基于岩石化学成分的火成岩分类研究

郑祎杰（组长）　吴冰玥　徐政委 郭宇欣　赵兴元　段小丫　林诚

引言

本文用于研究火成岩的化学变化，探讨不同类型火成岩（如花岗岩、玄武岩等）在化学成分、矿物组成、矿物结构等方面的差异和变化规律，反映火成岩形成的地质环境和地质过程；研究火成岩地化特征在采矿领域的应用，为矿业勘查和开发提供重要依据。

I.　数据集介绍

一、 数据集

该数据集来源于机器学习、数据、代码资源网站 Kaggle [①]，用于了解火成岩基本类型、基本性质与特征；数据集包含不同类型岩石的地球化学变化与成分（SiO_2、TiO_2、$Al2O_3$、Fe_2O_3 等）；该数据集的目标是按岩石化学成分进行数据分类——创建一个能够根据岩石化学识别岩石类型的模型。

二、 原始数据预览

rock_name	long	lat	SiO2n	TiO2n	Al2O3n	FeO*n	MnOn	MgOn	CaOn	Na2On	K2On	P2O5n
Basalt	-122.585	46.23	47.77	1.33	15.38	9.59	0.17	10.76	11.28	2.69	0.62	0.41
Basalt	-122.5806	46.2436	47.94	1.28	15.22	9.6	0.18	10.87	11.39	2.17	0.96	0.38
Basalt	-122.5925	46.2036	51.5	1.12	16.04	8.77	0.15	9.49	9.8	2.64	0.33	0.16
Basalt	-122.5797	46.2144	45.66	1.6	14.87	9.9	0.17	9.81	13.09	3.09	0.96	0.84
Basalt	-122.5839	46.21	48.24	1.34	15.91	10.16	0.17	10.06	10.78	2.51	0.56	0.26
Basalt	-122.5814	46.2222	45.42	1.57	14.97	10.13	0.19	9.94	13.01	3.3	0.76	0.7
Basaltic andesite	-122.5789	46.2078	52.54	1.26	16.45	8.38	0.14	8.16	9.2	3.06	0.57	0.22

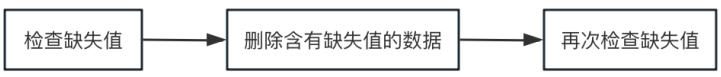
图 1 数据总览（部分）——火成岩类型分布

II.　数据预处理

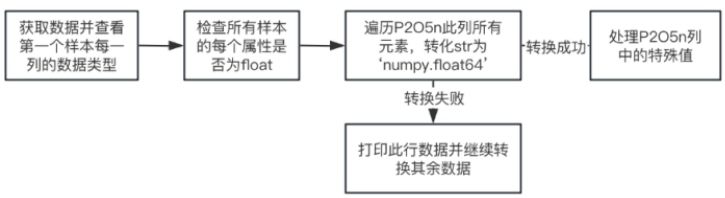
导入所需要的库、创建数据框完成数据的准备工作后，需要进行数据的预处理工作，可以分为以下几个环节。

一、 数据清洗

- 数据缺失处理流程：



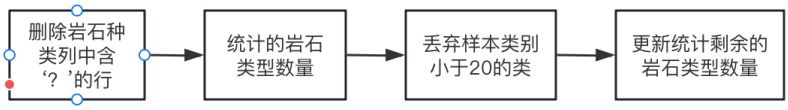
- 数据类型检查流程：



主要解决样本重复、检测和处理数据、构建新的属性并处理数据值的冲突问题等。

[①] https://www.kaggle.com/datasets/cristianminas/geochemical-variations-in-igneous-rocks-mining

➢ 丢弃不确定和样本过少的类别流程：



样本类别小于 20，其样本数量过少，统计意义不大。

表 1 数据清洗各个步骤的结果统计

处理步骤	删除数据量/条	剩余/条	删除岩石种类/类	剩余/类
初始数据		4162		98
删除含有缺失值的数据	275		2	
删除 rock_name 列包含 '?' 的行	10		8	
删除数据量小于 20 的类	312		72	
全部处理后		3575		16

二、 数据标准化与划分

统一数据的尺度，利于提取并构建有意义的特征。

➢ 数据标准化与划分流程：

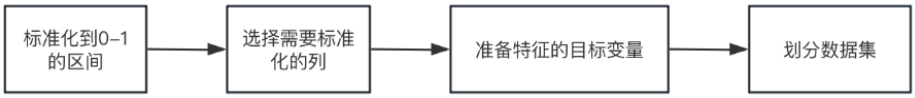


表 2 数据集划分

数据集	数据量/条	属性/个
Training set	2860	10
Testing set	715	10

III. 数据分布可视化

三、 样本类别分布

对数据集中不同岩石类型的样本数量进行统计，并且只保留那些样本数量大于 20 的岩石类型。结果发现样本分布相差较大（图 4）。

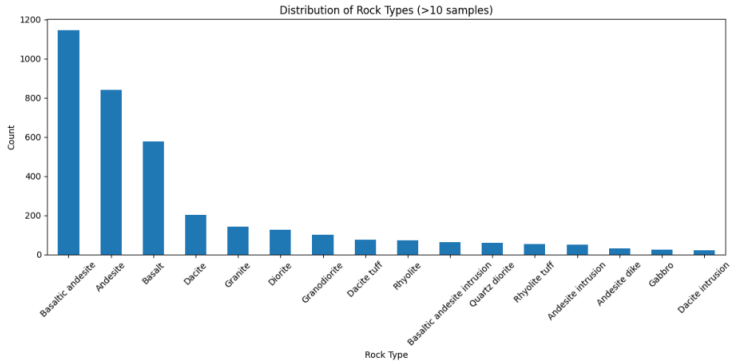


图 2 筛选后的岩石类型分布

四、 样本地理空间分布

目的是从数据集中随机采样 1000 个样本，并使用 Cartopy 库来可视化这些样本的地理空间分布

（图 5）。同一片区域出现了不同类别样本同时分布的情况。这与岩石分布的深度有关所以，仅靠地理位置进行分类是不可行的，要结合岩石化学成分特征进行分类。

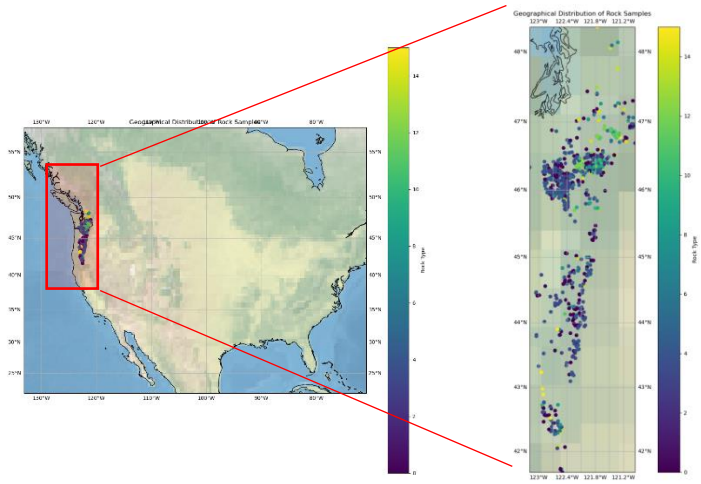


图 3 岩石样品的地理分布

三、 样本特征分布

1) 基于 PCA 方法

PCA 是一种线性降维技术，通过正交变换将一组可能相关的变量转换为一组线性不相关的变量。由于岩石的属性有很多，无法在高维空间可视化，使用主成分分析（PCA）算法来降低数据的维度，并在二维和三维空间中对数据点进行可视化。在 sklearn 中，主成分分析（PCA）可以通过 sklearn.manifold.PCA 对象来实现。

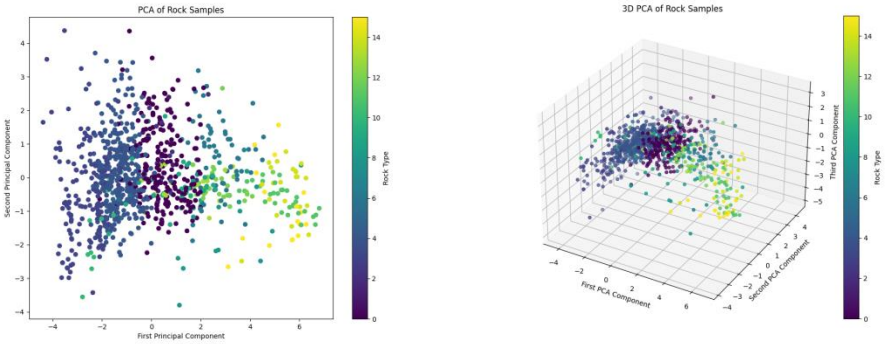


图 4 岩石样本的主成分分析可视化

2) 基于 t-SNE 方法

t-Distributed Stochastic Neighbor Embedding (t-SNE) (t-分布随机邻域嵌入) 是一种非线性降维算法，它通过最小化高维空间和低维空间中数据点之间的概率分布差异，将高维数据映射到低维空间，从而可以直观地展示数据的内在结构和聚类特征。

可以看到，同一类样本的特征降维后，在空间中具有聚集性，这给我们用机器学习方法进行准确分类提供了依据。两种降维方法可以得到相同的结论。

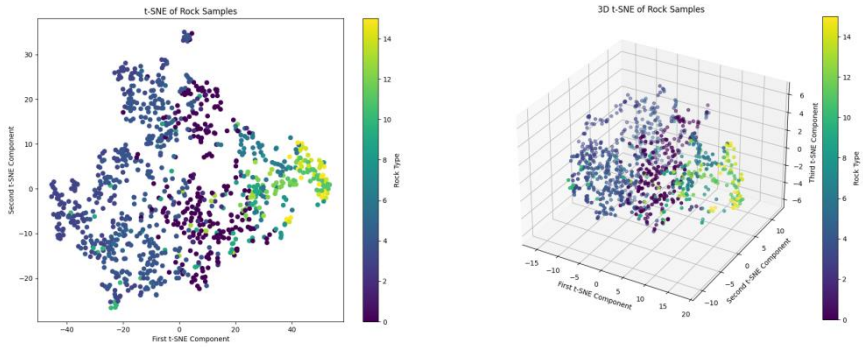


图 5 岩石样品的 t-SNE 降维可视化

IV. 训练

我们选择了三种机器学习模型：K-近邻算法（KNN），支持向量机（SVM），以及随机森林（Random Forest）。KNN（K-Nearest Neighbors）是一种基于实例的学习方法，通过计算样本之间的欧氏距离进行分类。我们使用了网格搜索交叉验证（GridSearchCV）方法来优化 KNN 模型的超参数。参数范围: [5, 10, 15]。通过五折交叉验证，我们发现最佳的超参数选择为 10。

表 3 KNN 模型搜索参数

参数	范围	最佳
n_neighbors	5, 10, 15	10

支持向量机（SVM）是一种分类器，通过构建最大间隔超平面来区分不同类别的数据。我们使用径向基函数（RBF）核，并通过网格搜索优化超参数 C 和 γ 。参数范围为 C: [0.1, 1, 10]；γ: [scale, auto]。五折交叉验证结果显示，最佳参数组合为 C=10 和 γ =auto。

表 4 SVM 模型搜索参数

参数	范围	最佳
C	0.1, 1, 10	10
γ	scale, auto	auto

随机森林（Random Forest）是一种基于树的集成学习方法，通过构建多棵决策树并将其预测结果进行筛选来提高分类性能。我们通过网格搜索优化了随机森林的超参数 n_estimators(树的数量)和 max_depth(树的最大深度)。参数范围为 n_estimators: [10, 50, 100]。max_depth: [5, 10, None]。最佳参数组合为 n_estimators=100 和 max_depth=10。

表 5 随机森林模型搜索参数

参数	范围	最佳
n_estimators	10, 50, 100	100
max_depth	5, 10, None	10

V. 结果可视化分析

在对火成岩地球化学数据进行机器学习模型训练后，我们对模型的性能进行了系统的评估，并通过以下三种可视化手段对结果进行了详细分析。分别为准确率（Accuracy）、混淆矩阵（Confusion Matrix）和分类报告（Classification Report）。

图 6　KNN, SVM, RF 模型分类结果的混淆矩阵

K-近邻（KNN）模型在测试集上的准确率为 71.33%。混淆矩阵显示，该模型在大多数类别上表现良好，但在某些特定类别存在一些的误分类现象。分类报告进一步证实了这一点，精确率和召回率在这些类别上相对较低。

支持向量机（SVM）模型在测试集上的准确率为 76.08%，总体表现优于 KNN 模型。

随机森林（Random Forest）模型在测试集上的准确率达到了 78.32%，显著高于前两种模型。

通过对三种模型的全面评估和结果可视化分析，我们发现随机森林模型在火成岩地球化学数据分类任务中表现最佳。

	precision	recall	f1-score	support
Basaltic andesite	0.845588	0.962343	0.900196	239
Andesite	0.737624	0.937107	0.825485	159
Basalt	0.930693	0.846847	0.886792	111
Dacite	0.673469	0.846154	0.75	39
Diorite	0.5	0.285714	0.363636	28
Granite	0.612903	0.863636	0.716981	22
Granodiorite	0.692308	0.428571	0.529412	21
Rhyolite	0.666667	0.555556	0.606061	18
Quartz diorite	0.5	0.076923	0.133333	13
Rhyolite tuff	0.5	0.230769	0.315789	13
Andesite intrusion	0	0	0	12
Dacite tuff	0.5	0.333333	0.4	12
Basaltic andesite intrusion	0	0	0	11
Andesite dike	0	0	0	7
Gabbro	0	0	0	7
Dacite intrusion	0	0	0	3

图 7　RF 模型的分类报告

对于分类错误的类别，例如安山岩和安山岩侵入体，他们岩性相似，只是形成的位置不同，所以区分难度较大。

观察分类报告可以发现，每个种类的分类准确率大致和样本数量成正相关。构建更大的样本库，才能训练泛化能力更强的模型。

参考文献

【1】Author, CRISTIAN CARTAGENA MATOS. Dataset_name, Geochemical Variations in Igneous Rocks – Mining. Platfrom, Kaggle, https://www.kaggle.com/datasets/cristianminas/geochemical-variations-in-igneous-rocks-mining，

【2】Scikit-learn: Machine Learning in Python, Pedregosa et al., JMLR 12, pp. 2825-2830, 2011.

【3】https://github.com/VoyagerXvoyagerx/RockClassification（代码已开源）