基于时间序列的校园网络流量分析

谢颖瑶[1] 黄 猛[1] 田累积[1] 任玺睿[1]
XIE Yingyao HUANG Meng TIAN Leiji REN Xirui

摘 要 准确分析校园网络的使用趋势和及时检测网络异常，是众多高校校园管理者提高网络资源利用率和应对网络故障做出判断和响应的关键。使用 STL 季节性分解算法对网络流量数据原始时间序列进行分离，获得长期趋势、季节差异和短期波动特征，分析校园网络使用情况。同时，将季节性分解算法与 ARIMA 模型相结合，建立 SARIMA 模型，通过获取某高校 8 个月的 168 万条校园网络数据，建立适应动态特征网络流量变化的复杂时间序列模型，对未来可能的网络流量值进行预测，当预测值与真实值的差值过大时，视为流量出现异常，产生网络预警信息。

关键词 时间序列；校园网络流量；季节性分解算法；SARIMA 模型；异常检测

doi：10.3969/j.issn.1672-9528.2024.06.040

0 引言

随着校园网络资源的不断扩充和应用场景的逐渐复杂化，日益增长的网络用户和接入设备都给校园网络带来各种挑战，比如校园网络占用率较高但是不能查明原因、带宽不足需优化而缺乏统计数据、网络突然中断不能查明原因等。因此，动态调节网络负载和准确发现网络流量异常，对于众多高校提高网络资源利用率和及时应对网络故障具有重要意义。

本文以基于时间序列 [1] 的数据为核心，使得校园管理者可以根据分析结果深入地了解网络流量分布、高峰时段、性别差异以及年级群体的流量使用情况，动态调节校园网络负载，提高网络资源的利用率。传统的设置静态阈值方法难以适应季节性、周期性变化的网络流量，可能导致预警系统无法准确捕捉到异常流量的出现 [2]。因此，本文将季节性分解算法 [3-4] 与 ARIMA 模型结合成适用复杂时间序列的 SARIMA 模型，针对动态特征网络流量变化进行网络流量预测 [5]。

1 数据概述

1.1 数据来源

时间序列数据是按照时间顺序排列的数据点的集合，具有周期性变化或符合某种趋势等特征。本文获取的某高校北区（获取平台：https://10.159.245.10/）2023 年 3 月到 10 月的 168 万条网络流量数据是单变量时序数据，由单一变量出或入流量随时间而变化，且这种变化存在明显的周期性，包括季度、寒暑假、毕业季等季节性变化。

1.2 数据处理

为了不影响分析结果的准确性，应该避免缺失值带来的时间间隔受损和周期性震荡经度受到影响的情况，使时间序列数据中的时间索引得以连续。因此，取缺失值的前一个和后一个观测值的平均值作为填补。

2 基于季节性分解算法分析流量趋势

校园网络流量数据的季节性分解分析是智慧校园建设中的重要一环。通过 STL 季节性分解算法，能够准确地将网络流量数据分解为长期趋势、季节差异和短期波动三个部分，从而更好地理解校园网络的使用情况。这项技术为实践数据治理提供了强有力的支持，让网络管理员能够基于数据的指引，实时调整网络负载，降低网络故障发生的风险。

以数据为核心，通过深入分析网络流量数据，网络管理员可以发现网络使用的高峰时段、不同性别间的流量差异以及各年级群体的使用特点。这种洞察力不仅有助于优化校园网络资源的分配，还能够为未来的网络规划和优化提供重要参考。在智慧校园建设中，数据驱动的决策能够更好地满足师生的需求，提升校园网络的整体运行效率，为教学和科研提供更好的支持。

1. 防灾科技学院信息工程学院 河北廊坊 065000

[基金项目] 防灾科技学院 2023 大学生创新创业项目（S202311775072）

2.1 STL 季节性分解算法原理

STL 季节性分解算法，将时间序列数据分解成趋势、季节性和残差三个成分。其原理如下。

2.1.1 趋势分解

首先，根据数据的采样频率和周期选择一个合适的窗口大小，应用局部加权回归算法，对于窗口内的每个数据点，使用多项式拟合局部趋势。然后，在整个数据集上重复进行，获得数据整体趋势

2.1.2 季节性分解

在趋势分解之后，通过计算滑动窗口内的局部平均值，对去除趋势后的时间序列进行季节性分解，可以捕捉数据在不同时间尺度上的周期性变化。

2.1.3 残差分解

在得到趋势和季节性成分之后，用原始数据减去趋势和季节性成分计算得到残差序列。残差序列包含了数据中无法被趋势和季节性解释的随机波动和不规则变动。

其算法公式为：

$$Y_t = T_t + S_t + I_t \quad (1)$$

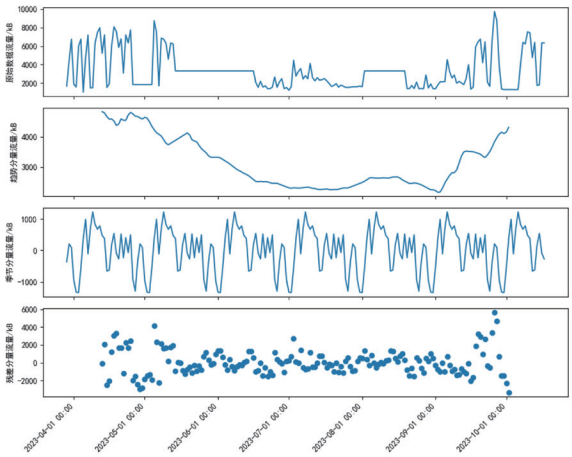
式中：Y_t 为原时间序列在 t 时刻的值，为 S_t 时间序列在 t 时刻的趋势量；I_t 为时间序列在 t 时刻的季节量，为时间序列在 t 时刻的趋随机量。

2.2 结果分析

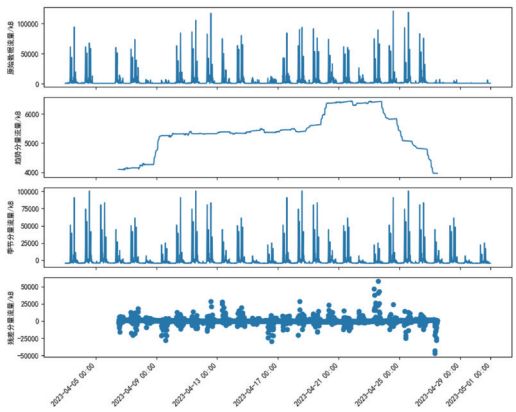
本文将某高校 3 月到 10 月教学区域以及学生区域的网络流量数据，基于教学区年月日的分析、教学区与学生区的对比、毕业生与非毕业生的对比、男生楼与女生楼的对比这四个角度进行季节性分析，得出如下结论。

2.2.1 教学区年周日季节性分析结果

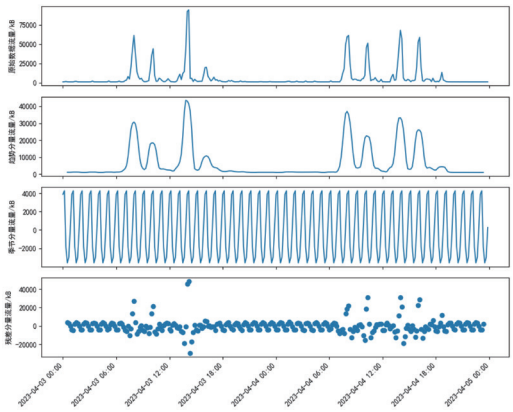
如图 1 所示，依次分别为教学区年分析结果图、教学区周分析结果图、教学区日分析结果图。由图 1 可知，在学期期间，网络流量呈现出一种正常的波动趋势。



（a）教学区年分析结果图



（b）教学区周分析结果图



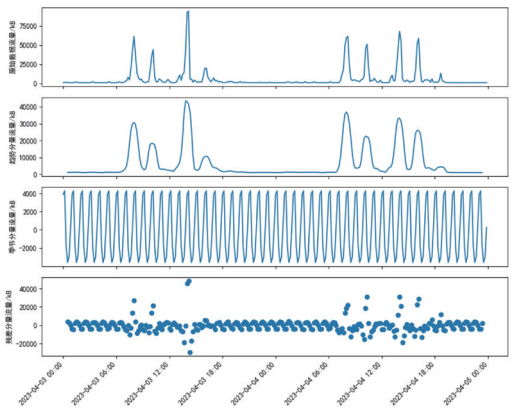
（c）教学区日分析结果图

图 1 教学区年周日季节性分析结果图

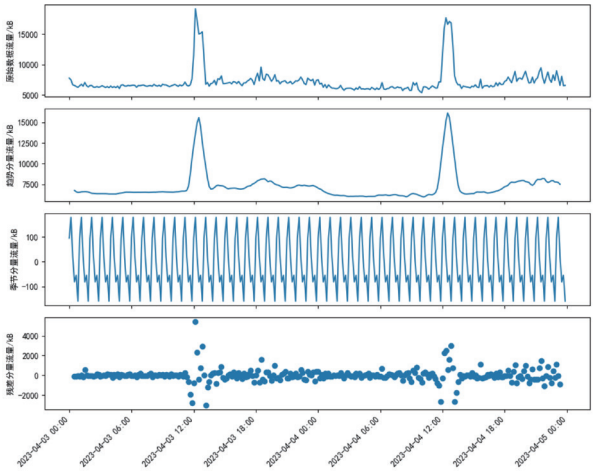
暑假期间，只有少数学生选择留校，但网络流量仍然保持在一定水平。这种波动在一周的不同时间段也有所体现。周一至周五，学生大多数时间都在教学区上课，因此网络流量在这些时间段内会有所波动。在周末，学生则主要在学生区域休息，导致教学区的网络流量较少。此外，一天中的不同时段也会对网络流量产生影响，在上课时间段，流量使用较为集中，而在午休或晚间，流量使用则相对较少。

2.2.2 教学区与宿舍区分析结果

如图 2 所示，依次分别是教学区日分析结果图、宿舍区日分析结果图。



（a）教学区日分析结果图



（b）宿舍区日分析结果图

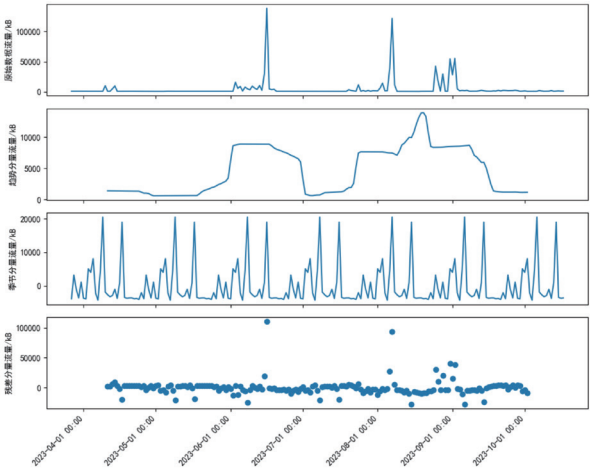
图 2 教学区与宿舍区日分析结果图

由图 2 可以看出，教学楼和宿舍区的网络流量使用存在着明显的时段性差异。在学生休息时间外，教学区的网络流量呈现出与宿舍区相反的趋势。这表明学生在下课后普遍选择返回宿舍，而不是在教学区停留。

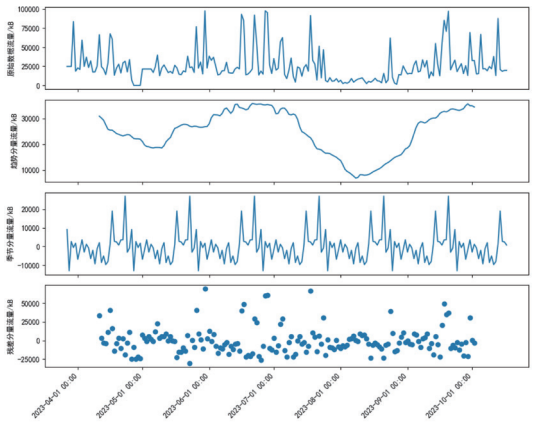
为了更好地适应校园网络用户的需求，管理员应当根据实际情况灵活调节和分配网络带宽资源。在教学楼网络流量较高的时候，可以考虑增加教学楼的带宽资源，以满足学生在上课期间的网络需求。相应地，在宿舍区网络流量较高时，应当将更多的带宽资源分配给宿舍区，以满足学生在课后自主学习时间的网络使用需求。

2.2.3 毕业生与非毕业生分析结果

如图 3 所示，依次为毕业生分析结果图、非毕业生分析结果图，由图 3 可知，在 4 月和 5 月这两个月份，大部分毕业生选择去实习，这导致毕业生宿舍的网络使用流量明显减少，与非毕业生相比有显著差异。然而，尽管毕业生的离校使得宿舍网络使用流量减少，但仍然存在着对校园网的需求。



（a）毕业生分析结果图



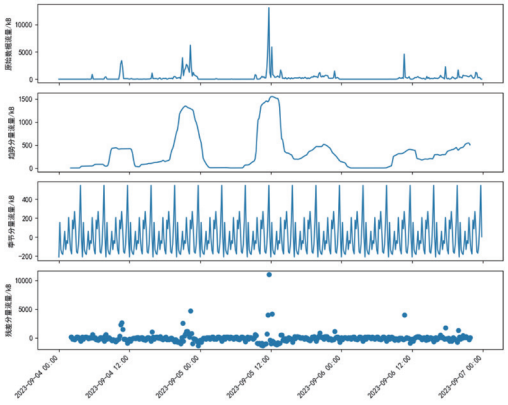
（b）非毕业生分析结果图

图 3 毕业生与非毕业生分析结果图

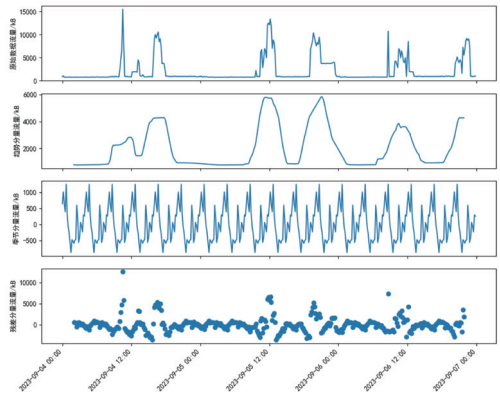
基于这些观察结果，学校可以考虑推出不同的网络流量套餐，以满足不同用户的需求。对于那些需要大量网络流量的用户，应提供高流量的套餐选项，以确保他们能够顺畅地使用校园网，而不会遭受网络拥塞或额外费用的困扰。相反，对于网络使用较少的用户，可以提供低流量的套餐选项。

2.2.4 男生楼与女生楼分析结果图

如图 4 所示，依次为男生楼分析结果图、女生楼分析结果图，由图 4 可观察到，男生楼的网络流量使用时间相对于女生楼更长。



（a）男生楼分析结果图



（b）女生楼分析结果图

图 4 男生楼与女生楼分析结果图

这表明男生们倾向于在更晚的时间段内使用网络资源，可能是因为他们更倾向于在晚间进行学习、娱乐或社交活动。值得注意的是，学生们的网络使用时间往往持续到深夜，甚至延伸至凌晨时分，这可能会对他们的健康和学习效率造成不利影响。

针对这一情况，学校可以考虑采取一系列措施，引导学生形成健康的作息习惯。除了在晚间一定时间后自动断网或限制访问特定网站，还可以加强与家长的沟通，共同制定并执行家庭作息时间表，以确保学生在家庭环境中也能够遵循良好的作息规律。

3 基于 SARIMA 模型预测校园网络流量

3.1 ARIMA 模型的局限性

传统的基于时间序列的未来值预测模型有 AR 模型（即自回归模型）和 MA 模型（即移动平均模型），但是二者都有其自身的问题。AR 模型通过假设过去值与未来值之间存在线性关系来预测未来值，但面对突发变化或者噪声很大的数据并不能很好地处理。而 MA 模型将一个时间序列看作是过去若干期噪声的加权平均，难以捕捉数据的“趋势”。为了结合 AR 模型和 MA 模型的优势，同时处理平稳和非平稳的时间序列问题，引入了 ARIMA 模型，其中 I 代表着差分过程，此过程将非平稳的时间序列变得平稳。但与此同时，时间序列中的周期性特征被弱化，甚至掩盖了。因此，ARIMA 模型并不能很好地预测具有季节性特征的时间序列的未来值 [6]。

3.2 构建 SARIMA 模型

SARIMA 模型（seasonal autoregressive integrated moving average）在 ARIMA(p,d,q) 模型的基础上结合 STL（seasonal-trend decomposition using loess）季节性分结算法，增加了 3 个超参数 (P,D,Q)，使得能够支持带有季节性成分的时间序列数据 [7-11]。

SARIMA(p,d,q)(P,D,Q,s) 总共 7 个参数，可以分成两类：3 个非季节参数 (p,d,q) 和 4 个季节参数 (P,D,Q,s)。

非季节参数：

p：AR(p)，非季节自回归的阶数

$$\chi_t = \theta_0 + \theta_1\chi_{t-1} + \cdots + \theta_p\chi_{t-p} + \epsilon_t \quad (1)$$

d：I(d)，一步差分的次数

q：MA(q)，非季节移动平均的阶数

$$\chi_t = \theta_0 + \theta_1\epsilon_{t-1} + \theta_2\chi_{t-2} + \cdots + \theta_q\chi_{t-q} + \theta_t \quad (2)$$

季节参数：

P：季节自回归的阶数

$$\chi_t = \alpha_0 + \alpha_1\chi_{t-s} + \cdots + \alpha_P\chi_{t-Ps} + \epsilon_t \quad (3)$$

D：季节差分的次数

Q：季节移动平均的阶数

$$\chi_t = \beta_0 + \beta_1\epsilon_{t-s} + \beta_2\epsilon_{t-2s} + \cdots + \beta_P\epsilon_{t-Qs} + \epsilon_t \quad (4)$$

s：季节长度，或者说是周期大小

为了使公式表示简化，引入延迟算子 B，$B^n y_t = x_{t-n}$ 表示延迟算子 B 作用在 y_t 这个时刻上，相当于对 y_t 后移 n 位，得到的结果就是 y_{t-n}，因此 $(1-B)y_t = y_t - x_{t-1}$ 表示差分，则 $(1-B)^d y_t$ 表示对 y_t 左 d 次差分。

定义延迟算子多项式：

$$\theta(B) = 1 - \theta_1 B - \cdots - \theta_P B^P \quad (5)$$

因此，可以将 ARIMA 模型简化为：

$$\theta(B)(1-B)^d y_t = \theta(B)\epsilon_t \quad (6)$$

对应可得 SARIMA 模型的简化公式：

$$\theta_{(P)}(B)\theta_{(P)}(B_s)(1-B)^d(1-B_s)^D y_t = \theta_{(q)}(B)\theta_{(Q)}(B_s)\epsilon_t \quad (7)$$

3.2.1 判断时间序列数据的平稳性

本文采用单位根的方式检验数据的平稳性，通过 augmented dickey-fuller（ADF）图展示检验结果。本文使用 Python 的 statsmodels 库中的 adfuller 函数进行 ADF 检验，先假设本文的时间序列数据为非平稳的。

```
import pandas as pd
from statsmodels.tsa.stattools import adfuller
data =history_flew
# 进行 ADF 检验
result = adfuller(data)
# 输出 ADF 统计量和 P 值
adf_statistic = result[0]
p_value = result[1]
# 输出临界值
critical_values = result[4]
print(f 'ADF Statistic: {adf_statistic}' )
print(f 'P-Value: {p_value}' )
print('Critical Values:')
for key, value in critical_values.items():
   print(f '{key}: {value}' )
if p_value < 0.05: # 选择显著性水平为 0.05
   print('Reject Null Hypothesis: Time series is stationary')
else:
    print('Fail to Reject Null Hypothesis: Time series is non-stationary')
```

控制台输出结果：

ADF Statistic: -1.827

P-Value: 0.366

Critical Values:

1%: -3.457

5%: -2.873

10%: -2.573

3.2.2 寻找原始差分步数

测试统计量（ADF statistic）大于所有的临界值（在 1%、5% 和 10% 的显著性水平下），表明不能拒绝原假设，即不能认为数据是平稳的。因此，在进行时间序列分析之前，需要进行差分，使时序数据平稳。

通过 pandas 库中的 diff() 函数用于计算一个时间序列的一阶差分。通过遍历不同的差分步数，对每个步数的结果进行 ADF 检验。

```
import pandas as pd
from statsmodels.tsa.stattools import adfuller
data = history_flew
time_series = pd.Series(data)
# 定义一个函数执行 ADF 检验
def adf_test(series, max_diff=5):
    for i in range(1, max_diff+1):
        differenced_series = series.diff(i).dropna()
        result = adfuller(differenced_series)
        adf_statistic = result[0]
        p_value = result[1]
        print(f 'ADF Statistic (Diff={i}): {adf_statistic}')
        print(f 'P-Value (Diff={i}): {p_value}')
        # 判断是否拒绝零假设
        if p_value < 0.05:
            print(f 'Reject Null Hypothesis (Diff={i}): Time series is stationary\n')
        else:
            print(f 'Fail to Reject Null Hypothesis (Diff={i}): Time series is non-stationary\n' )
# 执行 ADF 检验
adf_test(time_series)
```

在满足单位根检验的许多差分结果中，本文选择 ADF 最小或差分后数据方差最小的步数，根据代码运行结果，确定最佳差分步数是 12 步，用 12 步差分的数据进行研究。

3.2.3 使用多阶差分确定 *d* 值

本文通过绘制不同阶差分后的自相关函数（ACF）图确定 d 值。如果 ACF 图中的第一个滞后值是负的，那么可能会怀疑数据存在过差分问题，也就是对数据做了过多的差分。本文尝试 1 阶和 2 阶差分，打印 ACF 图。

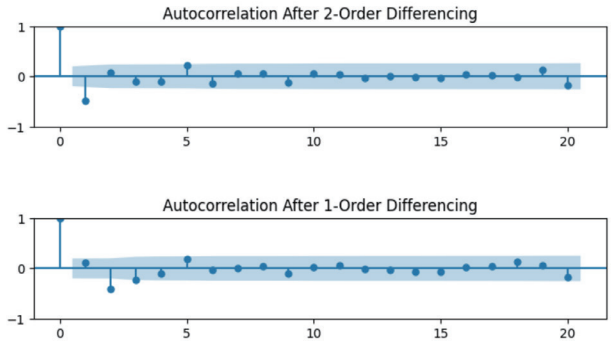


图 5 ACF 结果图

本文经过 12 步差分之后，是一个相对平稳的时间序列数据，当取 1 阶差分时，滞后为 1 的 ACF 值接近于 0；当取 2 阶差分时，滞后为 1 的 ACF 值为负，因此可以判断对 SARIMA 模型而言最佳的 *d* 取值为 1。

3.2.4 超参数择优

一般情况下，*p*、*q*、*D*、*D* 和 *Q* 不会取很大的值，因此本文尝试通过遍历的方式探索，先取 *p* 和 *q* 的范围为 1 ～ 4，使用每个值组合来拟合模型，比较每个模型的性能，探索剩余参数的最优组合。本文使用 2023 年 10 月 9 日至 2023 年 10 月 12 日，某教学楼网络流量数据分别建立模型，对接下来 12 小时网络流量数据进行预测的结果。按照每 10 分钟一条网络流量数据，设置其中季节长度为 144，即每天为一个周期（季节长度）。本文采用 AIC（Akaike Information Criterion）来自动择出最优的参数组合。

本文截取出最优的根据 AIC 值自动帮助选出来最优模型参数是：

SARIMA(1, 1, 2, 144)，AIC=1 231.715 3

图 6 为 SARIMA 模型的预测结果，蓝色线为真实值，橙色线为预测值 [12-15]。

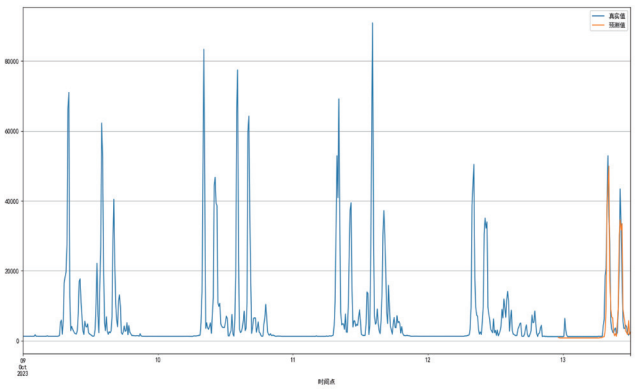


图 6 SARIMA 模型预测结果

由图 6 可得，预测值与真实值拟合程度较高。将本模型预测到的网络流量值和实际流量预测值进行比较，当差值过大时，视为流量出现异常。

4 结论

本文所提出的方法为校园网络管理者提供了一种有效的工具，能够更好地理解和应对校园网络的使用趋势和异常情况。根据本文的研究和分析，成功地使用 STL 季节性分解算法和 ARIMA 模型相结合，建立了适应动态特征网络流量变化的 SARIMA 模型。通过对某高校 8 个月的 168 万条校园网络数据进行分析，深入了解了网络流量的长期趋势、季节差异和短期波动特征。这使得校园管理者能够更准确地了解网络资源的利用情况，包括网络流量分布、高峰时段、性别差异以及年级群体的流量使用情况，对于提高网络资源利用率、动态调节校园网络负载以及及时解决网络故障具有重要的实际意义。

未来，将进一步优化模型，引入更多的因素和数据源，以提高网络流量预测的准确性和灵活性。此外，还可以探索机器学习等新技术的应用，结合大数据分析方法，进一步提升网络管理的智能化水平，为校园网络管理领域的进一步发展和提升提供有益的参考。

参考文献：

[1] 范振杰，罗娜 . 基于改进 VAE 的时间序列数据增强方法 [J/OL]. 华东理工大学学报 (自然科学版):1-11[2024-03-16]. https://doi.org/10.14135/j.cnki.1006-3080.20230315001.

[2] 张磊，巨能攀，何朝阳，等 . 滑坡裂缝计时序数据实时异常检测分析 [J]. 岩石力学与工程学报 ,2024,43(1):206-215.

[3] 姚汶伶，马蒙蒙，刘艳慧，等 . 基于季节性分解在广州市 2015—2022 年流感季节性及病原变迁分析中的应用 [J]. 中国公共卫生 ,2023,39(7):823-829.

[4]STEFENON S F, SEMAN L O, MARIANI V C, et al. Aggregating prophet and seasonal trend decomposition for time series forecasting of italian electricity spot prices[J]. Energies, 2023,16(3):1371.

[5] 胡凤婷 . 校园网流量的监测与分析研究 [D]. 桂林 : 桂林电子科技大学 ,2023.

[6] 周坤，许云飞，祁浩伟 . 基于改进 ARIMA 的新能源发电短期动态调度模型 [J]. 电脑与信息技术 ,2024,32(1):56-61.

[7] 冯隆基，楚成博，方磊，等 . 基于时序分解和 SARIMA-DSR 的台区可开放容量计算方法 [J]. 现代电子技术 , 2024, 47(2):127-132.

[8] 陈治霖，胡鸿韬，边迎迎 . 新冠疫情下基于 SARIMA 模型的上海港集装箱吞吐量预测 [J]. 工业工程与管理 , 2024, 29(1): 32-40.

[9] 董敏，娄峰 . 基于 SARIMA 和 ARIMA-GARCH 模型的共享单车用户量预测 : 以哈啰出行为例 [J]. 现代商业 , 2023(17): 46-49.

[10]CHATURVEDI S, RAJASEKAR E, NATARAJAN S, et al.A comparative assessment of SARIMA,LSTM RNN and Fb prophet models to forecast total and peak monthly energy demand for India[J]. Energy policy, 2022,168:113097.

[11]ZHAO Z, ZHAI M, LI G, et al.Study on the prediction effect of a combined model of SARIMA and LSTM based on SSA for influenza in shanxi province,china[J].BMC infectious diseases, 2023,23(1):1-14.

[12] 周浩，禹可，吴晓非 . 基于得分生成模型的时间序列异常检测方法 [J/OL]. 北京邮电大学学报 :1-7[2024-02-28]. https://doi.org/10.13190/j.jbupt.2023-093.

[13] 李昭怡，闫晓宇，马银翠，等 . 基于 AI 学习的校园网络异常流量检测 [J]. 现代信息科技 ,2023,7(22):15-19.

[14] 秦辉东，杨加，金建栋，等 . 校园网络实时监控平台设计及风险预警 [C]// 中国计算机用户协会网络应用分会 . 中国计算机用户协会网络应用分会 2023 年第二十七届网络新技术与应用年会论文集 . 北京 : 北京大学计算中心 , 2023:6.

[15]ZHANG L, SHI P, LAI X, et al.Air-conditioning load forecasting based on seasonal decomposition and ARIMA model[C]//2021 International Conference on Advance Computing and Innovative Technologies in Engineering,[v.1]. Piscataway: IEEE, 2021:664-669.

【作者简介】

谢颖瑶（2003—），女，广东广州人，本科，研究方向：计算机科学与技术。

黄猛（1976—），通信作者（email: hm@cidp.edu.cn），男，河南新乡人，硕士，教授，研究方向：自然语言处理、地理信息系统。

田累积（2002—），男，内蒙古包头人，本科，研究方向：计算机科学与技术。

任玺睿（2004—），女，仡佬族，贵州铜仁人，本科，研究方向：计算机科学与技术。

（收稿日期：2024-04-01）