融合学习模型的岩石光谱特征自动分类

贺金鑫[1]，任小玉[1]，陈圣波[2*]，熊　玥[1]，肖志强[1]，周　孩[1]

1. 吉林大学地球科学学院，吉林 长春　130061
2. 吉林大学地球探测科学与技术学院，吉林 长春　130061

摘　要　岩石光谱综合反映了岩石的物理化学性质、成分及其结构构造。岩石光谱数据已被应用于岩石分类的研究，但是不同于矿物光谱，岩石光谱并无标准数据库，且受较多干扰因素影响，例如矿物组分、结构构造、化学成分、风化力度，测量仪器的误差等。传统岩石光谱分类模型先是对岩石光谱进行预处理排除干扰，然后采用不同方法对部分光谱特征分析，以达到分类目的。但对光谱数据特征遗失较多，使得分类准确率低下且操作过程繁琐、效率不高。因此，建立一个简单、快速、准确的岩石光谱自动分类模型具有重要意义。机器学习能够对获得的所有数据进行学习，不存在遗漏，大大提高了分类精度，且是对原始数据直接操作，不需预处理，简化流程。为此，选取辽宁兴城地区作为研究区，采集了若干种典型岩石样本，利用美国ASD便携式光谱仪实测光谱，最终获得608条数据，依据岩石光谱特征分为三类进行研究。首先利用决策树(DT)及决策树的升级模型——随机森林(RF)对数据进行分类，但当数据噪音较大时随机森林容易陷入过拟合；因而利用对异常值不敏感的K-最近邻(KNN)建模，但KNN需要对每个样本都考虑，数据量大时计算量会很大，效率不高；所以通过支持向量机(SVM)来提升分类准确率。从实验结果可以看出，4种分类模型的准确率排序为：SVM＞KNN＞RF＞DT。为进一步提高岩石光谱特征的自动分类精度，采取了融合多个不同模型的办法，即对不同模型的分类结果进行投票，选择投票最多的作为最后分类结果。由于硬投票可在一定程度上减少过拟合现象的发生，更加适合分类模型，所以利用硬投票法融合了RF、KNN与SVM三个机器学习模型，最终的分类准确率可达到99.17%。综上所述，基于融合学习模型进行岩石光谱特征自动分类是切实可行且准确高效的。

关键词　岩石光谱分类；决策树；随机森林；K-最近邻；支持向量机；模型融合

中图分类号：TP79　　**文献标识码**：A　　**DOI**：10.3964/j.issn.1000-0593(2021)01-0141-04

引　言

在遥感地质领域，岩石光谱一直是热门研究方向之一，主要包括成像光谱岩矿识别、岩石光谱特征分析、影响岩石光谱的因素以及岩石光谱分类。在岩石光谱分类方面，吴辉等基于AVIRIS航空高光谱遥感数据，首先将预处理后的数据进行最小噪声分离，然后用PPI算法选择研究对象，最后用线性光谱混合分类模型进行岩性分类研究[1]；张翠芬等人将岩石单元的图形纹理特征及光谱特征进行协同分类研究，用面对对象方法进行图谱指数分割，然后用光谱指数提取岩石信息，划分精度较好[2]；徐清俊等人使用ASD光谱仪测量钻孔岩心数据，利用ViewSpecpro软件进行格式转换，输入到ENVI软件建成光谱库，与美国USGS光谱库中典型矿物光谱曲线进行对比分析，进而识别岩性得出结论[3]；周江将ASD光谱仪的光谱曲线与遥感影像通过ENVI软件相结合对岩石等地物进行分类，与用神经网络进行监督分类的结果相对比[4]。总之，该领域目前的主要问题在于要么是将数据进行一系列复杂预处理后利用传统模型进行分类；要么因为岩石光谱的特殊性，没有统一的光谱曲线标准，使得分类结果不够准确。因此，本文拟在不对岩石光谱数据进行复杂预处理的前提下，构建一种基于融合多种机器学习模型的岩石光谱特征自动分类方法；并与单一分类模型相对比，最终取得更高的分类准确率。

收稿日期：2019-12-15，**修订日期**：2020-04-11

基金项目：国家自然科学基金项目(41772346)资助

作者简介：贺金鑫，1979年生，吉林大学地球科学学院教授　　e-mail：hejx@jlu.edu.cn

*通讯作者　　e-mail：chensb@jlu.edu.cn

1 岩石光谱数据

1.1 研究区概况

研究区位于辽宁省兴城市，区域地貌属辽西山地黑山丘陵东部边缘的海滨丘陵，海拔在 20～500 m 之间，相对高差 200～350 m，地势总体呈西北高而东南低，区内河流发育，有六股河、烟台河等汇入辽东湾；气候属于北半球暖温带亚湿润气候，气候温和，干湿相宜[5]。

兴城地区出露的地层为典型的华北型，地层从太古宙到中—新元古界、古生界、中生界和新生界都有分布，发育较为齐全，主要岩石类型有花岗岩、砂岩、页岩、白云岩、灰岩、安山岩、玄武岩等[5]（如图 1 所示）。

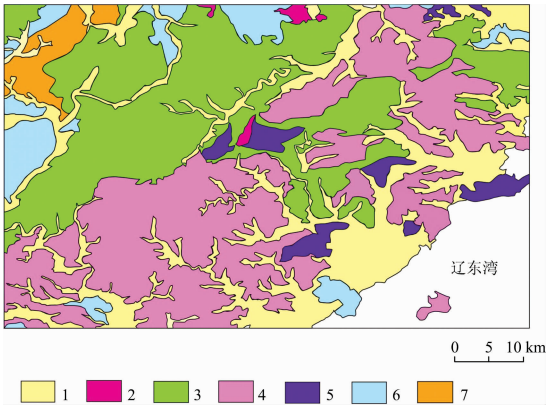

图 1 研究区岩性分布图

1：第四系：砾石、黄土、粉质粘土；2：闪长岩；
3：灰岩；4：花岗岩；5：砂岩；6：安山岩；7：玄武岩

Fig. 1 Distribution of rocks in the study area

1：Quaternary：Gravel，Loess，silty Clay；2：Diorite；3：Limestone；4：Granite 5：Sandstone；6：Andesite；7：Basalt

1.2 岩石光谱测量

用于测量岩石光谱的仪器为美国 FieldSpec-3 型便携式实测光谱仪，所测波长从 350 nm 的可见光范围分布到 2 500 nm 的短波红外范围。可见光的光谱测量间隔为 1.4 nm，分辨率约为 3 nm；短波红外的间隔为 2 nm，分辨率为 6.5～8.5 nm[6]。

目前取得已命名岩石光谱类型有二长花岗岩、花岗斑岩、石英砂岩、中粒岩屑长石砂岩、白云质灰岩、鲕状灰岩、燧石条带白云岩等。将测量得到的光谱数据进行整合，最终得到灰岩类 379 条数据、花岗岩类 147 条数据、砂岩类 82 条数据，其余类别由于数据量过少，暂不予以分类研究（如图 2 所示）。

1.3 岩石光谱特征

岩石光谱形状与其成分、含量等等因素都密切相关。而同种岩石光谱形态基本相似；实验所得数据中，花岗岩和砂岩在 1 400 nm 左右处都存在水汽吸收带（如图 3、图 5 所示），在 1 900 nm 处，三类岩石光谱都存在较强吸收谷（如图 3—图 5 所示）；花岗岩总体反射率在 0～0.5 之间，灰岩总体反射率在 0～0.7 之间，砂岩总体反射率在 0～0.6 之间（如图 3—图 5 所示）；砂岩在 900 nm 处存在铁离子吸收谱带，灰岩在 2 300 nm 处产生碳酸根离子的特征吸收，石英砂岩、白云岩等几类岩石在 2 200 nm 左右处有一个吸收谷，是由于羟基吸收所引起的[5]。

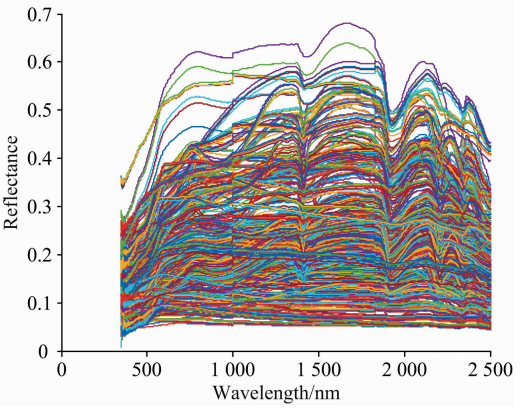

图 2 总样品数据集的岩石光谱反射率

Fig. 2 Reflectance spectra of the whole samples

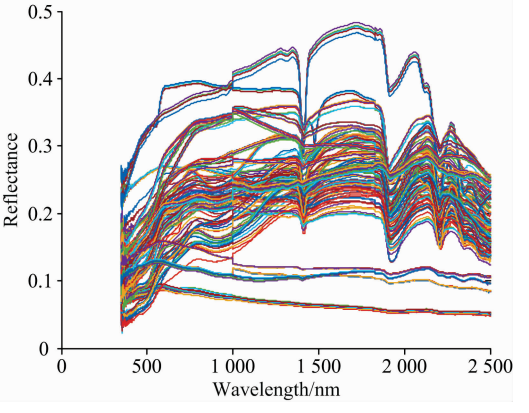

图 3 花岗岩光谱反射率

Fig. 3 Reflectance spectra of granite

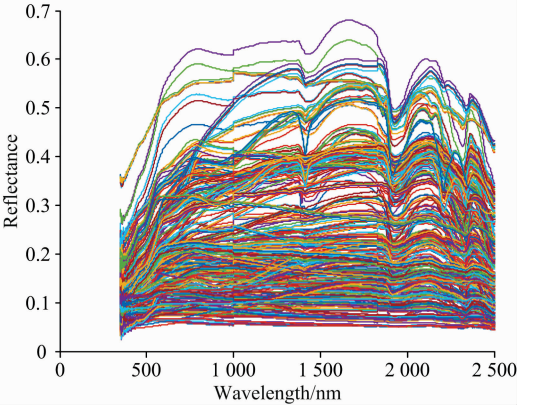

图 4 灰岩光谱反射率

Fig. 4 Reflectance spectra of limestone

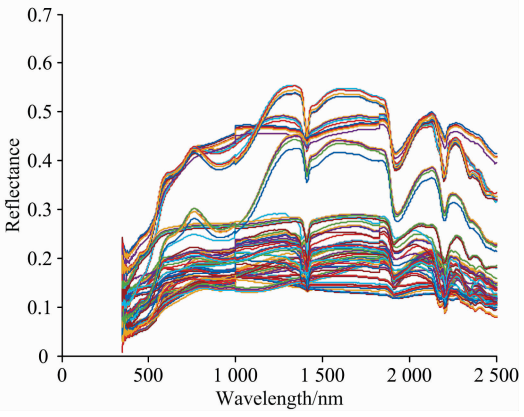

图 5　砂岩光谱反射率

Fig. 5　Reflectance spectra of sandstone

2　岩石光谱特征自动分类

2.1　决策树模型

决策树(decision tree，DT)是一个自上而下构建的树形模型，包括根节点，父节点和子节点，一个分支就代表一个测试输出。采用了决策树模型中的 CART 算法，相比传统数学统计方法分类更准确，且数据量越大越容易显现其优越性。CART 算法计算基尼系数来评判数据划分前后的分类效果，基尼系数越小证明数据纯度越高；所以选择能使分类后得到的基尼系数最小的特征，将其作为树中节点[7]。用 CART 决策树对三类岩石光谱数据的训练集建模，然后用测试集检验分类效果。将树的深度设置为 10；节点不纯度小于 0.02，即不再生成子节点，节点再划分所需最小样本数设为 2。

2.2　随机森林模型

为提高分类准确率，又选取了决策树的升级模型——随机森林(random forest，RF)，它是基于 bagging 策略的集成学习，通过多棵树对数据样本分类。包含两个随机过程：一是输入数据随机；二是分类特征随机选取。这样就得到了多颗 CART 决策树弱分类器，再将多个分类器采取投票法的策略，投出票数最多的作为最终结果[8]。RF 的参数也分为两部分：一是随机森林的 Bagging 框架参数，其中 CART 决策树的最大迭代次数设置为 1 000，划分 CART 决策树特征的评价标准选用基尼系数；二是决策树参数，深度 25，划分最大特征数为 45，节点再划分所需最小样本数设为 2。

2.3　K-最近邻模型

随机森林模型在数据噪音较大时易陷入过拟合，且数据特征过多时也会对模型准确率造成较大影响。而 K-最近邻模型(K-nearest neighbor，KNN)依据不同特征值间的距离进行分类，不存在训练过程，只是将最近的划分为一类。先将数据标准化；然后算出输入的数据与测试集的数据的距离，实验采取的计算距离方法为闵可夫斯基距离；找出距离最近的 k 个，这里 k 设置为 1；将出现最多的类别作为输入数据的类别[9]。但 KNN 需要对每个样本都予以考虑，当数据量大时计算量会很大，效率不高。

2.4　支持向量机模型

支持向量机模型(support vector machine，SVM)是通过在数据间找到距离最大处来工作的，而数据是否线性可分决定着是用硬间隔最大化还是软间隔最大化[10]。由于岩石光谱数据非线性可分，因而将数据映射到新空间，使之线性可分。核函数选高斯核函数；惩罚系数设为 10；gamma 值设定为 1。

2.5　多种模型相融合

为进一步提高岩石光谱特征自动分类的准确率，采取了融合多个不同模型的办法，即对不同模型的分类结果进行投票，选择投票最多的作为最后分类结果。在此基础上又分为硬投票和软投票，硬投票是直接对模型投票而软投票加入了权重，可以区分不同模型的重要度，但二者的基本原则都是少数服从多数。由于硬投票可在一定程度上减少过拟合现象的发生，更加适合分类模型，所以选用了硬投票方法。

3　结果与讨论

将岩石光谱数据特征分别导入 DT，RF，KNN，SVM 以及融合模型(全部基于 Python 语言编程实现)之中，分类结果如表 1 所示。可以看出在四种单一分类模型中：效果最好的是支持向量机，分类准确率为 98.76%；其次是 K-最近邻，准确率为 97.10%；然后是随机森林，准确率为 93.80%；最后是决策树模型，准确率为 88.84%。而将 RF，KNN 和 SVM 三种模型融合后得到的岩石光谱分类准确率可达到 99.17%。

表 1　不同模型的岩石光谱特征自动分类准确率

Table 1　Classification accuracy of rock spectra based on different models

模型	准确率/%
DT	88.84
RF	93.80
KNN	97.10
SVM	98.76
融合模型(RF+KNN+SVM)	99.17

4　结　论

在辽宁兴城地区实测的不同岩石反射光谱数据特征基础之上，分别利用 DT，RF，KNN，SVM 以及融合模型，进行了岩石光谱特征自动分类研究。从测试结果可以看出：第一，如果不考虑影响岩石光谱特征的各种因素，直接从光谱数据特征本身入手，可以发现机器学习模型的分类能力相对于传统的岩石光谱分类方式，效率更高、分类准确率更好；第二，四种单一机器学习模型的分类准确率高低排序为：SVM>KNN>RF>DT；第三，采用了多种模型融合学习的方法，进一步提高了岩石光谱特征自动分类的准确率，可达 99.17%。在后续研究工作中，将继续优化现有模型，使之不仅能划分岩石大类，还能准确地对细类岩性进行划分。

References

[1] WU Hui, YAN Xiao-tian, LIU Yang(吴 辉, 闫晓天, 刘 洋). Geomatics & Spatial Information Technology(测绘与空间地理信息), 2018, 41(10): 176.

[2] ZHANG Cui-fen, HAO Li-na, WANG Shao-jun, et al(张翠芬, 郝利娜, 王少军, 等). Earth Science(地球科学), 2020, 45(5): 1844.

[3] XU Qing-jun, YE Fa-wang, ZHANG Chuan, et al(徐清俊, 叶发旺, 张 川, 等). Journal of East China University of Technology(东华理工大学学报·自然科学版), 2016, 39(2): 184.

[4] ZHOU Jiang(周 江). Accuracy Analysis of Ground Objects Classification Based on ASD Portable Ground Object Hyperspectrometer and Multi-Spectral Satellite Image(基于 ASD 便携式地物高光谱仪与多光谱卫星影像对地物分类精度分析). Proceedings of Fifteen National Symposium on Mathematical Geology and Geological Information(第十五届全国数学地质与地学信息学术研讨会论文集), 2016.

[5] CHEN Sheng-bo, ZHANG Ying, GUO Peng-ju, et al(陈圣波, 张 莹, 郭鹏举, 等). Journal of Jilin University(吉林大学学报), 2015, 45(1): 320.

[6] ZHANG Ying, CHEN Sheng-bo, WANG Ming-chang, et al(张 莹, 陈圣波, 王明常, 等). Science Technology and Engineering(科学技术与工程), 2012, 12(24): 5966.

[7] Chen Wei, Li Yang, Xue Weifeng, et al. Science of the Total Environment, 2020, 701: 134979.

[8] Sun Rui, Wang Guanyu, Zhang Wenyu, et al. Applied Soft Computing Journal, 2020, 86: 105942.

[9] He Qiang, Li Xin, Nathan Kim D W, et al. Information Fusion, 2020, 55: 207.

[10] YANG Chang-bao, LIU Na, KUAI Kai-fu(杨长保, 刘 娜, 邰开富). Spectroscopy and Spectral Analysis(光谱学与光谱分析), 2019, 39(9): 2953.

Automatic Classification of Rock Spectral Features Based on Fusion Learning Model

HE Jin-xin[1], REN Xiao-yu[1], CHEN Sheng-bo[2*], XIONG Yue[1], XIAO Zhi-qiang[1], ZHOU Hai[1]

1. College of Earth Sciences, Jilin University, Changchun 130061, China

2. College of Geo-Exploration Science and Technology, Jilin University, Changchun 130061, China

Abstract The spectrum of rock is a comprehensive reflection of the physical chemistry properties, composition and structure of the rock. Rock spectral data have been applied to the study of rock classification. But unlike the mineral spectrum, the Rock spectrum has no standard database and is influenced by many disturbing factors, for example, mineral composition, structure, chemical composition, weathering strength, the error of measuring instrument, etc. The traditional rock spectrum classification model firstly preprocesses the rock spectrum to eliminate the interference. Then, some spectral features are analyzed by different methods to achieve the classification goal. However, the loss of spectral data features makes the classification of low accuracy and cumbersome operation process; efficiency is not high. Therefore, it is of great significance to establish a simple, fast and accurate automatic classification model of the rock spectrum. Machine learning can learn all the data obtained; there is no omission, greatly improving the classification accuracy. And is the direct operation to the original data, does not need the pretreatment, simplifies the process. Therefore, Xingcheng city of Liaoning Province, China was chosen as the study area, and several typical rock samples were collected. Based on the measured spectral data from the ASD Portable Spectrometer, 608 pieces of data were obtained. According to the spectral characteristics of rocks, the study is divided into three types. Firstly, the decision tree and the upgrade model of the decision tree are used as a the random forest, But when the data noise is large, random forest is easy to get into overfitting. Therefore, the knearest neighbor model, which is not sensitive to outlier is used. But KNN needs to consider every sample when the data is large, the computation will be very large, inefficient. So use Support vector machine to improve classification accuracy. The experimental results show that the order of accuracy of the four classification models is: SVM>KNN>Random Forest>Decision Tree. In order to further improve the automatic classification accuracy of rock spectral features. By fusing several different models. That is to vote on the classification results of different models, choose the most votes as the final classification results. Since hard voting can reduce the occurrence of over-fitting to a certain extent, it is more suitable for classification models. In this paper, we use a hard voting method to fuse three machine learning models: RF, KNN and SVM. The final classification accuracy can reach 99.17%. To sum up, it is feasible, accurate and efficient to classify rock spectral features automatically based on the fusion learning model.

Keywords Rock spectral classification; Decision tree; Random forest; K-nearest neighbor; Support vector machine; Model fusion

* Corresponding author

(Received Dec. 15, 2019; accepted Apr. 11, 2020)