

驱动智能教育奇点式发展的人工智能数据技术

——评《人工智能与大数据技术导论》



书名：

《人工智能与大数据技术导论》

作者：杨正洪 郭良越 刘玮

出版社：清华大学出版社

ISBN：978-7-3025-1798-6

出版时间：2018年12月

定价：98元

2017年由于人工智能技术的应用范围迅速拓展而被称为人工智能元年。但在教育领域，因为相关数据收集难度较高，人工智能技术驱动下的智能/智慧教育领域发展的速度一直差强人意。直到2020年，大数据技术支撑下的人工智能才真正以奇点模式迅速覆盖教育领域，彻底改变了人们的学习和教学工作方式。由杨正洪、郭良越和刘玮编著、清华大学出版社于2019年出版的《人工智能与大数据技术导论》一书，深入浅出地阐述了人工智能与大数据涉及的技术逻辑与理论基础，进而帮助读者全面理解人工智能技术的知识构架。

《人工智能与大数据技术导论》全书共包括十六章，主要内容包括人工智能的整体概述，AI产业、数据、机器学习概述，模型、机器学习算法、深度学习、TensorFlow、神经网络、知识图谱、数据挖掘以及银行业、医疗、工农业等行业人工智能的应用情况。同时，该书附录部分还给出了极富参考价值的大数据与人工智能产业参考资料。

数据作为推动AI智能发展的三大动力之一，是驱动AI智能发展最核心的生产要素。2020年以前一直是智能教育发展的瓶颈。主要原因有两点，一方面，教育数据的增长与金融、广告、影视等行业的大数据形成鲜明对比，教育数据的丰富与否制约着机器学习、深度学习在智慧教育中的应用。另一方面，数据产生价值的难度大、链条长，涉及数据的采集、整合与分析等多个平台的协作，只有运用合理的数据平台才能有效缩短这一链条。

大数据实现图像识别、语音识别等AI技术，需要依靠机器学习。机器学习基于概率、统计、优化等数学理论，其是基于一系列算法的解决问题的一种方法，能够使机器从大量样本数据中发现数据中暗藏的规律并自动学习规则，最终实现预测未知数据。机器学习的本质即通过大量样本的训练获得经验模型，最终产生预测新事物的能力，是对输入到输出的映射。

机器学习的任务流程一般包括数据导入、探索性分析、数据清洗、特征工程、算法选择和优化、模型训练和评估等六个步骤，并针对不同的具体任务会存在差异和侧重。在数据导入环节，以图片或音频等形式呈现的原始数据，需先转化成机器学习算法能够读入的结构化数据，这类数据的特征是以列的形式展开。在探索性分析环节，往往通过可视化工具对原始数据进行初步分析，挖掘数据集中有价值的信息并为数据清洗和特征工程（特征提取）提供方向，而特征提取是进行机器学习建模前的最为重要且极为耗时的一个环节。模型的训练即参数的求解，需通过算法来求解参数，以尽可能找到误差最小的参数为目标，误差函数的大小成为评估参数优劣的标准，可见参数的确定非常关键。

以海量数据积累为基础的大数据、算法和并行计算能力是构成人工智能发展的三要素。开源的机器学习平台能大大缩短开发时间，有效提升训练结果。机器学习首选的常用编程语言Python，内置许多预先写好的实用代码Library，这便于直接用来解决机器学习遇到的实际问题。Python具有完备的机器学习库，能够整合大量机器学习模型，使用者即使不懂模型的算法和原理，只需调动程序包，即可获得需要的结果。当然，受数据集、结构等多种因素的制约，没有一种算法能够解决所有问题，研究者需针对不同问题尝试不同的算法，然后通过保留的测试集对性能进行评估，进而选出最佳的算法。

随着训练深度的加深及神经网络新思路的出现，深度学习算法取得了颠覆性突破。它改变了传统机器学习通过算法对数据处理、分类、选取特征的路径，其模仿生物学神经元传递的过程，运用复杂模型建构多层神经网络，这种近似人脑工作原理的“特征学习器”使其能自主学习如何抓取特征，并通过训练神经网络中的参数，使神经网络模型具有预测能力。

数据智能技术驱动下的人工智能技术虽然与最初人们对生物智能技术的期待存在一定差异，但其越来越多地应用于人们日常生活的各方面。同时人工智能也正在改变着教育领域的发展，尤其是2020年大规模的线上教学必然产生大量的教育教学数据，海量的教育教学数据正加速驱动智能教育时代的来临。

（王巍，新疆师范大学，华东师范大学，讲师，博士研究生）