



专题：网络智能化与生成式人工智能

面向 IP 网络扩容应用的复杂网络流量预测方法

高勇, 陆钱春, 李锋

(移动网络和移动多媒体技术国家重点实验室, 广东 深圳 518055)

摘要: IP 网络扩容是通信运营商保持网络平稳运行的一种常见维护方式, 核心在于预测未来一段时间的网络流量走势。IP 网络流量非常复杂, 具有局部不确定性、突发性、异质性等, 给预测带来困难。提出了一种针对复杂网络流量的预测方法, 它采用编码-解码结构, 即在编码层增加全局特征、在解码层增加全局特征和局部特征解析的方式解决局部不确定性; 通过增加先验知识缓解突发性; 模型采用样本均衡、归一化等方式尽量提取数据的共性, 避开数据的异质性。模型整体的参数较少, 具有较强的泛化性能; 同时采用人工特征和自动特征结合方式保证了浅层网络的准确率。实验结果表明, 所提出的方法具有准确率高、泛化性能强的特性。目前该方法已经在工程中大规模应用。

关键词: 网络流量预测; IP 网络扩容; 复杂网络流量

中图分类号: TP393

文献标志码: A

doi: 10.11959/j.issn.1000-0801.2023175

A complex network traffic prediction method for IP network expansion applications

GAO Yong, LU Qianchun, LI Feng

State Key Laboratory of Mobile Network and Mobile Multimedia Technology, Shenzhen 518055, China

Abstract: IP network expansion is a common maintenance method for communication operators to keep the network running smoothly. The core is to predict the network traffic trend in the future period. IP network traffic is very complex, with local uncertainty, burstiness, heterogeneity, etc., which brings difficulties to prediction. A prediction method for complex network traffic was proposed. It adopted an encode-decode structure, adding global features to the encoding layer, and combining global features and local features in the decoding layer to solve local uncertainties. The model used sample balance, normalization and other methods to extract the commonality of the data as much as possible to avoid the heterogeneity of the data. And emergencies were alleviated by increasing prior knowledge. The overall model had fewer parameters and had strong generalization performance, at the same time, the combination of artificial features and automatic features ensured the accuracy of the shallow network. The experimental results show that the proposed method has the characteristics of high accuracy and strong generalization performance. At present, this method has been applied on a large scale in engineering.

Key words: network traffic prediction, IP network expansion, complex network traffic

收稿日期: 2023-06-19; 修回日期: 2023-09-01

通信作者: 高勇, seegaoyong@163.com



0 引言

随着互联网的飞速发展和人们日益增长的网络业务需求，IP 网络迫切需要智能网络维护，其中 IP 网络扩容的智能化是解决网络拥塞、提高服务质量的重要手段。网络扩容指因业务需要，通过在原有的拓扑网络中增加站点或对部分链路增加带宽等措施增加网络传输容量的工程。扩容是一个软/硬件结合的复杂操作过程。除了硬件上需要进行断开光纤、增加站点、重新连接光纤等操作，软件上还需要在网管上进行相关的数据配置。从实际反馈来看，通常需要 3 个月左右的时间才能完成一次扩容。传统的被动式扩容需要网络发生拥塞后才扩容，扩容期间的网络拥塞会给用户造成困扰。另外，热点区域流量增长迅速，传统的被动式扩容会出现反复扩容，造成资源浪费。因此，提前预测流量走势，采用主动扩容的处理方式具有重要的应用价值。

在上述扩容背景下，流量预测的需求相应产生。根据端口、链路、环等对象的历史流量数据，利用机器学习方法学习各对象的流量变化规律，以预测未来 3 个月的流量数据。通过预测的未来流量，判断流量拥塞情况，进而有针对性地进行扩容。网络流量的复杂性导致传统的预测方法难以预测准确。其复杂性包括如下几点。

- 异质性：网络节点的异质性体现在每个节点受历史时刻观察值的影响不一致，在数据层面上表现为数据分布不一致，例如，城市基站和农村基站、不同社区的基站的特性差异较大。
- 局部不确定性：IP 网络规模巨大、结构复杂，常存在新增节点、删除节点等连接关系的动态变化，在数据层面表现为数据阶段性的变化。另外，IP 网络运营涉及路由的变动，也会导致节点具有局部不确定性的流量。

- 突发性：包括用户行为的突发性和网络状态的突发性，如商场活动、假期及高考信号屏蔽等。

据笔者所知，关于扩容的研究目前仍然不够充分。文献[1]提出了一种针对高校的网络扩容方式，详细调研学校常用流量类型，采用最小二乘法建模契合流量变化的业务预测模型，但是该方法主要针对高校学生的流量预测，并不能代表其他群体的流量，不具有普适性。文献[2]仅通过带宽利用率一个维度进行判决扩容，不够精准，扩容目标可能被放大。

对于流量的预测，可以用常见的一维时间序列预测方法，这在学术界有大量的论文研究。差分自回归移动平均（autoregressive integrated moving average, ARIMA）模型^[3]及其变种的线性模型广泛应用于预测，此类模型更加擅长平稳序列的短期预测。Facebook 公司提出的时间序列算法模型 prophet^[4]是基于时间序列分解和机器学习的拟合，应用较为广泛，该模型更偏向于预测具有周期性变化的问题。XGBoost^[5-6]是一个较为成熟的基于梯度提升的回归树的模型，常在时间序列的比赛中见到，并能取得较好的效果。上述方法模型较为成熟，常在工程中应用，但是缺乏灵活性，预测精度也低于神经网络系列方法。在神经网络方面，长短期记忆（long short-term memory, LSTM）网络及其变种门控循环单元（gated recurrent unit, GRU）在预测领域占据优势^[7]，可以作为一个单独的网络，也常作为复杂网络的一部分。文献[8]提出了长短期时间网络（long-and short-term temporal network, LSTNet），LSTNet 的结构通过跳线、卷积等多种方式组合，使得模型具有自动适应长周期和短周期的能力，该模型对无周期变化的场景不适用。然而，这几种神经网络方法对数据只具有局部注意力，缺乏对全局特征的把握，Informer^[9]以采用全注意力机制的编码器-解码器为主要结构，并对注意力机制进行部

分改变,具体是改变解码器的输出、注意力计算方式等,捕捉了长时间序列输出和输入之间的长期依赖性,但是因为参数过多,容易出现过拟合场景。文献[10]提出了一种用于多元时间序列长期预测的分层注意力网络,通过为不相关的外生因素分配较小的权重来削弱它们对预测的负面影响,采用多模态融合门来控制时间因素的影响。也有学者做基于图网络的时空预测^[11-12]的研究,在一维时间序列的基础上增加了图信息,该方法综合利用时间和空间信息,对于提升准确率具有一定帮助,但是该方法往往需要稳定的拓扑关系,在 IP 网络中存在大量的路由业务变动,拓扑关系不够稳定。

在网络流量的复杂性处理方面,文献[13]提出一种针对网络突变性的异常检测方法,该方法针对网络流量的突发具有很好的检测性能,但是没有考虑其他特性。文献[14]提出一种利用经验模态分解对网络流量进行分解,获得流量数据的趋势特征和细节特征,构建模型矩阵后再通过支持向量进行预测的方法,该方法主要针对短期的用户侧流量预测,不适合超过一个月的网元数据的预测。文献[15]提出一种基于多任务学习的算法,根据网络流量的时空特征预测网络流量,但该方法依赖拓扑结构,在用于扩容场景的长期流量预测中拓扑结构是随时间变化的,该方法不适用。

针对面向 IP 网络扩容的复杂网络流量预测问题,本文提出了一种复杂网络流量预测方法,该方法充分考虑网络流量的异质性、局部不确定性、突发性,尽量缓解其对网络流量预测的影响。本文的主要贡献如下。

- 提出一种面向 IP 网络扩容应用的复杂网络流量预测方法,该预测方法支持长周期的流量预测,有效指导扩容方案的实施。
- 所提出的网络流量预测方法,不仅对网络流量进行预测,同时对预测结果进行分类评级,剔除低可信的预测,方便工程运维

人员使用。

- 针对网络的异质性、局部不确定性、突发性,在网络流量预测过程中提出缓解方式,使得预测尽量准确。

1 网络流量预测方法

1.1 问题描述

流量预测是指通过对历史数据进行观测,得到未来一段时间的流量结果。预测问题描述如图 1 所示,利用 M 个历史数据,预测未来 N 个长度的数据。将历史流量数据称为观测数据,记为 $\mathbf{Y}_{\text{obs}} = [y_{t-M+1}, \dots, y_t]$; 将未来的数据称为预测数据,记为 $\mathbf{Y}_{\text{pre}} = [y_{t+1}, \dots, y_{t+N}]$, 通用的预测往往是获得条件分布 P :

$$P(\mathbf{Y}_{\text{pre}} | \mathbf{Y}_{\text{obs}}) \quad (1)$$

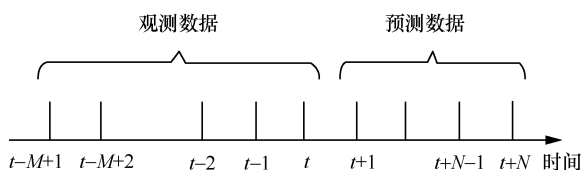


图 1 预测问题描述

但是对于流量预测问题,未来的流量隐含时间特征依赖,如星期几、是否节假日等,所以流量预测的条件分布中的条件可以为历史流量和日历特征,即:

$$P(\mathbf{Y}_{\text{pre}} | \mathbf{Y}_{\text{obs}}, \mathbf{X}_{(t-M+1):(t+N)}) \quad (2)$$

其中, $\mathbf{X}_{(t-M+1):(t+N)}$ 表示观测和预测区内的日历特征。

1.2 数据分析

两个端口流量的平均流速如图 2 所示,展示了两个网元端口在 2019 年 10 月 9 日到 2020 年 5 月 9 日之间的平均流速。样本 1 更靠近用户侧,受到用户周期规律的影响更大;样本 2 来自远离用户侧的网元端口,受网络内部管理的影响更大。从图 2 可以明显看到,样本 1 和样本 2 存在异质性,即数据分布不一致,样本 1 缺少大的突变流量,流量在一定范围波动,而样本 2 的数据大多数非



常平稳,波动较小,偶尔会出现非常大的流量突变。在图2的A处,样本1存在阶段性的跃进,体现出网络流量的局部不确定性,该处可能跟网络结构的变化有关,即在新的网络中有更多流量流经网元。在图2的B、C两处,样本2有明显的流量突变,此处可能跟春节有关,春节导致样本2流量的激增,但受到负载均衡的约束,第二天的流量迅速恢复。在图2的D处,样本1、样本2的流量比其他大多数时刻的流量高,体现网络流量的突发性和局部不确定性,此处应该是跟春节有关,也侧面说明引入节假日特征对于预测是应该有作用的。

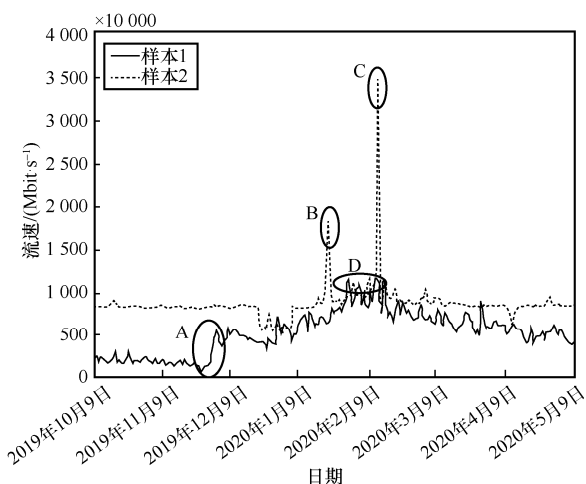


图2 两个端口流量的平均流速

通过分析可知,网络流量预测模型必须能够处理网络流量的异质性,如果不解决异质性问题,则无法在众多的网元上进行应用。模型同时要尽可能地缓解网络流量的突发性和局部不确定性的影响,这是提升网络流量准确率的关键。

1.3 流量预测模型

流量预测模型的整体架构如图3所示,以观测数据 X_{obs} 、观测期间日历特征 X_{obs} 、预测期间日历特征 X_{pre} 、观测数据的特征 Fea 为输入,分别获得流量预测的结果 Y_{pre} 和分类结果 C 。该网络流量预测模型针对性地解决网络流量的异质性、不确定性、突发性的影响。

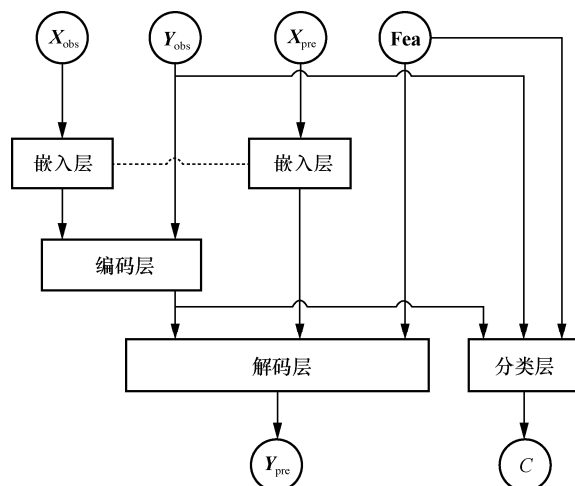


图3 流量预测模型的整体架构

对于异质性,本文主要通过以下方法解决:对数据进行样本级别的归一化,将所有流量预测数据归一化到0~1;进行样本均衡,等概率地选取每一个网元端口;网络参数较少,增加了适量的随机失活层,增强模型的泛化能力;增加分类层,对预测进行分类评级,剔除异质性过大导致的低可信预测。

对于局部不确定性,本文主要通过以下方法解决:在编码层增加全观测周期的特征汇聚层,减少LSTM导致的前期特征丢失;通过专家经验,增加人工全局特征,可以达到加强全局特征的效果;在解码层,通过全局解码层和局部解码层,充分结合全局特征和局部特征。

对于突发性,本文主要通过一些先验知识,获取预测数据内的节假日、周末等的特征。通过模型学习该部分特征对流量的影响,缓解突发性的影响。

首先所有日历特征通过嵌入层进行降维,将稀疏的特征维度变为稠密特征,观测部分和预测部分的嵌入层共享参数;然后将观测数据和对应的日历特征输入编码层,获得观测数据的特征编码。解码层输出预测流量的结果,解码层的输入分别是编码层的特征编码、输入的人工特征、待预测的日历特征。下面将对所有的组件分别阐述。

1.3.1 输入特征

对于输入特征, 主要对观测数据进行全局特征提取。因为编码层输出的潜在特征是完全黑盒的, 过于复杂的网络模型容易过拟合, 过于简单的网络模型无法有效提取特征。而人工提取的特征是基于专家经验提取的, 更加适用于该场景, 但是人工提取的特征较容易遗漏部分有效特征。将人工提取的有效特征和神经网络自动提取的特征结合, 可以综合提升模型能力。本文模型使用的输入特征是观测数据的均值、方差、中位数。均值、中位数对于流量预测结果是有帮助的, 方差对于分类结果是有帮助的。采用均值、方差等输入特征的方式, 可以以较小的计算成本获得性能提升。

1.3.2 网络结构

嵌入层主要对日历特征进行降维, 模型选取的日历特征主要为星期、节假日, 分别采用两个独热编码形成特征矩阵。星期采用 7 位长度的独热编码, 如星期一编码为 $[1,0,0,0,0,0,0]$; 节假日为了和星期对齐, 也采用 7 位长度编码, 但是只有前两位数据有效, $[1,0,0,0,0,0,0]$ 和 $[0,1,0,0,0,0,0]$ 分别表示是节假日和不是节假日; 最终日历特征维度为 2×7 , 分别表示星期和节假日的独热编码。通过示例可以看出独热编码需要 7 位表示, 且只有 1 位有效, 需要采用嵌入层对独热编码进行降维, 获得更加稠密的特征, 具体是采用一层全连接神经网络作为嵌入层, 输入为独热编码长度 7, 嵌入层的输出维度为 1, 将日历特征的最后一个维度输入, 则 2×7 日历特征转变为 2×1 的日历编码特征。需要强调的是, 此处所有日历特征共享一个嵌入层, 以减少模型参数。

编码层的主要作用是对输入进行编码, 以获得有价值的潜在特征, 消除冗余的属性。编码层示意图如图 4 所示, 该层的输入主要是经过嵌入层之后的日历特征 $X_{\text{obs-e}}$ 、观测数据 Y_{obs} 。首先两个输入通过融合层将两个特征合并为一个向量, 融合方式是时间维度不变, 在特征维度上进行拼

接。然后将融合之后的特征输入 LSTM 模块。对每一个时间步依次进行迭代, 最终可以获得一个对所有时间步进行编码的隐层特征 h_t 。因为 LSTM 是按照时间步的顺序处理的, 是有序的, 所以缺乏一个对历史流量信息的整体概括。常用的方法是增加一个反向的 LSTM 或者将历史流量输入新的全连接层, 本文提出的模型偏向于后者, 将 LSTM 输出的特征和历史流量进行拼接后输入全连接层, 通过该方式, 使用一层全连接既可以对 LSTM 和历史流量进行拼接, 又可以让隐层特征包含无序的全局概括, 同时可以控制编码层输出特征维度大小。

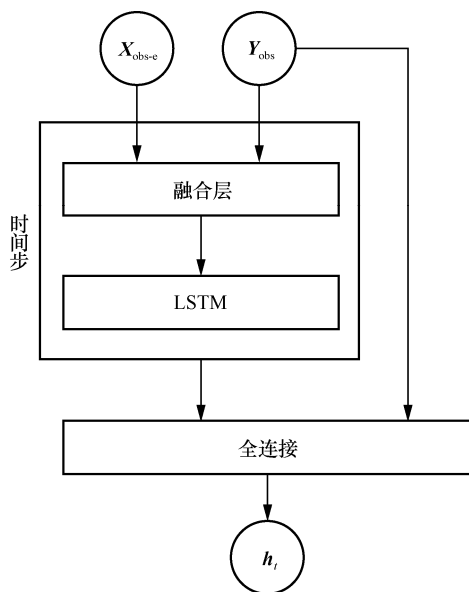


图4 编码层示意图

解码层的主要作用是对编码后的特征进行解码, 获得最终的流量预测结果。解码层示意图如图 5 所示, 该层的输入主要是经过嵌入层之后的日历特征、编码层输出数据、人工提取的特征。解码层主要分为两部分, 一部分是针对全局的特征解析, 另一部分是针对预测部分每个时间步的局部特征解析。全局特征层的输入分别是编码层的输出特征 h_t 、预测部分的日历特征 X_{pre} , 其中日历特征是经过嵌入层处理的, 将两者进行拼接后, 最终作为全局特征层的输入。全局特征是一层全



连接，它的输出分为两部分，一部分是所有预测的全局公共特征 $\mathbf{G}_{\text{com}}$ ，另一部分是对应于每个时间步的专有特征 $[\mathbf{G}_{t+1}, \dots, \mathbf{G}_{t+N}]$ 。该部分的优势是每一个时间步的预测都可以有全局的信息总览。局部特征层的输入是全局公共特征 $\mathbf{G}_{\text{com}}$ 、每个时间步的专有特征、人工提取的特征、对应的日历特征。局部特征也采用一层全连接的方式，主要专注于该时间步的预测，其输出为该时间步的预测结果。所有局部特征层的参数是共享的，减少了模型参数，提高了泛化能力。全局特征和局部特征结合的方式，使得模型可以同时关注细节和全局信息。

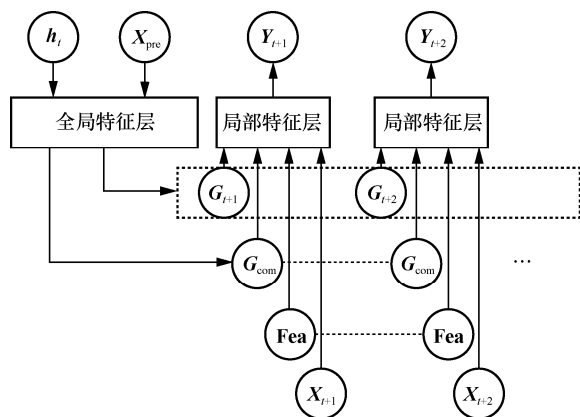


图5 解码层示意图

分类层的主要作用是对无法预测的样本进行识别。因为实际生活中采集的数据、流量存在波动，而有些样本因为一些人为因素或者外在因素发生流量突变，该部分是无法通过算法预测到的。分类层主要通过对历史流量信息进行分类，获得可预测和不可预测两个类别。该层输入为编码层特征、人工提取特征、历史流量数据。分类层是两层全连接网络，然后将信息输入 softmax 进行二分类。

1.3.3 损失函数与训练过程

预测的损失函数和分类的损失函数是分开的。预测损失函数采用 L1 范数，即：

$$L_{\text{predict}} = \sum_{i=1}^N \left| \frac{\hat{y}_i}{y_i} - 1 \right| \quad (3)$$

其中， $\hat{y}_i$ 和 y_i 分别是第 i 个时间步的预测值和真实值， $||$ 表示取绝对值。因为存在一些异常，损失突然增大，所以需要增加一个裁剪，使得异常的影响降低，即：

$$L_{\text{predict}} = \sum_{i=1}^N \text{Crop} \left(\left| \frac{\hat{y}_i}{y_i} - 1 \right| \right) \quad (4)$$

其中， $\text{Crop}(\cdot)$ 表示计算数值不能超过 1。

分类的损失函数采用经典的交叉熵损失函数，这里不赘述。

预测的训练过程和分类的训练过程是分开的，但是推理过程是同时进行的。先进行预测训练，等预测训练结束后，再进行分类训练。分开的原因如下：因为主要目的是预测流量，所以需要优先保证预测过程，不能让分类干扰预测；分类需要预测结束后提供标签，分类是针对预测的，所以对比预测结果与真实结果后，将对比结果作为分类的标签。

训练所需的所有样本需要进行样本级别的归一化，即取该样本的最大值线性映射到 1，该样本的最小值线性映射到 0，该方式可以加快模型收敛，消除网元端口流量数量级不同的影响。

训练过程中的所有批次是对所有样本等概率取样的，不会因为某个样本的数据长度长就增大概率，这样可以保证模型兼顾数据的异质性。

1.4 网络扩容应用

本文所提出的网络流量预测方法采用自监督的方式进行，不需要人工进行标记，只需要大量流量历史数据。考虑不同地市的差异，在新的地市往往需要重新训练，重新训练的初始化参数可以借助已训练完毕的地市参数。整体网络扩容方法如下。

步骤 1 收集新地市数据，并将网络预测方法进行部署。

步骤 2 重新训练流量预测网络，初始化参数可以选择已训练完的其他地市网络参数。

步骤 3 流量预测网络通过推理输出每个预

测对象的分类情况和预测情况。

步骤 4 丢弃分类结果为不可预测的样本；对于可预测的样本，统计其超过带宽 T 倍的天数 N ， T 一般取值 0.5。

步骤 5 按照天数 N 从大到小排序，此为扩容的优先顺序，依次安排运维人员扩容。同时展示网络端口流量的预测情况。

2 实验分析

本节通过对不同地市的端口流量数据进行实验分析，说明所提方法在预测准确率、分类准确率、泛化性能等方面的有效性。

2.1 样本采集

小型城市或城镇网元数量较少，且流量常往更大城市汇聚；为方便管理，常常对大型城市和超大型城市的网元数量进行区域划分，效果等同于中型城市。因此，中型城市的流量在一定程度上可以代表中国所有类型城市的流量，本文将中型城市作为实验数据集。选择两个中型城市的端口流量数据作为实验验证数据集，由于实际工程问题，在优先保证网络畅通的前提下，经常会遇到影响采集的情况，反映到数据就是数据缺失，为了实际解决网络扩容问题，选择两个有不同程度数据缺失的数据集，验证本文方法的泛化性能。由于不同地市设备部署存在先后关系，同时受工程业务的影响，不同地市的采集开始时间、结束时间未能保持一致。将每个地市作为一个单独的数据集，取最后 3 个月数据作为测试集，其他数据作为训练集。

数据集最终的呈现方式是按照天粒度。对于数据缺失的情况，训练集补缺方式是通过前 7 天的均值进行补偿；测试集不补偿，计算预测准确率时，跳过该数据。数据采集过程是将秒级数据汇聚为天粒度数据，即每秒采集的数据先汇聚为分钟级，再由分钟级汇聚为小时级，最后汇聚为天级别。

数据集描述见表 1，展示了不同数据的分布情况，包括开始采集时间、结束采集时间、端口数量。从表 1 可以明显看出，数据集 B 比数据集 A 的缺失严重，因此模型的验证主要基于数据集 A；而数据集 B 主要用于验证模型的泛化性能。

表 1 数据集描述

数据集名	开始采集时间	结束采集时间	端口数量/个	缺失数据占比
数据集 A	2019 年 11 月	2020 年 11 月	14 824	0.1%
数据集 B	2017 年 2 月	2019 年 12 月	35 464	10%

2.2 模型预测有效性

在扩容场景下，可以接受流量预测结果存在适当的误差，即并不需要保证百分百准确，在适当误差之内都可以认为该样本预测结果正确；反之，如果预测结果误差非常大，则认为该样本不准确。对于某一个预测结果来说，其偏差率小于 20% 则认为正确，偏差率 bias 定义为：

$$\text{bias} = \frac{|y_p - y_g|}{y_g} \cdot 100\% \quad (5)$$

其中， y_p 为预测结果， y_g 为真实值。基于偏差率，则预测的准确率定义为：

$$\text{accuracy} = \frac{\sum \text{sgn}(\text{bias} < 20\%)}{N} \cdot 100\% \quad (6)$$

其中， $\text{sgn}(\cdot)$ 表示满足条件时为 1，否则为 0。

对比方法采用 prophet^[4]、XGBoost^[5-6]、带自动调参的 XGBoost、LSTNet^[8]、Informer^[9]，其中 XGBoost 的自动调参选择 hyperopt^[16] 工具。在 prophet 方法里面，周期是一个关键参数，本文采用一周作为一个周期，模型选择加法的分解方式，节假日选择所有的阳历节日和阴历节日，对每一个样本建立一个单独的 prophet 模型。XGBoost 和带自动调参的 XGBoost 的区别主要是自动调参功能不同，其他处理相同，具体为 XGBoost 的输入是先将网络流量进行滑动窗的特征提取，然后再输入 XGBoost 中，得到一个



时间步的预测结果；再通过多次循环的一步预测，可以得到任意时间长度的预测结果；其中，XGBoost 的特征提取方法为不同滑动时间窗的均值、中位数、方差。LSTNet 的配置见文献[8]，Informer 的配置见文献[9]，除了数据输入的补缺值，其他完全按照开源代码。

数据集 A 上的不同模型预测准确率对比见表 2，可以看出，本文方法以 76.34% 的准确率远远领先于其他方法，这主要是因为本文方法综合针对数据的异质性、局部不确定性和突发性的特点，专门设计应对方案。XGBoost 和 XGBoost+hyperopt 的两种方法是次优的方法，这两种方法领先其他方法的一个原因就是，借助专家经验，提取人工特征作为特征输入；XGBoost+hyperopt 方法的准确率比 XGBoost 低，说明自动调参对预测性能提升不大。prophet、LSTNet 对数据的周期更加敏感，更适用于具有周期性的数据集，在流量预测问题中，很容易将数据的波动误识别为周期性的规律，但实际该规律不具有通用性，故容易造成准确率偏低。Informer 虽然是较为领先的时序预测方法，但是没有针对网络流量做出相应改进，对训练集出现过拟合，对数据中的突发性、不确定进行过度学习，无法适用于测试集。

2.3 预测结果分类有效性

本节主要验证模型对于无法预测的分类准确情况，对应网络结构中的分类层。因为实际生活中采集的数据、流量会存在波动，而有些样本因为一些人为的或者外在因素出现流量突变，该部分是无法通过算法预测到的。在扩容场景中，预测结果分类是一个重要的辅助功能，如果分类结果可信，则可以当作扩容建议；如果预测结果分类为不可预测类别，意味着无法准确预测流量，

则预测的流量不作为扩容建议，需要使用其他手段判断。预测结果分类可以让工程人员直接知道流量预测的可信度情况，而不需要工程人员对预测结果进行二次判断。

本文方法的分类模型主要将预测结果分为两类，分别为预测效果好和预测效果差的样本，为方便描述，定义为高样本和低样本。在实际工程中，很难找到通用的准确的定义来完全划分高/低样本，在不同的场景要求下，对高/低样本的要求不同，例如，非核心端口的全网占比多，预测未来流量超过扩容阈值的端口数量随之增多，全部扩容的成本高，因此只有预测效果非常好的端口才可以扩容；而核心端口的数量相对少，但是承载流量更多也更重要，需要适当放松对预测结果分类的要求。在上述例子中，可能存在同样预测效果的样本，在非核心端口被当作低样本，而在核心端口被当作高样本。为了更加准确地描述各个场景，本文采用动态阈值划分分类层 softmax 的输出，最终使用低样本占全部样本的比例表示这一划分。对于 softmax 的分类模型而言，其输出的是分类概率，范围为 0~1，分类概率越接近 1，说明样本越符合该类别。假设 $N_{\text{sort}} = [n_1, n_2, n_3, \dots]$ 表示按照 softmax 输出概率从小到大排序后的所有样本，总样本数量记为 N_{len} ，低样本占比为 rate_{low} ，则：

$$N_{\text{low}} = \lceil N_{\text{len}} \cdot \text{rate}_{\text{low}} \rceil \quad (5)$$

其中， $\lceil \cdot \rceil$ 表示向上取整。排序总样本的前 N_{low} 个样本定义为低样本，计算的准确率为低-准确率；剩余样本定义为高样本，计算的准确率记为高-准确率。

预测准确率随分类的变动如图 6 所示，展示不同低样本占比情况下，高-准确率和低-准确率

表 2 数据集 A 上的不同模型预测准确率对比

方法	prophet	XGBoost	XGBoost+hyperopt	LSTNet	Informer	本文方法
准确率	60.64%	68.78%	68.49%	58.56%	55.51%	76.34%

的变化情况。图6中横坐标低样本占比只控制低样本的占比变化,全部样本的准确率总值不会变,即图6中低样本占比0位置的高-准确率等于低样本占比100%位置的低-准确率,这两个位置分别表示全部样本为高样本和全部样本为低样本。随着低样本占比逐渐增大,如果预测结果分类有效,低样本集合中会有更好预测结果的样本加入,会拉高低样本集合的准确率。举个例子,假设4个样本排序后准确率为[10%,50%,80%,95%],则低样本占比为25%时,低样本集合为第一个样本,对应的低-准确率为10%;低样本占比为75%时,低样本集合为前3个样本,对应的低-准确率为46.6%;可以看出随着低样本占比增大,低-准确率也在上升。从图6明显可以看出,低-准确率随着低样本占比的增大而增大,这说明所有样本按照softmax输出概率从小到大排序是有效的,进而说明模型输出的分类概率是有效的,证明了预测结果分类的有效性。同理,随着低样本占比逐渐增大,如果预测结果分类有效,高样本集合中会减少预测结果差的样本,会拉高高样本集合的准确率。例如,上述例子中,低样本占比为25%时,高样本集合为后面3个,对应高-准确率为75%;低样本占比为75%时,高样本只有最后一个,对应高-准确率为95%。从图6明显可以看出,高-准确率随着低样本占比的增大而增大,同样证明预测结果分类的有效性。

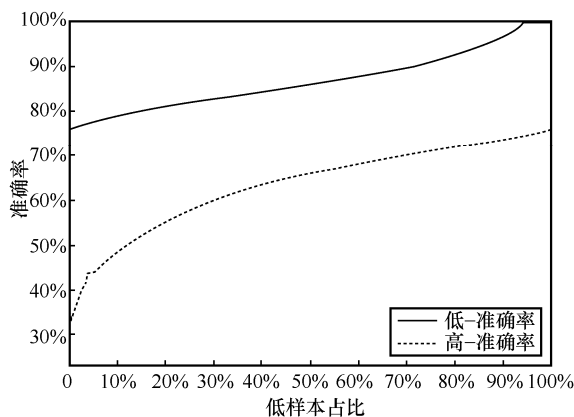


图6 预测准确率随分类的变动

2.4 泛化性能

在扩容应用中,泛化性能对于模型来说是非常重要的,尤其在冷启动期间,即某一城市的数据样本因采集时长等限制不足以支撑训练,需要在其他城市的数据上训练,然后在该城市进行推理。将模型迁移到不同数据集进行推理,可以验证模型的泛化性能。表1展示了两个不同数据集的差异,通过对比可以发现,数据集B的缺失较为严重。除此之外,不同数据集来自不同的地市,而不同地市的生活人群、网元等设备的差异,都会造成流量数据分布规律发生变化。本节选择两个差异较大的数据集,可以更加充分地验证模型迁移的泛化性能。

实验分为两个部分:在数据集A上训练,分别在数据集A和数据集B上测试;在训练集B上训练,分别在数据集A和数据集B上测试。实验选择两个数据集共同存在的流量指标。在不同数据集上的准确率对比见表3。

表3 在不同数据集上的准确率对比

训练情况	测试集为数据集A 的准确率	测试集为数据集B 的准确率
训练数据集A	76.34%	67.58%
训练数据集B	77.08%	68.96%
训练数据集A+修正值	77.40%	68.64%

通过实验可以发现,在数据集A上的准确率比在数据集B上的准确率更高,原因是数据集B有10%左右的数据缺失,以及缺失可能造成数据波动,另外还有数据集B的网元设备较旧。通过实验可以得到如下结论。

- 泛化性能强。本文方法可以适应不同测试集的测试,模型迁移造成的影响较小,在数据集B训练同时在数据集B测试,仅比在数据集A训练同时在数据集B测试的准确率高1.38%。
- 训练样本越多,覆盖时间范围越长,测试效果越好。数据集B的采集范围更加



广，样本数量也更多。数据集 A 在剔除测试数据集的 3 个月长度后，数据集 A 的训练集缺少一个国庆节期间的数据，导致在数据集 A 上的训练无法学习国庆节附近的流量走势，故将数据集 B 作为训练集比将数据集 A 作为训练集效果好，数据集 B 上的训练使得其无论是在数据集 A 还是数据集 B 上做测试都有优势，在数据集 B 上训练比在数据集 A 上训练的准确率平均高 1.06%。

- 修正数据差异后，更容易体现模型的泛化能力强，具体是将训练集 B 与训练集 A 准确率差值 1.06% 作为修正值增加到训练集 A 的结果上，得到表 3 的最后一行。对比表 3 最后 2 行，可以发现修正后的迁移差距很小，训练情况为训练集 A+修正值时，测试集为 B 的准确率为 68.64%，而训练集和测试集同时选择数据集 B 的准确率为 68.96%，二者差值小于 0.5%。综上，模型的泛化能力很强，在扩容场景下，完全可以在一个数据丰富的训练集上训练，然后迁移到其他城市上使用。

3 结束语

本文研究了实际工程中的中长期网络流量预测问题，所提方法采用编码—解码结构，其解码层综合考虑全局特征和局部特征，整个模型将人工特征和神经网络自动特征有效融合，具有高准确率、高泛化性能。通过与其他模型进行对比实验，证明所提出的模型非常适合中长期网络流量预测，即可以克服网络流量的高波动性；通过对比模型在不同数据集上的迁移，证明所提模型具有高泛化性能。未来研究演进方向可以是基于流量预测的路由计算等更多应用场景，以促进 IP 网络更加智能。

参考文献：

- [1] 王帅, 李毅, 陈茂林, 等. 高校不限流量套餐下 LTE 网络扩容[J]. 电信科学, 2020, 36(1): 101-108.
WANG S, LI Y, CHEN M L, et al. LTE network expansion under unlimited traffic packages in colleges and universities[J]. Telecommunications Science, 2020, 36(1): 101-108.
- [2] 黄圣红, 李洲. 一种基于用户体验的 4G 网络扩容策略研究[J]. 电信工程技术与标准化, 2019, 32(4): 42-45.
HUANG S H, LI Z. Research on TD-LTE capacity expansion strategy based on user experience[J]. Telecom Engineering Technics and Standardization, 2019, 32(4): 42-45.
- [3] ZARE M H, MASNADI-SHIRAZI M A. ARIMA model for network traffic prediction and anomaly detection[C]//Proceedings of 2008 International Symposium on Information Technology. Piscataway: IEEE Press, 2008: 1-6.
- [4] TAYLOR S J, LETHAM B. Forecasting at scale[J]. The American Statistician, 2018, 72(1): 37-45.
- [5] CHEN T Q, GUESTRIN C. XGBoost: a scalable tree boosting system[C]//Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM Press, 2016: 785-794.
- [6] MITCHELL R, FRANK E. Accelerating the XGBoost algorithm using GPU computing[J]. PeerJ Computer Science, 2017, 3: e127.
- [7] BANERJEE T, SINHA S, CHOUDHURY P. Long term and short term forecasting of horticultural produce based on the LSTM network model[J]. Applied Intelligence, 2022, 52(8): 9117-9147.
- [8] LAI G K, CHANG W C, YANG Y M, et al. Modeling long- and short-term temporal patterns with deep neural networks[C]//Proceedings of SIGIR'18: the 41st International ACM SIGIR Conference on Research & Development in Information Retrieval. New York: ACM Press, 2018: 95-104.
- [9] ZHOU H Y, ZHANG S H, PENG J Q, et al. Informer: beyond efficient transformer for long sequence time-series forecasting[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2021, 35(12): 11106-11115.
- [10] BI H J, LU L L, MENG Y Z. Hierarchical attention network for multivariate time series long-term forecasting[J]. Applied Intelligence, 2023, 53(5): 5060-5071.
- [11] JIANG W W, LUO J Y, HE M, et al. Graph neural network for traffic forecasting: the research progress[J]. ISPRS International Journal of Geo-Information, 2023, 12(3): 100.
- [12] LI Z L, YU J, ZHANG G W. Dynamic spatio-temporal graph

network with adaptive propagation mechanism for multivariate time series forecasting[J]. Expert Systems with Applications, 2023, 216: 119374.

- [13] 熊伟. 基于突变理论及协同学的网络流量异常检测方法研究[D]. 武汉: 华中科技大学, 2011.

XIONG W. Research on anomaly detection method of network traffic based on catastrophe theory and synergetics[D]. Wuhan: Huazhong University of Science and Technology, 2011.

- [14] 马佳艳, 王萍, 夏伟, 等. 复杂网络下的网络流量预测和预警研究[J]. 计算机与现代化, 2018(1): 102-106.

MA J Y, WANG P, XIA W, et al. Research on network traffic prediction and early warning in complex networks[J]. Computer and Modernization, 2018(1): 102-106.

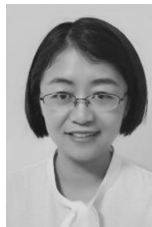
- [15] WANG S P, NIE L S, LI G J, et al. A multitask learning-based network traffic prediction approach for SDN-enabled industrial Internet of Things[J]. IEEE Transactions on Industrial Informatics, 2022, 18(11): 7475-7483.

- [16] BERGSTRA J, YAMINS D, COX D D. Making a science of model search: hyperparameter optimization in hundreds of dimensions for vision architectures[C]//Proceedings of the 30th International Conference on International Conference on Machine Learning - Volume 28. New York: ACM Press, 2013: 115-123.

[作者简介]



高勇(1995-), 男, 移动网络和移动多媒体技术国家重点实验室工程师, 主要研究方向为通信领域的人工智能算法。



陆钱春(1986-), 女, 移动网络和移动多媒体技术国家重点实验室工程师, 主要研究方向为通信网络智能化。



李锋(1978-), 男, 移动网络和移动多媒体技术国家重点实验室工程师, 主要研究方向为通信领域的人工智能控制闭环。