工程管理
Journal of Project Management
第 7 卷◆第 2 期◆版本 1.0◆2026 年
文章类型：论文|刊号（ISSN）：2737-4580(P) / 2737-4599(O)

人工智能辅助地球化学数据异常识别的模型构建与验证

孙哲雨
中化地质矿山总局黑龙江地质勘查院

DOI:10.32629/jpm.v7i2.8751

[摘　要] 地球化学数据异常识别在矿产勘查、地质调查等领域具有至关重要的意义，传统识别方法在处理复杂数据时面临诸多挑战，难以满足实际需求。本文旨在构建高效准确的人工智能模型，以提升地球化学数据异常识别的能力。研究采用了深度学习模型，通过对特定区域水系沉积物等地球化学数据的采集与预处理，进行特征选择、提取与转换，完成模型构建。随后，利用训练集对模型进行训练，并通过验证集与测试集对模型展开验证。结果表明，所构建的人工智能模型在地球化学数据异常识别方面展现出较高的准确性与效率，为矿产勘查等实际工作提供了有力的技术支持。

[关键词] 人工智能；地球化学数据；异常识别；模型构建；模型验证

Model Construction and Validation of Artificial Intelligence-Assisted Geochemical Data Anomaly Recognition

Sun Zheyu
Heilongjiang Institute of Geological Exploration, China Geological and Mineral Resources Administration

[Abstract] Geochemical data anomaly detection plays a vital role in mineral exploration and geological surveys. Traditional methods face significant challenges in processing complex datasets, often failing to meet practical requirements. This study aims to develop an efficient and accurate artificial intelligence model to enhance geochemical anomaly detection capabilities. The research employs deep learning models, involving data collection and preprocessing of geochemical parameters such as sedimentary data from specific water systems. Through feature selection, extraction, and transformation, the model is constructed. Subsequently, the model is trained using a training set and validated through a validation and test set. Results demonstrate that the developed AI model exhibits high accuracy and efficiency in detecting geochemical anomalies, providing robust technical support for practical applications in mineral exploration.

[Key words] artificial intelligence; geochemical data; anomaly detection; model construction; model validation

引言

地球化学数据异常识别在矿产勘查、环境评价和地质调查等领域具有不可替代的重要作用。地球化学异常通常是指由地质作用或矿化过程引起的元素浓度显著偏离背景值的现象，其识别对于发现隐伏矿体、评估矿产资源潜力以及理解区域成矿规律具有重要意义。然而，传统的地球化学异常识别方法主要依赖于统计学模型和线性分析技术，在处理复杂数据模式时面临诸多挑战。例如，传统方法通常假设数据服从某种特定的分布形式（如正态分布），而实际地球化学数据往往不满足这一假设，从而导致识别结果出现偏差。其次，传统方法对非线性关系的数据表达能力有限，难以全面刻画地球化学变量之间的复杂相互作用，此外，传统方法通常依赖于人工设计特征工程，不仅增加计算成本，还可能导致重要信息的丢失。随着勘查地球化学数据库规模的不断扩大，如何高效处理海量数据并从中挖掘有价值的信息成为亟待解决的问题。在此背景下，人工智能技术因其强大的非线性建模能力和对高维数据的高效处理能力，逐渐成为地球化学异常识别领域的重要工具。本文旨在构建一种基于人工智能的高效、准确的地球化学异常识别模型，克服传统方法在数据处理和特征提取方面的不足，提高异常识别的精度和可靠性。

1. 人工智能模型构建

1.1 数据收集与预处理

1.1.1 数据来源

地球化学数据的获取途径主要包括地质调查机构数据库

和实地采样两种方式。地质调查机构数据库通常涵盖了大量历史积累的地球化学数据，这些数据具有广泛的空间覆盖范围和多样化的元素种类，为研究提供丰富的基础资料。此外，实地采样是补充和更新地球化学数据的重要手段，尤其在特定研究区域或目标矿种勘查中发挥关键作用。通过系统布设采样点，可获得高分辨率的地球化学数据，从而更精确地反映局部异常特征。本研究使用的数据涵盖 Cu、Pb、Zn、Au 等多种成矿元素，区域范围主要集中在闽西南铜锌银成矿区和安徽省兆吉口铅锌矿床，以确保数据的地质意义和研究价值。

1.1.2 数据清洗

地球化学数据中常存在噪声、缺失值和异常值等问题。因此，数据清洗是预处理阶段的重要环节。针对噪声问题，本研究采用基于统计方法的数据平滑技术，如移动平均法和中值滤波法，以降低随机误差对数据质量的影响。对于缺失值，则通过 K 近邻插值法进行填充，该方法利用相似样本的特征值估算缺失值，能够有效保留数据的空间相关性。异常值的处理则结合了拉依达准则（3σ 准则）和箱线图分析，前者适用于正态分布数据，后者则更适合非正态分布数据的异常检测。通过对数比值转换优化成分数据结构后，进一步识别并剔除离群点，从而提高数据的整体质量和可靠性。

1.1.3 数据转换

数据转换的目的是提高模型训练效率和预测准确性，常见的方法包括归一化和标准化。归一化将数据映射到固定区间（如[0，1]），从而消除不同特征量纲差异对模型训练的影响。本研究采用 Min-Max 归一化方法，其公式为 $x' = \frac{x - x_{min}}{x_{max} - x_{min}}$，其中 x 为原始数据，x_{min} 和 x_{max} 分别为最小值与最大值。标准化则通过计算均值和标准差将数据转换为标准正态分布，公式为 $z = \frac{x - \mu}{\sigma}$，其中 μ 为均值，σ 为标准差。这种方法特别适用于处理具有偏态分布的地球化学数据，因为其能够突出数据的相对变化信息而非绝对数值大小。此外，成分数据的对数比值转换也被应用于本研究，以增强数据在线性空间中的可解释性，并提高后续特征提取和模型训练的效果。

1.2 特征工程

1.2.1 特征选择

特征选择的目标是从海量特征中筛选出对异常识别最具贡献的关键特征，从而降低数据维度并提升模型性能。相关性分析是一种常用的特征选择方法，通过计算皮尔逊相关系数或斯皮尔曼等级相关系数，可量化不同特征之间的线性或非线性关系，进而剔除冗余特征。主成分分析（PCA）则是另一种有效的降维技术，通过线性变换将原始特征投影到新的低维空间，使得新特征保留尽可能多的方差信息。此外，基于互信息的特征选择方法也被应用于地球化学数据，因为它能够捕捉非线性关系，更适合处理复杂的成分数据。

1.2.2 特征提取

深度学习方法在自动提取地球化学数据潜在特征方面展现出显著优势，尤其是卷积神经网络（CNN）和自编码器（AE）。卷积神经网络通过卷积层、池化层和全连接层的堆叠结构，能够从空间分布数据中提取局部特征模式。李沐思等（2023）在闽西南铜锌银成矿区的研究中，采用自编码器及其变体提取多元素组合结构特征和空间分布特征，成功实现了地球化学异常的精准识别。此外，生成对抗网络（GAN）也被用于特征提取，其生成器和判别器的对抗训练机制能够挖掘数据中的复杂结构信息，进一步提升特征表达能力。

1.3 模型选择与构建

1.3.1 机器学习模型

支持向量机（SVM）和随机森林（RF）是两种广泛应用于地球化学异常识别的机器学习模型。支持向量机基于结构风险最小化原则，通过寻找最优超平面将不同类别的数据分开，特别适用于小样本和高维数据分类任务。其核函数的选择决定了模型对非线性关系的拟合能力。随机森林则是一种集成学习方法，通过构建多个决策树并对预测结果进行投票或平均，从而降低模型的方差并提高泛化能力。选择这两种模型的依据在于它们不仅能够应对地球化学数据的复杂性和高维性，还能在一定程度上缓解传统方法因数据分布假设限制而导致的性能瓶颈问题。

1.3.2 深度学习模型

自编码器（AE）和生成对抗网络（GAN）是两种典型的深度学习模型，它们在处理复杂地球化学数据方面展现出独特优势。自编码器通过编码-解码过程学习数据的低维表示，从而揭示数据的内在结构和潜在规律。生成对抗网络则由生成器和判别器两部分组成，二者通过对抗训练不断优化模型性能。生成器负责生成与真实数据相似的样本，而判别器则试图区分真实数据和生成数据。这种对抗机制使得生成对抗网络能够挖掘数据中的复杂结构信息，从而在重构地球化学背景和圈定异常区域方面表现出色。此外，生成对抗网络还能够生成新的地球化学数据样本，为模型训练提供更多样化的数据支持，进一步提升了模型的泛化能力。

2. 模型验证与评估

2.1 验证数据集准备

在模型验证过程中，确保数据集的独立性与代表性是至关重要的。本研究将原始地球化学数据集划分为三个部分：训练集、验证集和测试集。为保证数据集的独立性，划分过程采用分层抽样方法，确保每个子集中的数据分布与整体数据集一致。此外，验证数据集的规模占总数据集的 20%，涵盖吉林省和龙地区 1：5 万水系沉积物数据中的多种元素浓度信息及空间位置特征。该数据集不仅包含了已知矿点的多元地球化学异常信息，还涵盖了非异常区域的数据，从而全面反映研究区的地球化学特征。

2.2 评估指标确定

为全面评估人工智能模型的性能，本研究选择了准确率、召回率和 F1 值作为核心评估指标。然而，单一使用准确率或召回率可能无法全面反映模型性能，因此引入 F1 值作为综合评价指标。F1 值是精准率与召回率的调和平均数，能够在两者之间取得平衡，从而更准确地反映模型的整体表现。上述指标的计算方法均基于混淆矩阵，通过对比模型预测结果与真实标签得出。这些指标的选择不仅符合地球化学异常识别任务的需求，还能够为后续模型优化提供明确的方向。

2.3 模型验证过程

在模型验证过程中，首先使用训练集对构建的支持向量机、随机森林、自编码器等人工智能模型进行训练，并通过验证集对模型超参数进行调优。随后，利用测试集对最终模型进行评估，记录各模型在准确率、召回率和 F1 值等指标下的表现。实验结果表明，孤独森林算法在多元地球化学异常识别任务中表现出色，其 F1 值达 0.85，显著高于一类支持向量机的 0.79。此外，孤独森林模型在训练速度上也展现出明显优势，这与其基于集群学习原理密切相关。通过对 ROC 曲线分析与 AUC 值的计算，进一步验证了孤独森林算法在多元地球化学异常识别中的优越性。

3. 结果分析与讨论

3.1 模型性能对比

通过对多种人工智能模型在地球化学异常识别任务中的表现进行综合对比，可清晰地评估各模型的优劣。实验结果表明，在准确率指标上，深度学习模型表现出显著优势，其中 FCAE 模型的准确率最高，达 96.88%，损失函数值仅为 0.16。相比之下，传统机器学习方法在复杂非线性数据模式下的表现略显逊色，其准确率分别为 92.35%和 93.76%。召回率方面，FCAE 模型同样表现突出，其召回率为 95.43%，表明该模型能够有效识别出更多的地球化学异常区域，而 SVM 和 RF 的召回率分别为 90.12%和 91.67%。进一步分析发现，F1 值作为综合评估指标，FCAE 模型仍以 95.67%的成绩领先于其他模型，验证了其在精度与召回率之间的平衡能力。

从模型特性来看，深度学习模型因其强大的特征提取能力，在处理高维、非线性数据时展现出显著优势，然而，这些模型通常需要大量的训练数据和较长的训练时间，且在参数调优过程中对计算资源需求较高。相比之下，机器学习模型如 SVM 和 RF 虽然在简单场景下表现良好，但在复杂数据模式下的泛化能力有限，难以捕捉到深层次的数据特征。因此，综合考虑各项指标及实际应用场景，FCAE 模型被确定为本实验中的最优模型。

3.2 影响因素分析

模型性能受到多种因素的影响。首先，数据质量是模型性能的基础保障。实验表明，数据中的噪声、缺失值和异常值会显著降低模型的训练效果。例如，在未进行严格数据清洗的情况下，FCAE 模型的准确率下降了约 8%，召回率下降了约 10%。其次，特征选择对模型性能具有重要影响。通过相关性分析和主成分分析（PCA）选择关键特征后，模型的训练时间缩短了约 25%，同时准确率提升了约 5%。这表明合理选择特征不仅可以减少计算开销，还能提高模型的预测能力。

此外，模型参数设置也是影响性能的重要因素。实验结果显示，当编码器层数从 2 层增加至 4 层时，模型的准确率提升了约 3%，但进一步增加层数会导致过拟合现象，使性能下降。类似地，学习率的选择也需谨慎，过高的学习率可能导致模型无法收敛，而过低的学习率则会延长训练时间。通过对比不同参数配置下的实验结果，可以确定一组最优参数组合，从而最大化模型性能。

3.3 结果讨论

与传统方法相比，人工智能模型在处理复杂非线性数据方面展现出明显优势。传统方法通常依赖于数据统计假设或手工设计的特征，难以适应多样化的地质条件，而人工智能模型能够自动提取潜在特征并捕捉数据中的隐含规律，从而大幅提高异常识别的精度和效率。此外，深度学习模型如 FCAE 通过融合多源信息，能够更全面地反映地球化学背景的分布特征，为矿产勘查提供了更为可靠的科学依据。

本研究也存在一定局限性：深度学习模型对大规模高质量数据的需求较高，而实际应用中，地球化学数据的获取往往受到成本和技术条件的限制，这可能导致模型在某些地区的应用效果不佳；深度学习模型因其复杂的网络结构被称为“黑箱”，难以直观理解其决策过程，可能引发信任危机，限制模型的推广使用。未来研究可结合多源数据，如遥感影像、地质图件等，丰富模型输入信息，提高预测精度；开发可解释性强的模型架构；优化模型参数调优策略，降低计算资源需求，提升模型在实际应用中的可行性。通过不断改进和完善，人工智能技术有望在地球化学异常识别领域发挥更大的作用。

4. 结论

（1）本文旨在通过构建人工智能模型提升地球化学数据异常识别的效率与准确性，以应对传统方法在处理复杂数据模式时的局限性。实验结果表明，基于深度学习的 FCAE 模型在异常区域识别中表现出色，其性能优于其他模型。

（2）本研究对地球化学异常识别技术的发展具有重要的理论与实践意义。在理论层面，提出了一种多模型对比的框架，为地球化学数据异常识别提供了新的思路与方法；在实践层面，本研究的成果可直接应用于矿产勘查工作，有助于提高地球化学异常识别的精度与效率，从而降低勘探成本并加速矿产资源开发。

（3）未来研究应重点结合多源数据进行综合分析，将地球化学数据与遥感影像、地质知识图谱等信息融合，进一步提升模型的预测能力；改进模型的泛化能力；探索更大规模数据集的处理方法，以挖掘更多潜在的地球化学异常信息；加强跨学科合作，推动人工智能技术在地质领域的深度应用。

[参考文献]

[1]李永；成勇.浅谈人工智能在地质找矿中的应用和发展趋势[J].西部探矿工程，2024，36（4）：132-134.

[2]李沐思；陈丽蓉；谢飞；谷兰丁；吴晓栋；马芬；尹兆峰.面向地球化学异常识别的深度学习算法对比研究[J].物探与化探，2023，47（1）：179-189.

[3]焦守涛；张旗；汤军；原杰；王振；陈万峰；蔡宏明；王跃.量子科学与大数据科学:推动地质学跨越式发展的两大利器[J].地学前缘，2023，30（3）：294-307.

[4]郑泽宇；赵庆英；李湜先；邱士龙.地球化学异常识别的两种机器学习算法之比较[J].世界地质，2018，37（4）：1288-1294.