https://doi.org/10.3799/dqkx.2022.006

基于机器学习的华南诸广山花岗岩体铀矿潜力评价

黄鑫怀[1,2],李增华[1,2,3*],邓 腾[2,3],刘志锋[1],陈冠群[2],曾皓轩[2],郭世超[2]

1. 东华理工大学江西省放射性地学大数据技术工程实验室,江西南昌,330013
2. 东华理工大学地球科学学院,江西南昌,330013
3. 东华理工大学核资源与环境国家重点实验室,江西南昌,330013

摘 要: 地学大数据和机器学习的结合,为矿床勘查提供了新的发展方向.华南广泛发育花岗岩体,是花岗岩型铀矿的重要产区,因此如何判断特定花岗岩体是否具有产铀矿的潜力,对于指导华南花岗岩型铀矿勘查具有重要意义.系统收集了前人已发表的华南花岗岩地球化学元素含量数据(不包括待评价的诸广山地区的九峰岩体、红山岩体和茶山岩体),共获得1 711条数据.然后按照7∶3的比例划分为训练集和测试集,进而分别建立了随机森林(random forest,RF)算法和K近邻(K-nearest neighbor,KNN)算法分类模型,并对两种分类模型的精确度、召回率、ROC(receiver operating characteristic curve)曲线进行评价,选出泛化能力较好的模型,最后利用泛化能力较好的模型对诸广山地区九峰岩体、红山岩体和茶山岩体进行成矿潜力评价.结果表明,随机森林分类模型对测试集的分类精确度、预测结果可靠度均高于K近邻分类模型,随机森林分类模型对测试集上的数据分类精确度达到了93%,利用上述创建的随机森林分类模型对九峰、红山和茶山岩体进行预测.预测结果表明,红山岩体和茶山岩体含矿的概率较高,而九峰岩体含矿概率较低.该研究为进一步缩小地质找矿勘查范围提供了可靠的依据,并且该模型可以作为地质找矿工作者的辅助工具.

关键词: 机器学习;随机森林算法;K近邻算法;花岗岩型铀矿;成矿潜力;岩石学.

中图分类号: P628;P595 **文章编号:** 1000-2383(2023)12-4427-14 **收稿日期:** 2021-12-11

Uranium Potential Evaluation of Zhuguangshan Granitic Pluton in South China Based on Machine Learning

Huang Xinhuai[1,2], Li Zenghua[1,2,3*], Deng Teng[2,3], Liu Zhifeng[1], Chen Guanqun[2], Zeng Haoxuan[2], Guo Shichao[2]

1. *Jiangxi Engineering Laboratory on Radioactive Geoscience and Big Data Technology, East China University of Technology, Nanchang* 330013, *China*
2. *School of Earth Sciences, East China University of Technology, Nanchang* 330013, *China*
3. *State Key Laboratory for Nuclear Resources and Environment, East China University of Technology, Nanchang* 330013, *China*

Abstract: The combination of geological data and machine learning provides a new direction for mineral exploration. The granitic pluton is widely developed in South China, which is an important producing area for granite-type uranium deposits. Therefore, whether the granitic pluton has the potential to produce uranium deposits is of great significance for guiding the exploration of granite-type uranium deposits in South China. In this paper, the geochemical data of granites in South China are systematically

基金项目: 东华理工大学江西省放射性地学大数据技术工程实验室开放基金(No.JELRGBDT202006)资助.

作者简介: 黄鑫怀(1995-),男,硕士研究生,研究方向为铀矿地质.ORCID:0000-0001-5511-4054. E-mail:huangxinhuai2020@163.com

* **通讯作者:** 李增华,教授,从事铀矿地质研究.ORCID: 0000-0003-2420-668X. E-mail: lizenghua@ecut.edu.cn

引用格式: 黄鑫怀,李增华,邓腾,刘志锋,陈冠群,曾皓轩,郭世超,2023.基于机器学习的华南诸广山花岗岩体铀矿潜力评价.地球科学,48(12):4427-4440.

Citation: Huang Xinhuai, Li Zenghua, Deng Teng, Liu Zhifeng, Chen Guanqun, Zeng Haoxuan, Guo Shichao, 2023. Uranium Potential Evaluation of Zhuguangshan Granitic Pluton in South China Based on Machine Learning. *Earth Science*, 48(12): 4427-4440.

collected (excluding the Jiufeng, Hongshan and Chashan granite plutons to be evaluated in Zhuguangshan area) from previous published papers, and a total of 1 711 data pieces are obtained. They are further divided into training set and test set according to the ratio of 7∶3. Then, the random forest (RF) algorithm and *K*-nearest neighbor (KNN) algorithm classification models were established respectively, and the accuracy, recall rate and ROC (receiver operating characteristic curve) curve of the two classification models were evaluated, and the models with good generalization ability were selected. Finally, the metallogenic potential of the Jiufeng pluton, Hongshan pluton and Chashan pluton in the Zhuguangshan area were evaluated using the models with good generalization ability. The results show that the classification accuracy and reliability of prediction results of random forest classification model are higher than those of *K*-nearest neighbor classification model, and the classification accuracy of the random forest classification model on the test set reached 93%. The random forest classification model created above was used to evaluate the metallogenetic potential of the Jiufeng, Hongshan and Chashan plutons. The prediction results show that the probability of metallogenetic potentiality in the Hongshan and Chashan plutons is high, whereas the probability in the Jiufeng pluton is low. This study provides a reliable basis for further geological prospecting, and the model can be used as an auxiliary tool for geological prospecting.

Key words: machine learning; random forest algorithm; *K*-nearest neighbor algorithm; granite-type uranium deposit; metallogenetic potentiality; petrology.

0 引言

目前,随着信息技术以及人工智能的高速发展,地学研究进入一个新的阶段.基于机器学习的地质矿产评价与数据挖掘现在已成为当前数字地球科学的热门领域,通过对前人积累的海量地球化学数据进行深一步的挖掘,并利用机器学习进行成矿潜力评价,为矿产勘查提供了新方向(张旗和周永章,2017;周永章等,2017,2018a,2021;郝慧珍等,2021;左仁广等,2021).机器学习是一个源于数据训练过程的模型,经过训练而最终给出一个最优的性能度量决策(周永章等,2018b).根据所处理数据类别的不同,机器学习可以分为监督学习和无监督学习,监督学习是告诉计算机在某个特定的情况下输出的正确结果,希望计算机从这些情况中面对没有见过的输入变量时也能给出正确的结果预测,常用的算法有支持向量机(陈永良等,2012;Rodriguez-Galiano *et al.*,2015)、随机森林(Vincenzi *et al.*,2011;Rodriguez-Galiano *et al.*,2015)和朴素贝叶斯分类器(Youn and Jeong,2009)等;无监督学习,是指在输入变量时没有给出特定的输出结果,希望计算机从数据中深度挖掘其中有价值的信息,算法以人工神经网络为代表(Brown *et al.*,2003;Harris *et al.*,2003;Izadi *et al.*,2013).随机森林算法可以处理输入特征变量的多维数据和数据中部分数值缺失的问题,可以自动识别出特征变量在模型中的占比重要性,抗过拟合能力较强.*K*-近邻算法为常用的监督学习方法,其算法思想比较简单,适用于样本数目较大、局部近邻区域类条件概率相同,不需要过多地调节就可以得到不错的性能,正是*K*-近邻的这种性质,使它成为分类问题中的重要方法(殷小舟,2009).人工神经网络的复杂度较高,构建模型复杂,训练时间较长,对数据的预处理要求较高;朴素贝叶斯分类器主要用于高维数据,常用于非常大的数据集.因此,随机森林算法和*K*-近邻算法在众多领域得到应用(洪瑾等,2018;章宝月等,2019).

花岗岩型铀矿床是我国重要的铀矿床类型,而华南地区花岗岩型铀矿是我国重要的铀成矿省,矿床数量占全国花岗岩型铀矿床总数的85.6%(邵飞等,2014).针对华南花岗岩体已经积累了大量的岩石元素分析数据,虽然已有文献利用对岩体进行精确的测年以及对两者地球化学以及年代学的对比,来探讨岩体产铀能力的强弱问题(田泽瑾,2014;Zhang *et al.*,2018),但尚未有研究利用机器学习算法来探索特定花岗岩体的铀成矿潜力,因此本文选取应用广泛的*K*-近邻算法以及可解释性突出的随机森林算法来探索特定花岗岩体的铀成矿潜力.本文在前人研究所得数据的基础上,首先对特征变量进行特征重要性和相关性分析,选取元素特征变量,然后通过机器学习的方法,研究岩体主、微量元素对岩体铀成矿的判别;运用随机森林和*K*-近邻算法分别构建岩体成矿的判别模型,对两种算法进行横向上的调整参数优化对比,选取最佳的机器学习模型.然后,运用建立的模型对诸广山的九峰岩体、

红山岩体以及茶山岩体进行判别,来评价这些岩体的铀成矿潜力,为以后的找矿勘查提供科学依据.

1　地质背景

发育在花岗岩外围地层中以及花岗岩体内的热液型铀矿床(一般不超过 2 km)称为花岗岩型铀矿.花岗岩型铀矿是我国主要的铀矿床类型之一,在我国铀矿分布上具有重大的意义.华南地区花岗岩分布广泛,出露总面积达 20 余万 km^2,是我国重要的铀矿产区(邵飞等,2014).这些花岗岩是新太古代－中生代多旋回岩浆活动的产物,其中以燕山期岩浆活动最为重要.华南产铀花岗岩体,除个别为古老的侵入期次简单的岩体外(晋宁期摩天岭花岗岩体),一般为自加里东至燕山期多期、多阶段侵入的复式岩体,但与铀成矿直接有关的花岗岩体则以较晚期的燕山期为主,次为印支期(邵飞等,2014).

华南地区铀资源又以桃山－诸广山成矿带为主,其中诸广山复式岩体是我国花岗岩型铀矿较多的地方,铀矿资源十分丰富,现在已经发现了很多大型的矿床,如 361 矿床、棉花坑矿床.诸广山复式岩体位于江西西南部、广东北部和湖南东南部三省接壤区域内,位于南岭东西向构造带和万洋山－诸广山南北向构造带的复合部位(Wang *et al.*,2020;Xiao *et al.*,2020),该研究区已经发现了大量铀矿床和许多富铀岩体,如三江口、长江、企岭、白云、江南等岩体(伍皓等,2020).

诸广山岩体是华南地区多期次多阶段的复式花岗岩体,出露面积大约可达 4 000 km^2,它的形成、展布以及铀成矿作用与该区域构造、岩浆活动有着极其密切的联系.诸广山岩体区域主要以花岗岩分布为特征,下古生界、上古生界和中新生界地层极少出露,地层的总体厚度大于 20 500 m.从晚侏罗世各地层都为角度假整合或不整合,反映了诸广山区域在晚侏罗世时期发生过多次构造运动(朱捌,2010).诸广山复式岩体形成于加里东时期,在印支期－燕山期活动频繁,诸广山复式岩体的主体由此构成.诸广山岩体的岩浆活动主要发生在印支－燕山期,在这一时期内主要发生中酸性岩浆活动,岩

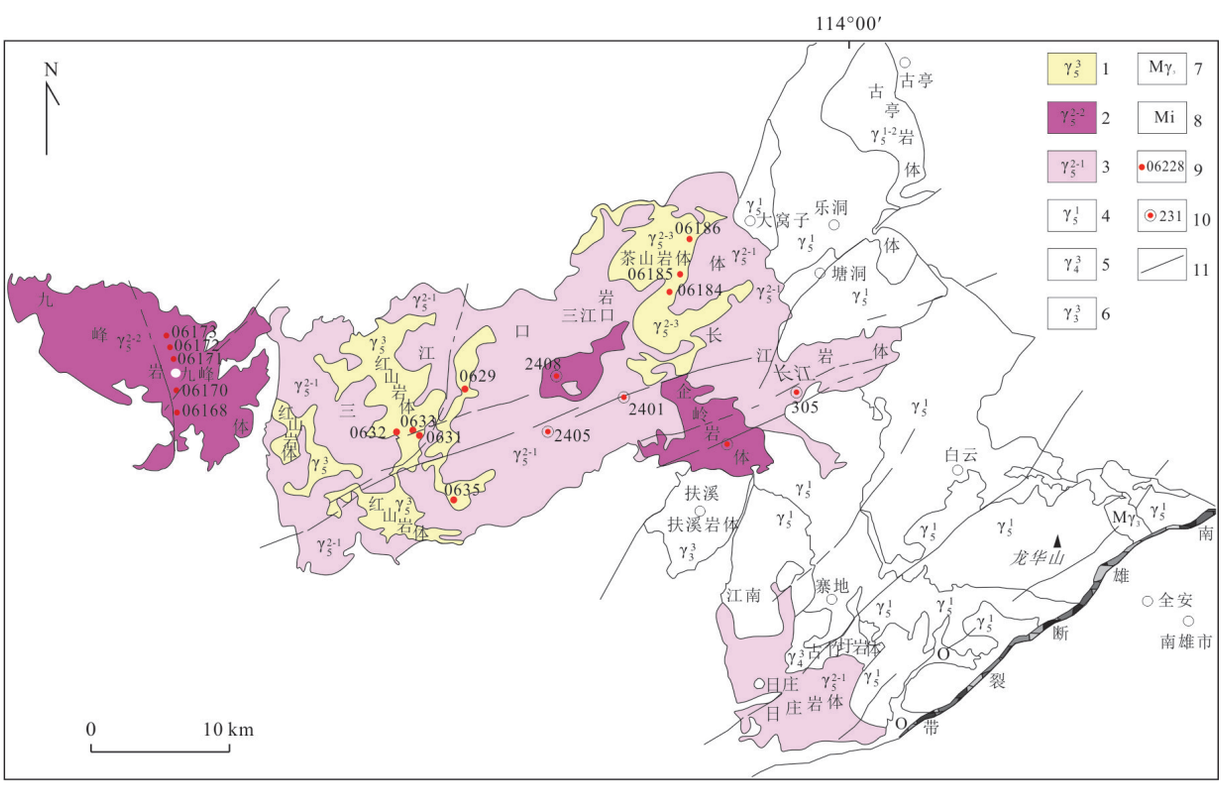

图 1　诸广山岩体分布及采样点位置图

Fig.1　Zhuguangshan pluton distribution and sampling point location map

1.中粒－细粒二云母花岗岩;2.中－中粗粒(斑状)黑云母花岗岩;3.粗粒斑状黑云母花岗岩;4.印支期花岗岩;5.海西期黑云母闪长岩;6.加里东期花岗闪长岩;7.条带状混合花岗岩;8.条带状混合岩;9.取样点及编号;10.铀矿床及编号;11.断裂;图改自朱捌(2010)

性主要为粗、中、细粒黑云母花岗岩和二云母花岗闪长岩.

2 花岗岩地球化学数据集

本次研究从已公开发表的文献中收集了来自华南不同花岗岩岩体的1 724条主、微量元素数据(花岗岩ω(SiO_2)>56%),建立了花岗岩主、微量元素地球化学数据集.具体选取的岩体主要包括:印支期的江南岩体、白云岩体、乐洞岩体、寨地岩体、龙华山岩体、棉土窝岩体、油洞岩体、大窝子岩体、古亭岩体、桃金洞岩体,燕山期的九峰岩体、红山岩体、茶山岩体、三江口岩体、长江岩体、企岭岩体、赤坑岩体、百顺岩体、日庄岩体,以及印支期和燕山期复式岩体,包括棉花坑岩体、龙源坝岩体、粤北贵东复式岩体、桃山复式岩体等.

这些花岗岩地球化学数据使用电子探针(EPMA)和(激光剥蚀—)电感耦合等离子体质谱((LA-) ICP-MS)等测试手段获得.由于数据来自不同的文献,所检测的元素也有所不同,考虑到并非所有的样品都测量了全部的元素含量,同时在主量元素中,Fe有+3价和+2价,由于在测量中很难准确测定,结合我们在收集数据的过程中,FeO数据缺失较多,所以我们选择Fe_2O_3作为研究数据,最终我们选定主量元素分别为SiO_2、TiO_2、Al_2O_3、Fe_2O_3、MnO、MgO、CaO、Na_2O、K_2O、P_2O_5 10种,微量元素分别为Rb、Sr、Y、Zr、Hf、Nb、Ta、Ba、Th、U 10种,共20种元素含量作为我们所研究的特征变量.据此对所收集的数据进行筛选,共筛选出1 711条完整数据用来作为建立机器学习模型的数据集.将收集的1 711条数据,根据研究者的采样点离铀矿床的远近以及采样岩体中是否含有铀矿床划分为866条含矿样本数据和845条不含矿样本数据,含矿的标为1,不含矿的标为0.另外单独收集九峰岩体、红山岩体和茶山岩体的数据(共13条),利用训练模型对九峰、红山和茶山岩体铀成矿潜力进行评价.本研究未对数据单位做出转换改变,主量元素单位为(%),微量元素单位为(10^{-6}).

通过对数据集建立箱线槽口图,我们从图2和图3可以看出,主量元素中SiO_2数据值的分布比较集中,含量较高,MnO含量较低,Al_2O_3、MnO和Na_2O数据值分布区间比较分散,数据之间相差比较大(图2).微量元素中,每种元素的数据分布比较均匀,其中Rb和Ba含量较高,Hf含量较低(图3).

3 研究方法

3.1 特征变量分析

理论上不同元素的性质不同,在构建分类或回归模型时,特征变量之间有着不同的作用,不同的特征变量对模型的预测准确率有着不同的重要性,有些特征之间可能存在着共性或差异性.为了探明

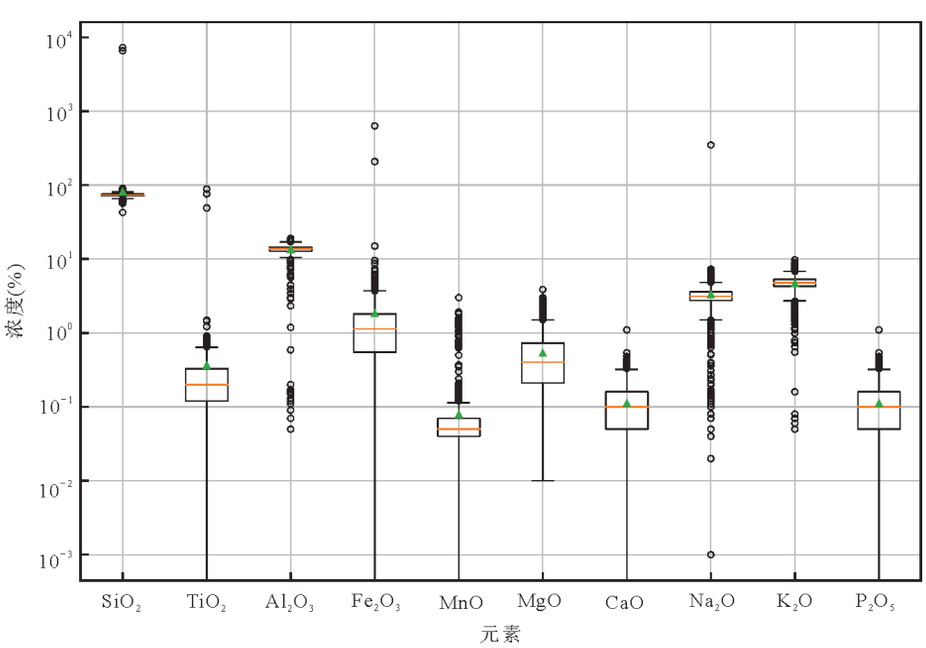

图2 花岗岩主量元素槽口箱线图

Fig.2 Box diagram of granite major element slot

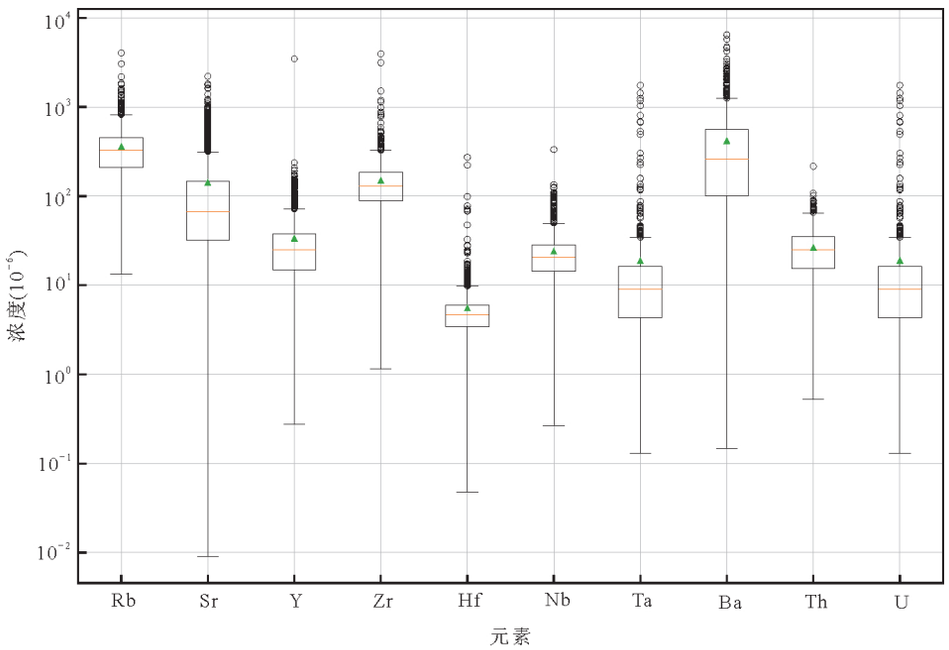

图3　花岗岩微量元素槽口箱线图

Fig.3　Box diagram of granite trace element slot

不同特征变量与构建模型之间的相关性,我们对上述的数据采取特征工程研究,以达到对重要特征变量的筛选.特征变量的选择对模型的构建十分重要,特征变量过多或者过少都会对模型的泛化能力产生影响,特征变量过多则会增加学习算法运算的时间和运行内存的需要,而且还很可能导致模型过拟合;反之,特征变量如若过少,模型则会欠拟合,导致模型泛化能力降低(Hong *et al.*, 2021; Sun *et al.*, 2021).本文对特征变量筛选的方法为随机森林算法的特征重要性度量和皮尔逊相关性分析.

3.1.1　特征重要性度量　特征重要性度量是一种十分重要的数据筛选方式,是模型建立识别的一个很关键的问题.在用随机森林算法建立分类或回归模型时(Strobl *et al.*, 2008),需要从众多的特征变量中对变量在分类或回归模型中占有的重要性识别,这就是特征重要性度量.在进行特征重要性度量时,通常利用袋外(out of bag, OOB)数据(Chehreh Chelgani *et al.*, 2016; Wang *et al.*, 2016),在袋外数据中可以先对某一个特征变量加入噪音,然后对比加入噪音前后模型的准确率,如果准确率大幅度下降,则可表明该特征变量的重要性较高.通过这个方法则可以计算出模型中所有特征变量的重要程度,如果特征变量的重要性值越高,则表明这个特征变量对模型的精确性能的影响越大,占的比重越大.对于每个特征变量来说,重要性值的范围都是在[0,1],0表示“特征变量在模型构建中根本没有用到”,1表示“特征变量在模型构建中完美预测目标值”,特征变量重要性求和始终为1.

3.1.2　皮尔逊相关性分析　随机森林模型对于各变量之间的相关性具有较高的敏感性,预测变量的强相关性会使得预测有偏差(Nicodemus and Malley, 2009; Altmann *et al.*, 2010).为了避免偏差的产生,对特征变量进行皮尔逊(Pearson)相关性分析.皮尔逊相关系数是用来计算两个特征变量之间的数值特征的,两个变量之间的协方差和标准差的商定义为两个特征变量之间的Pearson相关系数:

$$\rho_{X,Y}=\frac{COV(X,Y)}{\sigma_X\sigma_Y}=\frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X\sigma_Y},\quad(1)$$

式(1)中定义了总体相关系数,常用希腊小写字母作为代表符号.估算样本的协方差和标准差,可得到Pearson相关系数r:

$$r=\frac{\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i-\bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i-\bar{Y})^2}}.\quad(2)$$

皮尔逊相关系数反映了两个特征变量之间相关性的大小,取值范围为[-1,1].系数的值为1意味着所有的数据点都很好地落在一条直线上,表明两个变量之间正相关性极强,系数的值为-1意味着两个变量之间是呈负相关的,相关性极强会影响

模型的泛化能力.系数的值为0意味着两个变量之间没有线性关系,两个变量之间是独立存在的.一般而言,Pearson值大于0.6或小于－0.6时,都表明两者具有较强的相关性(Chehreh Chelgani *et al.*, 2016;Matin and Chelgani,2016).

3.2 随机森林算法

随机森林由Breiman(2001)提出的一种分类和回归算法的集成学习算法,它通过自助法(bootstrap)重采样技术,在原始训练样本集合N中有放回地重复随机选择N个样本,形成新的原始训练样本集合训练决策树,再经由上述过程产生的m棵决策树组成随机森林,新数据的划分结果按分类树投票数量产生的分值而定.其实质是对决策树算法的一个改良,将数个决策树整合在一块,每棵树的构建依赖于单独提取的样本.单棵树的分类能力可能很小,但在随机产生大量的决策树后,一个测试样本可以通过每一棵树的分类结果经统计后选择最可能的分类.建立该模型的过程可以分为以下几个步骤:(1)首先把数据集按照7:3的比例划分为训练集数据和测试集数据,从训练集中有放回地随机选取训练集2/3的数据作为样本集,剩下的1/3的数据则就是袋外数据;(2)在建立模型中设置要选取的最大特征参数和构建的决策树的个数,建立模型;(3)选取对模型性能最佳的参数,然后每棵树对测试集数据进行测试,评估结果便是预测结果.

随机森林比决策树对离群值和不平衡的数据集具有强大的性能,可扩展且能够处理数据集中的非线性趋势,它不需要对特征变量的数据特征分布和范围进行限制,所以随机森林不需要对特征进行缩放或修改,它可以适用于任何数据源.在随机森林中树的随机化的方法有两种:一种是通过选择用于构造树的棵树;一种是通过选择每次划分选取最大特征的个数.随机森林是以N个决策树为基本分类器,进行集成学习后得到的一个组合分类器.当输入待分类样本时,随机森林输出的分类结果由每个决策树的分类结果简单投票决定(董师师和黄哲学,2013).此外,随机森林还能自动计算出特征变量对于建立的模型的相对重要性(Breiman,2004).

3.3 *K*-近邻算法

K-近邻算法(K-nerest neighbor,KNN)是监督学习的一种,常用于分类和回归问题,K-近邻方法的核心思想是:在给定的已知标签的样本集上寻找与待测识别样本相距最近的K个样本,这些样本中的标签众数的那一个标签,即为该待测样本的分类标签(殷小舟,2009).本研究中,数据点间的远近定义为欧氏距离,即以X、y表征任意两数据点的位置,其在n维空间上的公式如下:

$$d(X,y)=\sqrt{\sum_{i=1}^{n}(X_i-y_i)^2}. \tag{3}$$

KNN的发现作为一种在不同领域使用的流行方法,被认为是大数据中的一种具有挑战性的方法,它基于在属性空间中发现K密切邻近的样本,KNN算法的思想相对于其他的机器学习方法比较简单,但是也具有很好的预测性能,并且还不需要对数据进行特别的假设.KNN算法的计算过程主要分为以下4个步骤:(1)计算待分类的数据点和已经知道类别的数据点之间的距离,并且按照距离的远近排序;(2)选取要分类点与已知类别点距离较小的N个数据点;(3)确定N个数据点所出现在类别中的频数;(4)前N个数据点出现次数最多的类别作为待分类点的预测类别.在K-近邻算法中,我们最需要注意的关键就是参数K值的选择,K值的选择对最终模型预测的结果会产生直接的影响,如果K值选择过小,就意味着整体模型变得复杂,容易产生过拟合(Song *et al.*,2007).如果选择的K值较大,就相当于用较大领域中的训练实例进行预测,就意味着模型过于简单,学习的近似误差会增大,是不可取的.因此,K值的选择对于模型的泛化能力十分重要.

3.4 模型评价

本研究采用测试集上的模型精确率(Precision)、召回率(Recall)、受试者工作特征(receiver operating characteristic curve,ROC)曲线来评价模型的分类效果(郑泽宇,2019),精确率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;ROC曲线是反映敏感性和特异性连续变量的综合指标,是用构图法揭示敏感性和特异性的相互关系,它通过将连续变量设定出多个不同的临界值,从而计算出一系列敏感性和特异性,再以敏感性为纵坐标、特异性为横坐标绘制成曲线,曲线下面积越大,诊断准确性越高.在ROC曲线上,最靠近坐标图左上方的点为敏感性和特异性均较高的临界值.AUC值表示的是ROC曲线下方X轴与Y轴所形成的面积,AUC值越大,表明该模型的泛化能力越好;反之,泛

化性能则越差.

本文采用的机器学习算法基于Python3.8和第三方模块scikit-learn编码实现,并使用jupyternotebook运行.

4　实验结果与讨论

4.1　特征变量选择

本文将上述的20种主、微量元素作为特征变量,并对特征变量进行相关的选择分析.首先对特征变量进行重要性度量,采用收集的1 711组数据生成随机森林模型,在生成模型的同时,已经对特征变量利用袋外数据进行特征重要性度量,特征变量重要性度量结果表明U、Rb、Ba、K_2O等元素对分类模型具有比较高的贡献值,P_2O_5对模型的性能产生的贡献值最低(图4),因此剔除P_2O_5的数据.

在对20种元素进行Pearson相关性分析,从结果可以看出,Rb和Ta具有较强的相关性(0.71),Ba和Sr具有较强的相关性(0.85),因此我们认为Rb和Ta、Ba和Sr元素之间两者的数据重叠区域是较大的,我们只需要采用具有强相关性元素之中的一种即可,对于强相关性的特征的选取,我们结合图4的特征变量重要性度量结果来做出选择,从图4中可以看出Rb的重要性大于Ta的重要性,Ba的重要性大于Sr的重要性,所以我们剔除Ta、Sr特征变量,保留Rb、Ba作为模型的特征变量.最后保留下来的特征变量元素有SiO_2、TiO_2、Al_2O_3、Fe_2O_3、MnO、MgO、CaO、Na_2O、K_2O、Rb、Y、Zr、Hf、Nb、Ba、Th和U.

4.2　随机森林算法

本文采用全部数据集的70%作为训练集,用来生成随机森林模型,30%的数据作为测试集,用来作为验证模型分类准确率的数据.对于随机森林模型,选择Gini指数和CART算法构建随机森林的决策树.

4.2.1　参数优化　在随机森林模型的建立中,通常特征数max_feature和决策数的数量n_eatimators会对预测结果的精度产生影响.本实验先对max_feature参数在1～6的范围内,对数据集进行训练,然后对模型的分类性能进行检测.图5表明最大特征个数为3的时候,该随机森林模型的分类准确率最佳.

在最大特征个数为3时,在决策树的数量为100～1 000的范围内,对随机森林模型进行分类性能测试.图6表明构造树的个数为400时,该随机森林模型的分类准确率最佳.因此,该随机森林模型在此参数下(max_feature=3,n_estimators=400)分类性能达到最佳,此时在测试集上的分类精确率达到了93%.

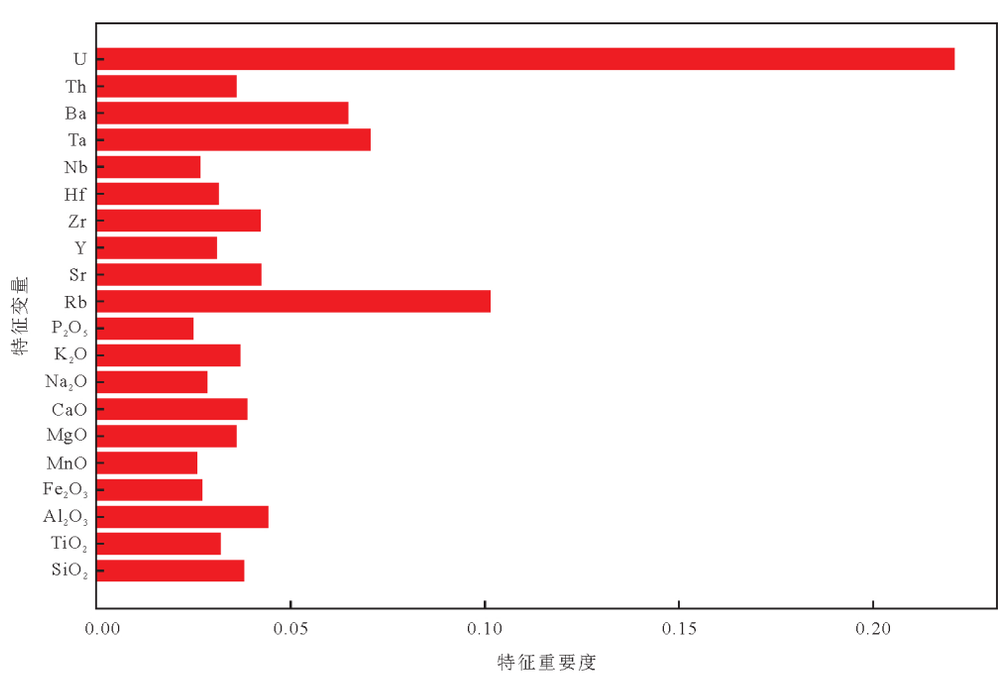

图4　基于随机森林模型的重要性度量

Fig.4　Importance measurement based on random forest model

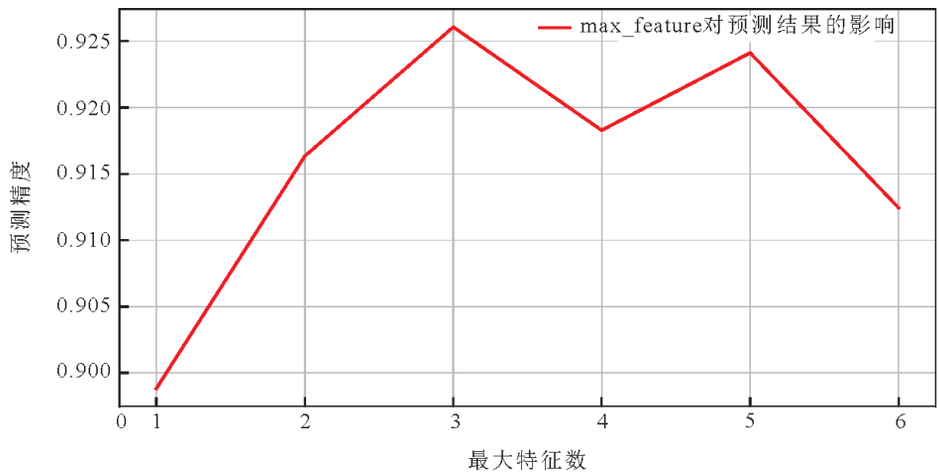

图 5 随机森林模型 max_feature 选取图

Fig.5 Max_feature selection diagram of random forest model

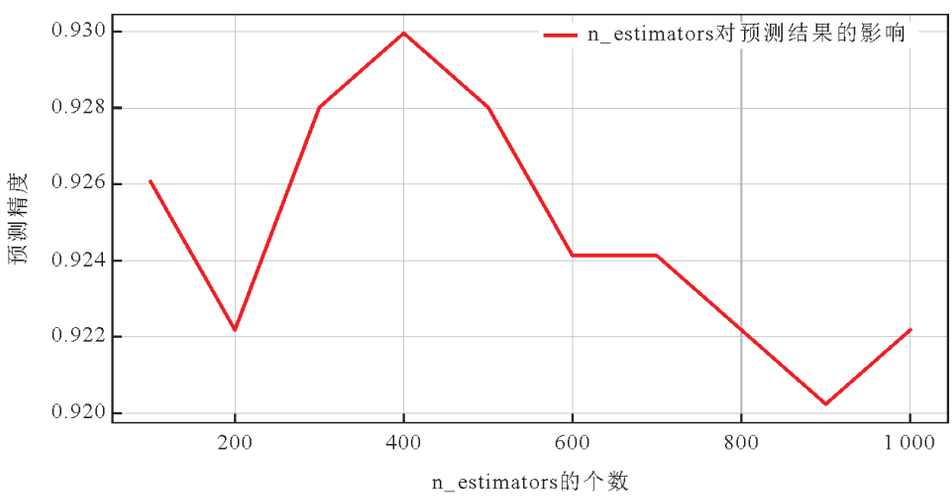

图 6 随机森林模型 n_estimators 选取图

Fig.6 Random forest model n_estimators selection diagram

4.2.2 随机森林模型测试集分类效果 采用上述性能最佳的最大特征数为 3,构造树个数为 400 的随机森林模型,对数据集分类所得到混淆矩阵结果(图 7).混淆矩阵是用来评价分类精度的一种标准格式,混淆矩阵的每一列代表了分类模型中的一种预测类别,每一列的总数表示预测为该类别的数据的数目;每一行代表了数据的真实归属类别,每一行的数据总数表示该类别的数据实例的数目.每一列中的数值表示真实数据被预测为该类的数目,如图 7 所示,第一行表示有 222 个含矿数据预测正确,21 个含矿数据预测错误;同理,第一行第二列的 21 表示有 21 个实际归属为含矿的实例被错误预测为第二类不含矿的实例.由图 6 知该模型此时分类准确率达到 93%.

4.3 *K*-近邻算法

在进行 *K*-近邻算法上,我们采用与随机森林算法一致的数据集(即进行过特征重要性度量和 Pearson 相关性分析筛选后保留的数据).

4.3.1 参数选择 在 *K*-近邻算法中,近邻数 *K* 的选择非常重要.本实验选取近邻数在 1~10 的范围内,对数据集在 *K*-近邻算法下进行分类正确率预测.结果如图 8 所示,在近邻数为 7 的时候测试集分类效果达到最佳,此时在测试集上的分类准确率达到了 81%.

4.3.2 *K*-近邻模型测试集分类效果 图 9 为在近邻数为 7 时的情况下对数据集进行 *K*-近邻分类所得的模型的测试结果.第一行表示在一共 242 个样本中,199 个预测正确,43 个样本预测错误,即第一行 199 表示有 199 个实际为含矿的样本预测正确,43 表示

有43个实际为含矿的样本被错误预测为不含矿.同理,第二行表示在272个样本中,218个不含矿样本被正确预测,54个不含矿样本预测错误,被预测为含矿样本.即第二行54表示有54个实际为不含矿的样本被预测为含矿,218表示有218个实际为不含矿的样本预测正确.由图8可知该模型此时的预测准确率为81%.

4.4　模型评价

召回率曲线适用于对二分类变量的模型评价,即为所有正例中被正确预测的比例.ROC曲线下面积(area under roc curve,*AUC*)作为一种单一的量化指标,它能很好地反映分类模型的分类效果,*AUC*的取值范围为0到1,*AUC*值越接近1,则该模型的分类性能越好,反之,*AUC*的值越小,则该模型的性能越差.

随机森林模型和*K*-近邻模型的精确率-召回率曲线和ROC曲线如图10和图11所示.综上可知,随机森林模型的分类精度为0.93,*AUC*值为0.96,而*K*-近邻模型的分类精确度为0.81,*AUC*值为0.89,随机森林模型在此数据集上的分类性能明显强于*K*-近邻分类模型.而召回率表示的是样本中的某类样本有多少被正确预测,召回率越高,特异性越小,也就是召回率曲线越靠近右上越好.所以,由图10和图11可知,随机森林模型在此数据集的分类性能优于*K*-近邻模型.

4.5　预测模型不确定性讨论

在最优模型下,随机森林模型和*K*-近邻模型都存在预测不正确的岩体,预测错误的数据主要存在于棉土窝岩体和乐洞岩体,两者均属于印支期花岗岩岩体.这些岩体具有较高含量的Th和Zr,导致岩浆中的U大量进入锆石等难溶矿物中,限制了花岗岩的成矿潜力.另外,数据点标签(含矿或非矿)是根据岩体内是否有已发现铀矿进行判别,标签的正确与否也影响最终的分类结果.一方面岩体可能仅仅是矿点或矿化,另一方面采样点的代表性也有偏差,具有人为因素影响,从而对数据的判别划分不当,造成了模型预测结果的不确定性.

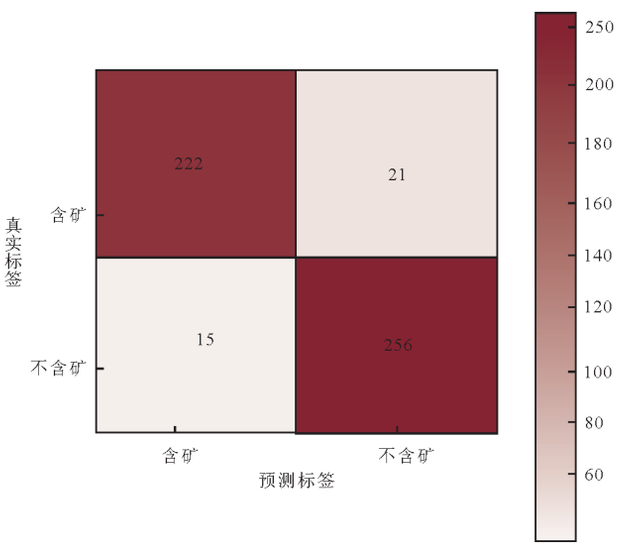

图7　随机森林分类模型混淆矩阵图

Fig.7　Random forest classification model confusion matrix diagram

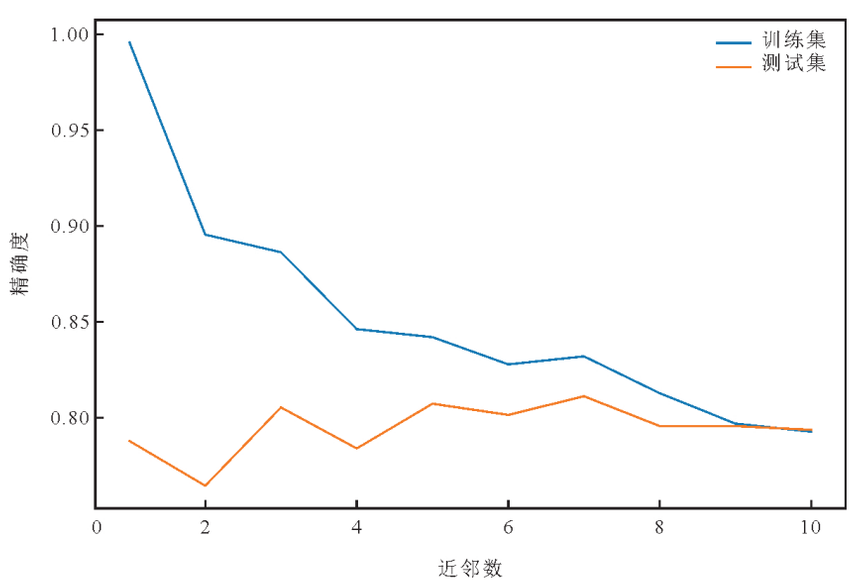

图8　KNN近邻数选取图

Fig.8　*K* value selection graph of KNN algorithm

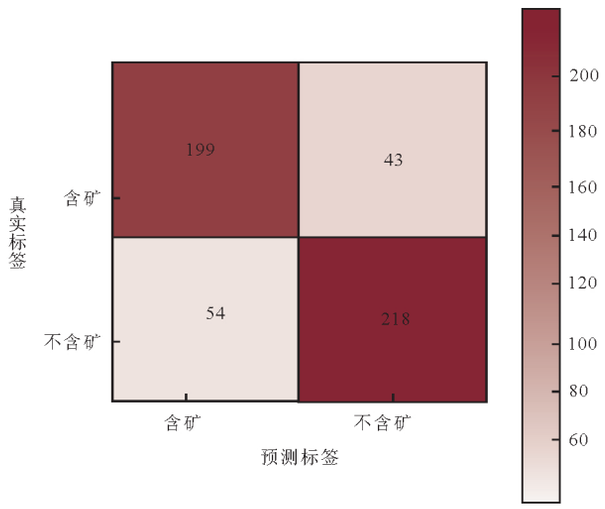

图 9 KNN 分类模型混淆矩阵图

Fig.9 KNN classification model confusion matrix diagram

另外数据精度和数据量也会造成预测结果的不确定性.数据集是所有机器学习系统的关键成分,数据量过少,数据收集不全面,特征变量不完整都可能造成预测结果的不确定性.同时,数据的精度会受到分析方法以及检出限高低的影响,如 LA-ICP-MS 数据的分析精度比 EPMA 要高,在数据集进行前处理的过程中会造成噪声,进而对预测结果的准确性也会产生影响.

5 成矿潜力评价分析

用上述建立的分类性能较好的随机森林模型对研究区九峰岩体、红山岩体以及茶山岩体采样点数据进行预测(采样点如图 1 所示),结果如表 1 所示.

九峰岩体采样点,不含矿概率分别为 64%、86%、89%、93% 和 84%,九峰岩体位于诸广山主成矿区的北东向,远离主成矿区,九峰岩体 5 个采样点中 4 个不含矿率都超过了 84%,认为该岩体的含矿率较低,而且已有前人研究认为九峰岩体碱度率低,不利于铀矿的富集,而现在在九峰岩体还未发现铀矿,所以认为九峰岩体为不产铀岩体(田泽瑾,2014;Zhang *et al.*,2017;张丽,2017).红山岩体采样点,含矿概率分别为 83%、89%、90%、90% 和 91%,红山岩体 5 个采样点中,每个采样点含矿概率都超过了 80%,表明该区的含矿概率较大.茶山岩体采样点含矿概率分别为 96%、97% 和 99%,茶山岩体的 3 个采样点中,每个采样点的含矿概率都超过了 95%,所以该区的含矿可能性极大(兰鸿锋等,2020).虽然当前在红山和茶山岩体中还未发现铀矿点,但未来应当作为找矿勘查的重点.

6 结论

本文基于大数据和机器学习的思维,利用已发表文献收集了华南花岗岩的地球化学数据,并对数据进行处理,利用随机森林方法和 K-近邻方法分别构建分类模型,并对比两种分类模型的分类精确度,选出最优模型,最终对诸广山地区九峰、红山和

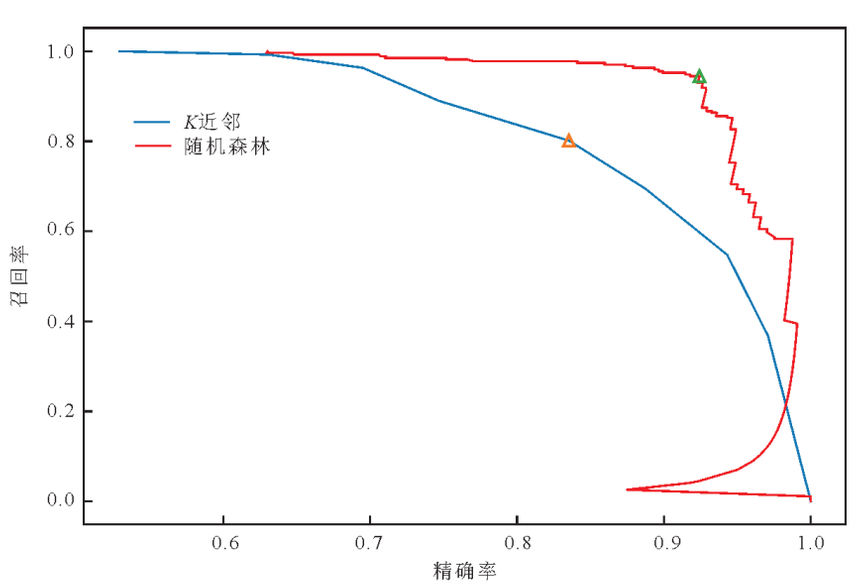

图 10 准确率—召回率曲线

Fig.10 Accuracy-recall curve

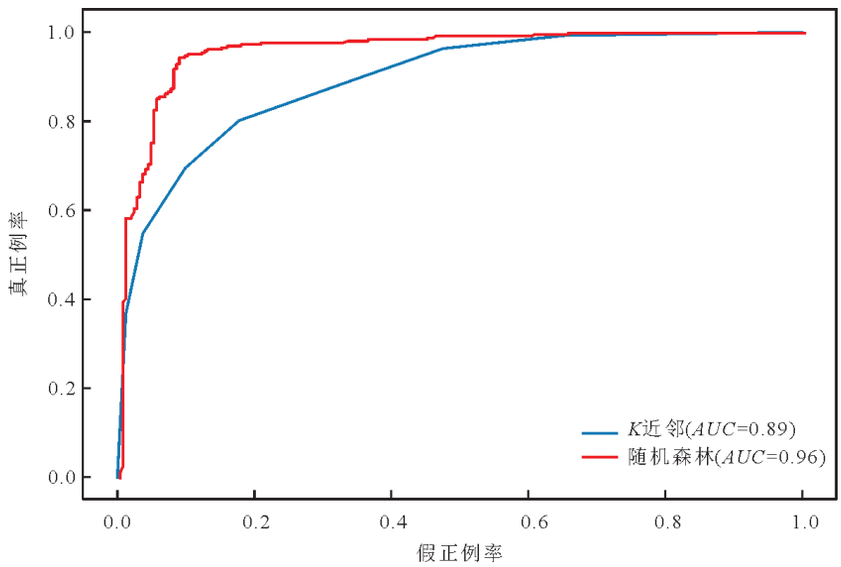

图 11　ROC 曲线

Fig.11　ROC curve

表 1　九峰、红山和茶山岩体随机森林模型预测结果

Table 1　Jiufeng, Hongshan and Chashan plutons random forest model prediction results

	编号	不含矿概率(%)	含矿概率(%)	预测结果
九峰岩体	06168	64	36	0
	06170	86	14	0
	06171	89	11	0
	06172	93	7	0
	06173	84	16	0
	0629	17	83	1
红山岩体	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

茶山岩体进行成矿潜力评价.

(1)本文以前人已发表文献中的花岗岩地球化学数据,建立以随机森林算法和K-近邻算法的机器学习分类模型,通过对两种算法的研究表明随机森林具有更好的性能,表现出更好的预测效果.证明这种模型在现有的数据种是可行的.

(2)通过对潜力区进行预测,表明九峰岩体含矿的概率较低,红山岩体的预测结果中含矿概率均为中高值,存在较高的成矿潜力,可作为矿产勘探潜力区.茶山岩体3个采样点的预测结果都为高含矿率,大概率存在未发现的矿体,可以优先作为找矿勘查的方向.

(3)基于华南花岗岩型铀矿的主微量地球化学大数据和随机森林算法构建的分类器泛化能力较强,是一种新的有效的矿床勘查预测分类模型,将来我们将进一步收集补充新的地球化学数据,建立更加完善的大数据集,并采用其他机器学习方法来建立分类模型,为地质人员提供一种便捷完善的矿产勘查预测工具.

References

Altmann, A., Toloşi, L., Sander, O., et al., 2010. Permutation Importance: A Corrected Feature Importance Measure. *Bioinformatics*, 26(10): 1340−1347. https://doi. org/10.1093/bioinformatics/btq134

Breiman, L., 2001. Random Forest. *Machine Learning*, 45: 5−32.

Breiman, L., 2004. RFtools-for Predicting and Understanding Data. Technical Report, Berkeley University, Berkeley, USA.

Brown, W. M., Groves, D. I., Gedeon, T. D., 2003. An Artificial Neural Network Method for Mineral Prospectivity Mapping: A Comparison with Fuzzy Logic and Bayesian Probability Methods. In: Sandham, W. A., Leggett, M., eds., Geophysical Applications of Artificial Neural Networks and Fuzzy Logic. Modern Approaches in Geophysics, 21. Springer, Dordrecht. https://doi. org/10.1007/978-94-017-0271-3_12

Chehreh Chelgani, S., Matin, S.S., Hower, J.C., 2016. Explaining Relationships between Coke Quality Index and Coal Properties by Random Forest Method. *Fuel*, 182: 754—760. https://doi.org/10.1016/j.fuel.2016.06.034

Chen, Y.L., Zhou, B., Li, X.B., 2012. Mineral Target Prediction Based on Boltzmann Machines. *Progress in Geophysics*, 27(1): 179—185(in Chinese with English abstract).

Dong, S.S., Huang, Z.X., 2013. A Brief Theoretical Overview of Random Forests. *Journal of Integration Technology*, 2(1): 1—7(in Chinese with English abstract).

Hao, H.Z., Gu, Q., Hu, X.M., 2021. Research Advances and Prospective in Mineral Intelligent Identification Based on Machine Learning. *Earth Science*, 46(9): 3091—3106(in Chinese with English abstract).

Harris, D., Zurcher, L., Stanley, M., et al., 2003. A Comparative Analysis of Favorability Mappings by Weights of Evidence, Probabilistic Neural Networks, Discriminant Analysis, and Logistic Regression. *Natural Resources Research*, 12(4): 241—255. https://doi. org/10.1023/b:narr.0000007804.27450.e8

Hong, J., Gan, C.S., Liu, J., 2018. Preliminary Study on the Relationship between Trace and Major Elements in Rocks Based on Machine Learning: A Case Study of Zr in OIB. *Chinese Journal of Geology*, 53(4): 1285—1299(in Chinese with English abstract).

Hong, S., Zuo, R.G., Huang, X.W., et al., 2021. Distinguishing IOCG and IOA Deposits via Random Forest Algorithm Based on Magnetite Composition. *Journal of Geochemical Exploration*, 230: 106859. https://doi.org/10.1016/j.gexplo.2021.106859

Izadi, H., Sadri, J., Mehran, N.A., 2013. Intelligent Mineral Identification Using Clustering and Artificial Neural Networks Techniques. 2013 First Iranian Conference on Pattern Recognition and Image Analysis (PRIA), IEEE. Birjand, Iran. https://doi. org/10.1109/pria.2013.6528426

Lan, H.F., Wang, H.Z., Ling, H.F., et al., 2020. Petrogenesis of the Chashan Granite in the Northern Guangdong Province and Its Implication for the Metallogenic Potential of Tungsten and Uranium Mineralization. *Acta Geologica Sinica*, 94(4): 1143—1165(in Chinese with English abstract).

Matin, S.S., Chelgani, S.C., 2016. Estimation of Coal Gross Calorific Value Based on Various Analyses by Random Forest Method. *Fuel*, 177: 274—278. https://doi. org/10.1016/j.fuel.2016.03.031

Nicodemus, K.K., Malley, J.D., 2009. Predictor Correlation Impacts Machine Learning Algorithms: Implications for Genomic Studies. *Bioinformatics*, 25(15): 1884—1890. https://doi.org/10.1093/bioinformatics/btp331

Rodriguez-Galiano, V., Sanchez-Castillo, M., Chica-Olmo, M., et al., 2015. Machine Learning Predictive Models for Mineral Prospectivity: An Evaluation of Neural Networks, Random Forest, Regression Trees and Support Vector Machines. *Ore Geology Reviews*, 71: 804—818. https://doi.org/10.1016/j.oregeorev.2015.01.001

Shao, F., Xu, J.J., Shao, S., et al., 2014. Geological Characteristics and Mineralization of the Granite-Type Uranium Deposits in South China. *Resources Survey and Environment*, 35(3): 211—217(in Chinese with English abstract).

Song, Y., Huang, J., Zhou, D., et al., 2007. IKNN: Informative *K*-Nearest Neighbor Pattern Classification. In: Kok, J.N., Koronacki, J., Lopez de Mantaras, R., eds., Knowledge Discovery in Databases: PKDD 2007, PKDD 2007. Lecture Notes in Computer Science, 4702. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-74976-9_25

Strobl, C., Boulesteix, A.L., Kneib, T., et al., 2008. Conditional Variable Importance for Random Forests. *BMC Bioinformatics*, 9: 307. https://doi. org/10.1186/1471-2105-9-307

Sun, Z.Y., Liu, H.Y., Ju, H.Y., et al., 2021. Assessment of Importance-Based Machine Learning Feature Selection Methods for Aggregate Size Distribution Measurement in a 3D Binocular Vision System. *Construction and Building Materials*, 306: 124894. https://doi. org/10.1016/j.conbuildmat.2021.124894

Tian, Z.J., 2014. Comparative Study on Chronology, Geochemistry and Mineralogical Characteristics of Uranium-Producing and Uranium-Non-Producing Granites in Zhiguang Mountain (Dissertation). China University of Geosciences, Beijing(in Chinese with English abstract).

Vincenzi, S., Zucchetta, M., Franzoi, P., et al., 2011. Application of a Random Forest Algorithm to Predict Spatial Distribution of the Potential Yield of *Ruditapes Philippinarum* in the Venice Lagoon, Italy. *Ecological Modelling*, 222(8): 1471—1478. https://doi.org/10.1016/j.ecolmodel.2011.02.007

Wang, H.Z., Yang, F., Luo, Z.Y., 2016. An Experimental Study of the Intrinsic Stability of Random Forest Variable Importance Measures. *BMC Bioinformatics*, 17: 60. https://doi.org/10.1186/s12859-016-0900-5

Wang, K.X., Sun, T., Yu, J.H., et al., 2020. Provenances of the Ediacaran Sedimentary Rocks in the Zhuguangshan Area and Their Implications for Granitoid-Related Uranium Mineralization in South China. *Ore Geology Reviews*, 124: 103588. https://doi. org/10.1016/j. oregeorev.2020.103588

Wu, H., Xia, Y., Zhou, K.K., et al., 2020. Highly Fractionated Granite Magmas maybe the Main Uranium Source of Granite-Type Uranium Deposits in South China: Evidence from the Uranium Content of Zircon in Southern Zhuguangshan Granitic Composite. *Acta Petrologica Sinica*, 36(2):589－600 (in Chinese with English abstract).

Xiao, Z.H., Xiong, S.B., Li, C.H., et al., 2020. Types of Uranium Deposits in Central Zhuguang Mountains in Hunan Province, South China and Their Metallogenic Regularity and Prospecting Directions. *China Geology*, 3(3):411－424.https://doi.org/10.31035/cg2020040

Yin, X.Z., 2009. An Ameliorated SVM Classifying Algorithm Combined with KNN. *Journal of Image and Graphics*, 14(11):2299－2303(in Chinese with English abstract).

Youn, E., Jeong, M.K., 2009. Class Dependent Feature Scaling Method Using Naive Bayes Classifier for Text Datamining. *Pattern Recognition Letters*, 30(5): 477－485. https://doi.org/10.1016/j.patrec.2008.11.013

Zhang, B.Y., Sun, J.K., Luo, X., et al., 2019. Data Analysis of Major and Trace Element of Gabbro Clinopyroxene from Different Tectonic Setting. *Earth Science Frontiers*, 26(4):33－44(in Chinese with English abstract).

Zhang, L., 2017. Mineralogical Characteristics of Representative Yanshanian Granites in Southern Zhuguang and Their Implications for Petrogenesis and Metallogenic Potential (Dissertation). Nanjing University, Nanjing(in Chinese with English abstract).

Zhang, L., Chen, Z.Y., Li, S.R., et al., 2017. Isotope Geochronology, Geochemistry, and Mineral Chemistry of the U-Bearing and Barren Granites from the Zhuguangshan Complex, South China: Implications for Petrogenesis and Uranium Mineralization. *Ore Geology Reviews*, 91: 1040－1065. https://doi. org/10.1016/j. oregeorev.2017.07.017

Zhang, L., Chen, Z.Y., Li, X.F., et al., 2018. Zircon U-Pb Geochronology and Geochemistry of Granites in the Zhuguangshan Complex, South China: Implications for Uranium Mineralization. *Lithos*, 308－309:19－33.https://doi.org/10.1016/j.lithos.2018.02.029

Zhang, Q., Zhou, Y.Z., 2017. Big Data will Lead to a Profound Revolution in the Field of Geological Science. *Chinese Journal of Geology*, 52(3): 637－648(in Chinese with English abstract).

Zheng, Z.Y., 2019. Comparison of Several Machine Learning Methods for Identification of Multiple Geochemical Anomalies in Helong Area, Jilin Province (Dissertation). Jilin University, Changchun (in Chinese with English abstract).

Zhou, Y.Z., Chen, S., Zhang, Q., et al., 2018a. Advances and Prospects of Big Data and Mathematical Geoscience. *Acta Petrologica Sinica*, 34(2): 255－263(in Chinese with English abstract).

Zhou, Y.Z., Wang, J., Zuo, R.G., et al., 2018b. Machine Learning, Deep Learning and Python Language in Field of Geology. *Acta Petrologica Sinica*, 34(11): 3173－3178(in Chinese with English abstract).

Zhou, Y.Z., Li, P.X., Wang, S.G., et al., 2017. Research Progress on Big Data and Intelligent Modelling of Mineral Deposits. *Bulletin of Mineralogy, Petrology and Geochemistry*, 36(2):327－331, 344 (in Chinese with English abstract).

Zhou, Y.Z., Zuo, R.G., Liu, G., et al., 2021. The Great-Leap-Forward Development of Mathematical Geoscience during 2010—2019: Big Data and Artificial Intelligence Algorithm are Changing Mathematical Geoscience. *Bulletin of Mineralogy, Petrology and Geochemistry*, 40(3):556－573, 777(in Chinese with English abstract).

Zhu, B., 2010. Research on Mantle Fluid and Uranium Mineralization (Dissertation). Chengdu University of Technology, Chengdu(in Chinese with English abstract).

Zuo, R.G., Peng, Y., Li, T., et al., 2021. Challenges of Geological Prospecting Big Data Mining and Integration Using Deep Learning Algorithms. *Earth Science*, 46(1): 350－358(in Chinese with English abstract).

附中文参考文献

陈永良,周斌,李学斌,2012.基于Boltzmann机的矿产靶区预测.地球物理学进展,27(1):179－185.

董师师,黄哲学,2013.随机森林理论浅析.集成技术,2(1):1－7.

郝慧珍,顾庆,胡修棉,2021.基于机器学习的矿物智能识别方法研究进展与展望.地球科学,46(9):3091－3106.

洪瑾,甘成势,刘洁,2018.基于机器学习的岩石微量元素与主量元素关系初探:以洋岛玄武岩中锆元素为例.地质科学,53(4):1285－1299.

兰鸿锋,王洪作,凌洪飞,等,2020.粤北荼山岩体岩石成因与

铀、钨成矿潜力探讨.地质学报,94(4):1143－1165.

邵飞,许健俊,邵上,等,2014.华南花岗岩型铀矿地质特征及成矿作用.资源调查与环境,35(3):211－217.

田泽瑾,2014.诸广山产铀与不产铀花岗岩的年代学,地球化学及矿物学特征对比研究(硕士学位论文).北京:中国地质大学.

伍皓,夏彧,周恳恳,等,2020.高分异花岗岩浆可能是华南花岗岩型铀矿床主要铀源:来自诸广山南体花岗岩锆石铀含量的证据.岩石学报,36(2):589－600.

殷小舟,2009.一种改进的结合K近邻法的SVM分类算法.中国图象图形学报,14(11):2299－2303.

郑泽宇,2019.吉林省和龙地区多元地球化学异常识别的几种机器学习方法比较(硕士学位论文).长春:吉林大学.

章宝月,孙建鹃,罗熊,等,2019.三类构造背景辉长岩单斜辉石主量元素和微量元素的数据分析研究.地学前缘,26(4):33－44.

张丽,2017.诸广南部燕山期代表性花岗岩的矿物学特征及对岩石成因和成矿潜力的指示意义(硕士学位论文).南京:南京大学.

张旗,周永章,2017.大数据正在引发地球科学领域一场深刻的革命:《地质科学》2017年大数据专题代序.地质科学,52(3):637－648.

周永章,陈烁,张旗,等,2018a.大数据与数学地球科学研究进展:大数据与数学地球科学专题代序.岩石学报,34(2):255－263.

周永章,王俊,左仁广,等,2018b.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,黎培兴,王树功,等,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,左仁广,刘刚,等,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学.矿物岩石地球化学通报,40(3):556－573,777.

朱捌,2010.地幔流体与铀成矿作用研究(博士学位论文).成都:成都理工大学.

左仁广,彭勇,李童,等,2021.基于深度学习的地质找矿大数据挖掘与集成的挑战.地球科学,46(1):350－358.