doi:10.3772/j.issn.1002-0470.2022.10.008

边缘大数据分析预测建模方法研究①

钟运琴②* ** 朱月琴③*** **** 焦守涛*** ****

(* 中国科学院大学中国科学院大数据挖掘与知识管理重点实验室 北京 100190)

(** 国务院发展研究中心信息中心 北京 100010)

(*** 中国地质调查局发展研究中心 北京 100037)

(**** 自然资源部地质信息工程技术创新中心 北京 100037)

摘 要 随着物联网大数据分析实时性要求的提高,中心控制的云端大数据分析方法无法满足实时性和准确性要求,表现为响应延迟大、成本开销大、特定环境下的预测准确性低。本文提出了在海量实时数据如传感器数据、流数据等场景下的边缘侧大数据分析预测建模方法,该方法在边缘侧训练小数据样本,根据特定的应用场景多接入边缘侧进行分布式建模学习,分而治之地训练模型和推理预测分析。首先,通过将大数据分析和边缘计算相结合提出了边缘侧和云端协同的大数据分析预测建模的理论范式框架;其次,在该标准范式框架的基础上,设计了边缘侧大数据分析预测的训练算法和调优机制;最后实现了边缘侧大数据分析的训练和评估系统原型。在百个节点测试环境的实验结果表明,在实时大数据场景,同云端训练相比,本文提出的边缘侧大数据训练的性能效率平均提升了3.95 倍,网络通信量减少了 88.7%,边缘侧协同训练模型的预测准确率、召回率和 F1 值比传统训练方法可以提升 3% ~9%,请求预测的响应延迟降低了 67.5%。本文方法可有效应用于科学计算、智能金融、自动驾驶、安防监控、数据安全、智能工厂和智慧城市等领域,具有一定的借鉴价值。

关键词 边缘计算; 大数据分析; 边缘大数据; 边缘机器学习; 边云协同

0 引 言

随着 5G 应用的推广,实时大数据分析应用场景不断增多,涌现了基于大数据训练的智能分析预测算法模型,这些算法模型的训练过程通常利用强大的云计算能力训练海量结构化和非结构化数据集,从而达到准确预测的目的[1-5]。云算力是远程云中心执行大数据分析任务,适用于非实时、长周期历史数据、全局决策的应用场景[6-8]。但是,针对实时性、短周期数据、本地化决策等场景,云上执行大数据分析任务的模式就表现不好,尤其在 5G 控制、物联网传感器数据、监控流数据、无人驾驶等应用领域[9-13],如果将实时数据回传到远程的云中心去处理,将会造成响应延迟大、分析预测准确率低,而且在大数据场景下的成本开销大[13-18]。针对这类应用问题,本文提出了边缘侧大数据分析预测建模方法,数据的分析预测需要在边缘侧本地数据计算后立即做出响应,而不需要通过网络传输等待云中心模型下发和服务反馈,充分利用边缘计算能力分而治之、化整为零地将分布在各个场景的数据进行本地化处理,在实时性、短周期数据和本地决策等场景

① 国家自然科学基金(41872253)和国家重点研发计划(2018YFC1505501)资助项目。

② 男,1985 年生,博士,副研究员;研究方向:大数据分析和智能边缘计算;E-mail: zhongyunqin@163.com。

③ 通信作者,E-mail: yueqinzhu@163.com。

(收稿日期:2021-07-20)

方面具有不可替代的作用。

本文针对复杂场景的大数据分析预测任务,提出边缘侧大数据分析方法既有大数据样本训练而成的通用模型预测能力,也具有边缘小样本训练而成的特定预测能力。云计算和边缘计算的范式不同,云计算的演化过程是将分布在各地的数据集汇聚到云上利用云的算力进行集中式分析处理,它是中央处理的模式。边缘计算的核心是将各地(分中心)的数据在边缘侧的服务器上处理,在实时数据处理方面显示出很大的性能优势。它可以将收集的实时数据集如传感器数据、监控流数据,在边缘计算平台上进行预处理、清洗、训练小数据样本模型和预测任务,无需将全量数据通过网络传输到远程的云平台上处理,仅需将各个边缘侧清洗处理后的小样本数据集共享到云平台。同时,多个边缘侧的数据集汇聚成了边缘大数据保存在远程的云端,通过边云协同网络下发到边缘平台,因而,边缘侧不仅获得了云端模型的通用分析预测能力,还具备了边缘侧特定小样本训练而构建的精确场景分析能力。

本文的主要贡献为:(1)提出了边缘侧和云端协同的大数据分析处理框架,定义了在处理大数据分析任务时边缘侧训练和云端训练的边界和边云协同大数据分析理论框架,该理论框架能够统一处理包括实时、短周期、本地决策分析任务,以及非实时、长周期和全局决策型分析任务。(2)提出了边缘侧大数据分析训练的分布式机器学习建模方法,分布在各个地方的边缘计算平台以多接入方式连接云平台,在云平台上训练大样本数据生成模型参数,然后将模型下发到边缘计算平台。边缘侧收集到的特殊的小样本数据通过迁移学习的方式共享云端模型参数,本文设计了迁移学习方式在边缘侧对小样本数据进行模型训练,将云端模型参数作为边缘侧分析训练的初始参数,边缘计算平台输入小样本数据,使用深度学习神经网络(deep learning neural network, DLNN)分类算法进行训练,在边缘侧更新模型参数生成新的更加准确的 DLNN 模型用于分析预测任务。(3)实现了边缘侧大数据分析预测系统原型(edge big data analysis and predicate system prototype,EDAP)。EDAP 系统将实时大数据预测任务在边缘侧就地执行,基于迁移学习方式在本地训练更新模型,无需将全量数据回传到云端,能有效降低网络开销和时间成本,也提高了预测精度。实验结果显示,EDAP 边缘侧大数据建模训练的效率平均提升了 3.95 倍,网络传输量平均减少了 88.7%,在特定场景下训练出来的模型 AUC 评估指标值从云端模型精度的 72.6% 提升到了 98.6%,边缘侧协同训练模型的预测准确率、召回率和 F1 值比传统训练方法可以提升 3% ~9%;请求预测的响应延迟平均降低了 67.5%。因此,本文方法在科学计算、智能金融、决策控制等领域具有一定的借鉴价值。

1 边缘侧和云端协同的大数据分析处理范式框架

本文面向多源、异构、区域分散的多个分中心采集到的大数据,设计了边缘侧和云端协同(简称“边云协同”)(edge-cloud processing, ECP)的大数据分析处理范式框架,分析预测响应时间通常为毫秒级到秒级。ECP 边云协同框架能够有效提升大数据分析应用性能,并在边缘侧添加小样本数据训练建模能力,由此更好地支撑实时分析和移动分析应用。

ECP 边缘大数据处理框架如图 1 所示。ECP 范式框架由 3 大部分组成:应用侧、边缘计算侧、云数据中心侧。这 3 个部分以边缘计算侧为桥梁,缩小应用侧和云数据中心侧的鸿沟。

边缘计算侧大数据分析平台有多个边缘计算节点和边缘管理节点组成的边缘大数据分析集群,边缘计算节点用来执行大数据分析处理任务,并且由边缘管理节点将训练任务和推理预测任务采用分治算法进行划分,按任务类型分阶段调度到边缘节点处理。边缘侧大数据分析平台的处理方法包括以下 4 个步骤。

(1)将采集到的物联网数据、传感器数据和实时数据,汇聚到边缘节点,一个边缘节点接入多个实时数据采集端,同时由多个边缘节点并行分析处理。

(2)边缘节点将所管理的实时数据信息的元数据发送到边缘管理节点,由边缘管理节点执行元操作(metaoperation)。元操作包括设置大数据分析训

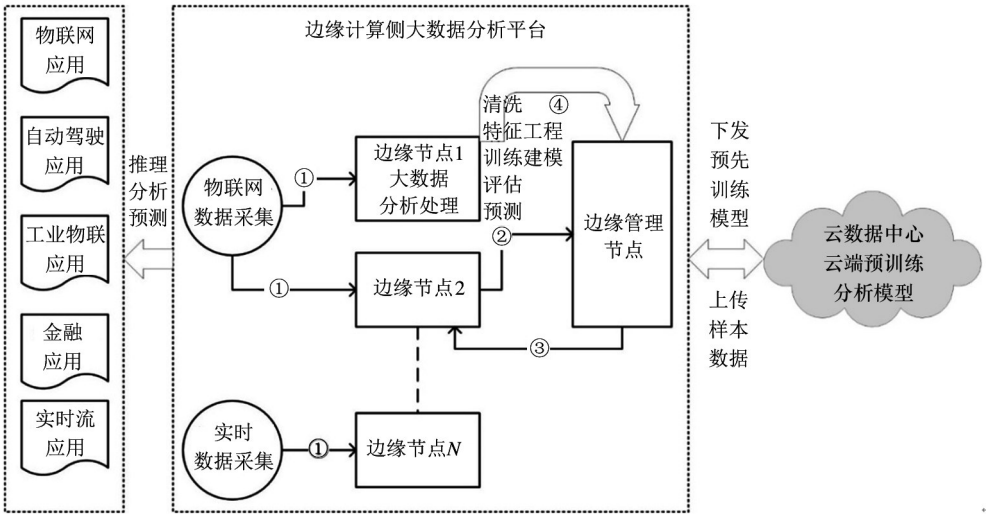

图 1　ECP 边缘大数据处理范式框架

练任务中机器学习算法的超参数和动态参数。超参数(super parameters)是用启发式算法在模型外部配置的变量参数;动态参数是历史数据训练学习到的变量参数。当在边缘侧针对特定问题调整机器学习算法时,使用网络搜索或随机搜索时将调整模型或命令的超参数,以发现一个可以使模型预测最熟练的模型参数。训练任务由边缘计算节点执行,边缘管理节点保存边缘节点每一步训练后的模型参数,让模型参数和超参数在所有边缘节点之间共享与复用。边缘计算节点的系统运行状态会以心跳方式主动报告给边缘管理节点。

(3)边缘管理节点承担任务划分、任务调度和容错处理。边缘管理节点将模型参数和超参数发送到边缘结算节点,边缘计算节点作为工作节点,在上一轮参数的基础上执行损失函数最优化算法训练运算得到本轮的参数。边缘管理节点根据评估样本数据集计算准确率(precision, P)、召回率(recall, R)、AUC 值(area under curve),以及损失函数值(loss function value,LFV),在分类模型和回归模型训练中,不断地更新模型参数和超参数,直到损失函数值 LFV 小于 δ(δ 趋近于零),P 值、R 值和 AUC 值无限接近于 1,其中 P, R, AUC ∈(0, 1),表示训练的模型的预测效果更优。

(4)边缘管理节点发送数据处理和建模训练程序指令给边缘计算节点,边缘计算节点执行数据加载、数据清洗、特征工程、特性属性选择、特征值处理、归一化、算法训练、建模、评估、推理、预测。多个边缘计算节点上存储了训练样本数据和评估样本数据。机器学习训练算法根据数据并行策略同时在多个边缘计算节点执行,每个边缘计算节点模型参数均共享自远程云数据中心的预训练模型参数,因而,边缘计算节点执行迁移学习算法,输入初始模型参数,训练后的新模型文件用于后续推理与预测。各个边缘计算节点训练后的模型文件传输至边缘管理节点,由边缘管理节点存储和分发模型。

经过以上 4 个步骤,边缘计算侧的大数据分析的最终任务执行结果——模型文件和评估结果,均汇聚到边缘管理节点,并由边缘管理节点作为入口与应用侧和云数据中心执行程序交互。

边缘计算侧的大数据分析和云数据中心侧的主要交互过程是:(1)边缘侧把数据质量审核后的小样本数据上传到云数据中心,多个边缘侧的样本数据回传到云数据中心形成大样本数据集。基于大样本数据集训练机器学习算法形成通用模型文件,该通用模型文件针对某类问题具有很好的泛化。(2)在云端进行训练算法,拟合出一个泛化能力较强的模型,并将通用模型保存在云端。(3)当有新增样本进入云端,云端自动启动训练程序迭代更新模型文件,并将模型下发到边缘侧,边缘侧的模型定期地同步云端的模型。因此,边缘侧和云端就形成了良好的交互,云端的预训练模型定期分发到边缘侧的管理节点,由管理节点分发至边缘计算节点,由边缘

计算节点根据预训练模型和特定小样本在本地训练形成特定的训练模型,最终每个边缘计算节点套用其训练模型执行推理和分析预测任务。

应用侧包括物联网应用、决策分析和实时流应用,交互过程是:(1)边缘侧将训练好的模型文件以接口形式进行封装,应用侧通过边缘网关调用接口进行推理和分析预测。(2)应用侧将实时采集到的数据输入到模型中运算得到实时结果,并以接口服务形式返回。

2 边缘侧大数据分析与分布式机器学习训练算法

针对边缘大数据分析训练任务提出了边缘侧机器学习(edge machine learning, EML)算法。EML 算法是一类在边缘侧实现的分布式机器学习算法,将机器学习任务运算分布在多个边缘计算节点协同执行,每个边缘计算节点均摊机器学习训练和推理预测工作负载。EML 算法解决了由于边缘侧设备算力性能低于云计算能力而不能训练大数据样本的问题。

在 EML 算法框架中,大样本数据集用于训练全局通用模型(global common model, GCM),由于边缘侧采集特定场景的新数据,因此边缘侧采集的小样本数据集在经过清洗、过滤和特征工程操作后,可以用来训练专用领域模型(specific domain model, SDM)。

全局的 GCM 模型和特定领域的 SDM 模型的关系是:两者的训练算法完全一致,并且都是有监督学习模型;SDM 是在 GCM 模型参数的基础上采用迁移学习训练而成;GCM 在多个边缘侧汇聚的大样本量下使用机器学习算法训练,而 SDM 通常利用单个边缘侧的小样本量使用机器学习算法训练而成;SDM 输入特定场景的标注数据集,根据有监督学习算法训练生成分类分析模型。

EML 边缘分布式机器学习算法框架如图 2 所示,该算法框架的主要执行步骤包括 7 步。

(1)将输入数据集通过分布式消息系统推送至边缘侧处理,边缘侧的多个边缘计算节点分布式地对各自分区的输入数据集进行清洗、转化、归一化、空缺值填充、特征选择等数据处理操作,同时计算预测目标字段和目标值分布,形成标注数据集。

(2)结合 IDS 数据集和 LDS 标注数据集,调用数据质量审核接口对 LDS 标注数据集的质量进行评估,评估的维度主要包括:缺失值比例要低于 10%,聚类法剔除异常值,指数平滑法对特征字段中的时序数据进行加权平均提升其预测的平稳程度。同时,对归一化的数据字段进行正态分布验证,离群值剔除操作和标准差检测,审核 LDS 标注数据集,并将评分在 85% 以上的 LDS 数据样本作为最终的标注数据集用于后续的机器学习训练。

(3)将审核后的标注数据集加载到边缘计算节点学习训练,向边缘管理节点注册并记录机器学习训练程序的任务执行状态。边缘侧机器学习算法主要有分类分析算法,卷积神经网络(convolutional neural networks, CNN)算法。

(4)通过边缘侧数据训练,结合云端的 GCM 模型参数作为边缘侧 SDM 模型的初始参数,再利用卷积神经网络中的反向传播(back propagation, BP)算法进行迭代,迭代至损失函数交叉熵值趋近于零收敛,生成边缘 SDM 模型。

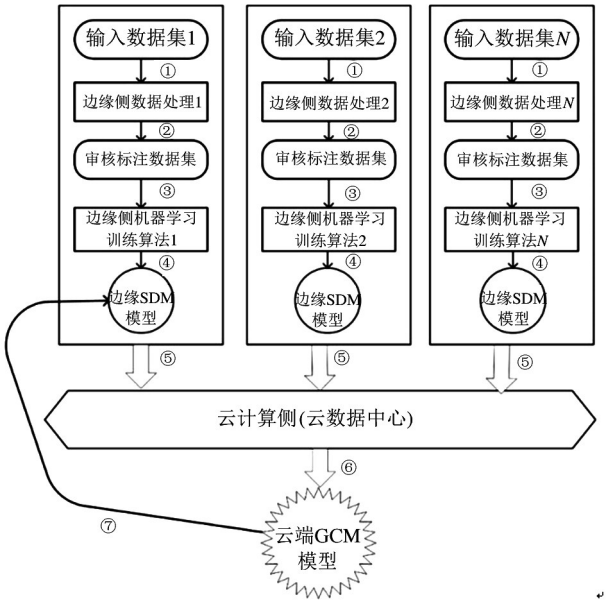

图 2 EML 边缘机器学习算法框架

(5)将多个边缘侧接入的边缘 SDM 模型和标

注数据集回传到云计算侧,数据在多个边缘侧周期性地回传汇聚到云端形成大样本数据集,云计算侧具有较强的运算能力,支持分布式机器学习任务的并行执行,能够对大数据样本进行高效训练。

(6)利用云平台算力加载大数据样本到神经网络深度学习训练程序,经过多轮迭代收敛,生成全局通用的 GCM 模型,GCM 模型是一类深度学习模型,兼容 CNN 模型。

(7)更新参数后的云端侧 GCM 模型实时同步到边缘侧,并且边缘模型的初始参数来源于更新后的 GCM 模型参数,边缘节点启动训练任务,将各个边缘侧收敛后的最新参数作为新的边缘 SDM 模型参数,用于边缘侧推理和预测分析应用。

EML 算法框架经过以上 7 个步骤的执行,从应用侧输入数据集,边缘侧的数据处理、数据标注、机器学习算法执行、边缘 SDM 模型训练,以及云端训练 GCM 模型。通过 GCM 模型做通用的分析预测,并且将特定场景的分析预测任务交由 SDM 模型完成。

3 边缘侧大数据分析预测原型 EDAP 系统

根据所设计边缘侧大数据分析预测建模方法实现了边缘侧大数据分析预测原型系统(edge big data analysis and predicate prototype system, EDAP),EDAP 系统实现了在边缘侧和云端协同训练 DLNN 深度学习算法模型。EDAP 系统实现了云端和边缘侧算法训练与预测程序。

EDAP 原型系统的 DLNN 深度学习训练算法工作原理如算法 1 所示。

算法 1 在 EDAP 系统中训练 DLNN 模型的算法步骤

1. 准备训练样本(S),包括 X 个特征和标签 L, 表示为 $S = \{X, L\}$
2. 输入: $S = \{X, L\}$
3. 在边缘计算节点上面运用 TensorFlow 和 Keras 深度学习框架创建 DLNN 模型,并将 CNN 深度学习模型表示成 $\boldsymbol{model^{cnn}}$
4. 定义损失函数, 损失误差 $\text{loss} = -\frac{1}{m}\sum_{j=1}^{m}\sum_{i=1}^{n} y_{ji}\log(\overline{y_{ji}})$

   调用 TensorFlow 接口方法为 *ross _entropy = TF. reduce _mean(- tf. reduce _sum(y _ * tf. log(y) + (y _ - 1) * tf. log(1 - y)))*
5. 定义循环次数的超参数 *num*,模型训练过程中最多迭代 *num* 次循环
6. 持续运行模型训练程序直到误差 loss 趋近于 10^{-8}, $loss \to \varepsilon$, where $\varepsilon < 10^{-8}$
7. 使用测试数据集验证模型 $model^{cnn}$ 的准确率 P 值、召回率 R 值和 AUC 值, $\langle P \in (0,1) \mid R \in (0,1) \mid AUC \in (0,1)\rangle$
8. 将训练后的模型 $model^{cnn}$, 如果模型评估的 P 值、R 值和 AUC 值均大于 0.5,也就是说 $\langle P \in (0.5,1) \mid R \in (0.5,1) \mid AUC \in (0.5,1)\rangle$, 则将模型文件保存在云计算节点上,在后续预测过程中就可以直接将模型 $model^{cnn}$ 文件直接从云节点下发同步到边缘计算节点,以支持边缘侧分析预测任务

EDAP 原型系统输入训练数据集,传输到算法 1 程序中进行训练执行,生成 DLNN 模型参数文件,训练的超参数由系统设定的组合自动选择一套最优的参数。多个边缘侧的数据集汇聚后的大样本数据集的训练过程在云端完成,模型保存在云端,并且定期将模型下发到边缘侧。在云端模型参数的基础上,边缘侧根据自身采集的数据集基于迁移学习方式进行优化模型训练,该过程在边缘计算节点完成,并在边缘管理节点保存边缘模型文件。

EDAP 原型系统的边缘智能预测通过迁移学习技术来实现,如图 3 所示的 EDAP 原型系统原理架构,首先基于云端服务器集群的基础数据集训练一个基础模型,将重要特征迁移到边缘侧目标模型,并以边缘计算节点上收集的目标标注数据集进行建模训练。在云端预先训练一个大规模通用预训练模型;然后通过迁移学习方式在边缘集群结合本地数据集与边缘计算节点资源进行轻量级的目标模型训练和部署。EDAP 原型系统能够显著降低 DLNN 深度学习模型在网络边缘训练的资源消耗,能够大幅减少网络通信量和计算资源消耗。

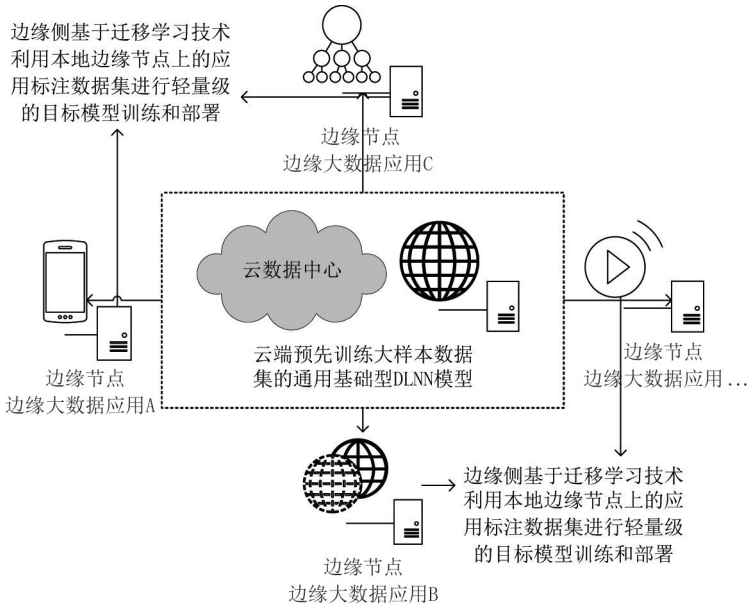

图 3　EDAP 原型系统原理架构

4　实验结果分析

本文设计了 5 个实验来验证边缘大数据分析方法的有效性和性能。实验平台为:(1)云端服务器 20 台,每个节点的配置为 16 核 CPU,32 GB 内存和 2 TB高速硬盘存储空间;(2)边缘节点 30 个,每个节点的物理配置是 4 核低功耗 CPU,8 GB 内存和 200 GB 固态硬盘;(3)软件包括 Linux 系统以及 scikit-learn 和 PyTorch 机器学习与深度学习框架,Python 3 编程语言。

本文实验对比的数据包括:(1) 训练样本为 59 307张图片标注数据集(image datasets),共 10 类标签,存储在图片键值存储系统;(2) 310 297 条结构化金融行为标注数据集(structured financial datasets),共 10 类标签,存储在对象关系数据库中。测试评估样本为 28 963 张标注图片数据集,120 393 条结构化数据集。

4.1　边缘侧大数据训练的性能效率

本实验使用结构化数据集训练随机森林(random forest,RF)算法模型,使用图片数据集训练卷积神经网络 CNN 算法模型。比较的实验环境包括云端(Cloud)和边缘侧协同(Edge-Cloud) 训练随机森林和卷积神经网络模型的性能效率,并且验证了训练学习的性能效率与边云集群节点规模的扩展性。训练性能的实验结果如图 4 所示,RF Cloud 表示云端训练随机森林,RF Edge-Cloud 表示边缘侧协同训练随机森林;CNN Cloud 表示云端训练 CNN 模型,CNN Edge-Cloud 表示边缘侧协同训练 CNN 模型。

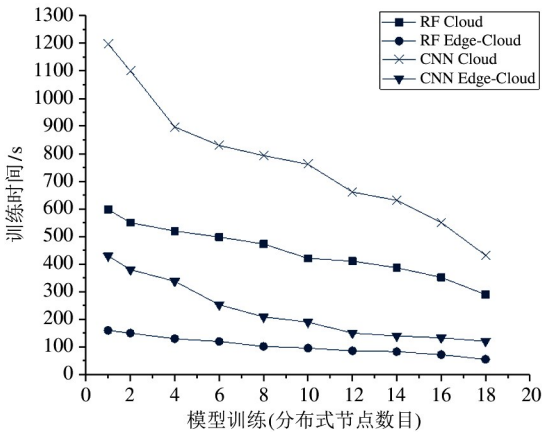

图 4　云端和边缘侧分别训练随机森林和卷积神经网络模型的性能效率

实验中随机森林 RF 算法的参数 n _ estimators 为 300,n _ jobs 设置为 16,有放回的随机抽样特征和样本,迭代执行 10 次取执行时间的平均值。CNN 算法参数设置为:原始图像数据做 3 次卷积,然后执行 1 次 max pooling 操作,循环 20 次;然后全连接层设置为 128 层,batch size 设置值为 64,损失函数是交叉熵。CNN 训练过程表示为:(3 × Conv→1 × Max

pooling) ×20→(128 × Full Connected Network)→Label。从图4性能对比看出,边缘侧RF训练性能比云端RF模型平均提升了3.95倍,边缘侧CNN模型训练效率比云端CNN训练效率提升了3.87倍。同时还验证了边缘侧分析建模方法的可扩展性,当云端与边缘训练节点数由2个扩展到18个,可见随着节点数的增加实验训练时间下降明显,RF Edge-Cloud实验的执行时间由159.3 s下降到53.6 s;CNN Edge-Cloud实验的执行时间由433.7 s下降到120.9 s,因此可证明在边缘侧训练大数据分析的随机森林模型和CNN模型具有良好的扩展性能。

4.2 边缘侧大数据分析模型的精度评估结果

本文运用测试样本数据集对边缘侧RF和CNN两类算法模型进行实验评估,评估指标均使用准确率(Precision, P),召回率(Recall, R)和F1值$\left(\frac{2 \times P \times R}{P + R}\right)$,RF和CNN模型文件生成后调用预测接口加载测试样本,将每个样本的预测值和实际值进行比较,统计得出评估指标值。如图5所示实验结果如下。

(1) RF Cloud模型的P值、R值和F1值分别为91.83%、93.27%、92.54%;

(2) RF Edge-Cloud模型的P值、R值和F1值分别为97.35%、98.32%、97.83%;

(3) CNN Cloud模型的P值、R值和F1值分别为93.78%、95.27%、94.52%;

(4) CNN Edge-Cloud模型的P值、R值和F1值分别为96.58%、96.37%、96.47%。

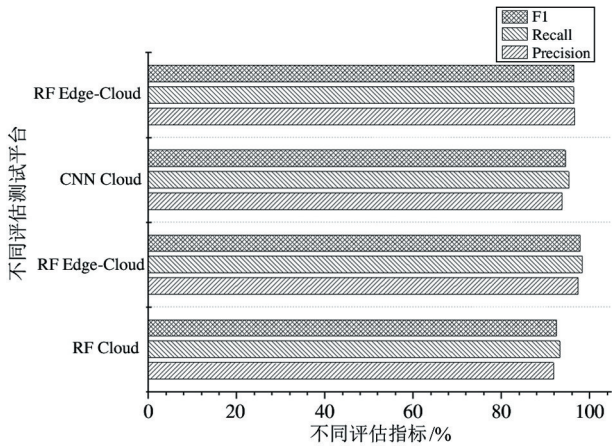

图5 边缘侧和云端训练的模型测试评估结果

从图5展示的模型评估结果可以看出,边云协同训练的算法模型,RF模型和CNN模型其准确率、召回率和F1值均比纯云端训练模型的预测效果要更好,模型的准确度提升了3% ~9%,在特定场景下训练出来的模型AUC评估指标值由72.6%提升到了99.6%。边云协同训练模型的准确度均在96%以上,而且边缘侧模型的泛化能力很好,验证数据集上泛化误差均小于2%。

4.3 不同方法训练模型任务的网络通信与I/O读写性能测试结果

本文设计实验对比纯云端训练模型和边云协同训练模型任务的网络通信与I/O读写量,如图6所示,边云协同训练(Edge Cloud)方法比纯云端(Cloud Only)训练方法的网络通信量与I/O读写量总和平均减少了93.7%。在大数据驱动的机器学习算法训练任务中,网络传输与I/O读写的时间占整个训练时间的比重通常比较大,因此边云协同训练方法优化了网络与I/O读写性能,训练任务执行时间降低了3.78倍,大幅提升了训练性能和效率。

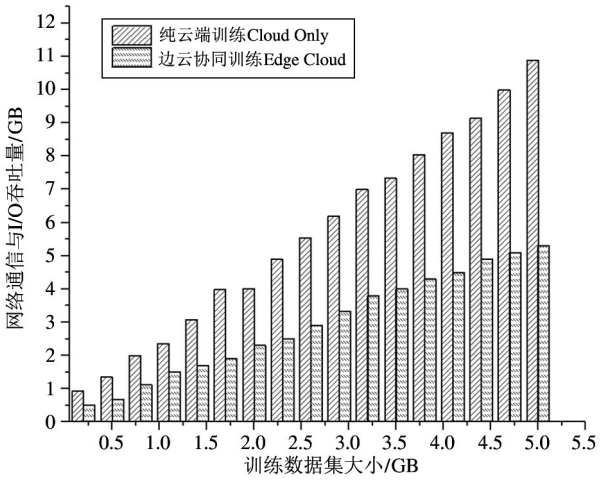

图6 不同训练方法的网络通信量与I/O读写量性能比较

4.4 模型预测任务在不同环境并发访问量下的响应性能结果

本文基于纯云端环境和边云协同环境设计了不同并发访问量场景下的模型预测任务,实验结果如图7所示。在10 ~160个并发用户请求模型预测场景下,边云协同方法比纯云端方法的并发访问响应时间降低了5.78 ~7.35倍,边云协同预测方法的平

均请求响应时间均在 10 ms 以内,因此对移动应用端使用模型预测具有很好的实用性。

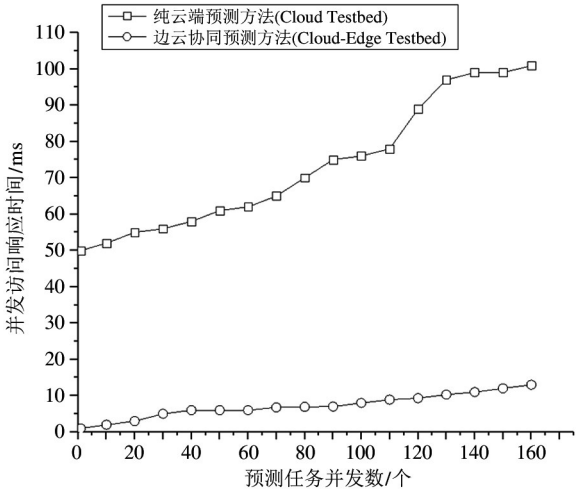

图 7 模型预测任务在不同环境并发访问量下请求响应性能比对

5 结 论

综上所述,边云协同大数据分析建模方法能够有效提升海量数据集的训练效率、预测性能和预测准确率。在大数据应用中海量异构多源的结构化数据、文本数据和图片数据集分布在各个数据中心,可以通过边云协同建模方法进行统一建模,形成逻辑上集中大数据分析平台和统一的分析预测模型调用服务平台。随着边缘计算、云计算和大数据分析的进一步融合,边缘智能和云端智能将是支撑泛在智能分析应用的重要基石。未来边云智能协同架构、算力调度、高速互联、边云 AI 模型自动设计、分布式 AI 模型机制等也将是重要的研究方向。

参考文献

[1] SUN Z, WANG P P. Big data, analytics and intelligence: an editorial perspective[J]. *Journal of New Mathematics and Natural Computation*, 2017,13(2):75-81

[2] POLYZOTIS N, ROY S, WHANG S E, ZINKEVICH M. Data management challenges in production machine learning[C]// Proceedings of the 2017 ACM International Conference on Management of Data, New York, USA, 2017:1723-1726

[3] SUZUKI L. Data as Infrastructure for Smart Cities[D]. London: University College London, 2016

[4] CHOWDHURY R R, ADNAN M A, GUPTA R K. Real-time principal component analysis[J]. *ACM/IMS Transactions on Data Science*,2020, 1(2):1-36

[5] 韩淑君, 李俊, 董谦. 车联网中基于服务的虚拟网络功能放置算法[J]. 高技术通讯, 2021, 31(4):341-349

[6] 于彤彤, 董婷婷, 肖创柏. 基于深度强化学习的舰载机在线调度方法研究[J]. 高技术通讯, 2021, 31(4):367-377

[7] 黄晓辉, 崔莉, 黄希. 基于规则实时性的端云动态分配方法研究[J]. 高技术通讯, 2021, 31(3): 223-231

[8] 杨乐, 李萌, 叶欣宇, 等. 融合边缘计算与区块链的工业互联网资源优化配置研究[J]. 高技术通讯, 2020, 30(12): 1253-1263

[9] 黄晓辉, 崔莉, 黄希. 基于监督学习的规则触发执行预测方法研究[J]. 高技术通讯, 2021, 31(2): 113-121

[10] 王璐, 张健浩, 王廷, 等. 面向云网融合的细粒度多接入边缘计算架构[J]. 计算机研究与发展, 2021, 58(6): 1275-1290

[11] 李凡长, 刘洋, 吴鹏. 元学习研究综述[J]. 计算机学报, 2021,44(2): 422-446

[12] 王鑫, 陈蔚雪, 杨雅君, 等. 知识图谱划分算法研究综述[J]. 计算机学报, 2021, 44(1): 235-260

[13] 任杰, 高岭, 于佳龙. 面向边缘设备的高能效深度学习任务调度策略[J]. 计算机学报, 2020, 43(3): 440-452

[14] 滕飞, 黄齐川, 李天瑞. 大规模时间序列分析框架的研究与实现[J]. 计算机学报, 2020, 43(7): 1279-1292

[15] 孟子尧, 谷雪, 梁艳春. 深度神经架构搜索综述[J]. 计算机研究与发展, 2021, 58(1):22-33

[16] 朱泓睿, 元国军. 分布式深度学习训练网络综述[J]. 计算机研究与发展, 2021, 58(1):98-115

[17] WANG F, ZHANG M, WANG X, et al. Deep learning for edge computing applications: a state-of-the-art survey [J]. *IEEE Access*, 2020, 8: 58322-58336

[18] XIA X, CHEN F, HE Q, et al. Cost-effective app data distribution in edge computing[J]. *IEEE Transactions on Parallel and Distributed Systems*, 2021, 32(1):31-44

Research on edge big data analysis and predictive modeling method

ZHONG Yunqin[*,**], ZHU Yueqin[***,****], JIAO Shoutao[***,****]

([*]Key Laboratory of Big Data Mining and Knowledge Management, Chinese Academy of Sciences, University of Chinese Academy of Sciences, Beijing 100190)

([**]Information Center, Development Research Center of the State Council, Beijing 100010)

([***]Development Research Center, China Geological Survey, Beijing 100037)

([****]Geological Information Engineering Technology Innovation Center, Ministry of Natural Resources, Beijing 100037)

Abstract

With the improvement of real-time requirements for big data analysis of the Internet of Things, the cloud big data analysis method controlled by the center cannot meet the real-time and accuracy requirements due to its large response delay, high cost, and low prediction accuracy in specific environments. This paper proposes an edge-side big data analysis and predictive modeling method under massive real-time data such as sensor data, streaming data and other scenarios. This method trains small data samples on the edge side, multi-accesses the edge side for distributed distribution according to specific application scenarios and conducts model learning, models training and inference predictive analysis. Firstly, by combining big data analysis and edge computing, a theoretical paradigm framework for big data analysis and prediction modeling on the edge side and cloud collaboration is proposed. Secondly, the edge side big data analysis and prediction training algorithm and tuning mechanism are designed. Finally, the prototype of the training and evaluation system for edge-side big data analysis is realized. Experimental results in a test environment with hundreds of nodes show that in real-time big data scenarios, compared with cloud training, the performance and efficiency of the edge-side big data training proposed in this paper is increased by an average of 3.95 times, and the network traffic is reduced by 88.7%. The prediction accuracy, recall rate and F1 value of the collaborative training model can be improved by 3% -9% compared with the traditional training method, and the response delay of request prediction is reduced by 67.5%. The method in this paper can be effectively applied to scientific computing, smart finance, autonomous driving, security monitoring, data security, smart factories, smart cities and other fields.

Key words: edge computing, big data analysis, edge big data, edge machine learning, edge-cloud collaboration