Advances in Applied Mathematics 应用数学进展, 2025, 14(2), 463-473
Published Online February 2025 in Hans. https://www.hanspub.org/journal/aam
https://doi.org/10.12677/aam.2025.142085

Hans汉斯

基于流数据的在线学习算法

张丽丽

青岛大学数学与统计学院，山东 青岛

收稿日期：2025年1月26日；录用日期：2025年2月19日；发布日期：2025年2月28日

摘　要

文章提出了一种针对流数据概念漂移现象的在线学习算法。为了提高预测的速度与精度，本文提出了多步预测回归集成模型，并详细描述了结合聚类算法的样本重抽样过程，以应对流数据的高维和大规模问题。通过将重抽样后的样本引入基于滑动窗口的在线自适应框架，结合多步预测回归模型组成本文的在线学习算法，该算法能够及时识别和处理概念漂移现象。此外，还提出了概念漂移的统计理论依据，确保了算法的准确性。针对路口车流量与网站浏览量数据，本文提出了概念漂移的类型，并针对突变漂移提出布尔因子，有效减少了突变漂移的不良影响。在实例评估中，本文方法在准确度和稳定性上均表现良好。

关键词

流数据，概念漂移，在线学习

Online Learning Algorithm Based on Streaming Data

Lili Zhang

School of Mathematics and Statistics, Qingdao University, Qingdao Shandong

Received: Jan. 26th, 2025; accepted: Feb. 19th, 2025; published: Feb. 28th, 2025

Abstract

This paper proposes an online learning algorithm for the concept drift phenomenon of streaming data. In order to improve the speed and accuracy of prediction, this paper proposes a multi-step prediction regression ensemble model and describes in detail the sample resampling process combined with the clustering algorithm to cope with the high-dimensional and large-scale problems of streaming data. By introducing the resampled samples into an online adaptive framework based on sliding windows and combining them with the multi-step prediction regression model to form the

文章引用: 张丽丽. 基于流数据的在线学习算法[J]. 应用数学进展, 2025, 14(2): 463-473.
DOI: 10.12677/aam.2025.142085

online learning algorithm of this paper, the algorithm can timely identify and handle the concept drift phenomenon. In addition, the paper also proposes a statistical theoretical basis for concept drift to ensure the accuracy of the algorithm. For the intersection traffic flow and website pageview data, this paper proposes the type of concept drift and proposes a Boolean factor for sudden drift, which effectively reduces the adverse effects of sudden drift. In the example evaluation, the method in this paper performs well in both accuracy and stability.

Keywords

Streaming Data, Concept Drift, Online Learning

1. 引言

在大数据时代，流数据的处理和分析成为了研究和应用中的重要课题。流数据通常表现为高速、连续、动态变化的数据流，其主要特点是数据量大且不断变化，要求系统具有较强的实时处理能力。随着信息技术和智能交通系统的快速发展，流数据的处理与分析在多个领域得到了广泛应用。路口车流量和网站浏览量作为两类典型的流数据，具有高速、连续、动态变化等特征。车流量数据对于交通管理和城市规划具有重要意义，而网站浏览量数据则广泛应用于网站优化、广告推荐以及用户行为分析等场景。如何有效、实时地分析流数据，成为了许多研究者面临的挑战。流数据分析不仅需要快速计算，还要求能够适应数据的概念漂移和动态变化。因此，如何设计高效且准确的在线学习算法，已成为流数据处理领域的热点研究方向。

由于流数据具有动态变化的特性，采用多步预测方法能够在一次预测中同时获得多个未来时间点的预测值，从而有效提高预测的效率。通过将多步预测与集成学习相结合，可以进一步增强模型的稳定性与准确性。多步预测不仅优化了流数据处理的速度，还能够提供更加精确的未来趋势预测，从而为决策提供更为可靠的支持。Dolgintseva 等人总结了多步预测[1]的关键挑战和方法，能够同时预测未来多个时间点的结果，从而提供更加全面的趋势分析。LightGBM [2]是一种基于决策树的梯度提升集成方法，它具有许多优势，包括稀疏优化、并行学习、多损失函数、正则化、批处理和提前停止，且在处理大规模数据时效率更高。由于 LightGBM 本身并不支持多步预测。递归、直接和直接递归等策略使 LightGBM 能够处理多步骤预测任务。直接法能提供比递归法更好的精度，与直接递归法相比具有更高的计算效率[3]，因此本文将直接法用于 LightGBM 模型的多步预测，提出了 LightGBM 与线性模型相结合的多步预测模型，提高了预测的速度，并利用集成方法提高了模型的准确度。

此外，针对流数据通常具有高维度和大规模的特点，Wu 等人提出了一种新颖的二阶在线特征选择算法[4]，旨在提高大规模高维稀疏数据流的处理效率，避免了传统方法中常见的高计算成本问题。Zhao 等人提出了一种基于分层抽样的大规模数据聚类算法[5]，通过选择抽样样本来提高计算效率，提升了处理大数据时的计算效率，并在多个大规模数据集上验证了其优越性。本文将响应变量视为关键变量，通过计算各个特征与响应变量之间的相关系数，进行特征选择，选出的特征随后用于样本重抽样，旨在减少冗余特征的影响，降低计算复杂度，提升模型的学习效率。

流数据中常常出现的概念漂移[6]现象，也对模型的准确性和稳定性提出了更高的要求。Ren Haojie 等

人提出了一种基于聚类模式下的最优抽样策略[7]，能够通过有限的数据流有效地检测到概念漂移。在线学习方法在应对流数据中的概念漂移现象方面表现出了显著的优势。例如在线 ARIMA 方法[8]通过采用在线梯度下降法(ODG)和在线牛顿步法(ONS)等流行的在线凸优化解，有效提高了对非平稳数据的预测精度。这些算法能够解决异常点检测问题，处理到达数据的观测顺序并及时更新模型参数。因此，它们在许多实际应用中展现了出色的适应性和效率。针对流数据中的概念漂移问题，Wu 等人[9]结合 OASW [10]滑动窗口方法，提出了一种在线自适应学习策略，结合聚类进行样本重抽样，显著提高了模型的准确性和稳定性。

基于上述背景，本文提出了一种面向回归问题的 OASW 滑动窗口方法的在线自适应学习框架。该框架能够在概念漂移发生时及时调整模型参数，从而保持较高的预测精度。同时，本文详细描述了样本重抽样的过程，首先通过特征选择提取出关键特征，然后将这些特征用于聚类算法，在聚类的基础上进行样本重抽样，获取发生概念漂移的重抽样样本用于在线自适应框架，重训练多步预测回归模型。这样既能克服流数据因规模庞大而导致的高计算成本问题，又能及时捕捉概念漂移的变化，适时调整模型参数，提升模型的准确性和稳定性。此外，本文还针对路口车流量和网站浏览量等数据，分析了不同类型的概念漂移。在处理突变漂移[11]时，本文提出添加布尔变量进行样本聚类，这大大减少了突变漂移对模型的影响，从而提高了模型的适用性和鲁棒性。针对多步预测模型，本文还提供了概念漂移(即数据分布变化)的理论依据，在统计层面证明了数据发生概念漂移，与在线自适应框架结果一致，从而实现了双重保障，增强了概念漂移的检测与应对。最后，本文通过集成学习方法对多步预测进行了增强，进一步提升了模型在处理流数据时的适应性、预测能力和效率。

2. 基于在线多步预测方法的理论介绍

2.1. 多步预测

假设样本量为 N，预测步长为 H，数据流可表示为：$D_t=\left(X_t,y_t\right)$，且$1\le t\le N$，数据包含 m 个解释变量，即 $X_t=\left(X_t^1,\cdots,X_t^m\right)$，$y_t$ 为响应变量。在 t 时刻，模型的输入为解释变量与响应变量的滞后值、解释变量的 H 个未来值，即 X_t 和 y_t 的历史值、X_t 的后续 H 个时刻的值。设 l 表示 X_t 和 y_t 的历史时间戳，对于预测层 h，直接预测的模型构建如下：

$$y_{t+h}=f_h\left(X_{t+H},\cdots,X_t,\cdots,X_{t-l+1}\right)+\beta_{h1}y_t+\cdots+\beta_{hl}y_{t-l+1}+\varepsilon_h$$

其中$t\in\left[1,N-H\right]$，$h\in\left(1,\cdots,H\right)$，$\varepsilon_h\sim\mathrm{N}\left(0,1\right)$。

模型表示解释变量与响应变量独立影响 y_{t+h}，将两部分之间的交互关系放入 ε_h 中。在 h 遍历整个预测步长 H 过程中，模型包含了两个函数关系，f_h 表示解释变量与 y_{t+h} 建立的函数关系，是解释变量与 y_{t+h} 平行建立的 H 个 Lightgbm 基模型；另一个函数关系是响应变量的历史值与 y_{t+h} 的线性函数关系。

本文考虑流数据的分布会随着时间而发生变化，发生概念漂移。针对流数据 y_t 可以表示如下：

$$y_t=\begin{cases}f\left(X_{t+H},\cdots,X_t,\cdots,X_{t-l+1}\right)+\beta_1y_t+\cdots+\beta_ly_{t-l+1}+\varepsilon,\ t=1,\cdots,\tau,\\ f\left(X_{t+H},\cdots,X_t,\cdots,X_{t-l+1}\right)+\beta_1y_t+\cdots+\beta_ly_{t-l+1}+\mu+\varepsilon,\ t=\tau+1,\cdots\end{cases}$$

τ 是一个未知的时间点，流数据在 τ 时间点上分布发生变化，即发生了概念漂移，μ 可看作 y_t 与其历史值的线性模型常数项，可以提出如下假设：$H_0:\mu=0$　$H_1:\mu\ne0$，通过 t 检验判断 μ 的显著性，从而判断数据的分布是否发生变化。t 统计量为：

$$t=\frac{\hat{\mu}}{SE\left(\hat{\mu}\right)}$$

将本文实例数据带入计算后，通过在线自适应方法捕捉到的概念漂移发生点，其 p 值均小于 0.05，表明我们有足够的证据拒绝原假设。因此，基于统计显著性的标准，可以得出结论：数据流在统计上发生了概念漂移。这样的结果验证了在线自适应方法在监测和识别概念漂移中的有效性，并进一步证明了数据流在不同时间点上出现了变化，从而影响了模型的预测性能。

2.2. 样本重抽样

Ren Haojie [6]提出了根据 kernel 函数得到的聚类统计量，能够有效检测到概念漂移的存在。本文将使用 MiniBatchKMeans 作为聚类算法的选择，通过最小化聚类中心与每个数据点之间的距离总和来聚类数据。MiniBatchKMeans 是 KMeans 的一个变种，使用小批量数据进行更新，从而显著提高了算法的计算效率，尤其适用于大规模数据集[12]。聚类性能会随着维数的增加而下降，因此需要采用降维技术来提高性能。本文使用 UMAP [13]降低聚类模型的输入维数。UMAP 通过其高效的计算速度、对局部和全局结构的保留、以及对大数据的可扩展性，使得它成为在高维数据和大数据处理中非常有效的工具。经过使用 UMAP 降维技术和 MiniBatchKMeans 聚类算法处理后，在线学习算法的完成时间显著减少，缩短了约 60%。

本文的样本重抽样流程如图 1 所示。

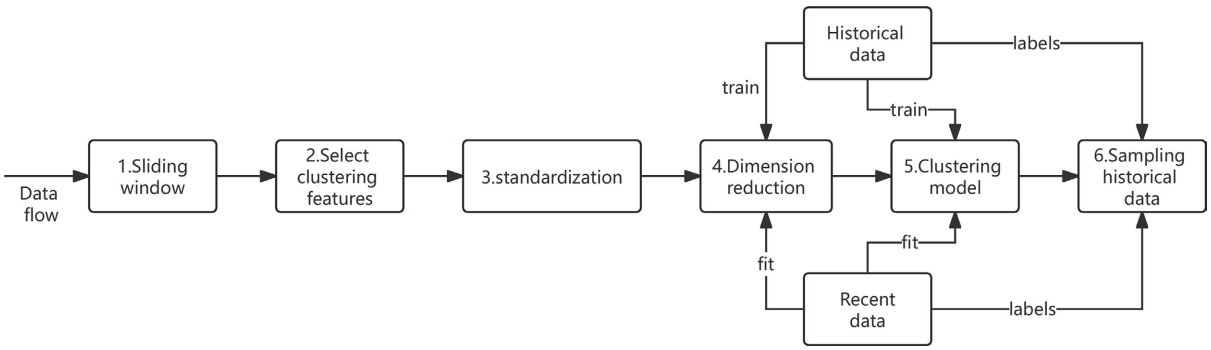

Figure 1. Sample resampling
图 1. 样本重抽样

为了更好地捕捉流数据分布的变化，提高模型的准确度，同时避免过拟合，我们引入了最近数据和历史数据的结合。最近数据被定义为从概念漂移开始的时间点到当前时间步之间的所有数据。由于最近数据的分布发生了变化，必须将其纳入模型进行重训练。然而，最近数据量较少且具有特殊性，因此我们需要对其进行数据跃迁处理，以增加重训练数据量并减弱其特殊性，从而避免模型过拟合。

为此，我们采用了样本重抽样的方法，从历史数据中提取与最近数据相似的部分，完成数据跃迁。这种方式不仅使得重训练的模型能够充分利用最近数据的最新信息，同时也有效地减少了由于数据量不足和特殊性引起的过拟合问题。

样本重抽样的过程为：首先通过滑动窗口方法遍历数据，并手动选择对聚类有重要作用的特征变量 S_t，其中 S_t 是挑选出的具有聚类作用的变量，考虑到本文提出的多步预测模型是基于线性关系构建响应变量之间的联系，且 LightGBM 模型也可以近似看作线性模型，因此在判断特征之间的相关性时，使用相关系数是合适的。通过计算数据中变量与变量之间的相关系数，找寻与因变量 y 相关系数最大的 2~3 个变量，与因变量 y 一起作为 S_t，完成变量的选择，后续用于样本聚类，为样本重抽样做准备。对数据进行标准化，减少量纲对聚类的影响；创建一个 UMAP 降维器实例，并将其 n_components 参数设置为多

步预测模型输入数据特征维度的 0.05 倍，使用历史数据对降维模型进行拟合训练。之所以选择 UMAP 进行降维，是因为在高维数据空间中进行聚类分析会极大增加计算负担，尤其是在在线学习算法中，可能导致实时性要求无法满足。使用 UMAP 降维能够显著提高计算效率，减少高维数据处理带来的时间消耗。使用 MiniBatchKMeans 对历史数据进行聚类分析并返回聚类标签 N；对于 MiniBatchKMeans 的参数设置，对于不同的数据集，将 n_clusters 参数指定为用于聚类特征变量 S_t 的个数，对于路口车流量数据，批次大小设置为一天的样本数量；对于网站浏览量数据，批次大小则设定为两天的样本数量，以便更好地适应不同的数据特性。将训练好的降维模型用于最近数据进行降维，将训练好的聚类模型用于最近数据中，得到最近数据的聚类标签 M，并得到每个标签所对应的最近数据样本数量，表示为 $k_1:m_1,k_2:m_2,\cdots,k_c:m_c$，其中 k_i 表示最近数据中的类标签，m_i 表示最近数据中对应类的样本数量，且 $m_1>m_2>\cdots>m_c$；从历史数据聚类标签 N 中，挑出最近数据聚类标签 M，以及其对应的历史数据样本数量，可以表示为 $k_1:n_1,k_2:n_2,\cdots,k_c:n_c$，$n_i$ 表示历史数据中类标签的样本数量。对于最近数据中样本量最多的聚类标签，直接将对应历史数据的同类样本纳入抽样中，也就是 n_1 直接放入抽样。对于其他剩余的聚类标签($k_2\cdots k_c$)，将采用如下方式确定抽样大小：

$$\begin{cases} a=\min\left(n_2,\cdots,n_c\right) \\ b_i=a*\dfrac{m_i}{size\left(y\right)},\ i=2,\cdots,c \end{cases}$$

取历史数据中其他聚类标签($k_2\cdots k_c$)的最小样本量作为每个标签的最大抽样量，再根据最近数据中聚类标签的样本量与总样本量的比例，随机抽取历史数据中的其他类标签样本，完成样本重抽样。通过这种基于聚类的样本重抽样方法，我们能够使重抽样后的历史数据分布更加接近最近数据的分布，从而充分利用最近数据中的漂移信息，同时避免模型过度拟合。这一过程为历史数据与最近数据之间的跃迁提供了有效支持，为后续在线自适应模型的构建奠定了坚实基础。

2.3. 在线自适应

Yang [9]提出了优化自适应滑动窗口(OASW)方法，该自适应 LightGBM 模型可以在没有人为干预的情况下对物联网数据流进行持续学习和漂移适应，具有较高的准确性和效率。作为一种与模型无关的自适应框架，OASW 也适用于时间序列预测任务，因此，本文提出了基于 OASW 的时间序列预测自适应框架并用于回归训练。

2.3.1. 自适应框架

Algorithm. Online adaptation (replacement)

Input:
$\boldsymbol{D_i}$, $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$: a data stream, the warning and drift thresholds,
$\boldsymbol{win_1}$, $\boldsymbol{win_2}$: size of sliding window and adaptive window,
$\boldsymbol{\theta}$: sampling function
$W \leftarrow \varnothing$
$S \leftarrow 0$
For $D_i \in$ Stream:
$\text{acc1}=\text{R2}\left(i-win_1:i\right)$
$\text{acc2}=\text{R2}\left(i-2win_1:i-win_1\right)$
If S==0 && $\text{acc1}<\alpha*\text{acc2}$:

续表

$W \leftarrow W \cup D_i$
$S \leftarrow 1$
end if
If S==1:
ta=Size(W)
If acc1 $< \beta *$ acc2 :
$S \leftarrow 2$
$f \leftarrow i$
$W \leftarrow W \cup D_i$
Regressor $\leftarrow$ retrain regressor on $W \cup \theta(D_i, W)$
else if acc1 $\geq \beta *$ acc2 || ta== win_2 :
$S \leftarrow 0$
$W \leftarrow \varnothing$

首先利用已有的历史数据来训练离线模型。随后，利用训练好的模型对流数据进行处理。如果在流过程中检测到漂移，则自适应系统部署专门设计的采样方法，从历史数据中获取采样。然后使用采样的历史数据和最新数据更新模型，以确保准确的预测。

大小为 win_1 的滑动窗口和大小为 win_2 的自适应窗口，α：警告阈值，β：漂移阈值。针对上述算法，计算相应的回归准确度，如果准确度下降到一定地步，将视为发生概念漂移，获取重抽样后的历史数据与最近数据进行模型的重训练，以减少概念漂移的损失，能够包容漂移的发生提高回归拟合准确度，减少误差。

2.3.2. 模型集成

在本文的在线多步预测模型的新方法中，我们基于基础模型权重映射提出集成模型，引入了一个模型权重来动态地改变模型，进而改变模型的输出预测，使预测更加准确。该策略针对多步预测中建立的 H 个基模型，每个时间点的数据均建立 H 个模型进行预测，那么该数据最终由基模型的加权集成模型来预测。因此，对于给定的基模型 i，模型权重即为：

$$w_i = acc(y_i, \hat{y}_i)$$

且权重 $w_i \in [0,1]$。考虑将多步预测的模型权重更改为其他模型评估的衡量标准，$w_i = 1/mae(y_i, \hat{y}_i)$，或 $w_i = 1/rmse(y_i, \hat{y}_i)$ 时，考察本文在线学习算法在实例评估部分的表现情况，综合衡量模型的决定系数和平均绝对误差，得出模型权重为 R 方时，实例结果表现最好。

集成模型的权重取决于基模型的准确度，随着流数据的遍历，动态改变着权重矩阵。通过对基模型的集成，使得预测值更加接近真实值，权重的动态改变及时适应概念漂移的出现，提高模型的预测准确度，减少损失。

2.3.3. 超参数优化

模型的性能受到 win_1、win_2、α、β 等超参数的影响。为了优化这些超参数，我们使用了粒子群优化算法(PSO)。在超参数优化方法中，粒子群优化算法是一种流行的基于群体的优化算法，它通过群体中个体之间的沟通和合作来检测最优值[14]。在 PSO 中，群体中的每个粒子将基于当前个体最佳位置和其他粒子共享的当前全局最佳位置，不断更新自己的位置和速度。因此，粒子将向候选全局最优位置移动，以检测最优解。

PSO 比大多数其他超参数优化方法更快，因为它的计算复杂度简单，并且支持并行执行。粒子群算

法适用于不同类型的超参数和大型超参数空间，如 LightGBM 的超参数空间。四个主要的超参数，通过 PSO 方法进行调优。通过检测这些超参数的最优值，可以获得优化的 LightGBM 模型，用于准确的流数据分析。

2.4. 辨别因子

本文提出了针对突变漂移的一个辨别因子。由于突变漂移发生变化迅速且突然，一般的自适应方法没有办法及时考虑到分布突变，于是本文从数据本身方面进行改进，标记数据。定义一个布尔类型的变量，针对目标变量值 y 的变化，可通过经验人士给出目标变量的合理值，如果超出此合理值，该定义的布尔类型变量为 False，其他合理值为 True。

$$x_i = \begin{cases} \text{True}, & y \in E \\ \text{False}, & y \notin E \end{cases}$$

其中 E 为目标变量 y 值的合理范围。定义此变量的目的是额外提供一个聚类变量，使自适应系统及时识别到样本本身发生的变化，能够有助于突变漂移的检测与自适应。

3. 实例评估

3.1. 数据

在本研究中，我们使用了两组不同类型的实际数据进行实例评估。首先，路口车流量数据涵盖了从 2015 年 11 月 1 日至 2017 年 7 月 1 日期间，三个路口每小时车辆通过的数量。该数据集包括时间戳、路口编号以及相应的车流量信息，为我们提供了关于城市交通状况的细致记录。其次，网站浏览量数据自 2015 年 7 月 1 日起，记录了 375 天内来自八个国家(如法国、英国等)的五大最受欢迎网页的每日浏览量。该数据集反映了不同国家用户的网络访问模式，能够为我们提供关于网站流量和用户行为的深刻见解。以上数据都可以从 kaggle 网站中获得。通过这些真实的流数据，我们能够对所提出的在线学习算法进行有效评估，分析其在实际应用中的表现与适应性。

3.2. 概念漂移类型检测

概念漂移主要是指解释变量和目标变量之间的关系随时间变化的在线监督学习场景。通常，概念漂移可以分为四种类型[6]，如图 2 所示。

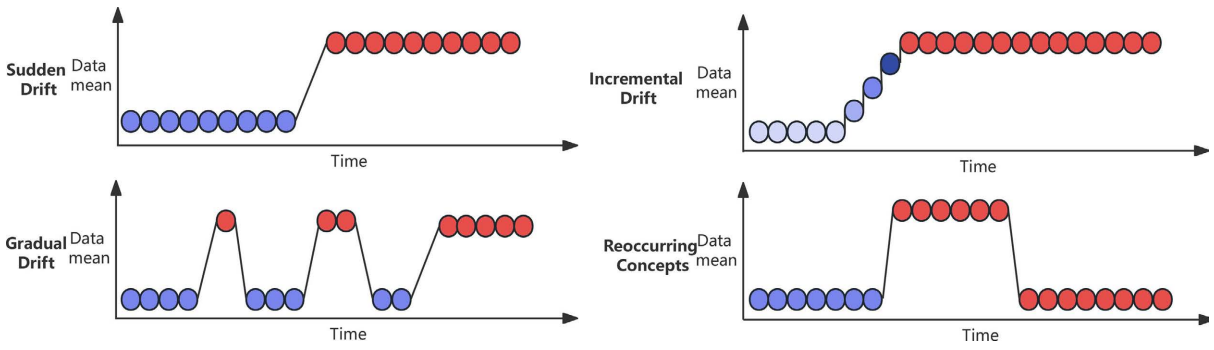

Figure 2. Concept drift types
图 2. 概念漂移类型

分析实例数据，根据数据均值的分布变化，分析数据中的概念漂移类型。

对于路口车流量数据，从本文方法监测到的第一个概念漂移时刻开始，随着时间的推移，等距样本的均值先趋于稳定，然后逐渐增大，最终趋于稳定，可以判断数据发生了增量漂移。

对于网站浏览量数据，随着时间的推移，等距样本的均值先趋于稳定，然后突然上升、下降，最后再次趋于稳定，且突变前后样本均值不同，可以判断此数据发生了突变漂移。

3.3. 实验设置

实验分为两个阶段。首先是离线阶段，在此阶段中，使用历史数据训练批处理模型，生成回归基模型，为接下来的在线阶段做好准备。接下来是在线阶段，模拟模型接收连续数据流的动态环境，使用本文提出的在线学习算法动态更新模型参数，以适应流数据中的概念漂移。为了评估所提出的在线学习算法的性能，我们使用路口车流量数据和网站浏览量数据进行实验，选取了前 20%的数据作为历史数据，用于训练模型。随后，使用剩余的数据模拟流数据进行重采样，并应用在线自适应算法进行评估。通过多步预测的准确度和平均绝对误差，来判断本文方法的适用性。

针对上述数据集，最优模型超参数的设置如表 1 所示。

Table 1. Hyperparameter settings
表 1. 超参数设置

数据集	模型	win_1	win_2	α	β
路口车流量数据	OASW	782	2360	0.978	0.941
路口车流量数据	Replacement	1300	3515	0.989	0.978
网站浏览量数据	OASW	640	3132	0.982	0.949
网站浏览量数据	Replacement	658	3153	0.985	0.957

3.4. 路口车流量数据集仿真

对于道路车流量数据集来说，一个时间戳是一小时，24*3 个时间戳代表一天。以 72 为 H(预测步长)，带入超参数，应用上述本文所提出的在线学习算法，采用 R_2 和 MAE 来评价模型的性能。R_2 和 MAE 的计算公式如下：

$$R_2\left(y,\hat{y}\right)=1-\frac{\sum_{i=1}^{72}\left(y_i-\hat{y}_i\right)^2}{\sum_{i=1}^{72}\left(y_i-\bar{y}_i\right)^2}$$

$$\mathrm{MAE}\left(y,\hat{y}\right)=\frac{1}{72}\sum_{i=1}^{72}\left|y_i-\hat{y}_i\right|$$

R_2 和 MAE 的结果表现如表 2 所示。

Table 2. Simulation results
表 2. 仿真结果

Model	Replacement	Offline	OASW	ARD
R_2	0.89	0.48	0.82	0.86
MAE	2.73	4.02	3.66	5.06

自动相关性确定(ARD)回归[15]是一种基于贝叶斯方法的回归算法，旨在解决回归任务。该模型与贝

叶斯岭回归具有相似的结构，但通过实施 ARD 回归算法，能够自动选择并调整特征的相关性和重要性，从而提高预测精度。值得注意的是，ARD 回归也可以与 MultiOutputRegressor 函数相结合，完成多步预测任务，进一步增强了其在复杂回归任务中的应用潜力。绘制出在样本剩余 80%的测试数据中，随着流数据的增加，在每个时间戳上，离线学习与在线学习算法(本文方法)的 R_2 变化曲线，如图 3 所示。

Figure 3. Accuracy curve
图 3. 准确度曲线

可以看出随着流数据样本量的增加，样本数据发生概念漂移，样本分布发生变化，变得不稳定，普通的离线回归模型已经满足不了准确度的需求。采用本文提出的在线学习算法，在发生概念漂移时，对历史数据重采样，将重采样的数据带入本文提出的在线自适应算法中，在样本分布发生变化的时候动态地改变回归模型，及时将准确度提升，减少损失。

图中显示流数据存在大范围的增量漂移，随着样本量越来越大，样本的分布变化越来越明显，准确度逐渐下降，与上述根据样本均值分布展示的增量漂移一致。因此，本文提出的方法可以解决增量漂移，提高准确度。

图 4 为四种方法的 R_2 箱线图，表示不论样本流数据的分布发生什么变化，本文方法的拟合稳定性都为最佳。

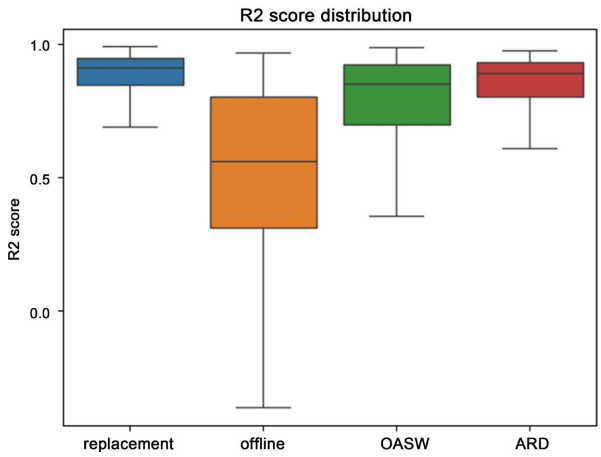

Figure 4. Box plot
图 4. 箱线图

3.5. 网站浏览量数据集仿真

针对网络浏览量数据，一个时间戳就是一天，但是数据中含有 40 个来自不同国家的最热网址，也就

是说 40 个时间戳代表一天。以 40 为 H (预测步长)，带入超参数，应用上述本文所提出的在线学习算法，计算 R_2 和 MAE 的值，结果如表 3 所示。

Table 3. Simulation results
表 3. 仿真结果

Model	Replacement	Offline	OASW	ARD
R_2	0.92	0.75	0.90	0.91
MAE	0.190	0.191	0.465	0.64

随着流数据的增加，在每个时间戳上，离线学习与在线学习算法(本文方法)的 R_2 变化曲线如图 5 所示。

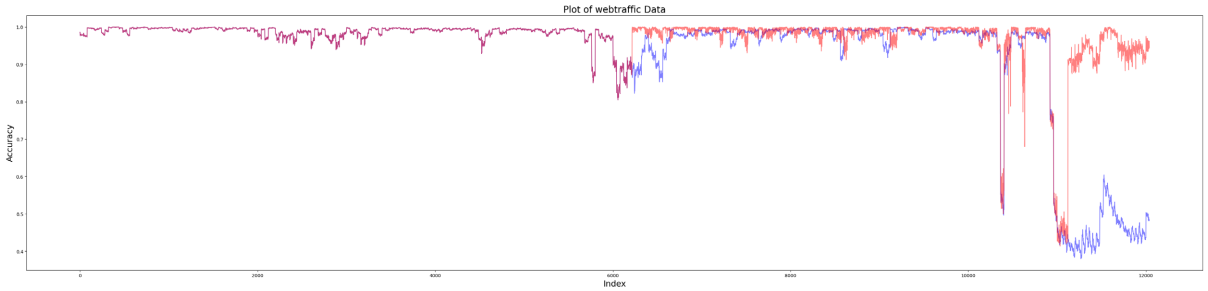

Figure 5. Accuracy curve
图 5. 准确度曲线

针对网站浏览量数据，离线学习模型的准确度在某个时刻出现了显著下降，而在此之前，准确度一直保持较高且稳定的状态。这一变化表明该流数据发生了突变漂移，与上述根据样本均值分布展示的突变漂移一致。实验结果证明，本文提出的方法能够有效应对突变漂移，从而提高模型的准确度。

图 6 为四种方法的 R_2 箱线图，表示不论样本流数据的分布发生什么变化，本文方法的拟合稳定性都为最佳。

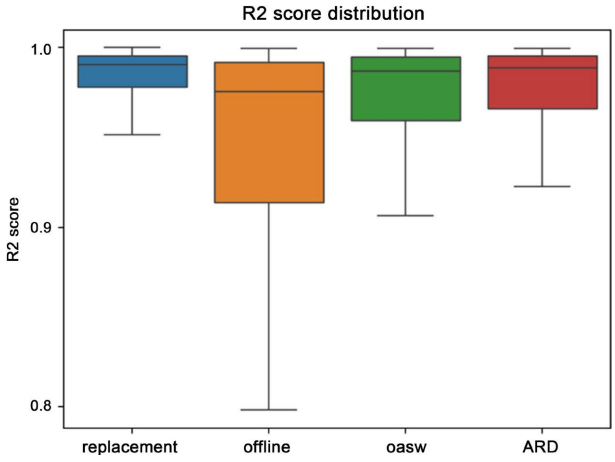

Figure 6. Box plot
图 6. 箱线图

4. 结论

本文提出了一种高效的在线学习算法，旨在解决流数据中的概念漂移问题，从而提升流数据处理的准确性与稳定性。同时给出了结合多步预测回归模型与样本重抽样的在线自适应框架，并基于统计理论提供了概念漂移的理论依据。在分析路口车流量与网站浏览量数据时，本文探讨了概念漂移的不同类型，并针对突变漂移提出了创新的布尔因子，旨在提高对概念漂移的适应能力。在实例评估中，本文方法展示了较高的精确度与适应能力，显著优于其他算法。但是还有待于判断本文算法对其他类的概念漂移的适应能力。未来的研究可以进一步探讨如何优化算法的计算效率，还可以考虑结合深度学习等技术，探索更加复杂的概念漂移类型和更广泛的应用场景。通过这一研究，本文为流数据分析和预测任务中的在线学习提供了更具实用性的理论支持与算法框架。

参考文献

[1] Dolgintseva, E., Wu, H., Petrosian, O., Zhadan, A., Allakhverdyan, A. and Martemyanov, A. (2024) Comparison of Multi-Step Forecasting Methods for Renewable Energy. *Energy Systems*. https://doi.org/10.1007/s12667-024-00656-w

[2] Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., *et al*. (2017) LightGBM: A Highly Efficient Gradient Boosting Decision Tree. *Proceedings of the* 31*st International Conference on Neural Information Processing Systems*, Long Beach, 4-9 December 2017, 3149-3157.

[3] Ben Taieb, S., Bontempi, G., Atiya, A.F. and Sorjamaa, A. (2012) A Review and Comparison of Strategies for Multi-Step Ahead Time Series Forecasting Based on the NN5 Forecasting Competition. *Expert Systems with Applications*, **39**, 7067-7083. https://doi.org/10.1016/j.eswa.2012.01.039

[4] Wu, Y., Hoi, S.C.H., Mei, T. and Yu, N. (2017) Large-Scale Online Feature Selection for Ultra-High Dimensional Sparse Data. *ACM Transactions on Knowledge Discovery from Data*, **11**, 1-22. https://doi.org/10.1145/3070646

[5] Zhao, X., Liang, J. and Dang, C. (2019) A Stratified Sampling Based Clustering Algorithm for Large-Scale Data. *Knowledge-Based Systems*, **163**, 416-428. https://doi.org/10.1016/j.knosys.2018.09.007

[6] Gama, J., Žliobaitė, I., Bifet, A., Pechenizkiy, M. and Bouchachia, A. (2014) A Survey on Concept Drift Adaptation. *ACM Computing Surveys*, **46**, 1-37. https://doi.org/10.1145/2523813

[7] Ren, H., Zou, C., Chen, N. and Li, R. (2020) Large-Scale Datastreams Surveillance via Pattern-Oriented-Sampling. *Journal of the American Statistical Association*, **117**, 794-808. https://doi.org/10.1080/01621459.2020.1819295

[8] Kozitsin, V., Katser, I. and Lakontsev, D. (2021) Online Forecasting and Anomaly Detection Based on the ARIMA Model. *Applied Sciences*, **11**, Article 3194. https://doi.org/10.3390/app11073194

[9] Wu, H., Elizaveta, D., Zhadan, A. and Petrosian, O. (2023) Forecasting Online Adaptation Methods for Energy Domain. *Engineering Applications of Artificial Intelligence*, **123**, Article 106499. https://doi.org/10.1016/j.engappai.2023.106499

[10] Yang, L. and Shami, A. (2021) A Lightweight Concept Drift Detection and Adaptation Framework for IoT Data Streams. *IEEE Internet of Things Magazine*, **4**, 96-101. https://doi.org/10.1109/iotm.0001.2100012

[11] Lu, J., Liu, A., Dong, F., Gu, F., Gama, J. and Zhang, G. (2018) Learning under Concept Drift: A Review. *IEEE Transactions on Knowledge and Data Engineering*, **31**, 2346-2363. https://doi.org/10.1109/tkde.2018.2876857

[12] Peng, K., Leung, V.C.M. and Huang, Q. (2018) Clustering Approach Based on Mini Batch Kmeans for Intrusion Detection System over Big Data. *IEEE Access*, **6**, 11897-11906. https://doi.org/10.1109/access.2018.2810267

[13] McInnes, L., Healy, J., Saul, N. and Großberger, L. (2018) UMAP: Uniform Manifold Approximation and Projection. *Journal of Open Source Software*, **3**, Article 861. https://doi.org/10.21105/joss.00861

[14] Yang, L. and Shami, A. (2020) On Hyperparameter Optimization of Machine Learning Algorithms: Theory and Practice. *Neurocomputing*, **415**, 295-316. https://doi.org/10.1016/j.neucom.2020.07.061

[15] Paper, D. (2019) Scikit-Learn Regression Tuning. In: Paper, D., Ed., *Hands-on Scikit-Learn for Machine Learning Applications*, Apress, 189-213. https://doi.org/10.1007/978-1-4842-5373-1_7