Contribution ID: **8**　　Type: **not specified**

面向科学大数据的深度分析与智能计算平台研究与应用

Wednesday, 17 July 2019 16:30 (20 minutes)

科学数据是国家科技创新和发展的基础性和战略性资源，是科研创新最根本、最活跃、影响面最广的科技资源。大数据时代，科技创新越来越依赖科学数据的综合分析，从气象预测到生物基因分析，从牧草选育变化到生态数据监测，科学数据的意义及重要性不言而喻。随着人工智能技术的蓬勃发展，其在各学科领域的大规模交叉应用必将成为促进科研创新，推动社会经济发展，提升国家科技竞争力的重要手段。

近年来，深度学习和机器学习等技术已经在高能物理、生物医药、空间地理等学科的研究工作中得到了广泛的应用，例如高能物理领域利用卷积神经网络、递归神经网络和对抗生成网络等深度学习算法模型应用于粒子鉴别、事例分类、事例重建、异常检测等多个科研场景。但是将这些新技术应用于学科研究也存在很多障碍，其中最大的问题在于机器学习等技术对学科专家来说往往是一个黑盒，使用这些技术和算法需要对算法模型和软件工具具有深入的理解，并且具备一定的编程能力。对于科学家们来说，存在着极高的学习成本和不确定因素。

我们通过分析大数据背景下科学大数据分析的方法和工具，整理大数据环境下科学大数据分析工作的流程和需求，设计并实现了面向科学大数据的深度分析与智能计算平台，让学科专家无需编写程序便可以轻松使用各种数据分析算法，采用数据流模式自由创建组合各种算法流程来完成复杂的科学大数据分析任务。

平台的核心由数据管理、数据计算和数据可视化三个模块组成。

数据管理模块
数据管理模块用来存储和管理科学数据分析的相关数据，由元数据管理、数据引接、数据整合以及数据管理四部分组成。
（1）元数据管理：元数据管理负责平台多源异构数据格式的定义和管理。
（2）数据引接：数据引接负责平台多源异构数据的导入。
（3）数据整合：数据整合负责对引接进来的数据进行清洗、加工、整合等功能。
（4）数据管理：数据管理负责对平台存储的非文本数据进行管理。

数据计算模块
数据管理模块用来存储和管理大数据分析算法和科学数据分析流程，并为用户提供可视化建模环境用于创建自定义的科学数据分析流程，创建完成的流程可提交至平台进行计算并获得每一步的计算结果。
（1）算法管理：算法管理负责对平台的大数据算法进行管理，平台内置了包括自然语言处理、分类回归、推荐、结果评价等多种类型的数十种算法，同时支持用户上传基于 Scala、Java、Python 等语言编写的算法包来完成个性化的科学数据分析任务。
（2）任务调度：任务调度为用户内置了多种科学数据分析模型，并为用户提供了一个可视化建模环境。在该环境中，每种科学数据分析模型都被表示为一个有向无环图（DAG）[16]，算法和数据集都被作为图中的一个节点。要分析的数据由根节点进入，通过每个算法节点进行计算，并将结果发送给其后代节点，最终结果从终节点流出。每个节点的运算状态采用不同颜色表示，白底灰边框表示等待执行，绿色表示执行成功，白底绿边框表示正在执行，红色表示执行失败。用户
可以无需编写任何代码，直接将算法和数据集（节点）采用拖拽方式进行流程构建，最终创建个性化的科学数据分析流程。

数据可视化模块
数据可视化模块用来将平台中存储的数据和科学数据分析计算后的最终结果进行可视化。对于平台中存储的数据，可以通过选择行、列和对比字段，创建各种不同类型的图表，包括折线图、柱状图、雷达图、词云图等十余种可视化图表。

本文从科学大数据分析工作的流程和框架入手，针对科学大数据分析工作的实际需求，设计并实现了面向科学大数据的深度分析与智能计算平台。并且通过实际案例，展示了平台的功能和效果。事实证明，该平台可以很好地满足大数据时代各个学科领域科学大数据分析工作的要求，促进大数据挖掘和知识发现在科学大数据分析领域的应用。

Summary

本文从科学大数据分析工作的流程和框架入手，针对科学大数据分析工作的实际需求，设计并实现了面向科学大数据的深度分析与智能计算平台。并且通过实际案例，展示了平台的功能和效果。事实证

明，该平台可以很好地满足大数据时代各个学科领域科学大数据分析工作的要求，促进大数据挖掘和知识发现在科学大数据分析领域的应用。

Primary authors: Prof. 文, 奕 (中国科学院成都文献情报中心); Mr 杨, 宁 (中国科学院成都文献情报中心)

Presenter: Mr 杨, 宁 (中国科学院成都文献情报中心)

Session Classification: 科学计算与数据管理 II

Track Classification: 科学计算技术与平台