Machine Learning in Geology: Challenges and Prospects

Yuheng Liu*

School of Ocean and Earth Science Tongji University Shanghai, China

*Corresponding author email: 2050337@tongji.edu.cn

Abstract. Machine Learning methods, along with high-performance computing, is creating chances for data intensive science in various domains such as Geology. Known for its precision and efficiency, Machine learning has renovated many fields of science thoroughly. However, when it comes to Geology, Machine Learning methods aren't fully applied, lacking an overall frame to arrange those works with scattered aims and methods. This paper presents a comprehensive review of research focusing on application of Machine Learning methods in Geology studies. The works analyzed were categorized in three aspects of Geology studies: mineral prospecting and mapping, land-cover monitoring; geochemical anomalies detection; earthquake monitoring and geological sample identification. The selection and classification of the presented articles demonstrate how Geology benefits from Machine Learning methods.

Keywords: machine learning; geology.

1. Introduction

Geology is the rule of the evolution of Earth. The primary target of Geology is the lithosphere under the surface whose composition and inner structure reveals the history of Earth [1]. Geology is the key to grasping the rule of our planet which helps us to untap the greatest potential as well as secure our well-being from underlying risk. Since the target geology studies are large and complicated, traditional geological research tend to be troubled by multiple data sources and low accuracy [2]. As technology advances many new methods have been introduced to geological studies covering geophysical, geochemical, remote sensing and other fields. These technologies deal with more data and provide better accuracy.

Geological big data are a new notion introducing big data technology to Geological fields. As novel technologies are applied, data storage costs less and more data are gathered, geological data are gradually characterized by elements of big data involving quantity, value, diversity, and timeliness [2]. The conventional ways to process data are increasingly uncapable of meeting the requirements in terms of processing methods and efficiency. In big data technologies, Machine Learning methods, along with the development of high-performance computing, have offered a more effective approach which guarantees a good prospect for models in data processing to be utilized in the Geology industry.

Geological science is a rule about observation in which geologists deduce the whole face from the observed existing phenomena[3].Traditional geological research methods, limited by researchers' subjective experience and data from certain locality, sometimes fail to draw a conclusive result, however, Machine Learning methods can integrate and use the great amount of geological data to conclude geological features; view geological phenomena objectively and efficiently; explore the rules behind geological activity and draw more scientific conclusions[4]. Therefore, geological big data technologies involving Machine Learning methods sure to bring more vigor into the development of Geology

To trace, efforts have been made to apply Machine Learning methods to Geology in aspects including mineral prospecting and mapping (MPM), land-cover monitoring, geochemical anomalies detection, geological sample identification and earthquake monitoring. This paper will discuss the application of several Machine Learning methods in the above-mentioned fields including Artificial Neural Networks (ANNs), Decision Trees (DTs), Random Forest (RFs), and Support Vector Machines (SVMs), etc.

2. Machine Learning algorithms

2.1. Artificial Neural Networks

The most popular method in nonparametric and nonlinear classification involves Artificial Neural Networks (ANNs), as shown in Fig.1. There are many kinds of ANNs. This section will briefly introduce a popular ANN: the feed-forward propagation neural network.[5] In the core, neutrons (units or nodes) function as the fundamental processing sections of an artificial neural network. In a neural network, units in the form of layers are connected for information to flow unidirectionally, starting from the input unit to the output layer passing through units positioned in the hidden layer. A neuron, with the help of a nonlinear function, perform linear regression. Weights are used to connect neurons of multiple layers.

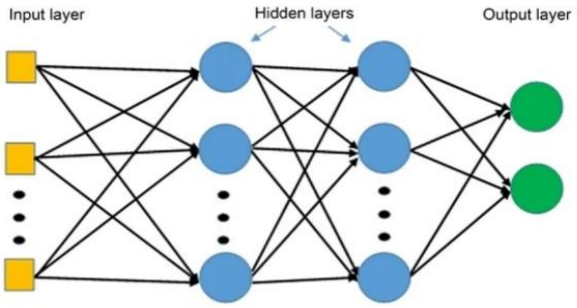

Figure 1. Simplified mechanism of ANNs [5]

2.2. Decision Trees

Decision trees (DTs), accompanied with neural networks, are the most popular machine learning algorithms in geosciences, as shown in Fig.2 [6]. Owing to their simplicity, readability (possibility to be demonstrated graphically) and relatively low computation cost, DTs are being used increasingly. A DT organizes restrictions or limiting factors hierarchically which are continuously applied from the root to the final node/leaf of the tree. Decision trees differs from ANNs in terms of its transparency and convenience to interpret. To conclude the DT from a dataset, an evaluation measure of every valuable feature is applied to maximize the inter node difference. DTs are classified into two methodologies: classification trees and regression trees (RTs).

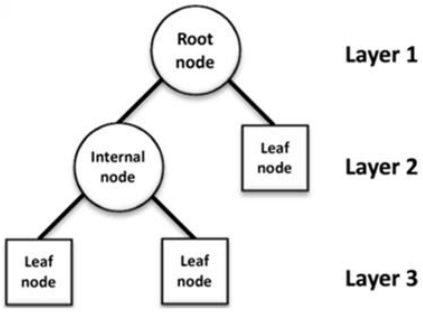

Figure 2. Sample process of DTs [6]

2.3. Random Forests

Random Forests (RFs) is a regression method combining the utility of plenty of DT algorithms to categorize or predict the value of a variable, as shown in Fig. 3 [7]. When an (x) input vector was given containing values with different evidential features analyzed for a certain training area, RF builds a number K of regression trees and averages the results.

In order to reduce the impact of anomalies in the original data, RF adds the diversity of the trees by bagging through which data are trained by resampling randomly with no deletion of the original data to create different subsets $\{h(x,\Theta k), k = 1,\ldots,K\}$, where $\{\Theta k\}$ are independent random vectors

distributed in the same pattern. Therefore, some data may appear repeatedly in the training, while others might never be applied. In this way, the system is more stable as it acts robust when facing subtle changes in the input data and increases accuracy. Meanwhile, when a tree grows in the RF, the best split point within a subset of randomly selected input evidential features from the overall set is used. Although this reduces the strength of every single tree, the generalization error is declined as well. Another feature is that there is no pruning in RF which enables lightness from the perspective of computing.

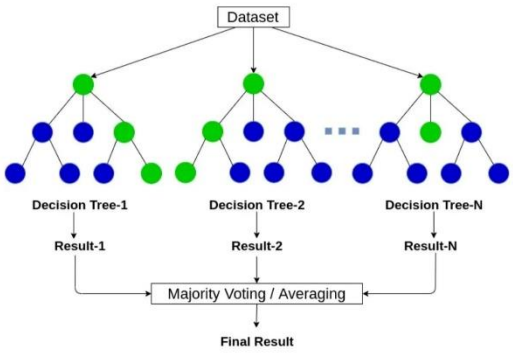

Figure 3. Working process of RFs [7]

2.4. Support Vector Machines

In machine learning, support vector machines (SVMs) are categories as supervised learning methods with relevant algorithms to analyze data for classification and regression purposes. Developed Vladimir Vapnik and his colleagues, as shown in Fig.4 [9], SVMs are known for its stability, dependent on the VC theory put forward by Vapnik(1982,1995) and Chervonenkis(1974). Based on training sets, which each fall into two categories, SVMs set up a model in which new examples are labeled into two categories, functioning as a non-probabilistic binary linear classifier. SVMs converts data into mapping points to maximize the length of the gap. New examples are then presented into the same part in space and belong to the category which that side stands for.

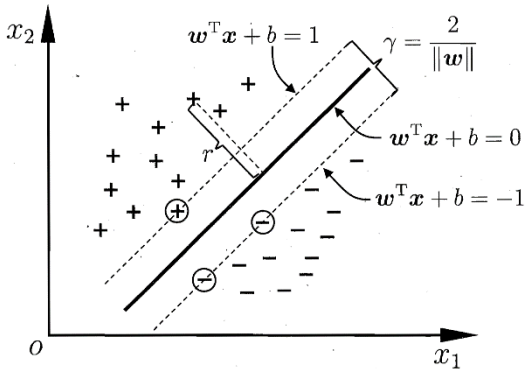

Figure 4. Hyper-plane in SVMs [9]

3. Applications of AI methods in Geology

3.1. Applications in MPM and Land-cover Monitoring

Machine learning algorithms (MLAs) are strong data-driven processing tools providing automatic approaches for identifying patterns of high-dimensional data. Therefore, when it comes to mineral prospecting and mapping (MPM) and land-covering monitoring (LCM) problems which rely heavily on the remotely sensed geophysical data with rapidly expanding data capacity, MLAs have a wide space to perform its superiority.

3.1.1. Application in mineral prospecting and mapping (MPM)

Recently, researchers attached great importance to the application of core machine learning technologies in MPM and LCM. To data, there has been multiple examples of supervised classifications methods in these two fields. Embrown et al. [10] first indicated the possibility of using neural network methods in integrating large multi-source datasets and mineral prospectively mapping, claiming its superiority over other conventional methods. Porwal et al. [11] specified a case using a certain type of Artificial Neural Network (RBFLN) to map the prospectively of SEDEX-type base metal deposits in a spot in western India. The result proved to be a success; the spatial distribution of the high favorability zones matches the theoretical models of base metal metallogeny in the province.

V.F. Rodriguez-Galiano et al. [12] used Random Forest regression to predict gold deposits using more than 50 known occurrences of epithermal Au deposits. The results showed that RF regression was more accurate and efficient than the conventional LR method, in addition, RF regression was able to predict the importance of variables precisely from GIS layers and was capable of n handling complex data of different statistical distributions other than linear distribution. Carranza and Laborte [13] utilized Random Forest technique to predict prospectively of hydrothermal Au-Cu deposits in a study area in Philippines where there are less than 20 mineral deposit occurrences. The results indicated that RF outperformed EB (evidential belief) modelling and was able to handle missing values instead of simply represented these as maximum uncertainties, proving RF's potential in data-driven predictive modelling of mineral prospectively.

G.M.Foody et al. (2004)[14] indicated that a single multi-class SVM classification can be used to generate very precise classifications in multi-class image classification. Overall, the SVM classification were more accurate that other comparable classifications. The result was based on a small number of training sites, when training sites increased into great amount, proper training data were needed. M.Abedi et al. (2012) [15] first used SVM to predict mineral prospectively of copper deposit indicating SVM's potential in prospectively mapping, especially in high potential areas and increased the resolution of prospectively mapping from binary classification into multi-classification. The Cracknell and Reading [16] made a comparison between five machine-learning algorithms for geological mapping in which they implemented a logical and persuasive comparison of five MLAs: k-Nearest Neighbors, Naïve Bayes, Random Forests, Support Vector Machines, and Artificial Neural Networks, under the background of a supervised lithology classification mission. The result indicated that Random Forests were recommended as the priority for lithological supervised classification based on remotely sensed geophysical data which are efficient in training and computing and enjoy high stability considering changes in parameter values and are at least of the same accuracy that other MLAs experimented. Rodriguez-Galiano et al. [17] respectively applied Artificial Neural Networks (ANNs), Decision Trees (DTs), Random Forests (RFs)and Support Vector Machines (SVMs) to MPM. The conclusion indicated that, RFs outperform other methods. Meanwhile, SVMs accompany RFs in the highest accuracy surpassing ANNs with relatively good degree of accuracy in the bare context of a certain combination of adjusting parameters. Overall, the performance of RFs regarding different parameter combination stands out in terms of stability and accuracy.

Considering the results in each category from the aspect of positive or negative appearance, accuracy of different MLAs tends to vary. RFs were able to depict both areas with same accuracy, while ANNs and SVMs classified both areas differently, underestimating deposit areas. Other statistical measures comparing mapping quality also draw the conclusion that RFs were the best.

3.1.2. Application in land-cover monitoring

X.Song et al. [18] presented a conclusive comparison of ANN and SVM implemented in remote sensing for land cover classification in a study area in the north of China and found that well-trained SVMs performs slightly better than those of ANNs and SVMs better perform on small training set and avoid inadequate training. M.A.Friedl and C.E.Brodley (1997)[19] indicated that in land cover mapping, from the perspective of classification accuracy, the various decision tree algorithms discussed all performs better than the maximum likelihood and linear discriminant function

classification methods. J.Rogan et al. (2003)[20] presented a case study of the potential of classification trees for land-cover change monitoring in regions with different vegetation types and disturbance in a study area in California.

P.O. Gislason et al. [21] explored the application of Random Forest classifier in land cover categorization and compared the precision of the Random Forest classifier with other overall approaches on multi-source remote-sensed and geographic data. The results indicated that accompanied by helpful analytical methods, the Random Forest classifier is optimal for classifying remote-sensed and geographic data from different sources where traditional methods tend to fail. Yang [22] further discussed factors in parameter settings that can affect the accuracy of SVM in land-cover classification and found that kernel types and error penalty can affect the classification accuracy greatly and found that selecting parameter settings carefully can help improve the performance of the SVM in land-cover classification. Rodriguez-Galiano and M. Chica-Rivas [23] compared different machine learning methods (ANN,CT, RF, SVM) for land covering mapping of a Mediterranean area. The result indicated that similar high accuracies were achieved for all the four algorithms, nevertheless, some differences were also exposed: RF was the most precise classifier with much simpler parameterization and SVM was the most stable classifier to noise and data reduction.

3.2. Applications in Geochemical Anomalies Detection

Due to complicated geological environment where distribution patterns of geochemical data are mostly unknown, conventional statistical methods tend to cost huge time and effort when recognizing geochemical anomalies. If MLAs could be applied to this field which are good at handling the nonlinear relationship without assumptions about how the data are distributed, the effectiveness and efficiency of geochemical anomalies detection can be greatly improved. For example, Beucher et al. (2014) [24] used RBFLN (an ANN model) to discover soils' distribution in a study area in southwest Finland and this model is recommended for acid sulfate soil mapping which helps to create reliable and referenceable maps and represents its strength within the acid sulfate soil mapping process, increasing speed and efficiency.

Twarakavi et al. (2006) [25] applied a SVM and a LS-SVM to depict graphically the concentration of arsenic trioxide with the help of that of gold in Alaskan sediments. The results indicate that SVM and robust LS-SVM improve model's performance and predicting abilities compared with neural networks and kriging techniques and the robust LS-SVM outperforms the SVM. Chen et al. (2014) [26] successfully recognized geochemical anomalies in southern Jilin, China with the help of the continuum-limited Boltzmann machine method in the context with complicated geological conditions in which unexplored complicated distribution pattern of probability of geochemical sample population fails the traditional method involving Gaussian distribution. Gonbadi et al. (2015) [27] used Kerman Province, Iran, as a study area to recognize Cu-related porphyry geochemical anomalies with the assistance of supervised machine learning. Results indicate that feature-selecting algorithms is crucial for improving the precision and capacity of generalization of the applied classifiers. The best performance is achieved by wrapper mode subset selection method combined with a genetic algorithm (GA) approach. Involved classification methods surpass Gaussian linear discriminant analysis (GLDA) and in accuracy, stability, and reliability. Chen et al. (2019) [28], to raise the ability of a neural network to implement multi-dimensional anomaly detection, used a spatially limited Neural Network approach based on multi-autoencoder.

Chen et al. (2019) [29] used a non-interactive net-structure in MCAE (multi-convolutional auto-encoder approach) to precisely discover geochemical anomalies. The results indicate that the recognition domain determination greatly improves the quality of anomaly detection and MCAE surpassed different exist approaches comprehensively in handling missing spatial structural patterns of geochemical context and achieved results with great consistency to the reality.

The above studies indicate that MLAs are helpful for recognizing multivariate geochemical anomalies.

3.3. Applications in Geological Sample Identification and Earthquake Monitoring

When it comes to geological sample identification, there exists great scope for Machine Learning methods to perform since the degree to which data should be classified is clear and data comes from a wide range. History identification can be well-referenced for current studies which can serve as the training material for AI model to analyze and then to classify the samples more efficiently and accurately since mistakes caused by human sorting are reduced [31]. To trace, deep learning models are proved to be well-applied in the field of mineral sample identification. The application of Machine Learning methods has a broad prospect in the future.

Closely linked to a country's economy and people's living standard, earthquake monitoring has great practical value. In the era of information, application of earthquake monitoring with the help of big data technologies involving Machine learning methods are sure to add more accuracy and efficiency which will greatly benefit related operations.

4. Discussion

Machine learning methods are widely applied in computer and data science, however, when it comes to geology, they are in there prime. So far, there lacks a comprehensive combination of theories and system to solve the theoretical problems of geology with the aid of Machine Learning methods.

The relevant achievements of geology include mineral prospecting and mapping, land-cover monitoring [10-23], geochemical anomalies detection [24-29] and other areas which proves the big data technologies have a wide scope for application in geology. This passage mainly focuses on the discussion of applications of Machine Learning methods in mineral prospecting and mapping, land-cover monitoring and geochemical anomalies detection where there have been sufficient case studies and comparison of AI methods to discuss, for the other fields like geological sample identification and earthquake monitoring, where there is currently a lack of relevant research in application of Machine Learning methods, this passage only analyzes its potential conceptually.

In geology, multi-scale effect, spatiotemporal heterogeneity and spatiotemporal correlation are important features [32]. As the research in geology derives from single scale to multi-scale and from static to dynamic [33], the demand for new ways and concepts is heavy, which provide great space for Machine Learning methods to satisfy this need, also, as data validity, data bias and model reliability gain more and more attention, relevant geological studies need specialization in big data domain which fits the demand as well. Considering the four features of big data---massive scale, diverse types, low value density and rapid data generation [33]— big data analysis requires greater computing capacity and more effective process methods. Therefore, to better apply big data technologies to building spatiotemporal database of geology based on big data platforms, strong computing power is necessary, also there are other inherent issues in scientific and technical aspects to be solved in the future.

5. Conclusions

In conclusion, recent upgrade in technology has allowed for success in Geological study. These include mineral prospecting and mapping, land-cover monitoring, geochemical anomalies detection, geological sample application and earthquake monitoring. In the mineral prospecting and mapping and land-cover monitoring aspects, every one of the four methods are already successfully applied and comprehensively speaking, RF has a comparative edge over other three AI methods. When it comes to geochemical anomalies detection, the four methods are all implemented in different specific examples. However, in the two remaining fields, there only exists theoretical possibilities with currently no further investigation and experiment of those methods in these fields.

Considering the characteristics of big data and evolving trend in Geology, the application of Machine Learning methods in Geology has a quite promising future, especially when some methods like RF has been proven to be feasible and accurate in certain Geology field like ore mineralogy studies. It is safe to believe that more AI models will be applied, it is just a matter of time. However, for AI

methods to fully perform in Geology field, a universal AI platform based on spatiotemporal database is to be built which is of much difficulty. But once it is built, it will open up new possibilities for the new epoch of Geology.

References

[1] W. R. Muehlberger, Earth Science[J]. Science, 1966, 152 (3724):950-951.

[2] S. F. Huang, X. H. Liu, Thinking about the application of geological big data and geological information development[J]. China Mining Magazine, 2016, 25(08):166-170.

[3] L. N. Huynh, Y. Lee, R. K. Balan, DeepMon: Mobile GPU-based Deep Learning Framework for Continuous Vision Applications[A]. In: The International Conference, 2017[C].

[4] Q. Zhang, X. Y. Jia, Z. Wu, J. R. Wang, S. T. Jiao, W. F. Chen, Big data will lead to a great change in Geological Science Research[A].2015 China Geoscience union annual meeting, Beijing, China, 2015 [C].

[5] Rumelhart, D.E., Hinton, G.E., Williams, R.J., 1986. Learning representations by backpropagating errors. Nature 323, 533–536.

[6] Friedl, M.A., Brodley, C.E., 1997. Decision tree classification of land cover from remotely sensed data. Remote Sens. Environ. 61, 399–409.

[7] Guo, L., Chehata, N., Mallet, C., Boukir, S., 2011. Relevance of airborne lidar and multispectral image data for urban scene classification using random forests. ISPRS J. Photogramm. Remote Sens. 66, 56–66.

[8] Breiman, L., 2001. Random forests. Mach. Learn. 45, 5–32.

[9] Cortes, Corinna; Vladimir Vapnik (1995). "Support-Vector Networks". Machine Learning. 20 (3): 273–297.

[10] Brown, W.M., Gedeon, T.D., Groves, D.I., Barnes, R.G., 2000. Artificial neural networks: a new method for mineral prospectivity mapping. Aust. J. Earth Sci. 47, 757–770.

[11] Porwal, A., Carranza, E.J.M., Hale, M., 2003. Artificial neural networks for mineral-potential mapping: a case study from Aravalli Province, Western India. Nat. Resour. Res. 12, 155–171.

[12] Rodriguez-Galiano, V.F.; Chica-Olmo, M.; Chica-Rivas, M. (2014). Predictive modelling of gold potential with the integration of multisource information based on random forest: a case study on the Rodalquilar area, Southern Spain. International Journal of Geographical Information Science, 28(7), 1336–1354.

[13] Carranza, Emmanuel John M.; Laborte, Alice G. (2016). Data-Driven Predictive Modeling of Mineral Prospectivity Using Random Forests: A Case Study in Catanduanes Island (Philippines). Natural Resources Research, 25(1), 35–50.

[14] Foody, G.M., and A. Mathur, 2004a. A relative evaluation of multiclass image classification by support vector machines, IEEE Transactions on Geoscience and Remote Sensing, 42(6):1335–1343

[15] Abedi, M., Norouzi, G.H., Bahroudi, A., 2012. Support vector machine for multi-classification of mineral prospectivity areas. Comput. Geosci. 46, 272–283.

[16] M. J. Cracknell, A. M. Reading, Geological mapping using remote sensing data: A comparison of five machine learning algorithms, their response to variations in the spatial distribution of training data and the use of explicit spatial information[J]. Computers & Geosciences, 2014,63(1):22-33.

[17] V. G. Rodriguez, M. C. Sanchez, M. O. Chica, et al. Machine learning predictive models for mineral prospectivity: An evaluation of neural networks, random forest, regression trees and support vector machines[J]. Ore Geology Reviews, 2015,71:804-818.

[18] Song, Xianfeng; Duan, Zheng; Jiang, Xiaoguang (2012). Comparison of artificial neural networks and support vector machine classifiers for land cover classification in Northern China using a SPOT-5 HRG image. , 33(10), 3301–3320.

[19] Friedl, M.A., Brodley, C.E., 1997. Decision tree classification of land cover from remotely sensed data. Remote Sens. Environ. 61, 399–409.

[20] Rogan, J., Miller, J., Stow, D., Franklin, J., Levien, L., Fischer, C., 2003. Land-cover change monitoring with classification trees using Landsat TM and ancillary data. Photogramm. Eng. Remote Sens. 69, 793–804.

[21] Gislason, P.O., Benediktsson, J.A., Sveinsson, J.R., 2006. Random forests for land cover clas sification. Pattern Recogn. Lett. 27, 294–300.

[22] Yang, X., 2011. Parameterizing support vector machines for land cover classification. Photogramm. Eng. Remote Sens. 77, 27–37

[23] Rodriguez-Galiano, V.F., Chica-Rivas, M., 2012. Evaluation of different machine learning methods for land cover mapping of a Mediterranean area using multi-seasonal Landsat images and digital terrain models. Int. J. Digit. Earth 7, 492–509.

[24] A. Beucher, P. Österholm, A. Martinkauppi, et al. Artificial neural network for acid sulfate soil mapping: Application to the Sirppujoki River catchment area, south-western Finland[J]. Journal of Geochemical Exploration, 2013,125(1):46-55.

[25] N. K. C. Twarakavi, D. Misra, S. Bandopadhyay, Prediction of Arsenic in Bedrock Derived Stream Sediments at a Gold Mine Site Under Conditions of Sparse Data[J]. Natural Resources Research, 2006, 15(1):15-26

[26] Y. Chen, L. Lu, X. Li, Application of continuous restricted Boltzmann machine to identify multivariate geochemical anomaly[J]. Journal of Geochemical Exploration, 2014,140(4):56-63.

[27] A. M. Gonbadi, S. H. Tabatabaei, E. J. M.Carranza, Supervised geochemical anomaly detection by pattern recognition[J]. Journal of Geochemical Exploration, 2015, 157:81-91.

[28] L. Chen, Q. Guan, Y. Xiong, J. Liang, Y. Wang, Y. Xu, A Spatially Constrained Multi-Autoencoder Approach for Multivariate Geochemical Anomaly Recognition. Computers and Geosciences, 2019, 125: 43–54.

[29] L. Chen, Q. Guan, B. Feng, H. Yue, J. Wang, F. Zhang, A Multi-Convolutional Auto-encoder Approach for Multivariate Geochemical Anomaly Recognition, Minerals, 2019, 9(5): 270.

[30] Q. Zhang, X. Y. Jia, Z. Wu, J. R. Wang, S. T. Jiao, W. F. Chen, Big data will lead to a great change in Geological Science Research[A].2015 China Geoscience union annual meeting, Beijing, China, 2015 [C].

[31] M. J. Cracknell, A. M. Reading, Geological mapping using remote sensing data: A comparison of five machine learning algorithms, their response to variations in the spatial distribution of training data and the use of explicit spatial information[J]. Computers & Geosciences, 2014,63(1):22-33.

[32] K. Regenauerlieb, M. Veveakis, T. Poulet, et al. Multiscale coupling and multiphysics approaches in earth sciences: Applications[J]. Journal of Coupled Systems & Multiscale Dynamics, 2013,1(3):281-323.

[33] M. Hilbert, Big Data for Development: A Review of Promises and Challenges[J]. Development Policy Review, 2016,34(1):135-174