Advances in Geosciences 地球科学前沿, 2024, 14(7), 923-934
Published Online July 2024 in Hans. https://www.hanspub.org/journal/ag
https://doi.org/10.12677/ag.2024.147086

Hans汉斯

基于残差图神经网络的成矿远景区预测研究

张　鑫，薛子如，高　乐*

五邑大学电子与信息工程学院，广东　江门

收稿日期：2024年2月28日；录用日期：2024年7月2日；发布日期：2024年7月10日

摘　要

研究针对地球化学元素数据的成矿远景区预测问题，提出了一种基于残差图神经网络的深度学习框架。文章以广东省庞西垌研究区作为案例研究对象，针对地质数据稀缺、数据不平衡和深度学习模型构建难度等问题，文章采取了以下关键步骤：首先，对地球化学元素数据进行了“去闭合化”处理，以适应后续的分析；其次，针对矿区样本不足的问题，文章引入了生成对抗网络来进行数据增强，并证明了其有效性；文章提出了一种自适应阈值的皮尔森相关系数方法，将地球化学元素数据构建为图数据；最后，文章提出一种基于残差图神经网络模型，对数据进行特征提取和分类。实验结果与传统机器学习方法和其他图神经网络方法相比，文章方法在成矿远景区预测任务中表现出显著的优势。

关键词

图神经网络，成矿远景区预测，生成对抗网络，地球化学元素

Research on Prediction of Mineralized Prospective Areas Based on Residual Graph Neural Network

Xin Zhang, Ziru Xue, Le Gao*

School of Electronics and Information Engineering, Wuyi University, Jiangmen Guangdong

Received: Feb. 28th, 2024; accepted: Jul. 2nd, 2024; published: Jul. 10th, 2024

Abstract

This study proposes a deep learning framework based on residual graph neural network for pre-

*通讯作者。

文章引用: 张鑫, 薛子如, 高乐. 基于残差图神经网络的成矿远景区预测研究[J]. 地球科学前沿, 2024, 14(7): 923-934.
DOI: 10.12677/ag.2024.147086

diction of mineralized prospective areas using geochemical element data. This paper takes the Pangxidong research area in Guangdong Province as the case study object. In response to issues such as scarce geological data, imbalanced data, and difficulty in constructing deep learning models, this paper has taken the following key steps: first, this paper has "de closed" the geochemical element data to adapt to subsequent analysis; Secondly, to address the issue of insufficient samples in mining areas, this paper introduced Generative Adversarial Networks (GAN) for data augmentation and proved their effectiveness; This paper proposes an adaptive threshold Pearson correlation coefficient method to construct geochemical element data into graph data; Finally, this paper proposes a residual graph based neural network model for feature extraction and classification of data. Compared with traditional machine learning methods and other graph neural network methods, our method shows significant advantages in prediction of mineralized prospective areas.

Keywords

Graph Neural Network, Prediction of Mineralized Prospective, Generative Adversarial Network, Geochemical Elements

1. 引言

随着人类社会的进步与科技的飞速发展，对于矿产资源的需求量和重视程度与日俱增，然而，当下矿产资源的勘查的形式已经发生了巨大变化，找矿环境由简单变得越来越复杂。随着计算机科学的高速发展，人工智能技术被广泛应用于地球科学领域，传统的统计方法寻找成矿远景区逐渐转变为基于人工智能方法的智能找成矿远景区[1]-[3]。由于地质条件的复杂性，地质数据的非线性特征较强，机器学习、深度学习算法能更好地刻画矿化点和证据要素间的复杂非线性关系，比统计方法的适用性更强[4]。基于机器学习的分类器和回归模型能够根据地质、地球物理和遥感数据的特征，自动学习并建立预测模型，从而识别出具有潜在矿化可能性的区域[5] [6]。Carranza 等探索了随机森林在成矿预测领域的应用效果，相较于证据权法、证据置信和逻辑回归建模，随机森林方法表现出了更优秀的拟合和泛化性能，并开展了菲律宾 Baguio 地区金矿成矿预测[7]；Ghezelbash 等人使用遗传算法优化 K-means 和 SVM 算法预测伊朗西北部 Varzaghan 地区斑岩型铜矿床[8]；毕晨曦等人融合了动力学模拟结果的机器学习对安徽铜山铜矿的找矿潜力进行三维定量预测，测试集和验证集上的 AUC 值分别能达 0.998 和 0.999，其前 7%的高概率区基本能包含全部已知矿体[9]。深度学习算法，如卷积神经网络(CNN)和循环神经网络(RNN)，能够通过多层次的特征提取和学习，对复杂的地质关系进行建模，进一步提高了矿靶区预测的准确性[10]。Li 等提出了一种基于卷积神经网络算法的方法，并采用迁移学习方法减少了该地区有限数量的已知矿床和矿点的影响、加快了收敛速度、提高了模型的准确性[11]；Xu 等构建深度回归神经网络绘制甘肃省大桥金矿的矿产前景图谱，神经网络使用多源数据进行训练，包括研究区域的地质、地球物理和地球化学数据，揭示了矿产前景图与地质、地球物理和地球化学特征之间的复杂关系，提高了预测结果[12]。黄勇杰等提出一种多尺度特征交互框架，使用元网络为多尺度分类网络生成卷积权重并使用自蒸馏挖掘多尺度分类网络中的隐知识用于预测[13]。

尽管目前众多学者在智能找成矿远景区方向取得了很大的成功，但是成矿远景区预测仍面临以下两

大挑战：

1) 矿区样本不足。由于已知矿床稀少，训练数据不均衡，容易导致深度学习模型的泛化能力低，最终影响模型分类效果。

2) 深度学习网络模型构建困难。基于深度学习网络模型进行成矿远景区预测多依赖于 CNN，卷积核只能处理规则分布的网格或栅格图片。然而，由于复杂多样的地质环境，在野外采集的地球化学样品在空间的位置往往是不规则的，为了处理不规则区域内不规则分布的样品的地球化学数据集(非欧几里得数据)，必须应用内插法从原始样品中生成一个规则的网格化数据集，这种插值不可避免地将不确定性引入到数据中，降低了模型学习的准确性。图神经网络的出现给深度学习建模带来了一种新思路，图神经网络能够有效地处理非规则分布的样品地球化学数据集，同时能够充分利用样品的空间分布信息，避免了数据插值所引入的不确定性。

本研究利用广东庞西垌研究区的地球化学元素数据作为实验数据集，提出一种基于残差图卷积网络的网络模型。本文的主要贡献如下：

1) 本文提出一种基于残差图神经网络的成矿远景区预测模型，对图结构数据进行整图特征提取。我们以广东庞西垌研究区的水系沉淀物为实验数据集进行验证，实验结果表明本文模型在准确度、敏感性、F1 分数均比现有模型有所提升。表明本文模型在地质领域的应用潜力。

2) 本文针对矿区含矿样本少的问题，使用生成对抗网络来对实验数据进行扩充，最终使样本数据达到平衡，减少模型过拟合的风险，使模型更具泛化性能。

3) 在深度学习模型构建过程中，地球化学元素的高维度数据和图数据的问题是主要难题。为了克服这一难题，本文采用一种创新的方法，我们利用水系沉淀物中各元素之间的相关性来构建图数据，这样既可以避免常用的数据插值所带来的不确定性又能更准确地捕捉地球化学元素之间的关联。这一方法为模型提供了更丰富的信息，从而提高了成矿远景区预测准确性。

2. 实验数据处理

本研究的实验数据来源于中国华南地区广东省庞西垌研究区，图 1 所示为广东庞西垌研究区的地质图概况，其中蓝点表示已知矿床、矿点。化探数据为该研究区 1:50,000 水系沉积物资料中提取的地球化学元素，水系沉积物采样面积为 1694 km^2，水系沉淀物样品共计 7236 件，平均采样密度为 4.27 个/km^2，从水系沉积物测量样品分析 Au、B、Sn、Cu、Ag、Ba、Mn、Pb、Zn、As、Sb、Bi、Hg、Mo、W、F 共 16 种化学元素。

2.1. 地球化学数据变换处理

地球化学数据是典型的成分数据，即各个元素含量的总和为一定值[14]。然而，成分数据会产生闭合效应，导致各元素之间相互制约表现出伪相关关系[15]，使得一些统计方法在对地球化学数据进行分析时会忽略数据的“定和”约束，导致产生虚假的相关性。针对以上问题，Aitchison 基于成分数据的比值不受“定和”约束影响以及比值的对数通常服从正态分布的特点提出了中心对数比变换(centered log ratio transformations, clr) [16]，该方法以变量的几何平均值作为分母，用其它变量除以该变量再取自然对数。

根据图 2 可以观察到原始数据在空间尺度上存在较大的差异，数据分布也较为离散，且多数元素存在大量高值异常点。由于原始数据之间量级和尺度的差异较大，累计密度曲线未在此处展示。然而，经过 clr 变换的数据，各元素的空间分布尺度差异显著减小，数据箱图的分布更加均匀，对应的密度曲线也近似呈现单峰/多峰正态分布。

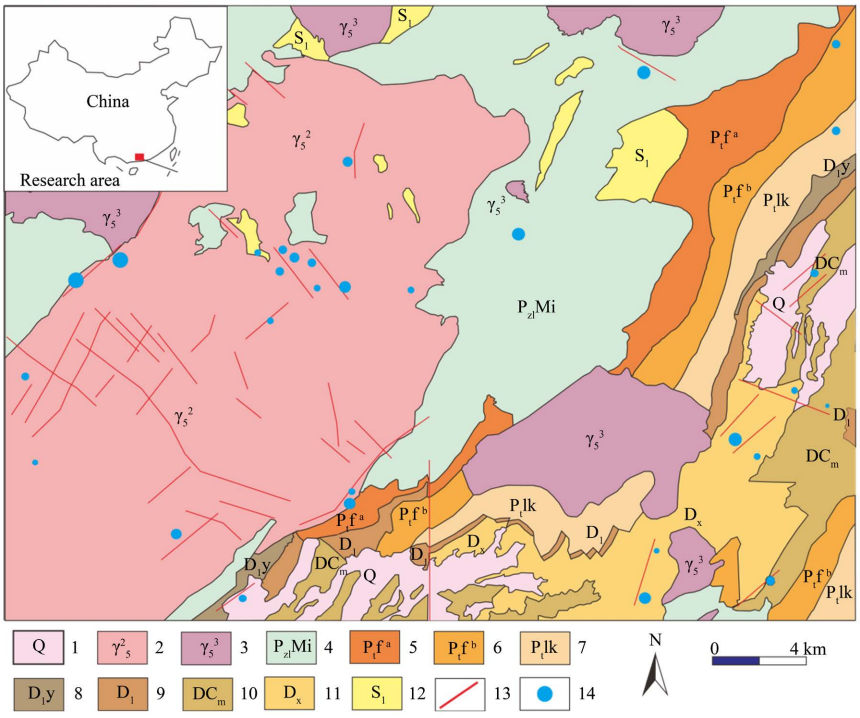

Figure 1. Simple geological map of the Pangxidong study area, Guangdong. (1: Quaternary; 2: Early Yanshanian granite; 3: Late Yanshanian granite; 4: Upper Proterozoic Migmatite; 5: Lower Member of Middle-Upper Proterozoic Fengdongkou Formation; 6: Upper Member of Middle-Upper Proterozoic Fengdongkou Formation; 7: Middle-Upper Proterozoic Lankeng Formation; 8: Devonian Yangxi Formation; 9: Devonian Laohutou Formation; 10: Devonian-Carboniferous Maozifeng Formation; 11: Devonian Xindu Formation; 12: Silurian Liantan Formation; 13: Faults; 14: Deposit)

图 1. 广东庞西垌研究区地质图 (1：第四纪；2：燕山早期花岗岩；3：燕山晚期花岗岩；4：加里东期混合岩；5：中晚元古代丰垌口组下段；6：中晚元古代丰垌口组上段；7：中晚元古代兰坑组；8：泥盆纪杨溪组；9：泥盆纪天子峰组；10：晚泥盆世－早石炭世帽子峰组；11：泥盆纪信都组；12：早志留纪；13：断层；14：已知矿床、矿点)

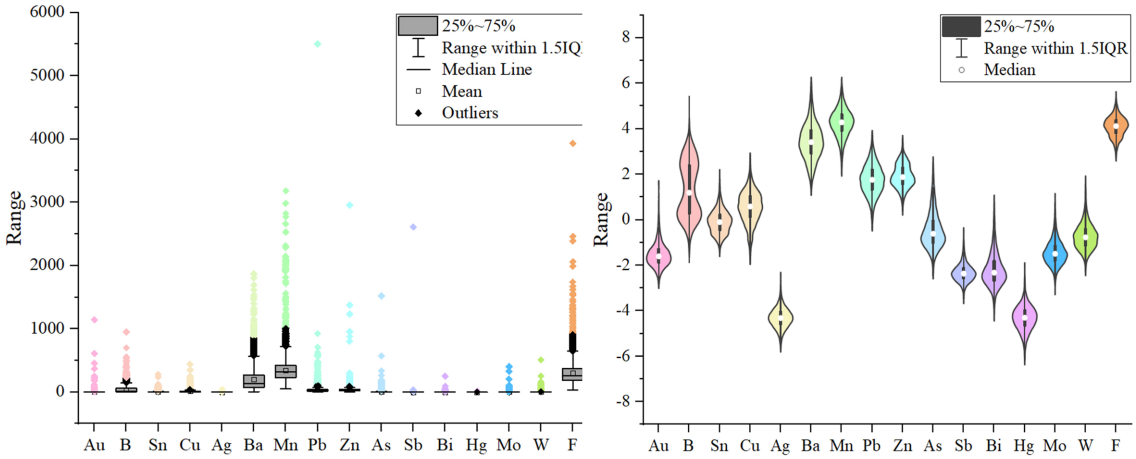

Figure 2. Distribution of data before and after transformation (a) Box plot of raw data, (b) Violin plot of clr-transformed data

图 2. 数据变换前后的分布情况(a) 原始数据箱图，(b) clr 变换数据小提琴图

2.2. 数据样本不平衡处理

矿产的分布在地质空间中常表现很强的不均衡性，导致在一定的研究区域内，含矿数据样本的数量

只占总样本量的很小一部分。成矿远景区预测数值建模所涉及的数据是典型的不平衡数据集，即非含矿类的样本要比含矿类的样本数目多得多。使用对抗生成网络对少数类样本进行扩充。

生成对抗网络(Generative Adversarial Networks, GAN)是一种深度学习模型[17]，如图 3 所示，GAN 由两个神经网络共同组成：生成器(Generator)和判别器(Discriminator)。生成器旨在从随机噪声中生成逼真的数据样本，而判别器是判别输入是生成器生成的样本还是真实数据样本得二分类器。这两个网络相互博弈、相互对抗，通过不断的训练迭代，使生成器逐渐生成更逼真的样本，同时判别器变得更加准确。最终，生成器能够生成与真实数据样本相似度很高的新样本。

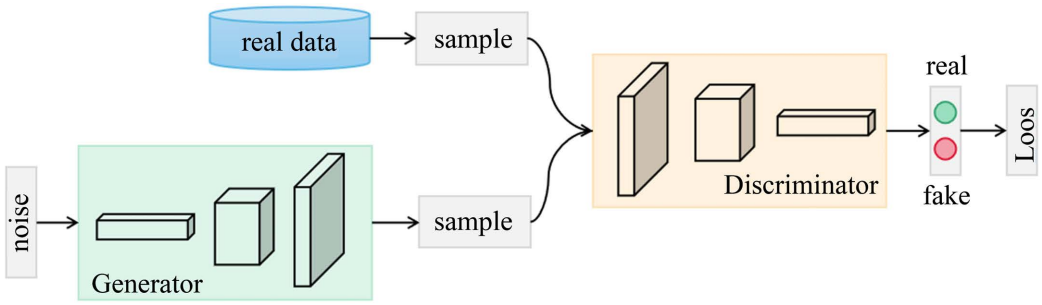

Figure 3. GAN network model diagram
图 3. GAN 网络模型图

2.3. 地球化学元素数据的图构造

无向连通图的定义为 G = (V, E)，其中 V，E 分别表示相应的节点集和边集，$v_i, v_j \in V$ 表示图中的节点，$(v_i, v_j) \in E$ 表示图中的边。我们考虑地球化学元素的相关性在其扩散、迁移、聚集等过程中普遍存在。可以运用统计学原理来描述变量(元素)之间内在联系程度的强弱，同时用定量(−1, 1)的数据指标来表征变量(元素)间的相关程度。相关系数的计算公式为：

$$r_{(u,v)} = \frac{Cov(v_i, v_j)}{\sigma_{v_i}\sigma_{v_j}} \tag{1}$$

其中，$Cov(v_i, v_j)$ 为 v_i 与 v_j 的协方差，σ_{v_i} 与 σ_{v_j} 分别为 v_i 与 v_j 的标准差。

$r > 0$ 表明两种元素正相关，$r < 0$ 则表示两种元素负相关。r 的绝对值与 1 越接近，则表示两变量间的线性相关程度越高。

当我们筛选出具有相关性高于阈值的地球化学元素对时，我们随即建立这些元素对之间的联系，从而形成了一个图数据结构。在这个图中，每个地球化学元素都被视为图中的一个节点(node)，而那些相关性较高的元素对则被视为图中的一条边(edge)。我们引入了一个超参数 α 来灵活地控制图数据的构建。这个超参数的作用是在生成图数据的边时调整阈值，以确保节点之间能够保持一定的连接，从而提高图分类任务的准确度。

3. 方法

3.1. 残差图神经网络

本文使用图神经网络来对地球化学元素数据进行特征提取及分类，我们使用图卷积神经网络(Graph Convolutional Network, GCN)和图注意力网络(Graph Attention Networks, GAT)作为基线模型[18] [19]，现有的大部分图神经网络在模型层数加深时会出现节点信息过平滑，所有节点的特征会朝向一个相同的值，

还会导致梯度消失出现的风险，所以一般图卷积神经网络层数不超过 3 层，浅层图卷积无法使它们从高阶邻居提取信息，因此性能受到限制。如图 4 所示本文在图卷积网络中加入残差网络的思想[20]，在邻接矩阵的更新过程中加入一定权重的上层邻接矩阵，通过残差连接，节点的浅层特征信息将会在图卷积传播过程中一直保留下来以提高分类模型的稳定性，缓解过拟合等问题。

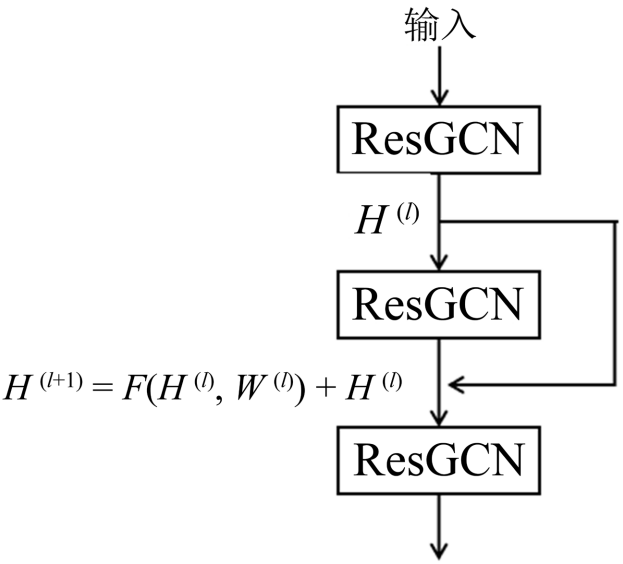

Figure 4. ResGCN network model diagram
图 4. ResGCN 网络模型图

当 l > 0 时，残差图神经网络的节点更新公式如下式(2)：

$$H^{(l+1)} = F\left(H^{(l)}, W^{(l)}\right) + H^{(l)} \tag{2}$$

$H^{(l)}$ 是直接映射，$F\left(H^{(l)}, W^{(l)}\right)$ 是对 $H^{(l)}$ 的图卷积操作。

3.2. 全局池化

该模型使用了全局池化结构，适用于处理不同结构的图数据。使用全局池化常用的求和方法聚合各个节点在每层的特征表示，得到整个图的特征表示 Z_g，最后通过 softmax 激活函数得到分类结果：

$$Z_g = \text{READOUT}\left[H_i^1, H_i^1, \cdots, H_i^j\right], \forall i \in g \tag{3}$$

其中 j 表示层数，任意的节点 i 都属于图 g，$G = \left\{G_1, G_2, \cdots, G_g\right\}$，所有的图 g 都属于数据集 G。

3.3. 目标函数

我们使用了交叉熵损失函数作为优化的目标函数。预测概率和交叉熵损失函数可表示为：

$$\hat{Y} = \text{softmax}\left(W \cdot Z_g + b\right) \tag{4}$$

$$L_{class} = -\sum_{g \in G} \sum_{c=1}^{C} Y_{gc} \ln \hat{Y}_{gc} \tag{5}$$

其中 W 和 b 分别为权重矩阵和偏差矩阵，整图表示 Z_g 经过线性变换和 softmax 激活函数得到预测概率 $\hat{Y}$。C 表示图的标签类型数量，Y_{gc} 表示图 g 实际的标签类型，$\hat{Y}_{gc}$ 表示图 g 预测的标签类型，模型通过最小化交叉熵损失函数对参数进行拟合，得到图的分类结果。

4. 实验

4.1. 实验设置

实验平台为Tesla V100 32GB GPU，深度学习框架为Pytorch。GCN模块，Learning Rate设置为0.001，dropout设置为0.5，使用Adam作为参数优化器，损失函数为交叉熵损失函数。使用十折交叉验证方案计算分类性能。

4.2. 评价指标

为了验证我们方法的可行性，我们从准确度、敏感度、特异度和F1分数对其进行评估。计算公式如下：

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN} \tag{6}$$

$$\text{Sensitivity} = \frac{TP}{TP + FN} \tag{7}$$

$$\text{Specificity} = \frac{TN}{TN + FP} \tag{8}$$

$$\text{F1} - \text{score} = \frac{2 \times P \times R}{P + R}, P = \frac{TP}{TP + FP}, R = \frac{TP}{TP + FN} \tag{9}$$

其中，*TN*表示正确分类的无矿样本，*FP*表示被错误分类的无矿样本，*FN*表示被错误分类的有矿样本，*TP*表示被正确分类的有矿样本。

4.3. 交叉验证

为了更好地评估我们的模型，我们采用十折交叉验证。我们将全部数据分为10个等份，每个等份包含相等数量的样本。在每一轮的验证过程中，我们选择其中一份作为验证集，而将其他9份作为训练集。这个过程将重复进行10次，确保每个等份都曾被用作验证集。这样，我们可以综合考虑不同的验证集组合，更全面地评估我们模型的性能。最终，我们将十轮验证的结果进行平均，以获得最终的性能指标，这有助于减小随机性对评估结果的影响。实验结果如表1所示。

Table 1. Cross-validation results
表1. 交叉验证结果

方法	评价指标	1	2	3	4	5	6	7	8	9	10	平均值
ResGCN	准确度	94.78	94.9	95.01	94.38	96.02	94.72	96.15	95.72	95.43	96.29	95.34
	敏感度	87.93	89.71	88.51	87.25	90.72	88.96	89.71	89.85	90.73	89.93	89.33
	特异度	99.91	99.95	99.92	99.93	99.91	99.86	99.93	99.82	99.88	99.89	99.9
	F1分数	93.52	94.45	93.84	94.85	93.81	95.45	94.73	93.88	95.18	94.59	94.43
ResGAT	准确度	95.01	95.96	95.78	96.58	95.85	96.61	96.15	95.56	95.93	95.77	95.92
	敏感度	89.43	90.54	88.87	90.87	91.06	88.79	89.86	90.54	91.02	91.12	90.21
	特异度	99.85	99.89	99.93	99.86	99.91	99.93	99.92	99.91	99.87	99.93	99.9
	F1分数	94.32	94.83	94.62	95.43	94.92	95.33	94.76	95.52	94.79	94.88	94.94

表2表示使用ResGCN和ResGAT对广东庞西垌地区进行矿区分类的十折交叉验证结果，可以看出

两个模型在负类别(无矿区)的分类上都表现得非常出色，几乎没有将负类别样本错误分类为正类别(矿区)的情况，而它们对于正类别(矿区)的分类略差一点，但也能识别出大部分的有矿样本，造成这种结果的原因可能是因为在原始数据集中有矿样本较少而经过数据增强后数据达到平衡，但经过数据增强的有矿样本的特征不够明显或不够充分。模型 ResGAT 的效果略好于模型 ResGCN，这是由于在 ResGCN 中每个节点的邻居节点都平等地参与信息聚合，而 ResGAT 通过引入注意力权重，使每个节点能够以不同的权重聚合其邻居的信息，这允许 ResGAT 关注对当前节点更重要的邻居，从而更好地捕获图的结构。

4.4. 对比实验

在之前的实验中，直接使用地球化学元素来进行成矿远景区预测时，通常使用随机森林、SVM、Kmeans、KNN、AE 等经典算法进行分类实验，经典的图分类网络有 GIN、DiffPool、GraphSAGE，由于地质数据的特殊性不能获得，我们使用这些方法在自己的数据集上进行实验，实验结果如表 2 所示。

Table 2. Experimental results of different models
表 2. 不同模型实验结果

方法		准确度	敏感度	特异度	F1 分数
传统方法	SVM	73.2	69.72	75.33	69.63
	随机森林	72.26	66.50	76.91	68.56
	Kmeans	77.03	76.10	77.81	76.79
	KNN	76.32	75.59	77.04	75.88
	AE	80.65	76.66	84.35	77.63
图神经网络方法	GIN	90.89	85.77	95.26	90.03
	DiffPool	90.08	85.36	94.99	89.01
	GraphSAGE	89.37	83.75	94.73	88.62
本文方法	ResGCN	95.34	89.33	99.9	94.43
	ResGAT	95.92	90.21	99.9	94.94

尽管这些算法在许多领域都表现出色，但在由于地质数据的特殊性，这些数据可能包含了复杂的地质结构和多种地质元素的信息，传统算法难以充分捕捉这些复杂性。从表 2 中的实验结果可以看出，在这个特定的成矿远景区预测任务中，这些传统算法的性能相对较低，其他图神经网络的效果虽然比传统方法的效果好，但还是低于我们所提出的 ResGCN 和 ResGAT，说明基于残差图神经网络方法的在成矿远景区预测有着一定的优势。

5. 讨论

5.1. α 对分类性能的影响

我们采用了超参数 α 作为阈值，这个阈值用于判断元素对之间的相关性是否显著。计算皮尔森相关系数后，我们将其与 α 进行比较。如果相关系数的绝对值大于 α，那么我们认为这两个元素之间存在显著的相关性，于是在图的邻接矩阵中创建一条边来连接它们；反之，如果相关系数不足以表明显著的相关性，则我们不在图中创建对应的边。不同 α 取值的选择直接影响着邻接矩阵的构建，进而对图神经网络的输入产生深远的影响，进而对分类效果造成重要影响。如图 5 所示，我们呈现了在不同 α 取值($\alpha = \{0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8\}$)下的分类准确度(ACC)。

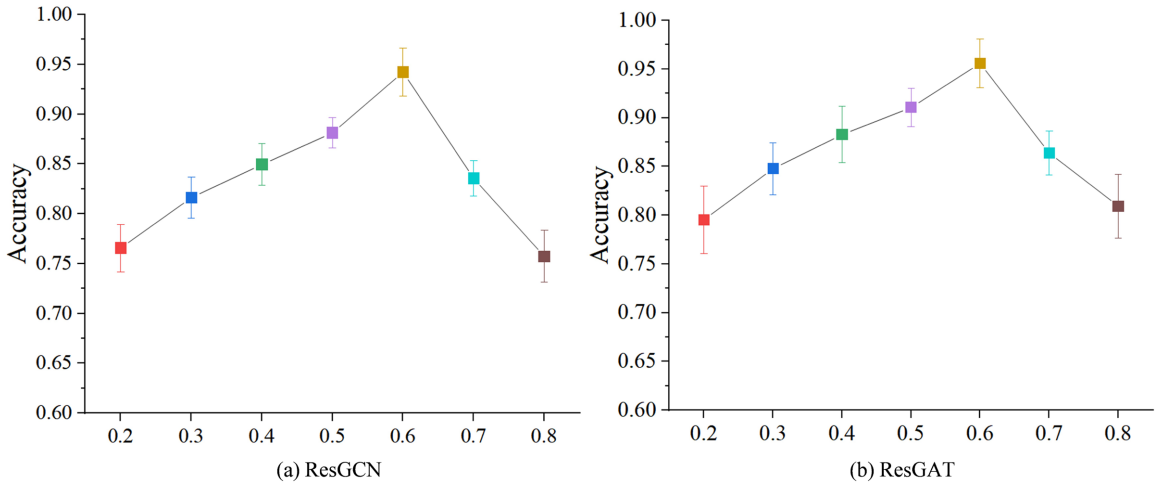

Figure 5. Accuracy for different values of α
图 5. 不同 α 取值的准确度

α 取 0.6 的选择显示出对于 ResGCN 和 ResGAT 模型在分类任务中表现最佳。这可能是因为在这个 α 取值下，模型能够保留关键的地球化学元素相关性，同时避免引入过多的噪声信息。因此，这个 α 值的选择在保持信息丰富性的同时，有助于提高分类的准确性。

5.2. 不同数据增强方法对分类性能的影响

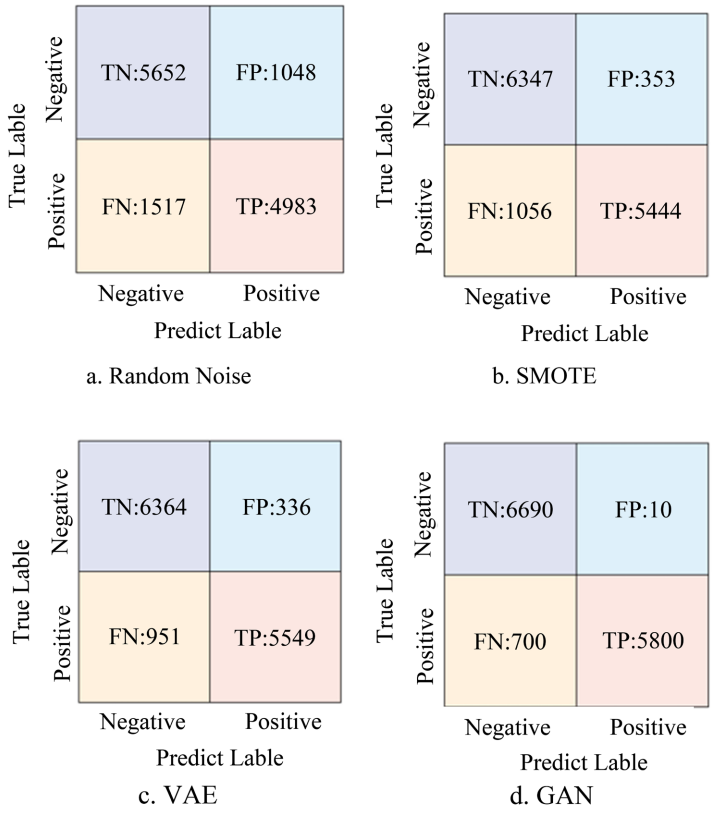

Figure 6. Confusion matrix under various enhancement techniques
图 6. 各种增强技术下的混淆矩阵

在成矿远景区预测任务中，由于矿区样本数量远远少于无矿样本，这导致了数据不平衡的情况。数据不平衡可能会对模型的性能产生不利影响，因为模型可能会倾向于对数量众多的无矿样本进行分类，而对矿区样本的分类性能较差。因此在进行模型训练之前要进行数据增强操作，常用的数据增强操作有直接添加噪声、SMOTE、VAE、GAN 等方法，本文采用上述方法对数据集进行处理，然后进行分类，并获得如图 6 所示的混淆矩阵，其中 *TN* 表示正确分类的无矿样本，*FP* 表示被错误分类的无矿样本，*FN* 表示被错误分类的有矿样本，*TP* 表示被正确分类的有矿样本。

综合分析了不同的数据增强方法在本次实验中的表现，可以得出明显的结论。首先，采用直接添加噪声的方法，虽然能够识别矿区的样本，但同时也增加了误分类无矿区样本的风险，导致了相对较高的假正例(FP)。SMOTE 方法在减少假正例方面表现出色，但相对较高的假负例(FN)可能会导致一些矿区样本被错误分类。与之相比，VAE 和 GAN 方法在各方面都表现出色。它们不仅降低了假正例和假负例的数量，还提高了分类的准确性。尤其是 GAN 方法，其在减少 FP 和 FN 方面表现出色，这意味着它可以更好地区分矿区和无矿区的样本。因此，从这些实验结果来看，GAN 是更适合本次成矿远景区预测实验的数据增强方法，它能够显著提高模型的分类性能，为成矿远景区预测任务提供了更可靠的解决方案。

5.3. 成矿预测远景区圈定

在对广东庞西垌研究区的地球化学元素数据进行特征提取和分类后，我们认为预测为有矿的数据密度超过 5 为 A 类成矿预测远景区，小于等于 5 为 B 类成矿预测远景区。如图 7 所示，蓝色圆点为已知矿区，红色三角形为预测为有矿的采样点，黑色实线为圈定的 A 类成矿预测远景区，黑色虚线为圈定的 B 类成矿预测远景区。我们在广东庞西垌研究区共圈定成矿预测远景区 19 个，其中有 10 个 A 类成矿预测远景区，9 个 B 类成矿预测远景区，圈定的成矿远景区完全覆盖已知矿区。结果说明基于残差图神经网络的成矿远景区预测模型，对广东庞西垌研究区成矿远景区的确定有着很高的准确性，为研究区下一步矿区勘查工作提供了重要依据，也为深部矿、隐伏矿的成矿远景区预测提供了新的思路和方法。

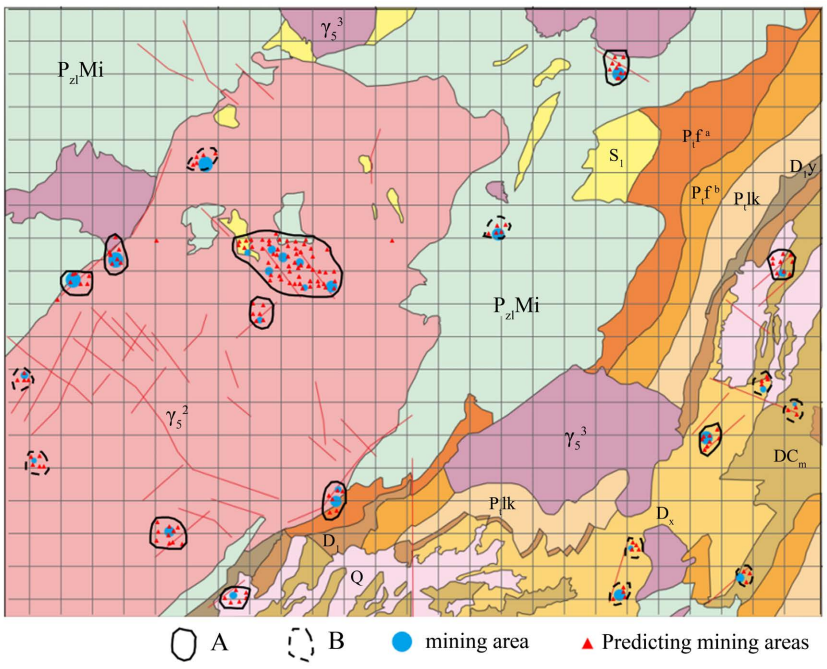

Figure 7. Mineralization prediction prospective area circled
图 7. 成矿预测远景区圈定

6. 结论

本文的研究旨在应对矿区样本稀缺、深度学习模型构建困难等问题，提出了一种基于残差图神经网络的深度学习框架，以广东省庞西垌研究区为案例进行成矿远景区的智能预测。通过与其他方法的比较，研究结果强调了该方法的有效性。主要结论包括：

1) 数据预处理：针对地球化学元素数据的典型成分特点，本研究在使用数据前进行了“去闭合化”处理，采用 clr 方法对原始数据进行了有效处理，以更好地适应后续的分析和建模过程。

2) 数据不平衡问题：鉴于矿区样本较少的挑战，研究采用了 GAN (生成对抗网络)对少数类样本进行扩充。通过与其他数据扩充方法进行对比，实验证明了 GAN 数据增强方法的有效性，有助于解决不平衡数据问题。

3) 深度学习构建挑战：针对深度学习模型构建的难题，研究提出了一种基于自适应阈值的皮尔森相关系数法，将地球化学元素数据构建成图数据。随后，利用残差图神经网络进行特征提取和分类。通过与传统机器学习方法以及其他图神经网络方法的对比，结果表明残差图神经网络在成矿远景区预测任务中表现出了显著的有效性，为解决深度学习建模难题提供了新的思路。

综上所述，本研究在数据预处理、数据增强和深度学习模型构建方面提出了一系列有效的方法，并在实际案例中展示了它们的效用，为成矿远景区预测任务提供了更可靠的解决方案。

参考文献

[1] Sun, T., Li, H., Wu, K., Chen, F., Zhu, Z. and Hu, Z. (2020) Data-driven Predictive Modelling of Mineral Prospectivity Using Machine Learning and Deep Learning Methods: A Case Study from Southern Jiangxi Province, China. *Minerals*, **10**, Article 102. https://doi.org/10.3390/min10020102

[2] Hsieh, C., Zheng, K., Lin, C., Mei, L., Lu, L., Li, W., *et al*. (2021) Automated Bone Mineral Density Prediction and Fracture Risk Assessment Using Plain Radiographs via Deep Learning. *Nature Communications*, **12**, Article No. 5472. https://doi.org/10.1038/s41467-021-25779-x

[3] Liu, C., Wang, W., Tang, J., Wang, Q., Zheng, K., Sun, Y., *et al*. (2023) A Deep-Learning-Based Mineral Prospectivity Modeling Framework and Workflow in Prediction of Porphyry-Epithermal Mineralization in the Duolong Ore District, Tibet. *Ore Geology Reviews*, **157**, Article ID: 105419. https://doi.org/10.1016/j.oregeorev.2023.105419

[4] Abedi, M., Norouzi, G. and Fathianpour, N. (2013) Fuzzy Outranking Approach: A Knowledge-Driven Method for Mineral Prospectivity Mapping. *International Journal of Applied Earth Observation and Geoinformation*, **21**, 556-567. https://doi.org/10.1016/j.jag.2012.07.012

[5] Liu, Y., Cheng, Q., Xia, Q. and Wang, X. (2014) Multivariate Analysis of Stream Sediment Data from Nanling Metallogenic Belt, South China. *Geochemistry*: *Exploration*, *Environment*, *Analysis*, **14**, 331-340. https://doi.org/10.1144/geochem2013-213

[6] 左仁广. 基于数据科学的矿产资源定量预测的理论与方法探索[J]. 地学前缘, 2021, 28(3): 49-55.

[7] Carranza, E.J.M. and Laborte, A.G. (2015) Data-Driven Predictive Mapping of Gold Prospectivity, Baguio District, Philippines: Application of Random Forests Algorithm. *Ore Geology Reviews*, **71**, 777-787. https://doi.org/10.1016/j.oregeorev.2014.08.010

[8] Ghezelbash, R., Maghsoudi, A., Shamekhi, M., Pradhan, B. and Daviran, M. (2022) Genetic Algorithm to Optimize the SVM and K-Means Algorithms for Mapping of Mineral Prospectivity. *Neural Computing and Applications*, **35**, 719-733. https://doi.org/10.1007/s00521-022-07766-5

[9] 毕晨曦, 刘亮明, 周飞虎. 融合动力学模拟的机器学习三维成矿预测: 以安徽铜山铜矿为例[J/OL]. 大地构造与成矿学: 1-16. https://doi.org/10.16539/j.ddgzyckx.2023.01.104, 2024-02-28.

[10] Puzyrev, V., Zelic, M. and Duuring, P. (2023) Applying Neural Networks-Based Modelling to the Prediction of Mineralization: A Case-Study Using the Western Australian Geochemistry (WACHEM) Database. *Ore Geology Reviews*, **152**, Article ID: 105242. https://doi.org/10.1016/j.oregeorev.2022.105242

[11] Li, H., Li, X., Yuan, F., Jowitt, S.M., Zhang, M., Zhou, J., *et al*. (2020) Convolutional Neural Network and Transfer Learning Based Mineral Prospectivity Modeling for Geochemical Exploration of Au Mineralization within the Guandian-Zhangbaling Area, Anhui Province, China. *Applied Geochemistry*, **122**, Article ID: 104747. https://doi.org/10.1016/j.apgeochem.2020.104747

[12] Xu, Y., Li, Z., Xie, Z., Cai, H., Niu, P. and Liu, H. (2021) Mineral Prospectivity Mapping by Deep Learning Method in Yawan-Daqiao Area, Gansu. *Ore Geology Reviews*, **138**, Article ID: 104316. https://doi.org/10.1016/j.oregeorev.2021.104316

[13] 黄勇杰, 高乐, 杨田, 等. 基于多尺度特征和元学习的智能预测找矿靶区实验研究[J]. 计算机应用研究, 2022, 39(6): 1772-1778.

[14] Bronstein, M.M., Bruna, J., LeCun, Y., Szlam, A. and Vandergheynst, P. (2017) Geometric Deep Learning: Going Beyond Euclidean Data. *IEEE Signal Processing Magazine*, **34**, 18-42. https://doi.org/10.1109/msp.2017.2693418

[15] Karacan, C.Ö., Erten, O. and Martín-Fernández, J.A. (2023) Assessment of Resource Potential from Mine Tailings Using Geostatistical Modeling for Compositions: A Methodology and Application to Katherine Mine Site, Arizona, USA. *Journal of Geochemical Exploration*, **245**, Article ID: 107142. https://doi.org/10.1016/j.gexplo.2022.107142

[16] Aitchison, J. (1982) The Statistical Analysis of Compositional Data. *Journal of the Royal Statistical Society Series B*: *Statistical Methodology*, **44**, 139-160. https://doi.org/10.1111/j.2517-6161.1982.tb01195.x

[17] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., *et al*. (2020) Generative Adversarial Networks. *Communications of the ACM*, **63**, 139-144. https://doi.org/10.1145/3422622

[18] Wu, Z., Pan, S., Chen, F., Long, G., Zhang, C. and Yu, P.S. (2021) A Comprehensive Survey on Graph Neural Networks. *IEEE Transactions on Neural Networks and Learning Systems*, **32**, 4-24. https://doi.org/10.1109/tnnls.2020.2978386

[19] Veličković, P., Cucurull, G., Casanova, A., *et al*. (2017) Graph Attention Networks. arXiv: 1710.10903.

[20] 李钢, 陈太兵, 杨之博, 等. MBRNet: 融合残差连接的多分支手写字符识别网络[J/OL]. 计算机工程与应用: 1-12. http://kns.cnki.net/kcms/detail/11.2127.tp.20240104.1338.036.html, 2024-02-28.