

AI技术在运营数据低代码分析方面的应用研究

文 | 操文煜

一、引言

随着信息技术特别是互联网技术的迅猛发展，社交网络、移动互联网、物联网、云计算等相继进入人们的日常工作和生活中，用户、企业和政府等个人和组织都在产生大量的数据，数据信息量正呈指数式爆炸增长。为了更好地理解用户需求，提高服务质量，以及洞悉用户的潜在价值，研究AI技术在运营数据低代码分析方面的应用，对帮助业务进行深入的数据挖掘和解决方案显得尤为重要。

二、相关工作

机器学习算法的应用日益广泛，相关的开源平台和工具也如雨后春笋般涌现，其中不乏一些非常优秀的产品，如Powerbi、Stata、Orange3等，在实践处理业务需求时所面临的关键痛点有：

数据的多样性与复杂性：随着数据类型的多样性和算法复杂性的提升，开源的算法在计算性能、稳定性和可扩展性方面面临着巨大挑战。特别是在运营领域，对于复杂多变的业务环境，需要更灵活、高效的解决方案。

安全与隐私的重要性：在实际业务场景中，数据分析涉及大量敏感数据。因此，确保数据操作的安全性和隐私性显得尤为重要，尤其是在运营领域，对于客户隐私的保护至关重要。

研发效率与专业门槛：传统的编译代码建模开发周期较长，对于人才的专业技术要求也较高。一些数据分析产品在扩展能力上存在不足，无法快速扩展算子并共享使用。同时，缺乏统一的模型管理和模型服务平台，导致业务部门在应用模型服务时时效性不佳。在运营中，迅速响应变化、快速优化模型以适应新的业务场景是至关重要的。

瞄准上述痛点，研究专业易用、深度定制的机器学习开放平台，以快捷低代码方式帮助业务进行深入的数据

挖掘和解决方案显得尤为重要。

三、企业级数据分析低代码平台

（一）平台特点

低代码数据分析平台，其独特之处在于提供了全生命周期、自助式的大数据挖掘平台，为用户提供了高度灵活和易用的数据分析工具。通过低代码数据分析平台，用户能够在整个数据处理生命周期中进行自助式操作，从数据的收集和清洗，到模型的构建和结果的可视化，实现了一体化的数据分析体验。

1.海量数据处理

低代码数据分析平台构建于分布式计算的基础上，汇集分布式算法和数据分析方法，包括数据预处理、特征工程、机器学习算法等，可高效的完成海量、高维度数据的复杂计算，满足企业级数据挖掘需求，帮助客户从海量数据中挖掘出业务价值。

2.自动化模型构建

(1)自助机器学习，支持与特征、模型、优化、评价有关的重要步骤进行自动化地学习，使得机器学习模型无需人工干预即可被应用。(2)模型自动学习，支持对已保存的模型，自动根据最新的数据进行训练，并能自动将符合一定要求的模型发布或者部署到生产中。(3)模型自动调参，支持对模型参数进行自动搜索，选择最优的模型参数。

3.完善的生态体系

工具支持从数据上传、数据预处理、特征工程、模型训练、模型评估到模型发布、模型和服务管理的机器学习全流程。并且拥有银行级别的数据安全权限管理。

4.拖拽式可视化建模

封装经典的机器学习算法，(1)支持在WEB 界面，通过拖、拉、拽等方式即可完成复杂数据挖掘流程。(2)

支持使用内置的自助机器学习功能进行自动建模,实现模型自动选择、参数自动探索、模型效果自动评估、模型自动向下传导及模型自动优化。

5.可视化的无缝集成

能够与BI工具的可视化功能无缝集成,将挖掘结果通过多样丰富的BI报表可视化手段进行分析展现。

6.一键式模型部署服务

将挖掘流程以服务的方式进行发布,用户直接调用服务链接访问,实现模型到业务的无缝衔接,支持Restful API数据接入到实验中,支持WEB服务。

(二) 技术架构

低代码数据分析平台的总体可分为五层:

数据源层:包括关系型数据库、HDFS、流式数据kafka、MPP数据库以及本地文件CSV、TXT等。

调度技术框架层:大数据计算平台Spark集群、高可用平台ZooKeeper、扩展开发平台Anaconda(Python)、DAG调度引擎集群、服务引擎集群。

分布式算法模型库:分类、聚类、回归、推荐、特征工程、统计、文本挖掘等。

数据分析管理层:DAG操作框架、实验管理、服务管理、数据可视化、资源监控、自定义管理。

应用层:流失预警、客户细分、异常值检测、价值分析、产品推荐、流量分析等。

其中,作为低代码数据分析平台的核心智能驱动内核,其计算引擎构件主要包含实验引擎、服务引擎、Python计算代理、Spark图计算代理等。其中,实验引擎负责接收实验执行请求,通过解析实验定义,生成Spark计算任务或Python计算任务,并分别发送给代理到对应的计算集群。

服务引擎可以将训练好的模型发布成Restful接口,便于提供模型预测服务给第三方系统调用。平台特有的大部分组件则由Spark分布式计算代理完成,承担着平台的核心计算任务。而对于Python计算任务,则发送到Python计算代理中完成,主要弥补平台组件的不足,增加扩展算法的能力。

(三) 技术应用场景

在数据分析低代码平台中,常见的技术场景是相对有限的,如分类、回归、聚类、异常检测等,但落实到实际项目的业务层出现的场景却各一。例如,分类这个技术场景,在电力行业是窃漏电检测,在工业设备检测为是否故障,在客户运营方面则为流失用户预测。因此,掌握常

见的技术场景,是进行下一步业务场景建模的核心关键。

在金融领域,通过对海量的交易数据进行分析,实现更精准的风险评估和预测,同时利用其强大的可视化功能,更直观地了解业务状况,做出更明智的决策。在电力行业,通过对能源生产和消耗数据的深度挖掘,帮助企业实现对电力系统的全面监控与管理。从发电设备的状态监测到电网运行的实时分析,为电力企业提供了全方位的数据支持,优化能源利用效率,提升系统的可靠性。在制造业方面,通过对生产数据的实时监控和分析,制造企业能够实现生产过程的精细化管理,提高生产效率,减少资源浪费,并优化供应链的运作。在政府行业,可以对各种社会经济数据的整合和分析,更好地了解社会发展趋势,为政策制定提供科学依据,推动社会治理的智能化升级。

1.聚类分群

“物以类聚,人以群分”即充分解释了特征分群,它是指将具有相似特征的群体或者对象聚合在一起成为一个类,在该类中的对象都是彼此相似的。因此,可根据不同群体的特征利用数据挖掘给客户、地区、品种等进行特征分群,然后依据不同分群的特点制定相应的策略。例如:对客户进行科学的分群,可为客户提供适配的产品、制定针对性的营销活动和管理用户,最终提升产品的客户满意度,实现商业价值。

聚类算法常用于无监督学习问题,对大量未知标注的数据集,按数据的内在联系将数据集划分为多个类别,使类别内的数据相似度较大,而类别间的数据相似度较小。常见算法:K均值、高斯混合模型。

应用场景:对客户进行分层营销、学生消费水平分群、电网用户增加画像维度、航空公司客户价值分析。

2.分类预测

类别预测指的是根据对象已知的类别及特征进行建模训练,再使用模型对其它未知的对象进行类别划分预测。例如:根据客户本身属性和历史行为特征等(年龄、性别、工作类型、婚姻状况、文化程度、个人贷款、收入情况)结合历史划分的类别进行建模,预测其他或将来的客户是否愿意办理相关业务,也可预测办理业务的意愿程度。

分类算法是根据已经分好类的一些数据,分析每一类的潜在特征并建立分类模型,用于识别未知数据的类别。常见算法:逻辑回归、朴素贝叶斯、决策树、随机森林、支持向量机、梯度提升决策树、多层感知机。

(下转60页)

理的先决条件,能够融合不同模态的语义信息,更全面分析及利用不同形态的信息理解与表达,优于单个模态信息的方法。这一过程中,自然语言处理技术能够通过自身语言分析整合能力,剔除多模态信息间的冗余信息,以跨媒体感知与分析技术的跨模态转化能力来实现多模态信息统一表达,从而获取更全面的语义特征。

四、基于自然语言处理技术的跨媒体感知与分析技术实施路径

首先,加强智能芯片创新研发力度。跨媒体感知与分析技术能够促使计算机对外界的感知与信息进行转化分析,形成对信息的内部整合与统一理解,从而将所理解内容进行统一表征,使各类信息系统能够实现智能化对答。为实现这一技术在现有计算平台上的应用,科研人员应以跨媒体感知与分析技术的能力为模型,构造模拟和超越生物感知的智能芯片和系统,进一步推动人工智能实现机器感知。

其次,构建应用跨媒体感知与分析技术的智能引擎。为实现人工智能的高效表达,研究人员可以构建独特

感知机制跨媒体智能引擎,建立从视频、图像等复杂数据中快速搜索兴趣目标的理论、模型和算法,实现具备自适应、自学习能力的智能感知系统。例如,研究超越人类感知能力的视觉、听觉感知系统,以及突破真实自然交互环境中的言语感知系统。

最后,推动跨媒体感知与分析技术在各行业的应用。跨媒体感知与分析技术能推进企业智能制造转型,推动中国制造业核心竞争力发生着前所未有的变化,促进我国智能产业沿着“共融机器人”的路径演进。“共融机器人”基于跨媒体感知与分析技术产生,可用于构建能适应不同环境,且能够更好地理解人行为意图,在一定的规则下实现人机、机器人群体间的自然交互与协同。随着智能科技不断发展并渗透至各行各业,跨媒体感知与分析技术的应用在医疗、教育、智能家居等行业已获得卓越成果,在未来将逐渐参与更多行业的发展,为我国经济社会发展提供动力。

作者单位:广州铁路职业技术学院 信息工程学院

(上接62页)

应用场景:银行客户流失预测、个人贷款违约预测、银行理财产品预定预测、用户窃漏电预测,山体滑坡预测等。

3. 回归预测

回归预测指的是根据某事件以往的历史数据记录(营销数据、交易数据、市场供需等),对未来该事件的目标走势(如价格等指标)进行预测,从而指导相关单位提前做出应对方案(规划资源调度或者价格调整等)。例如:根据某个区域以往的电力或供水系统的负荷消耗能源的情况,对未来一段时间某个区域的电力或供水系统的负荷消耗能源总量做预测。以此指导相关单位提前规划产能投入和相关基础设施建设,从而降低企业运营成本提高客户满意度。常见算法:线性回归、梯度提升回归树。

应用场景:服装门店销售额预测、用电需求预测、二手车交易价格预测、房子价格预测、超市商品销量预测。

4. 关联推荐

关联规则是反映一个事物与其他事物之间的相互依存性和关联性,用于从大量数据中挖掘出有价值的数项之间的相关关系,可从数据中关联分析出“由于某些事件

的发生而引起另外一些事件的发生”之类的规则。如平时在超市购买洗发水大多数也会购买沐浴露的情况,关联规则常用于做推荐问题分析。常见算法:FP-Growth。

应用场景:商品零售购物篮关联分析、客户理财产品关联分析、客户书籍喜好关联分析。

四、结语

本文先是介绍了企业级数据分析低代码平台的特点和技术架构,以此作为介绍逐步引入机器学习技术在低代码数据分析中的体现。而后,针对一些行业的典型应用场景展开举例,试图通过一些简单的案例介绍,能让读者对技术在工程上的应用有个清晰的认识。最后,结合我所情况做了一些探索性的分析介绍。随着增强分析与AI技术不断的更替,未来必然能引领数据分析走向智能化的道路,帮助我所提升各方面的能力,响应国家数字化改革的号召。

作者单位:南京电子技术研究所