

基于时空特征的移动网络流量预测模型

张 杰¹ 白光伟¹ 沙鑫磊¹ 赵文天¹ 沈 航^{1,2}

(南京工业大学计算机科学与技术学院 南京 211816)¹

(南京邮电大学通信与网络技术国家工程研究中心 南京 210003)²

摘 要 研究表明,历史流量数据可以用于移动网络流量的预测,同时周边区域的流量信息可以提高流量预测的准确性。为此,文中提出一种基于时空特征的移动网络流量预测模型 STFM。STFM 模型利用目标区域及周围区域的历史移动网络流量对目标区域的流量进行预测。其核心思想是,首先利用三维卷积网络(3D CNN)从流量中提取移动网络流量空间上的特征,再利用时间卷积网络(TCN)提取移动网络流量时间上的特征,最后全连接层对提取的特征与实际的流量值建立映射关系,产生预测的流量值。根据实验的验证与分析,STFM 在移动网络流量预测上的标准均方根误差(NRMSE)相比 TCN,CNN 和 CNN-LSTM 分别减少了 28%,21.7%和 10%。因此,STFM 模型能够有效提高移动网络流量预测的准确率。

关键词 移动网络,流量预测,时空特征,卷积网络,全连接层

中图法分类号 TP393 文献标识码 A DOI 10.11896/jsjcx.181102207

Mobile Traffic Forecasting Model Based on Spatio-temporal Features

ZHANG Jie¹ BAI Guang-wei¹ SHA Xin-lei¹ ZHAO Wen-tian¹ SHEN Hang^{1,2}

(College of Computer Science and Technology,Nanjing Tech University,Nanjing 211816,China)¹

(National Engineering Research Center for Communication and Network Technology,Nanjing University of Posts and Telecommunications,Nanjing 210003,China)²

Abstract Research shows that historical traffic data can be used for the prediction of mobile network traffic,and traffic information in surrounding areas can improve the accuracy of traffic prediction. To this end,this paper proposed the traffic prediction model STFM for mobile network based on spatio-temporal features. STFM uses the historical mobile traffic of the target area and surrounding areas to predict the traffic of the target area. Firstly,3D convolutional neural network(3D CNN) is used to extract the spatial features of the mobile network traffic,then time convolutional network(TCN) is used to extract the temporal features of the mobile network traffic. Finally,fully connected layers establish a mapping relationship between the real traffic and extracted features and generate a predicted traffic value. Validation and analysis of experiments show that the STFM reduce the normalized root mean square error (NRMSE) by 28%,21.7% and 10%,compared to TCN,CNN and CNN-LSTM. Consequently,STFM can effectively improve the accuracy of mobile network traffic prediction.

Keywords Mobile network, Traffic prediction, Spatio-temporal feature, Convolution neural network, Fully connected layers

1 引言

随着智能设备的发展,移动网络流量迅猛增长。根据相关机构的预测,到 2021 年,移动网络流量将在 2016 年的基础上增长 7 倍^[1]。同时,移动网络的架构、协议等也在不断发

展,第 5 代移动通信技术已经逐渐成熟^[2]。为了满足移动网络流量增长和移动网络发展的需求,移动网络在资源分配、异常检测等方面需要更加智能化^[2-3]。准确地分析流量和预测流量,是提高移动网络管理智能化的基础。移动网络管理系统可以根据预测的流量动态地配置网络,从而对资源进行更

到稿日期:2018-11-28 返修日期:2019-04-13 本文受国家自然科学基金项目(61502230,61073197,61501224),江苏省自然科学基金项目(BK20150960),江苏省普通高校自然科学研究项目(15KJB520015),南京市科技计划项目(201608009),南京大学计算机软件新技术国家重点实验室资助项目(KFKT2017B21),江苏省研究生科研与实践创新计划项目(KYCX18_1074)资助。

张 杰(1993—),男,硕士生,主要研究方向为网络流量预测等,E-mail:njutzhangjie@njtech.edu.cn;白光伟(1961—),男,教授,博士生导师,CCF 杰出会员,主要研究方向为移动互联网、多媒体网络服务质量、网络系统性能分析和评价等,E-mail:bai@njtech.edu.cn(通信作者);沙鑫磊(1995—),男,硕士生,主要研究方向为智能移动网络、智能路由算法等;赵文天(1995—),男,硕士生,CCF 学生会会员,主要研究方向为绿色移动网络;沈 航(1984—),男,博士,讲师,CCF 会员,主要研究方向为移动互联网、无线多媒体通信协议等。

加合理的分配,以提高能源的利用率^[4]。根据预测的流量值,管理系统也能够进行异常检测,增强移动网络的安全性。

许多文献已经对移动网络流量预测算法进行了研究。文献[5]通过分析 9000 个移动网络基站的数据,将移动网络流量分成可预测部分和不可预测部分,证明了移动网络流量在时间上的相关性,并且提出了基于时间序列分析的预测方法。由于移动网络流量的预测是时间序列问题,文献[6]使用自回归积分滑动平均模型 (ARIMA) 来预测网络流量。ARIMA 模型需要将不平稳的序列数据转为平稳数据,这可能会导致预测准确性的下降。文献[7]证明了移动网络基站的流量会受到周围基站的影响,并基于此提出了模型,从时间和空间的角度预测移动网络流量。

随着 GPU 和优化算法的发展,机器学习和深度学习在近几年得到了飞速的发展。以卷积神经网络 (CNN)、循环神经网络 (RNN) 等网络结构为代表的神经网络在图像识别、自然语言处理等领域取得了显著的效果。为了能够抓取移动网络流量数据中复杂且非线性的特征^[8],研究者使用一些机器学习和深度学习的方法来预测移动网络的流量。文献[9]研究了多个基于人工神经网络的预测算法,包括多层感知机 (MLP)、栈自编码器 (SAE) 和循环神经网络 (RNN),并证明了 RNN 在移动网络流量预测上的效果优于 MLP 和 SAE。文献[10]提出了基于深度信念网络 (DBN) 的预测算法,其取得了较好的预测效果。但文献[9-10]提出的方法都是基于流量数据的时间特征对移动网络流量进行预测,根据文献[7],一个区域的移动网络流量会受到周围区域流量的影响,所以不能仅考虑移动网络流量在时间上的特征。文献[11]提出了基于 CNN 和 RNN 的预测模型,证明了 CNN-RNN 模型的预测效果优于 CNN 和 RNN,但是 CNN-RNN 模型的预测不够准确。

本文第 2 节具体描述了移动网络流量的预测问题;第 3 节详细介绍所提出的模型,并且给出模型中的神经网络结构;第 4 节介绍实验过程并对实验结果进行详细分析;最后对全文进行总结,并且对下一步的研究工作进行展望。

2 移动网络流量预测

移动网络流量的预测是根据指定大小的区域内已知的流量来预测该区域内未来的流量。对于一个较大的区域,将其划分为 $M \times N$ 个区域,如图 1 所示,每个方格代表一个区域。也就是说,预测模型需要根据 $M \times N$ 个区域内从 $t-S-1$ 时刻到 t 时刻的流量来预测 $t+1$ 时刻 $N \times N$ 个区域的流量。

(M,1)	(M,2)	...	...	(M,N)
...	...	...	...	...
...	...	...	...	...
(2,1)	...	...	...	...
(1,1)	(1,2)	...	(1,N-1)	(1,N)

图 1 区域的划分

Fig. 1 Division of area

对于图 1 中的 $M \times N$ 个区域,任意一个时刻的流量情况可以采用矩阵来表示。如式(1)所示, $\mathbf{X}_t$ 表示 t 时刻 $M \times N$

个区域的流量矩阵,其中 $x_t^{(m,n)}$ 表示区域 (m,n) ($m \in \{1, \dots, M\}, n \in \{1, \dots, N\}$) 在 t 时刻的流量值。

$$\mathbf{X}_t = \begin{bmatrix} x_t^{(1,1)} & \dots & x_t^{(1,N)} \\ \vdots & & \vdots \\ x_t^{(M,1)} & \dots & x_t^{(M,N)} \end{bmatrix} \quad (1)$$

已知一个流量序列 $X = (\mathbf{X}_{t-S-1}, \dots, \mathbf{X}_{t-1}, \mathbf{X}_t)$, 如果需要预测 $t+1$ 时刻的流量, 对于一个预测模型来说, 其目标就是求解式(2)。

$$\tilde{\mathbf{X}}_{t+1} = \arg \max_{\mathbf{X}_{t+1}} p(\mathbf{X}_{t+1} | \mathbf{X}_{t-S-1}, \dots, \mathbf{X}_t) \quad (2)$$

其中, $p(\mathbf{X}_{t+1} | \mathbf{X}_{t-S-1}, \dots, \mathbf{X}_t)$ 表示条件概率。

矩阵 $\tilde{\mathbf{X}}_{t+1}$ 可以通过求解其中的所有元素求出, 因此式(2)可以写成:

$$\tilde{\mathbf{X}}_{t+1} = \arg \max_{\mathbf{X}_{t+1}} \prod_{m=1}^M \prod_{n=1}^N p(x_{t+1}^{(m,n)} | \mathbf{X}_{t-S-1}, \dots, \mathbf{X}_t) \quad (3)$$

根据文献[7,12], 区域 (m,n) 在 $t+1$ 时刻的流量 $x_{t+1}^{(m,n)}$ 会受到相近区域流量的影响, 但是能够影响 (m,n) 的区域距离是有限的, 与 (m,n) 距离较远的区域的影响可以忽略不计, 所以当预测 (m,n) 当预测区域的流量时, 并不需要考虑所有的区域。因此, 式(3)可以写成:

$$\tilde{\mathbf{X}}_{t+1} = \arg \max_{\mathbf{X}_{t+1}} \prod_{m=1}^M \prod_{n=1}^N p(x_{t+1}^{(m,n)} | f_{t-S-1}^{(m,n)}, \dots, f_t^{(m,n)}) \quad (4)$$

其中, $f_t^{(m,n)}$ 是 $(2r+1) \times (2r+1)$ 区域的流量情况, r 表示能够影响 (m,n) 的范围。该矩阵表示能够对 (m,n) 产生影响的区域的流量。 $f_t^{(m,n)}$ 如式(5)所示:

$$f_t^{(m,n)} = \begin{bmatrix} x_t^{(m-r,n-r)} & \dots & x_t^{(m-r,n)} & \dots & x_t^{(m-r,n+r)} \\ \vdots & & \vdots & & \vdots \\ x_t^{(m,n-r)} & \dots & x_t^{(m,n)} & \dots & x_t^{(m,n+r)} \\ \vdots & & \vdots & & \vdots \\ x_t^{(m+r,n-r)} & \dots & x_t^{(m+r,n)} & \dots & x_t^{(m+r,n+r)} \end{bmatrix} \quad (5)$$

$t+1$ 时刻的预测值 $\tilde{\mathbf{X}}_{t+1}$ 可通过求解所有的 $\tilde{x}_{t+1}^{(m,n)}$ ($m \in \{1, \dots, M\}, n \in \{1, \dots, N\}$) 而得到。而区域 (m,n) 的流量 $\tilde{x}_{t+1}^{(m,n)}$ 可以通过式(6)得到。

$$\tilde{x}_{t+1}^{(m,n)} = \arg \max_{x_{t+1}^{(m,n)}} p(x_{t+1}^{(m,n)} | f_{t-S-1}^{(m,n)}, \dots, f_t^{(m,n)}) \quad (6)$$

本文提出的模型就是为了求解式(6)。该模型利用前 S 个时刻的 (m,n) 周围区域的流量序列 $f_{t-S-1}^{(m,n)}, \dots, f_t^{(m,n)}$, 预测出 $t+1$ 时刻流量矩阵中心区域 (m,n) 的流量。

3 预测模型

移动网络的流量同时具有时间和空间上的特征, 在提取流量在时间上的特征时, 需要同时充分地获取流量在空间上特征, 以提高模型预测的准确度。因此, 本文提出了基于时空特征的预测模型 STFM (Spatio-Temporal Forecasting Model), 来解决由前 S 个时刻的流量预测 $t+1$ 时刻流量的问题。该模型通过使用三维卷积网络 (3D CNN) 和时间卷积网络 (TCN) 来提取流量在空间和时间上的特征。

3.1 时间卷积网络

在处理时间序列数据时, 基于 RNN 改进的 LSTM 是被广泛使用且效果比较好的神经网络^[13]。LSTM 能够在训练某一层时同时考虑到上一层的输出, 因此在序列问题上得到了广泛应用, 也被用于解决网络流量预测问题^[11]。但是以

LSTM 为代表的 RNN 存在参数较多、训练时间较长等缺点,所以需要考虑采用其他方法来获取移动网络流量的时间特征。

RNN 由于能够有效地利用历史信息,因此能够在时间序列问题上得到广泛应用。卷积网络必须有效地利用历史信息才能解决时间序列问题。TCN 主要使用因果卷积(Casual Convolution)和扩张卷积(Dilated Convolution)^[14]来解决对历史信息的利用问题,因果卷积能够使 TCN 充分利用历史信息。图 2 给出因果卷积和扩张卷积进行卷积操作的过程。从图 2(a)可以看出,因果卷积不断地利用一维卷积对特征进行卷积操作。图 2(a)中所示的因果卷积可以使得每个输出单元利用前 4 个时刻的历史信息。

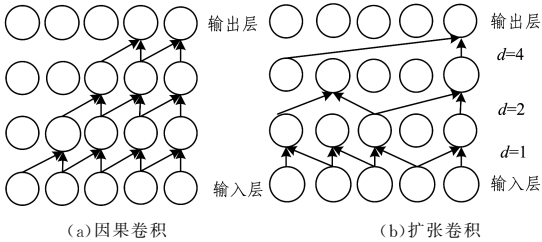


图 2 因果卷积和扩张卷积
Fig. 2 Casual convolution and dilated convolution

虽然因果卷积解决了历史信息的依赖问题,但如果需要的历史信息比较长,将导致网络层数增加,从而使模型变得比较复杂。模型的复杂性可能会导致训练变得困难,甚至出现过拟合的问题。因此,TCN 使用扩张卷积来解决该问题。

扩张卷积如图 2(b)所示。其中, d 表示每一层的扩张率,扩张率指卷积时扩张的大小。 d 随着网络层数的增多呈指数增长,图 2(b)中 $d=2^i$,表示神经网络的层数。扩张卷积所涉及的神元范围随着网络层数的增加呈指数增长。

从图 2 中因果卷积和扩张卷积的过程可以看出,虽然输入层所涉及的神元的范围是一样的,但是在隐藏层中,扩张卷积使用的神元的范围要比因果卷积大。

TCN 在连接神经网络各层时采用了残差连接。残差连接能够解决在网络层数增加时训练结果退化的问题^[15],其公式如下:

f_output = ReLu(x + f(x)) (7)

其中, $f(x)$ 表示卷积操作的输出, x 表示本层的输入。这里采用 ReLu 函数作为激活函数。Relu 函数具有计算简单、收敛快等优点,其计算公式如下:

f_act(x) = max(0, x) (8)

3.2 三维卷积网络

CNN 由于能够有效地抓取空间特征,因此在图像和文本分析等领域被广泛应用。但是,传统的 CNN 是二维的,只能在二维平面上抓取空间特征,无法直接用于视频等三维空间特征的获取。对于移动网络流量序列 $X = (X_{t-S}, \dots, X_{t-1}, X_t)$,如果采用二维 CNN,在提取空间特征时只能对 $X_t (t-S-1 \leq i \leq t)$ 进行卷积操作。3D CNN 相对于二维 CNN,增加了时间维度,可以在进行卷积操作的同时考虑历史信息。在对流量序列 X 进行卷积时,可以同时多个流量矩阵进行卷积操作。三维卷积的公式如下:

H_{ij}^{xyz} = ReLu(b_{ij} + \sum_m \sum_{p=0}^{P_i-1} \sum_{q=0}^{Q_i-1} \sum_{r=0}^{R_i-1} W_{ijm}^{pqr} H_{(i-1)m}^{(x+p)(y+q)(z+r)}) (9)

其中, H_{ij}^{xyz} 表示在 (x, y, z) 位置上的神经元在第 i 层的第 j 个特征图。3D CNN 中用于三维卷积操作的特征图是三维特征图。本文中,初始的特征图是由流量序列 $f_{t-S-1}^{(m,n)}, \dots, f_t^{(m,n)}$ 组成的三维流量矩阵。 W_{ijm}^{pqr} 表示第 i 层的 j 个特征图中位置 (p, q, r) 的卷积核连接到第 $i-1$ 层 m 个特征图的权值。 P_i, Q_i, R_i 表示卷积核的 3 个维度。 b_{ij} 表示偏置。

在卷积层后通常有池化层(Pooling Layer),池化层的输入是前一个卷积层的输出。池化层的作用是通过降低特征面的分辨率来获得具有空间不变性的特征^[16]。池化层起到二次提取特征的作用,它的每个神经元对局部接受域进行池化操作。与 2D 卷积神经网络相同,常用的三维池化方法包括最大池化、平均池化等。本文所提出的 STFM 模型使用三维最大池化。三维最大池化的公式如下:

H_{x,y,z} = \max_{0 \leq i \leq S_1, 0 \leq j \leq S_2, 0 \leq k \leq S_3} (U_{x * s + i, y * t + j, z * r + k}) (10)

其中, U 为池化层的输入; H 为池化层的输出; s, t 和 r 为池化操作的采样步长; S_1, S_2 和 S_3 表示池化操作的窗口大小。

3.3 基于时空特征的预测模型

为了能够利用 TCN 和 3D CNN 抓取移动网络流量在时间和空间上的特征,从而准确地预测移动网络流量,本文设计了如图 3 所示的模型。STFM 模型主要由 Spatial Learning Phrase, Temporal Learning Phrase 和 Output Phrase 这 3 个部分组成,它们分别表示空间特征提取层、时间特征提取层和输出层。

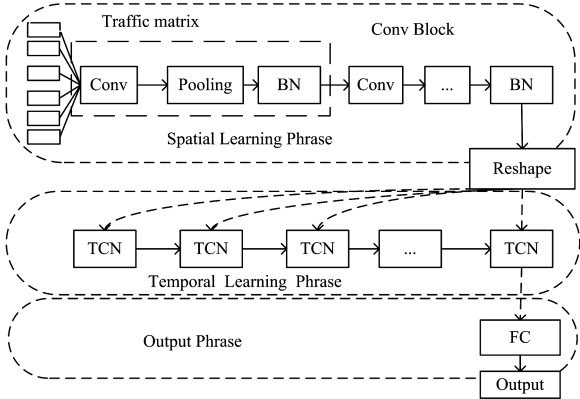


图 3 基于时空特征的预测模型
Fig. 3 Forecasting model based on spatio-temporal features

空间特征提取层(Spatial Learning Phrase)的结构如图 4 所示,其主要由卷积单元构成,也就是图 3 和图 4 中的 Conv Block。每个卷积单元都由卷积层(Conv 层)、池化层(Pooling 层)以及批归一化层(BN 层)组成。空间特征提取层的主要作用是获取移动网络流量的空间特征。

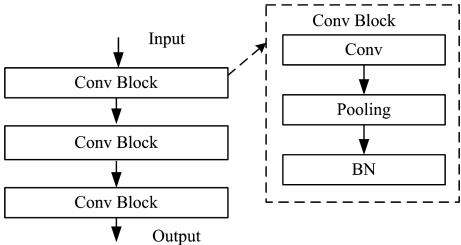


图 4 空间特征提取层的结构
Fig. 4 Architecture of spatial learning phrase

卷积层由三维卷积组成,三维卷积的计算公式如式(9)所示。池化层采用最大池化,其计算公式如式(10)所示。卷积单元的最后一层是批归一化层,其作用是调整数据的分布,减少模型出现过拟合的情况,从而增强模型的非线性表达能力^[17]。批归一化(Batch Normalization)的计算公式如下:

$$y_i = \gamma \frac{x_i - \mu}{\sqrt{\sigma^2 + \epsilon}} + \beta \tag{11}$$

其中, μ 和 σ^2 表示一个训练批次中训练数据的均值和方差; x_i 表示每个训练批次中的一个数据; γ 和 β 是模型学习的参数; ϵ 的作用是防止分母为零,可以取为 1×10^{-8} 。

空间特征提取层的输入是由网络流量组成的三维矩阵 $(f_{i-s-1}^{(m,n)}, \dots, f_t^{(m,n)})$,其中 $f_t^{(m,n)}$ 为二维矩阵,如式(5)所示。空间特征提取层的输出也是三维矩阵 $(y_{t-s-1}(a,b), \dots, y_t(a,b))$,其中 $y_t(a,b)$ 表示维度为 (a,b) 的二维矩阵。

时间特征提取层(Temporal Learning Phrase)的主要作用是提取移动网络流量的时间特征。时间特征提取层主要由TCN构成,而TCN的主要组成单元是扩张卷积(见图2(b))。扩张卷积的计算公式如式(12)所示:

$$F(s) = \sum_{i=0}^{k-1} f(i) \cdot (X_{s-d \cdot i}) \tag{12}$$

其中, k 表示卷积核的大小, f 表示卷积核, d 表示扩张系数, X 是输入序列, s 表示序列中的第 s 个元素。

TCN层的连接采用残差连接,如图5所示。

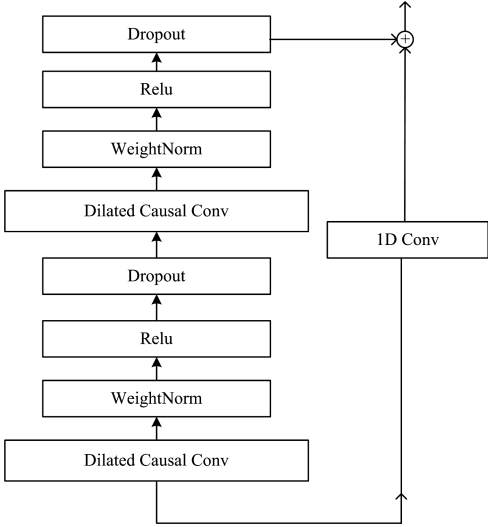


图 5 TCN 中的残差连接

Fig. 5 Residual connection of TCN

图5中的Dilated Causal Conv是扩张卷积层,每个残差块中都有两个扩张卷积,在每个扩张卷积后WeightNorm层会进行权重的归一化。经过两次扩张卷积后,将输出与输入相加,如式(7)所示。因为经过卷积操作后输出向量的形状与输入不同,所以需要输入向量进行一维卷积,以保证输入和输出能够相加。

时间特征提取层的输入是空间特征提取层的输出,也就是三维矩阵 $(y_{t-s-1}(a,b), \dots, y_t(a,b))$ 。因为时间特征提取层的输入是二维矩阵,所以需要加入Reshape层将空间提取层的输出矩阵转化为二维矩阵,如图3所示。时间特征提取层的输出是一维矩阵。

输出层(Output Phrase)的作用是得到最终的预测结果。时间特征提取层的输出作为此层的输入。输出层由全连接层组成。全连接层的计算公式如下:

$$Y_{t+1} = g(W_{t+1}X_{t+1} + b_{t+1}) \tag{13}$$

其中, W_{t+1} 和 b_{t+1} 表示模型需要学习的参数; X_{t+1} 表示时间特征层的输出; $g(\cdot)$ 表示激活函数,本文中的激活函数采用ReLU函数。

预测流程如算法1所示。

算法 1 STFM 的预测流程

输入:历史流量矩阵 $f=(f_{t-s-1}^{(m,n)}, \dots, f_t^{(m,n)})$

输出:预测的流量值 $\hat{x}_{t+1}^{(m,n)}$

1. For $i=1,2,\dots,n$ //n 表示卷积单元数

2. $C=Conv_i(f)$ //三维卷积操作

3. $C=MaxPooling_i(C)$

4. $C=BN_i(C)$

5. End For

6. $T=C$

7. For $i=1,2,\dots,m$ // m 表示 TCN 的层数

8. $T_i=T$

9. For $j=1,2$ //两次扩张卷积

10. $T=Dilated_j(T)$ //扩张卷积

11. $T=WeightNorm_j(T)$ //权重归一化

12. $T=ReLU(T)$

13. $T=Dropout(T)$

14. End For

15. $T=ReLU(Conv(T_i)+T)$ //残差连接

16. End For

17. $\hat{x}_{t+1}^{(m,n)}=ReLU(W_{t+1}T+b_{t+1})$

18. Return $\hat{x}_{t+1}^{(m,n)}$

在使用STFM模型预测之前,需要进行训练,以得到较优的预测模型。训练过程如算法2所示。

算法 2 STFM 的训练过程

输入:历史流量矩阵 $\{f_t^{(m,n)}|t=0,1,2,\dots,s-1\}$,时间特征模块的序列长度 S,Mini Batch 的大小 b

输出:训练得到的预测模型

1. $D=\emptyset$ // D 表示训练集,初始为空

2. For all available t ($S+1 \leq t \leq s-1$)

3. $Z_x=[(f_{t-s-1}, \dots, f_t)]$

4. $Z_y=[f_{t+1}]$

5. $(Z_x,Z_y) \rightarrow D$

6. End For

7. 初始化模型中所有参数 θ //开始模型训练

8. Repeat

9. 从训练集 D 上选取大小为 b 的 batch D_b

10. 在 D_b 上最小化损失函数来找到参数 θ 的最优解

11. Until $L(\theta)$ 达到最小值

12. Return 预测模型

算法2中步骤2—步骤6的作用是从数据集中生成每个时间段的训练样本。在获得训练样本后,从训练集中选取batch,然后通过反向传播算法找到最优的参数。模型训练时,需要指定损失函数作为训练的目标。本文采用均方误差(MSE)作为损失函数。均方误差的计算为:

$$L(\theta) = \|Y_t - \hat{Y}_t\|_2^2 \tag{14}$$

其中, Y_t 和 $\hat{Y}_t$ 分别表示实际值和预测值, θ 表示模型需要学习的参数。STFM 模型的训练就是最小化 $L(\theta)$ 的过程。

4 实验与结果分析

4.1 数据集

本文实验所用的数据集是意大利米兰地区的移动网络数据^[18]。该数据集收集了米兰地区 2013 年 11 月 1 日到 2014 年 1 月 1 日的移动网络数据。在该数据集中,米兰地区被划分为 100×100 的区域,每个区域有唯一的编号,即图 1 中 $M=100$ 且 $N=100$ 。每个区域表示 $235\text{ m} \times 235\text{ m}$ 的实际区域。数据集中的每一条记录包含区域编号(Square ID)、记录的开始时间(Begin Time)以及此记录的流量信息(Traffic)。每条记录中的流量是开始时间后 10 min 的流量信息。记录的示例如表 1 所列,其中, $Square\ ID$ 是一个整数, $1 \leq Square\ ID \leq 10\,000$,每个 ID 对应于图 1 中的一个区域。

表 1 米兰地区的移动网络流量信息

Table 1 Mobile traffic information in milano

Square ID	Begin Time	Traffic
1	2013-11-01 00:00:00	11.028366381681023
1	2013-11-01 00:10:00	11.127100875673673

4.2 实验参数

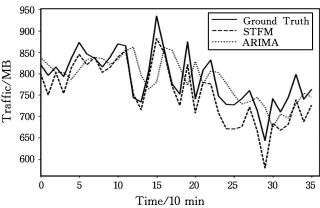
对于输入的流量序列 $f_t^{(m,n)}, \dots, f_t^{(m,n)}$, S 取值为 8, r 取值为 7,即模型的输入是形状为 $6 \times 15 \times 15$ 的矩阵。

STFM 所用的参数如表 2 所列。

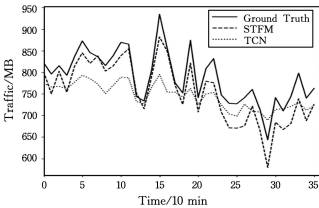
表 2 模型参数

Table 2 Parameters of model

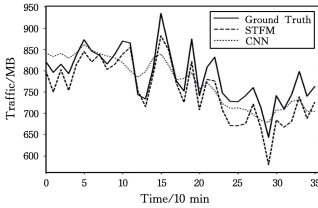
网络层	Kernels	Kernel Size	Stride
Conv1	32	3,5,5	1,2,2
Pooling1		1,2,2	1,1,1
Conv2	32	3,5,5	1,2,2
Pooling 2		1,2,2	1,1,1
Conv3	64	3,5,5	2,2,2
Pooling 3		1,2,2	1,1,1



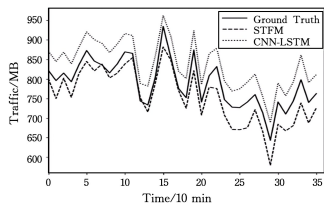
(a) ARIMA 与 STFM 的对比



(b) TCN 与 STFM 的对比



(c) CNN 与 STFM 的对比



(d) CNN-LSTM 与 STFM 的对比

图 6 各模型的流量预测结果

Fig. 6 Traffic prediction results of models

图 6 中,Ground Truth 表示实际值。可以看出,STFM 的拟合曲线相比其他 4 个模型的曲线更加准确。特别是在高峰时段时,如图 6 中的 15 时刻,STFM 能够较好地反映移动网络流量的变化规律,而 ARIMA 对高峰时段的流量预测性能较差。CNN 和 TCN 能够预测出高峰时段的流量变化规律,但是预测的流量值相比实际值误差较大。STFM 对高峰时段流量值的预测效果比 CNN-LSTM 差。

为了衡量模型的预测性能,采用 NRMSE 作为衡量指标。

表 2 中,Kernels 表示卷积操作中滤波器的输出数量;Kernel Size 表示窗口的大小;Stride 在卷积中表示卷积的步长,在池化层中表示特征图缩小的比例。池化层采用 Max Pooling。

在 TCN 层中,扩张系数设为(1,2,4,8),一维卷积滤波器的数量设为 24。

ARIMA 方法中,选择参数 $p=5, d=1, q=3$ 。 p 表示依赖的历史值数量, d 表示差分的阶数, q 表示依赖过去最近的 q 个历史的预测误差值。

表 3 列出了 CNN 的部分参数,表 4 列出了 LSTM 的部分参数。

表 3 CNN 参数

Table 3 Parameters of CNN

网络层	Kernels	Kernel Size	Stride
Conv	64	3,5,5	1,1,1
Max Pooling		1,2,2	1,2,2
Conv	128	3,5,5	1,1,1
Avg Pooling		2,2,2	2,1,1
Conv	64	3,5,5	1,1,1
Avg Pooling		2,2,2	2,1,1

表 4 LSTM 参数

Table 4 Parameters of LSTM

	Cells	Steps
LSTM	256	6

网络层为 LSTM 时,Cells 表示 LSTM 各层的神经元数,Steps 表示步长。

4.3 实验结果与分析

首先选取区域(49,67)作为研究对象,分别使用 STFM, ARIMA,TCN,CNN 和 CNN-LSTM 预测区域(49,67)的流量情况。图 6 表示 STFM 与 ARIMA,TCN,CNN,CNN-LSTM 进行对比的实验结果。图中横轴以 10 min 作为单位(如横轴上的 5 表示 50 min)。

NRMSE 的计算公式为:

$$NRMSE = \frac{1}{\bar{d}} \sqrt{\frac{\sum_{k=1}^N (\bar{d}_k - d_k)^2}{N}} \tag{15}$$

其中, $\bar{d}_k$ 表示预测值, d_k 表示实际值, N 表示用于评估的数据量, $\bar{d}_k$ 表示平均值。

我们计算了 STFM,ARIMA,CNN,TCN 和 CNN-LSTM 的 NRMSE 值,结果如表 5 所列。

表 5 不同模型的 *NRMSE*

Table 5 *NRMSE* in different models

方法	<i>NRMSE</i>
STFM	0.18
ARIMA	0.31
TCN	0.25
CNN	0.23
CNN-RNN	0.20

由表 5 可以看出,上述算法中,使用 ARIMA 预测的准确度是最差的;使用 TCN 神经网络模型预测的准确度要优于 ARIMA,但是 TCN 模型只考虑了网络流量时间上的特征,因此预测的准确率较低;使用 CNN-RNN 模型预测的准确度要优于 CNN 或者 TCN,但 CNN-RNN 模型预测的准确率仍然低于 STFM 模型。

结束语 本文提出了一种基于移动网络流量时空特征的流量预测模型 STFM。该预测模型采用 TCN 和 3D CNN 提取移动网络流量的时间和空间特征,进而实现对流量的预测。实验表明,STFM 具有较好的预测能力。但 STFM 只对短期流量进行了预测,下一步我们将对预测长期流量进行研究。

参 考 文 献

[1] Cisco Visual Networking Index:Forecast and Methodology 2016—2021[EB/OL]. <https://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni-complete-white-paper-c11-481360.html>.

[2] SUNDARESAN K. 5G:An Evolution towards a Revolution [C] // Proceedings of ACM International Conference on Mobile Computing and Networking, New Delhi,India,2018:659-659.

[3] IMRAN A,ZOHA A. Challenges in 5G:How to Empower SON with Big Data for Enabling 5G [J]. IEEE Network, 2014, 28(6):27-33.

[4] NIU Z,WU Y,GONG J,et al. Cell Zooming For Cost-Efficient Green Cellular Networks [J]. IEEE Communications Magazine, 2010,48(11):74-79.

[5] XU F,LIN Y,HUANG J,et al. Big Data Driven Mobile Traffic Understanding and Forecasting:A Time Series Approach [J]. IEEE Transactions on Services Computing, 2016, 9(5):796-805.

[6] SHU Y,YU M,LIU J,et al. Wireless Traffic Modeling and Prediction Using Seasonal ARIMA Models[C] // Proceedings of IEEE International Conference on Communications. Anchorage: IEEE,2003:1675-1679.

[7] LI R,ZHAO Z,ZHENG J,et al. The learning and prediction of

application-level traffic data in cellular networks [J]. IEEE Transactions on Wireless Communications, 2017, 16(6):3899-3912.

[8] ZHANG C,PAUL P. Long-term mobile traffic forecasting using deep spatio-temporal neural networks[C]// Proceedings of ACM International Symposium on Mobile Ad Hoc Networking and Computing, Angeles:ACM,2018:231-240.

[9] OLIVEIRA T P,BARBAR J S,SOARES A S. Computer Network Traffic Prediction:A Comparison between Traditional and Deep Learning Neural Networks [J]. International Journal of Big Data Intelligence,2016,3(1):28-37.

[10] NAREJO S,PASERO E. An Application of Internet Traffic Prediction with Deep Neural Network [J]. Multidisciplinary Approaches to Neural Computing,2018,69(1):139-149.

[11] HUANG C W,CHIANG C T,LI Q. A Study of Deep Learning Networks on Mobile Traffic Forecasting[C] // Proceedings of IEEE Personal, Indoor, and Mobile Radio Communications. Montreal:IEEE,2017:1-6.

[12] LI R,ZHAO Z,ZHOU X,et al. The Prediction Analysis of Cellular Radio Access Network Traffic:From Entropy Theory to Networking Practice [J]. IEEE Communications Magazine, 2014,52(6):234-240.

[13] FU R,ZHANG Z,LI L. Using LSTM and GRU neural network methods for traffic flow prediction[C] // Proceedings of IEEE Youth Academic Annual Conference of Chinese Association of Automation. Wuhan:IEEE Press,2016:324-328.

[14] YU F,KOLTUN V. Multi-scale context aggregation by dilated convolutions[C] // Proceedings of International Conference on Learning Representations. San Juan:ICLR,2016.

[15] HE K,ZHANG X,REN S,et al. Deep residual learning for image recognition[C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Nevada:IEEE,2016:770-778.

[16] ZHOU F Y,JIN L P,DONG J. Review of Convolutional Neural Network[J]. Chinese Journal of Computers,2017,40(6):1229-1251.

[17] IOFFE S,SZEGEDY C. Batch normalization:accelerating deep network training by reducing internal covariate shift[C]// Proceedings of International Conference on International Conference on Machine Learning. Lille:ACM,2015:448-456.

[18] GIANNI B,MARCO D N,ROBERTO L,et al. A multi-source dataset of urban life in the city of Milan and the Province of Trentino[J]. Scientific Data,2015,2:150055.