机器学习算法在提升大数据处理效率中的应用研究

张晖

山东济宁烟草有限公司，山东济宁，272000；

摘要：大数据时代的数据处理需求呈现指数级增长态势，传统计算架构在应对海量数据实时分析任务时面临严峻挑战。机器学习算法的创新应用为提升数据处理效能提供了新的技术路径，其核心价值在于通过智能决策优化计算资源分配与任务调度机制。当前实践中，数据预处理效率低下、特征工程耗时过长、模型训练资源浪费等问题普遍存在，反映出算法设计与工程实践之间的协同性不足。技术效能的突破需要建立算法优化与系统架构改进的联动机制，在数据处理全流程中实现智能决策与计算资源的动态适配。

关键词：机器学习算法；大数据；效率；应用

DOI：10.69979/3041-0673.25.10.096

引言

数据规模与复杂度的持续攀升，推动数据处理技术向智能化方向加速演进。机器学习算法在提升数据处理效率方面的潜力，不仅体现在计算过程的加速优化，更在于其对数据处理范式的根本性革新。传统批处理模式在应对高并发实时分析需求时，暴露出的响应延迟、资源僵化配置、异常处理能力薄弱等缺陷，亟待通过算法创新实现处理流程的动态优化。算法效能的释放需要突破静态规则约束，构建基于数据特征自感知的弹性处理机制。

1 机器学习算法概述

机器学习算法身为人工智能领域的关键支撑，正在极大地改变我们对于数据分析以及模式识别的认知和实践，展开来说，机器学习算法是一系列可使计算机系统能从数据里自动学习并提升性能的数学模型与计算程序，这些算法不需要明确的编程指令，而是借助分析大量数据中的模式和关系，自行调整参数，达成预测、分类、聚类、回归等多种任务。机器学习算法的种类丰富多样，各有特点，监督学习算法，像线性回归、决策树以及神经网络，依靠标注好的训练数据，凭借学习输入与输出之间的映射关系，对新的未知数据做出准确预测，无监督学习算法，例如聚类分析，可在无标签数据中挖掘潜在结构，揭示数据内在规律，强化学习算法依靠智能体与环境交互，依据奖励信号持续优化策略，实现复杂决策任务[1]。

2 提升大数据处理效率的重要性

2.1 加速决策制定，抢占市场先机

在如今这个处于信息爆炸状态的时代之中，数据成为了企业极为宝贵的资产之一，有高效的大数据处理能力，企业有能力快速地从海量的数据里提取出有价值的信息，为决策层给予及时且准确的洞察，这样一种快速响应的机制，可让企业在面对市场发生变化的时候迅速地调整自身策略，抢占市场的先机。就像在电商领域，借助于对用户行为数据展开实时分析，企业可精准地推送个性化商品，以此提升用户体验，并且促进销售增长，提升大数据处理的效率对企业的战略决策以及市场竞争力而言有着相当关键的意义[2]。

2.2 优化资源配置，降低运营成本

大数据并非仅仅与决策相关联，实际上它涉及到企业运营的各个方面，有高效的数据处理能力，可帮助企业更深入地了解自身内部的运营状况，精准识别出存在资源浪费以及效率低下的环节，借助数据分析，企业可对库存进行合理规划，优化生产流程，调整人力资源配置，达成资源的最大化利用效果。这可削减运营成本，而且还可以提高整体运营效率，提高企业的盈利能力，在当下资源变得日益紧张的形势下，这种优化资源配置的能力变得日益凸显其关键性。

2.3 推动创新发展，引领行业变革

大数据作为创新的源头活水，高效的数据处理是把这一源头转化为切实创新成果的关键环节，借助对大数据展开挖掘与分析，企业可寻觅到新的市场契机，研发新产品，开创全新服务模式，这种依靠数据的创新，可以契合消费者变得日益多样的需求，还可引领整个行业发生变革与发展。以医疗健康领域为例，大数据分析正促使精准医疗不断发展，为个性化治疗方案的制定给予

科学依据，提升大数据处理效率，对企业的持续创新以及行业的长远发展有着意义[3]。

3 机器学习算法在提升大数据处理效率中的应用

3.1 算法优化与定制化

大数据有高维特性，传统机器学习算法常面临维度灾难问题，使得计算复杂度迅速上升，处理效率降低，以支持向量机算法来讲，在高维数据空间里，寻找最优超平面的计算量会随特征维度增加呈指数级增长，要解决此问题，可采用核技巧结合特征选择的办法，核技巧把原始数据映射到高维特征空间，让数据在高维空间线性可分，不过直接在高维空间计算会产生巨大计算负担。凭借特征选择，选出对分类或回归任务最有影响力的特征子集，能减少数据维度，降低计算复杂度，比如在文本分类任务中，借助信息增益、卡方统计等方法评估和选择文本特征，去除冗余及无关特征，之后将筛选后的特征输入支持向量机算法，利用核函数把文本特征映射到高维空间进行分类。如此保留了数据关键信息，又提升了算法处理效率[4]。

对于时间序列数据而言，传统的机器学习算法或许难以有效地捕捉数据里的时间依赖关系，鉴于此，可进行定制化开发基于循环神经网络也就是RNN及其变体像长短期记忆网络 LSTM、门控循环单元 GRU 的算法，这些算法借助引入记忆单元，可以保存并传递历史信息，处理时间序列数据当中的长期依赖问题。以股票价格预测作为例子，股票价格数据有十分突出的时间序列特征，会受到历史价格、交易量等诸多因素的影响，运用 LSTM 模型，可以把历史股票价格、交易量等数据当作输入序列，依靠 LSTM 单元的记忆以及遗忘机制，捕捉数据中的时间依赖关系，输出未来股票价格的预测数值。和传统的线性回归模型相比较，LSTM 模型可更为准确地捕捉股票价格的动态变化，提升预测的准确性[5]。

3.2 分布式计算与机器学习算法的深度集成

随着大数据规模持续不断地扩大，单机计算已经没办法契合数据处理提出的需求了，像 Apache Hadoop、Apache Spark 这类分布式计算框架，为大数据处理给予了强大的计算能力，把机器学习算法和分布式计算框架深度整合在一起，可充分施展两者的优势，提升大数据处理的效率。

以 Apache Spark 来说，它给出了丰富的机器学习库 MLlib，这里面有多种常用的机器学习算法，像分类、回归以及聚类等，在处理大规模数据集之际，可借助 Spark 的分布式计算能力，把数据分布于多个节点上并行处理，比如在训练一个大规模的分类模型时，可以把数据集划分成多个子集，分配到不同的计算节点上，每个节点独立计算梯度，之后凭借参数服务器进行模型参数的更新。这种并行计算方式极大地缩短了模型的训练时间，并且 Spark 还有内存计算功能，把数据缓存到内存里，减少了磁盘 I/O 操作，提高了数据处理效率。

在实际应用里，以图像分类任务作为例子，当面对数量众多的海量图像数据时，传统的单机训练方式耗费的时间较长，可能需要数天甚至数周，要是采用分布式机器学习算法并与 Spark 集成，那么可把图像数据分布于多个计算节点上，同时开展特征提取以及模型训练工作，每个节点负责处理一部分图像数据，提取特征之后把特征向量发送至主节点进行模型聚合以及参数更新。借由这样的方式，训练时间大幅缩短，图像分类任务的效率得以提升[6]。

3.3 特征工程优化

特征工程是机器学习中的关键环节，它直接影响模型的性能和处理效率。在大数据处理中，由于数据维度高、噪声大，特征工程变得更加复杂和重要。优化特征工程可以从特征选择、特征提取和特征转换等方面入手，挖掘数据中的核心价值，提升大数据处理效率。

特征选择属于去除冗余以及无关特征的进程，可削减数据的维度，让模型的计算复杂度有所降低，常见的特征选择方法囊括过滤法、包装法以及嵌入法，过滤法借助统计指标像方差、相关系数等对特征开展排序和筛选，包装法依靠构建模型并评估特征子集的性能来挑选特征，嵌入法把特征选择过程融入到模型训练过程里，例如决策树算法中的特征关键性评估。比如在文本分类任务当中，采用 TF-IDF 也就是词频-逆文档频率作为特征权重，接着借助过滤法去除低权重的特征，减少特征维度，提升模型训练效率，特征提取是从原始数据里提取更具意义的特征表示，对提高模型的泛化能力以及处理效率有帮助，在图像处理领域，卷积神经网络即 CNN 凭借卷积层和池化层自动提取图像的局部特征与全局特征，避免了手工设计特征的繁杂与局限。例如在人脸识别任务中，CNN 可自动学习人脸的关键特征，像眼睛、鼻子、嘴巴等部位的形状和位置信息，提高了人脸识别的准确性和效率，特征转换是将原始特征转变为更适宜模型学习的形式，比如标准化、归一化等，在大数据处

理过程中，因为数据分布存在差异，特征转换可以改善数据的分布特性，提高模型的训练效率。例如在回归分析里，对特征进行标准化处理，使其均值为 0，方差为 1，可提高梯度下降算法的收敛速度，减少训练时间[7]。

3.4 自动化与智能化流程构建

在大数据处理的复杂生态环境里，传统流程对人工操作有着高度的依赖性，而这已然成为限制处理效率以及质量提升的一个关键妨碍因素，就数据源头的清洗环节而言，人工需要花费大量的时间去逐个排查缺失值以及异常值，在面对数据量极为庞大的场景时，这样的操作很容易因为人为的疏忽而致使关键数据被误删，或者保留错误的数据，对后续的分析结果产生影响。到了特征工程阶段，人工进行特征选择往往是依据经验来开展的，很难全面地考虑到数据多维度特征之间的关联情况，这有可能造成模型训练的时候出现特征冗余或者关键特征缺失的问题，使得模型性能有所降低，在模型选择环节，人工要评估不同模型在特定数据集上的表现，耗费时间又耗费精力，并且很难准确判断不同模型在复杂数据分布状况下所有的潜在优势[8]。

要突破当前面临的这一困境，构建自动化以及智能化的处理流程是十分必要的，自动化流程借助整合多种工具以及技术来达成数据处理各个环节的自动化操作，在数据清洗这个方面，依靠 Python 的 Pandas 库，可以编写灵活的脚本来达成批量数据处理，借助定义规则可自动识别缺失值，依照特定的统计方法比如均值填充、中位数填充来进行填充，或者依据异常值检测算法像是基于标准差、箱线图法来自动标记并且处理异常值。在特征选择环节，自动化特征选择工具运用多种评估指标例如信息增益、卡方统计、递归特征消除等，依据数据特征自动筛选出最优特征子集，减少人工干预所带来的主观性以及不确定性，在模型训练与评估方面，自动化机器学习框架如 Scikit-learn，可自动遍历多种模型参数组合，借助交叉验证等方法评估模型性能，快速选择出最优模型，极大地缩短了模型开发周期。智能化流程则融合深度学习、强化学习等前沿技术，实现更高级别的自动化与智能化，深度学习模型凭借其强大的特征学习能力，可自动从原始数据中提取高层特征，无需人工开展复杂的特征工程，比如在图像识别任务中，卷积神经网络凭借多层卷积和池化操作，自动学习图像的边缘、纹理、形状等特征，较大提高了图像识别的准确性以及效率。强化学习算法则依靠智能体与环境的交互，根据奖励信号自动优化模型参数，在模型超参数调优场景中，强化学习算法可自动探索不同的参数组合，找到最优的超参数设置，提高模型的泛化能力以及处理效率。

4 结束语

在大数据时代，机器学习算法为提升数据处理效率开辟了新路径。通过对本文诸多研究内容的梳理可知，合理运用机器学习算法能有效应对大数据的复杂性、多样性与海量性。未来，随着算法不断优化与创新，其将在大数据处理领域发挥更大作用，推动各行业实现智能化转型，为社会发展注入新动力，让我们共同期待这一美好前景。

参考文献

[1]周匀茜,焦鹏,邓正万,等.大数据分析中的机器学习算法应用[J].集成电路应用,2025,42(01):292-293.

[2]王佳倩,陈佳骏.机器学习和大数据分析在电气设备识别及故障预警中的应用[J].电子技术,2024,53(11):427-429.

[3]王强,刘海德,牛清娜,等.基于场景化的大数据+AI算法仓平台研究[J].电脑知识与技术,2024,20(14):73-75.

[4]金鹏.大数据技术和机器学习算法在热网集控系统中的应用[J].电动工具,2022,(01):27-29.

[5]张仕斌,黄曦,昌燕,等.大数据环境下量子机器学习的研究进展及发展趋势[J].电子科技大学学报,2021,50(06):802-819.

[6]刘言林.基于条件生成对抗网络的小样本机器学习数据处理算法研究[J].宁夏师范学院学报,2021,42(10):66-73.

[7]辛忠洋.基于机器学习的气体传感器数据处理算法解析[J].数字通信世界,2021,(08):128-129+144.

[8]孙开华,代余杰,封晴.基于机器学习的大数据分析和处理[J].中国新通信,2021,23(13):65-66.

作者简介：姓名:张晖，出生年月：1982 年 3 月，性别：女，民族：汉,籍贯：山东济宁，单位名称：山东济宁烟草有限公司，学历：本科，职称：无，主要研究方向：数据安全。