1000-0569/2018/034(11)-3173-78 *Acta Petrologica Sinica* 岩石学报

地质领域机器学习、深度学习及实现语言*

周永章[1,2,3] 王俊[1,2,3] 左仁广[4] 肖凡[1,2,3] 沈文杰[1,2,3] 王树功[1,2,3]

ZHOU YongZhang[1,2,3], WANG Jun[1,2,3], ZUO RenGuang[4], XIAO Fan[1,2,3], SHEN WenJie[1,2,3] and WANG ShuGong[1,2,3]

1. 广东省地质过程与矿产资源探查重点实验室,广州 510275
2. 中山大学地球环境与地球资源研究中心,广州 510275
3. 中山大学地球科学与工程学院,广州 510275
4. 中国地质大学,武汉 430074

1. *Center for Earth Environment & Resources, Sun Yat-sen University, Guangzhou* 510275, *China*
2. *Guangdong Provinical Key Laboratory of Mineral Resources and Geological Processes, Sun Yat-sen University, Guangzhou* 510275, *China*
3. *School of Earth Sciences & Engineering, Sun Yat-sen University, Guangzhou* 510275, *China*
4. *China University of Geology, Wuhan* 430074, *China*

2018-05-30 收稿, 2018-08-10 改回.

Zhou YZ, Wang J, Zuo RG, Xiao F, Shen WJ and Wang SG. 2018. Machine learning, deep learning and Python language in field of geology. *Acta Petrologica Sinica*, 34(11):3173-3178

Abstract Geological big data is exponentially expanding. It is the only way to catch up with its extraordinary growing to develop intelligent data processing. As the core of artificial intelligence, machine learning is a fundamental way to endow computer with intelligence. Machine learning has been becoming the front hotspot of geological big data mining. It will attach wings to geological big data mining, and thereby bring revolution to geological research. Machine learning is a data adaptive training process and model, resulting in giving a good performance decision. As a subclass of machine learning, deep learning develops machine learning model with various hidden layers, and makes iterative evolution of the model through massive data training, and finally extracts essential features to help more exactly classing and predicting. The convolution neural network is one of the most frequently used deep learning algorithms. It is widely used in image recognition and speech analysis. Python language is playing an increasingly important role in science research. The Python Scikit-Learn is a machine learning-oriented library to provide with data preprocessing, classification, regression, clustering, prediction, model analysis and other modules. The Keras is a Python deep learning library based on Theano and Tensorflow, and can be used to construct concise artificial neural network.

Key words Geological big data; Machine learning; Deep learning; Artificial neural network; Intelligent geology; Python

摘 要 地质大数据正在以指数形式增长。只有发展智能数据处理方法才有可能追上大数据的超常增长。机器学习是人工智能的核心,是使计算机具有智能的根本途径。机器学习已成为地质大数据研究的前沿热点,它将让地质大数据插上翅膀,并因此改变地质。机器学习是一个源于数据的模型的训练过程,最终给出一个面向某种性能度量的决策。深度学习是机器学习研究中的一个重要子类,它通过构建具有很多隐层的机器学习模型和海量的训练数据,来学习更有用的特征,从而最终提升分类或预测的准确性。卷积神经网络算法是最为常用的一种深度学习算法之一,它广泛用于图像识别和语音分析等。Python 语言在科学领域的地位占据着越来越重要。其下的 Scikit-Learn 是一个机器学习相关的库,提供有数据预处理、分类、回归、聚类、预测、模型分析等算法。Keras 是一个基于 Theano/Tensorflow 的深度学习库,可以应用来搭建简洁的人工神经

* 本文受国家重点研发计划重点专项项目(2016YFC0600506)、国家自然科学基金项目(41273040)、中国地质调查局项目(12120113067600)和广东省地质过程与矿产资源探查重点实验室基金联合资助.

第一作者简介: 周永章,男,1963 年生,教授,大数据、数学地球科学与地球化学专业,E-mail: zhouyz@mail.sysu.edu.cn

网络。

关键词 地质大数据;机器学习;深度学习;人工神经网络;智能地质学;Python

中图法分类号 P628

地质大数据正在以指数形式增长。它们大量产生于基础地质、矿产地质、水文地质、工程地质、环境地质、灾害地质的调查、勘查和相应的地质科学研究过程中,能源、矿产的开发利用和环境、地灾的监测、防治过程中,以及各类天基、空基对地遥感观测活动中。地质大数据获得的途径包括地球物理、地球化学、钻探测井、遥感遥测、传感监测,还可以来自各种拓展应用,如图件编绘、分析计算、模拟仿真、预测评价、智能管控等。地质大数据可以是结构化的,如地球化学分析和地球物理探查获得的数据;还有更多的非结构化、半结构化的数据,如古生物、矿物、岩石、矿床、岩心照片,海啸音频、地震视频,构造、遥感光谱图件,标本、野外记录、地质图表等。

在这一背景下,任何个人以传统方式处理地质大数据,就如同人用腿跟汽车、飞机、火箭赛跑,越往前走,差距越大,最终被先进交通工具所被抛弃。只有发展智能数据处理方法,才有可能追上大数据的超常增长,因此可以说,人工智能地质学应是重要的发展方向。

机器学习被认为是人工智能的核心,是使计算机具有智能的根本途径。目前,机器学习与人工智能各种基础问题的统一性观点正在形成(图 1)。

尽管具有历史使命感的科学家在严肃、认真地探索(Mayer-Schonberger and Cukier, 2013; Carranza and Laborte, 2015; de Mulder *et al.*, 2016; Aryafar and Moeini, 2017; Ross *et al.*, 2018),但依托大数据的人工智能地质学还远不够成熟(张旗和周永章, 2017; 周永章等, 2018a)。

本期作为大数据专辑,聚焦机器学习(含深度学习)建模和应用的论文相对比较集中(徐述腾, 2018; 周永章, 2018; 韩帅等, 2018; 焦守涛等, 2018; 刘艳鹏等, 2018; 王怀涛等, 2018)。从中亦折射了,机器学习已是当前地质大数据研究的重要热点之一。作者认为,机器学习将为地质大数据插上翅膀,强有力地处理海量数据,挖掘它们背后有价值的丰富信息,并因此改变地质。

图 1 人工智能、机器学习与深度学习之间的关系

Fig. 1 Relationship between artificial intelligence, machine learning and deep learning

1 机器学习的分类

从本质上讲,机器学习是一个源于数据的模型的训练过程,最终给出一个面向某种性能度量的决策。

机器学习可以分为有监督学习和无监督学习。在监督式学习(Supervised Learning)下,每组训练数据都有一个标识值或结果值。在建立预测模型的时候,监督式学习建立一个学习过程,将预测的结果与训练数据的实际结果进行比较,不断的调整预测模型,直到模型的预测结果达到一个预期的准确率。监督式学习的常见方法如图 2 所示。

在无监督式学习(Unsupervised Learning)中,数据并不被特别标识,学习模型是为了推断出数据的一些内在结构。常见的无监督学习方法如图 3 所示。

在机器学习中,SVM(Support Vector Machine)是一种比较有代表性的方法。它的基础是二元分类算法,核心思维是升维和线性化。很多在低维样本空间无法线性处理的样本集,在高维特征空间中却可以通过一个线性超平面实现线性划分(或回归)。SVM 通过一个非线性映射 p,把样本空间映射到一个高维的特征空间中,使得在原来的样本空间中非线性可分的问题转化为在特征空间中的线性可分的问题。

2 深度学习

深度学习(Deep learning)是机器学习研究中的一个子类。它的目的是建立、模拟人脑进行分析学习的神经网络,模仿人脑的机制来解释数据,例如图像、声音和文本。深度学习的实质,是通过构建具有很多隐层的机器学习模型和海量的训练数据,来学习更有用的特征,从而最终提升分类或预测的准确性(Hinton *et al.*, 2006, 2012; Brenden *et al.*, 2015; LeCun *et al.*, 2015; Schmidhuber, 2015; Bianco *et al.*, 2017)。"深度模型"是手段,"特征学习"是目的。

表 1 列出了目前常见的深度学习模型或方法。

其中,卷积神经网络算法是目前最为常用的一种深度学习算法。它广泛用于图像识别和语音分析。它本质上是一种输入到输出的映射,能够学习大量的输入与输出之间的映射关系,而不需要任何输入和输出之间的精确的数学表达式,只要用已知的模式对卷积网络加以训练,网络就具有输入输出对之间的映射能力。

卷积神经网络是一个多层的神经网络,每层由多个二维

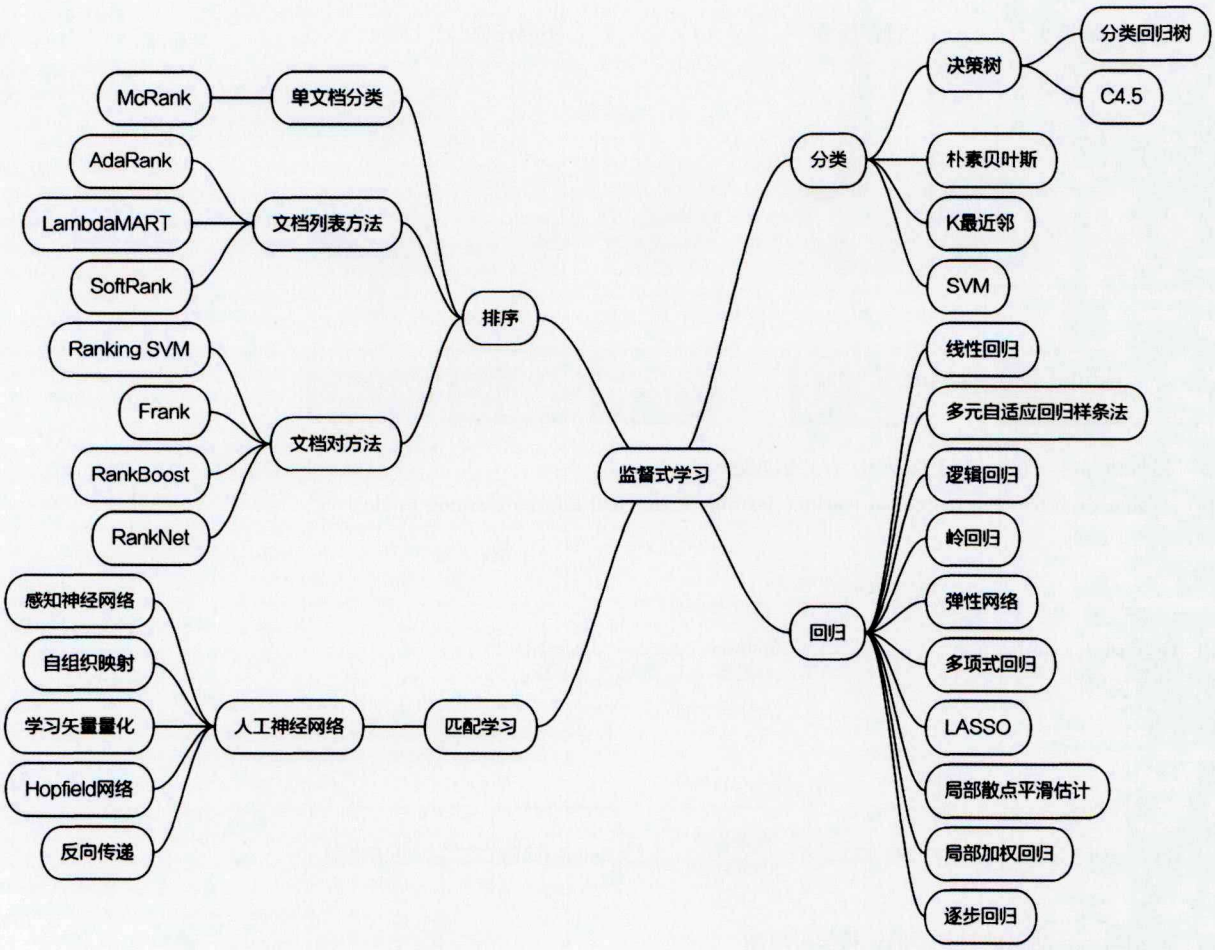

图 2 常见的监督式学习方法(据周永章等, 2018b)

Fig. 2 Common supervised learning algorithms (after Zhou *et al.*, 2018b)

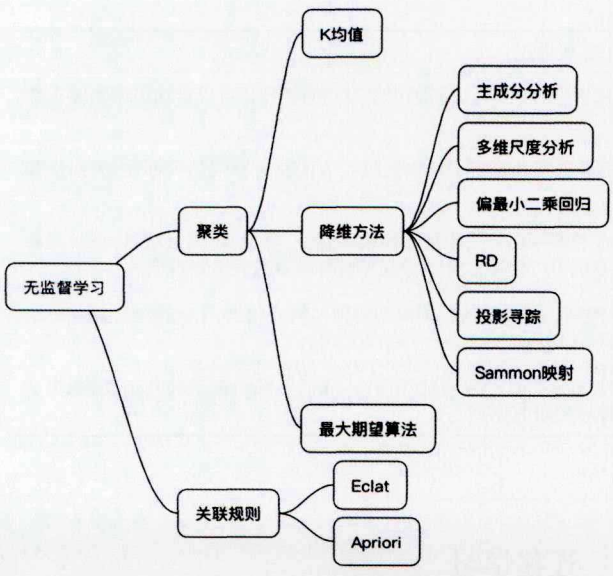

图 3 常见的无监督学习方法

Fig. 3 Common unsupervised learning algorithms

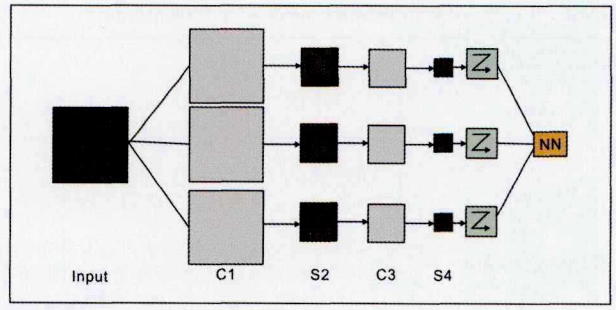

图 4 卷积神经网络的概念示范

Fig. 4 Conceptual diagram of convolution neural network

平面组成,而每个平面由多个独立神经元组成(图 4)。卷积网络是为识别二维形状而特殊设计的一个多层感知器,这种网络结构对平移、比例缩放、倾斜或者共他形式的变形具有高度不变性。

在本期论文中,徐述腾和周永章(2018)以吉林夹皮沟金矿和河北石湖金矿的黄铁矿、黄铜矿、方铅矿、闪锌矿等硫化物矿物为例,设计了有针对性的 Unet 卷积神经网络模型,实

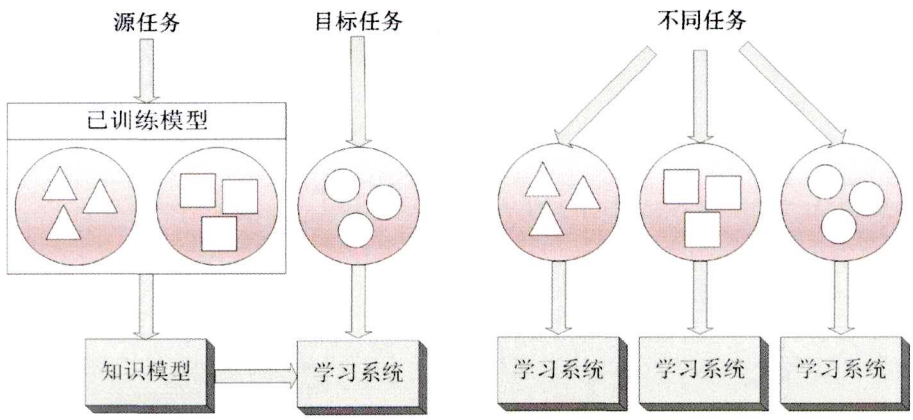

图 5 传统机器学习(左)与迁移学习(右)的过程差别

Fig. 5 Comparison between traditional machine learning (left) and transfer learning (right)

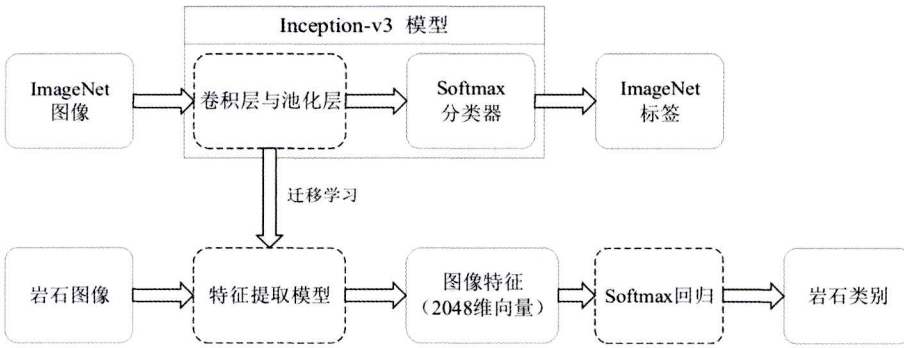

图 6 矿物岩石自动识别与分类的迁移学习模型

Fig. 6 Transfer learning model for automatic identification and classification of minerals

表 1 深度学习常用模型或方法及其描述

Table 1 Common deep learning models and algorithms

常用模型或方法	算法描述
自编码器	一种无监督的神经网络模型. 可以学习到输入数据的隐含特征(编码),同时用学习到的新特征可以重构出原始输入数据(解码)。自动编码器被用于降维或特征学习
稀疏编码	一种无监督学习方法。用来寻找一组"超完备"基向量来更高效地表示样本数据. 方法具有空间的局部性、方向性和频域的带通性,是一种自适应的图像统计方法
受限波尔兹曼机	RBM,一种可通过输入数据集学习概率分布的随机生成神经网络. 受限玻兹曼机在降维、分类、协同过滤、特征学习和主题建模中得到了应用. 根据任务的不同,受限玻兹曼机可以使用监督学习或无监督学习的方法进行训练
深信度网络	DBNs,由多个受限玻尔兹曼机层组成的一个概率生成模型. 与传统的判别模型的神经网络相对,生成模型是建立一个观察数据和标签之间的联合分布. 可拓展为卷积 DBNs(CDBNs)
卷积神经网络	CNNs,人工神经网络的一种,它的权值共享网络结构使之更类似于生物神经网络,降低了网络模型的复杂度,减少了权值的数量. CNNs 已成为当前语音分析和图像识别领域的研究热点

现了基于深度学习算法的镜下矿石矿物自动识别与分类。实验显示,训练出的模型对测试集的显微镜镜下矿石矿物照片的识别成功率均高于 90%,说明实验所建立的模型,具有很好的图像特征提取能力,能完成镜下矿石矿物智能识别的任务。

3 迁移学习

迁移学习(Transfer learning)是把已学训练好的模型参数迁移到新的模型,以便帮助新模型训练(Yosinski *et al.*,

2014)。考虑到大部分数据或任务是存在相关性的，所以通过迁移学习可以将已经学到的模型参数(也可理解为模型学到的知识)通过某种方式来分享给新模型从而加快并优化模型的学习效率不用像大多数网络那样从零学习。

传统机器学习与迁移学习的过程有实质性的差别，如图5所示。

迁移学习被用来研究矿物岩石的自动识别与分类，可以为岩石岩性自动分类提供了一种新的手段。张野等(2018)选取花岗岩、千枚岩、角砾岩三种岩石图像来进行测试识别分析。实验中用到的岩石图像样本是通过照片、岩石数据库和网络搜索等不同手段采集得到，岩石类型主要由实验室岩石标本、现场岩石标本及现场大范围岩石三种图像组成。为了使整个过程更加智能化，对于岩石图像的缩放、裁剪的处理均在训练中自动完成，输入的图像只保证固定的格式，对于图像大小，尺寸和像素均无具体要求。作者建立了基于Inception-v3的岩石图像深度学习迁移模型，如图6所示，对花岗岩、千枚岩和角砾岩三种岩石的自动识别率可以达到80%以上，部分结果甚至可以达到95%以上。训练过程对于岩石图像的大小、成像距离及光照强度要求低，充分证明了其鲁棒性和泛化能力。

4 算法实现

随着NumPy、SciPy、Matplotlib、Pandas等众多程序库的开发，Python在科学领域占据着越来越重要的地位(周永章等，2018a)。

Scikit-Learn是一个机器学习相关的库，是Python下强大的机器学习工具包。它提供了完善的机器学习工具箱，包括数据预处理、分类、回归、聚类、预测、模型分析等。

人工神经网络是功能相当强大的、但是原理又相当简单的模型，在图像识别、语言处理等领域都有重要的作用。Theano也是Python的一个库。它是由深度学习专家Yoshua Bengio带领的实验室开发出来的，用来定义、优化和高效地解决多维数组数据对应数学表达式的模拟估计问题。它具有高效地实现符号分解、高度优化的速度和稳定性等特点，最重要的是它还实现了GPU加速，使得密集型数据的处理速度是CPU的数十倍。用Theano可以搭建起高效的神经网络模型，然而对于普通读者来说门槛是相对较高的。

因此，可以应用Keras库来搭建神经网络。应用Keras库可以幅度简化了搭建各种神经网络模型的步骤，允许普通用户轻松地搭建并求解具有几百个输入节点的深层神经网络。Keras并非简单的神经网络库，而是一个基于Theano的强大的深度学习库，利用它不仅仅可以搭建普通的神经网络，还可以搭建各种深度学习模型，如自编码器、循环神经网络、递归神经网络、卷积神经网络等等。由于它是基于Theano的，因此速度也相当快。

用Keras搭建神经网络模型的过程相当简洁，也相当直观，它纯粹地就像搭积木一般。通过短短几十行代码，就可以搭建起一个非常强大的神经网络模型，甚至是深度学习模型。

在作者的教学与科研中，推荐搭建Python开发平台，应用Python语言可以很好实现机器学习算法。

5 结语

通过上述讨论，可以形成如下认识：

(1)地质大数据正在以指数形式增长。有且只有发展智能数据处理方法才有可能追上大数据的超常增长，因此而发展的人工智能地质学应是重要的发展方向。

(2)机器学习是使计算机具有智能的根本途径。它本质是一个源于数据的模型训练过程，最终给出一个面向某种性能度量的决策。

(3)深度学习的目的是建立、模拟人脑进行分析学习的神经网络，模仿人脑的机制来解释数据。它的实质是通过构建具有很多隐层的机器学习模型和海量的训练数据，来学习更有用的特征，从而最终提升分类或预测的准确性。

(4)NumPy、SciPy、Matplotlib、Pandas等众多程序库的开发，使Python在科学领域占据着越来越重要的地位。Scikit-Learn和Keras是利用Python构建实现机器学习和人工神经网络的重要工具包。

(5)尽管依托大数据的人工智能地质学还远不够成熟，但机器学习算法的突破和发展，使迅速处理海量地质大数据，挖掘它们背后有价值的丰富信息成为可能，并因此将改变地质。

References

Aryafar A and Moeini H. 2017. Application of continuous restricted Boltzmann machine to detect multivariate anomalies from stream sediment geochemical data, Korit, East of Iran. Journal of Mining and Environment, 8(4): 673 - 682

Bianco S, Buzzelli M, Mazzini D and Schettini R. 2017. Deep learning for logo recognition. Neurocomputing, 245: 23 - 30

Brenden M, Ruslan S and Joshua B. 2015. Human-level concept learning through probabilistic program induction. Science, 350(6266): 1332 - 1338

Carranza EJM and Laborte AG. 2015. Random forest predictive modeling of mineral prospectivity with small number of prospects and data with missing values in Abra (Philippines). Computers & Geosciences, 74: 60 - 70

de Mulder EFJ, Cheng Q, Agterberg F and Goncalves M. 2016. New and game-changing developments in geochemical exploration. Episodes, 39(1): 70 - 71

Han S, Li MC, Liu CZ and Ren QB. 2018. Determination and analysis of tectonic setting based on basalts and intelligent algorithms. Acta Petrologica Sinica, 34(11): 3207 - 3216 (in Chinese with English abstract)

Hinton GE, Osindero S and Teh Y. 2006. A fast learning algorithm for deep belief nets. Neural Computation, 18: 1527 - 1554

Hinton GE, Deng L, Yu D, Dahl GE, Mohamed AR, Jaitly N, Senior A, Vanhoucke V, Nguyen P, Kingsbury B and Sainath T. 2012.

Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. IEEE Signal Processing Magazine, 29(6): 82 - 97

Jiao ST, Zhou YZ, Zhang Q, Jin WJ, Liu YP and Wang J. 2018. Study on intelligent discrimination of tectonic settings based on global gabbro data from GEOROC. Acta Petrologica Sinica, 34(11): 3189 - 3194 (in Chinese with English abstract)

LeCun Y, Bengio Y and Hinton GE. 2015. Deep learning. Nature, 521 (7553): 436 - 444

Liu YP, Zhu LX and Zhou YZ. 2018. Application of Convolutional Neural Networkin prospecting prediction of ore deposits: Taking the Zhaojikou Pb-Zn ore deposit in Anhui Province as a case. Acta Petrologica Sinica, 34(11): 3217 - 3224 (in Chinese with English abstract)

Mayer-Schonberger V and Cukier K. 2013. Big Data: A Revolution that will Transform How We Live, Work and Think. New York: Houghton Mifflin Harcourt Publishing Company

Ross ZE, Meier MA and Hauksson E. 2018. P-wave arrival picking and first-motion polarity determination with deep learning. Journal of Geophysical Research: Solid Earth, 123. https://doi. org/10. 1029/2017JB015251

Schmidhuber J. 2015. Deep learning in neural networks: An overview. Neural Networks, 261: 85 - 117

Wang HT, Luo JM, WangJR, Yang J, Song BT, Wang YX, Wang XW and Zhou YQ. 2018. Quantitative classification and metallogenic prognosis of basic-ultrabasic rocks based on big data: Taking its application in Beishan area for example. Acta Petrologica Sinica, 34 (11): 3195 - 3206 (in Chinese with English abstract)

Xu ST and Zhou YZ. 2018. Artificial intelligence identification of ore minerals under microscope based on Deep Learning algorithm. Acta Petrologica Sinica, 2018, 34(11): 3244 - 3252 (in Chinese with English abstract)

Yosinski J, Clune J, Bengio Y and Lipson H. 2014. How transferable are features in deep neural networks? In: Ghahramani Z, Welling M, Cortes C, Lawrence ND and Weinberger KQ (eds.). Proceedings of Neural Information Processing Systems. 28th Annual Conference on Neural Information Processing Systems. Montreal, 2014. USA, MA: MIT Press, 3320 - 3328

Zhang Q and Zhou YZ. 2017. Big data will lead to a profound revolution in the field of geological science. Chinese Journal of Geology, 52 (3): 1 - 12 (in Chinese with English abstract)

Zhang Y, Li MC and Han S. 2018. Automatic identification and classification in lithology based on deep learning in rock images. Acta Petrologica Sinica, 34 (2): 333 - 342 (in Chinese with English abstract)

Zhou YZ, Chen S, Zhang Q, Xiao F, Wang SG, Liu YP and Jiao ST. 2018a. Advances and prospects of big data and mathematical geoscience. Acta Petrologica Sinica, 34(2): 256 - 263 (in Chinese with English abstract)

Zhou YZ, Zhang LJ, Zhang AD and Wang J. 2018b. Big Data Mining & Machine Learning in Geoscience. Guangzhou: Sun Yat-sen University Press, 1 - 360 (in Chinese)

附中文参考文献

韩帅, 李明超, 刘承照, 任秋兵. 2018. 基于玄武岩大数据的大地构造环境智能挖掘判别与分析. 岩石学报, 34(11): 3207 - 3216

焦守涛, 周永章, 张旗, 金维浚, 刘艳鹏, 王俊. 2018. 基于GEOROC数据库的全球辉长岩大数据的大地构造环境智能判别研究. 岩石学报, 34(11): 3189 - 3194

刘艳鹏, 朱立新, 周永章. 2018. 卷积神经网络及其在矿床找矿预测中的应用研究——以安徽省兆吉口铅锌矿床为例. 岩石学报, 34(11): 3217 - 3224

王怀涛, 罗建民, 王金荣, 杨婧, 宋秉田, 王玉玺, 王晓伟, 周煜祺. 2018. 基于大数据的基性-超基性岩定量分类及成矿预测研究——以北山地区为例. 岩石学报, 34(11): 3195 - 3206

徐述腾, 周永章. 2018. 基于深度学习的镜下矿石矿物的智能识别实验研究. 岩石学报, 34(11): 3244 - 3252

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, 52(3): 1 - 12

张野, 李明超, 韩帅. 2018. 基于岩石图像深度学习的岩性自动识别与分类方法. 岩石学报, 34(2): 333 - 342

周永章, 陈烁, 张旗, 肖凡, 王树功, 焦守涛, 刘艳鹏. 2018a. 大数据与数学地球科学研究进展. 岩石学报, 34(2): 256 - 263

周永章, 张良均, 张奥多, 王俊. 2018b. 地球科学大数据挖掘与机器学习. 广州: 中山大学出版社, 1 - 360