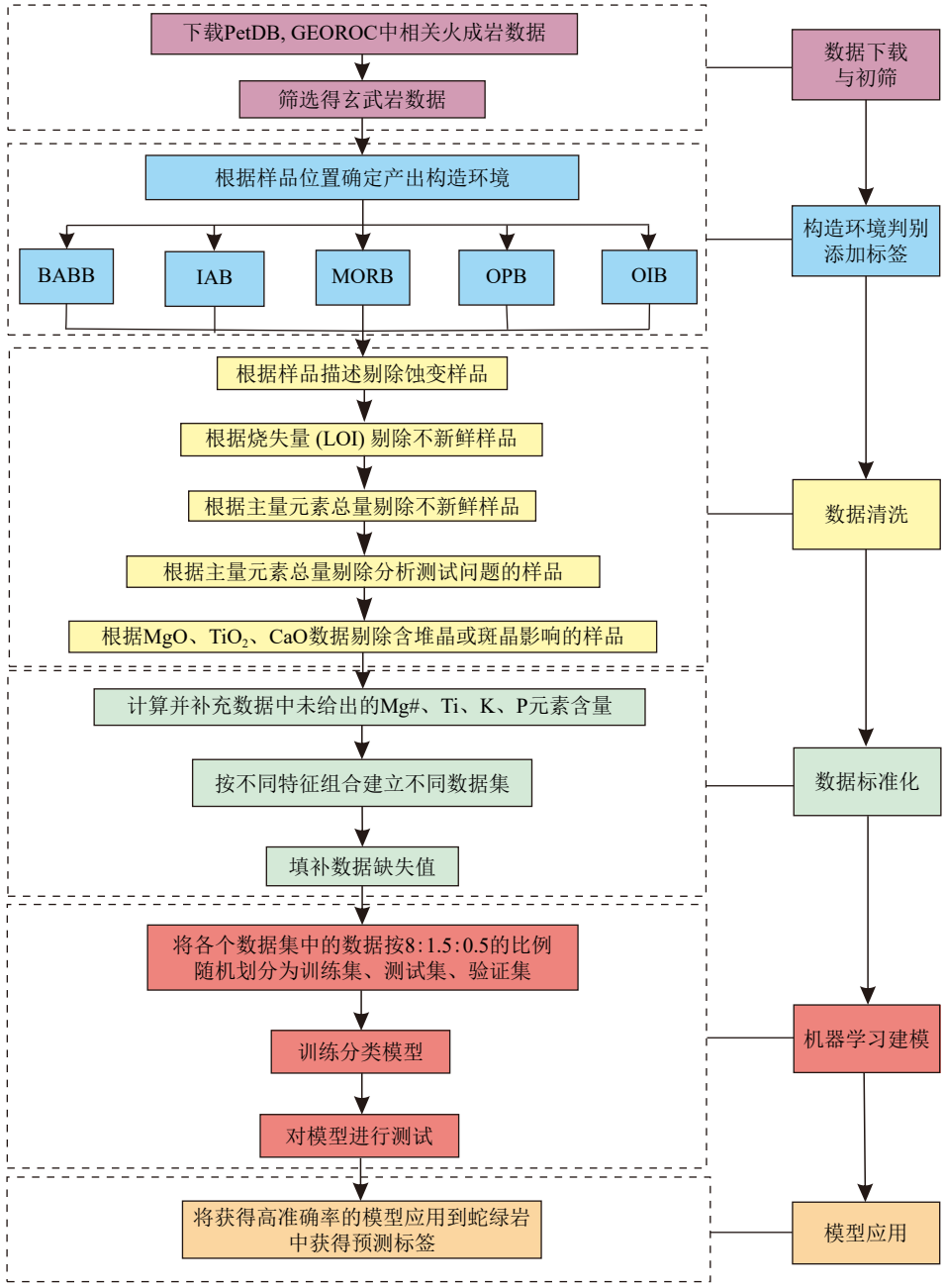

图 1 基于机器学习的大洋玄武岩构造环境判别总思路流程图

Fig.1 Tectonic discrimination of oceanic basalt using machine learning

30 594 条(图 3a)。

1.3 数据清洗

在这些数据中,岩石地球化学成分是最关键的信息,每种成分指标可以看作机器学习时输入的一个特征,计算机通过对特征的学习建立玄武岩地球化学成分和构造环境的联系,从而生成判别模型。所以,数据的质量非常重要,直接关系到机器学习的结果。大洋玄武岩形成后,极易遭受海底热液改造及海底风化作用影响,导致其成分发生变化,降低机器学习模型的准确率,所以在建模过程中应剔除蚀变的样品。除此之外,由于玄武岩是喷出岩,有时会含有大量斑晶,过量的斑晶也会影响岩石成分,因此含有过量斑晶的样品也不适用于建模。以上这些不符合建模要求的样品可以通过数据样品的描述(样品蚀变程度)以及样品成分(烧失量及主量元素总量)有效筛查出来。但是,由于不同文献来源数据对于氧化铁(FeO, Fe_2O_3, FeO^T, $Fe_2O_3^T$)的表达有所差别,所以在数据质量筛查前需统一所有样品数据的主量元素列表。本文统一使用 FeO^T,其转化公式如下:$FeO^T = FeO + 0.8998 * Fe_2O_3$;$FeO^T = 0.8998 * Fe_2O_3^T$,主量元素总量 $SUM = SiO_2 + TiO_2 +$

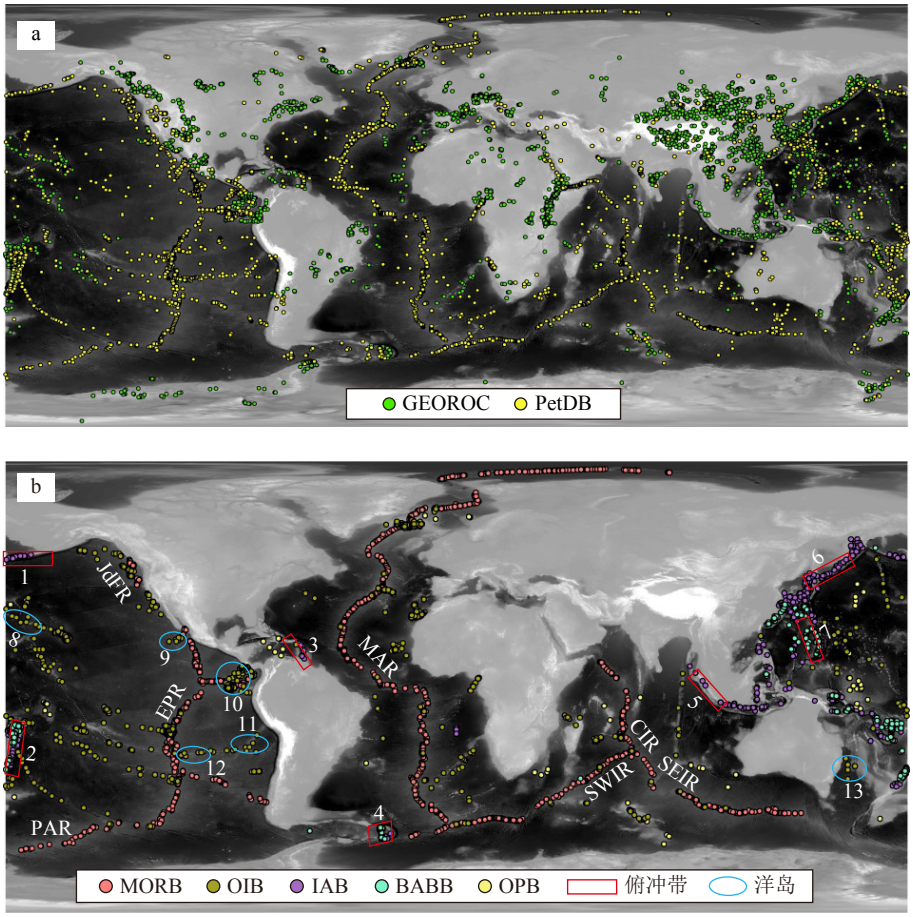

图 2　下载的全球玄武岩数据分布

a: 两个不同数据库的原始玄武岩数据分布图, b: 数据筛选后以构造环境为分类标准的玄武岩分布图。MAR: 大西洋中脊, EPR: 东太平洋海隆, CIR: 印度洋中脊, SWIR: 西南印度洋脊, SEIR: 东南印度洋脊, PAR: 太平洋-南极洋脊, JdFR: 胡安德富卡洋脊, 1: 阿留申俯冲带, 2: 汤加-克马德克俯冲带, 3: 安的列斯俯冲带, 4: 南桑德维奇俯冲带, 5: 爪哇俯冲带, 6: 日本-千岛俯冲带, 7: 伊豆-小笠原-马里亚纳俯冲带, 8: 夏威夷群岛, 9: 索科隆群岛, 10: 加拉帕戈斯群岛, 11: 圣菲利克斯群岛, 12: 复活节岛, 13: 塔斯曼群岛。底图高程数据来自 ETOPO1。

Fig.2　Global distribution of basalt in the world used in this study

a: Global distribution of original basalt samples downloaded from two databases of GEOROC and PetDB, b: global distribution of selecting basalt samples of different tectonic setting. MAR: Mid-Atlantic Ridge, EPR: East Pacific Rise, CIR: Central Indian Ridge, SWIR: Southwest Indian Ridge, SWER: Southeast Indian Ridge, PAR: Pacific Antarctic Ridge, JdFR: Juan de Fuca Ridge, 1: Aleutian Trench, 2: Tonga Trench, 3: Antilles Trench, 4: South Sandwich Subduction Zone, 5: Sunda Trench, 6: Japan-Kuril Subduction Zone, 7: Izu-Bonin-Mariana Subduction Zone, 8: Hawaii Islands, 9: Socorro Islands, 10: Galapagos Islands, 11: San Felix Islands, 12: Easter Island, 13: Tasman Islands. The topography data are from the ETOPO1.

$Al_2O_3+FeO^T+CaO+MgO+MnO+K_2O+Na_2O+P_2O_5$。

数据质量筛查包括以下方面: ①通过样品描述, 将发生蚀变的样品剔除; ②通过数据中的烧失量(LOI)信息判断岩石新鲜程度, 将 LOI＞2% 的样品判断为经历了蚀变作用的不新鲜样品并剔除; ③通过样品主量元素总量, 判断样品新鲜程度, 如对于 MORB、OIB、OPB 这些“干”岩浆系统, 剔除主量元素总量 ＜98.5% 的样品, 而对于 IAB、BABB 这些“湿”岩浆系统, 则剔除主量元素总量＜97% 的样品; ④将主量元素总量＞101% 的样品判断为分析测试有问题并剔除; ⑤将 MgO＞10.5%, TiO_2＞6%, CaO＞15% 的玄武岩数据剔除, 排除堆晶或斑晶的影响。最终筛选得到的 GEOROC 数据共 9346 条, PetDB 数据共 24916 条(图 3a)。图 3b、c 展示了数据处理前后各类别大洋玄武岩的数据量, 图 1b 展示了数据处理后用于下一步机器学习的大洋玄武岩分布图。

1.4　数据预处理——数据标准化

由于样品的 Mg#及 Ti、K、P 元素含量是指示岩石成因的重要地球化学指标, 所以在建立最终机器学习数据集前还需要进行相关计算并补充。最后

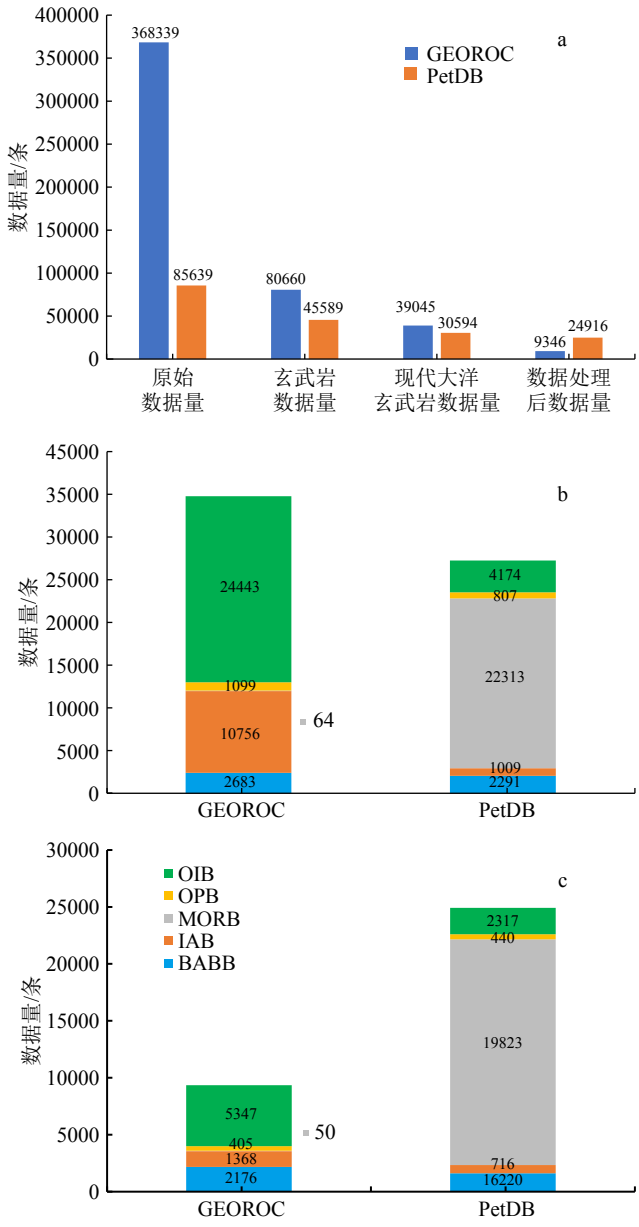

图 3 数据处理前后数据统计图

a: 数据处理过程大洋玄武岩数据统计图, b: 数据清洗前不同构造类别大洋玄武岩统计图, c: 数据处理后不同构造类别大洋玄武岩统计图。OIB: 洋岛玄武岩, OPB: 洋底高原玄武岩, MORB: 大洋中脊玄武岩, IAB: 岛弧玄武岩, BABB: 弧后盆地玄武岩。

Fig.3 Histograms of the number of samples before and after data processing

a: Histograms of the number of oceanic basalt sample, b: histograms of the number of oceanic basalt samples from different tectonic settings before data cleaning, c: histograms of the number of oceanic basalt sample from different tectonic settings after data cleaning. OIB: ocean island basalt, OPB: ocean plateau basalt, MORB: mid-ocean ridge basalt, IAB: island arc basalt, BABB: back-arc basin basalt.

用于机器学习的玄武岩数据包括以下特征: 主量元素（SiO_2, TiO_2, Al_2O_3, FeO^T, CaO, MgO, MnO, K_2O, Na_2O, P_2O_5）、Mg#、微量元素（Ti, K, P, Li, Be, B, Sc, V, Cr, Co, Ni, Cu, Zn, Ga, Rb, Sr, Y, Zr, Nb, Cs, Ba, La, Ce, Pr, Nd, Sm, Eu, Gd, Tb, Dy, Ho, Er, Tm, Yb, Lu, Hf, Ta, Pb, Th, U）和同位素（$^{87}Sr/^{86}Sr$, $^{143}Nd/^{144}Nd$, $^{206}Pb/^{204}Pb$, $^{207}Pb/^{204}Pb$, $^{208}Pb/^{204}Pb$, $^{176}Hf/^{177}Hf$）。

我们将两个数据库中 5 个不同标签的数据分别进行合并生成总数据集（Basic_Data）。为了对比不同成分数据模型的构造判别能力, 本文还分别提取主量元素、微量元素、主微量元素和同位素整理生成主量元素数据集（Basic_Data_ME）、微量元素数据集（Basic_Data_TE）、主微量元素数据集（Basic_Data_M&TE）、同位素数据集（Basic_Data_RI）。随后将每个数据集中的数据按照 8∶1.5∶0.5 的比例随机分配形成训练集、测试集、验证集。

在实现机器学习的过程中, 程序是不允许数据的特征中有缺失值存在的, 但每个数据集中的数据都或多或少存在特征包含缺失值的情况。对于缺失值, 本文的处理方法是: ①找出数据缺失值所对应的特征, 将这些特征按缺失值的数量从小到大排列, 编写程序建立循环; ②每个循环内把这次循环要填补的特征当作一个标签, 剩下的所有特征（包含已填补的特征和尚未进行填补的特征）与标签构成一个特征矩阵, 用随机森林回归填补进行数据填充, 关于随机森林回归填补的原理前人[20]已做过详细解释; ③特征矩阵里的缺失值暂时用中位数填充, 在后面的循环中会将随机森林回归得到的数值填进用中位数填补的数据当中, 之后重复以上步骤, 直到所有数据都填补完成。数据填补之后, 需进行特征工程处理。特征工程处理后的数据可以决定机器学习模型分类的上限, 使机器学习结果更加准确[21]。Python 下 Scikit-Learn 库提供了特征工程处理工具, 可对数据进行归一化和标准化处理。

1.5 机器学习建模及应用

高效的算法是建立高准确率判别模型的关键, 本文选择了具有简单、容易实现、计算成本小[13]的随机森林算法（RF）和作为机器学习主流技术的支持向量机算法（SVM）构建模型。随机森林（RF）是在结合多个决策树算法的基础上, 数据集被随机有放回过程的选出新的子数据集, 并在子数据集中选出部分特征作为判别子节点从而实现数据分类的一种集成学习算法[22]; 而支持向量机（SVM）是一种将数据特征非线性映射到高纬度特征空间, 并通过在特征空间中形成构造线性决策曲面从而实现数据分类的算法[23]。

确定算法后, 将训练集的数据代入算法中训练

并生成模型。为了使模型具有更好的分类效果, 本文利用验证集数据调整模型的超参数, 最后利用测试集检验模型在实际使用过程中的泛化能力, 使用训练集准确率作为评估模型最终分类效果的指标。准确率即被正确分类的数据量与数据总量之比。

建立现代大洋玄武岩的机器学习模型后, 本文尝试将这些模型应用于古老大洋玄武岩(产于蛇绿岩中玄武岩)构造环境判别来评判模型的应用前景。用于机器学习预测的蛇绿岩数据来自 PetDB 数据库中的 Ophiolite 数据集, 原始蛇绿岩数据共 4 129 条, 筛选出玄武岩类型的数据共 140 条, 数据处理的过程按照本文 1.2—1.4 节的方法进行操作, 经过数据处理后符合要求的蛇绿岩数据为 34 条。我们将蛇绿岩数据代入不同数据集和不同算法形成的模型中, 并将预测结果与原文献结果进行对比, 计算出蛇绿岩按照现代大洋玄武岩构造环境细分类(MORB、IAB、BABB、OIB、OPB)的预测准确率和按照构造环境大类分类(大洋中脊的玄武岩 MORB、俯冲相关的玄武岩 IAB 和 BABB、板内玄武岩 OIB 和 OPB)的预测准确率。

2　结果

本文采用准确率(表 1)评判模型对大洋玄武岩构造环境进行判别。准确率(Accuracy)代表的是被正确预测的样本数与样本总数的比值。从表 1 可以看出, 不同的数据集有不同的准确率, 但 SVM 和 RF 算法的准确率都可以达到 0.9 以上。表 2 展示了用机器学习判别模型对蛇绿岩按照现代大洋玄武岩细分类和构造环境大类分类标准判别的准确率。可以看出, 对于细分类模型, 不管哪一种算法, 其准确率都很低; 而对于构造大类分类模型, 其准确率有所提高, 但无法达到现代大洋玄武岩构造判别的水平。

表 1　SVM、RF 算法下现代大洋玄武岩分类模型准确率

Table 1　Accuracy of modern oceanic basalt classification models using SVM and RF algorithms

	Basic_Data	M&TE	ME	TE	RI	Average
SVM	0.94	0.949	0.979	0.952	0.975	0.959
RF	0.997	0.994	0.914	0.993	0.982	0.976

表 2　蛇绿岩中玄武岩构造环境预测准确率

Table 2　Prediction accuracy of ophiolite using SVM and RF algorithms

		Basic_Data	M&TE	ME	TE	RI	Average
细分类模型	SVM	0.206	0.235	0.059	0.294	0.206	0.200
	RF	0.118	0.265	0.088	0.294	0.265	0.206
大类分类模型	SVM	0.765	0.824	0.765	0.794	0.765	0.783
	RF	0.706	0.794	0.706	0.853	0.765	0.765

3　讨论

3.1　构造判别图与机器学习对现代大洋玄武岩构造环境判别结果对比

为了探讨构造判别图解与机器学习对于现代大洋玄武岩的形成构造环境判别能力, 本文选择了两个较为常用的构造判别图解: Shervais[6] 提出的 Ti/1 000-V 判别图解和 Wood[5] 提出的 Th-Hf/3-Ta 判别图解, 将本文生成的现代大洋玄武岩数据集中的部分数据投图到这两幅判别图解上, 对比这两个构造判别图解与机器学习生成的判别模型的准确率。

Ti/1 000-V 图解的原理在于 V 元素在硅酸盐体系中可以以 V^{3+}、V^{4+}或 V^{5+}存在, 而 Ti 仅以 Ti^{4+}的形式存在, V 元素的矿物-熔体分配系数会随氧逸度变化, 因此岩石的 V 含量可以大致反映岩浆形成及演化过程的氧逸度。更重要的是, Ti 和 V 在热液蚀变过程中是不活动的, 并且在中高级变质过程中也是稳定的。而 Th-Hf/3-Ta 判别图解的原理在于 Th、Hf 和 Ta 这些不相容元素在不同构造环境中其分配系数不同。这些构造判别图解所使用的元素为流体不活动元素, 对风化和蚀变甚至一些变质过程都相对不敏感。所以这两个图解被认为可以有效区分由于源区成分及熔融环境不同而产生的不同成分的玄武岩, 如 MORB、IAB 和 OIB。本文选取了数据处理后的现代大洋玄武岩数据集中 IAB、MORB、OIB 中 Ti-V 和 Th-Hf-Ta 未出现空值的数据进行投图(图 4), 并计算了各图解不同构造环境的准确率。可以看出, Ti/1 000-V 图解中 IAB 和 OIB 都有一部分落在 MORB 的区域内, 因此准确率相对较低; Th-Hf/3-Ta 图解中大部分 IAB 落在对应区域内, 部分 OIB 落在 MORB 区域内, 两个图解相比 Th-Hf/3-Ta 图解准确率相对较高。

判别图解的准确率代表落入区域的数据个数除以该类别数据总数, 机器学习的准确率等于正确分类的数据总量除以全部数据总量。为使机器学习建立的模型可以更有效地与判别图解对比, 本文采用 RF 算法下各数据集的精确率(Precision)、召回

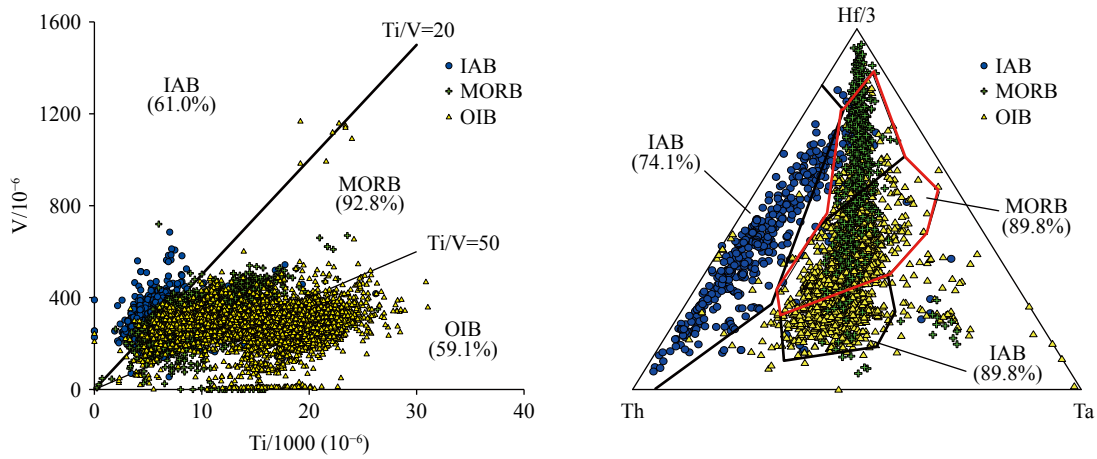

图 4 现代大洋玄武岩数据集 IAB、MORB、OIB 数据在 Ti/1000-V 图解和 Th-Hf/3-Ta 图解上的投影

括号中列出图解对现代大洋玄武岩的判别准确率。IAB: 岛弧玄武岩, MORB: 大洋中脊玄武岩, OIB: 洋岛玄武岩。

Fig.4 Projection of IAB, MORB, and OIB data on the Ti/1000-V and Th-Hf/3-Ta diagrams

The accuracy is shown in bracket. IAB: island arc basalt, MORB: mid-ocean ridge basalt, OIB: ocean island basalt.

率(Recall)和二者调和平均数 F1 分数(F1 Score)评价模型分类效果。精确率等于被模型正确判断为某一类别的样本数除以被模型判断为该类的总样本数,体现了模型对阴性样本的区分能力,即精确率越高,模型区分能力越强;召回率等于被模型正确判断为某一类别的样本数除以该类别的总样本数,体现了分类模型对阳性样本的判别能力,即召回率越高,模型判别能力越强;F1 分数是精确率和召回率的调和平均数,F1 得分越高,模型对数据分类状况越稳健。从表 3 总体来看,除 ME 数据集外,其他数据集对各类别玄武岩的判别能力比图解更强,且区分程度高,分类状况稳健。与传统的二维或三维图相比,机器学习利用了更高维度的数据对构造环境进行判别,且使用的数据来自于全球各地,与构造判别图本身相比,机器学习判别适用范围更广。因此机器学习更优于判别图。

3.2 蛇绿岩预测结果分析

机器学习在现代大洋玄武岩的构造环境判别中具有较高的准确率,下文将探讨机器学习判别模型在蛇绿岩中应用结果。在机器学习对蛇绿岩的准确率判别结果中可以看出,用机器学习判别模型对蛇绿岩按照现代大洋玄武岩细分类标准判别的准确率与文献结果有极大的出入(表 2)。最可能出现这种情况的原因是由于部分文献将蛇绿岩的原构造环境划分为弧前盆地玄武岩(FAB),但在本次机器学习中并未让计算机训练此种类别,这是由于经过数据预处理环节后符合建模要求的 FAB 数据仅剩 153 条,与其他 5 种类型玄武岩相比数据较少,

表 3 RF 算法下各数据集的分类精确率、召回率与 F1 分数

Table 3 Classification accuracy, recall, and F1 score of each dataset using RF algorithm

	BABB	IAB	MORB	OPB	OIB	数据集
精确率	1	1	0.99	1	0.99	
召回率	0.99	0.99	1	0.96	0.99	Basic_Data
F1分数	0.99	0.99	1	0.98	0.99	
精确率	1	0.99	0.99	0.98	0.99	
召回率	0.98	0.99	1	0.92	0.99	M&TE
F1分数	0.99	0.99	0.99	0.95	0.99	
精确率	0.82	0.84	0.9	0.92	0.91	
召回率	0.59	0.78	0.97	0.5	0.91	ME
F1分数	0.69	0.81	0.93	0.65	0.91	
精确率	0.99	0.99	0.99	0.98	0.99	
召回率	0.97	0.99	1	0.93	0.99	TE
F1分数	0.98	0.99	0.99	0.96	0.99	
精确率	0.96	0.97	0.99	0.99	0.98	
召回率	0.97	0.96	0.99	0.94	0.97	RI
F1分数	0.97	0.96	0.99	0.96	0.96	

这将导致模型对 FAB 的判别能力、对 FAB 和其他 5 种类型玄武岩的区分程度、判别过程中模型的稳健度这 3 个指标下降,进而使模型整体准确率下降。为了解决这个问题,我们将建立的现代大洋玄武岩数据集按构造环境大类进行合并重新归类,划分为以下 3 种类型:大洋中脊玄武岩(MORB)、俯

冲相关玄武岩(IAB 和 BABB)以及板内玄武岩(OIB 和 OPB)。文献中的 FAB 很明显也是属于俯冲相关玄武岩,其成分与俯冲相关的其他玄武岩相似,富集大离子亲石元素,亏损高场强元素[24-25]。合并后机器学习判别模型的准确率明显上升(表 2)。

机器学习判别模型在蛇绿岩中应用的准确率低的另外一个原因可能是,本次预测准确率计算是假定数据来源的研究所恢复的蛇绿岩形成环境为真实值。然而,这些研究部分是通过传统构造判别图对蛇绿岩中玄武岩的构造环境进行判别而得到的[26-28]。原文献中利用的判别图放到其他文献的数据中并不能完全适用(图 5),因此机器学习和文献本身出现的判别差异也有可能是使用了局限性较强的传统构造判别图导致。同时,蛇绿岩作为残留的古老大洋岩石圈,在构造环境的恢复上依然存在很大困难,其原因可能是洋壳从形成到闭合的过程中经历了不同构造体制下的软流圈熔融、壳幔分异、地表风化、热液蚀变甚至变质等多重作用,从而导致其在化学成分上的改变[29],使其难以与现代大洋玄武岩成分直接对比。

3.3　存在问题及可能改进方法

本文在建模过程中发现,特征在机器学习中占据重要作用,但特征数量的增加往往不一定能给模型带来性能上的提升。高维的数据可能会导致维度灾难,即随着维度的增高,模型效果会随着维度的增高而降低的现象。高维度数据集会使构建的模型复杂度增加而导致准确率下降。增加模型的复杂度的确可以提高拟合度,但容易导致数据的过拟合,致使该模型的预测力大幅降低。特征数量过多会使得部分特征成为冗余特征。为了避免过多的特征带来的问题,需要对特征进行筛选,选出重要特征,消除无用、冗余特征,这被称为特征选择[30]。

特征选择首先需要进行特征重要性评估,特征重要性评估的一个重要指标是基尼不纯度,即根据节点中数据的分布对其进行分类时,从节点中随机选择的数据被分错的概率。随着分类节点的增多,平均加权基底不纯度将会减少,也就是分类错误率减少。随机森林中的特征重要性表示在该特征上拆分的所有节点的基尼不纯度减少的总和。重要性度量数值越高代表基尼不纯度减少总和越高,这个特征对应分类节点的错误率也就越低。特征重要性评估能够选择稳健的特征,减少待提取特征数量,同时还能够提高分类算法泛化能力[31]。

以 RF 算法下准确率 Basic_Data 数据集为例进行特征重要性度量,表 4 列出了 Basic_Data 数据集所有特征的重要性度量指数。从该表可以看出排名第 25 的特征往后,重要性度量指数已经小于 0.01,应该舍弃,否则可能会对训练造成噪声干扰,造成准确率下降。

4　结论

本文利用 GEOROC 和 PetDB 数据库的数据,经过一系列的数据预处理,建立了现代大洋玄武岩的数据集,并通过机器学习方法建立了大洋玄武岩构造背景判别模型。建模结果显示,不同数据集下模型的准确率会有不同,但综合来看,机器学习方法

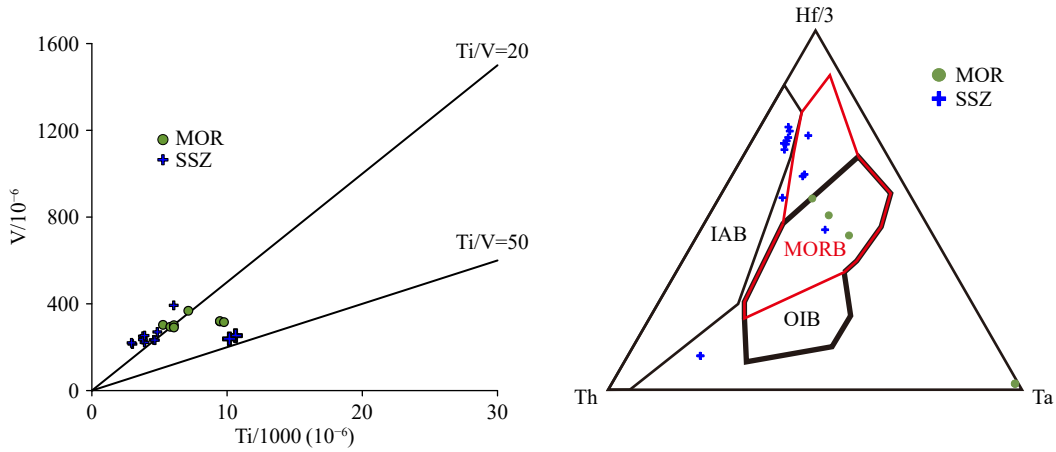

图 5　本文使用的蛇绿岩数据在 Ti-V 及 Th-Hf/3-Ta 判别图上的投影图

MOR: 大洋中脊, SSZ: 超俯冲带。

Fig.5　The Ti-V and Th-Hf/3-Ta diagrams for ophiolite data used in this study

MOR: mid-ocean ridge, SSZ: supra subduction zone.

表 4 特征重要性度量

Table 4 The feature importance metric

排名	特征	特征重要性度量	排名	特征	特征重要性度量	排名	特征	特征重要性度量
1	Pb^{208}/Pb^{204}	0.0997	19	Gd	0.0182	37	Ba	0.0013
2	Pb^{206}/Pb^{204}	0.0848	20	Pr	0.0157	38	Ni	0.0012
3	Sr^{87}/Sr^{86}	0.0845	21	La	0.0143	39	Sm	0.0011
4	Nd^{143}/Nd^{144}	0.0826	22	Nd	0.0124	40	Zr	0.0011
5	Ta	0.0689	23	Nb	0.0112	41	FeO^{T}	0.0011
6	Li	0.0529	24	Dy	0.0108	42	SiO_2	0.001
7	Ga	0.0492	25	Eu	0.0104	43	P_2O_5	0.0009
8	Cs	0.0458	26	Sc	0.0096	44	Al_2O_3	0.0009
9	Lu	0.0443	27	Ce	0.0072	45	K_2O	0.0009
10	Pb	0.0432	28	Co	0.0063	46	K	0.0008
11	Tm	0.0352	29	Cu	0.0057	47	MgO	0.0008
12	Yb	0.03	30	Sr	0.0048	48	CaO	0.0006
13	U	0.0264	31	Zn	0.0046	49	Rb	0.0006
14	Tb	0.0209	32	Th	0.002	50	Y	0.0006
15	Pb^{207}/Pb^{204}	0.0204	33	Ti	0.0017	51	$Mg^{\#}$	0.0005
16	Er	0.0193	34	Cr	0.0017	52	Na_2O	0.0005
17	Hf	0.0189	35	TiO_2	0.0015	53	P	0.0004
18	Ho	0.0187	36	V	0.0014	54	MnO	0.0004

下模型对于现代大洋玄武岩的判别能力、区分能力、判别的准确度都是优胜于传统的判别图解。其原因可能是机器学习相比较于传统的构造判别图解利用了更高维度的数据且数据具有全球性。

为了探讨机器学习方法建立的大洋玄武岩构造背景判别模型在蛇绿岩中的应用前景，本文还利用模型对 PetDB 搜集的来自全球蛇绿岩中的玄武岩形成构造环境进行预测。预测结果与文献研究恢复的蛇绿岩形成构造环境有一定的差异，这可能与蛇绿岩中玄武岩经历的地质作用复杂，导致其成分发生变化等因素有关。该工作还有待进一步补充更多的蛇绿岩数据并利用特征重要性评估后挑选重要特征进行建模。

致谢： 在本文数据处理及机器学习建模过程中奇安信安全技术有限公司陈海健工程师给予了建设性意见，两位审稿人为本文的改进提供了很大的帮助，在此一并表示感谢。

参考文献 (References)

[1] White W M. Probing the Earth's deep interior through geochemistry [J]. Geochemical Perspectives, 2015, 4(2)：95-96.

[2] Doucet L S, Tetley M G, Li Z X, et al. Geochemical fingerprinting of continental and oceanic basalts: A machine learning approach[J]. Earth-Science Reviews, 2022, 233：104192.

[3] Pearce J A. Role of the sub-continental lithosphere in magma genesis at active continental margins[M]//Hawkesworth C J, Norry M J. Continental Basalts and Mantle Xenoliths. Nantwich, Cheshire: Shiva Publications, 1983: 230-249.

[4] Pearce J A. Geochemical fingerprinting of oceanic basalts with applications to ophiolite classification and the search for Archean oceanic crust[J]. Lithos, 2008, 100(1-4)：14-48.

[5] 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.

[6] 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.

[7] Pearce J A. Trace element characteristics of lavas from destructive plate boundaries[M]//Thorpe R S. Orogenic Andesites and Related

Rocks. Chichester, England: John Wiley and Sons, 1982: 528-548.

[8] Rollinson H, Pease V. Using Geochemical Data: To Understand Geological Processes[M]. 2nd ed. Cambridge: Cambridge University Press, 2021: 226-278.

[9] 第鹏飞, 王金荣, 张旗, 等. 玄武岩构造环境判别图评估—全体数据研究的启示 [J]. 矿物岩石地球化学通报, 2017, 36(6)：891-896,879.［DI Pengfei, WANG Jinrong, ZHANG Qi, et al. The evaluation of basalt tectonic discrimination diagrams: Constraints on the research of global basalt data[J]. Bulletin of Mineralogy, Petrology and Geochemistry, 2017, 36(6)：891-896,879.］

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

[11] 周永章, 王俊, 左仁广, 等. 地质领域机器学习、深度学习及实现语言 [J]. 岩石学报, 2018, 34(11)：3173-3178.［ZHOU Yongzhang, WANG Jun, ZUO Renguang, et al. Machine learning, deep learning and Python language in field of geology[J]. Acta Petrologica Sinica, 2018, 34(11)：3173-3178.］

[12] Bergen K J, Johnson P A, De Hoop M V, et al. Machine learning for data-driven discovery in solid Earth geoscience[J]. Science, 2019, 363(6433)：eaau0323.

[13] 周志华. 机器学习 [M]. 北京: 清华大学出版社, 2016.［ZHOU Zhihua. Machine Learning[M]. Beijing: Tsinghua University Press, 2016.］

[14] 刘坤, 刘文波. 机器学习与大陆板内玄武岩构造环境判别 [J]. 工程技术与管理, 2017, 1(2)：188-191.［LIU Kun, LIU Wenbo. Machine learning and identification of the tectonic environment of basalt in the continental plate[J]. Engineering Technology & Management, 2017, 1(2)：188-191.］

[15] 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.［JIAO Shoutao, ZHOU Yongzhang, ZHANG Qi, et al. Study on intelligent discrimination of tectonic settings based on global gabbro data from GEOROC[J]. Acta Petrologica Sinica, 2018, 34(11)：3189-3194.］

[16] 任秋兵, 李明超, 李玉琼, 等. 基于全球橄榄石数据的玄武岩构造环境智能判别方法及其验证 [J]. 大地构造与成矿学, 2020, 44(2)：212-221.［REN Qiubing, LI Mingchao, LI Yuqiong, et al. An intelligent method for geochemical discrimination of tectonic settings of basalt based on olivine composition: GWO-SVM method and its verification[J]. Geotectonica et Metallogenia, 2020, 44(2)：212-221.］

[17] Guo P, Yang T, Xu W L, et al. Machine learning reveals source compositions of intraplate basaltic rocks[J]. Geochemistry, Geophysics, Geosystems, 2021, 22(9)：e2021GC009946.

[18] 余星. 海底岩石地球化学研究中的"大数据": PetDB 及其应用 [J]. 地球科学进展, 2014, 29(2)：306-314.［YU Xing. The big data tool for seabed petrogeochemistry research-PetDB and its application in geoscience[J]. Advances in Earth Science, 2014, 29(2)：306-314.］

[19] 葛粲, 汪方跃, 李永东, 等. 基于 GEOROC 大数据分析地壳厚度地球化学指标 [J]. 岩石学报, 2018, 34(11)：3179-3188.［GE Can, WANG Fangyue, LI Yongdong, et al. Analysis of geochemical indices of crustal thickness based on GEOROC big data[J]. Acta Petrologica Sinica, 2018, 34(11)：3179-3188.］

[20] 张晓琴, 程誉莹. 基于随机森林模型的成分数据缺失值填补法 [J]. 应用概率统计, 2017, 33(1)：102-110.［ZHANG Xiaoqin, CHENG Yuying. Imputation of missing values for compositional data based on random forest[J]. Chinese Journal of Applied Probability and Statistics, 2017, 33(1)：102-110.］

[21] 朱紫怡, 周飞, 王瑀, 等. 基于机器学习的锆石成因分类研究 [J]. 地学前缘, 2022, 29(5)：464-475.［ZHU Ziyi, ZHOU Fei, WANG Yu, et al. Machine learning-based approach for zircon classification and genesis determination[J]. Earth Science Frontiers, 2022, 29(5)：464-475.］

[22] Breiman L. Using iterated bagging to debias regressions[J]. Machine Learning, 2001, 45(3)：261-277.

[23] Cortes C, Vapnik V. Support-vector networks[J]. Machine Learning, 1995, 20(3)：273-297.

[24] Pearce J A. Immobile element fingerprinting of ophiolites[J]. Elements, 2014, 10(2)：101-108.

[25] Dai J G, Wang C S, Stern R J, et al. Forearc magmatic evolution during subduction initiation: Insights from an Early Cretaceous Tibetan ophiolite and comparison with the Izu-Bonin-Mariana forearc[J]. GSA Bulletin, 2021, 133(3-4)：753-776.

[26] Clarke D B, Cameron B I, Muecke G K, et al. Early Tertiary basalts from the Labrador Sea floor and Davis Strait region[J]. Canadian Journal of Earth Sciences, 1989, 26(5)：956-968.

[27] Deng H, Peng S B, Polat A, et al. Neoproterozoic IAT intrusion into Mesoproterozoic MOR Miaowan Ophiolite, Yangtze Craton: evidence for evolving tectonic settings[J]. Precambrian Research, 2017, 289：75-94.

[28] Güneş A, İlbeyli N, Rasimgil S, et al. Petrological and geochemical characteristics of the diabase and metasomatised dikes from the Tekirova ophiolite (SW Anatolia, Turkey): Tectonomagmatic evolution of the southern Neotethys[J]. Geochemistry, 2021, 81(3)：125767.

[29] 熊庆. 蛇绿岩记录的大洋地幔内熔体迁移过程 [J]. 矿物岩石地球化学通报, 2021, 40(5)：999-1011.［XIONG Qing. Ophiolitic records of melt migration processes in oceanic mantle[J]. Bulletin of Mineralogy, Petrology and Geochemistry, 2021, 40(5)：999-1011.］

[30] 卢泓宇, 张敏, 刘奕群, 等. 卷积神经网络特征重要性分析及增强特征选择模型 [J]. 软件学报, 2017, 28(11)：2879-2890.［LU Hongyu, ZHANG Min, LIU Yiqun, et al. Convolution neural network feature importance analysis and feature selection enhanced model[J]. Journal of Software, 2017, 28(11)：2879-2890.］

[31] 赵庆媛, 叶春茂, 鲁耀兵. 基于随机森林的微动特征重要性评估研究 [J]. 现代防御技术, 2022, 50(4)：124-131.［ZHAO Qingyuan, YE Chunmao, LU Yaobing. A micro-motion feature importance evaluation algorithm based on random forest[J]. Modern Defence Technology, 2022, 50(4)：124-131.］