文章编号:1672-3961(2024)04-0086-09 DOI:10.6040/j.issn.1672-3961.0.2023.287

基于自适应线性模型的环境数据预测算法

王凤娟[1],王语睿[2],卫兰[3,4*],范存群[3,4],徐晓斌[2]
(1. 山东省东明县气象局综合气象业务科, 山东 菏泽 274500; 2. 北京工业大学计算机学院, 北京 100124; 3. 中国气象局中国遥感卫星辐射测量和定标重点开放实验室/国家卫星气象中心(国家空间天气监测预警中心), 北京 100081; 4. 许健民气象卫星创新中心, 北京 100081)

摘要:针对环境大数据在智慧城市应用中的实时性和准确性问题,提出一种基于自适应线性模型的环境数据预测算法。根据气象数据的实时变化情况对模型进行训练,自适应调整训练窗口大小,并在训练态与预测态之间动态实时切换,使模型具有较强的适应环境的能力。该算法具有较低的时延和较小的计算开销,可以在传感器节点上直接部署,满足数据预测的实时性需求。在真实环境数据集的基础上构建仿真试验,相比固定窗口模型,该算法数据预测误差降低 17.4%以上,环境数据采集能耗降低 80%以上,平均时延降低超过 50%;相比已有的机器学习算法,训练及预测时间降低 37%以上。

关键词:智慧城市;环境大数据;边缘服务;线性预测;节能减排
中图分类号:TP3 **文献标志码**:A
引用格式:王凤娟,王语睿,卫兰,等. 基于自适应线性模型的环境数据预测算法[J]. 山东大学学报(工学版),2024,54(4):86-94.
WANG Fengjuan, WANG Yurui, WEI Lan, et al. Environmental data prediction algorithm based on adaptive linear model[J]. Journal of Shandong University (Engineering Science), 2024, 54(4):86-94.

Environmental data prediction algorithm based on adaptive linear model

WANG Fengjuan[1], WANG Yurui[2], WEI Lan[3,4*], FAN Cunqun[3,4], XU Xiaobin[2]
(1. Comprehensive Meteorological Operation Section, Dongming County Meteorological Bureau of Shandong Province, Heze 274500, Shandong, China; 2. College of Computer Science, Beijing University of Technology, Beijing 100124, China; 3. Key Laboratory of Radiometric Calibration and Validation for Environmental Satellites, National Satellite Meteorological Center (National Center for Space Weather), China Meteorological Administration, Beijing 100081, China; 4. Innovation Center for FengYun Meteorological Satellite (FYSIC), Beijing 100081, China)

Abstract: To address the issues of real-time performance and accuracy in the application of environmental big data in smart cities, an environmental data prediction algorithm based on an adaptive linear model was proposed. The model was trained according to the real-time changes in meteorological data, with the training window size being adaptively adjusted. A dynamic and real-time switch between training and prediction states was implemented, enhancing the model's adaptability to environmental changes. The algorithm featured lower latency and reduced computational overhead, allowing for direct deployment on sensor nodes to meet the real-time requirements of data prediction. Simulation experiments constructed on real environmental datasets showed that, compared to fixed-window models, the proposed algorithm reduced data prediction error by more than 17.4%, decreased the energy consumption of environmental data collection by over 80%, and reduced the average latency by more than 50%. When compared to existing machine learning algorithms, the training and prediction time of the proposed algorithm was reduced by more than 37%.

Keywords: smart city; environmental big data; edge service; linear prediction; energy saving

0 引言

环境大数据可以为智慧城市提供强有力的数据支撑[1]。然而,频繁环境数据采集往往导致数据采集设备能耗较高,城市碳排放量急剧增大。对于环境采集设备,数据传输能耗通常远大于数据处理能耗,因此如何降低数据传输次数成为智

收稿日期:2023-12-09
基金项目:国家重点研发计划资助项目 (2021YFB3901000, 2021YFB3901005);风云星应用先行计划资助项目(FY-APP-2021.0501)
第一作者简介:王凤娟(1973—),女,山东东明人,工程师,主要研究方向为气象数据挖掘及气象数据处理。E-mail: dmxwangfengjuan@163.com
*通信作者简介:卫兰(1980—),女,江苏姜堰人,高级工程师,主要研究方向为气象数据挖掘及卫星数据处理。E-mail: weilan@cma.cn

慧城市场景中的重要挑战。在大多数实际智慧城市应用场景中,用户及应用并不需要完全精确的环境数据,通常对环境数据的误差具有一定程度的容忍。训练环境数据得到预测模型,并对环境数据进行预测,在一定误差范围内以预测数据代替实际采集数据,仅上传预测准确度不足的数据,能够有效降低数据传输次数,节省数据采集设备能耗[2],对国家“双碳”政策的实施起到重要的促进作用。

相比数据的精确度,当前多种环境数据预测算法更加关注数据误差的可控性,即用户可以任意确定可接受的误差范围,数据预测算法可以在此误差范围内训练模型,决定数据的收集策略[3]。在同样的误差范围内,数据模型精确度越高,数据采集频率越低。已有研究表明,大部分环境数据(例如温度、湿度、气压、光照强度等)的变化具有连续、缓慢的特点,线性模型在处理数值型环境数据时具有开销小、时延低等特点[4]。基于机器学习的预测算法能够在特定的场景与数据模式中提供较为精确的数据预测[5-13],然而复杂的机器学习模型通常需要基于海量数据进行模型训练,难以在环境采集设备中直接部署;在云端基于历史数据进行模型训练通常需要较长时间,难以满足数据采集的实时性需求。当前已有多种时序预测模型及时序数据预测算法能够对环境数据进行预测[14-19],然而已有时序预测算法往往存在模型灵活性不足,难以适应环境实时变化的问题,预测精度有待提升。

为解决以上问题,本研究将机器学习与线性模型的特点融合,提出一种自适应的线性模型,用于对温度、湿度、气压、光照强度等具有线性特性的数值型环境数据进行预测。该模型将训练与预测过程合二为一,分为训练态与预测态 2 个状态:当模型处于训练态时,采取实时增量式训练策略,满足训练精度要求或时延要求时进入预测态;当模型处于预测态时,如果预测精度不符合预设要求,则重新返回训练态。模型通过 2 种状态的动态实时调整,增强模型的环境适应性,降低开销,提升预测准确度。

1　相关工作

为了实现绿色低碳发展,减少数据采集频率成为智慧城市的重要需求。时序数据预测能够在较低的数据采集频率基础上获得较丰富的环境数据,成为智慧城市数据收集中的关键技术。

1.1　基于机器学习的预测算法

近年来,基于机器学习的预测算法在时序预测中发挥着重要作用。对于单个物联网设备,文献[4]提出一种基于多变量时间序列预测的物联网网络自适应数据传输周期控制(prediction based adaptive data transmission period control, PBATPC)算法和基于时间序列距离测度的多元时间序列数据编码方案,减少物联网传感器不必要的数据传输。针对海冰运动的预测,文献[5]提出一种基于深度学习算法,由具有卷积长短期记忆(long-term and short-term memory, LSTM)单元的编码器-解码器网络组成,该网络能够学习运动时间序列内的长期依赖关系,其卷积结构有效捕获了相邻运动向量之间的空间相关性。针对长期时间序列预测问题,文献[6]提出一种基于反向传播神经网络和信息粒度的长期预测算法,利用信息粒度中的合理粒度原则,得到时间序列的各个数值区间,开发一种自动线性趋势提取算法提取趋势变化,以此构建一个以信息粒度为输入的神经网络层次结构模型。文献[7]提出一种基于趋势聚类驱动的物联网系统联邦学习算法,利用聚类站点的时间序列数据训练出一个 LSTM 网络,对聚类站点的能量需求进行预测,大幅度降低能源需求预测误差。在智能交通领域,文献[8]提出利用图的概念,将机场建模为具有时间序列特征的节点,对图结构数据进行数据挖掘,以应对交通扰动在机场之间的传播问题;文献[9]提出一种新的用于交通流预测的时空切比雪夫图神经网络模型,采用 LSTM 模型学习交通状态变化,获取时间依赖性,与空间特征结合实现交通流预测;文献[10]提出深度学习算法 DeepTrack,使用时态卷积网络提供更可靠的时间预测算法,使计算量更小,更适合实际的嵌入式物联网设备,基于这一算法对实时车辆轨迹开展预测。在智慧城市领域,针对社会突发事件和事故引起的瞬时城市人流量变化,文献[11]提出一个基于线上到线下交互的膨胀因果卷积神经网络框架预测城市人群流。针对现有多变量时间序列预测算法不能保障局部变量预测精度的局限性问题,文献[12]提出一种基于自演化预训练的多变量时间序列预测算法,构建和训练单变量时间序列模型,通过拓展时序卷积网络和 LSTM 单元建模变量间复杂的时序依赖关系,在模型融合再训练过程中,该算法比现有算法获得相对较高的预测精度,在保障局部变量预测精度上具有更好的性能。文献[13]对图神经算子进行改进,提出傅里叶神经算子,在一定程度上摆脱网格分辨率限制,提高数据预测模型的训练效率。基于机器

学习的预测算法具有较高的预测准确度,但往往采用复杂的模型结构,具有大量的参数和计算层级,导致模型训练和实时计算的复杂度增加。

在实际的智慧城市场景中,环境感知节点需要进行密集部署,每 km^2 可能需要部署数十万感知设备。成本的限制使这些感知设备通常仅具备基础的数据处理能力,其硬件能力难以部署复杂的机器学习模型。复杂的机器学习模型往往需要基于海量数据进行训练,模型的训练也具有较高开销。因此,在实际的智慧城市应用中,复杂的机器学习模型通常部署在云端,基于历史数据进行训练,往往难以实时更新。当前机器学习模型尽管能够较好地表示环境变化情况,具有较高的预测准确度,但难以满足环境数据采集的实时性需求。

1.2 基于轻量级线性模型的预测算法

由于环境数据在一定时间范围内通常呈现线性特点,数据预测算法的另一方向是基于少量环境数据建立轻量级、实时的线性模型,通过模型的频繁更新保障预测的准确度[14]。这类算法往往结构简单、计算量少、复杂度相对较低,可以在环境采集设备中部署并实时运行。近年来,许多研究者在这些经典算法上进行创新,应用于不同领域中。针对异构物联网应用于网络边缘通信场景,文献[15]提出一个时间转换器模型和一个统一的系统,预测几个服务质量(quality of service, QoS)指标,该算法具有较高的 QoS 预测准确性及鲁棒性。针对交通运行及减排管控措施正朝动态、精细化方向发展的现况,文献[16]应用局部加权线性回归预测公交车路段平均速度和运行模式分布,建立排放因子、耗电因子预测模型,贴合当下智慧交通的特点。针对北京市用电量短期预测,文献[17]提出用时间序列分析算法构建模型,可实现对北京市未来用电量的预测,为政府的短期电力政策提供参考。针对交通客流量分析,文献[18]提出基于季节分类模型的轨道交通客流预测算法,建立季节分类模板和季节时间序列,采用乘法季节自回归差分滑动平均模型构建,能有效预测轨道交通客流,较好地避免预测误差波动性问题。对于具有强非线性的快动态批次过程,文献[19]提出一种高效迭代学习预测函数控制策略,将原非线性系统沿参考轨迹线性化,得到二维跟踪误差预测模型,加强优化计算效率,有效降低计算负担。当前基于线性模型的预测算法具有训练简单、实时性强等特点,但模型往往仅在较短时间内有效,对环境变化的适应能力有所不足,特别是在环境变化较为频繁时,往往存在预测精度不足等问题。频繁的模型训练导致开销增大,如何使模型具有更好的环境适应性,灵活调整训练与预测过程成为提高线性预测算法精确度的关键。参考已有机器学习算法的特性,设计自适应的线性预测算法,对训练和预测过程进行实时动态调整,成为本研究所解决的关键问题。

2 问题模型

2.1 线性数据预测模型

线性数据预测模型是最经典的时序数据预测算法之一,常用于对变化缓慢、具有线性特征的环境数据进行建模。该算法通常在感知设备中临时存储短期内采集到的环境数据,并基于这些数据进行训练,得到线性模型。感知设备无需上传原始数据,仅上传训练后的模型。云中心接收到模型后,与感知设备同步进行数据预测,当预测误差小于用户可接受的阈值时,则认为预测成功,使用预测值代替实际环境数值;当预测误差大于该阈值时,则认为预测失败,对模型进行实时更新,并上传新的模型。

假设环境数据 $e=f(t)$,环境数据的线性模型可以表示为一个分段的线性函数:

$$\hat{e}=\hat{f}(t)=\begin{cases}a_1t+b_1, & t_1\leqslant t<t_2\\ a_2t+b_2, & t_2\leqslant t<t_3\\ \cdots\\ a_{n-1}t+b_{n-1}, & t_{n-1}\leqslant t_{n-1}<t_n\end{cases}, \tag{1}$$

式中:t_j 为预测模型的更新时间,$\forall j\in[1,n-1]$,j 为正整数;a_j、b_j 为模型参数。t_j 与 t_{j+1} 之间的时间段就是一组模型参数 a_j、b_j 的有效时间。

线性数据预测模型以一定误差为代价,降低数据采集频率,节省能耗。模型的训练过程是影响预测模型性能的关键。

2.2 误差及时延限制下的线性预测问题

线性预测模型通常具有一定误差,模型的训练及更新过程也将带来额外时延。在实际应用场景中,用户通常对数据精确度及数据收集时延具有一定要求,为此本研究对误差及时延限制下的线性预测问题进行建模。假设用户可容忍的预测模型误差为 ε,可容忍的时延上限为 T,则线性预测模型的建立可表示为求解式(1),使得:

$$\frac{\int_{t_{\text{start}}}^{t_{\text{end}}}|\hat{f}(t)-f(t)|}{t_{\text{end}}-t_{\text{start}}}\leqslant\varepsilon, \tag{2}$$

$$\max(t_{n+1}-t_n)\leqslant T, \tag{3}$$

式中,t_{start}为模型生效时间,t_{end}为模型失效时间,t_{n+1}和t_n分别为第n+1 次和第n次预测模型更新的时间。式(2)表明t_{start}~t_{end}期间,模型预测值与实际值之间的平均误差小于或等于ε;式(3)表示任意 2 次连续模型更新的最大时间间隔不应超过T。

3　环境数据预测算法

3.1　自适应线性模型框架

为求解误差及时延限制下的线性预测问题,本研究参考已有机器学习算法及传统线性预测模型算法的特性,提出一种自适应线性模型。该模型设定训练态与预测态 2 种状态,根据特定条件,模型在训练态与预测态之间动态切换。

(1)当前无可用线性方程时,模型处于训练态。当自适应线性模型处于训练态时,模型维护一个自增长训练窗口,存储近期实时数据,并基于这些数据进行线性模型训练。当模型误差超过指定上限ε或窗口大小达到指定上限n_b时,存储此时的线性方程,并切换至预测态。

(2)存在可用的线性方程时,模型处于预测态。当自适应线性模型处于预测态时,模型存储实时采集的环境数据并将之与存储的线性方程所生成的数据进行比对,若误差超过指定上限ε,删除此时的线性方程,并切换至训练态。

由以上框架可知,当模型处于训练态时,训练误差不超过ε,训练时延不超过$n_b\Delta t$,其中Δt为数据采集的时间间隔;当模型处于预测态时,预测误差不超过ε,预测时延不超过Δt。显然,通过设定ε、n_b、Δt,即可得到式(1)的一组可行解。

3.2　模型训练算法

在任意一组参数有效期$[t_i,t_{i+1})$内,预测模型需要先对模型参数进行训练,再基于训练结果开始预测。在训练阶段,预测模型通过均匀采样获得n次原始环境数据,基于这些数据开始训练,得到预测模型。

在线性预测模型中,通常采取最小二乘法获得最小残差平方和,计算过程如下。

用$e_i(i\in\{t_j,t_{j+1},\cdots,t_{j+m-1}\})$表示任意一个训练阶段的原始环境数据,其中$m$为训练所用的数据个数。采取最小二乘法得到的参数结果为:

$$\begin{cases} a=\dfrac{m(\sum ie_i)-(\sum i)(\sum e_i)}{m(\sum i^2)-(\sum i)^2} \\ b=\dfrac{(\sum i^2)(\sum e_i)-(\sum i)(\sum ie_i)}{m(\sum i^2)-(\sum i)^2} \end{cases}。\tag{4}$$

当基于 2 个原始环境数据训练线性函数时,上述算法训练生成的线性模型是精确的。当原始环境数据数量增多时,误差开始增大。自适应的线性模型训练算法采取实时逐步尝试策略,维护一个自增长的训练窗口,由 2 个原始环境数据开始建立线性模型。每次采集得到新的环境数据后,训练窗口加 1,将此时的环境数据存入训练窗口,进行线性模型训练。如果训练后误差小于上限,则继续训练;反之,则将上次训练的模型进行输出。

由于模型的训练过程需要等待数据的采集,此时将带来额外的时延。为了满足用户所设定的时延要求,可以通过设定训练窗口的上限控制时延上限。若用户可接受的时延上限为T,数据采集的频率为f,则当训练窗口的上限设定为$n_b=Tf+1$时,训练时延上限即可满足用户要求。

通过训练误差上限及训练窗口上限,自适应的线性模型可保证当处于训练态时,误差及时延满足式(2)(3)。自适应的线性模型训练算法详细步骤见算法 1。

算法 1　自适应的线性模型训练算法

输入:ε、n_b。

输出:预测模型(a,b)。

(1) 初始化$n=1$,$Q_b=2\varepsilon^2$;

(2)　WHILE 新的数据e_i产生 DO;

(3)　n++;

(4)　使用式(4)计算a'、b';

(5) $Q=\sum_{i=1}^{n}(e_i-a'_i-b')^2$;

(6)　IF $Q>Q_b$　RETURN(a,b);

(7)　ELSE $a=a'$, $b=b'$, $Q_b+=\varepsilon^2$;

(8)　IF $n==n_b$ RETURN (a,b);

(9)　END WHILE。

在这一算法中,Q_b为偏差平方和上限,Q为当前模型的偏差平方和,a'、b'分别为 2 个临时变量。当感知节点收集到新的数据时,计算开销为$O(n)$。由于n最小为 2,最大为n_b,在最差情况下,计算开销为$O(n_b)$,最好情况下,仅需一步运算即可得出结果。

3.3　误差及时延限制下的环境数据采集

在实际的智慧城市数据采集应用场景中,可以通过如下步骤,获得符合误差及时延限制条件的环境数据。

步骤 1:用户根据应用需求,为不同类型的环境采集设备设定误差上限及训练窗口上限,这些参数将通过智慧城市中心控制节点发送至指定设备。

步骤 2:环境采集设备初始状态为训练态,根据

算法 1 建立预测模型,并把模型发送至智慧城市中心控制节点。

步骤 3:智慧城市中心控制节点进入预测态,与环境采集设备同时使用预测模型得到感知数据的近似值。当采集到新的环境数据时,将实际环境数据与近似值进行比较,若其绝对误差大于误差上限,则判断此时模型不再符合环境变化情况,环境采集设备进入训练态,并退回步骤 2,重新进行模型训练。

步骤 4:当智慧城市中心控制节点收到一个预测模型时,根据这个模型重构该节点的数据,并退回步骤 3。

智慧城市中心控制节点所执行的自适应周期环境数据预测算法用于根据收到的预测模型构建数据。详见算法 2。

算法 2 自适应环境数据预测算法

输入:(t,a,b)、Δt。

输出:$\{e_t,e_{t+1},\cdots\}$。

(1) WHILE 收到新的预测模型 DO;
(2) $e_t=at+b$;
(3) FOR($i=t$; $i\leqslant t_c c$; i++)
(4) e_i+=a;
(5) SLEEP(Δt);
(6) END FOR;
(7) END WHILE;
(8) WHILE 未收到新的预测模型;
(9) e_{i+1}+=a;
(10) SLEEP(Δt);
(11) END WHILE。

在这一算法中,t_c 为当前的时间序号,e_t 为预测模型在时间 t 处的预测误差。通过以上步骤得到的预测数据,满足式(2)(3)中的误差约束及时延约束,因此,采取自适应的环境数据预测算法可以得到式(1)的一组可行解。

算法 2 开销与算法 1 类似。每当中心控制节点收到新的数据时,仅需一步运算即可得出结果。由于 n 最小为 2,最大为 n_b,在最差情况下,计算开销为 $O(n_b)$,最好情况下,计算开销为 $O(1)$。由于中心控制节点在服务器中部署,其计算能力和电能均较为充裕,算法 2 易于实现,具有较低的额外开销,能够满足实时性需求。

4 仿真试验

为了验证本研究算法在数据预测准确度、时延、能耗等方面的性能,在真实数据集基础上构建仿真试验。基于固定窗口进行模型训练是当前线性预测算法中经典的常用算法,具有较强的代表性。本研究算法为线性预测算法增加了窗口大小及训练状态的灵活性,为此选择基于固定窗口线性模型的预测算法进行对比。

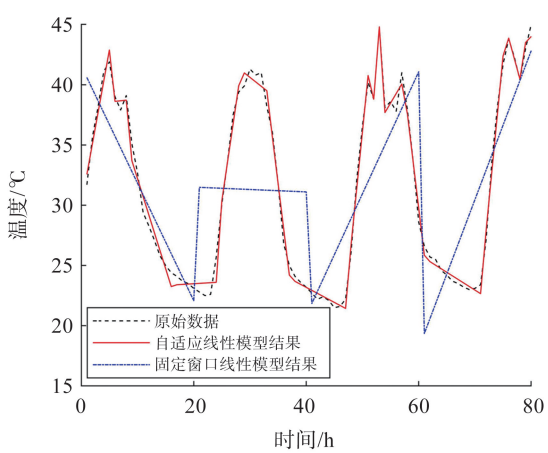
4.1 仿真设置

本研究在 MATLAB 2019b 中建立仿真试验。为了更加全面评估本研究算法,基于中国气象局真实大气探测数据构建仿真试验。在全国范围内随机选择 10 000 个站点,截取 2021-08-01—2021-08-31 的真实气象数据。在该真实数据集中,气象监测设备每小时采集一次环境数据,数据类型包括空气温度、地面温度、风速、风向等。基于真实数据集采用本研究所提的自适应线性模型进行预测,并与现有算法进行对比。

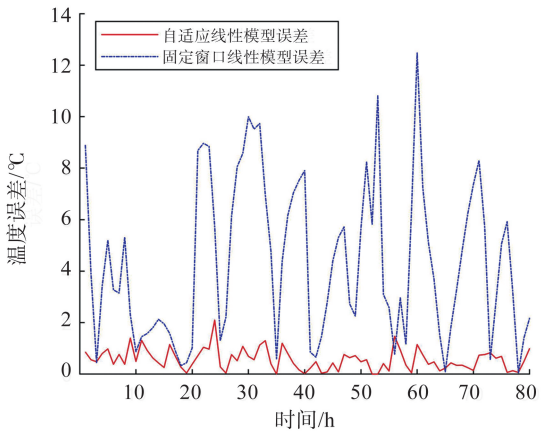
由于传统的线性预测模型通常具有轻量化的特点,可以在感知节点中部署,而机器学习算法通常需要基于海量数据进行模型训练及预测,通常在云端部署。因此,本研究基于 54 908 号站点真实数据模拟单个感知节点中自适应线性模型训练及预测过程,并与固定窗口的线性模型进行对比,主要比较数据预测误差、感知节点的节能情况、实时训练的时延等指标。基于 10 000 个站点的真实数据采取自适应线性模型及常见的几种机器学习模型分别进行模型训练,比较数据训练的时间开销。

4.2 误差对比

试验用于评估数据预测的准确度。采用本研究提出的自适应线性模型与预测窗口固定为 20 的固定窗口线性模型进行对比。自适应线性模型参数设定为:最大误差为 1 ℃,最大训练窗口为 20。固定窗口线性模型始终使用 20 个数据建立线性模型。为了更直观地展示数据预测结果,试验选取54 908号站点前 80 个数据进行对比,如图 1、2 所示。



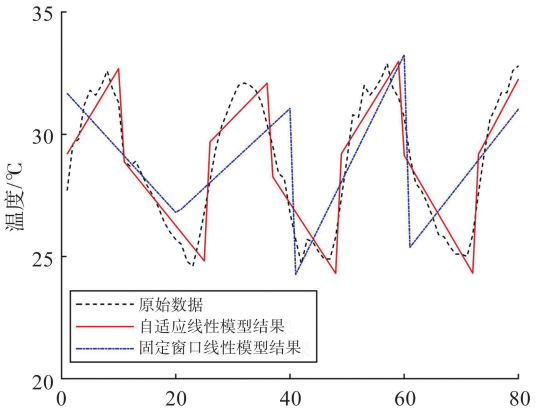
(a) 预测空气温度对比



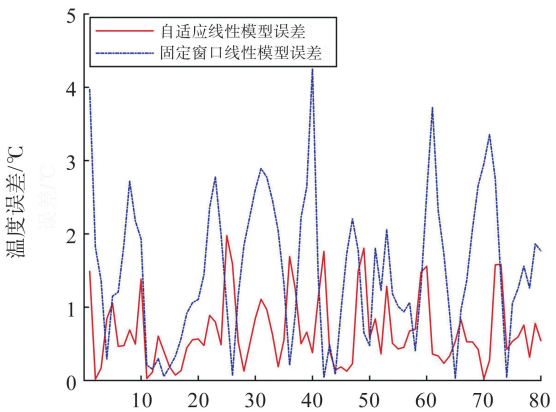
(b) 空气温度预测误差对比

图 1　在空气温度下 2 种模型误差对比

Fig.1　Error comparison between the two models at air temperature



(a) 预测地面温度对比



(b) 地面温度预测误差对比

图 2　在地面温度下 2 种模型误差对比

Fig.2　Error comparison between the two models at ground temperature

如图 1(a)、2(a)所示,本研究提出的自适应周期预测算法温度曲线与真实数据曲线非常接近,仅在部分变化较频繁的数据中失去了部分细节,整体符合数据的变化情况。由图 1(a)可以看出,采取固定窗口预测算法时,由于模型固化,不适应环境较为剧烈的变化,图形失真较为严重。为了进一步量化对比预测误差,本研究对 2 种算法的绝对误差进行计算,结果如表 1 所示。本研究算法最大误差仍在2 ℃范围以内,总体误差较小;对比算法预测误差大部分高于本研究算法,空气温度最大误差超过 10 ℃,地面温度最大误差超过 4 ℃,这一误差在实际环境数据采集中是不可接受的。以上结果表明,本研究算法能够较好地适应温度变化情况。

表 1　预测误差对比

Table 1　Comparison ofprediction errors

单位:℃

训练窗口大小	空气温度自适应模型误差	空气温度固定窗口模型误差	地面温度自适应模型误差	地面温度固定窗口模型误差
10	0.527 5	1.643 3	0.497 0	0.581 7
12	0.557 9	1.972 9	0.565 3	0.729 5
16	0.618 2	2.590 2	0.620 3	1.087 4
20	0.610 1	3.500 9	0.672 7	1.419 9
24	0.606 5	2.784 1	0.671 9	1.266 5

为了进一步对 2 种算法的误差进行对比,本研究构建更多参数设定下的对比试验。由于线性预测算法训练窗口决定了训练所用的最大数据个数及训练频率,如果基于 10 个以下数据点进行训练,训练过于频繁,降低能耗的效果不再明显,采取预测算法与不采取预测算法效果接近。因此,本研究将自适应线性模型训练窗口最大值分别设置为 10、12、16、20、24。相应地,固定窗口模型也设定为相同的数值。对不同预测算法的平均绝对误差进行对比,结果如表 1 所示。

由表 1 可知,本研究算法平均绝对误差相对比较稳定,对参数设定依赖较小,误差始终小于固定窗口模型误差。特别是在空气温度数据集中,本研究算法最小误差能达到固定窗口误差的 17.4%。这一结果较为全面地展示了自适应模型具有较好的环境适应能力,能够达到较高的预测准确度。随着训练窗口大小降低,基于固定窗口的预测算法失真情况也会降低,其预测误差将逐渐接近本研究算法。

4.3　能耗分析

在典型的感知设备中,处理一条指令的能耗通常是传输 1 bit 数据所需能耗的 1/1 000,在轻量级的线性预测算法中,模型训练及预测的计算能耗通常可以忽略不计。感知节点节省能耗的比例通常用数据传输的降低比例进行估算。本研究将对所提算法在降低能耗方面的效果进行评估。基于不同的最大窗口值对 2021 年 8 月真实数据集进行数据预测,统计预测后数据采集次数降低的比例,得到节能效果比例,结果如图 3 所示。

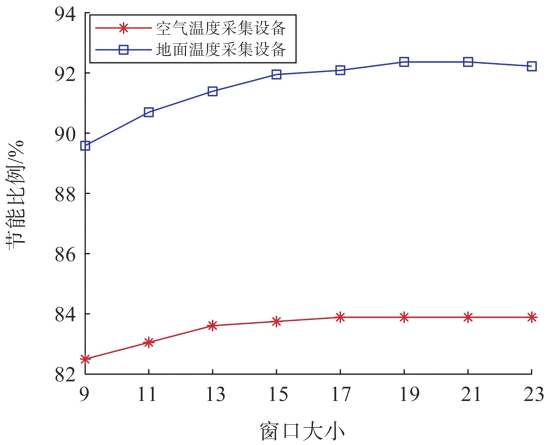


图 3 算法能耗分析

Fig.3 Energy consumption analysis of the algorithm

由图 3 可以看出,采取本研究提出的环境数据预测算法,在 2021 年 8 月,可以使空气温度采集设备能耗降低 80%以上,地面温度采集设备最大能耗可以降低 92%以上。随着最大训练窗口增大,能耗降低比例逐渐增大,增大幅度总体较小,这是由于本研究算法能够较好地适应环境,受参数设定影响较小。

在已有真实数据集中,温度采样周期通常以小时为单位,这时由于温度感知设备电能通常有限,难以支撑较高频率的数据收集。如果每小时采集温度数据,一旦某处存在环境异常等情况,可能最长 1 h 才能发现。本研究算法可以以更高的频率采集数据,以更低的频率上传数据,在环境数据出现异常时重新训练模型并上传,确保环境异常及时发现,降低传输能耗。

4.4 时延分析

本研究提出的自适应周期预测算法在预测过程中将带来额外的数据采集时延。为了全面评估本研究算法的时延特性,在真实数据集基础上对时延进行分析。限定用户容忍误差为 1 ℃,时延上限设定为 8~22 h,与固定窗口模型进行对比,结果如图 4 所示。

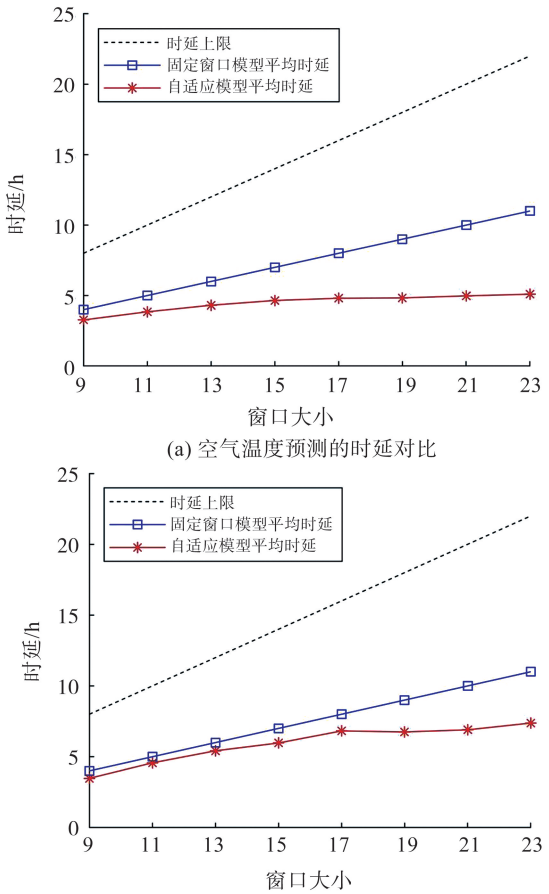


图 4 算法时延对比

Fig.4 Delay comparison of algorithms

由图 4 可以看出,当限定时延上限时,固定窗口模型平均时延始终为时延上限的一半;采取自适应周期预测算法的平均时延始终低于固定窗口模型的平均时延,且随窗口大小的提高而降低,在空气温度预测场景中,窗口大小为 24 时,平均时延降低 50%以上。需要说明的是,这一时延是数据收集的平均时延,加入了数据符合线性特点时未及时上传数据的时延。当出现异常数据时,新模型的训练时间决定了系统的响应时间。

综合以上结果可以看出,本研究算法无论是在预测准确度还是数据采集时延方面,均优于固定窗口线性模型。这是由于自适应模型能够更好地适应环境的实际变化情况,对模型进行实时调整。采用本研究提出的自适应模型能够将数据采集能耗降低 80%以上,有效降低智慧城市建设带来的额外能耗,为节能减排提供较好的支撑。本算法具有较低的运算开销,可以封装为边缘服务,并在实际的气象监测设备中部署实现。

4.5 训练时间对比

为了更全面地对比本研究算法在模型训练效率方面的优势,基于全国 10 000 个站点真实数据集模拟云端训练场景。在该场景中,采取经典的机器学习算法,基于前 6 d 的历史空气温度数据进行模型训练,并对第 7 天的数据进行预测。为了对比的公平性,本研究算法可接受误差设为 1 ℃,最大窗口设为 24。

本研究算法是在感知设备中基于少量数据进行实时预测的算法,与云端基于海量数据进行训练的算法并不具备直接可比性。为此,本研究在云端模拟感知设备内部的环境,基于前 6 d 的数据进行

模型实时训练及预测。为了保证对比的公平性,将本研究所用算法在 6 d 中训练及预测所用的全部运算处理时间计入训练时间,并与已有机器学习算法的训练时间进行对比。由于机器学习算法基于云端存储的 6 d 内的全部数据进行一次性训练,而本研究算法基于实时的少量数据进行多次训练,因此数据误差不再具有可比性。由于在云端训练的算法不存在降低节点数据上传次数的效果,因此不再对比能耗。

在机器学习算法方面,选取傅里叶神经算子算法[13]、弹性网络回归算法、支持向量回归(support vector regression, SVR)高斯核算法、SVR 线性核算法、随机森林算法、线性回归算法、反向传播(back propagation, BP)神经网络算法进行对比,主要对比模型训练时间。对比算法选择经典参数设定,试验结果对比如表 2 所示。

表 2　训练时间对比

Table 2　Comparison of training time

预测算法	训练时间/s
自适应线性预测模型	18.296 9
傅里叶神经算子	42.861 8
随机森林	4 572.306 9
弹性网络回归	34.339 2
BP 神经网络	40 170.884 0
线性回归	30.443 4
SVR 高斯核	29.397 4
SVR 线性核	29.211 9

由表 2 可以看出,本研究提出的自适应线性预测模型训练时间最短,仅为 18.296 9 s,其次为 SVR 高斯核及 SVR 线性核算法,训练时间分别为 29.397 4、29.211 9 s。事实上,本研究算法的训练时间是 10 000 个节点 6 d 内全部训练及预测的总时间,每个节点单次训练时间仅为 10 ms 级别,这也意味着当环境出现异常时,感知设备可以在 10 ms 左右训练得出新的模型并上传,在实时环境数据收集场景中具有较好的性能。该结果表明本研究算法训练及预测总时间相比当前机器学习算法的训练时间更短,训练及预测时间降低超过 37%。自适应线性预测模型训练及预测更快捷,能够快速得到合适的模型。

5　结论

本研究提出一种基于自适应线性模型的环境数据预测算法,能够较好适应环境的变化,对预测模型进行实时调整,有效提升模型的准确度,降低模型训练时延,节省数据采集设备的能耗。基于真实气象数据集的仿真试验结果验证了本研究算法的有效性。后续该算法将进行实物测试,进一步验证其降低能效的实际效果。

参考文献:

[1] EI M O, RACHID S, ABDELLAH C, et al. 6G enabled smart environments and sustainable cities: an intelligent big data architecture[C]//Proceedings of the 2022 IEEE 95th Vehicular Technology Conference (VTC2022-Spring). Helsinki, Finland: IEEE, 2022: 1-5.

[2] SHIDROKH G, MOHAMMAD H A, SEYED A S, et al. An IoT-based prediction technique for efficient energy consumption in buildings[J]. IEEE Transactions on Green Communications and Networking, 2021, 5(4): 2076-2088.

[3] LI J Z, LI G H, GAO H. Novel ε -approximation to data streams in sensor networks[J]. IEEE Transactions on Parallel and Distributed Systems, 2015, 26(6): 1654-1667.

[4] HUNG V, JEUNG. H, ABERER K. Anevaluation of model-based approaches to sensor data compression[J]. IEEE Transactions on Knowledge and Data Engineering, 2013, 25(11): 2434-2447.

[5] ZISIS I P, TIAN Y L. Prediction ofsea ice motion with convolutional long short-term memory networks[J]. IEEE Transactions on Geoscience and Remote Sensing, 2019, 57(9): 6865-6876.

[6] WANG W N, LIU W Q, CHEN H. Information granules-based BP neural network for long-term prediction of time series[J]. IEEE Transactions on Fuzzy Systems, 2021, 29(10): 2975-2987.

[7] DYLAN P, WANG N, SHEN S H. Energy demand prediction with optimized clustering-based federated learning[C]//Proceedings of the 2021 IEEE Global Communications Conference (GLOBECOM). Madrid, Spain: IEEE, 2021: 1-6.

[8] JIANG Y S, NIU S T, ZHANG K, et al. Spatial-temporal graph data mining for IoT-enabled air mobility prediction[J]. IEEE Internet of Things Journal, 2022, 9(12): 9232-9240.

[9] YAN B W, WANG G J, YU J G, et al. Spatial-temporal Chebyshev graph neural network for traffic flow prediction in IoT-based ITS[J]. IEEE Internet of Things Journal, 2022, 9(12): 9266-9279.

[10] VINIT K, MOHAMMADREZA B, NICHOLE M, et al. DeepTrack: lightweight deep learning for vehicle trajectory prediction in highways[J]. IEEE Transactions on Intelligent Transportation Systems, 2022, 23(10): 18927-18936.

[11] ZENG Y Y, ZHOU S J, XIANG K. Online-offline interactive urban crowd flow prediction toward IoT-based smart city [J]. IEEE Transactions on Services Computing, 2022, 15(6): 3417-3428.

[12] 万晨, 李文中, 丁望祥, 等. 一种基于自演化预训练的多变量时间序列预测算法[J]. 计算机学报, 2022, 45(3): 513-525.
WAN Chen, LI Wenzhong, DING Wangxiang, et al. A multivariate time series forecasting algorithm based on self-evolution and pretraining [J]. Chinese Journal of Computers, 2022, 45(3): 513-525.

[13] LI Z, KOVACHKI N, AZIZZADENSHELI K, et al. Fourier neural operator for parametric partial differential equations[C]// Proceedings of the 9th International Conference on Learning Representations (ICLR). [S.l.]: ICLR, 2021: 1-6.

[14] XU X B, ZHAO H, YAO H P, et al. A blockchain-enabled energy-efficient data collection system for UAV-assisted IoT[J]. IEEE Internet of Things Journal, 2021, 8(4): 2431-2443.

[15] AROOSA H, JOHN V, ARIS L, et al. Toward QoS prediction based on temporal transformers for IoT applications [J]. IEEE Transactions on Network and Service Management, 2022, 19(4): 4010-4027.

[16] 杨鹏史, 丁卉, 陈同, 等. 基于局部加权线性回归的城市公交车排放能耗预测[J]. 中山大学学报(自然科学版), 2019, 58(6): 111-118.
YANG Pengshi, DING Hui, CHEN Tong, et al. Estimation of emissions or electricity consumptions of urban buses based on locally weighted linear regression [J]. Acta Scientiarum Naturalium Universitatis SunYatseni, 2019, 58(6): 111-118.

[17] 郭松亮, 闫鹏君, 鄂浩坤. 基于 ARIMA 模型的北京市全社会用电量短期预测[J]. 北京信息科技大学学报(自然科学版), 2020, 35(5): 93-96.
GUO Songliang, YAN Pengjun, E Haokun. Short-term forecast of the total electricity consumption in Beijing based on ARIMA model[J]. Journal of Beijing Information Science & Technology University (Natural Science Edition), 2020, 35(5): 93-96.

[18] 唐继强, 钟鑫伟, 刘健, 等. 基于时间序列季节分类模型的轨道交通客流短期预测[J]. 重庆交通大学学报(自然科学版), 2021, 40(7): 31-38.
TANG Jiqiang, ZHONG Xinwei, LIU Jian, et al. Short term forecast of rail transit passenger flow based on time series seasonal classification model [J]. Journal of Chongqing Jiaotong University (Natural Science Edition), 2021, 40(7): 31-38.

[19] 马乐乐, 刘向杰. 非线性快速批次过程高效迭代学习预测函数控制[J]. 自动化学报, 2022, 48(2): 515-530.
MA Lele, LIU Xiangjie. A high efficiency iterative learning predictive functional control for nonlinear fast batch processes[J]. Acta Automatica Sinica, 2022, 48(2): 515-530.

(编辑:孙亚彤)