Current Status and Trends of Intelligent Analysis Software for Scientific Big Data (Postprint)

Authors: Zhong Hua, Liu Jie, Wang Wei

Date: 2023-03-19T00:00:00+00:00

Abstract

In recent years, the field of artificial intelligence has achieved breakthrough progress. How to adopt novel artificial intelligence technologies in natural sciences to accelerate scientific discovery has become a focal point for both scientists and the industrial community. Under the backdrop of multidisciplinary and cross-domain integration, scientific big data mining, analysis, and knowledge discovery rely on constructing an efficient, user-friendly, and scalable intelligent analysis software system for scientific big data, which provides support through learning models, algorithms, and development tools for complex data processing, analysis, pattern extraction, and knowledge discovery. This article conducts a comprehensive investigation of representative intelligent analysis software systems in typical scientific domains, comparatively analyzes their commonalities and differences, and explores their development trends. Building upon this foundation, the article proposes an integrated, customizable intelligent analysis framework oriented toward scientific big data, which supports scientists in interactively constructing intelligent analysis models and executing them efficiently, thereby providing system and tool support for rapidly advancing scientific discovery research.

Full Text

Current Situation and Trends of Intelligent Analysis Software for Scientific Big Data

Abstract

In recent years, artificial intelligence has achieved breakthrough progress. How to adopt new AI technologies in natural sciences to promote scientific discovery has become a focal point for both scientists and industry. In the context of interdisciplinary and cross-domain research, scientific big data mining, analysis, and knowledge discovery depend on constructing an efficient, user-friendly,

and scalable intelligent analysis software system that provides learning models, algorithms, and development tool support for complex data processing, analysis, pattern extraction, and knowledge discovery. This paper selects representative intelligent analysis software systems from typical scientific domains for comprehensive investigation, compares and analyzes their commonalities and differences, and discusses development trends. On this basis, we propose an integrated and customizable intelligent analysis framework for scientific big data that supports scientists in interactively building intelligent analysis models and executing them efficiently, providing system and tool support for rapid scientific discovery research.

Keywords: scientific big data, intelligent analysis, data-intensive scientific discovery, software system

---

The Fourth Paradigm of Scientific Discovery

In 2007, Turing Award winner Jim Gray delivered his famous speech "The Fourth Paradigm: Data-Intensive Scientific Discovery," categorizing scientific research into four paradigms: experimental induction, model deduction, simulation, and data-intensive scientific discovery, thereby proposing the new perspective of "scientific big data" widely known as the "fourth paradigm" [1]. After a decade of technological development, advanced technologies such as deep learning have achieved breakthrough progress in AI fields including image, speech, and natural language processing. In recent years, scientists in natural sciences have also followed this trend, employing new technologies like deep learning based on the new model of scientific big data-driven research, achieving a number of significant scientific discoveries published in authoritative academic journals such as *Science* and *Nature*. However, big data-driven scientific research remains challenging for most scientific teams due to its heavy reliance on advanced information technologies.

Data-intensive scientific discovery is inseparable from software system support. This paper focuses on typical software systems for scientific big data intelligent analysis developed over the past decade. From the perspective of applicability, scientific big data intelligent analysis software can be simply divided into two categories: general-purpose and domain-specific. Based on deployment models, these software systems can be classified into three types—single-machine environment, distributed environment, and cloud computing environment—which also represent three stages of intelligent analysis software development.

Classification by Deployment Environment

Single-Machine Environment Intelligent Analysis Software In commercial data analysis software, Matlab provides a high-level programming lan-

guage and interactive environment for algorithm development, data visualization, data analysis, and numerical computation, with widespread application across numerous scientific fields. Among many open-source free data analysis software, R [3], Scikit-Learn [4], and Weka [5] are typical representatives. R is a language for statistical analysis and graphics, providing rich statistical analysis functions, and users can enhance R's functionality by developing and installing extension packages. Python has numerous scientific data analysis algorithm libraries, including Scikit-Learn, which is widely used in machine learning and data mining. The Weka data mining platform, developed based on Java, provides a visual, drag-and-drop analysis process design interface and integrates a large number of data preprocessing and machine learning algorithms. These software systems were originally designed to run in single-machine mode and cannot process large-scale data based on distributed storage, presenting inherent limitations in big data scenarios. Additionally, these software systems lack effective support for deep learning technologies.

Distributed Environment Intelligent Analysis Software In distributed environments, big data analysis software provided by the open-source community has become mainstream, with Hadoop Mahout and Spark MLlib [6] being typical representatives. Researchers have solved distributed parallel mining problems by leveraging Hadoop and Spark frameworks, providing typical machine learning algorithms and models. In recent years, a batch of open-source deep learning frameworks has emerged, such as TensorFlow, Caffe, CNTK, and MXNet, for building and training deep neural network models, supporting distributed computing and heterogeneous computing . Although these open-source software systems provide rich algorithm libraries and efficient distributed computing platforms, they require professional programming development and system configuration skills with steep learning curves, making them difficult for scientific teams to use.

Cloud Computing Environment Intelligent Analysis Software Providing big data intelligent analysis services through cloud platforms has become a standard offering of large public cloud platforms, and "machine learning as a service" (MLaaS) has become a development trend among leading cloud platform vendors. Azure Machine Learning (Azure ML) is a machine learning analysis service provided by Microsoft Azure [7], offering not only a large number of general machine learning analysis algorithms but also an interactive graphical development interface for data scientists. Similar MLaaS platforms include Aliyun PAI. These systems typically support only a specific development language and application programming interface (API), and users cannot independently expand algorithm libraries, resulting in platform lock-in issues. In addition to big data intelligent analysis services provided by public cloud vendors, some scientific teams have deployed interactive analysis software with "browser/server" architecture on public or private clouds, implementing a "simplified" MLaaS. For example, Jupyter Notebook is interactive analysis software

supporting "browser/server" architecture, enabling editing and running multiple programming languages through a browser, with data processing, numerical simulation, statistical modeling, machine learning, and visualization performed on the server side.

General-Purpose vs. Domain-Specific Software

General-Purpose Scientific Big Data Intelligent Analysis Software With the rapid development of big data and AI technologies, numerous software systems have emerged. This paper selects commonly used, representative intelligent analysis software systems and classifies them according to deployment mode as described above.

Domain-Specific Scientific Big Data Intelligent Analysis Software Natural sciences include numerous subdomains, each with specialized scientific data analysis software. This paper selects several representatives for analysis and divides them into two categories:

Classic Domain-Specific Scientific Data Analysis Software. These systems are specifically developed by scientists in particular domains and are suitable for specialized processing, computation, and analysis of data in those fields. ROOT is open-source software developed by CERN, primarily used for data processing, scientific computing, and visual analysis in particle physics experiments, providing mathematical and statistical tools, parallel processing, neural networks, and multivariate analysis packages, making it a typical tool for data analysis in high-energy physics. AstroML is a machine learning and data mining algorithm package for astronomy [8], built on Python algorithm libraries such as NumPy, SciPy, and Scikit-Learn, providing loaders for multiple open astronomical datasets and numerous case studies for astronomical data analysis and visualization. Currently, such domain-specific software still adopts single-machine deployment, cannot perform distributed parallel big data processing and analysis, and has not yet integrated or supported deep learning technologies.

Emerging Domain-Specific Scientific Data Analysis Software. This category refers to analysis software employing new technologies such as big data, machine learning, and cloud computing. SDAP is currently an incubation project of the Apache Software Foundation, a scientific big data analysis platform for geophysical oceanography. SDAP relies on the NEXUS system for big data processing, a software project developed by NASA's Jet Propulsion Laboratory (NASA/JPL) that employs Map/Reduce distributed parallel computing technology to conduct scientific analysis on large datasets collected by various NASA missions. The National Energy Research Scientific Computing Center (NERSC) hosts the primary scientific computing facilities for the U.S. Department of Energy's Office of Science. Recently, NERSC has supported applying

deep learning to climate research, neutrino experiments, and neuroscience research, achieving a batch of breakthrough scientific discoveries. Researchers at Verily Life Sciences (formerly Google Life Sciences) developed a deep learning software tool called DeepVariant , which converts genomic information into images for analysis, significantly improving the accuracy of genetic variant identification. Google Earth Engine is a cloud platform provided by Google for online visual analysis and processing of massive global-scale Earth science data (especially satellite data), enabling scientists in relevant fields to utilize the platform's long-term near-Earth satellite data and thousands of cloud servers for online data processing and analysis, already achieving a number of high-visibility research results. It is evident that Google Earth Engine's characteristics—domain-specific massive data, cloud-based distributed parallel computing, online mining analysis algorithm libraries, and real-time map display—represent the development trend of emerging scientific big data intelligent analysis software.

Development Trends of Scientific Big Data Intelligent Analysis Software

The development trends of scientific big data intelligent analysis software exhibit five important characteristics: AI empowerment, integration, cloud services, open sharing, and customization.

(1) AI Empowerment. Scientists' demand for employing new AI technologies for scientific discovery in their research fields is increasingly high. Therefore, in addition to providing domain-related basic operations and traditional algorithms, intelligent analysis software needs to support the integrated application of new AI technologies such as deep learning, natural language understanding, and knowledge graphs, providing full-lifecycle tool support for training, testing, deployment, and operation of AI models.

(2) Integration. Scientific big data intelligent analysis involves complex data processing, analysis, pattern extraction, and knowledge discovery processes, while existing big data frameworks and platforms suffer from high learning curves and development costs. Therefore, beyond traditional "programmatic" development models, it is necessary to provide domain scientists with simple and easy-to-use "assembly-style" visual mining analysis environments that leverage high-quality, reusable model and algorithm libraries for innovative design of scientific big data analysis models, achieving integrated support covering data source integration, code editing, process design, model and algorithm reuse, as well as execution and visualization.

(3) Cloud Services. Cloud-service-based scientific big data intelligent analysis software does not require local software installation and maintenance. Consequently, on one hand, the browser becomes a unified portal interface for the entire mining analysis process and management; on the other hand, models, algorithms, and data sources will be shared and reused in the form of online

APIs, a model also known as "function as a service."

(4) Open Sharing. Major discoveries in cross-disciplinary science require comprehensive application of analysis models and algorithms from multiple domains. Aggregating common models across domains to form a rich, high-performance model and algorithm library will become the foundation for reducing development difficulty and improving efficiency of comprehensive analysis models for domain intersection. Meanwhile, sharing high-quality models and algorithms by scientific teams from various domains will also promote the continuous evolution of software systems, making them more vital. For example, the R language algorithm library CRAN is a model for cross-domain algorithm sharing, currently 收录了收录了 (Note: This appears to be a typo in the original, should be "currently 收录了") over 4,000 algorithms contributed by scientists from various domains, attracting a large number of users.

(5) Customization. Data analysis patterns vary dramatically across different scientific domains, and general, fixed big data analysis software cannot meet the personalized analysis needs of specific domain scientific teams. Such personalized needs exist at various levels including analysis processes, data sources, algorithm models, and visualization. Therefore, an ideal scientific big data intelligent analysis software should support domain customization and extension in multiple aspects including data, model algorithms, and visualization views, enabling domain scientists and software engineers within the domain to develop unique components.

Reference Architecture for Scientific Big Data Intelligent Analysis Software

Our team has completed the development of big data systems for multiple scientific and industry domains in recent years and is currently undertaking the development of the Earth Big Data Mining Analysis System (Big Earth Data Miner) for the Chinese Academy of Sciences' Strategic Priority Research Program "Earth Big Data Science Project." Through investigating the big data analysis requirements of multiple domain scientific teams and analyzing current situations and trends, we propose a reference architecture for next-generation scientific big data intelligent analysis software (Figure 1 [Figure 1: see original paper]).

The software system is deployed on cloud platforms, employing general big data systems and machine learning systems as underlying computing support. On this foundation, it provides scientific big data distributed computing processing engines and machine learning engines that meet domain-specific requirements, supporting special processes for scientific big data analysis and processing. Since mining and analysis tasks exhibit characteristics of being both data-intensive and resource-intensive, with significantly different service response requirements such as real-time analysis, online analysis, and offline analysis, it is necessary to

explore efficient resource management and task scheduling mechanisms to meet the differentiated support needs of large-scale concurrent users.

The data resource library provides management of public data resources and personal data resources, supporting users in conveniently and quickly searching, importing personal data resources, and performing data sharing. The algorithm and model library provides management of general algorithms and models as well as domain-specific algorithms and models, supporting secondary development, sharing, and performance optimization of algorithms and models. For models trained based on big data, techniques such as transfer learning can be explored to achieve cross-domain sharing.

The intelligent analysis environment provides multiple intelligent analysis modes. The workflow mode primarily targets relatively fixed analysis scenarios within domains; the code development mode mainly serves scientific teams with research capabilities and flexible analysis needs; the visual interactive analysis mode primarily serves application scenarios that rely on visual observation and analysis. Future extensions could include virtual reality, augmented reality, and additional analysis modes.

The software system provides online mining and analysis services through a browser, enabling users to conduct one-stop analysis work after registering an account. Throughout this process, the cloud service needs to ensure data security for scientists and isolation of user analysis work. Furthermore, it is necessary to explore the use of microservices architecture to achieve domain-specific customization oriented toward different scientific domain requirements.

Science and technology constitute primary productive forces, and intelligent analysis software for scientific big data is an important supporting tool for scientific research. Domestic scientific teams have achieved world-renowned results in many subdomains but have not released open intelligent analysis software with global influence. Therefore, there is an urgent need for domestic scientific teams to unite with information technology research teams, targeting cross-domain scientific exploration and knowledge discovery, fully considering the big data analysis needs of scientific teams from different domains, and designing and developing more suitable intelligent analysis software systems for scientific big data to contribute to human scientific and technological progress.

References

1. Tony H, Stewart T, Kristin T. The Fourth Paradigm: Data-Intensive Scientific Discovery. Translated by Pan Jiaofeng, et al. Beijing: Science Press, 2012.

2. Gorelick N, Hancher M, Dixon M, et al. Google Earth Engine: Planetary-scale geospatial analysis for everyone. *Remote Sensing of Environment*, 2017, 202: 18-27.

3. Ihaka R, Gentleman R. R: A Language for Data Analysis and Graphics. *Journal of Computational and Graphical Statistics*, 1996, 5(3): 299-314.

4. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: Machine learning in python. *The Journal of Machine Learning Research*, 2011, 12: 2825-2830.

5. Hall M, Frank E, Holmes G, et al. The WEKA data mining software: an update//SIGKDD. New York: ACM, 2009: 10-18.

6. Meng X, Bradley J, Yavuz B, et al. Mllib: Machine learning in apache spark. *The Journal of Machine Learning Research*, 2016, 17(34): 1-7.

7. Barga R, Fontama V, Tok W H. *Predictive Analytics with Microsoft Azure Machine Learning*. Berkeley: Apress, 2015.

8. VanderPlas J, Connolly A J, Ivezić Ž, et al. Introduction to astroML: Machine learning for astrophysics. [2018-08-06]. https://ieeexplore.ieee.org/document/6382200/?tp=&arnu

---

Author Biographies

ZHONG Hua is Deputy Director of the Institute of Software, Chinese Academy of Sciences (ISCAS), Director of the Technology Center of Software Engineering (TCSE), and a researcher and doctoral supervisor. He is a council member of the China Computer Federation and an executive council member of the China Software Industry Association. He has long been engaged in research on distributed systems, software engineering, cloud computing, and big data, publishing over 70 papers in well-known international journals and conferences. He has won two second prizes of the National Science and Technology Progress Award, one first prize of the Chinese Academy of Sciences Science and Technology Progress Award, one first prize of the Beijing Science and Technology Award, and one second prize of the Military Science and Technology Progress Award. E-mail: zhonghua@iscas.ac.cn

Corresponding author

Responsible editor: Wen Yanjie

Note: Figure translations are in progress. See original paper for figures.