科学发现中的机器学习方法研究

孟小峰[1)] 郝新丽[1)] 马超红[1)] 杨 晨[2),3)] 艾山·毛力尼亚孜[1)]
吴 潮[4)] 魏建彦[4)]

[1)](中国人民大学信息学院 北京 100872)
[2)](中国人民银行清算总中心 北京 100048)
[3)](清华大学计算机科学与技术系 北京 100084)
[4)](中国科学院国家天文台 北京 100101)

摘 要 大规模科学装置与重大科学实验使得科学发现进入了数据密集型的第四范式,借助蓬勃发展的人工智能技术促进智能科学发现势在必行.机器学习作为人工智能中的一项重要技术,已广泛应用于各个科学领域.然而,现有工作仅研究特定任务下的机器学习方法,没能抽象出一个通用的智能科学发现研究框架.本文首先总结了科学发现任务中常用的机器学习方法,并将科学任务归类为五大机器学习问题.其次,提出了基于机器学习的智能科学发现研究框架,作为"AI for Science"的典型范例,阐述了一种高效的智能科学发现模式.再次,本文以时域天文学中发现瞬变事件这一科学任务为例,通过实验证明了唯有恰当地结合领域知识后,机器学习算法才能更好地服务于智能科学发现,验证了该框架的有效性.最后进行总结与展望,以期对各领域进行智能科学发现形成参考意义.

关键词 科学发现;机器学习;科学大数据;瞬变事件发现;智能科学发现
中图法分类号 TP18 **DOI 号** 10.11897/SP.J.1016.2023.00877

Research on Machine Learning for Scientific Discovery

MENG Xiao-Feng[1)] HAO Xin-Li[1)] MA Chao-Hong[1)] YANG Chen[2),3)] MAOLINIYAZI Ai-Shan[1)]
WU Chao[4)] WEI Jian-Yan[4)]

[1)](*School of Information, Renmin University of China, Beijing* 100872)
[2)](*China National Clearing Center, Beijing* 100048)
[3)](*Department of Computer Science and Technology, Tsinghua University, Beijing* 100084)
[4)](*National Astronomical Observatories, Chinese Academy of Sciences, Beijing* 100101)

Abstract Probing valuable scientific phenomena is very important for revealing the laws of the universe and verifying the proposed scientific hypothesis. The rare scientific phenomena prompt people to build many large-scale scientific devices or carry out large-scale scientific experiments to collect a lot of scientific data for analysis, which is called data-intensive scientific discovery. In this paradigm, relying solely on the expertise of scientists is no longer feasible and scientific discovery needs a kind of more efficient method. As a result, a kind of key artificial intelligence (AI) technique, machine learning, plays a more and more important role in it. In other words, "AI for Science" is booming. Scientific big data and scientific discovery tasks are different from general big data and tasks on the Internet. For example, scientific big data has a more long lifecycle, more uncertainty, and hard to get be repeatedly. Scientific discovery tasks are not only innovative

收稿日期:2022-02-28;在线发布日期:2022-10-14.本课题得到国家自然科学基金项目(62172423,91846204,U1931133)资助.**孟小峰**(通信作者),博士,教授,博士生导师,中国计算机学会(CCF)会员,主要研究领域为科学数据管理、云数据管理、隐私保护以及社会计算等交叉性研究.E-mail:xfmeng@ruc.edu.cn.**郝新丽**,博士研究生,主要研究方向为智能科学发现、时间序列分析和可解释机器学习.**马超红**,博士研究生,主要研究方向为科学数据管理、机器学习化数据库系统和学习化索引.**杨 晨**,博士,博士后研究员,主要研究领域为大数据实时分析、大数据系统性能分析和智能运维.**艾山·毛力尼亚孜**,博士研究生,主要研究方向为自然语言处理、知识图谱、机器学习.**吴 潮**,博士,副研究员,主要研究方向为数据挖掘与瞬变源搜索.**魏建彦**,博士,研究员,博士生导师,主要研究领域为瞬变源观测与科学.

but also rigorous. Because of the above characteristics, there are a lot of tough and common problems when different machine learning methods meet scientific discovery. However, the existing work only focused on specific machine learning algorithms to accomplish specific scientific discovery tasks, rather than giving a general research framework of AI-driven scientific discovery to solve these common problems. In this paper, we first summarize the latest development of intelligent scientific discovery in six scientific fields, in which machine learning has been widely used. On the one hand, we analyze frequently-used methods in scientific discovery tasks from machine learning and deep learning two perspectives. On the other hand, we classify scientific discovery tasks into 5 kinds of machine-learning problems from basic science and applied science two perspectives. Secondly, we propose a general research framework for intelligent scientific discovery as an example of "AI for Science". It describes an efficient mode of applying machine learning to scientific discovery and helps scientists make sense of how to use machine learning efficiently in scientific tasks. Corresponding to the scientific discovery pipeline, this framework is composed of six components. Every component solves several challenges when scientific discovery meets machine learning. These six components are scientific data integration and sharing, scientific discovery task transformation, scientific data pre-processing, scientific discovery method, scientific discovery verification, and domain knowledge constraints, respectively. Thirdly, we verify this framework through a series of experiments. We choose time-domain astronomy as a typical scientific field of "Big Data+AI". In this field, we aim at discovering a kind of transient event, which is called a stellar flare. To compare different discovery methods, we use seven machine-learning methods and a classical method in time-domain astronomy. One of the most important conclusions is that machine learning is not omnipotent. Only when combined with domain knowledge, will machine learning reach its full potential. Lastly, we summarize three challenges that need to be solved in the future and three lessons learned. Machine learning has its advantages and disadvantages for scientific discovery. Scientists should make more efforts in science-oriented machine learning, not only developing machine learning applications for scientific discovery.

Keywords scientific discovery; machine learning; scientific big data; transient event discovery; intelligent scientific discovery

1 引 言

大规模科学装置的建设与重大科学实验的开展,使得科学发现研究无法完全依赖专家经验从海量数据中捕捉并研究稀有的科学现象.近年来,人工智能技术(Artificial Intelligence,AI)蓬勃发展,机器学习(Machine Learning,ML)作为其重要的研究领域之一,被科学家广泛应用于科学发现任务,促成"AI for Science"的重大发展机遇.

从生物学、化学到天文学,机器学习在科学发现任务中日益盛行,利用机器学习技术从海量科学数据中发现稀有科学现象、研究复杂难题成为了科学领域的首选方案.AlphaFold2[1]人工智能算法可以预测人类 98.5%的蛋白质结构,准确度达到原子级别.同样发表于《Nature》的研究成果[2]利用机器学习技术以前所未有的速率进行逆向合成反应,具有化学界的"AlphaGo"之称.2019 年首张黑洞照片的合成同样离不开机器学习算法的巨大贡献[3-4].此外,利用机器学习探测引力波,相较传统的模板匹配,速度提升多个数量级[5]……诸如此类的成功案例不胜枚举,机器学习凭借在预测精度、时间效率等方面的优异表现,在科学发现任务中发挥重要作用.2020 年美国能源部高级科学计算咨询委员会(ASCAC)讨论了科学数据急剧增加带来的挑战和机遇,批准了关于机器学习应用于科学的报告①,并呼吁制定一个为期十年的 AI 计划②.可见,科学发

① https://science.osti.gov/ascr/ascac/Meetings/202009
② https://www.ciste.org.cn/index.php?m=content&c=index&a=show&catid=148&id=2233

现已经进入“科学大数据＋人工智能”的新范式.

在作者已发表的论文中[6]，我们从数据管理的角度提出了大规模科学数据的智能发现与管理框架，分为智能分析、知识融合、数据存储三个方面解决智能科学发现问题.本文主要选取第一个方面，探讨机器学习方法应用于科学数据智能分析、实现智能科学发现的挑战与机遇.

科学大数据的特点.科学大数据是指以海量科学证据形式存在的事实，包括观测与监测数据、实验与模拟数据等原始及衍生数据[7]，数据类型包括栅格点云等空间数据、时序数据、时空数据以及图像数据等[8].相比于互联网大数据，科学大数据不仅拥有4V 特点，而且还具备独特的科学特征.主要体现在如下三个方面：

(1)生命周期长.相比于互联网大数据“重分析、轻存储”的短暂生命周期，科学大数据的生命周期包含“采集与实时分析-存储与处理-发布与共享-再分析与重用-归档与长期保存”的全过程[7].科学大数据的生命周期更为完整且长久，价值具有长效性.因此注重对科学大数据进行实时处理的同时，离线分析同样具有重要的科学意义.

(2)不确定性强.科学大数据是对客观世界的描述，但由于观测条件以及设备的限制，其中普遍存在环境噪声与系统观测误差，甚至数据缺失，导致数据在质量和产生速率等方面存在较高的不确定性[7].另外科学数据通常来源于非人工系统[9]，如海洋、大气、天体，人类不能完全确定并控制其运作机理，自然系统本身的高度不确定性加剧了科学大数据的不确定性.

(3)不可重复与随意更改.对科学事件的观测通常是不可重复的[10]，尤其是具有时间属性的科学对象.例如2019 年1 月29 日美国国家航空航天局通过凌日系外行星勘测卫星捕捉到了罕见的“潮汐干扰”，黑洞撕碎恒星，该观测数据是无法重复获得的.另外，科学数据是对真理的客观描述，用于探索科学问题，对科学数据的每一处理步骤都会对科学发现的结果产生影响，因此对科学数据的预处理需严格遵循科学原理，不可随意更改与变换.

科学发现任务的特点.随着科学数据演化成为科学大数据，基础科学与应用科学领域中的科学发现任务也形成了新的特点.基础科学是探索自然界最普遍规律的科学，包括物理学、天文学等分支，其以自然界中某种物质的形态及运动形式为研究对象，以揭示自然界的基本规律为科学发现任务，从而使人类更好地“认识世界”.应用科学是运用基础科学的理论成果从而创造性地解决人类在生产实践中具体问题的科学领域，包括新药研发、材料设计等分支，其科学发现任务为在特定领域内发现新技术、新产品，从而使人类更好地“改造世界”.本文将科学发现任务的特点总结为如下三个方面：

(1)数据密集型科学发现.早在2007 年，图灵奖得主吉姆格雷(Jim Gray)指出，科学研究的范式经历了经验科学(实验归纳)、理论科学(模型推演)、计算科学(模拟仿真)三阶段的演化过程，当代已经进入了数据密集型科学发现(Data-intensive Scientific Discovery)的第四范式[11]，从以计算为中心转化为以数据处理为中心.随着科学数据的大规模采集和积累，科学数据不再仅仅作为科学研究的成果，而成为科学研究的对象和工具，基于数据来设计和实施科学研究成为科学发现的一般方法.

(2)依赖先进的技术手段.新技术手段拓展科学发现的途径，而科学发现又可以促进新技术手段的诞生，二者呈现出相互依赖、交替发展的关系.例如计算机模拟和仿真技术开辟了新的科学实验和研究方式，半导体材料科学的进步又提升了计算机芯片的性能.当今蓬勃发展的机器学习技术将会对科学研究产生重大影响，成为科学发现的有力工具.

(3)兼具创新性与严谨性.科学发现的目标为发现新现象、新物质、新原理或创造新技术，因此创新是科学发现的生命.同时，科学创新还须具备严谨性，包括可解释性和可复现性两方面.科学家不仅要有新的发现，还要用严谨的逻辑解释其背后的原因，从而实现干预世界的目标.可复现性是科学研究的基本属性，这意味着其他研究人员可以重复同样的过程，一个不可重复的偶然科学发现无法作为其他科学研究的基础，因而失去价值.

机器学习应用于科学发现的挑战.由于上述特殊性，将机器学习应用于科学发现时面临很多挑战，本文将其总结为任务、数据、模型三个方面：

(1)科学任务转化困难.将科学发现任务转化为机器学习问题具有挑战性①，例如：如何将物理定律参数化从而使用机器学习发现新的定律?目前很多科学任务仍停留在本领域而未被转化为适合机器学习解决的问题，这限制了机器学习相关学者研究这些科学问题.将蛋白质结构预测这一科学任务转化为从序列到三维结构的映射问题是 AlphaFold2

① https://ai4sciencecommunity.github.io/

成功的重要因素之一.因此巧妙地将科学任务转化为能够被机器学习所解决的问题和场景是首要挑战.

(2)数据集成与预处理困难.科学数据分散在不同国家的科研机构,经过集成并共享给其他研究人员才能发挥更大的价值.然而目前科学数据面临着缺乏有效融合、大型标准数据集建设不足以及质量参差不齐等问题[12-13].另外科学数据通常是多源、异构、高维、低信噪比且不均衡的[9],需要进行复杂的预处理来保证数据质量.并且数据的科学属性对处理过程产生了更多的限制,使流程更为复杂[14].因此通过集成与共享获取高质量的科学数据,进而结合领域知识对数据进行合适的预处理是应用机器学习的基础.

(3)模型科学性验证困难.科学发现具有严谨性,因此需要验证机器学习模型的可解释性和可复现性,从而保证模型的科学性.但深度神经网络具有高度的非线性和复杂度,人们很难解释模型原理及在相应领域的科学意义,一些保守的科学家对其持观望态度,称其为"黑盒"模型[12].另外,目前深度学习方法的可复现性受到争议,利用机器学习所形成的科研成果在很多情况下是不可复现的[15].对于一个严谨的科学领域,提高机器学习模型的可解释性和可复现性,让科学家确认科学产出的真正价值,具有重要意义.

为了深入探究机器学习在科学发现任务中的优缺点,为各科学领域学者使用机器学习解决问题提供参考方案,本文第 2 节梳理科学发现中常用的机器学习方法,分析其适用场景,并阐述机器学习擅长处理的科学发现任务,将其归类为五大机器学习问题;第 3 节提出基于机器学习的智能科学发现研究框架,阐述一种高效的智能科学发现模式;第 4 节通过实验验证框架的有效性,即以时域天文学这一典型的"大数据+AI"的科学领域为例,使用 7 种机器学习方法和科学领域的传统方法完成发现天体瞬变事件的科学任务;第 5 节对机器学习应用于科学发现任务的经验教训和发展方向进行总结与展望.

本文与其他相关综述性文章[12,16-20]的主要区别为:本文综合分析了机器学习在各个科学领域的研究现状,探讨其中的共性问题;并提出一个通用的研究框架,用以指导各科学领域学者使用机器学习进行高效的科学发现研究,同时促进机器学习相关学者快速了解科学发现任务;最后通过案例验证该框架的有效性.

2 研究现状

科学发现正在被机器学习所改变,从天文学、物理学、化学等基础科学,到生物制药、材料科学、气象科学等应用科学,越来越多的科学家利用机器学习进行科学发现、解决复杂难题.在科学发现任务中常用的机器学习方法有很多,根据模型复杂度可分为基于统计的传统机器学习和基于神经网络的深度学习两大类,根据任务目标又可分为分类、回归、聚类、异常检测、数据生成等.相对于传统机器学习,深度学习方法被越来越多地应用于科学发现研究,辅助科学家实现了更加重大的科学突破.

本节首先根据第一种分类方式讨论在科学发现任务中常用的机器学习方法,分析传统机器学习与深度学习方法在科学发现任务中的优缺点与适用场景.其次以基础科学和应用科学领域中的部分学科为例,阐明机器学习所擅长处理的科学发现任务,并根据第二种分类方式,将不同的科学发现任务归类为五大机器学习问题.

2.1 传统机器学习方法

在机器学习应用于科学发现的初期,由于数据量以及计算能力的限制,科学家普遍使用传统机器学习方法[21],并在经典模型的基础上,根据具体问题对模型的输入、超参数、结构等做出适当调整,从而在实际科学任务中达到最好的效果.

其中,分类与回归是两种应用最为普遍的方法.二者均属于预测方法,区别在于预测数据的类型不同,因此很多分类算法同时可以担任回归任务[22].朴素贝叶斯(Naive Bayes)、逻辑回归(Logistic Regression,LR)、K近邻(K-Nearest Neighbor,KNN)、决策树(Decision Tree)是较为简单的方法,具有模型直观、容易实现的特点[23-25].支持向量机(Support Vector Machines,SVM)、随机森林(Random Forest)与极度梯度提升(eXtreme Gradient Boosting,XGBoost)是更为复杂同时更为有效的方法[24-29],其中Random Forest与XGBoost是两种典型的基于Bagging与Boosting的集成方法,通常准确率更高,因此在科学发现任务中的应用也更为普遍.

当数据无标签时,科学发现任务中还常用聚类和异常检测两种无监督方法来解决科学问题.例如在天文领域,瞬变科学事件具有偶发性和不可预知性,因此可通过孤立森林(Isolated Forests)、一类支持向量机(One-Class SVM)等异常检测算法发现瞬

变源[30-31]. 当需要挖掘未知类别与属性时，还可以利用聚类算法进行分析[31-32]，例如时域天文学家对光变曲线提取特征后，通过 HDBSCAN(Hierarchical Density-Based Spatial Clustering of Applications with Noise)等聚类算法分析类之间的异质性，并进一步分析其中的异常[31].

由于传统机器学习依赖人工提取特征，因此在科学发现中的相关研究侧重于根据领域知识提取更为有效的特征表示，进而利用相应模型完成科学发现任务，特征质量对最终的结果具有决定性作用. 例如在时域天文学领域，天文学家根据领域知识从光变曲线中提取偏度峰度等标准统计量、以及峰的宽度等领域特征，进而对天体进行分类[33-34]. 在化学领域，科学家将非数值类型的化学式转化为机器学习可以识别的形式，如通过物理化学描述符、分子指纹、分子简写式等方式对分子进行表示，进而通过分类或回归模型对化合物性质进行建模与预测[35]. 在高能物理领域，将探测器响应的时间、位置和幅度信息作为输入，利用回归模型可以重建物理事件的主要产物在探测器中的运动轨迹[36]. 在新药研发领域，通过对已知药物与靶点的对应数据进行建模，可以预测其他药物与靶点之间的复杂关系[37].

随着研究的深入，提取的特征数目越来越多，例如从光变曲线中可以提取出数百种特征，但并非所有特征都同等重要. 因此科学家利用主成分分析、t 分布随机近邻嵌入(t-SNE)等方法对高维数据进行降维，从中筛选出更有效的特征子集，并可视化数据间的聚簇关系，提升后续模型的效果[38-39].

小　结. 传统机器学习方法的输入特征表示由科学家结合领域知识人工提取，具有较强的物理意义；模型具有较强的理论支撑、相对简单，因此通常认为具有较好的可解释性. 但上述特点具有两面性. 人工提取特征容易引入偏见，特征具有局限性. 更多的特征并未带来效果的显著提升；并且模型相对简单，导致其拟合能力受限. 因此传统机器学习适用于人工提取特征已十分充分、对可解释性要求高、数据和计算资源受限的情况下使用. 本文对科学发现中常用的传统机器学习方法与相应科学文献的总结见表 1.

表 1　科学发现中常用的机器学习方法

	模型	分类	聚类	回归	异常检测	数据生成	参考文献	可解释性
传统机器学习	Naive Bayes、Logistic Regression、SVM	✓		✓			[23-27,33]	具有一定可解释性
	KNN、Decision Tree、Random Forest、XGBoost	✓		✓			[24-25,28-29,35-37]	
	K-Means、HDBSCAN		✓				[31-32]	
	One-Class SVM、Isolated Forests				✓		[30-31]	
深度学习	CNN、LeNet、AlexNet、FCN、GCN	✓		✓	✓		[40-49]	可解释性较差
	RNN、GRU、LSTM	✓		✓	✓		[48,50-57]	
	GAN、VAE					✓	[58-63]	

注：✓表示机器学习模型常用于解决该类任务，无符号表示不常用于解决该类任务.

2.2　深度学习方法

深度学习可以从原始数据中逐层自动提取特征，是一种重要的表示学习方法. 其网络结构复杂，对大数据具有更好的拟合能力，适用于解决图像、序列、图等数据中的复杂难题. 相对于传统机器学习方法，近年来越来越多的研究使用深度学习进行科学发现，实现了一些重大的科学突破.

(1) 卷积神经网络

卷积神经网络(Convolutional Neural Network，CNN)是一种含有卷积层的前馈神经网络，以擅长处理图像数据著称，经典网络包括 LeNet[64]、AlexNet[65]、ResNet[66]等. 由于图像数据是一种重要的科学数据类型，因此 CNN 在科学领域得到广泛应用. CNN 不仅适用于图像数据，其在时序数据上也具有良好的效果，避免了循环神经网络不可并行的缺点. 全卷积网络(Fully Convolutional Network，FCN)和时域卷积网络(Temporal Convolutional Network，TCN)是两种典型的用于处理序列数据的 CNN 模型. FCN 起初用于对图像进行像素级的分类[67]，Wang 等人[68]将其应用于时间序列分类任务中，并取得了良好效果. 时域卷积网络 TCN[69]相较于 FCN 加入了因果卷积，从而使得每个时间点的输出只与更早的输入有关，防止了信息泄露.

在科学发现任务中，CNN 模型主要用于对图像数据或可以转化为图像数据的信号进行分类与回归，但也不乏用其处理序列数据的研究[40]. 在天文领域，CNN 模型被广泛用于处理观测图像和光谱，从而完成天体分类[41]、红移估计[42]等任务. 在高能物理领域，科学家将探测器的输出映射为图片中的像素值，进而使用 CNN 实现粒子喷注的分类[43]. 超过 90%的医学数据为影像数据[20]，因此在医学领域广泛使用 CNN 模型进行辅助诊断，例如皮肤癌分

类[44]、COVID-19 肺炎的诊断和预测[45]等.

除分类与回归问题以外,CNN 模型也可用于异常检测.在时域天文学中,通过对连续两帧观测图像相减,可以检测到短暂出现的科学事件[46],完成暂现源检测任务.另外,图卷积神经网络作为卷积神经网络在图领域的推广,在化学、材料等具有图结构数据的领域实现新的突破.例如科学家可以将原子和化学键分别表示为分子图中的节点和边,利用图卷积神经网络预测反应物到产物的每个原子对之间的化学键变化[47].

(2)循环神经网络

循环神经网络(Recurrent Neural Network,RNN)是另一种重要的深度学习模型,与善于处理图像数据的 CNN 不同,RNN 是为了更好地处理序列数据而设计的.其通过引入状态变量存储过去时间信息,让网络具有记忆.在科学发现任务中,长短期记忆网络(Long Short-Term Memory,LSTM)[70]和门控循环神经单元(Gated Recurrent Unit,GRU)[71]是两种常用的循环神经网络,二者均解决了在时间步数过大时 RNN 的梯度衰减问题,可以更好地捕捉时间序列中的长期依赖关系[72].GRU 相对于 LSTM 更为简单,参数量更少,加快了训练速度.

在科学发现任务中,RNN 主要用于序列数据的分类与预测问题.在时域天文学领域,RNN 常用于对光变曲线等序列进行建模分析,如基于光变曲线的星系分类[50]、引力透镜参数估计[48]等.在高能物理领域,基于 RNN 的 Jet 分类模型解决了因探测器大小不同而导致 CNN 模型中部分信息丢失的问题[51].在有机化学领域,通过分子简写式对 RNN 进行训练可以生成新型分子结构[52-53].在生物领域,RNN 可用于处理基因和蛋白序列数据,例如预测转录因子识别位点[54]、利用氨基酸序列预测酶的生化功能[55]、预测非编码基因的功能[56]等.另外在时域天文学中,可利用 RNN 完成暂现源检测任务,根据实时预测的光变曲线与真实接收的光变曲线的差距,判断是否检测到暂现源候选体[57].

(3)深度生成模型

CNN 和 RNN 均为判别模型,另一种重要模型为生成模型,在科学领域也普遍应用.判别模型根据观察变量 X 直接学习条件概率分布 $P(Y|X)$或决策函数 $F(X)$,用于直接推断,学习准确率高.而生成模型对输入与输出数据的联合概率分布 $P(X,Y)$进行建模,相比于判别模型,其训练难度更大、模型结构更复杂[73],但因其出色的生成能力得到广泛研究.变分自编码器(Variational Auto-Encoder,VAE)[74]和生成式对抗网络(Generative Adversarial Network,GAN)[75]是两类常用于科学发现的深度生成模型.与传统的自编码器通过数值描述潜在空间不同,VAE 以概率的方式描述潜在空间,在数据生成方面具有优势.GAN 由生成器和判别器组成,生成器通过调节参数试图生成判别器无法辨别真假的伪样本,在二者博弈的过程中,生成器可以生成更为接近真实样本的数据,从而实现数据的自动生成与信息补全.

在科学发现任务中,生成模型主要用于生成新数据或重建旧数据.在天文领域,可以利用 VAE 生成高质量星系图像进而对暗物质进行研究[58]、利用 CosmoGAN 生成引力透镜[59]、利用 GalaxyGAN 重建星系图像[60]等.在高能物理领域,利用 GAN 模拟粒子经过每层探测器后形成的"物理照片",可以实现样本的快速生成[61].在化学领域,利用 VAE 将离散的分子简写式编码为隐空间的连续变量,进而通过随机解码、扰乱或插入的方式可以产生新的分子结构[62].在新药研发领域,结合 VAE 与 GAN 可以生成具有特定抗癌活性的分子[63].

小 结.深度学习方法直接对原始数据进行挖掘,擅长探索高维数据的隐含结构与相关性,可以学习出科学家暂时所无法明确提取的复杂特征与规律、突破人工提取特征的局限性、避免偏差、减少计算特征的时间开销,相比于传统的机器学习实现了更加卓越的效果.但与此同时,其消耗的数据资源与计算资源更多,在资源敏感的科学场景下受到了限制.并且由于模型结构复杂、不可解释,一些保守的科学家对其持怀疑态度,这在一定程度上限制了深度学习在科学领域的应用[12,76].因此深度学习模型适用于计算与数据资源充足、问题复杂、人工特征提取困难或者效果不佳、对可解释性要求不高的场景.本文对科学发现中常用的深度学习算法与相应科学文献的总结见表 1.

2.3 科学发现的任务分类

机器学习与科学发现的交叉研究主要集中在数据积累充足、数字化程度高的学科中,而在基础物理等基础理论研究中,由于缺乏大量样本而难以应用.本小节以天文学、高能物理、化学、新药研发、材料设计、气象科学 6 个应用程度较高的领域为例,梳理机器学习在各领域所擅长处理的科学发现任务,并将不同的科学发现任务总结为分类、聚类、回归、异常检测和数据生成 5 种机器学习问题.

(1) 基础科学领域

随着大型观测设备的发展，飞速增长的数据促进了机器学习在天文学中的应用，主要包括以下几种任务[12]：① 目标检测与分类[33,35,41,50,57]，如暂现源检测和星系分类，可转化为机器学习异常检测、分类或聚类问题，其中暂现源检测在离线分析的基础上，还要求实时捕捉短暂的异常现象；② 天体参数估计[29,42,48]，如根据光谱与测光数据估计天体的质量、元素丰度等物理量，属于机器学习的回归问题；③ 观测数据的降噪与重建[59-61]，如对观测图像进行超分辨率重建，从而在硬件成本一定的情况下获得更高的数据精度，可对应于数据生成问题. 机器学习减轻了天文学家的负担，提升数据处理的效率，尤其在处理图像与光变曲线等特征复杂的数据类型时，深度学习成为了天文学家的首选方案.

在高能物理领域，以对撞机为代表的一系列大型科学装置产生了海量的科学数据，使得该领域具有应用机器学习的数据基础. 由于传统算法的开发难度大、优化困难、难以并行，无法胜任处理海量数据的科学任务，因此促使了物理学家使用机器学习进行科学发现研究[16]. 主要任务包括[77]：① 探测器粒子径迹重建[36]，涉及回归与聚类问题；② 物理对象鉴别[43,51]，如粒子种类鉴别和粒子喷注(Jet)标记，可对应于机器学习的分类问题；③ 动力学测量[28]，如簇射动力学参数估计、重离子碰撞喷注横动量估计，属于回归问题；④ 物理仿真[61]，如生成喷注图像、电磁簇射，对应于机器学习的数据生成问题.

随着化学信息学的发展，机器学习在化学领域同样展现出显著优势，来自柏林自由大学的科学家利用深度学习计算薛定谔方程的基态解，实现了准确度和计算效率的突破[78]. 当传统的化学研究理论面对复杂体系时，预测能力有限，大部分新发现依靠大量的实验“试错”，偶然性与不确定性强、效率低、成本高[17]. 而机器学习凭借其强大的学习能力和计算能力，提升了研究效率. 主要任务包括：① 化合物性质预测[35]，如活性、毒性、溶解度等，根据预测数据类型不同可分为分类与回归问题；② 分子设计[52-53]，如设计具有特定性质的分子，涉及机器学习的预测与生成问题；③ 前向反应预测与逆合成分析[47]，如根据反应物预测生成物，或从产物出发预测可能的前体，可对应于机器学习回归问题.

(2) 应用科学领域

近年来利用机器学习进行药物研发成为一种行业新趋势. 传统的新药研发方式面临着费用高、成功率低、耗时长的困境，因此全球多家制药企业与人工智能企业开展了深度合作. 机器学习在新药发现和临床前研究两个阶段实现了重大突破[18,79]，具体任务包括：① 药物靶点发现[37]，确定药物靶点是药物研究的基础，主要对应于机器学习的分类问题；② 化合物筛选[26]，即选择对某一特定靶点活性较高的化合物，属于分类或回归问题；③ 分子生成[63]，即根据已知化合物分子的结构和成药性等规律，合成新的化合物作为候选药物分子，对应于数据生成问题；④ 临床结果预测[27]，如预测新靶点和候选药物的性质和作用，属于机器学习回归问题. 事实证明，利用机器学习技术辅助新药研发，可以大幅缩短研发周期、降低研发成本、提高研发效率.

随着材料数据的积累，机器学习被用于有机材料、光伏、半导体等材料设计的各个领域. 面对巨大的设计空间，基于理论研究、实验分析、计算仿真的传统方法无法高效研发出新材料，大量试错实验导致研发效率很低[19]，这一问题与传统新药研发所面临的困境具有相似之处. 因此材料科学家将数据驱动的机器学习应用于新材料开发的过程中，主要任务包括：① 材料属性预测[25]，即从成分、能量特征等参数出发，研究材料性质的变化规律，属于机器学习的分类或回归任务；② 新材料合成[24]，涉及机器学习的分类、回归与数据生成问题；③ 缺陷识别[49]，如根据材料薄板裂纹图像识别缺陷，可对应于图像分类问题. 机器学习的应用避免了成本昂贵的实验和大量计算，极大地提高了新材料的研发效率.

在气象科学领域，DeepMind 与气象科学家合作利用机器学习实时预测降雨量[80]，开辟了实时降雨量预报的新途径，该任务可对应于机器学习回归问题. 类似的，还可以进行飓风、风暴等极端天气的检测与分类[81]，可对应于异常检测与分类问题. 如此的研究不胜枚举，机器学习方法已经广泛应用于各个科学领域，并在科学探索与发现任务中发挥着重大作用.

小　结. 综上所述，本文将科学发现任务与对应的机器学习问题总结为表 2. 大部分科学发现任务都可以转化为机器学习领域的分类与回归问题，依据不同科学数据类型，可进一步分为图像分类或序列分类问题等. 聚类、数据生成与异常检测问题相对较少. 此外，一些科学任务通过离线挖掘实现，如天体参数估计，对实时性没有要求；而另一些科学发现任务，由于科学现象稀有且短暂，例如天文暂现源检测，不仅需要离线分析，实时检测更有意义.

表 2 科学发现任务分类

领域		科学发现任务	分类	聚类	回归	异常检测	数据生成
基础科学	天文学	目标检测与分类[33,35,41,50,57]	✓	✓		✓	
		天体参数估计[29,42,48]			✓		
		观测数据降噪与生成[59-61]					✓
	高能物理	探测器粒子径迹重建[36,77]		✓	✓		
		物理对象鉴别[43,51]	✓				
		动力学测量[28]			✓		
		物理仿真[61]					✓
	化学	化合物性质预测[35]	✓		✓		
		分子设计[52-53]	✓		✓		✓
		前向反应预测与逆合成分析[47]			✓		
应用科学	制药学	药物靶点发现[37]	✓				
		化合物筛选[26]	✓		✓		
		药物分子生成[63]					✓
		临床结果预测[27]			✓		
	材料科学	材料属性预测[25]	✓		✓		
		新材料合成[24]	✓		✓		✓
		材料缺陷识别[49]	✓				
	气象科学	实时降雨预测[80]			✓		
		极端天气的检测与分类[81]	✓			✓	

注：✓表示科学发现任务常对应的机器学习任务，无符号表示不常对应于该类机器学习任务.

当然以上的划分边界并不总是清晰明确，需要根据具体的科学问题与场景具体分析，才能将科学任务转化为合适的机器学习问题.综上，机器学习突破了传统方法的瓶颈，被广泛用于获得新的科学视角，成为智能科学发现的有效途径.

3 智能科学发现

综合分析机器学习在各科学领域的研究现状，本文认为，当前大部分研究聚焦于特定的科学任务、着力解决具体科学问题，而机器学习在各科学领域面临着诸多共性的困难与挑战，亟需一个通用的智能科学发现框架用以指导科学家进行高效的科学发现研究.

因此，本文提出了基于机器学习的智能科学发现研究框架，作为“AI for Science”的典型范例，阐述一种高效的智能科学发现模式，为各领域科学家使用机器学习进行科学发现提供解决方案.框架分为科学数据集成共享、科学发现任务转化、科学数据预处理、科学发现方法、科学发现验证以及领域知识约束六个部分，具体如图 1 所示.各部分之间并非独立，而是相互反馈，提高了框架的智能性.本节按照框架组成的顺序，依次分析机器学习在科学发现任务中面临的共性问题，并提出解决方案.

3.1 科学数据集成共享

大规模的科学数据是应用机器学习的基础，然而科学数据分散在各国科研机构，因此将科学数据进行集成与共享为智能科学发现框架保证了数据基础.当前，国内外有诸多科学数据中心，如国际科学联合会世界数据中心①、中国科学院科学数据中心、美国国家航空航天局空间科学数据中心等.

但目前，科学数据集成与共享普遍面临着缺乏有效融合、标准数据集建设不足以及质量参差不齐三个方面的问题，阻碍了机器学习发挥更大潜能.第一，虽然有多种集成的数据库供科学家使用，但其对数据间关联挖掘不足，缺乏有效的知识融合机制[13]，如天文学家需手动查找多种波段的数据库才能获取一个天体的完整信息，不利于科学家高效检索与使用；第二，目前科学领域侧重于将分散的科学数据集中起来，但缺乏针对共性问题而建立的统一大型标准数据集[12]，不利于对解决同一问题的各种方法进行比较；第三，科学数据具有不确定性强、多源异构等特点[13]，目前公开发布的科学数据存在着精度不同、格式不同、处理方法不同等数据质量参差不齐的问题.

因此，为了提高数据检索和使用的效率，需对多源异构数据进行知识融合，利用知识图谱构建与补全技术，构建并完善科学领域知识图谱.具体可通过两个层面实现：(1) 从开源科学数据库、历史文献中获取大量结构化/非结构化信息，挖掘多种科学数据类型在时空范围、内容属性、主题分类、类型格式等方面的关联，自动抽取三元组，从而在各学科构建领

① http://wdc.org.ua/.

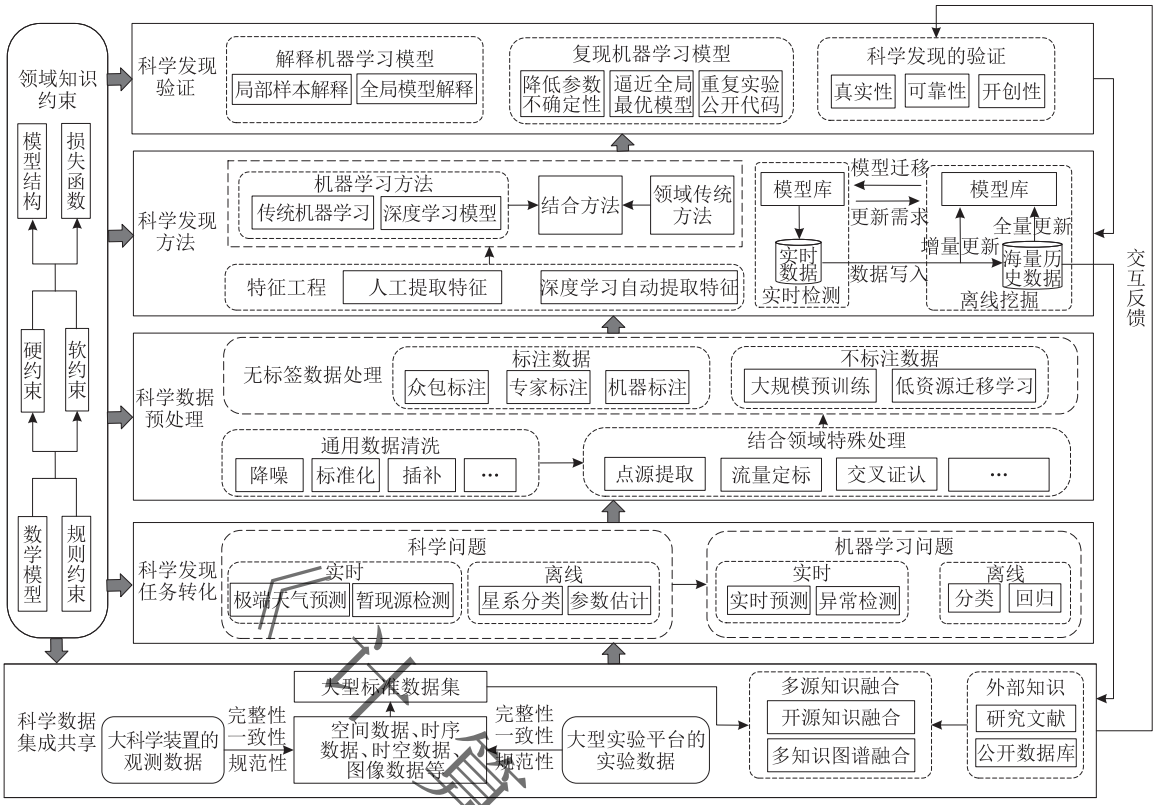

图 1　基于机器学习的智能科学发现研究框架

域知识图谱；(2) 融合多个已有领域知识图谱，不断丰富和完善语义信息，扩大知识图谱的应用范围. 通过知识图谱技术促进物理目标之间、物理目标及其文献描述之间、文献描述之间的数据与知识融合，有助于科学家快速检索，并服务于对科学发现结果的高效验证.

此外，各个科学领域内需针对共性问题建立高质量的统一大型标准数据集，从而便于对不同方法进行统一测评；数据集成方需提升对数据内容的完整性、一致性、规范性的检查力度，发布数据的精度、不确定度和适用范围，从而提升数据质量.

3.2　科学发现任务转化

在获取科学数据后，将科学发现任务转化为合适的机器学习问题是实现智能科学发现的重要环节. 在第 2.2 节本文已经较为详细地讨论了各种科学发现任务与分类、回归、聚类、异常检测、数据生成 5 种机器学习问题的对应关系. 除此以外，科学发现任务还包括离线挖掘和实时检测两大类：一些科学发现任务通过分析历史数据完成，如天文学的星系分类与参数估计，大量的历史数据中包含着全面的信息，无需随新数据的产生而实时分析；而另一些研究短暂科学现象的任务，通过实时的方式完成更有意义. 如天文学的暂现源检测，实时进行异常检测可以及时调用精度更高的观测设备，从而更为清晰地观测稍纵即逝却稀有的科学现象.

实时的科学发现任务需要将离线和实时两种机器学习问题结合考虑：离线训练的数据量大，可以涵盖全局特征，模型精度高，但训练时间长；而实时收集的数据代表了最新的数据特征，当数据分布发生变化时，已有模型难以发挥作用. 因此，实时的科学发现任务不仅需要利用离线挖掘的模式，还需要根据实时收集的数据对已挖掘的模式进行补充和更新.

确定了机器学习的研究问题，可以有针对性地指导后续的数据预处理. 如分类问题通常需要均衡的数据集，而异常检测问题则不需要，因此二者所需的数据预处理方法不尽相同. 另外不同的机器学习问题也对应着不同的机器学习方法. 将科学发现任务转化为机器学习问题是承接科学大数据集成共享与科学大数据预处理及后续流程的关键步骤.

3.3　科学数据预处理

科学大数据通常采集于大规模科学装置或大型实验设备，由于观测条件以及设备的限制，科学数据具有不确定性强、信噪比低等特点，因此在科学数据飞速增长的同时，其中的噪声和冗余信息也会随之增加.

为了使机器学习方法不受数据噪声和冗余信息

的干扰,需要根据不同的研究问题,对数据进行有针对性的预处理.另外由于科学数据具有不可重复、不可随意更改等特性,因此相比于互联网大数据,对科学大数据的预处理过程具有更多的限制,流程更为复杂.

首先,对科学数据的预处理包含基本操作,如格式转换、数据降噪、数据插补等常规清理.其次,在每一个科学领域还具有其独特的预处理方法,不恰当的预处理方法容易引入误差,影响机器学习算法的性能,进而阻碍科学发现的进程.例如在时域天文学领域,光学观测设备采集的原始数据为图像数据,在进行基本清洗处理后,还需要经过点源提取、交叉证认、流量定标等预处理步骤才能形成最终的光变曲线数据[7,82-83].

另外,科学大数据的标签非常有限[12],可在预处理阶段增加标注量,也可采取相反的方式,即仅利用标签有限的数据.对于第一种方式,科学领域普遍通过生成模拟数据实现,或众包标注、专家标注.星明天文台公众超新星搜寻项目是一个典型的众包标注项目,自 2010 年开始运行已发现 30 颗超新星①.然而,对科学数据进行标注相对于常识标注更为困难,如干旱现象比动物类别更难界定,因此有很多科学项目是业余爱好者难以完成的,从而需要科学家亲自标注.但人工标注的成本较高,另一方面,模拟数据不能完全反映数据的真实情况,因此使用机器自动标注是一种值得探索的方式,可以更为经济高效地扩充科学数据的标签.

直接利用标签有限的数据是另一种解决思路,如迁移学习和预训练方法.在低资源情境下利用迁移学习方法可以把开放领域中的研究模型迁移到科学领域中来,从而降低对科学数据的标注需求.基于自监督的大规模预训练模型使用大量未标注的样本学习数据中的共性,不针对特定的下游任务,预训练完成后再将其应用到特定任务上,利用有限的标注数据进行微调.其优点在于可以降低对标注数据的需求,帮助下游任务模型更好地初始化、加速收敛、避免过拟合,缺点在于大规模的预训练往往需要大量的训练样本以及庞大的计算资源,这对于普通学者而言具有难度.而近年逐渐发展的科学计算平台能够带来解决方案,其整合了世界上各科学项目的数据资源,同时拥有很强的计算基础设施,如虚拟天文台[84].利用科学计算平台可以实现科学大数据的大规模预训练,从而解决科学大数据标注样本稀少的问题,为普通学者利用机器学习方法进行智能科学发现助力.

3.4 科学发现方法

经过数据预处理后,采用合适的方法是科学发现任务的核心,包括领域传统方法与机器学习方法两大类.领域传统方法通常为科学模型,具有极强的理论保证,并经过了大量实践的验证,具有较强的外推能力,但通常具有效率较低、建模困难等缺点;机器学习方法包括传统机器学习和深度学习方法两大类,极大地提高了科学研究的效率,但缺乏严谨的理论证明和可解释性.

领域传统方法和机器学习方法各有优缺点,有效的手段是通过一定的方式将二者进行结合,相互补充,扬长避短.例如,很多科学模型需要参数,但这些参数难以根据基本原理获得[85],因此可以利用机器学习方法从一个候选集合中学习最优参数,从而实现动态智能地参数化科学模型;当一个理论模型的子模型具有半经验性质,同时拥有足够数量的观察样本时,可利用机器学习模型替代该子模型;或通过"并行"与"串行"的方式将机器学习方法与领域传统方法"级联"使用,从而纠正传统方法的建模残差.

对于实时科学发现任务,可通过机器学习的在线学习或增量学习完成:在线学习模型[86]以数据流为输入,并立即更新模型,但不注重对旧知识的保持能力;增量学习[87]也能够不断处理连续的信息流,并在学习新知识的同时保持旧知识,任务增量学习和类增量学习是研究较为广泛的两种类型.但由于仅利用增量样本进行学习,因此逼近"原样本+增量样本"的全局最优点成为研究难点.

本文提出实时-离线交互反馈机制来应对实时科学发现任务.该机制在离线层利用全部历史数据进行训练进而形成模型库,能够涵盖更全面的特征;实时层收集的科学数据以批量的形式合并到离线数据库,并定时对离线模型库进行增量更新.当实时层模型库难以应对最新数据时,实时层提出更新需求,将离线学习的模型库迁移至实时层.由于增量更新难以逼近全局最优模型,因此离线层会定期进行全量更新,并将模型同步至实时层进行科学发现.实时-离线交互反馈机制提高了实时科学发现的智能性.

3.5 科学发现验证

当前大部分研究聚焦于利用机器学习算法提高科学发现的效率和准确性,但对机器学习模型的验证不足,这关乎结论是否严谨且符合科学意义.本框

① http://xjltp.china-vo.org/pspjs.html

架提出从模型可解释性、模型可重复性，结果真实性三方面对科学发现进行验证.

可解释性. 科学家所熟知的基于概率统计的建模方法能够根据数据分布的假设求出预测偏差与置信区间，模型可解释性强[12]. 而众多机器学习方法仅从结果上说明其有效性而缺乏严格的理论证明，系统误差较难估计，牺牲了部分可解释性和科学严谨性，受到部分科学家的质疑.

因此，可使用机器学习的解释技术对模型进行解释，包括对样本的局部解释和对模型的全局解释两大类：对样本的局部解释可以清晰地呈现模型对每一个样本的决策依据，全局解释可以呈现模型对整体的评估而降低样本间差异的影响. 通过对模型进行解释，可以揭示成功的原因，或分析失败的教训用以改进模型，甚至有可能帮助发现新的科学规律. 近年来有科学家尝试对机器学习模型进行初步解释，判断其是否符合科学意义. 例如在天文学领域，通常利用基于梯度、灵敏度分析的可解释方法找出对模型重要的特征，进而判断其是否符合领域知识[88,96].

可复现性. 可复现性是科学方法的基本属性，要求其他研究人员可以重复同样的过程，一个不可重复的科学发现无法作为其他科学研究的基础，因而失去价值. 但目前机器学习方法的可复现性受到争议[15]，这其中的原因包括代码和数据未公开、模型结果对训练条件敏感（算力、数据划分、初始化和超参数不同）等.

因此提高机器学习模型的可复现性的具体做法包括：减少模型在参数初始化过程中的不确定性、通过设置周期学习率等方法避免局部最优参数从而逼近全局最优模型[91-92]、多次重复实验过程以验证结果的可复现性，并公开代码、数据和训练参数设置等信息以供他人验证和使用.

真实性. 最后，还需要结合历史文献和公开数据库等开源数据验证科学发现的真实性、可靠性、开创性. 以时域天文学为例，当机器学习算法对光变曲线进行处理产生疑似事件后，科学家需要通过天文数据库、历史文献等多种数据源，综合分析天体位置、历史观测信息、光变曲线质量等多方面信息，才能发现异常信号背后的真相、确定疑似事件的真伪、确定是否首次发现该天体发生耀发现象等. 而利用“数据集成共享”阶段构建的领域知识图谱，可以帮助科学家进行更为高效的后期验证.

3.6　领域知识约束

在机器学习过程中整合科学领域知识是实现智能科学发现的必备手段，也是与其他机器学习应用的本质区别. 领域知识有多种形式，通常为数学方程的形式，如解析表达式和微分方程，或者以规则约束的形式表示实例或类之间的关系，如对称性、不变性、时空相关性、渐进极值、守恒定律等. 领域知识可通过硬约束和软约束的形式与机器学习模型整合：硬约束为不可违背的约束，用约束问题替代无约束优化问题，在训练过程中强制执行；软约束的执行程度可变，如损失函数.

领域知识可以通过模型结构和损失函数两方面与机器学习模型进行结合[93]. 其一，通过修改模型结构，引入强归纳偏置，使模型遵守领域约束，从而产生更具一般性和可解释性的模型，如在模型中增加物理变量[94]、编码对称性约束[95]、基于物理信息进行模型结构搜索[96]等；其二，通过修改损失函数，使其包含领域知识，鼓励模型与先验知识保持一致，这是一种软约束，可以提高模型的收敛性、泛化性，并减少所需的训练数据量.

领域知识贯穿整个智能科学发现框架，在科学发现任务转化、科学数据预处理、科学发现方法、科学发现验证等方面都起到至关重要的作用：在任务转化时结合领域知识才能保证等价性；在数据预处理时结合领域知识，可以生成模拟数据、减少对数据量或数据标签的需求，也可以在特征工程中提取或选择具有物理意义的特征，提高可解释性；在科学发现的方法中结合领域知识，可以加快机器学习模型的训练速度、纠正模型的优化目标、提高准确性；在验证过程中结合领域知识，可以判断模型结果是否与已知的科学原理相一致，验证模型的可解释性和结果的可靠性.

唯有紧密围绕科学属性与领域知识约束，才能保证模型产出的科学价值. 抛开领域知识、盲目使用通用算法的方式在科学发现任务中是行不通的.

4　实验分析

时域天文学是一个典型的“大数据＋AI”的科学领域，大视场短时标巡天设备是该领域的观测利器，该设备具有超大视场覆盖和高时间分辨率的数据采样特性，奠定了该领域的大数据基础. 地基广角相机阵（Ground-based Wide Angle Cameras，GWAC）[47,83,88]是我国自主研发的大视场短时标时域天文观测设备. 本节以 GWAC 实际产生的光变曲线（星等 Mag 序列）数据为研究对象，以恒星耀发

(Stellar Flare)[97]这种典型的瞬变事件为科学发现目标(如图 3 所示),以本文提出的"基于机器学习的智能科学发现研究框架"为指导,通过实验分析各种科学发现方法优劣,验证框架的有效性.本文的代码和数据公开于 GitHub 开源社区①和阿里天池实验室②.

首先,本案例将发现瞬变事件这一科学任务转化为机器学习的时间序列分类问题.其次,为了应对数据存在"间断性"和"类别不均衡性"两方面的挑战,本文通过数据截断和数据增强两种方式进行预处理.具体地,本文根据文献[98]所揭示的类太阳恒星耀发活动的特征时间,对间隔过大的序列进行截断处理;在文献[99]提出的耀发模型的基础上,通过更改模型参数实现对耀发幅度、持续时间、耀发位置的多种变换来生成耀发信号,并与已有负样本进行叠加实现数据增强,使训练集的正负样本比例达到 1∶4.测试集仍保持实际数据分布.经过上述预处理,最终的实验数据集包含 152 635 条光变曲线子序列,其中训练集、验证集与测试集的比例约为 6∶2∶2.接下来我们在第 4.1 节介绍科学发现方法,第 4.2 节分析各方法的优劣,第 4.3 节进行科学发现验证.

4.1 实验方法

本文使用领域传统方法、机器学习方法以及二者结合的方法完成瞬变科学事件的发现任务.

(1)领域传统方法

模板匹配(Template Matching)是一种经典的传统方法,其基于模板这一专家知识,广泛应用于目标检测、分类以及参数估计任务中[11].本文将该方法作为时域天文学中传统方法的代表,与机器学习方法进行对比分析.该方法具体分为两步:为了提高匹配效率,首先基于一定的规则对光变曲线序列进行截断处理,其次使用领域专家给定的耀发模板,计算子序列与该模板之间的距离,从而判断该子序列是否与耀发模板匹配.

(2)机器学习方法

本文选择 7 种科学发现任务中常用的机器学习方法作为代表,对光变曲线进行二分类,其中包括 3 种传统机器学习模型(KNN、Decision Tree、SVM)和 4 种深度学习模型(CNN、GRU、FCN、TCN).

由于序列数据可以互相度量距离,因此本文直接将光变曲线作为 KNN 模型的输入.Decision Tree 和 SVM 无法直接处理序列数据,本文按照文献[100]的做法,从光变曲线中提取 13 种特征值,进而输入到 Decision Tree 和 SVM 中进行训练与预测.此 3 种传统机器学习模型使用 Scikit-learn[101]实现,最佳参数通过网格搜索得到.

为了捕捉耀发曲线的形状特征,本文绘制出每一条光变曲线的散点图,再利用 CNN 模型进行分类.RNN 模型擅长处理序列数据,本文选用结构较为简单的 GRU 作为 RNN 模型的代表.另外,由于在许多情况下使用 CNN 对序列建模可以比 RNN 取得更好的性能、更高的效率,因此本文使用 TCN 和 FCN 两种模型作为 CNN 的代表直接处理序列数据.此 4 种深度学习模型使用 Keras③实现,通过经验不断调整模型结构和参数至最佳.

(3)结合方法

本文"级联"领域传统方法与机器学习方法,即依次使用模板匹配和机器学习方法,从而将二者结合.具体地,首先利用模板匹配对测试数据进行初选,将预测为正例的数据输入到已训练好的机器学习模型中做进一步判断,将两种方法均预测为正例的数据作为最终的预测正例,其余样本作为负例.

4.2 实验结果

本文从两个角度分析实验结果,分别为领域传统方法(模板匹配)与各种机器学习方法的优劣对比、不同"级联"方式的效果对比(模板匹配"级联"机器学习方法,以及机器学习方法"级联"机器学习方法).

本文计算精确率(*Precision*)、召回率(*Recall*)以及 F_2*score* 作为模型的评价指标.由于科学事件具有重大意义,因此科学发现任务在注重精确率的前提下,更加注重高召回率,即不希望遗漏重要的科学事件.因此本文选取较为常用同时更侧重于召回率的 F_2*score* 作为评价指标对各模型进行评估.

(1)领域传统方法与机器学习方法的对比

表 3 显示,模板匹配方法和机器学习方法并没有明显优劣.具体地,模板匹配、决策树、支持向量机三种方法召回率最高,但精确率均较低,在真实场景下,这会导致天文学家后期验证的工作量过大.CNN 和 FCN 两种卷积神经网络的模型取得了相类似的效果,召回率较高、精确率不足、F_2*score* 欠佳.GRU 和 KNN 模型的精确率和召回率均不能满足要求.TCN 模型效果最佳,其余模型与之差距较大.

实验证明,模板匹配或机器学习方法虽然可以发现瞬变科学事件,但大部分模型仍有提升空间.

① https://github.com/915466648/gwac_flare
② https://tianchi.aliyun.com/competition/entrance/531805/introduction
③ https://github.com/fchollet/keras

表 3　传统方法与机器学习方法的 *Precision*、*Recall*、F_2 *score*

评价指标	Template Matching	KNN	Decision Tree(DT)	SVM	CNN	FCN	TCN	GRU
Precision	0.3218	0.0357	0.2667	0.3011	0.6279	0.6923	**0.9310**	0.7576
Recall	**1.0000**	0.8929	**1.0000**	**1.0000**	0.9643	0.9643	0.9643	0.8929
F_2 *score*	0.7035	0.1539	0.6452	0.6829	0.8710	0.8940	**0.9574**	0.8621

(2) 不同“级联”方法的对比

本实验将上述 3 种高召回率的模型与其余模型进行“级联”结合，从而在精确率和召回率之间取得最佳平衡，并对比 3 种“级联”方式的优劣. 具体地，分别利用模板匹配、Decision Tree、SVM 三种方法对样本进行初筛，再“级联”其余模型做进一步判断.

图 2 表明，3 种“级联”方式下，各组合模型的 F_2 *score* 相对于未“级联”的单一模型均有所提升. 总体而言，对传统机器学习模型提升幅度较大，对深度学习模型的提升幅度较小，原因在于各单一深度学习模型已经达到较好的效果. 具体数值如表 4 所示，由于 3 种初筛方法的召回率为1，对后续模型的召回率没有影响(即组合模型的召回率与表 3 中各单一模型的召回率相同)，同时提升了组合模型的精确率，从而使得各组合模型的 F_2 *score* 均基本满足实际科学发现的需要.

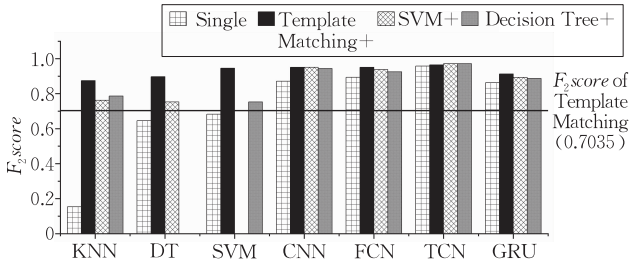

图 2　未“级联”与三种“级联”方式下各模型 F_2 *score* 对比

表 4　三种“级联”方式下各组合模型的 *Precision*、F_2 *score* 以及相对单一模型的提升幅度平均值

“级联”方式	评价指标	KNN	Decision Tree(DT)	SVM	CNN	FCN	TCN	GRU	△Average
Template Matching+	*Precision*	**0.8065**	**0.6364**	**0.7778**	**0.9000**	**0.9000**	0.9643	**1.0000**	**+0.3390**
	F_2 *score*	**0.8742**	**0.8974**	**0.9459**	**0.9507**	**0.9507**	0.9643	**0.9124**	**+0.2042**
SVM+	*Precision*	0.4808	0.3784	—	**0.9000**	0.8438	**1.0000**	0.8929	+0.1975
	F_2 *score*	0.7622	0.7526	—	**0.9507**	0.9375	**0.9712**	0.8929	+0.1473
Decision Tree+	*Precision*	0.5319	—	0.3784	0.8710	0.7941	**1.0000**	0.8621	+0.1820
	F_2 *score*	0.7862	—	0.7527	0.9441	0.9246	**0.9712**	0.8865	+0.1407

表 4 中最后一列表示在 3 种“级联”方式下，各个组合模型在多个指标上相对单一模型(表 3)的提升幅度平均值. 对比可知，虽然使用 SVM 和 Decision Tree 进行初筛同样可以提高各组合模型的精确率，但总体不及使用模板匹配效果明显，仅在 TCN 一种模型上的初筛效果优于模板匹配，即模板匹配“级联”机器学习模型比机器学习模型相互“级联”效果更好. 由此表明：基于专家知识的模板匹配方法可以过滤掉数据驱动的机器学习方法所不能分辨的假正样本.

实验证明，在模板匹配具有高召回率的前提基础上，通过将模板匹配与机器学习模型相结合，可以最大幅度提高组合模型的精确率，有效地提升科学发现的准确性. 因此，将基于专家知识的传统方法与基于数据驱动的机器学习方法相配合可以更好地完成科学发现任务，二者起到相互补充的作用.

4.3　模型解释

深度学习模型由于高度非线性和复杂度，通常难以解释. 为了提高本文采用的深度学习模型的可解释性，本小节选择 FCN 模型为代表，使用类激活映射(Class Activation Mapping，CAM)的方式，对 FCN 模型的分类结果进行可视化分析[68]. CAM 是一种适用于 CNN 模型的解释方法，其通过计算特定目标类别与各特征图的加权和，求得一个全局特征图，再通过重采样至原始数据大小并将二者叠加，可以解释每一个输入维度对该类别的重要程度.

图 3 所示为 FCN 正确分类的耀发样本 CAM 热力图. 由图可知，相对于其余较为平坦的部分，光变曲线中上升及下降部分对模型结果贡献更大. 由此可得，本文训练的模型捕捉到了耀发序列“快速上升、缓慢下降”的特点，具有一定的物理意义及可解释性.

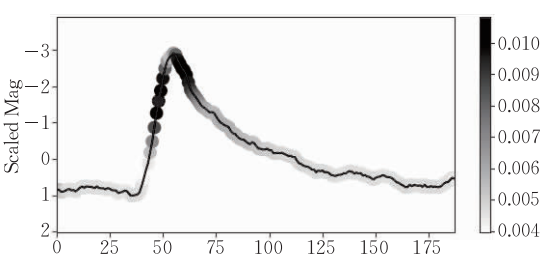

图 3　FCN 正确分类为耀发样本的 CAM 热力图

5 总结与展望

综上所述，机器学习已经在科学发现任务中发挥了重大作用.尽管本文提出了基于机器学习的智能科学发现框架，并通过实验验证了其有效性，但仍有一些挑战问题值得探讨，具体包括：

虚假关联问题.在一般情况下，通过绘制热力图等方式对机器学习模型进行解释，可以判断模型的决策依据是否为虚假关联，如文献[102]通过 LIME 分析机器学习模型将雪地上的哈士奇错误地分类为冰原狼，结果表明分类错误的原因为模型关注的是雪地而非犬类特征，由此可知模型利用了“雪-冰原狼”的虚假关联.然而，科学数据中蕴含着未知知识，因此当模型的决策依据与已有先验知识不符时，科学家难以判断模型的决策依据为虚假关联还是未知的科学知识，从而难以判断科学发现的可靠性.

高效验证问题.目前科学领域对模型进行解释的研究工作大部分面向单一样本，即利用可视化的方法对样本逐个绘制热力图，科学家需要依次检查模型的关注点是否与科学意义相符.当面对海量数据时，此种方式会耗费科学家大量的时间和精力，带来巨大的负担.因此未来需要结合领域特点，基于概率统计理论模型，设计统计层面上的高效解释方法，这让科学家对数据和模型的整体情况有更加直观的了解，节省科学家的时间与精力.

模型易用性问题.在科学发现过程中利用机器学习方法可以提高科学发现的效率.但实际应用时，机器学习尤其是深度学习对于科学家仍然存在较高的门槛.尽管 TensorFlow[103]、PyTorch[104] 等机器学习框架提高了机器学习的易用性，但其通常面向通用任务，尚未考虑科学领域的需求.

在这一方面，生物领域已经领先迈出一步.PaddleHelix① 是基于百度飞桨深度学习框架开发的生物计算平台，用于新药研发、疫苗设计、精准医疗等多种任务，成为学术界与工业界结合的典范.未来学界与业界需要秉持更加开放的精神，共同建设新的科研基础设施，打造智能科学发现的科研共同体，帮助科学家使用强大的工具实现重大的科学突破.

最后，在应用机器学习完成科学发现任务的过程中，本文总结以下三点经验教训：

机器学习方法并非万能.在机器学习的过程中整合科学领域知识是实现智能科学发现的必备手段，也是与其他机器学习应用的本质区别.虽然机器学习方法愈加流行，但基于专家知识的领域传统方法仍具有较强的优势，不可摒弃，需要通过合适的方式将领域传统方法与机器学习相结合，才能更好地完成科学发现任务.

机器学习需要扬长避短.基于统计的传统机器学习与基于神经网络的深度学习各有优缺点，适用于不同的应用场景.领域科学家需要根据不同的场景选择合适的机器学习方法，扬长避短，合理利用机器学习技术突破传统方法的局限性，实现高效的科学发现.

领域与方法具有倾向性.在不同的学科领域中，机器学习技术的研究与应用程度相差甚远.在拥有海量科学数据同时面临维度灾难问题的科学领域中，机器学习得到更为广泛和充分的研究.同时，深度学习方法相较于传统机器学习方法，也得到越来越多的关注与研究，实现了更为重大的科学突破.

机器学习能够以自动化或半自动化的模式帮助科学家产出科学成果，提高科学发现的效率.但面对最前沿、最复杂的问题时，仍需要人工智能学者与各领域的科学家汇聚在一起，形成统一的计算思维，同时恰到好处地利用机器学习工具，才能形成重大的科学发现与研究成果.

参 考 文 献

[1] Jumper J, Evans R, Pritzel A, et al. Highly accurate protein structure prediction with AlphaFold. Nature, 2021, 596(7873): 583-589

[2] Segler M H S, Preuss M, Waller M P. Planning chemical syntheses with deep neural networks and symbolic AI. Nature, 2018, 555(7698): 604-610

[3] Akiyama K, Algaba J C, Alberdi A, et al. First M87 event horizon telescope results. VIII. Magnetic field structure near the event horizon. The Astrophysical Journal Letters, 2021, 910(1): L13

[4] Akiyama K, Algaba J C, Alberdi A, et al. First M87 event horizon telescope results. VII. Polarization of the ring. The Astrophysical Journal Letters, 2021, 910(1): L12

[5] Huerta E A, Khan A, Huang X, et al. Accelerated, scalable and reproducible AI-driven gravitational wave detection. Nature Astronomy, 2021, 5(10): 1062-1068

[6] Meng Xiao-Feng. Scientific data intelligence: AI for scientific discovery. Bulletin of National Natural Science Foundation of China, 2021, 35(3): 419-425(in Chinese)
(孟小峰. 科学数据智能：人工智能在科学发现中的机遇与挑战. 中国科学基金, 2021, 35(3): 419-425)

① https://paddlehelix.baidu.com/

[7] Li Jian-Hui, Shen Zhi-Hong, Meng Xiao-Feng. Scientific big data management: Concepts, technologies and system. Journal of Computer Research and Development, 2017, 54(2): 235-247(in Chinese)
(黎建辉，沈志宏，孟小峰. 科学大数据管理：概念、技术与系统. 计算机研究与发展，2017，54(2)：235-247)

[8] Deiana A M C, Tran N, Agar J, et al. Applications and techniques for fast machine learning in science. Frontiers in Big Data, 2022, 5: 787421

[9] Guo Hua-Dong, Wang Li-Zhe, Chen Fang, et al. Scientific big data and digital Earth. Chinese Science Bulletin, 2014, 59(12): 1047-1054(in Chinese)
(郭华东，王力哲，陈方等. 科学大数据与数字地球. 科学通报，2014，59(12)：1047-1054)

[10] Guo Hua-Dong. Scientific big data—A footstone of national strategy for big data. Bulletin of the Chinese Academy of Sciences, 2018, 33(8): 768-773(in Chinese)
(郭华东. 科学大数据——国家大数据战略的基石. 中国科学院院刊，2018，33(8)：768-773)

[11] Hey T, Tansley S, Tolle K, et al. The Fourth Paradigm: Data-intensive Scientific Discovery. Redmond, USA: Microsoft Research, 2009

[12] Tao Yi-Han, Cui Chen-Zhou, Zhang Yan-Xia, et al. The application of deep learning in astronomy. Progress in Astronomy, 2020, 38(2): 168-188(in Chinese)
(陶一寒，崔辰州，张彦霞等. 深度学习在天文学中的应用与改进. 天文学进展，2020，38(2)：168-188)

[13] Zhu Yun-Qiang, Pan Peng, Shi Lei, et al. Progress and challenge of scientific big data integration and sharing. China Science & Technology Resources Review, 2017, 49(5): 2-11(in Chinese)
(诸云强，潘鹏，石蕾等. 科学大数据集成共享进展及面临的挑战. 中国科技资源导刊，2017，49(5)：2-11)

[14] Wang Cheng-Bin, Ma Xiao-Gang, Chen Jian-Guo. The application of data pre-processing technology in the geoscience big data. Acta Petrologica Sinica, 2018, 34(2): 303-313(in Chinese)
(王成彬，马小刚，陈建国. 数据预处理技术在地学大数据中应用. 岩石学报，2018，34(2)：303-313)

[15] Hutson M. Artificial intelligence faces reproducibility crisis. Science, 2018, 359(6377): 725-726

[16] Wang Lu. The application of deep learning in high energy physics. Physics, 2017, 46(9): 597-605(in Chinese)
(汪璐. 深度学习在高能物理领域中的应用. 物理，2017，46(9)：597-605)

[17] Liu Yi-Di, Yang Qi, Li Yao, et al. Application of machine learning in organic chemistry. Chinese Journal of Organic Chemistry, 2020, 40(11): 3812-3827(in Chinese)
(刘伊迪，杨骐，李遥等. 机器学习在有机化学中的应用. 有机化学，2020，40(11)：3812-3827)

[18] Huang Fang, Yang Hong-Fei, Zhu Xun. Progress in the application of artificial intelligence in new drug discovery. Progress in Pharmaceutical Sciences, 2021, 45(7): 502-511(in Chinese)
(黄芳，杨红飞，朱迅. 人工智能在新药发现中的应用进展. 药学进展，2021，45(7)：502-511)

[19] Mi Xiao-Xi, Tang Ai-Tao, Zhu Yu-Chen, et al. Research progress of machine learning in material science. Materials Review, 2021, 35(15): 15115-15124(in Chinese)
(米晓希，汤爱涛，朱雨晨等. 机器学习技术在材料科学领域中的应用进展. 材料导报，2021，35(15)：15115-15124)

[20] Qiu Chen-Hui, Huang Chong-Fei, Xia Shun-Ren, et al. Application review of artificial intelligence in medical images aided diagnosis. Space Medicine & Medical Engineering, 2021, 34(5): 407-414(in Chinese)
(邱陈辉，黄崇飞，夏顺仁等. 人工智能在医学影像辅助诊断中的应用综述. 航天医学与医学工程，2021，34(5)：407-414)

[21] Zhou Zhi-Hua. Machine Learning. Beijing: Tsinghua University Press, 2016(in Chinese)
(周志华. 机器学习. 北京：清华大学出版社，2016)

[22] Li Hang. Statistical Learning Methods. 2nd Edition. Beijing: Tsinghua University Press, 2019(in Chinese)
(李航. 统计学习方法(第 2 版). 北京：清华大学出版社，2019)

[23] Sendek A D, Yang Q, Cubuk E D, et al. Holistic computational structure screening of more than 12000 candidates for solid lithium-ion conductor materials. Energy & Environmental Science, 2017, 10(1): 306-320

[24] Raccuglia P, Elbert K C, Adler P D F, et al. Machine-learning-assisted materials discovery using failed experiments. Nature, 2016, 533(7601): 73-76

[25] Shandiz M A, Gauvin R. Application of machine learning methods for the prediction of crystal system of cathode materials in lithium-ion batteries. Computational Materials Science, 2016, 117: 270-278

[26] Xie Q Q, Zhong L, Pan Y L, et al. Combined SVM-based and docking-based virtual screening for retrieving novel inhibitors of c-Met. European Journal of Medicinal Chemistry, 2011, 46(9): 3675-3680

[27] Song D, Chen Y, Min Q, et al. Similarity-based machine learning support vector machine predictor of drug-drug interactions with improved accuracies. Journal of Clinical Pharmacy and Therapeutics, 2019, 44(2): 268-275

[28] Haake R, Loizides C. Machine-learning-based jet momentum reconstruction in heavy-ion collisions. Physical Review C, 2019, 99(6): 064904

[29] Márquez-Neila P, Fisher C, Sznitman R, et al. Supervised machine learning for analysing spectra of exoplanetary atmospheres. Nature Astronomy, 2018, 2(9): 719-724

[30] Malanchev K L, Pruzhinskaya M V, Korolev V S, et al. Anomaly detection in the Zwicky Transient Facility DR3. Monthly Notices of the Royal Astronomical Society, 2021, 502(4): 5147-5175

[31] Webb S, Lochner M, Muthukrishna D, et al. Unsupervised machine learning for transient discovery in deeper, wider, faster light curves. Monthly Notices of the Royal Astronomical Society, 2020, 498(3): 3077-3094

[32] Weber L M, Robinson M D. Comparison of clustering methods for high-dimensional single-cell flow and mass cytometry data. Cytometry Part A, 2016, 89(12): 1084-1096

[33] McWhirter P R, Steele I A, Al-Jumeily D, et al. The classification of periodic light curves from non-survey optimized observational data through automated extraction of phase-based visual features//Proceedings of International Joint Conference on Neural Networks. Alaska, USA, 2017: 3058-3065

[34] Gabruseva T, Zlobin S, Wang P. Photometric light curves classification with machine learning. Journal of Astronomical Instrumentation, 2020, 9(1): 2050005

[35] Yang Q, Li Y, Yang J D, et al. Holistic prediction of the pKa in diverse solvents based on a machine learning approach. Angewandte Chemie, 2020, 132(43): 19444-19453

[36] Liu B, Xiong X, Hou G, et al. Applications of machine learning at BESIII//Proceedings of the EPJ Web of Conferences. High Tatra Mountains, Slovakia, 2019, 214: 06033

[37] Kumari P, Nath A, Chaube R. Identification of human drug targets using machine-learning algorithms. Computers in Biology and Medicine, 2015, 56: 175-181

[38] Kobak D, Berens P. The art of using t-SNE for single-cell transcriptomics. Nature Communications, 2019, 10(1): 1-14

[39] Techa-Angkoon P, Tanakul N, Bootkrajang J, et al. Identification of discriminative features from light curves for automatic classification of variable stars//Proceedings of the 18th International Joint Conference on Computer Science and Software Engineering(JCSSE). Lampang, Thailand, 2021: 1-6

[40] Visser K, Bosma B, Postma E. A one-armed CNN for exoplanet detection from light curves. arXiv preprint arXiv: 2105.06292, 2021

[41] Mahabal A, Sheth K, Gieseke F, et al. Deep-learnt classification of light curves//Proceedings of IEEE Symposium Series on Computational Intelligence. Honolulu, USA, 2017: 1-8

[42] Parks D, Prochaska J X, Dong S, et al. Deep learning of quasar spectra to discover and characterize damped Lyα systems. Monthly Notices of the Royal Astronomical Society, 2018, 476(1): 1151-1168

[43] Komiske P T, Metodiev E M, Schwartz M D. Deep learning in color: Towards automated quark/gluon jet discrimination. Journal of High Energy Physics, 2017, 2017(1): 1-23

[44] EstevaA, Kuprel B, Novoa R A, et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature, 2017, 542: 115-118

[45] Zhang K, Liu XH, Shen J, et al. Clinically applicable AI system for accurate diagnosis, quantitative measurements, and prognosis of COVID-19 pneumonia using computed tomography. Cell, 2020, 181(6): 1423-1433

[46] Xu Yang. Automatic recognition of optical transient in short timescale and super large field of view: Research and application[Ph. D. dissertation]. National Astronomical Observatories, Chinese Academy of Sciences, Beijing, 2020(in Chinese)
(徐洋. 超大视场短时标光学暂现源自动识别方法研究与应用[博士学位论文]. 中国科学院国家天文台, 北京, 2020)

[47] Coley C W, Jin W, Rogers L, et al. A graph-convolutional neural network model for the prediction of chemical reactivity. Chemical Science, 2019, 10(2): 370-377

[48] Morningstar W R, Hezaveh Y D, Levasseur L P, et al. Analyzing interferometric observations of strong gravitational lenses with recurrent and convolutional neural networks. arXiv preprint arXiv:1808.00011, 2018

[49] Yun J P, Shin W C, Koo G, et al. Automated defect inspection system for metal surfaces based on deep learning and data augmentation. Journal of Manufacturing Systems, 2020, 55: 317-324

[50] Naul B, Bloom J S, Pérez F, et al. A recurrent neural network for classification of unevenly sampled variable stars. Nature Astronomy, 2018, 2(2): 151-155

[51] Louppe G, Cho K, Becot C, et al. QCD-aware recursive neural networks for jet physics. Journal of High Energy Physics, 2019, 2019(1): 1-23

[52] Segler M H S, Kogej T, Tyrchan C, et al. Generating focused molecule libraries for drug discovery with recurrent neural networks. ACS Central Science, 2018, 4(1): 120-131

[53] Olivecrona M, Blaschke T, Engkvist O, et al. Molecular de-novo design through deep reinforcement learning. Journal of Cheminformatics, 2017, 9(1): 1-14

[54] Shen Z, Bao W, Huang D S. Recurrent neural network for predicting transcription factor binding sites. Scientific Reports, 2018, 8(1): 1-10

[55] Li Y, Wang S, Umarov R, et al. DEEPre: Sequence-based enzyme EC number prediction by deep learning. Bioinformatics, 2018, 34(5): 760-769

[56] Quang D, Xie X. DanQ: A hybrid convolutional and recurrent deep neural network for quantifying the function of DNA sequences. Nucleic Acids Research, 2016, 44(11): e107-e107

[57] Sun Y, Zhao Z, Ma X, et al. Short-timescale gravitational microlensing events prediction with ARIMA-LSTM and ARIMA-GRU hybrid model//Proceedings of International Conference on Big Scientific Data Management. Beijing, China, 2018: 224-238

[58] Ravanbakhsh S, Lanusse F, Mandelbaum R, et al. Enabling dark energy science with deep generative models of galaxy images//Proceedings of the 31st AAAI Conference on Artificial Intelligence. San Francisco, USA, 2017: 1488-1494

[59] Mustafa M, Bard D, Bhimji W, et al. CosmoGAN: Creating high-fidelity weak lensing convergence maps using Generative Adversarial Networks. Computational Astrophysics and Cosmology, 2019, 6(1): 1

[60] Schawinski K, Zhang C, Zhang H, et al. Generative adversarial networks recover features in astrophysical images of galaxies beyond the deconvolution limit. Monthly Notices of the Royal Astronomical Society: Letters, 2017, 467(1): L110-L114

[61] Van der Maaten L, Hinton G. Visualizing non-metric similarities in multiple maps. Machine Learning, 2012, 87(1): 33-55

[62] Gómez-Bombarelli R, Wei J N, Duvenaud D, et al. Automatic chemical design using a data-driven continuous representation of molecules. ACS Central Science, 2018, 4(2): 268-276

[63] Gilmer J, Schoenholz S S, Riley P F, et al. Neural message passing for quantum chemistry//Proceedings of the 34th International Conference on Machine Learning. Sydney, Australia, 2017: 1263-1272

[64] LeCun Y, Bottou L, Bengio Y, et al. Gradient-based learning applied to document recognition. Proceedings of the IEEE, 1998, 86(11): 2278-2324

[65] Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks. Communications of the ACM, 2017, 60(6): 84-90

[66] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition//Proceedings of the 29th IEEE Conference on Computer Vision and Pattern Recognition. Nevada, USA, 2016: 770-778

[67] Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation//Proceedings of the 28th IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA, 2015: 3431-3440

[68] Wang Z, Yan W, Oates T. Time series classification from scratch with deep neural networks: A strong baseline//Proceedings of the International Joint Conference on Neural Networks (IJCNN). Alaska, USA, 2017: 1578-1585

[69] Bai S, Kolter J Z, Koltun V. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. arXiv preprint arXiv:1803.01271, 2018

[70] Hochreiter S, Schmidhuber J. Long short-term memory. Neural Computation, 1997, 9(8): 1735-1780

[71] Cho K, Van Merriënboer B, Bahdanau D, et al. On the properties of neural machine translation: Encoder-decoder approaches//Proceedings of the SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation. Doha, Qatar, 2014: 103-111

[72] Zhang Aston, Li Mu, Lipton Z C, et al. Dive into Deep Learning. Beijing: Posts & Telecom Press, 2019(in Chinese)
(张·阿斯顿, 李沐, 立顿·扎卡里等. 动手学深度学习. 北京: 人民邮电出版社, 2019)

[73] Hu Ming-Fei, Zuo Xin, Liu Jian-Wei. Survey on deep generative model. Acta Automatica Sinica, 2022, 48(1): 40-74(in Chinese)
(胡铭菲, 左信, 刘建伟. 深度生成模型综述. 自动化学报, 2022, 48(1): 40-74)

[74] Kingma D P, Welling M. Auto-encoding variational Bayes. arXiv preprint arXiv:1312.6114, 2013

[75] Goodfellow I, Pouget-Abadie J, Mirza M, et al. Generative adversarial nets//Proceedings of the 27th International Conference on Neural Information Processing Systems. Montreal, Canada, 2014: 2672-2680

[76] Chen Yuan-Qiong, Zou Bei-Ji, Zhang Mei-Hua, et al. A review on deep learning interpretability in medical image processing. Journal of Zhejiang University (Science Edition), 2021, 48(0): 18-29+40(in Chinese)
(陈园琼, 邹北骥, 张美华等. 医学影像处理的深度学习可解释性研究进展. 浙江大学学报(理学版), 2021, 48(1): 18-29+40)

[77] Ai Peng-Cheng. Online feature extraction algorithms for high energy physics based on convolutional neural networks[Ph. D. dissertation]. Central China Normal University, Wuhan, 2020(in Chinese)
(艾鹏程. 基于卷积神经网络的高能物理事例特征信息在线提取算法研究[博士学位论文]. 华中师范大学, 武汉, 2020)

[78] Hermann J, Schätzle Z, Noé F. Deep-neural-network solution of the electronic Schrödinger equation. Nature Chemistry, 2020, 12(10): 891-897

[79] Liu Xiao-Fan, Sun Xiang-Yu, Zhu Xun. Current situation and challenges facing artificial intelligence in its application in new drug research and development. Progress in Pharmaceutical Sciences, 2021, 45(7): 494-501(in Chinese)
(刘晓凡, 孙翔宇, 朱迅. 人工智能在新药研发中的应用现状与挑战. 药学进展, 2021, 45(7): 494-501)

[80] Ravuri S, Lenc K, Willson M, et al. Skilful precipitation nowcasting using deep generative models of radar. Nature, 2021, 597: 672-677

[81] Racah E, Beckham C, Maharaj T, et al. ExtremeWeather: A large-scale climate dataset for semi-supervised detection, localization, and understanding of extreme weather events//Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach, USA, 2017: 3405-3416

[82] Wan M, Wu C, Wang J, et al. Column store for GWAC: A high-cadence, high-density, large-scale astronomical light curve pipeline and distributed shared-nothing database. Publications of the Astronomical Society of the Pacific, 2016, 128(969): 114501

[83] Yang Chen, Weng Zu-Jian, Meng Xiao-Feng, et al. Data management challenges and real-time processing technologies in astronomy. Journal of Computer Research and Development, 2017, 54(2): 248-257(in Chinese)
(杨晨, 翁祖建, 孟小峰等. 天文大数据挑战与实时处理技术. 计算机研究与发展, 2017, 54(2): 248-257)

[84] Cui C Z, Zhao Y H. Worldwide R&D of virtual observatory. Proceedings of the International Astronomical Union, 2007, 3(S248): 563-564

[85] Baker N, Alexander F, Bremer T, et al. Workshop report on basic research needs for scientific machine learning: Core technologies for artificial intelligence. USDOE Office of Science, Washington, USA: Technical Report: 1478744, 2019

[86] Hoi S C H, Sahoo D, Lu J, et al. Online learning: A comprehensive survey. Neurocomputing, 2021, 459: 249-289

[87] Masana M, Liu X, Twardowski B, et al. Class-incremental learning: Survey and performance evaluation on image classification. arXiv preprint arXiv:2010.15277, 2020

[88] Zhang C, Wang C, Hobbs G, et al. Applying saliency-map analysis in searches for pulsars and fast radio bursts. Astronomy & Astrophysics, 2020, 642: A26

[89] Yip K H, Changeat Q, Nikolaou N, et al. Peeking inside the black box: Interpreting deep learning models for exoplanet atmospheric retrievals. The Astronomical Journal, 2021, 162(5): 195

[90] Sedaghat N, Romaniello M, Carrick J E, et al. Machines learn to infer stellar parameters just by looking at a large number of spectra. Monthly Notices of the Royal Astronomical Society, 2021, 501(4): 6026-6041

[91] Vidal R, Bruna J, Giryes R, et al. Mathematics of deep learning. arXiv preprint arXiv:1712.04741, 2017

[92] Smith L N. Cyclical learning rates for training neural networks// Proceedings of the IEEE Winter Conference on Applications of Computer Vision. CA, USA, 2017: 464-472

[93] Moseley B. Physics-informed machine learning: From concepts to real-world applications [Ph. D. dissertation]. University of Oxford, Oxford, UK, 2022

[94] Daw A, Thomas R Q, Carey C C, et al. Physics-guided architecture (PGA) of neural networks for quantifying uncertainty in lake temperature modeling//Proceedings of the 2020 SIAM International Conference on Data Mining. Ohio, USA, 2020: 532-540

[95] Udrescu S M, Tan A, Feng J, et al. AI feynman 2.0: Pareto-optimal symbolic regression exploiting graph modularity// Proceedings of the 34th International Conference on Neural Information Processing Systems. Vancouver, Canada, 2020: 4860-4871

[96] Ba Y, Zhao G, Kadambi A. Blending diverse physical priors with neural networks. arXiv preprint arXiv: 1910.00201, 2019

[97] Liang En-Si. Detection of stellar flares via photometry and confirmation of an exoplanet [Ph. D. dissertation]. Nanjing University, Nanjing, 2020(in Chinese)
(梁恩思. 恒星耀发的测光观测及系外行星的证认[博士学位论文]. 南京大学, 南京, 2020)

[98] Yan Y, He H, Li C, et al. Characteristic time of stellar flares on Sun-like stars. Monthly Notices of the Royal Astronomical Society: Letters, 2021, 505(1): L79-L83

[99] Davenport J, Hawley S L, Hebb L, et al. Kepler flares II: The temporal morphology of white-light flares on GJ 1243. The Astrophysical Journal, 2014, 797(2): 122

[100] Richards J W, Starr D L, Butler N R, et al. On machine-learned classification of variable stars with sparse and noisy time-series data. The Astrophysical Journal, 2011, 733(1): 10

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

[102] Ribeiro M T, Singh S, Guestrin C. "Why Should I Trust You?" Explaining the predictions of any classifier//Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Francisco, USA, 2016: 1135-1144

[103] Abadi M, Barham P, Chen J, et al. TensorFlow: A system for large-scale machine learning//Proceedings of the 12th USENIX Conference on Operating Systems Design and Implementation. Savannah, USA, 2016: 265-283

[104] Paszke A, Gross S, Massa F, et al. PyTorch: An imperative style, high-performance deep learning library//Proceedings of the 33rd International Conference on Neural Information Processing Systems. Vancouver, Canada, 2019: 8026-8037

MENG Xiao-Feng, Ph. D., professor, Ph. D. supervisor. His main research interests include scientific data management, cloud data management, privacy protection and interdisciplinary researches like social computing.

HAO Xin-Li, Ph. D. candidate. Her main research interests include intelligent scientific discovery, time series analysis, interpretable machine learning.

MA Chao-Hong, Ph. D. candidate. Her main research interests include scientific data management, machine learning for database systems and learned indexes.

YANG Chen, Ph. D., postdoctoral fellow. His primary research focus is real-time analysis of big data, performance profiling of big data system and artificial intelligence for IT operations.

MAOLINIYAZI Ai-Shan, Ph. D. candidate. His main research interests include natural language processing, knowledge graph and machine learning.

WU Chao, Ph. D., associate professor. His research interest covers data mining and astronomical transient search.

WEI Jian-Yan, Ph. D., professor, Ph. D. supervisor. His research interest covers observation and science of astronomical transients.

Background

"AI for Science" has received extensive attention from academia and industry and has become a research hotspot. With the construction of large-scale scientific devices and the development of major scientific experiments, scientific discovery research cannot completely rely on expert experience to capture and study rare scientific phenomena from massive data. Using machine learning technology to discover rare scientific phenomena, study complex objects, and solve complex problems from massive scientific data has become the preferred solution in the scientific field.

However, the current research only focuses on specific machine learning algorithms and has not abstracted a general machine learning research framework for AI-driven scientific discovery tasks.

Therefore, this paper first comprehensively summarizes the research status of machine learning in various scientific fields and discusses the common problems and challenges.

Secondly, on this basis, this paper proposes a general research framework of intelligent scientific discovery based on machine learning. It describes an efficient mode of applying machine learning to scientific discovery and can guide scholars in various scientific fields. At the same time, it opens a window for scholars related to machine learning to further understand scientific discoveries.

Finally, this paper conducts a case study to verify the effectiveness of the framework. It is a time-domain astronomy scientific discovery mission: celestial transient event discovery. A series of experiments demonstrate that only when domain knowledge is properly combined, machine learning algorithms can better serve intelligent scientific discovery. Before this paper, the authors analyzed and managed scientific big data since 2016 in the National Key R&D Program "Scientific Big Data Management System". They have developed a relatively deep understanding of scientific discoveries during working with astronomers. A series of related works are published in ICDE, TKDE, EDBT, etc. In the process of participating in the National Natural Science Foundation of China project "Intelligent Analysis of Astronomical Big Data for Large Field-of-View Short-Timescale Sky Survey", the authors have further thought about the scientific discovery of intelligence, thus completing this paper.

This work was partially supported by grants from the National Natural Science Foundation of China (62172423, 91846204, U1931133)