科学大数据智能分析软件的现状与趋势

钟 华* 刘 杰 王 伟

中国科学院软件研究所 北京 100190

摘要 人工智能领域近年来取得突破进展，如何在自然科学领域采用人工智能新技术促进科学发现，成为科学家和产业界的关注焦点。多学科、跨领域交叉背景下的科学大数据挖掘分析与知识发现，依赖于构建一套高效、易用、可扩展的科学大数据智能分析软件系统，为复杂数据处理、分析、模式提取和知识发现提供学习模型、算法及开发工具支持。文章选取典型科学领域内代表性的智能分析软件系统进行充分的调研，对比分析这类软件的共性和差异，并探讨其发展趋势。在此基础上，文章提出一个面向科学大数据的一体化、可定制的智能分析框架，支撑科学家交互式构建智能分析模型并高效执行，为快速开展科学发现研究提供系统和工具支撑。

关键词 科学大数据，智能分析，数据密集型科学发现，软件系统

DOI 10.16418/j.issn.1000-3045.2018.08.007

2007年图灵奖得主吉姆·格雷（Jim Gray）发表了著名演讲《科学方法的革命》，将科学研究分为4类范式（paradigm），即实验归纳、模型推演、仿真模拟和数据密集型科学发现（data-intensive scientific discovery），从而提出了被广泛称为“第四范式”的“科学大数据”新视角[1]。经过10年的技术发展，深度学习等先进技术在图像、语音、自然语言等人工智能领域均取得突破进展。在自然科学领域，近年来科学家们也紧跟趋势，基于科学大数据驱动的新模式，采用深度学习等新技术，取得了一批重大科学发现成果，发表在*Science*、*Nature*等权威学术刊物。然而，大数据驱动的科学研究工作因为严重依赖于先进的信息技术，对于大多数科学家团队而言仍具有一定门槛。

多学科、跨领域交叉背景下的科学大数据挖掘分析与知识发现，依赖于构建一套高效、易用、可扩展的科学大数据智能分析软件系统，为复杂数据处理、分析、模式提取和知识发现提供学习模型、算法及开发工具支持。通过分析该领域发展现状，我们发现，一些分析软

*通讯作者

资助项目：中国科学院战略性先导科技专项（XDA19020500）

修改稿收到日期：2018年8月6日

件因为运行在单机环境而无法处理大规模数据，一些分析软件因需要较高的编程开发技能而令科学家团队望而却步。随着云计算、大数据和人工智能技术的发展，利用云计算平台承载人工智能技术进行大数据智能分析已经成为趋势，而开放共享与个性化定制也成为软件发展的主流方向。从中可以总结出科学大数据智能分析软件的五大发展趋势：AI 赋能、一体化、云服务、开放共享和可定制。

笔者通过对众多科学家进行需求调研，结合大数据智能分析技术及软件的发展趋势，提出了一个面向科学大数据的一体化、可定制的智能分析框架，支持科学家交互式的构建智能分析模型，并基于云平台分布式计算引擎实现分析模型的高效执行，为快速开展科学发现研究提供系统和工具支撑。期望通过该智能分析框架的研发与应用，为下一代科学大数据智能分析软件提供参考方案。

1 发展现状

数据密集型科学发现离不开软件系统的支撑，本文的研究对象聚焦于近 10 年来面向科学大数据智能分析的典型软件系统。从适用范围来看，科学大数据智能分析软件可以简单分为通用型和领域专用型两类。通用型智能分析软件是大数据、人工智能等领域的通用分析软件，并被科学家团队应用于特定领域的研究工作，如 Matlab①。领域专用型智能分析软件是指针对特定科学领域的专有分析软件，如地学、资源环境科学领域流行的 Google Earth Engine[2]。

1.1 通用型科学大数据智能分析软件

大数据和人工智能技术发展迅速，涌现了大量软件系统，本文选取科学家团队较为常用、具有代表性的智能分析软件，并依据软件系统的部署模式，将这些软件分为 3 类——单机环境、分布式环境和云计算环境，同时这也是智能分析软件发展的 3 个阶段。

（1）单机环境智能分析软件。在商业数据分析软件方面，Matlab 提供了用于算法开发、数据可视化、数据分析以及数值计算的高级编程语言和交互式环境，在众多科学领域应用广泛。在众多开源免费数据分析软件中，R 语言[3]、Scikit-Learn[4]、Weka[5]是典型代表。R 语言是一种用于统计分析和绘图的语言，提供了丰富的统计分析功能，用户还可以通过开发并安装扩展包增强 R 的功能。 Python 语言拥有大量科学数据分析的算法库，其中就包括被广泛应用于机器学习和数据挖掘的 Scikit-Learn。Weka 数据挖掘平台基于 Java 语言开发，提供了可视化、拖拽式的分析流程设计界面，并集成了大量数据预处理和机器学习算法。这些软件系统在设计之初是以单机模式运行，无法针对基于分布式存储的大数据进行处理，在大数据场景下存在先天不足。此外，这些软件系统还缺乏对深度学习技术的有效支持。

（2）分布式环境智能分析软件。在分布式环境下，开源社区提供的大数据分析软件成为主流，Hadoop Mahout、Spark MLlib[6]是其中的典型代表，研究人员借助于 Hadoop、Spark 框架，解决了分布式并行挖掘问题，并提供了典型的机器学习算法和模型。近年来，涌现出一批开源深度学习框架，例如 Tensor Flow、Caffe、CNTK、MXNet 等，用于深度神经网络模型的构建及训练，支持分布式计算和异构计算②。尽管这些开源软件提供了丰富的算法库和高效的分布式计算平台，但仍需要专业的编程开发和系统配置技能，且学习曲线陡峭，不利于科学家团队使用。

（3）云计算环境智能分析软件。通过云平台提供大数据智能分析服务已成为大型公有云平台的标配服务，“机器学习即服务”（machine learning as a service，

① MATLAB: https://ww2.mathworks.cn/products/matlab.html.

② Comparison of deep learning software: https://en.wikipedia.org/wiki/Comparison_of_deep_learning_software.

MlaaS）也成为多家领先云平台厂商的发展趋势。Azure Machine Learning（Azure ML）是微软Azure云平台提供的机器学习分析服务[7]，在提供大量通用机器学习分析算法基础上，Azure ML还面向数据科学家用户提供了交互式的图形化开发界面。类似的MLaaS还包括Aliyun PAI等。这些系统通常仅支持某种特定开发语言和应用程序编程接口（API），用户无法自主扩充算法库，存在平台锁定（lock-in）问题。除了上述公有云厂商提供的大数据智能分析服务，一些科学家团队将具有“浏览器/服务器”架构模式的交互式分析软件部署在公有云或私有云，实现了“简化版”的MLaaS。例如，Jupyter Notebook③是支持“浏览器/服务器”架构的交互式分析软件，支持通过浏览器编辑运行多种编程语言，在服务器端进行数据处理、数值模拟、统计建模、机器学习以及可视化等。

1.2 领域专用型科学大数据智能分析软件

自然科学包括大量细分领域，每个领域都存在专用的科学数据分析软件，本文选取其中若干代表进行分析，并将这些软件分为两类进行介绍：经典的领域专用科学数据分析软件和新兴的领域专用科学数据分析软件。

（1）经典的领域专用科学数据分析软件。这类软件是特定领域科学家专门研发的系统，适合对该领域的科学数据进行专门处理、计算和分析。ROOT④是欧洲核子研究中心（CERN）开发的开源软件，主要用于粒子物理实验的数据处理、科学计算和可视化分析，提供数学及统计工具、并行处理、神经网络及多变量分析软件包，是目前高能物理领域数据分析的典型工具。AstroML是面向天文领域的机器学习和数据挖掘算法包[8]，建立在NumPy、SciPy、Scikit-Learn等Python算法库基础上，提供了多个开放天文数据集的加载器，以及大量天文领域的分析与可视化数据集案例。目前，这类领域专用软件仍采用单机部署，无法进行分布式并行的大数据处理分析，并且尚未对深度学习技术提供集成与支持。

（2）新兴的领域专用科学数据分析软件。这类软件指采用了大数据、机器学习和云计算等新技术的分析软件。SDAP目前是Apache软件基金会的孵化项目，是面向地球物理海洋学领域的科学大数据分析平台。SDAP⑤依赖于NEXUS系统进行大数据处理，NEXUS是由美国国家航空航天局喷气推进实验室（NASA/JPL）开发的一个软件项目，采用Map/Reduce分布式并行计算技术，旨在对NASA各种任务收集的大型数据集进行科学分析。美国国家能源研究科学计算中心（NERSC）⑥，具有美国能源部科学局的主要科学计算设备。最近NERSC支持将深度学习应用到气候研究、中微子实验以及神经科学研究，并取得了一批突破性科学发现。Verily Life Sciences（原谷歌生命科学公司）的研究人员开发了一种深入学习软件工具DeepVariant⑦，该工具可将基因组信息转换成图像进行分析，可显著提升基因变异的识别准确率。Google Earth Engine是Google提供的对大量全球尺度地球科学资料（尤其是卫星数据）进行在线可视化分析处理的云平台，相关领域的科学家团队可以利用该平台提供的长时序近地卫星数据以及数千台的云服务器进行在线数据处理和分析，目前已经取得了一批有显示度的研究成果。可以看出，Google Earth Engine的特定领域海量数据、云端分布式并行计算、在线挖掘分析算法库、地图即时展现等特点，正代表了新兴科学大数据智能分析软件的发展趋势。

③ Jupyter: http://jupyter.org/.
④ ROOT: https://root.cern.ch/.
⑤ Science Data Analytics Platform (SDAP): https://sdap.apache.org/.
⑥ NERSC: http://www.nersc.gov/.
⑦ DeepVariant: https://github.com/google/deepvariant.

2 发展趋势

科学大数据智能分析软件的发展趋势呈现出 AI 赋能、一体化、云服务、开放共享和可定制的重要特征。

（1）AI 赋能。科学家在其研究领域尝试使用人工智能新技术进行科学发现的需求日益高涨。因此，智能分析软件除了提供领域相关的基础运算操作和传统算法，还需要支持深度学习、自然语言理解、知识图谱等新型人工智能技术的集成应用，为人工智能模型的训练、测试、部署和运行提供全生命周期的工具化支持。

（2）一体化。科学大数据智能分析包含复杂的数据处理、分析、模式提取和知识发现过程，而现有的大数据框架和平台存在学习曲线高、开发代价大等问题。因此，在传统“编程式”的开发模式基础上，还需要为领域科学家提供简单易用的“拼装式”可视化挖掘分析环境，并利用高质量、可复用的模型与算法库，进行科学大数据分析模型的创新设计，实现涵盖数据源集成、代码编辑、流程设计、模型算法复用以及执行与可视化的一体化支撑。

（3）云服务。云服务化的科学大数据智能分析软件不需要本地进行软件安装和维护。因此，一方面，浏览器成为挖掘分析全流程操作和管理的统一门户界面；另一方面，模型、算法以及数据源将以在线 API 的形式进行共享和复用，这一形式也被称为“功能即服务”（function as a service）。

（4）开放共享。交叉科学的重大发现需要综合应用多领域的分析模型和算法。汇聚跨领域的共性模型，形成类型丰富、性能优异的模型和算法库，这将成为降低领域交叉综合分析模型开发难度、提升开发效率的基础。同时，各领域科学家团队通过共享高质量的模型和算法，也将促进软件系统持续演化，使软件系统更具生命力。例如，R 语言算法库 CRAN 是交叉领域算法共享的典范，该算法库目前收录了各领域科学家贡献的 4 000 多种算法，吸引了大量的用户。

（5）可定制。不同科学领域的数据分析模式千差万别，通用的、固化的大数据分析软件无法满足特定领域科学家团队的个性化分析需求，这种个性化需求存在于分析流程、数据源、算法模型、可视化等各个层面。因此，一个理想的科学大数据智能分析软件应该支持数据、模型算法和可视化视图等多个方面的领域定制与扩展，支持领域科学家以及领域内的软件工程师进行特有组件的开发。

3 科学大数据智能分析软件参考方案

笔者所在团队近年来完成了多个科学、行业领域的大数据系统研发，目前正在承担中国科学院战略性先导科技专项“地球大数据科学工程”的地球大数据挖掘分析系统（Big Earth Data Miner）研发任务。通过对多个领域科学家团队的大数据分析需求进行调研，结合现状及趋势分析，笔者提出下一代科学大数据智能分析软件的参考方案（图1）。

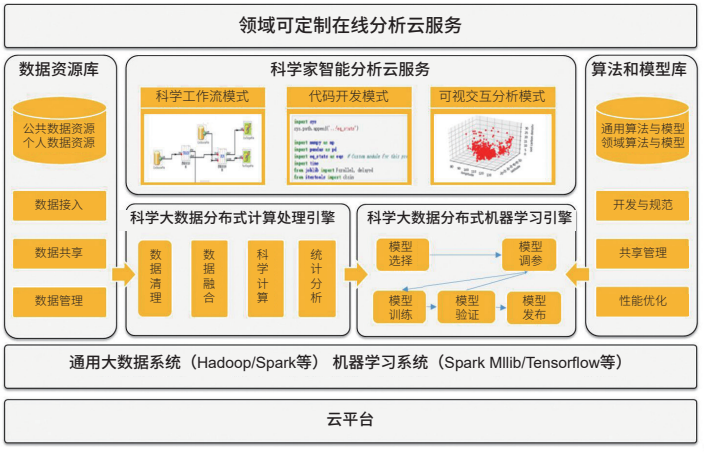

图 1　科学大数据智能分析软件参考方案

该软件系统基于云平台部署，采用通用大数据系统和机器学习系统作为底层计算支撑；在此基础上，提供满足领域特性需求的科学大数据分布式计算处理引擎和机器学习引擎，支持科学大数据分析处理的特殊过程。同时，挖掘分析任务具有数据密集型与资源密集型相结合的特征，也存在即时分析、在线分析以及离线分析等

差异明显的服务响应需求，因此需要探索提供高效的资源管理和任务调度机制，以满足大规模并发用户的差异化支撑需求。

数据资源库提供公共数据资源和个人数据资源管理，支持用户在数据资源库方便快捷地查找、导入个人数据资源，并进行数据共享。算法与模型库提供通用算法及模型、领域算法及模型管理，支持算法和模型的二次开发、共享与性能优化。其中，针对基于大数据训练得到的模型，可探索采用迁移学习等技术实现跨领域共享。

智能分析环境提供多种智能分析模式。其中，工作流模式主要面向领域内相对固化的分析场景；代码开发模式主要面向具有研发能力和灵活分析需求的科学家团队；可视交互式分析模式主要面向依赖可视化观察分析的应用场景。未来还可以扩展到虚拟现实、增强现实等更多的分析模式。

该软件系统通过浏览器提供在线的挖掘分析服务，用户通过注册账户就可开展一站式的分析工作，在此过程中云服务需要确保科学家数据安全和用户分析工作的隔离。此外，需要探索利用微服务架构，实现面向不同科学领域需求的领域化定制。

4 结语

科学技术是第一生产力，而科学大数据的智能分析软件则是科学研究的重要支撑工具。国内科学家团队在很多细分领域都取得了世界瞩目的成果，但是并没有发布具有世界影响力的开放的智能分析软件。因此，迫切需要国内科学家团队与信息技术研究团队联合起来，瞄准交叉领域的科学探索与知识发现，充分考虑不同领域科学家团队的大数据分析需求，设计研发出更适用于科学大数据的智能分析软件系统，为人类科技进步贡献力量。

参考文献

1 Tony H, Stewart T, Kristin T. 第四范式：数据密集型科学发现. 潘教峰, 等 译. 北京: 科学出版社, 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=&arnumber=6382200.

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

ZHONG Hua* LIU Jie WANG Wei

(Institute of Software, Chinese Academy of Sciences, Beijing 100190, China)

Abstract The field of artificial intelligence has made a breakthrough in recent years. How to promote scientific discovery in the field of natural science, especially the field of Earth Science with mass and multi-source data, has become the focus of scientists and industry. The scientific 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 scientific data analysis software system for scientific data. It provides learning models, algorithms and development tools for complex data processing, analysis, pattern extraction and knowledge discovery. In this study, the representative intelligent analysis software system in the typical scientific field is selected to make a full investigation and comparison on the generality and difference of this kind of software, and the development trend is also discussed. On this basis, this study proposes an integrated and customizable intelligent analysis framework for scientific big data, which supports the interactive construction of intelligent analysis models, and provides systems and tools supporting for the rapid development of scientific discovery research.

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

钟 华 中国科学院软件研究所副所长，软件工程技术研发中心主任，研究员、博士生导师，中国计算机学会理事，中国软件行业协会常务理事。长期从事分布式系统、软件工程、云计算、大数据等方面的研究工作，在国际知名期刊和会议发表论文70余篇，曾获国家科技进步奖二等奖2次、中国科学院科技进步奖一等奖1次、北京市科学技术奖一等奖1次、军队科技进步奖二等奖1次。E-mail: zhonghua@iscas.ac.cn

ZHONG Hua Researcher at Institute of Software, Chinese Academy of Sciences (ISCAS). He is the Deputy Director of the institute, also the Director of Technology Center of Software Engineering (TCSE). He has been engaged in research on distributed system, software engineering, cloud computing, big data, and so on. He has published more than 70 papers in well-known international journals such as *IEEE Transactions on Knowledge and Data Engineering*, *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, and *Journal of Statistical Software*, and international conferences such as International Conference on Software Engineering (ICSE), IEEE/ACM International Conference on Automated Software Engineering (ASE), IEEE International Parallel & Distributed Processing Symposium (IPDPS), IEEE International Conference on Web Services (ICWS), International Conference on Database Systems for Advanced Applications (DASFAA), and so on. He has won two second prizes of the National Science and Technology Progress Award, one first prize of CAS Science and Technology Progress Award, one first prize of Beijing Science and Technology Award, and one second prize of Military Science and Technology Progress Award. E-mail: zhonghua@iscas.ac.cn

■责任编辑：文彦杰

*Corresponding author