DOI：10.12090/j.issn.1006-6616.2021.27.03.031
文章编号：1006-6616（2021）03-0339-11

一种基于机器学习算法的岩性填图方法

冀全伟[1,2,3]，王文磊[1,2]，刘治博[4]，祝茂强[3]，袁长江[3]
JI Quanwei[1,2,3]，WANG Wenlei[1,2]，LIU Zhibo[4]，ZHU Maoqiang[3]，YUAN Changjiang[3]

1. 自然资源部古地磁与古构造重建重点实验室，北京 100081；
2. 中国地质科学院地质力学研究所，北京 100081；
3. 中国地质大学（北京），北京 100083；
4. 中国地质科学院矿产资源研究所，北京 100037

1. *Key Laboratory of Paleomagnetism and Tectonic Reconstruction of Natural Resources*, *Beijing* 100081, *China*;
2. *Institute of Geomechanics*, *Chinese Academy of Geological Sciences*, *Beijing* 100081, *China*;
3. *China University of Geosciences*, *Beijing* 100083, *China*;
4. *Institute of Mineral Resources*, *Chinese Academy of Geological Sciences*, *Beijing* 100037, *China*

JI Q W, WANG W L, LIU Z B, et al., 2021. A machine learning-based lithologic mapping method [J]. Journal of Geomechanics, 27 (3): 339-349. DOI: 10.12090/j.issn.1006-6616.2021.27.03.031

Abstract: In this study, a gradient boosting decision tree (GBDT)-based lithologic mapping method constituted by field survey and machine learning is introduced. The Duolong mineral district, Tibet, China is currently chosen for model test. During the practical application, geochemical data at a 1∶50000 scale is analyzed to identify lithologic units, while a geological map at the same scale currently provides lithologic units identified by field survey. Lithologic units within a small area are firstly collected from the geological map. Correspondence between geochemical data and lithologic units within the small area can consequently be marked, by which the GBDT method is applied to reclassify the geochemical data and further predict lithologic units in the Duolong district. Transforming the result to a probability distribution, areas with low probability can be identified, and further investigation will be implemented to update geological knowledge and correspondence between geochemical and lithologic units. Iteration of the process will lead a reasonable lithologic mapping result. It is shown that the model accuracy increases with iteration growing, and reaches 87% after 7 iterations. The currently proposed method highlights deep integration of field survey and machine learning algorithm, and emphasizes importance of field work in the whole modeling process. Useful geo-information can be deeply mined from existing data and further updates former geological understandings. Meanwhile, lithologic units within un-explored areas can be identified based on the knowledge in explored areas. The GBDT-based method which effectively reduces field work is a meaningful exploration in lithologic mapping and will provide a new reference and supplementary way to geological mapping.
Key words: data mining; information fusion; geologic unit; decision tree; geological mapping

摘　要：通过野外地质调查与机器学习方法的有机融合，提出了一种基于梯度提升决策树算法的岩性单元填图方法。研究以多龙矿集区为模型试验区，选择1∶5万勘查地球化学数据为基础预测数据，以1∶5万区域地质图为参考，进行基于梯度提升决策树算法的岩性预测填图模型试验。首先选择研究区内小范围空白区开展

基金项目：国家自然科学基金项目（41822206，41772353）
This research is financially supported by the National Natural Science Foundation of China (Grant No. 41822206, 41772353)
第一作者简介：冀全伟（1996—），男，在读硕士，从事定量地学研究。E-mail：jqwcug@163.com
通讯作者：王文磊（1983—），男，研究员，从事数学地质研究。E-mail：wenleiw@163.com
收稿日期：2020-11-09；**修回日期**：2021-01-10；**责任编辑**：范二平
引用格式：冀全伟，王文磊，刘治博，等，2021. 一种基于机器学习算法的岩性填图方法［J］. 地质力学学报，27（3）：339-349. DOI：10.12090/j.issn.1006-6616.2021.27.03.031

野外填图，建立原始数据集并初步构建岩性单元与预测数据对应关系；其次利用机器学习方法对预测数据进行多分类任务，进而开展目标填图区预测填图工作；最后通过概率选区选定概率较低目标区，开展进一步的小范围野外地质调查填图，对原始数据和知识库进行补充，迭代循环以上流程，直至预测填图达到要求。试验显示，随着迭代次数的增加，模型精度不断提高，并在7次迭代后模型准确率达到87%。该方法强调在实际应用中野外地质调查与基于机器学习预测填图的深度融合，以及野外实地工作在整个流程中的重要性和不可或缺性；同时能够充分挖掘已有数据资料的有用信息，用于辅助修正已有岩性填图内容，或根据已勘探区资料对邻近的未勘探区进行岩性分类，有效减少野外填图工作量，是对岩性填图方法、地质单元定量预测识别的有益探索，为区域地质填图工作提供了新的参考思路和辅助手段。

关键词：数据挖掘；信息融合；地质单元；决策树；地质填图

中图分类号：P628 **文献标识码**：A

0 引言

信息化时代，社会经济发展与生态环境治理对地质调查工作提出了新的要求，地质调查工作面临新的机遇与挑战。例如，在特殊地质地貌区开展区域地质调查工作将有助于特殊地质景观区基础地质问题的研究，服务于多门类自然资源与生态环境问题的解决（胡健民和陈虹，2019）。随着地质调查工作的持续开展，基础地质研究程度不断提高，成果数据资料保持快速积累与更新。如何系统整合已有地质、地球化学、地球物理、遥感等多元、多尺度地质调查数据资料，发展能够提高工作质量与效率的方法，深度挖掘有用信息，进而优化提升基础地质、矿产地质、水文地质、灾害地质等调查评价技术（杨星辰等，2020；张鑫刚等，2020），被认为是地质调查工作手段升级，提高社会经济服务能力的突破口之一。亟需学习吸收并引进数学、信息学等学科先进的数据与信息挖掘技术，创新发展地质调查评价思路与方法。

地质填图作为区域地质调查工作最基本的核心工作内容之一，其效率和精度将直接影响后续研究工作的开展。传统地质填图工作主要包括前期资料收集整理、工作方案编制、野外实地勘查、样品测试分析及数据处理、成图及报告编写等阶段。其中，前期资料收集整理工作多停留在基本资料了解阶段，基础资料及数据的应用程度不高；而野外工作依靠地质工作者的主观判断来确定填图单元，受限于填图技术人员的业务水平不同，填图质量受到一定影响。因此，为保证填图成果质量，野外实地勘查工作需投入较高的人力、财力和物力成本来完成大量路线调查及剖面测量等实物工作量。此外，在偏远山区、无人区、高原地区开展野外工作还存在一定风险性。

随着机器学习方法的快速发展，基于机器学习的岩性填图方法的提出，取得了较好的研究成果与进展。相较传统地质填图技术，机器学习方法中的分类模型或组合算法在岩性分类识别方面具有高效、智能化的特点，可作为具有巨大潜在优势的辅助手段来提高传统地质填图技术方法体系的工作效率与能力。已有基于机器学习方法的岩性填图研究（吴俊等，2016；陈松等，2017），通过系统整合多源遥感、地震、物探、化探、航磁等数据，建立岩性分类的基础数据集，利用度量学习、支持向量机（SVM）、人工神经网络（ANN）、随机森林（RF）等机器学习分类算法，开展了岩性识别、岩性单元填图等相关分类问题的研究（Cracknell and Reading，2014；Harris and Grunsky，2015；郑阳，2017；Othman and Gloaguen，2017；Kuhn et al.，2018；张艳等，2019；段友祥等，2020；朱明永等，2020；Wang et al.，2020a，2020b；Wu et al.，2021）。已有研究表明，这一岩性填图思路在特定地质条件下具有特殊优势（严昊伟等，2017）。

文章主要通过野外基础地质调查和机器学习分类算法的有机融合，在填图空白地区或工作程度较低地区开展基于勘查数据分析预测的岩性单元填图方法探索性研究。选取西藏多龙矿集区开展模型试验主要是考虑到两方面原因。首先，多龙矿集区是中国重要成矿区带班公湖-怒江成矿带内已发现最大的斑岩型Cu-Au矿产地，具有巨大资源潜力。区内近年来已完成了1∶5万区域与矿产地质调查工作，对岩性单元划分具有较为清晰的

认识，有利于预测结果的验证与应用效果评价。其次，多龙矿集区积累了大量基础图件和勘查数据资料，可供研究通过选取不同基础预测数据组合，构建不同工作基础条件下的模型方法试验。同时，文中提出的数据填图方法需要开展多批次小范围野外填图支撑岩性单元预测的迭代算法。在模型试验过程中，已有地质图件能够代替野外填图直接为预测模型提供原始数据和现有知识补充。换而言之，通过从已有地质图中提取迭代算法所需的小范围岩性单元分布来实现数据集与知识库的更新，为模型试验节省了实际野外填图的时间成本。因此，研究以多龙矿集区为模型试验区，选择 1∶5万勘查地球化学数据为基础预测数据，以 1∶5万区域地质图为参考，进行基于梯度提升决策树算法的岩性预测填图模型试验。

1　研究区概况

多龙矿集区位于西藏阿里地区改则县境内，所处的大地构造位置为班公湖-怒江成矿带西段，班公湖-怒江缝合带北侧、羌塘-三江复合板片南缘（郭娜等，2018；李兴奎等，2018；任纪瞬等，2019）。地层分区属于羌南-保山地层区多玛地层分区，区内地层（图 1）以中生界为主，主要有中侏罗统曲色组（J_2q）和色哇组（J_2s）浊积岩建造、下白垩统美日切错组（K_1m）火山碎屑岩建造以及新生界新近系康托组（N_1k）陆源碎屑岩建造和第四系残坡积物（Q_4）（江少卿等，2014；陈红旗等，2015）。其中，J_2q 组岩性为粉砂质板岩夹变长石石英砂岩（李云强等，2020），J_2s 组的岩石主要由砂岩、砂砾岩和变长石石英砂岩等组成（符家骏等，2014），同时两组地层也是含矿岩体的主要围岩（王勤等，2018）。K_1m 组的岩石主要为安山岩、英安岩、玄武岩、火山角砾岩和碎屑岩等。N_1k 组以砾岩、含砾砂岩、红色泥岩为主要岩性（韦少港等，2017）。多龙矿集区岩浆活动极为发育，总体上以喷发、喷溢和浅成、超浅成侵入为主，具多期活动特征，形成时代为燕山中—晚期（江少卿等，2014；李红梅，2017）。喷出岩主要由玄武岩、安山岩和流纹岩组成，（孙嘉等，2019）。侵入岩主要为基性、中酸性侵入岩，基性岩主要为辉长岩和辉绿岩，中酸性浅成岩主要为闪长岩、英安岩、花岗闪长斑岩，侵入时代以早白垩为主（陈红旗等，2015；王勤等，2018）。区内接触变质岩变质程度不高，岩体周边广泛发育热液蚀变及少量石英脉（王继斌，2018）。

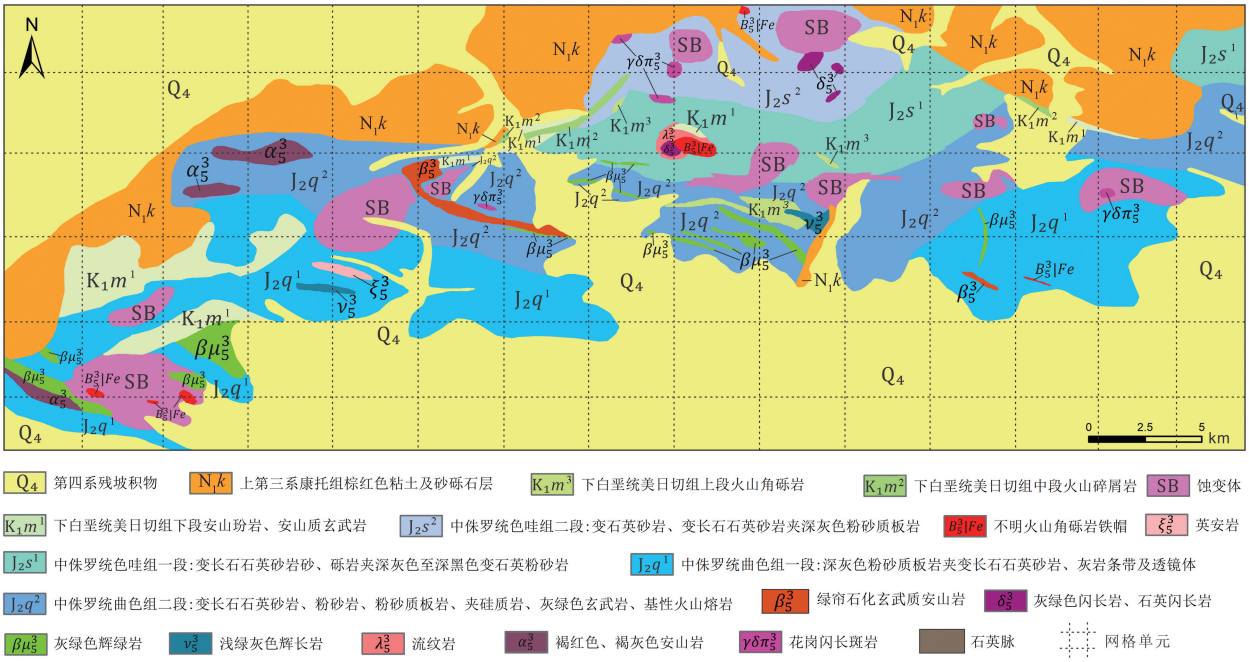

图 1　多龙矿集区岩性分布图

Fig. 1　Spatial distribution of the lithologic units in the Duolong mineral district, Tibet, China

自 20 世纪 70 年代以来，先后有多家地勘和研究单位对多龙矿集区开展了 1∶100 万、1∶25 万和

1∶5万图幅的区域地质调查工作。该区的区域物探、化探、遥感、矿床勘查工作以及相关岩石地球化学（韦少港等，2019）、年代学（王勤等，2015）、控矿构造识别（刘治博等，2017）、遥感异常信息提取（代晶晶等，2013；别小娟等，2013）、蚀变矿物学（赵子欧等，2020）等方面研究取得了较好的成果进展。通过近些年多方面研究，对多龙矿集区的地质背景、成矿规律、矿床模型等有了新的认识（杨欢欢等，2019；王勤等，2019；石洪召等，2019；孙嘉等，2020），目前正根据已有资料开展进一步综合研究。

2 基于机器学习的岩性填图方法

2.1 基于机器学习的岩性填图思路

基于机器学习方法的岩性填图对研究区的基础地质数据积累与研究程度具有较高要求，大多针对特定的数据资料类型且依赖高质量数据集，在空白区或数据资料不充分地区开展工作，将会面临缺乏基础地质支撑的困难。文中通过野外地质调查与机器学习方法的有机融合，提出了一种基于梯度提升决策树（Gradient boosting decision tree，GBDT）算法的岩性单元填图方法（图 2）：①选择研究区内小范围已填图区作为假想野外填图区，建立原始数据集并初步构建岩性单元与预测数据（遥感、化探、物探）对应关系；②利用机器学习方法对预测数据进行多分类任务，进而开展目标填图区预测填图工作；③通过概率选区选定概率较低目标区，开展进一步的小范围野外地质调查假想填图，对原始数据和现有知识进行补充；④迭代循环以上流程，直至预测填图达到要求。

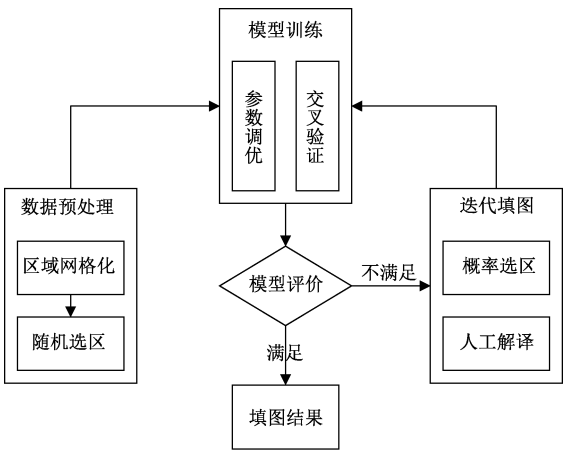

图 2 基于机器学习的岩性填图思路

Fig. 2 Flowchart of machine learning-based lithologic mapping

2.2 数据预处理

数据预处理时，若将研究区整体定义为单一栅格作为目标选区基本单元，代表性较弱，同时易受模型分类过程中分类准确率的影响。因此，需要通过对研究区进行网格化划分（图 1），将基本单元由单一栅格分解为 $w \times h$ 个网格单元，并在此基础上进行概率均值的统计，以此作为迭代填图目标选区的评判基础。文中将多龙矿集区内填图范围划分成 90 个网格单元，网格单元面积为 3.5 km×3.5 km。

针对研究区进行网格化处理后，通过随机选区的采样策略完成初始数据集的创建。从研究区划分网格单元中随机选取若干单元作为目标采样区。通过野外地质调查在选区内开展地质填图，获取区内岩性单元分布情况。模型试验将通过从已有地质图中直接提取选区内的岩性单元分布来代替野外实际填图工作。

通过距离反比权重法（IDW）对试验区 3200 个地球化学数据点进行空间插值，得到的栅格数据作为模型试验的预测数据。将初始选区的岩性填图结果与对应的地球化学数据进行标签化整合，完成初始数据集的创建。最后，通过模型训练建立岩性算法分类模型，根据模型评价标准实施迭代填图，预测全区岩性分布结果，进而探索基于 GBDT 算法的岩性填图方法。

2.3 梯度提升决策树（GBDT）

梯度提升决策树（Gradient boosting decision tree，GBDT）算法（Friedman，2001）是一种采用集成学习思想的迭代决策树算法。所谓集成学习，即通过对多个学习器（如决策树）的组合得到比单一学习器性能更好的算法模型训练策略。一般情况下，GBDT 以决策树（Quinlan，1986）为基础分类器，并利用损失函数的负梯度作为提升树残差的近似值进行算法实现。其中，提升树 $f_M(x)$ 可表示为：

$$f_M(x) = \sum_{m=1}^{M} \gamma_m T_m(x) \tag{1}$$

其中，$T_m(x)$ 为弱学习器，即决策树；γ_m 为每个弱学习器最优拟合的权重；M 为树的个数，即迭代次数。

模型的训练过程是损失函数 L 的最小化过程。假设训练样本数据量为 N，第 i 条数据的变量与真值分别为 x_i 和 y_i，则参数调优的目标函数为：

$$\hat{f} = \arg\min_{f} \sum_{i=1}^{N} L(y_i, f_{m-1}(x_i) + \gamma_m T_m(x_i)) \tag{2}$$

其中，$\hat{f}$ 表示训练完成的预测模型；L 为训练过程中的损失函数；$argmin$ 则表示最小化损失函数 L 时 f 的取值；其他变量同公式（1）。

2.4　归一化指数函数

归一化指数函数（Softmax）是逻辑函数在多分类任务上的一种推广，其目的是将多分类结果以概率的形式展现出来。若以 D_T 表示样本训练集，则 $D_T = \{(x_i, y_i), i = 1, \cdots, n_T\}$。其中，$x_i$ 是模型输入的数据，如用来预测岩性单元的遥感、地球化学等数据；y_i 是对应地质目标名称，如岩性单元标签。假设训练集岩性单元种类数为 K，则一般情况下 $n_T > K$。在分类问题上，GBDT 的作用是计算 x_i 与 y_i 之间的映射函数 f：$R^{15} \rightarrow R^K$。对于输入的 x，输出 P 维特征向量 $\boldsymbol{\nu}$，并代入 Softmax 函数计算分类概率值：

$$p_k = \frac{e^{\nu_k}}{\sum_{j=1}^{P} e^{\nu_j}} \tag{3}$$

其中，p_k 表示属于第 k 类岩性的预测概率值。根据 Softmax 计算公式可知，对于任一数据 x，各岩性预测概率之和必为 1。

3　模型训练与迭代

3.1　模型训练

文中采用 GBDT 作为核心算法对区内地球化学数据与岩性单元的对应关系开展信息挖掘与拟合工作。针对小样本数据集，特别是当前基础预测数据小于 10^4 数量级的情况下，GBDT 算法在训练的过程中可能会出现过拟合问题。目标函数在机器学习过程中将会过度依赖训练样本集，将所有样本（包括噪声）都拟合到函数当中，从而只在训练集中表现优异，对于未知样本则无法正确预测。因此，为客观判断训练参数对训练集以外数据的符合程度，论文采用交叉验证的思想对模型整体分类能力进行评估。将样本数据集随机分为 F 个不相交子集，从 F 个子集中逐次选取一个子集定义为测试集，其余 $F-1$ 个子集定义为训练集，基于训练集进行训练得到 GBDT 模型。利用测试集对模型进行分类器性能评价，将 F 次测试结果的均值定义为 F 折交叉验证下模型性能指标，并以此来评估模型精度。此外，需要在交叉验证基础上进行多次参数调优，得到更为合理的模型参数，以保证训练得到的 GBDT 模型具备较强的分类能力。

根据每次迭代过程中对模型进行多次训练的结果（图 3）可知，经过 300 次训练后模型表现趋于平稳，损失值基本稳定在 0.2。这说明即使对于较为复杂的多分类问题，该模型仍具有较强的有效性和稳定性。

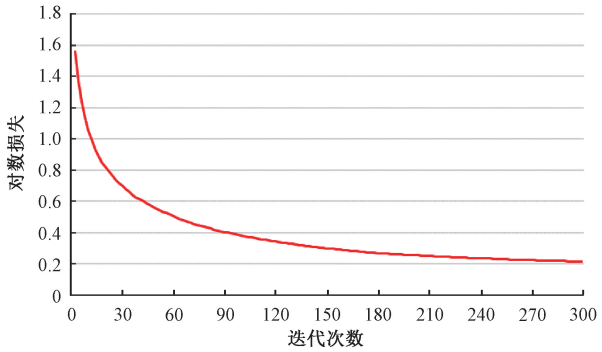

图 3　模型损失函数统计图

Fig. 3　Statistical diagram of the loss function

3.2　迭代填图

从概率角度选定网格单元，将其作为目标填图区进行针对性的迭代填图，并逐步更新预测分类数据集是此次研究思路核心之一。迭代填图这一思路作为整套方法流程中最主要的数据与知识补充过程，其准确性高低将对最终出图结果造成直接影响。与传统岩性填图相结合，通过专家野外填图的方式完成概率选区范围内的信息采集工作，在保证结果精度的前提下减少传统岩性填图的野外实际工作量。在具体实施过程中，根据研究区预测概率分布结果（图 4），以网格为基本单元进行概率均值计算。按概率高低对全部单元进行排序，选取其中概率最低的若干网格单元（图 4 中黑框位置）作为目标区域，开展野外局部实地填图。将填图区岩性分类结果与对应的地球化学数据进行整合，并更新至样本数据库。

4　岩性填图结果分析

4.1　模型评价结果

模型评价主要包括适用性和实用性评价两个方面。模型适用性评价主要是从算法角度评价 GBDT 模型对地质问题的适用性。针对从区内网格中选取的野外填图区，根据野外填图获得岩性分布，按比例划分出预测评价区。训练模型应用于预测区获得相应的岩性分类结果。以地质图为真

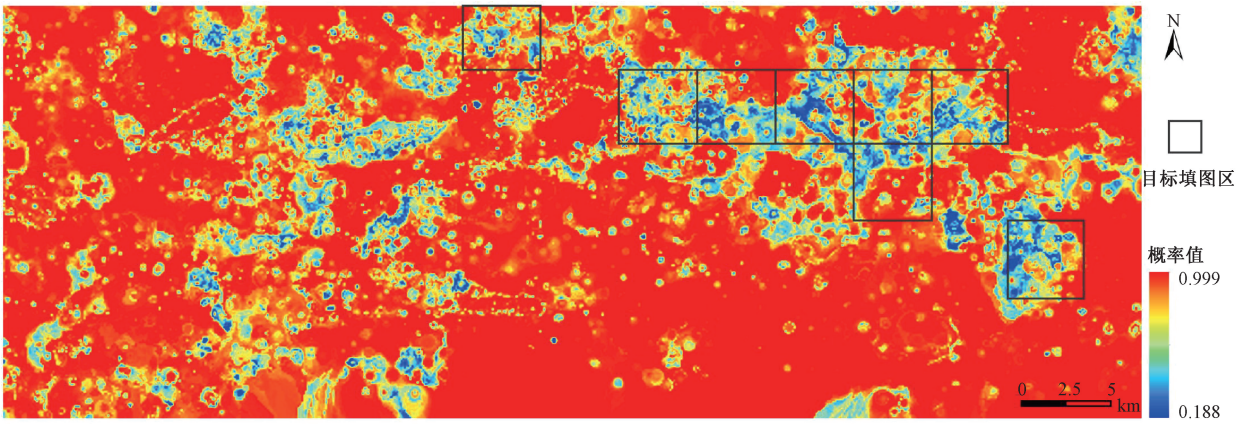

图 4 概率分布选区示意图

Fig. 4 Schematic diagram of probability distribution-based area selection

值统计分类结果的准确率、宏平均 F_1 分数等模型评价指标，并根据各类指标情况进行模型修正。

模型实用性评价主要是从预测概率角度评价预测结果对预期分类结果的满意程度。预测概率值是将模型输出值与各类岩性单元特征向量之间的残差通过 Softmax 函数进行归一化计算获得。概率值高低代表当前地球化学数据分类结果与各岩性单元类型的相近程度。假设已知专家填图区岩性单元种类集合为 S，则概率分布高值区通常代表当前第 i 区域分类 $k_i \in S$，低值区表示当前分类范围较大可能存在实际岩性单元种类 $k_i \notin S$ 的情况。基于以上原则，将模型预测概率与预期结果进行对比。若满足，则将模型应用于全区地球化学数据并预测全区岩性单元分类，否则，进行迭代填图，直至满足预期。

文中采用准确率（Accuracy，简记 Ac）、宏平均精确率（Macro Average Precision，简记 Pr）、宏平均召回率（Macro Average Recall，简记 Re）以及宏平均 F_1 分数（Macro Average F_1，简记 F_1）等指标对基于机器学习方法的岩性单元分类任务进行性能评估。其中，准确率表示正确预测的样本比例，宏平均精确率表示预测为正样本中正确的比例，宏平均召回率表示正样本中预测正确的比例；宏平均 F_1 分数是兼顾宏平均精确率与宏平均召回率的调和平均数。

假设混淆矩阵 $\boldsymbol{G}$：

$$\boldsymbol{G} = \begin{bmatrix} g_{11} & g_{12} & \cdots & g_{1K} \\ g_{21} & g_{22} & \cdots & g_{2K} \\ \vdots & \vdots & \ddots & \vdots \\ g_{K1} & g_{K2} & \cdots & g_{KK} \end{bmatrix} \tag{4}$$

其中，K 表示岩性种类数。

在混淆矩阵 $\boldsymbol{G}$ 中准确率、宏平均精确率、宏平均召回率以及宏平均 F_1 分数的计算公式：

$$Ac = \frac{1}{\sum_{a=1}^{K}\sum_{b=1}^{K} g_{ab}} \times \sum_{a=1}^{K} g_{aa} \tag{5}$$

$$Pr = \frac{\sum_{b=1}^{K} \dfrac{g_{aa}}{\sum_{a=1}^{K} g_{ab}}}{K} \tag{6}$$

$$Re = \frac{\sum_{a=1}^{K} \dfrac{g_{aa}}{\sum_{b=1}^{K} g_{ab}}}{K} \tag{7}$$

$$F_1 = \frac{2 \times Pr \times Re}{Pr + Re} \tag{8}$$

其中，g_{aa} 表示 a 类岩性预测正确的数量；g_{ab} 表示 a 类岩性预测为 b 类的数量。

利用上文所述方法在多龙矿集区开展岩性单元预测分类模型试验，获得了迭代过程各阶段的模型评价指标。结果显示，采用概率选区原则进行数据样本逐步更新的思路具有良好表现，各指标随迭代均保持递增（表 1）。以准确率为例，该指标表示当前分类结果与该区实际填图获得的岩性单元的匹配程度。模型经过 7 次迭代更新，准确率从初始 47.3%增加至 87%，性能提升近一倍。

同时，结果显示 7 次迭代后野外实际填图的累计范围占研究区面积的 62.2%（表 2），即，在全区约 2/3 范围内开展野外填图的情况下，获得了与传统填图方法相近的岩性分类结果，说明文中提出的预测填图方法在岩性填图工作中的效率。根据已有地质图可知，区内岩性单元种类数为 20。

表 1　模型迭代性能统计表

Table 1　Performance of model iteration

迭代次数	宏平均精确率	宏平均召回率	宏平均 F_1 分数	准确率
1	0.348	0.193	0.185	0.473
2	0.472	0.362	0.361	0.571
3	0.633	0.472	0.507	0.635
4	0.737	0.600	0.638	0.707
5	0.761	0.683	0.701	0.768
6	0.797	0.715	0.746	0.821
7	0.827	0.765	0.789	0.870

岩性单元预测种类数在 7 次迭代过程中由 13 类增加至 19 类，覆盖率达到 95%。经统计发现，由于石英脉在研究区面积占比极少，仅为 0.007%，缺少足够的数据样本，未能在研究中成功预测分类。由此可见，该研究方法从概率的角度定义迭代填图范围具有较高可行性。

表 2　迭代分析结果信息统计表

Table 2　Statistics table of iteration results

迭代次数	面积占比	岩性种类数
1	0.088	13
2	0.171	17
3	0.262	18
4	0.357	19
5	0.445	19
6	0.531	19
7	0.622	19

4.2　预测分类结果

从 7 次迭代后的预测分类结果来看（图 5），在岩性单元分布较为复杂且多类型交替出现的场景下，相应的岩性单元边界仍能被有效地划分。该方法通过机器学习算法进行分类，提高了岩性单元填图的工作效率。同时，与野外填图结果对比发现具有较高的吻合度，体现了对岩性单元预测分类的准确性。

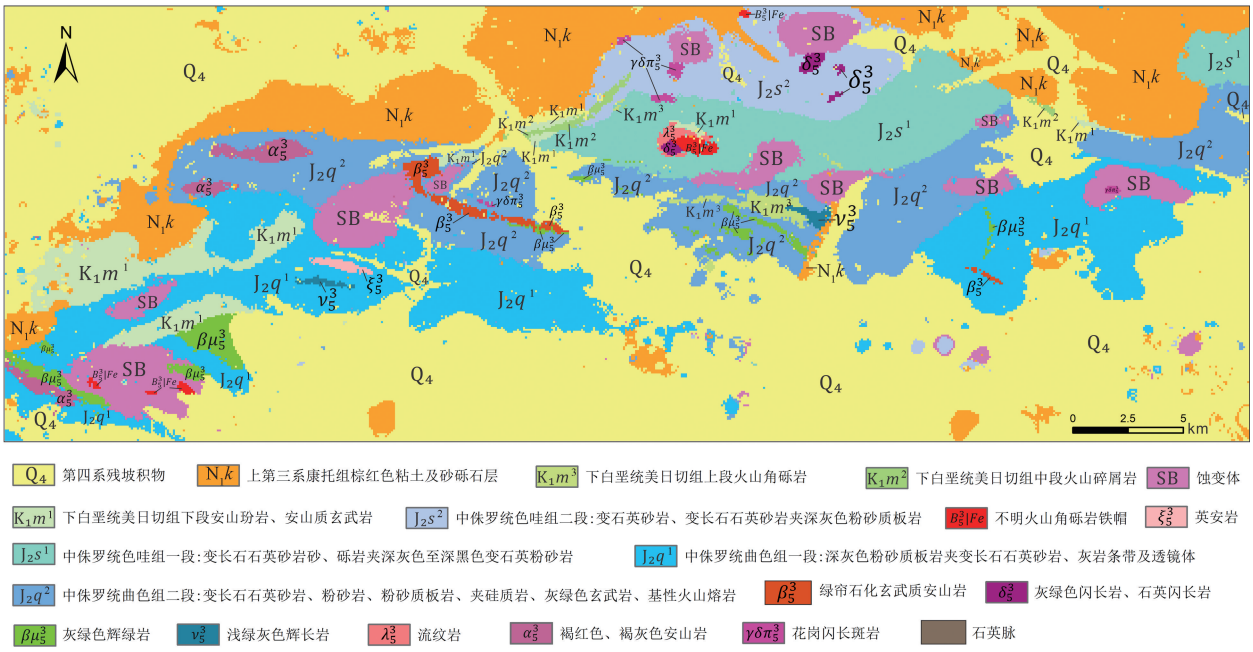

图 5　多龙矿集区岩石单元预测结果

Fig. 5　Prediction results of lithologic units in the Duolong mineral district

模型试验经过 7 次迭代后，预测概率达到预期要求，分类结果涉及 19 类不同岩性单元。采用宏平均 F_1 分数对各类单元进行精度评价（表 3），模型分类精度整体表现优秀，各类预测精度均值达到 0.845，其中 5 类超过 0.9，仅有 1 类不足 0.7，占比 5%。最高为 δ_5^3，达到 0.935，且该岩性单元仅占全区面积的 0.47%，这说明该方法对于研究区面积占比较低的岩性单元仍具备较高的识别能力。最低为 $\beta\mu_5^3$，宏平均 F_1 分数仅有 0.683，但具备同等地球化学元素组成的 β_5^3，其宏平均 F_1 分数达到了 0.8，反映了该方法虽然对以岩石结构特征命名的地质单元无法有效区分，但对具备相同地球化学元素特征的岩性大类仍具备较高准确度。此外，通过预测结果与已知结果对比发现，第四纪区域预测与原地质图有一定差别。经遥感查证，在排除第四系冲积扇区域之后，原 1∶5万图幅的第

四系分布范围内局部显示了露头出露，表明该方法对已有填图工作有部分修正作用。由于地球化学元素反演矿化蚀变的天然优势，该方法对蚀变区域的有效识别，可产生重要的经济价值。

表 3 模型分类精度表

Table 3 Table of classification accuracy of the current model

岩性单元	岩性符号	宏平均 F_1 分数
流纹岩	λ_5^3	0.880
第四系残坡积物	Q_4	0.926
上第三系康托组棕红色粘土及砂砾石层	N_1k	0.794
下白垩统美日切组上段火山角砾岩	K_1m^3	0.834
下白垩统美日切组中段火山碎屑岩	K_1m^2	0.810
下白垩统美日切组下段安山玢岩、安山质玄武岩	K_1m^1	0.843
中侏罗统色哇组二段：变石英砂岩、变长石石英砂岩夹深灰色粉砂质板岩	J_2s^2	0.909
中侏罗统色哇组一段：变长石石英砂岩砂、砾岩夹深灰色至深黑色变石英粉砂岩	J_2s^1	0.933
中侏罗统曲色组二段：变长石石英砂岩、粉砂岩、粉砂质板岩、夹硅质岩、灰绿色玄武岩、基性火山熔岩	J_2q^2	0.870
中侏罗统曲色组一段：深灰色粉砂质板岩夹变长石石英砂岩、灰岩条带及透镜体	J_2q^1	0.849
不明火山角砾岩铁帽	$B_5^3 \mid Fe$	0.839
褐红色、褐灰色安山岩	α_5^3	0.739
浅绿灰色辉长岩	ν_5^3	0.841
灰绿色辉绿岩	$\beta\mu_5^3$	0.683
灰绿色闪长岩、石英闪长岩	δ_5^3	0.935
英安岩	ξ_5^3	0.906
花岗闪长斑岩	$\gamma\delta\pi_5^3$	0.756
绿帘石化玄武质安山岩	β_5^3	0.882
蚀变体	SB	0.834

5 结论

文中提出了一种基于 GBDT 算法的岩性单元预测分类方法，将西藏多龙矿集区作为试验区，以 1∶5万勘查地球化学数据为例，对岩性填图方法进行了有益的探索。研究强调了野外地质填图与基于机器学习预测分类方法的深度融合，以及野地质调查工作在岩性预测填图工作中的重要性和不可或缺性。在强调野外地质调查重要性的基础上，将岩性填图工作融入了机器学习方法。通过小范围野外人工填图迭代更新数据与知识库，从而对全区进行岩性单元预测分类工作。该方法是对岩性单元填图工作思路和流程的探索，是对现有工作模式的一种有益补充与辅助优化；体现了“基于大数据理论方法来促进地质问题的解决，并不意味着取代或摒弃地学传统方法，而在于激活、提升和创新发展传统方法”这一大数据科学范式在地质科学研究中的特点和优势。

传统岩性填图方法通常要求对穿越地质体最多、地质构造复杂的路线进行复杂详尽的野外调查工作，文中采用概率选区的方式来确定迭代填图过程中的目标填图区，使整个岩性填图过程更具有针对性与高效性。根据试验结果对比研究区地质图，该方法基于 62.2% 的已知研究区信息，有效实现了 87% 研究区范围内的岩性单元分类。这一结果证明该方法不仅具有良好的填图效果，而且能够有效减轻野外填图工作量。对在新疆、青海、西藏等野外环境条件艰苦地区的岩性填图工作具有积极的参考作用。此外，为验证该方法的通用性，未来可开展除化探数据以外其他数据资料，如遥感、航磁、航放、钻井等数据资料的适用性研究，从而共同为地质资料相对匮乏或单一的研究区开展岩性填图工作提供有效支撑。

References

BIE X J, ZHANG T B, SUN C M, et al., 2013. Extraction of remote sensing anomaly and metallogenic prediction in Duolong ore-concentrated area of Tibet [J]. Journal of Guilin University of Technology, 33 (2): 252-258. (in Chinese with English abstract)

CHEN H Q, QU X M, FAN S F, 2015. Geological characteristics and metallogenic-prospecting model of Duolong porphyry copper-gold ore concentration area in Gerze County, Tibet [J]. Mineral Deposits, 34 (2): 321-332. (in Chinese with English abstract)

CHEN S, CHEN C J, WU J, et al., 2017. Application and exploration of geophysical methods in geological mapping in strongly weathered area [J]. Journal of Geomechanics, 23 (2): 206-213. (in Chinese with English abstract)

CRACKNELL M J, READING A M, 2014. Geological mapping using

remote sensing data: A comparison of five machine learning algorithms, their response to variations in the spatial distribution of training data and the use of explicit spatial information [J]. Computers & Geosciences, 63: 22-33, doi: 10.1016/j. cageo. 2013.10.008.

DAI J J, WANG R J, QU X M, et al., 2013. Application of remote sensing alteration information in prospecting of Duolong ore concentration area in Tibet [J]. Acta Mineralogica Sinica, 33 (S2): 753-754. (in Chinese)

DUAN Y X, ZHAO Y S, MA C F, et al., 2020. Lithology identification method based on multi-layer ensemble learning [J]. Journal of Data Acquisition and Processing, 35 (3): 572-581. (in Chinese with English abstract)

FRIEDMAN J H, 2001. Greedy function approximation: a gradient boosting machine [J]. The Annals of Statistics, 29 (5): 1189-1536.

FU J J, ZHAO Y Y, GUO S, 2014. Geochemical characteristics and significance of granodiorite porphyry in the Duolong ore concentration area, Tibet [J]. Acta Petrologica et Mineralogica, 33 (6): 1039-1051. (in Chinese with English abstract)

GUO N, SHI W X, HUANG Y R, et al., 2018. Alteration mapping and prospecting model construction in the Tiegelongnan ore deposit of the Duolong ore concentration area, northern Tibet, based on shortwave infrared technique [J]. Geological Bulletin of China, 37 (2-3): 446-457. (in Chinese with English abstract)

HARRIS J R, GRUNSKY E C, 2015. Predictive lithological mapping of Canada's North using Random Forest classification applied to geophysical and geochemical data [J]. Computers & Geosciences, 80: 9-25, doi: 10.1016/j. cageo. 2015.03.013.

HU J M, CHEN H, 2019. Innovation and exploration of the guiding ideology and method system of geological mapping in the coverage area-an overview of the results of the pilot project of mapping special geological and geomorphic areas [J]. Journal of Geomechanics, 25 (5): 1001-1002. (in Chinese)

JIANG S Q, SUN X G, YANG T Z, et al., 2014. Integrated anomaly model and metallogenic prediction of the Duolong porphyry copper-gold ore concentration area in northern Tibet [J]. Geology in China, 41 (2): 497-509. (in Chinese with English abstract)

KUHN S, CRACKNELL M J, Reading A M, 2018. Lithologic mapping using Random Forests applied to geophysical and remote-sensing data: A demonstration study from the Eastern Goldfields of Australia [J]. Geophysics, 83 (4): B183-B193, doi: 10.1190/geo2017-0590.1.

LI H M, 2017. Hydrothermal alteration mineral (groups) mapping and distribution characterization analysis based on multispectral remote sensing in Duolong ore concentration area, Tibetan [D]. Chengdu: Chengdu University of Technology. (in Chinese with English abstract)

LI X K, LI C, WANG M, et al., 2018. Nature and evolution of crustal basement beneath the Duolong ore concentration area, northern Tibet, and their constraints on the metallogenesis: Insights from U-Pb ages of inherited zircons from the Bolong volcanic-intrusive rocks [J]. Geological Bulletin of China, 37 (8): 1439-1449. (in Chinese with English abstract)

LI Y Q, FEI G C, WEN C Q, et al., 2020. Characteristics of Ce^{4+}/Ce^{3+} ratio and oxygen fugacity of Zircon in porphyry from Bolong and Duobuza porphyry Cu-Au deposits in Duolong ore district, Tibet [J]. Mineralogy and Petrology, 40 (2): 59-70. (in Chinese with English abstract)

LIU Z B, WANG W L, SONG Y, et al., 2017. Geo-information extraction and integration of ore-controlling structure in the Duolong ore concentration area of Tibet [J]. Acta Geoscientica Sinica, 38 (5): 803-812. (in Chinese with English abstract)

OTHMAN A A, GLOAGUEN R, 2017. Integration of spectral, spatial and morphometric data into lithological mapping: A comparison of different Machine Learning Algorithms in the Kurdistan Region, NE Iraq [J]. Journal of Asian Earth Sciences, 146: 90-102, doi: 10.1016/j. jseaes. 2017.05.005.

QUINLAN J R, 1986. Induction of decision trees [J]. Machine Learning, 1 (1): 81-106.

REN J S, NIU B G, ZHAO L, et al, 2019. Basic ideas of the multisphere tectonics of earth system [J]. Journal of Geomechanics, 25 (5): 607-612. (in Chinese with English abstract)

SHI H Z, LI Y C, HUANG H X, et al., 2019. Genesis of early Cretaceous Meiriqiecuo Formation volcanic rocks in the Duolong ore concentration area, southern margin of Qiangtang, Tibet, China [J]. Journal of Chengdu University of Technology (Science & Technology Edition), 46 (4): 421-434. (in Chinese with English abstract)

SUN J, MAO J W, LIN B, et al., 2019. Comparison of ore geology and ore-forming processes of ore deposits (ore spots) in Duolong area, Tibet [J]. Mineral Deposits, 38 (5): 1159-1184. (in Chinese with English abstract)

SUN J, MAO J W, WANG J X, et al., 2020. Timing of Cu-Au mineralization in Nadun Cu-Au deposit of Duolong district, Tibet, and its implication for mineral exploration [J]. Mineral Deposits, 39 (6): 1091-1102. (in Chinese with English abstract)

WANG J B, 2018. Mineral assemblages mapping of porphyry copper deposits based on normalized multispectral remote sensing data in the Dulong ore concentrating area [D]. Chengdu: Chengdu University of Technology. (in Chinese with English abstract)

WANG Q, LIN B, TANG J X, et al., 2018. Diagenesis, lithogenesis and geodynamic setting of intrusions in Senadong Area, Duolong district, Tibet [J]. Earth Science—Journal of China University of Geosciences, 43 (4): 1125-1141. (in Chinese with English abstract)

WANG Q, TANG J X, CHEN Y C, et al., 2019. The metallogenic model and prospecting direction for the Duolong super large copper (gold) district, Tibet [J]. Acta Petrologica Sinica, 35 (3): 879-896. (in Chinese with English abstract)

WANG Q, TANG J X, FANG X, et al., 2015. Petrogenetic setting of andsites in Rongna ore block, Tiegelong Cu (Au-Ag) deposit, Duolong ore concentration area, Tibet: Evidence from zircon U-Pb LA-ICP-MS dating and petrogeochemistry of andsites [J]. Geology in China, 42 (5): 1324-1336. (in Chinese with English abstract)

WANG Z Y, ZUO R G, DONG Y N, 2020a. Mapping Himalayan leucogranites using a hybrid method of metric learning and support vector machine [J]. Computers & Geosciences, 138: 104455, doi:

10.1016/j.cageo.2020.104455.

WANG Z Y, ZUO R G, JING L H, 2020b. Fusion of geochemical and remote-sensing data for lithological mapping using random forest metric learning [J]. Mathematical Geosciences, doi: 10.1007/s11004-020-09897-8.

WEI S G, SONG Y, TANG J X, et al., 2019. Geochemistry, Si-O isotopic compositions and its tectonic significance of the siliceous rocks in the Duolong deposit, Tibet [J]. Acta Geologica Sinica, 93 (2): 428-439. (in Chinese with English abstract)

WEI S G, TANG J X, SONG Y, et al., 2017. Zircons LA-MC-ICP-MS U-Pb Ages, Petrochemical, petrological and its significance of the potassic monzonitic granite porphyry from the Duolong Ore-concentrated district, Gaize County, Xizang (Tibet) [J]. Geological Review, 63 (1): 189-206. (in Chinese with English abstract)

WU G P, CHEN G X, CHENG Q M, et al., 2021. Unsupervised machine learning for lithological mapping using geochemical data in covered areas of Jining, China [J]. Natural Resources Research, 30 (2): 1053-1068, doi: 10.1007/s11053-020-09788-z.

WU J, BU J J, XIE G G, et al., 2016. Application of regional geochemical data in geological mapping in strongly weathered area in southern China [J]. Journal of Geomechanics, 22 (4): 955-966. (in Chinese with English abstract)

YAN H W, LIU J, TIAN Y, 2017. Magnetic characteristics and airborne radioactive anomaly characteristics of the shallow coverage in Baiyintuga area in Inner Mongolia [J]. Modern Mining, 33 (3): 35-39, 45. (in Chinese with English abstract)

YANG X C, YE M N, YE P S, et al, 2020. Information construction method of geological survey projects based on digital mapping technology [J]. Journal of Geomechanics, 26 (2): 263-270. (in Chinese with English abstract)

YANG H H, SONG Y, DILLES J H, et al., 2019. The thermal-tectonic history of the Duolong ore district: evidence from apatite (U-Th) [J]. Acta Petrologica Sinica, 35 (3): 867-878. (in Chinese with English abstract)

ZHANG X G, LI Y C, SUN R B, 2020. State, features, trends and enlightenment of geological mapping in major countries of the world [J]. Mineral Exploration, 11 (2): 301-310. (in Chinese with English abstract)

ZHANG Y, SUN J, YU C C, et al., 2019. Classification of Quaternary Coverings in desert grassland shallow cover area based on multi-source remote sensing data: a case of 1∶50000 pilot geological mapping in Qigandianzi, Inner Mongolia [J]. Geological Science and Technology Information, 38 (2): 281-290. (in Chinese with English abstract)

ZHAO Z O, QIAO D H, ZHAO Y Y, 2020. Alteration mineralogical and geochemical features of the Rongna deposit in Duolong mining district of Tibet and their deep prospecting significances [J]. Acta Petrologica Sinica, 36 (9): 2785-2798. (in Chinese with English abstract)

ZHENG Y, 2017. Research on lithology recognition based on deep learning [D]. Beijing: China University of Petroleum (Beijing). (in Chinese with English abstract)

ZHU M Y, LI B Q, FU H Z, et al., 2020. SVM lithological classification based on multi-source data collaboration: a case study in Jianggalesayi area [J]. Uranium Geology, 36 (4): 288-292, 317. (in Chinese with English abstract)

附中文参考文献

别小娟, 张廷斌, 孙传敏, 等, 2013. 西藏多龙矿集区遥感异常提取与成矿预测 [J]. 桂林理工大学学报, 33 (2): 252-258.

陈红旗, 曲晓明, 范淑芳, 2015. 西藏改则县多龙矿集区斑岩型铜金矿床的地质特征与成矿-找矿模型 [J]. 矿床地质, 34 (2): 321-332.

陈松, 陈长敬, 吴俊, 等, 2017. 物探方法在强风化区填图中的应用探索 [J]. 地质力学学报, 23 (2): 206-213.

代晶晶, 王瑞江, 曲晓明, 等, 2013. 遥感蚀变信息在西藏多龙矿集区找矿中的应用研究 [J]. 矿物学报, 33 (S2): 753-754.

段友祥, 赵云山, 马存飞, 等, 2020. 基于多层集成学习的岩性识别方法 [J]. 数据采集与处理, 35 (3): 572-581.

符家骏, 赵元艺, 郭硕, 2014. 西藏多龙矿集区花岗闪长斑岩地球化学特征及其意义 [J]. 岩石矿物学杂志, 33 (6): 1039-1051.

郭娜, 史维鑫, 黄一入, 等, 2018. 基于短波红外技术的西藏多龙矿集区铁格隆南矿床荣那矿段及其外围蚀变填图-勘查模型构建 [J]. 地质通报, 37 (2-3): 446-457.

胡健民, 陈虹, 2019. 覆盖区区域地质填图指导思想与方法体系的创新与探索: 特殊地质地貌区填图试点项目成果概述 [J]. 地质力学学报, 25 (5): 1001-1002.

江少卿, 孙兴国, 杨铁铮, 等, 2014. 藏北多龙斑岩铜金矿集区综合信息找矿模型研究 [J]. 中国地质, 41 (2): 497-509.

李红梅, 2017. 多龙矿集区遥感蚀变矿物 (组合) 提取及蚀变分带特征研究 [D]. 成都: 成都理工大学.

李兴奎, 李才, 王明, 等, 2018. 藏北多龙矿集区地壳基底性质、演化及其对成矿的制约: 来自波龙火山-侵入岩中继承锆石 U-Pb 年龄的信息 [J]. 地质通报, 37 (8): 1439-1449.

李云强, 费光春, 温春齐, 等, 2020. 西藏多龙矿集区波龙、多不杂斑岩铜金矿床岩体锆石 Ce^{4+}/Ce^{3+} 比值及氧逸度特征 [J]. 矿物岩石, 40 (2): 59-70.

刘治博, 王文磊, 宋扬, 等, 2017. 多龙矿集区控矿构造信息提取、识别与融合 [J]. 地球学报, 38 (5): 803-812.

任纪舜, 牛宝贵, 赵磊, 等, 2019. 地球系统多圈层构造观的基本内涵 [J]. 地质力学学报, 25 (5): 607-612.

石洪召, 李玉昌, 黄瀚霄, 等, 2019. 西藏多龙矿集区早白垩世美日切错组火山岩成因 [J]. 成都理工大学学报 (自然科学版), 46 (4): 421-434.

孙嘉, 毛景文, 林彬, 等, 2019. 西藏多龙矿集区典型矿床 (点) 矿化特征与成矿作用对比研究 [J]. 矿床地质, 38 (5): 1159-1184.

孙嘉, 毛景文, 王佳新, 等, 2020. 西藏多龙矿集区拿顿铜金矿床成矿时代的厘定及其找矿指示意义 [J]. 矿床地质, 39 (6): 1091-1102.

王继斌, 2018. 基于归一化多光谱遥感数据的多龙矿集区斑岩铜矿蚀变矿物组合提取 [D]. 成都: 成都理工大学.

王勤, 唐菊兴, 方向, 等, 2015. 西藏多龙矿集区铁格隆南铜 (金银) 矿床荣那矿段安山岩成岩背景: 来自锆石 U-Pb 年代学、岩石地球化学的证据 [J]. 中国地质, 42 (5): 1324-1336.

王勤, 林彬, 唐菊兴, 等, 2018. 多龙矿集区色那东岩体年龄、成因

与动力学背景［J］. 地球科学—中国地质大学学报，43（4）：1125-1141.

王勤，唐菊兴，陈毓川，等，2019. 西藏多龙超大型铜（金）矿集区成矿模式与找矿方向［J］. 岩石学报，35（3）：879-896.

韦少港，唐菊兴，宋扬，等，2017. 西藏改则多龙矿集区地堡那木岗矿床钾玄质二长花岗斑岩锆石 LA-MC-ICP-MS U-Pb 年龄、地球化学特征及其地质意义［J］. 地质论评，63（1）：189-206.

韦少港，宋扬，唐菊兴，等，2019. 西藏多龙矿集区硅质岩岩石地球化学、Si-O 同位素特征及其构造意义［J］. 地质学报，93（2）：428-439.

吴俊，卜建军，谢国刚，等，2016. 区域化探数据在华南强烈风化区地质填图中的应用［J］. 地质力学学报，22（4）：955-966.

严昊伟，刘君，田野，2017. 内蒙古白音图嘎浅覆盖区地层磁性及航空放射性异常特征［J］. 现代矿业，33（3）：35-39，45.

杨星辰，叶梦旎，叶培盛，等，2020. 地质调查成果信息化建设方法探索：基于数字填图技术［J］. 地质力学学报，26（2）：263-270.

杨欢欢，宋扬，DILLES J H，et al.，2019. 西藏多龙矿集区热构造演化历史：来自磷灰石（U-Th）/He 的证据［J］. 岩石学报，35（3）：867-878.

张鑫刚，李仰春，孙仁斌，2020. 世界主要国家地质填图现状、特点、趋势及启示［J］. 矿产勘查，11（2）：301-310.

张艳，孙杰，于长春，等，2019. 基于多源遥感数据的第四系覆盖物分类方法研究：以内蒙古旗杆甸子幅 1∶5万填图试点为例［J］. 地质科技情报，38（2）：281-290.

赵子欧，乔东海，赵元艺，2020. 西藏多龙矿集区荣那铜金矿床蚀变矿物学和地球化学及找矿意义［J］. 岩石学报，36（9）：2785-2798.

郑阳，2017. 基于深度学习的岩性识别研究［D］. 北京：中国石油大学（北京）.

朱明永，李炳谦，付翰泽，等，2020. 基于多源数据协同的 SVM 岩性分类研究：以江尕勒萨依地区为例［J］. 铀矿地质，36（4）：288-292，317.

开放科学（资源服务）标识码（OSID）：

可扫码直接下载文章电子版，也有可能听到作者的语音介绍及更多文章相关资讯