EarthDataMiner: A Cloud-Based Big Earth Data Intelligence Analysis Platform

Jie Liu, Wei Wang, Hua Zhong

Institute of Software, Chinese Academy of Sciences, Beijing, China
ljie@otcaix.iscas.ac.cn

Abstract. Big Earth data mining analysis involves complex data processing, pattern extraction and knowledge discovery processes, and requires easy to use integrated software tools to deal with large-scale data. With the support of Big Earth Data Science Engineering Project (CASEarth) of the Chinese Academy of Sciences Strategic Priority Research Program, our team are ongoing developing EarthDataMiner, a cloud-based Big Earth data intelligence analysis platform for scientists in the field of Earth sciences, which brings together cross-domain common models, introduces advanced artificial intelligence algorithms, supports online satellite imagery visually interactive analysis, and achieves efficient task processing through a series of distributed execution systems.

1. Motivation

Big Earth data has multi-domain crossover characteristics and diverse analytical processing requirements [1]. The field of artificial intelligence has made a breakthrough in recent years [2]. How to promote scientific discovery in the field of Big Earth data with mass and multi-source data, has become the focus of scientists and industry. The Big Earth data mining analysis and knowledge discovery in the multidisciplinary and cross field intersecting background depend on building a set of efficient, easy to use and extensible software system. As we know, Google Earth Engine [3] is an excellent cloud system provided by Google for online visual analysis and processing of a large number of global-scale geoscientific data (especially satellite data). However, Google Earth Engine is not designed to deal with multi-domain various Earth data and it is difficult to integrate emerging artificial intelligence algorithms.

With the support of CASEarth project, our team are ongoing developing EarthDataMiner, which is an integrated and customizable intelligent analysis framework for Big Earth data. EarthDataMiner focus on how to share data and algorithm together crossover scientists. It supports the interactive construction of intelligent analysis models, and provides systems and tools for the rapid development of analytical algorithms for Earth science research.

2. System Architecture

The system architecture of EarthDataMiner is shown in Figure 1. EarthDataMiner use cloud computing technology to provide distributed big data processing engine and machine learning engine for the special process of large scientific data analysis and processing. Data repository supports scientists to find, import and share public and personal data resources conveniently. Algorithms repository provides common used algorithms, and support scientists to publish and share new algorithms and models. Intelligent analysis environment provides a variety of intelligent analysis tools, such as python code IDE (integrated development environment), Jupyter notebook, and workflow tool...
developed by our team. Data visualization tool with charts and GIS is also provided for online satellite imagery visually interactive analysis.

The system provides online mining and analysis services through web browsers, and users can carry out whole analysis work through registered accounts. In this process, the system need to ensure the security and isolation of scientists’ data and analysis tasks.

3. System Progress
At present, EarthDataMiner can provide online analysis code development for remote-sensing data in a similar way to Google Earth Engine. Meanwhile, EarthDataMiner has the following featured innovations:

- Using EarthDataMiner, scientist can online access and analyse about million gigabytes of data related to Earth sciences in the five-year CASEarth project.
- The important cross-domain discoveries require not only data sharing, but also models and algorithm sharing. EarthDataMiner defines standard specifications to support scientist share well-trained models or algorithm codes, and to accumulate more and more excellent algorithm models through the power of the community.
- It integrates advanced intelligent analysis technologies such as deep learning and automated machine learning (AutoML) to help scientists develop new algorithms and verify them quickly online.
- It supports scientists to use the popular Python language for online algorithm development verification, and provide integrated tool support, including code management, data management, and task management.

Based on EarthDataMiner, some cases have been developed with the cooperation with scientists, such as forest detection, water bird identification, and remote sensing object recognition. EarthDataMiner is planned to be published as an online cloud service using the cloud computing resources provided by CASEarth.

References
[1] Huadong Guo et al, International Journal of Digital Earth, Volume 10, 2017 - Issue 1.
[2] LeCun, Y., Bengio, Y. & Hinton, G, Nature, 521, 436–444 (2015).
[3] Gorelick N, Hancher M, Dixon M, et al, Remote Sensing of Environment, 2017, 202:18-27.