

机器学习在岩矿地球化学研究中的应用

——综述与思考

谢玉芝, 汪洋

中国地质大学(北京)地球科学与资源学院, 北京, 100083



Pre-pub. on line: www.
geojournals.cn/georev

内容提要: 岩石与矿物的地球化学成分数据具有高维度特征。传统的岩矿地球化学成分研究主要采用二元/三元图解判别法, 准确率不高, 在数理统计方法上有欠缺。机器学习方法非常适用于对大样本高维度的岩矿成分数据进行数理统计处理。本文在介绍机器学习常见算法基本原理的基础上, 总结近5年来国内外学者将机器学习方法应用于岩石矿物成分数据研究的实例, 包括: ① 根据矿物成分溯源其母岩(源岩)、判别矿床类型, ② 新生代火山岩溯源, ③ 判别变质岩原岩, ④ 依据岩浆岩成分判别大地构造环境等。已有的研究实例显示, 机器学习方法的准确度明显优于传统的低维度判别法。机器学习本质是分析大样本数据的高维度变量之间的相关、归类等多元统计问题。推广机器学习的应用需要建设开放获取(Open Access)的矿物、岩石成分数据库, 同时全面实施开放研究(Open Research)的发表策略。

关键词: 机器学习; 岩矿成分; 判别图解; 大数据

地球科学是数据驱动型(data-driven)的学科, 随着信息技术的进步和普及, 伴随人工智能研究的发展, 应用地球科学大数据发掘数据之间的联系以提高找矿勘探效率、探究新的地学认知已成为当下的现实。(Petrelli and Perugini, 2016; 周永章等, 2017, 2018a, 2021)。机器学习是大数据时代进行数据信息化处理的重要手段之一, 其实质是通过大样本观察数据的学习和训练得到最优性能度量决策模型, 基于此模型对未知数据做出最佳决策或预测(周志华, 2016; 周永章等, 2018b)。现代岩石学、矿物学、矿床学和地球化学研究的发展过程中已经积累了大量的分析资料(全岩化学成分、同位素、矿物成分等)(张颖慧等, 2020), 已广泛应用于岩石起源演化、大地构造背景、成矿潜力等方面的研究。传统的岩矿地球化学成分数据研究的一个主要手段是各类判别图解, 例如: 基于岩浆岩化学成分的花岗岩类型判别图解(Bonin et al., 2020; 邓晋福等, 2015)、玄武岩或花岗岩主微量元素大地构造环境判别图解(Pearce and Cann, 1973; 邓晋福等, 2015; Xia Linqi and Li Xiangmin, 2019)、变质岩原岩判别图解(王仁民等, 1987)、造岩矿物源岩判别图

解(Schrön et al., 1988; Rusk, 2012)等。这些早期研究结果存在数据集样本数量少、分区界线依赖制作者的主观经验等缺点, 而且只能以二维图解的形式表达二元或三元数据, 在数据处理缺乏坚实的数理统计学基础, 无法处理多于三个变量的高维度数据(Snow, 2006; Verma, 2020)。岩矿地球化学数据是高维度的数据, 仅岩石的主要氧化物成分(SiO_2 、 TiO_2 、……、 P_2O_5)就是一个8或9维的变量。机器学习是基于计算技术对大样本数据进行数理统计处理的技术手段, 应用机器学习的方法对大样本的岩矿成分数据进行研究无疑可以有效地推进岩矿地球化学研究的深度和广度(周永章等, 2018a, b)。本文简介机器学习常见算法的基本原理, 综述机器学习在岩石、矿物地球化学研究方面的应用现状, 在此基础上对相关的科学问题提出少许浅见, 以求抛砖引玉。

1 机器学习算法原理简介

机器学习是利用概率论、数理统计、统计学和复杂算法等知识通过计算机来模拟人的学习过程(陈海虹等, 2017)。在大数据环境下, 机器学习可以从

注: 本文为中国地质调查局地质调查项目(编号:D1912)的成果。

收稿日期: 2022-09-27; 改回日期: 2023-01-12; 网络首发: 2023-01-20; 责任编辑: 章雨旭。Doi: 10.16509/j.georeview.2023.01.095

作者简介: 谢玉芝, 女, 1993年生, 硕士研究生, 地质工程专业; Email: 2215703425@qq.com。通讯作者: 汪洋, 男, 1969年生, 主要从事岩石学和地热学研究工作; Email: allen_thalassa@sina.com。

大数据中提取隐藏、有效的信息,实现对数据的分类和预测。机器学习分为监督学习和非监督学习两大类;监督学习是通过已知标签的学习(训练)样本数据的运算得到最优分类模型,再依据该模型对未知数据进行分类、预测。常用的监督学习算法包括:K 邻近、朴素贝叶斯、支持向量机、决策树、随机森林等。非监督学习以人工神经网络算法为代表,应用算法训练无标签样本数据找到其潜在结构,实现数据分类和预测。

1.1 K 邻近算法(KNN)

K 邻近算法(K-Nearest Neighbour, KNN)的基本原理是:将在特征空间内距离某一类别样本集最近的样本归属为该类别,判别函数是欧氏距离(公式 1):

$$D(X, Y) = \sqrt{\sum_{i=1}^n (x_i - y_i)^2} \quad (1)$$

该算法常用来处理分类与回归等常见问题(周永章等, 2018b)。

1.2 朴素贝叶斯算法(NB)

朴素贝叶斯算法(Naive Bayes, NB)是基于贝叶斯定理且假设特征条件之间相互独立的一种分类算法。其基本原理:通过数据训练求得各子集的概率分布,对于需要被分类的未知样本,计算其对应于各个已知分类子集的贝叶斯概率值,据此实现未知样本的归类(Pignatelli et al., 2021)。该算法逻辑简单,算法稳定,要求数据属性之间相互独立;当数据属性之间存在相关性时分类效率会降低。

1.3 支持向量机算法(SVM)

支持向量机(Support Vector Machine, SVM)是 Cortes (1995)提出的应用统计学习理论和结构风险最小化原理来寻找最优决策超平面的算法。通过超平面划分 N 维空间来对未知样本分类。SVM 算法将分类问题处理为函数 $f(x)$ 的最优化问题(韩帅等, 2018)。对于线性问题,将寻求最优分类平面转化为求二次规划问题,用拉格朗日乘子法解决约束法问题:

$$\begin{cases} \min f(x) & x \in E^n \\ \text{s. t. } \varphi(i) \geq 0 & x \in \{1, 2, 3, \dots, m\} \end{cases} \quad (2)$$

式中: $f(x)$ 表示目标函数, s. t. 表示 subject to, 受限的意思, φ_i 表示约束条件。

对于非线性问题,引入合适的核函数 $k(x_i, x)$, 可以将输入变量转换为某个高维空间实现线性分离,来达到求最优分类超平面,分类函数表达为

$$g(x) = \sum_{i=1}^n \varphi_i y_i k(x_i, x) + b \quad (3)$$

式中: $g(x)$ 是分类函数, φ_i 是拉格朗日因子, (x_i, y_i) 是已知样本, x 是未知数; $k(x_i, x)$ 是核函数。

SVM 算法具有泛化能力强的特点,但存在对缺失数据较敏感,会受到核函数和参数选取影响的缺点(周志华, 2016)。

1.4 决策树算法(DT)

决策树(Decision Tree, DT)的算法结构类似于树,包含“根”节点、“枝”节点和“叶”节点,计算流程是从“根”节点出发,依次判断各层节点的数据属性,实现数据的逐层分类,直至最后的“叶”节点。对于连续数据变量,算法在各节点设置阈值,将阈值两侧的数据分成两组,以此类推。其本质是多级次的二分类问题。该算法简单,结果易于解释,但容易过度拟合。

1.5 随机森林算法(RF)

随机森林(Random Forest, RF)是由多个决策树组成的算法分类器(Breiman, 2001),通过独立、随机的多个决策树(并行)计算的方式对已知样本学习(训练)。该算法能有效处理高维度的离散型和连续性数据,训练速度快,不容易过度拟合,容易实现并行运算(Belgiu and Dragut, 2016; Ueki et al., 2018)。

1.6 人工神经网络算法(ANN)

人工神经网络(Artificial Neural Network, ANN)由输入层、隐藏层和输出层的多个神经元组成,神经元之间相互连接构成网络。输入层用于获取输入信息,隐藏层用于特征提取,输出层输出隐藏层的处理结果。ANN 算法学习能力强,对噪声数据的鲁棒性(robustness)和容错性强,但涉及网络通路的权值和神经元的阈值等参数数量大,运算过程中需要不断调整各类参数来达到优化分类的效果。

2 机器学习算法在矿物岩石成分方面的应用

应用机器学习对多维数据进行统计建模、对未知数据归属进行预测,这是当前机器学习在岩矿地球化学方面的应用重点。现有的研究实践表明,机器学习适合处理广泛的岩矿地球化学问题,其结果准确度较传统图解法有相当大的提升。相较于传统的二元/三元判别图解,机器学习方法可以快速处理大样本数据集(“大数据”),尤其是对高维数据进行统计分析,这是传统方法所无法比拟的。图 1

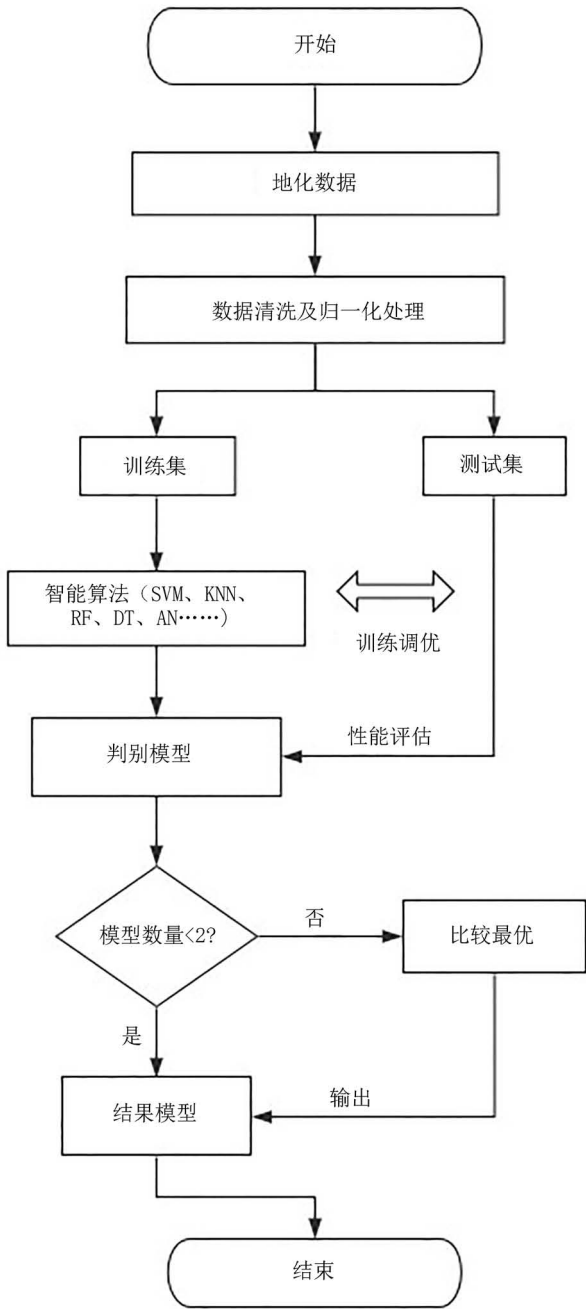


图1 应用于岩石、矿物成分数据的机器学习流程图

Fig. 1 The machine learning flow chart for the studies on rock and mineral composition data

是处理岩石、矿物成分数据的机器学习流程。

2.1 基于岩石化学成分的岩浆岩溯源及恢复变质岩原岩

Petrelli 等(2017)、Pignatelli 和 Piochi (2021)、Bolton 等(2020)应用机器学习分别对意大利中部、南部以及阿拉斯加第四纪火山岩(灰)的成分进行

了成分溯源。Bolton 等(2020)的研究结果表明,ANN 和 RF 算法的判别模型表现最好,精度达 96%。Ouzounis 和 Papakostas (2021)利用多种机器学习算法对爱琴海新生代火山岩进行溯源,训练结果表明极限梯度提升模型取得效果最好,准确率达 93.07%。Valeitich 等(2021)基于机器学习方法对伊豆—小笠原—马利亚纳(Izu—Bonin—Mariana)俯冲带的玻古安山质火山玻璃成分进行了归类研究,在此基础上探讨了玻古安山质熔体的分离结晶演化过程。Guo Peng 等(2021)采用 SVM 和 DDN(深度神经网络)算法对中国东北地区新生代玄武岩的源岩类型进行了判别。

Hasterok 等(2019)基于全球变质岩成分数据,应用 RF、ANN、SVM、KNN 和回归树 5 种算法,根据变质岩的主量元素组成判别其原岩,结果显示决策树算法类中的 RustBoosts 算法可以有效判别变质岩样品的原岩,其准确度优于传统的成分鉴别图解。Matsuno 等(2022)采用属于决策树算法大类的轻量梯度提升算法(light gradient boosting machine, LightGBM)构建了判别变质玄武岩原岩成分的成分模型,用于研究变质过程中的质量迁移。

2.2 应用矿物成分判别矿床成因类型或母岩类型

通过矿石或脉石矿物成分判别矿床成因类型是矿床学研究者关注的课题之一(Schrön et al., 1988; Rusk, 2012; Breiter et al., 2020)。传统上使用的判别方法是二元和三元图解,不能展示多维度信息特征(王瑀等, 2022)。全球矿床矿物成分数据的积累为应用机器学习算法判别矿床成因提供了可能(周永章等, 2017)。Hong Shuang 等(2021)利用 RF、SVM 和 XGBoost(极端梯度提升)算法研究 EPMA(电子探针显微分析仪)和 LA-ICP-MS(激光烧蚀电感耦合等离子体质谱)测量的磁铁矿化学成分来区分 IOCG 和 IOA 矿床,认为磁铁矿可以有效区分这两类矿床,其中 LA-ICP-MS 数据集的分类精度为 91%,EPMA 数据集的分类精度为 76%。Sun Guotao 等(2022)基于 SVM、DT 和 KNN 算法对辽宁青城子铅锌矿田的黄铁矿成分进行分类判别,结果表明机器学习方法可以成功判别该矿床的成因类型,其中 DT、KNN 和 SVM 算法的判别精度分别达到 98.2%、96.4%和 93.6%。Zhong Richen 等(2021)应用机器学习方法研究了不同成矿阶段的黄铁矿成分变化。王瑀等(2022)根据全球七类典型矿床 1220 个石英微量元素数据,应用机器学习的监督学习算法构建了 Ti/Ge—P 判别图解,对产自不同类

型矿床的石英进行分类。周统等(2021)运用穷举法构建了新的磷灰石源岩 Eu/Y—Ce 判别图解。Itano 等(2020)采用机器学习与多项回归分析的方法构建了不同来源独居石的成分判别模型,以解决独居石母岩溯源问题。Schönig 等(2021)运用 RF 算法研究碎屑石榴子石主量元素的化学成分来预测其母岩,结果准确率高、可解释性好。

2.3 花岗岩成矿潜力评价

黄鑫怀等(2022)应用 KNN 和 RF 算法研究了南岭中段花岗岩的铀矿成矿潜力。根据研究区花岗岩的主、微量元素数据集构建判别模型,结果显示 RF 算法分类模型的判别准确率达 93%,优于 KNN 算法模型。同时,对训练样本以外的红山、茶山和九峰岩体进行评价预测,结果显示:红山和茶山含矿概率较高,而九峰山岩体含矿概率较低(表 1)。

表 1 九峰、红山和茶山岩体铀矿成矿概率随机森林分类预测结果表(黄鑫怀等, 2022)

Table 1 Predicted uranium ore potential of the Jiufeng, Hongshan, and Chashan batholith by random forest classification method (Huang Xinhuai et al., 2022&)

	编号	不含矿概率 (%)	含矿概率 (%)	预测结果
九峰岩体	06 168	64	36	0
	06170	86	14	0
	06171	89	11	0
	06172	93	7	0
	06173	84	16	0
红山岩体	0629	17	83	1
	0631	11	89	1
	0632	10	90	1
	0633	10	90	1
	0635	9	91	1
茶山岩体	06184	4	96	1
	06185	3	97	1
	06186	1	99	1

2.4 岩浆岩大地构造环境判别

根据岩浆岩的成分特征判别其形成的大地构造环境是岩石学、地球化学研究的一个重要议题(Snow, 2006;Li Chusi et al., 2015;邓晋福等,2015;任秋兵等, 2019; Doucet et al., 2022)。岩浆岩构造环境判别图解得到研究者的广泛应用,如;Pearce 和 Cann (1973)的经典论文已得到 2914 次引用(最近 5 年的引用次数达 584 次,来源:Scopus, 2022 年 9 月 21 日)。但是传统二元/三元判别图解存在准确度欠佳、结果具有多解性、使用范围局限等问题

(Snow, 2006; Li Chusi et al., 2015; 韩帅等, 2018)。基于此,有研究者应用机器学习方法研究岩浆岩构造环境判别问题。

Petrelli 和 Perugini (2016) 利用 PetDB 和 GEOROC 数据库的 3095 个火山岩的全岩主、微量元素与 Sr—Nd—Pb 同位素数据(29 维),应用 SVM 算法判别其形成的八类大地构造环境——大陆弧(continental arc)、岛弧(island arc)、洋内弧(intra-oceanic arc)、弧后盆地(back arc basin)、大陆溢流玄武岩省(continental flood)、大洋中脊(mid-ocean ridge)、海洋高原(oceanic plateau)、洋岛(oceanic island)。判别准确率平均 93%,其中洋岛环境火山岩的判别正确率高达 99%,弧后盆地构造环境的判别率最低(65%)。Ueki 等(2018)同样应用 PetDB 和 GEOROC 数据库的 2074 个火山岩全岩 17 种主、微量元素以及 Sr—Nd—Pb 同位素数据,采用 SVM、RF 和稀疏多元逻辑回归(sparse multinomial regression, SMR)算法判别其形成的八类大地构造环境,所得研究结果与 Petrelli 和 Perugini (2016)类似。

韩帅等(2018)应用多种机器学习算法(NB、KNN、SVM、RF)建立了洋中脊玄武岩(MORB)、洋岛玄武岩(OIB)和岛弧玄武岩(IAB)的全岩化学成分判别模型,其中 RF 算法的判别准确率可达 88%。任秋兵等(2019)基于遗传算法来优化神经网络的耦合判别法(GA-NNDM 法)判别玄武岩构造环境,以求降低因高维数据的特征筛选和众多复杂的未知参数影响分类准确性的问题。方思源(2021)采用改进型随机森林算法对陆内、汇聚板块边缘和洋岛玄武岩进行判别,结果的 F1-score 值(精度和召回率的综合平均数)平均值为 84.32%。焦守涛等(2018)基于 GEOROC 数据库,利用 SVM、RF 和 KNN 算法分析辉长岩的地球化学数据特征,以求判断其产出的四种大地构造环境——大陆溢流玄武岩省、汇聚边界、板内、大洋岛弧。结果表明,RF 算法判别模型的效果最好,精确率达 97%,而 SVM 判别效果最差(表 2);据此认为辉长岩成分数据也能用于判断岩浆岩大地构造背景。

Ren Qiubing 等(2019)融合神经网络的多层感知器与灰狼优化算法,依据玄武岩中的橄榄石成分判别三大类玄武岩,模型判别准确率在 85%以上。Han Shuai 等(2019)应用 20 种机器学习算法对火成岩中的尖晶石成分进行分类建模,以判断其母岩形成的大地构造环境,报道的判别准确率大于 84%。

Liu Baoshun 和 Shi Junxia (2022) 根据 11331 条玄武岩全岩化学、微量元素和 Sr—Nd—Pb 同位素 (29 维) 数据,应用粒子群优化支持向量机 (particle swarm optimized support vector machine, PSO-SVM) 算法判别大地构造环境。对四类大地构造背景 (汇聚大陆边缘、洋岛、板内、大陆溢流) 玄武岩的判别准确率多在 80% 以上, F1-score 值大于 90%。

Doucet 等 (2022) 根据中生代以来的 29407 条玄武岩成分数据,采用链式方程多元插值 (Multivariate Imputation by Chained Equation, MICE) 法估算不完备数据项,应用监督学习与投票系统 (Voting System) 相结合的方法构建并预测玄武岩的大地构造环境,准确率高于 96%。

Saha 等 (2022) 利用岩浆岩的黑云母成分进行原岩大地构造环境判别,采用的机器学习算法是 XGBoost 和 LightGBM。该研究给出的判别准确率平均值约 90%,据此认为黑云母的主量元素成分可以判别其岩浆岩母岩形成的大地构造环境。

表 2 辉长岩大地构造环境判别的三种
机器算法输出结果 (焦涛涛等, 2018)

Table 2 Comparison of three machine algorithms for
tectonic discrimination of gabbroic rocks (Jiao Shoutao et
al., 2018&)

SVM 分类报告	准确率	召回率	F1-score	数据量
大陆溢流玄武岩省	0.40	0.50	0.45	252
汇聚边界	0.73	0.33	0.46	673
板内	0.00	0.50	0.01	2
大洋岛弧	0.23	0.34	0.28	108
平均数/总计	0.60	0.37	0.43	1035
KNN 分类报告	准确率	召回率	F1-score	数据量
大陆溢流玄武岩省	0.92	0.93	0.92	312
汇聚边界	0.98	0.94	0.96	318
板内	0.83	0.91	0.87	235
大洋岛弧	0.94	0.88	0.91	170
平均数/总计	0.92	0.92	0.92	1035
RF 分类报告	准确率	召回率	F1-score	数据量
大陆溢流玄武岩省	0.95	0.98	0.96	302
汇聚边界	0.99	0.97	0.98	312
板内	0.95	0.94	0.95	262
大洋岛弧	0.99	0.98	0.98	159
平均数/总计	0.97	0.97	0.97	1035

3 讨论与思考

3.1 数据——机器学习的基石

机器学习的基石是大样本数据集。数据集的样

本数、数据完备性、数据质量和多维数据各变量之间的自相关性、时空涵盖程度对研究结果具有决定性影响。

(1) 目前已发表的岩矿地球化学方面的机器学习研究实例中,一些数据样本并不大,例如:玄武岩大地构造背景研究 (样本数 755 个,韩帅等, 2018); 应用石英成分判别矿床类型的研究 (样本数 1220 个,王瑀等, 2022); 磷灰石源岩判别研究 (样本数 1925 个,周统等, 2022); Ueki et al (2018)、Petrelli 和 Perugini (2016) 火山岩构造环境判别机器学习研究的数据量为 2000 ~ 3000 量级。尚不足以称之为“大数据”研究。

(2) 现有岩矿数据库 (GEOROC、PetDB、EarthChem) 的不同数据条目记录的元素 (和同位素) 种类并非完全一致,即:对于涉及高维 (> 10) 数据的研究问题而言 (如:岩浆岩大地构造环境判别),数据完备性存在欠缺。此问题的一种解决方法是进行数据筛选 (filtering) (如:Ueki et al., 2018; Petrelli and Perugini, 2016)。其不足之处是:有可能 (大幅) 缩减样本数量、同时加剧数据的地理分布偏差 (bias) 程度。另一种解决方法是对“残缺”数据进行估算 (imputation) (Doucet et al., 2022)。其本质仍然是统计估计/插值方法,插值数据的可靠性低于实测数据。

(3) 岩矿地球化学数据库中收录数据的质量对机器学习结果的影响也不应被忽视。以 GEOROC 为代表的岩石成分数据库对收录的数据采取与原始文献相一致的原则,即:不修订原始数据、直接复制原文的地质背景等关键信息。在出现印刷错误、重复录入 (引用前人数据)、对某些数据样本所处构造环境的认识存在不同认识等情况下,若不进行数据清洗,直接应用数据进行机器学习会对结果产生影响 (Verma, 2020)。不同的数据清洗原则或算法导致不同研究者处理同一问题时所用的样本数量与质量存在差异,机器学习的结果难以直接对比 (对比: Petrelli and Perugini, 2016; Ueki et al., 2018; Liu Baoshun and Shi Junxia, 2022)。遗憾的是对数据清洗算法进行详细描述的研究并不多 (Liu Baoshun and Shi Junxia, 2022)。

(4) 对于矿物岩石主量元素数据而言,其主要氧化物的含量需要满足加和为 100% 的约束。这意味着这些氧化物的含量之间不是相互独立的,例如:岩浆岩 SiO₂ 含量增加必然导致其 MgO 等的含量减少 (Chayes, 1960; Verma, 2020)。这就需要采取先

求元素(氧化物)比值然后对各比值进行对数变换的方法实现各变量之间的统计独立性(Aitchison, 1986; Verma, 2020)。目前只有很少的研究实例考虑了这一点(Schönig et al., 2021)。

(5) 样本数据的时空偏差。地质科学数据具有明确的时空属性。现有岩矿地球化学成分数据库中涉及大陆地区的样品,绝大多数来自地质研究程度高的欧洲、北美、亚洲东部(中国、日本)以及澳大利亚,而非洲、南美洲、极地地区的数据量以及地理覆盖程度则远远不及(参见:Hasterok et al., 2019)。另外,主要岩矿地球化学数据库的数据源是英文文献,其它语种文献收录的很少。对研究程度较低地区开展研究是解决这一问题的关键。同时,发掘“长尾数据”和非英语(中文、西班牙文、俄文等)文献的数据可以为构建地理偏差程度更小的岩矿地球化学数据集提供帮助。

3.2 效率与准确度的平衡

对岩矿地球化学研究而言,机器学习的计算效率是传统(低维)统计方法(图解)所不能比拟的,已

有的研究结果显示机器学习的准确度明显优于传统方法。大多数研究实例采用了多种机器学习算法,比较而言监督学习的 SVM 算法准确率普遍优于 NB、DT、KNN 等,因此得到最广泛的应用(Petrelli and Perugini, 2016; Ueki et al., 2018; Doucet et al., 2022; Liu Baoshun and Shi Junxia, 2022)。无监督学习的 ANN 算法具有较好的普适性,颇受国内研究者重视(Ren Qiubing et al., 2019; 任秋兵等, 2019; 王瑀等, 2022)。但已有的研究表明,ANN 算法的超参数调整颇为复杂,在某些实例中,若不限定随机权重每次得到的结果不稳定(王瑀等, 2022)。不少研究者采用了多种算法的组合进行建模,如:将遗传算法、灰狼优化算法与 ANN 结合(Ren Qiubing et al., 2019; 任秋兵等, 2019)、粒子群优化(particle swarm optimized)与 SVM 结合(Liu Baoshun and Shi Junxia, 2022),以求提高模型预测的准确度。

采用多种算法进行机器学习建模可以比较不同算法在处理某一具体问题的准确度和效率。我们认

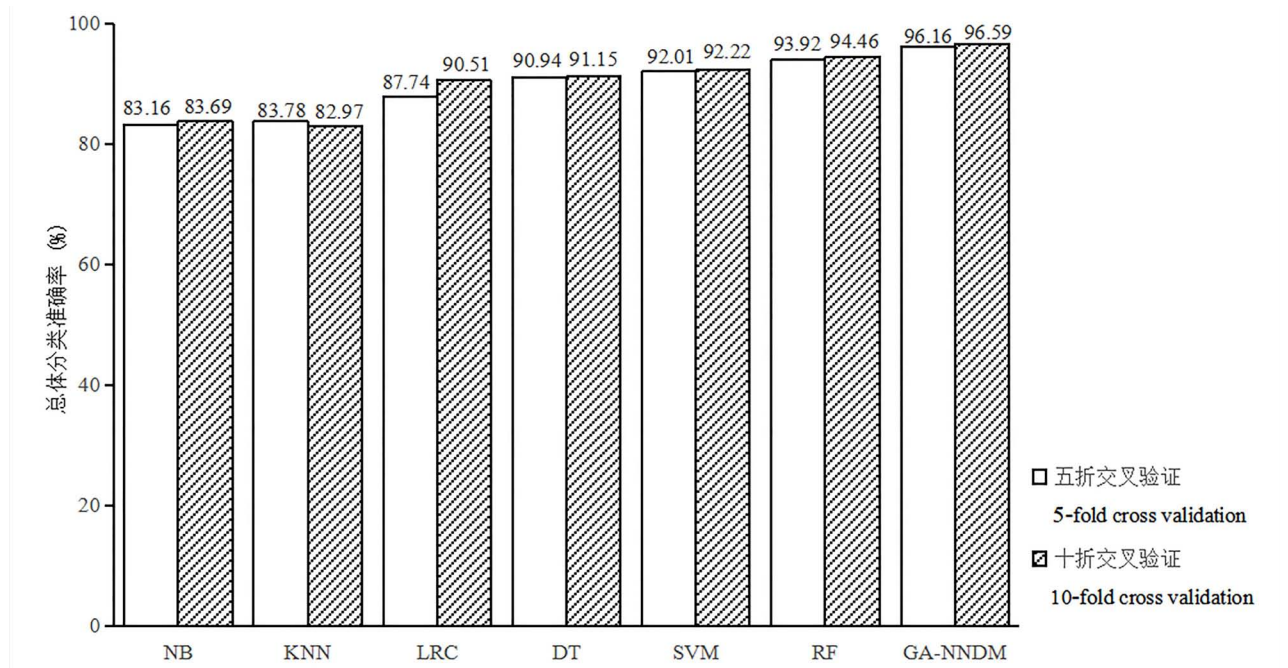


图2 任秋兵等(2019)所用机器学习算法准确度对比

Fig. 2 The accuracy of different machine-learning algorithms used by Ren Qiubing et al. (2019&)

十折交叉验证:一种常用测试算法准确性的方法,将数据分成10份,每次取其中9份用于训练,1份用于测试,将10次训练结果的正确率进行求均值表示算法精度的估计。五折交叉验证:将数据分成5份,每次取其中4份用于训练,1份用于测试,将5次训练结果的正确率进行求均值表示算法精度的估计

10-fold cross validation: a commonly used method to test the accuracy of algorithms. It divides the data into 10 copies, and takes 9 copies for training and 1 copy for testing each time, and averages the accuracy of 10 training results to represent the estimation of algorithm accuracy. 5-fold cross validation: it divides the data into 5 copies, and take 4 copies for training and 1 copy for testing each time, and averages the accuracy of 5 training results to represent the estimation of algorithm accuracy

为机器学习研究需要在算法的准确度与编程、计算效率之间取得平衡为宜。图 2 表示的是任秋兵等(2019)不同算法的准确度校验结果,可以看出 GA-NNDM、SVM、RF 等算法的准确率超过 90%,明显优于左侧准确率 83%的 KNN、NB 算法。但是,高准确率的各算法优劣差异并不明显。具体而言就是相较于 SVM、RF 等相对简单的算法,该研究采用的遗传算法——神经网络算法(GA-NNDM)对结果准确率的提升不明显。另一个实例是 Liu Baoshun 和 Shi Junxia (2022)的研究(图 3)。该研究表明,SVM、RF 与 KNN 算法的准确率显著优于 NB 和 AB (AdaBoost)算法,但是采用不同的优化策略(格网算法或粒子群优化算法)在大多数情况下没有明显改进算法的准确率。所以,我们认为没必要为追求过高的准确率而牺牲算法的简洁性。过于复杂的算法和代码并不利于交流和维护。

3.3 研究手段还是目的?

作为大数据研究主要方法之一的机器学习,可以有效揭示高维数大样本岩石、矿物成分数据的内在相关性(Saha et al., 2022)。因此,有学者认为“大数据时代”地质科学研究应该“从对因果关系的追求转变为对相关关系的追求”(张旗等, 2018)。我们认为利用机器学习乃至其他大数据研究方法得

到的数据相关性只是科学研究的第一步,是探究相关性背后因果关系的基础。以观察自然界现象为基础的地质学、生物学、天文学的起点就是对自然对象(现象)的描述、观测,在数据积累的基础上寻找不同事物(现象)之间的相关性,最终得到基于因果关系的逻辑认知。只不过受时代制约,科学早期发展阶段观测数据的精度和数量远不及现代而已。科学研究并非张旗等(2018)所认为的那样,以前追求因果关系,当下“大数据时代”需要追求相关关系。机器学习的最大优势在于高效揭示大样本高维数据的复杂相关关系,这才是机器学习研究的落脚点。若不能深入挖掘研究对象内在的因果关系,则研究结果的科学意义相当局限。例如:有关玄武岩构造环境判别的机器学习研究,国外学者(Ueki et al., 2018; Doucet et al., 2022)在建模、预测的基础上,进一步讨论了不同构造环境玄武岩地球化学特征及其成因;而相当数量的国内研究只是满足于算法的实现或不同算法的比较,缺乏对研究结果的地质解释(焦守涛等, 2018; 韩帅等, 2018; Ren Qiubing et al., 2019; 任秋兵等, 2019; Liu Baoshun and Shi Junxia, 2022)。

作为处理岩矿地球化学数据的一种很有前景的研究工具。机器学习方法推广中的障碍主要来自其

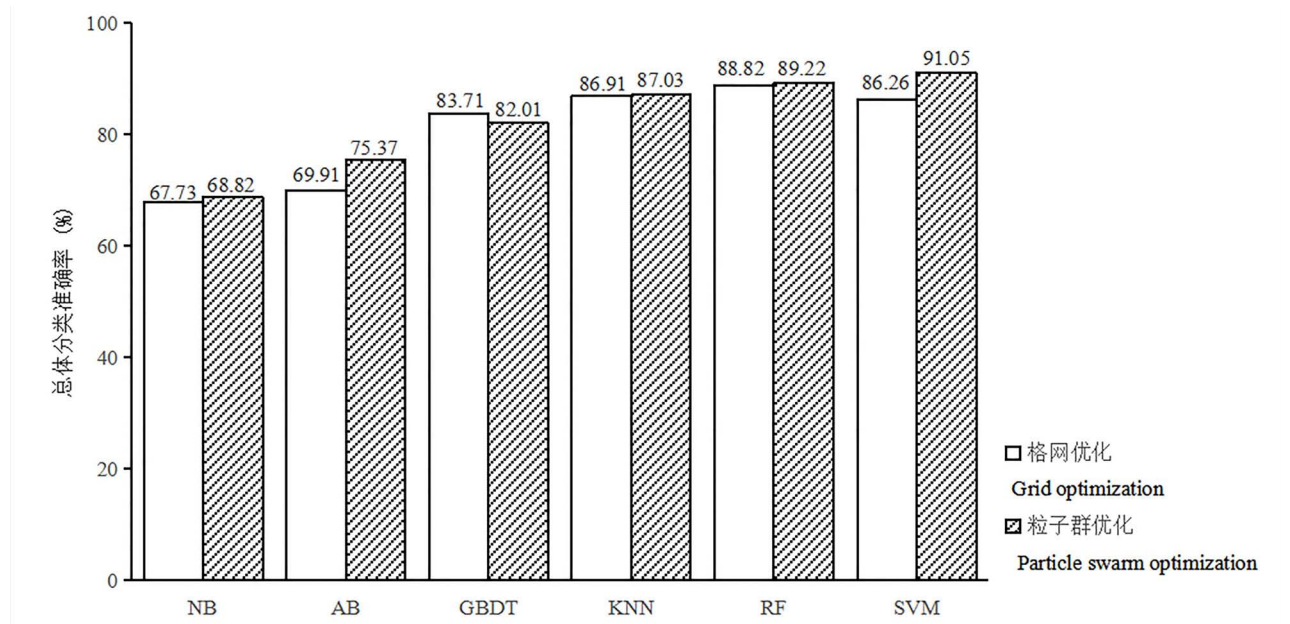


图 3 Liu Baoshun 和 Shi Junxia (2022) 所用 6 种机器学习算法准确率对比

Fig. 3The accuracy of six machine-learning algorithms used by Liu Baoshun and Shi Junxia (2022)

格网优化:对所有参数进行穷举搜索,得到使结果最优的参数;粒子群优化:通过随机搜索算法来找到全局最优解

Grid optimization: search all parameters exhaustively to obtain the parameters with the best results; Particle swarm optimization: find the global optimal solution through random search algorithm

陡峭的学习曲线(learning curve),由于相当数量的地质专业人员的计算机应用水平有限,难以将机器学习应用于工作之中。雪上加霜的是,相当数量发表的研究论文不提供机器学习的源程序代码,研究结果缺少可重复性,其他研究者也无法有效使用。在当前开放研究(Open Research)策略的推动下,欧美学者机器学习论文多附有源代码,对推进机器学习在岩矿地球化学领域的应用大有裨益。相反,亚洲学者的相关论文很少公开源代码,给人以“自娱自乐”之感。秉持开放研究思维,公开源代码有助于推动机器学习研究的可重复性、可移植性和规范性,是避免研究工具“黑箱”效应的不二法门。

4 结论和建议

应用机器学习方法处理岩石、矿物成分数据,可以在较高准确度的基础上高效实现对大样本高维度岩矿地球化学数据的多元统计建模与预测,近五年已被广泛用于岩浆岩大地构造环境判别、造岩矿物母岩识别、火山岩溯源、变质岩原岩判别、矿床成因类型判别、花岗岩成矿潜力预测等方面的研究。其判别准确度明显优于传统的二元/三元判别图解。

机器学习的科学实质仍然是分析样本变量之间的相关、归类等统计问题。在此意义上,机器学习是对传统二元/三元统计建模、判别方法的改进和升级而非“革命”。我们认为,机器学习在岩矿地球化学领域应用的进一步推广需要:

(1)建设更多开放获取(Open Access)的全球或区域矿物、岩石成分数据库。

(2)研究论文应该全面详细介绍其采用数据筛选/清洗(data filtering / cleaning)的原则、步骤和算法。

(3)研究论文需要公开机器学习程序的源代码。

致谢:感谢中山大学周永章教授对本文提出宝贵的意见和建议。

参 考 文 献 / References

(The literature whose publishing year followed by a “&” is in Chinese with English abstract; The literature whose publishing year followed by a “#” is in Chinese without English abstract)

邓晋福,刘翠,冯艳芳,肖庆辉,狄永军,苏尚国,赵国春,段培新,戴蒙. 2015. 关于火成岩常用图解的正确使用: 讨论与建议. 地质论评, 61(4): 717 ~ 734.

方思恩. 2019. 基于大数据和机器学习方法的玄武岩构造背景自动判别及其应用. 导师:董少春. 南京: 南京大学硕士学位论文: 1 ~ 60.

韩帅,李明超,任秋兵,刘承照. 2018. 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. 岩石学报, 34(11): 3207 ~ 3216.

黄鑫怀,李增华,邓腾,刘志锋,陈冠群,曾皓轩,郭世超. 2022. 基于机器学习的华南诸广山花岗岩体铀矿潜力评价[J][OL]. 地球科学: 1 ~ 23 [2022-08-19] (2022-01-28) <http://kns.cnki.net/kcms/detail/42.1874.P.20220128.0947.002.html>

焦守涛,周永章,张旗,金维浚,刘艳鹏,王俊. 2018. 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. 岩石学报, 34(11): 3189 ~ 3194.

任秋兵,李明超,韩帅. 2019. 基于改进遗传算法——神经网络的玄武岩构造环境判别及对比实验. 地学前缘, 26(4): 117 ~ 124.

任秋兵,李明超,李玉琼,韩帅,张野,张旗. 2020. 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. 大地构造与成矿学, 44(2): 212 ~ 221.

王仁民,贺高品,陈珍珍,郑松彦,耿元生. 1987. 变质岩原岩图解判别法. 北京: 地质出版社: 1 ~ 199.

王瑀,邱昆峰,侯照亮,于皓丞. 2022. 石英 Ti/Ge—P: 基于机器学习的矿床类型判别新图解. 岩石学报, 38(1): 281 ~ 290.

张旗,焦守涛,卢欣祥. 2018. 论地质研究中的因果关系和相关关系——大数据研究的启示. 岩石学报, 34(2): 275 ~ 280.

张颖慧,王涛,焦守涛,郭磊,范润龙,王杨刚,张建军. 2020. 国内外岩浆岩数据库现状与应用前景. 高校地质学报, 26(1): 11 ~ 26.

周统,邱昆峰,王瑀,于皓丞,侯照亮. 2022. 磷灰石 Eu/Y—Ce: 基于大数据的源区类型判别新图解. 岩石学报, 38(1): 291 ~ 299.

周永章,黎培兴,王树功,肖凡,李景哲,高乐. 2017. 矿床大数据及智能矿床模型研究背景与进展. 矿物岩石地球化学通报, 36(2): 327 ~ 331 + 344.

周永章,王俊,左仁广,肖凡,沈文杰,王树功. 2018a. 地质领域机器学习、深度学习及实现语言. 岩石学报, 34(11): 3173 ~ 3178.

周永章,张良均,张奥多,王俊. 2018b. 地球科学大数据挖掘与机器学习. 广州: 中山大学出版社: 180 ~ 183.

周永章,左仁广,刘刚,袁峰,毛先成,郭艳军,肖凡,廖杰,刘艳鹏. 2021. 数学地球科学跨越发展的十年: 大数据、人工智能算法正在改变地质. 矿物岩石地球化学通报, 40(3): 556 ~ 573 + 777.

周志华. 2016. 机器学习. 北京: 清华大学出版社.

Aitchison J. 1986. The Statistical Analysis of Compositional Data. London: Chapman and Hall: 1 ~ 416.

Belgiu M, Dragut L. 2016. Random forest in remote sensing: A review of applications and future directions. ISPRS Journal of Photogrammetry and Remote Sensing, 114: 24 ~ 31.

Bolton M S M, Jensen B J L, Wallace K, Praet N, Fortin D, Kaufman D, De Batist M. 2020. Machine learning classifiers for attributing tephra to source volcanoes: an evaluation of methods for Alaska tephra. Journal of Quaternary Science, 35(1~2): 81 ~ 92.

Bonin B, Vojtěch J, Moyen J F. 2020. Chemical variation, modal composition and classification of granitoids. Geological Society, London, Special Publications, 491: 9 ~ 51.

Breiman L. 2001. Random forests. Machine Learning, 45(1): 5 ~ 32.

Breiter K, Urišová J, Dosbaba M. 2020. Chemical signature of quartz from S- and A-type rare-metal granites——A summary. Ore Geology Reviews, 125: 103674

Chayes F. 1960. On correlation between variables of constant sum.

- Journal of Geophysical Research, 65: 4185 ~ 4193.
- Cortes C., Vapnik V. 1995. Support-Vector Networks. Machine Learning, 20: 273 ~ 297.
- Deng Jinfu, Liu Cui, Feng Yanfang, Xiao Qinghui, Di Yongjun, Su Shangguo, Zhao Guochun, Duan Peixin, Dai Meng. 2015&. Correct use of commonly used diagrams of igneous rocks: discussion and suggestions. Geological Review, 61 (4): 717 ~ 734.
- Doucet L S, Tetley M G, Li Zheng-Xiang, Liu Yebo, Gamaleldien H. 2022. Geochemical fingerprinting of continental and oceanic basalts: A machine learning approach. Earth-Science Reviews, 233: 104192
- Fang Siyuan. 2021&. Automatic Identification of Basalt Tectonic Setting Based on Big Data and Machine Learning Method and Its Application. Tutor: Dong Shaochun. Nanjing: Nanjing University Master Degree Thesis: 1 ~ 60.
- Guo Peng, Yang Ting, Xu Wenliang, Chen Bin. 2021. Machine learning reveals source compositions of intraplate basaltic rocks. Geochemistry Geophysics Geosystems, 22(9): e2021GC009946.
- Han Shuai, Li Mingchao, Ren Qiubing, Liu Chengzhao. 2018&. Intelligent mining discrimination and analysis of basalt geotectonic environment based on big data method. Acta Petrologica Sinica, 34 (11): 3207 ~ 3216.
- Han Shuai, Li Mingchao, Ren Qiubing. 2019. Discriminating among tectonic settings of spinel based on multiple machine learning algorithms. Big Earth Data, 3(1): 67 ~ 82.
- Hasterok D, Gard M, Bishop C M B, Kelsey D. 2019. Chemical identification of metamorphic protoliths using machine learning methods. Computers and Geosciences, 132: 56 ~ 68.
- HongShuang, Zuo Renguang, Huang Xiaowen, Xiong Yihui. 2021. Distinguishing IOCG and IOA deposits via random forest algorithm based on magnetite composition. Journal of Geochemical Exploration, 230: 106859.
- Huang Xinhui, Li Zenghua, Deng Teng, Liu Zhifeng, Chen Guanqun, Zeng Haoxuan, Guo Shichao. 2022&. Evaluation of uranium potential of Zhuguangshan granite body in South China based on machine learning [J][OL]. Earth Science: 1 ~ 23 ; [2022-08-19] (2022-01-28) <http://kns.cnki.net/kcms/detail/42.1874.P.20220128.0947.002.html>
- Itano K, Ueki K, Iizuka T, Kuwatani T. 2020. Geochemical discrimination of monazite source rock based on machine learning techniques and multinomial logistic regression analysis. Geosciences, 10(2): 63.
- Jiao Shoutao, Zhou Yongzhang, Zhang Qi, Jin Weijun, Liu Yanpeng, Wang Jun. 2018&. Research on intelligent discrimination of geotectonic environment based on global gabbro big data of GEOROC database. Acta Petrologica Sinica, 34(11): 3189 ~ 3194.
- Li Chusi, Arndt N T, Tang Qingyan, Ripley E M. 2015. Trace element indiscriminability diagrams. Lithos, 232: 76 ~ 83.
- Liu Baoshun, Shi Junxia. 2022. A machine learning-based approach to discriminating basaltic tectonic settings. International Journal of Computational Intelligence and Applications, 21(2): 2250012.
- Matsuno S, Uno M, Okamoto A, Tsuchiya N. 2022. Machine-learning techniques for quantifying the protolith composition and mass transfer history of metabasalt. Scientific Reports, 12(1): 1385.
- Ouzounis A G, Papakostas G A. 2021. Machine learning in discriminating active volcanoes of the Hellenic Volcanic Arc. Applied Sciences, 11: 8318.
- Pearce J A, Cann J R. 1973. Tectonic setting of basic volcanic rocks determined using trace element analyses. Earth and Planetary Science Letters, 19(2): 290 ~ 300.
- Petrelli M, Perugini D. 2016. Solving petrological problems through machine learning: the study case of tectonic discrimination using geochemical and isotopic data. Contributions to Mineralogy and Petrology, 171(10): 81.
- Petrelli M, Bizzarri R, Morgavi D, Baldanza A, Perugini D. 2017. Combining machine learning techniques, microanalyses and large geochemical datasets for tephrochronological studies in complex volcanic areas: new age constraints for the Pleistocene magmatism of Central Italy. Quaternary Geochronology, 40: 33 ~ 44.
- Pignatelli A, Piochib M. 2021. Machine learning applied to rock geochemistry for predictive outcomes: The Neapolitan volcanic history case. Journal of Volcanology and Geothermal Research, 415: 107254.
- Ren Qiubing, Li Mingchao, Han Shuai. 2019. Tectonic discrimination of olivine in basalt using data mining techniques based on major elements: a comparative study from multiple perspectives. Big Earth Data, 3(1): 8 ~ 25.
- Ren Qiubing, Li Mingchao, Han Shuai. 2019&. Basalt tectonic environment discrimination and comparison experiment based on improved genetic algorithm neural network. Earth Science Frontiers, 26(4): 117 ~ 124.
- Ren Qiubing, Li Mingchao, Li Yuqiong, Han Shuai, Zhang Ye, Zhang Qi. 2020&. Intelligent discrimination method of basalt tectonic environment based on global olivine data and its verification. Geotectonica et Metallogenia, 44(2): 212 ~ 221.
- Rusk B. 2012. Cathodoluminescent textures and trace elements in hydrothermal quartz. In: Götze J and Möckel R (eds.). Quartz: Deposits, Mineralogy and Analytics. Berlin, Heidelberg: Springer: 307 ~ 329.
- Saha R, Upadhyay D, Mishra B. 2021. Discriminating tectonic setting of igneous rocks using biotite major element chemistry——A machine learning approach. Geochemistry, Geophysics, Geosystems, 22 (11): e2021GC010053.
- Schönig J, von Eynatten H, Tolosana-Delgado R, Meinhold G. 2021. Garnet major-element composition as an indicator of host-rock type: a machine learning approach using the random forest classifier. Contributions to Mineralogy and Petrology, 176(12): 98.
- Schrön W, Schmadicke E, Thomas R, Schmidt W. 1988. Geochemische untersuchungen an pegmatitquarzen. Zeitschrift für Geologische Wissenschaften, 16(3): 229~244.
- Snow C A. 2006. A reevaluation of tectonic discrimination diagrams and a new probabilistic approach using large geochemical databases: moving beyond binary and ternary plots. Journal of Geophysical Research, 111(6): B06206.
- Sun Guotao, Zeng Qingdong, Zhou Jiayi. 2022. Machine learning coupled with mineral geochemistry reveals the origin of ore deposits. Ore Geology Reviews, 142: 104753.
- Ueki K, Hino H, Kuwatani T. 2018. Geochemical discrimination and characteristics of magmatic tectonic settings: A machine-learning-based approach. Geochemistry, Geophysics, Geosystems, 19(4): 1327 ~ 1347.
- Valetich M J, Losq C L, Arculus R J, Umuro S, Mavrogenes J. 2021. Compositions and classification of fractionated boninite series melts from the Izu - Bonin - Mariana Arc: A machine learning approach. Journal of Petrology, 62: egab013.
- Verma S P. 2020. Road from Geochemistry to Geochemometrics. Berlin: Springer: 1 ~ 669.
- Wang Renmin, He Gaopin, Chen Zhenzhen, Zheng Songyan, Geng Yuansheng. 1987#. Discrimination Diagrams for Metamorphic Petrology. Beijing: Geological Publishing House: 1 ~ 199.

- Wang Yu, Qiu Kunfeng, Hou Guangming, Yu Haocheng. 2022&. Quartz Ti/Ge—P: a new diagram for ore deposit type discrimination based on machine learning. *Acta Petrologica Sinica*, 38(1): 281 ~ 290.
- Xia Linqi, Li Xiangmin. 2019. Basalt geochemistry as a diagnostic indicator of tectonic setting. *Gondwana Research*, 65: 43 ~ 67.
- Zhang Qi, Jiao Shoutao, Lu Xinxiang. 2018&. On causality and correlation in geological research—enlightenment from big data research. *Acta Petrologica Sinica*, 34(2): 275 ~ 280.
- Zhang Yinghui, Wang Tao, Jiao Shoutao, Guo Lei, Fan Runlong, Wang Yanggang, Zhang Jianjun. 2020&. Current situation and application prospects of magmatic rock databases at home and abroad. *Geological Journal of China Universities*, 26(1): 11 ~ 26.
- Zhong Richen, Deng Yi, Li Wenbo, Danyushevsky L V, Cracknell M J, Belousov I, Chen Yanjing, Li Lamei. 2021. Revealing the multi-stage ore-forming history of a mineral deposit using pyrite geochemistry and machine learning-based data interpretation. *Ore Geology Reviews*, 133: 104079.
- Zhou Tong, Qiu Kunfeng, Wang Yu, Yu Haocheng, Hou Liangliang. 2022&. Apatite Eu/Y—Ce: a new diagram for source area type discrimination based on big data. *Acta Petrologica Sinica*, 38(1): 291 ~ 299.
- Zhou Yongzhang, Li Peixing, Wang Shugong, Xiao Fan, Li Jingzhe, Gao Le. 2017&. Research background and progress of deposit big data and intelligent deposit model. *Bulletin of Mineralogy, Petrology and Geochemistry*, 36(2): 327 ~ 331 + 344.
- Zhou Yongzhang, Wang Jun, Zuo Renguang, Xiao Fan, Shen Wenjie, Wang Shugong. 2018a&. Machine learning, deep learning and implementation language in geological field. *Acta Petrologica Sinica*, 34(11): 3173 ~ 3178.
- Zhou Yongzhang, Zhang Liangjun, Zhang Aoduo, Wang Jun. 2018b#. *Earth Science Big Data Mining and Machine Learning*. Guangzhou: Sun Yat-sen University Press: 180 ~ 183.
- Zhou Yongzhang, Zuo Renguang, Liu Gang, Yuan Feng, Mao Xiancheng, Guo Yanjun, Xiao Fan, Liao Jie, Liu Yanpeng. 2021&. Ten years of leapfrog development of mathematical geosciences: big data and artificial intelligence algorithms are changing geology. *Bulletin of Mineralogy, Petrology and Geochemistry*, 40(3): 556 ~ 573 + 777.
- Zhou Zhihua. 2016#. *Machine Learning*. Beijing: Tsinghua University Press.

A review on the machine learning approach to rock and mineral geochemistry research

XIE Yuzhi, WANG Yang

School of Earth Science and Resources, China University of Geosciences (Beijing), Beijing, 100083

Abstract: The geochemical composition data of rocks and minerals have high dimensional characteristics. The conventional study on the geochemical composition of rocks and minerals mainly adopts the binary/ternary graphical discrimination method, which has low accuracy and lacks solid basis of mathematical statistics. The machine learning method is very suitable for the statistical processing of large scale high dimensional data as the rock and mineral compositions. On the basis of introducing the basic principles of common machine learning algorithms, this paper summarizes the case studies of the machine learning approach to the rock and mineral geochemistry in the past five years, including: (1) discrimination of the source rock of the minerals from their compositions, (2) distinguishing the type of deposit from the mineral compositions, (3) identifying the provenance of Cenozoic volcanic rocks, (4) distinguishing the proto-lithology discrimination for metamorphic rocks, and (5) tectonic discrimination of magmatic rocks, etc. Compared with conventional low-dimensional discrimination method, the machine learning approach provides higher accuracy and the ability to process the high-dimensional data. The nature of machine learning approach is to perform the multivariate statistical analysis, such as the correlation and classification among the high-dimensional variables of large sample data. For popularizing the machine-learning approach in petrological community, more open accessed databases of mineral and rock compositions are needed, and the *Open Research* policy should be fully implemented in academic publications.

Keywords: machine learning; rock and mineral compositions; discriminant diagram; big data

Acknowledgements : This paper is a result of geological survey project (No. D1912) of China Geological Survey. We would like to thank Professor ZHOU Yongzhang of Sun Yat-sen University for his valuable comments and suggestions on this article

First author : XIE Yuzhi, female, born in 1993, master candidate, is mainly engaged in the application of machine learning to the studies of mineralogy and petrology; Email: 2215703425@qq.com

Corresponding author: WANG Yang, Email: allen_thalassa@sina.com

Manuscript received on: 2022-09-27; Accepted on: 2023-01-12; Published online on: 2022-12-20

Doi: 10.16509/j. georeview. 2022. 12. 095

Edited by: ZHANG Yuxu

