DOI:10.19651/j.cnki.emt.2108286

基于注意力机制的 Bi-GRU 内容流行度预测算法*

许 阅 刘光杰
(南京信息工程大学电子信息工程学院 南京 210044)

摘 要: 内容资源流行度预测是内容分发网络提高缓存与调度效率的主要依据之一。针对当前流行度预测算法特征表征能力和适应性较差,准确率低等不足,提出一种基于深度学习的内容资源流行度预测算法,该算法基于融合注意力机制的双向 GRU 模型可以更好地挖掘资源访问历史中蕴含的信息及其相关性,提高特征提取的效率和质量,并具有更为包容的泛化能力。相关不同数据集上的实验结果表明该算法各项指标均优于已有的多种主流算法,且准确率高达 96.20%和 98.03%。

关键词: 双向 GRU;注意力机制;流行度预测;内容分发网络
中图分类号: TP393.0 **文献标识码:** A **国家标准学科分类代码:** 510.4030

Bi-GRU content popularity prediction algorithm based on attention mechanism

Xu Yue Liu Guangjie
(College of Electronical and Information Engineering, Nanjing University of Information Science and Technology, Nanjing 210044, China)

Abstract: Content resource popularity prediction is one of the main basis for content delivery network to improve the efficiency of caching and scheduling. In view of the poor feature representation ability, adaptability and low accuracy of current popularity prediction algorithms, this paper proposes a content resource popularity prediction algorithm based on deep learning. The algorithm is better based on a two-way GRU model of the fusion attention mechanism, which can better mine the information contained in the resource access history and its correlation, improve the efficiency and quality of feature extraction, and has a more tolerant generalization ability. The experimental results on different data sets show that the various indicators of the algorithm are better than the existing mainstream algorithms, and the accuracy rates are as high as 96.20% and 98.03%.

Keywords: bidirectional GRU;attention mechanism;popularity prediction;content delivery network

0 引 言

用户生成内容(UGC)营销正在以显著的速度增长,每天都有数千亿的内容被用户制作和上传。随着互联网内容的爆发式增长[1],网络流量大幅度增加,对内容的访问需求量也在增加。传统的以主机为中心的传输网络已经无法满足暴涨的并发用户需求。内容分发网络(CDN)的提出和应用有效缓解了这些问题[2]。为减少用户端的响应时间,提升服务质量(QoS)[3],诸如 YouTube、抖音、爱奇艺等内容服务商利用缓存策略和 CDN 技术,通过提前将热点内容资源缓存至离用户更近的数据中心。由于 CDN 缓存服务器的容量有限,如何在有限的条件下更加合理地对内容进行缓存、更新是内容分发边缘计算面临的主要问题,其关键是对网络中的内容资源流行度进行精准预测[4]。

按照预测采用的技术手段,目前已有的方法大致分为传统流行度预测算法和基于深度学习的流行度预测算法两种。传统预测方法又可以分为特征驱动方法和点过程建模方法两类。其中,特征驱动方法通过提取人工定义特征信息(如类别、上传者的社交网络和上下文信息等),将流行度预测问题转换为回归或分类问题,并引入机器学习等方法进行求解。Hassine 等[5]研究了多种 ARMA 模型的预测能力,揭示了 ARMA 模型难以对内容提供最优预测;朱琛刚等[6]利用主成分分析法(PCA)对所选特征实施降维处理,再基于随机森林(random forests,RF)算法构建了流行

收稿日期:2021-11-07
* 基金项目:国家自然科学基金(U61772281,61702235,61931004)、中央高校基础研究基金(30918012204)项目资助

度预测模型,实现了对视频流行度的快速预测;徐冬等[7]利用基于 logistic 机器学习算法计算用户行为信息,设计了适用于消费数据较为稀疏的场景的内容预测,难以适应长历史数据下的内容预测;方元武等[8]提出了一种基于 ARIMA 和多元对数回归的混合内容资源的流行度预测算法,既可以适应稀疏数据的局部性特征也能适应长历史数据的季节性变化特征。

基于点过程的预测方法是将信息传播过程建模为用户转发分享行为的到达过程,将流行度预测转化为到达过程的速率函数的计算。Gao 等[9]利用不同的时间衰减函数,扩展增强泊松过程用于内容资源流行度预测;Bao 等[10]研究了用户活动等候时间的分布情况,使用自激励 Hawkes 过程对 YouTube 视频的浏览量进行建模,其中自激励过程认为过去时刻的每一次转发分享对当前时刻的速率函数都有一次新的激励;Mishra 等[11]融合了前两种方法,对信息传递过程使用自激励过程建模,学习参数并手工提取特征,合并后运用回归模型预测内容地流行度,与前两者相比,取得了较好的效果。

上述的传统流行度预测方法模型依赖于设计的启发式特征质量,或者依赖于某些特定的统计假设,所设计的算法具有明确的显式机理,但是受限于特征维度和统计模型的表示能力,在一些样本和场景中虽然具有较好的准确率,但是在开放的实际工程场景中,由于方法的适应性较差,预测准确率始终难以令人满意。基于深度学习来构建基于大量业务数据的预测模型能更好地挖掘数据中的潜在规律,受到了学术界和工业界的广泛关注。陈亮等[12]联合用户行为信息和视频内容数据,提出一种基于深度信念网络的视频热度预测方法,充分利用了深度神经网络对特征信息的选择优势;Nia 等[13]在深度信念网络的基础上加入了内容聚类,,取得了更有竞争力的实验结果;Qiang 等[14]基于 LSTM 网络能够有效的捕捉时序数据的演化趋势,利用 LSTM 网络对内容的流行动态和流行度进行建模和预测,取得了较好的效果;宋旭鸣等[15]提出一种三层堆叠的 LSTM 预测模型,选取内容文件流行度特征后经过特征提取层,非线性映射层和流行度预测层,最终得到了较单层 LSTM 模型效果更佳的预测输出;鲍鹏等[16]利用图注意力机制学习内容不同阶段的级联结构,再放入时序卷积网络获取级联传播的时序特征,最后通过全卷积层映射出在线内容流行度的变化;武维等[17]融合了文本信息和时序信息,使用 Attention-LSTM 和神经网络因子分解机(NFM)分别挖掘时序信息和文本特征,采用 concatenate 方法将两者结合得到最终的预测模型。

已有研究结果表明基于 LSTM 能较为准确地内容预测资源的流行度,但是依然存在模型训练准确率不高且算法适应性不足等问题。为了进一步提高算法的综合性能,本文采用更易于训练的 GRU 作为基础的模型部件,引入双向(Bidirectional)循环机制和注意力(Attention)机制强化学习能力,以充分利用历史信息和未来信息,并进一步突出了影响流行度变化的信息,使得流行度预测准确率更高,误差更小,适应性更强。本文的主要贡献如下:

1)为了更好地度量每个视频内容资源在某段时间内的受欢迎程度,本文选取内容资源的请求次数作为流行度评价标准,流行度预测问题即为时间序列预测问题,定义未来某时间段内访问量预测值为流行度增量。

2)为了充分挖掘学习时序相关性,进一步突出影响流行度的关键信息,提出在 GRU 网络的基础上加入双向循环机制和注意力机制,模型将编码化的流行度特征向量输入 Bi-GRU 层,再经过注意力机制层,大大提高了流行度预测的准确率。

3)本文在真实公开的 Facebook 网站数据集和 YouTube 视频数据集上进行预测实验,并选择了 5 种已有预测模型作对比试验,结果表明该模型在 4 个评价指标上都有较好表现。

1　内容资源流行度预测算法

1.1　问题描述

设内容资源集合为 $\boldsymbol{C}=\{c_1,c_2,c_3,\cdots,c_n\}$,时间 t 表示为 $t=1,2,3,\cdots,T$;与文献[2,13,15]类似,本文将内容资源流行度视为时间序列变量,选择其请求次数即访问量 $\boldsymbol{V}=[v_1,v_2,v_3,\cdots,v_T]$作为流行度评价标准 P_C。内容资源流行度预测模型可以概括性地表示为:

$$P_{c_k}=g(P_{c_k}(1),P_{c_k}(2),P_{c_k}(3),\cdots,P_{c_k}(t-1)),k=1,2,3,\cdots,n \tag{1}$$

流行度预测问题本质上即为时间序列预测问题,即给定某内容资源 c_k 在观测窗口[1,$t-1$]内的流行度大小,来预测 c_k 下个时间段的预计访问量(即其流行度增量)。

1.2　模型框架

本文提出的基于注意力机制的 Bi-GRU(BiGRU-Attention)模型分为 3 个组成部分:流行度特征向量输入层、预测算法层和输出层。其中,预测算法层由 Bi-GRU 层、Attention 层和 Dense 层 3 个部分组成,本文的模型框架如图 1 所示。

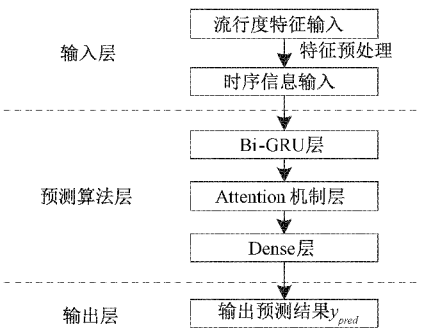

图 1　模型总结构

算法的输入是不同时刻经过数据预处理的内容资源历

史流行度即历史访问次数，不同时刻的历史流行度与未来内容资源流行度的相关联程度是不同的。利用 Bi-GRU 层可充分引入时序信息的双向知识，更有利于特征的提取与识别。由于内容资源流行度的动态性是复杂且非平稳的，预测过程中更应该关注变化较为剧烈的环节，发现其中潜在的规律。反映到深度学习中，采用注意力机制(Attention)层来表征这种某段时序的重视，对多个重要时段的数据进行加权求和。输出层的输入为 Attention 机制层的输出，最终得到预测结果 y_{pred}。具体地，各层的设计及其功能如下。

1)输入层

输入层主要是对数据集进行预处理，使数据成为能够被 Bi-GRU 层接受并处理的特征向量模式。预处理步骤如下：

(1)读取数据集并对数据进行清洗，选取本文需要的特征。

(2)对流行度即内容资源访问次数特征数据进行归一化处理。数据归一化能够消除数据量纲影响，提高网络模型的收敛效率以及模型的精度。常用的数据归一化方式为离差标准化，是对原始数据的线性变换，将结果映射到[0,1]之间。转换公式如下表示，式(2)中 min$p$ 为各维特征 p 的最小值；max$p$ 为各维特征 p 的最大值：

$$p' = \frac{p - \min p}{\max p - \min p} \tag{2}$$

(3)划分训练集和测试集，具体划分形式将在实验部分详细介绍。

经过上述 3 个步骤之后，采集到的流行度特征数据(时序信息)就变成预测算法层直接接受并处理的特征向量形式。

2)Bi-GRU 层

门控循环单元(gate recurrent unit，GRU)是一种带有循环结构的神经网络，也是长短期记忆(long short term memory，LSTM)模型的变体。LSTM 网络在捕捉事件变化过程方面很有优势，如交通事故预测[18-19]、股票走势预测[20]等。GRU 在保持 LSTM 效果的同时简化其结构，且计算方便，训练速度更快。在 GRU 中只有更新门和重置门，更新门决定了时刻 t 之前的信息对当前时刻 t 产生的影响，重置门决定怎样将当前时刻 t 的输入信息与 t 时刻之前的累积信息相结合。它可以对多个时间的输入特征进行自动提取，不同时刻的数据可以共享相同的权值矩阵。由于本身独特的自循环机制，隐层特征可以在不同时刻间单向传递，并记录下先前时刻的“历史数据”，再融合历史数据与新数据得到当前时刻的隐藏层特征值。与 LSTM 一样，GRU 亦采用“门”结构克服了短时记忆的影响，使得信息有选择性的在隐藏层中传递，在记忆重要信息的同时还可以改善循环神经网络(RNN)中存在的“梯度消失”问题。单向 GRU 网络的结构如图 2 所示。

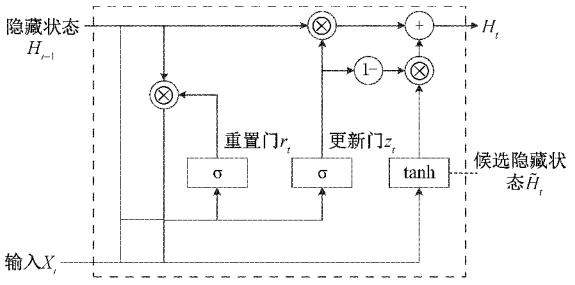

图 2　GRU 结构

$$\boldsymbol{z}_t = \sigma(\boldsymbol{W}_z \boldsymbol{x}_t + \boldsymbol{U}_z \boldsymbol{h}_{t-1} + \boldsymbol{b}_z) \tag{3}$$

$$\boldsymbol{r}_t = \sigma(\boldsymbol{W}_r \boldsymbol{x}_t + \boldsymbol{U}_r \boldsymbol{h}_{t-1} + \boldsymbol{b}_r) \tag{4}$$

$$\widetilde{\boldsymbol{h}_t} = \tanh(\boldsymbol{W}_h \boldsymbol{x}_t + \boldsymbol{U}_h (\boldsymbol{h}_{t-1} \otimes \boldsymbol{r}_t) + \boldsymbol{b}_n) \tag{5}$$

$$\boldsymbol{h}_t = (1 - \boldsymbol{z}_t) \otimes \boldsymbol{h}_{t-1} + \boldsymbol{z}_t \otimes \widetilde{\boldsymbol{h}_t} \tag{6}$$

$$\boldsymbol{h}_t = [\overrightarrow{\boldsymbol{h}_t}, \overleftarrow{\boldsymbol{h}_t}] \tag{7}$$

其中，$\boldsymbol{z}_t$ 和 $\boldsymbol{r}_t$ 分别为 GRU 的更新门和重置门，更新门控制了上一时刻状态信息传递至当前时刻的程度，重置门则控制上一时刻的状态信息被遗忘的程度。$\boldsymbol{W}_z$，$\boldsymbol{W}_r$，$\boldsymbol{W}_h$ 和 $\boldsymbol{U}_z$，$\boldsymbol{U}_r$，$\boldsymbol{U}_h$ 分别为神经元当前时刻的输入权重矩阵和循环输入的权重矩阵，$\boldsymbol{b}_z$，$\boldsymbol{b}_r$，$\boldsymbol{b}_h$ 为偏置向量。σ 为 sigmoid 非线性激活函数，用于增强模型对于非线性数据的处理能力，其表达式为 $\sigma(x) = 1/(1+e^{-x})$。$\otimes$表示哈达玛积即对应元素相乘；$\tanh(x) = (e^{x} - e^{-x})/(e^{x} + e^{-x})$ 为一非线性映射函数。首先，通过上一时刻的隐藏状态信息 $\boldsymbol{h}_{t-1}$ 和当前时刻的节点输入 $\boldsymbol{x}_t$ 来获取 2 个门控的状态。得到门控信号后，利用重置门来获取遗忘后的状态 $\boldsymbol{h}_{t-1} \otimes \boldsymbol{r}_t$；然后与当前时刻的输入 $\boldsymbol{x}_t$ 相加并通过 tanh 激活；最后用更新门对当前节点的输入选择记忆。

传统的 GRU 是单向的，数据只能沿着一个方向处理，重要信息容易丢失，网络只能结合当前时刻的输入及之前时刻的隐藏状态信息计算新的隐藏层状态信息 $\boldsymbol{h}_t$。为了充分利用上下文信息，本文采取双向 GRU 网络结构，可以同时处理双向的顺序信息，双向 GRU 对序列分别采用前向和反向计算得到两个不同的隐藏层状态，再将两个向量相加得到最终的结果，充分利用序列信息，有利于特征提取的进行。在数据量较少时，双向 GRU 的表现也会比双向 LSTM 模型更加优异。

3)Attention 层

注意力(Attention)机制从本质上讲类似于人类的选择性视觉注意力机制，其主要目的是从大量的信息中选择出对当前任务更关键的信息。通过计算不同时刻 Bi-GRU 网络中的输出特征向量权重，突出对预测结果中占比更大的特征，从而使得整个模型表现出更好的性能。目前 Attention 机制已经广泛应用于如自然语言处理中的机器翻译、图像处理、语音识别、情感分析[21-22]、目标检测[23]等领域。在本文的流行度预测中，神经网络在训练过程中通过 Attention 机制来选择关注一些关键的特征，可以增强时

间序列信息中时间相隔较远信息之间的相关性。注意力机制的核心是权重系数。在注意力机制上,预测任务和自然语言处理的思想基本一致,都是为了加强某个时间节点信息与其他节点信息时间的关联性。向神经网络中引入注意力机制,可以选取一部分的关键信息进行处理,提升整个网络的效率,提高预测的精度。Attention 机制模型如图 3 所示。

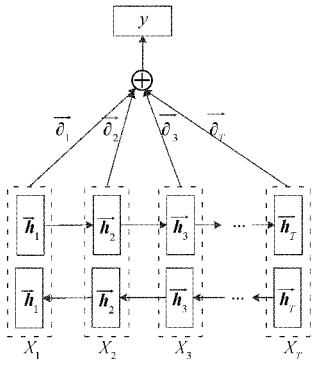

图 3　Attention 结构

其中 $\overrightarrow{h_t}$ 为 Bi-GRU 网络输出的第 t 个特征向量,经过注意力机制层后得到初始的状态向量 $\overrightarrow{s_t}$,然后再和权重系数 ∂_t 对应相乘,累加求和得出最终的输出向量 y,最后将 y 与 Dense 层整合到一起形成输出值输入至输出层。计算过程如下:

$$\overrightarrow{e_t} = \tanh(\overrightarrow{W_t s_t} + \overrightarrow{b_t}) \tag{8}$$

$$\overrightarrow{\partial_t} = \frac{\exp(\overrightarrow{e_t})}{\sum_{t=1}^{T}\exp(\overrightarrow{e_t})} \tag{9}$$

$$y = \sum_{t=1}^{T}\overrightarrow{\partial_t s_t} \tag{10}$$

4)输出层

输出层的输入为注意力机制层的输出,利用多分类 Softmax 函数进行相应的计算,得出流行度预测结果。计算公式如下:

$$y_{pred} = \mathrm{softmax}(\omega_1 y + b_1) \tag{11}$$

式中:ω_1 代表注意力机制层到输出层所需的训练权重系数矩阵;b_1 为相应的偏置;y_{pred} 为输出层的输出预测标签。

2　实验分析

2.1　实验数据集

由于预测场景以及预测对象的不同,在流行度预测的领域中,已经采用了各种不同的数据集,包括新浪微博、推特等社交网络数据集、论文引用数据集以及 YouTube、Facebook 等视频内容数据集。其中只有部分的数据集是公开的,具体如表 1 所示。本文使用两个数据集进行验证,一是从 Facebook 网站收集的真实视频数据集[15,24],该数据集收集了 2015 年 8 月 1 日～2015 年 10 月 15 日期间由几家出版商上传到 Facebook 的 1 820 个视频信息,Trzciński 等[24]向公众发布了这个数据集,以便进一步研究媒体内容的流行度预测这一主题;二是选自 Kaggle 网站的 YouTube 视频数据集,该数据集包括视频每小时的观看次数变化。本文选取数据集一中 1 820 条视频每个小时被访问的次数即 168 个流行度特征数以及数据集二中 5 000 条视频每小时观看次数即 240 个流行度特征数,并按比例将样本数据集分成训练数据集(60%)和测试数据集(40%)。在训练阶段使用 Dropout 技术,用于避免预测模型出现过拟合的状况。

表 1　流行度预测数据集

作者	数据集	是否公开
方元武等[8]	网站爬取数据集	否
Nia 等[13]	YouTube 数据集	是
宋旭鸣等[15]	Facebook 数据集	是
Trzciński 等[24]		
Cao 等[25]	2016 年 6 月 1 日新浪微博数据集	是
	APS 论文引用数据集	否
Cheng 等[26]	推特数据集	否
Tan 等[27]	MovieLens 20M	是
Liao 等[28]	微信公众号的文章数据	否

2.2　实验环境

硬件环境:CPU 为 AMD Ryzen 7 4800H with Radeon Graphics 2.90 GHz,RAM 为 16.0 GB。

软件环境:Windows10 系统,Python3.7。

2.3　评价标准

本文利用均方误差(MSE)、均方根误差(RMSE)、平均绝对误差(MAE)和准确率(Accuracy)对流行度预测模型效果进行评价,具体计算公式如下:

$$MSE = \frac{1}{n}\sum_{t=1}^{N}(predicted_t - original)^2 \tag{12}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{t=1}^{N}(predicted_t - original)^2} \tag{13}$$

$$MAE = \frac{1}{n}\sum_{t-1}^{N}\left|(predicted_t - original)\right| \tag{14}$$

$$Accuracy = \frac{TP + TN}{P + N} \times 100\% \tag{15}$$

其中,$predicted_t$ 和 $original$ 分别代表预测值和真实值,$TP+TN$ 表示所有预测正确的数量,$P+N$ 表示总数。

2.4　实验设置

以目前主流的流行度预测效果较好的 LSTM 模型为基础实验,分别和已有的 ARIMA、SVR、堆叠 LSTM 模型[15]、双向 LSTM 模型、BiLSTM-Attention 模型和本文模型进行对比。表 2 所示为本文模型的参数设置。

2.5　对比试验结果及分析

本文提出的基于注意力机制的 Bi-GRU 流行度预测模

表 2　BiGRU-Attention 模型参数值

参数	值
损失函数	MAE
优化器	Adam
隐藏层层数	3
隐藏层节点数	128,64,64
Batch Size	50
Epoch Num	1 000
Dropout 比率	0.2

型，将擅长学习长期依赖信息的双向 GRU 模型加入到预测中，然后用注意力机制提高对流行度预测结果有决定作用的特征权重，实验结果表明本文的方法准确率达到了 96.20%和 98.0%。同时，与其他已有模型进行对比，均方误差(MSE)，均方根误差(RMSE)，平均绝对误差(MAE)的值都显著降低。如表 3 和 4 所示为不同数据集下 MSE，RMSE 和 MAE 三个评价标准结果对比，表 5 所示为不同数据集下的准确率结果对比。

表 3　数据集一不同模型 MSE,RMSE,MAE 结果对比

模型	MSE	RMSE	MAE
ARIMA	0.000 172	0.013 1	0.003 2
SVR	0.000 157	0.012 5	0.007 8
LSTM	0.000 108	0.010 4	0.002 4
堆叠 LSTM	0.000 095	0.009 7	0.004 0
BiLSTM-Attention	0.000 077	0.008 8	0.004 6
本文模型	0.000 036	0.006 0	0.000 86

表 4　数据集二不同模型 MSE,RMSE,MAE 结果对比

模型	MSE	RMSE	MAE
ARIMA	0.000 161	0.012 7	0.004 1
SVR	0.000 140	0.011 8	0.005 3
LSTM	0.000 107	0.010 3	0.005 6
堆叠 LSTM	0.000 074	0.008 6	0.000 9
BiLSTM-Attention	0.000 061	0.007 8	0.002 8
本文模型	0.000 059	0.007 7	0.001 1

表 5　不同数据集下各模型准确率　%

数据集	模型						
	ARIMA	SVR	LSTM	堆叠 LSTM	Bi-LSTM	BiLSTM-Attention	本文模型
一	81.69	83.22	88.43	89.88	90.39	91.54	96.20
二	82.85	85.05	88.60	92.15	92.34	93.48	98.04

从表 3 和 4 可以看出，对于基本主流算法而言，LSTM 模型的 MSE，RMSE 和 MAE 相对较低，预测效果较好；由此可以验证深度学习方法的确优于传统机器学习方法，机器学习需要通过大量的样本进行模型训练使预测准确度得到提高，这样对于刚上传或者数据周期短的内容资源预测效果有很大降低。另外，就均方根误差(RMSE)的值而言，数据集一(Facebook 数据集)下本文模型较主流模型降低了 31.8%～54.2%；数据集二(YouTube 数据集)下降低了 2.02%～39.58%。表中本文模型的 MSE，RMSE 和 MAE 的值都较低，误差较小，效果最好。

由表 5 可以看出本文的模型准确率高达 96.20%和 98.0%，相比 ARIMA、SVR、LSTM、堆叠的 LSTM 模型、双向 LSTM 模型和 BiLSTM-Attention 模型，数据集一中预测准确率分别提升了 15.08%、13.50%、8.07%、6.57%、6.04%和 4.62%；数据集二中预测准确率分别提升了 15.49%、13.25%、9.63%、6.01%、5.81%和 4.65%。

对于其他深度模型，本文提出的流行度预测模型中，尽管 GRU 模型的作用与 LSTM 模型类似，然而相对于长序列来说更能有效抑制梯度消失或爆炸，其效果明显优于传统的 RNN 模型。数据集二中，选取了较多的流行度特征以及样本数量，从结果可以看出预测的效果优于较少的流行度预测特征，这是因为特征越多，模型可以更好地学习上下文信息，对结果有一定优化作用。

图 4 中，准确率值再一次验证深度学习方法较传统机器学习方法更好。引入的双向循环机制使得预测准确率分别提升了 2.21%和 4.22%，因为其在获取历史信息的同时，也不忽略未来的信息，考虑到了不同特征之间的时序相关性，充分捕捉前后信息特征；再引入注意力机制后，判断每个特征对流行度预测的重要程度，提高预测的效果，自适应地对重要的特征赋予更高的权重。从实验结果也可以看出，加入 Attention 机制的模型准确率分别提高了 1.51%和 1。23%。并且在实验过程中，双向 GRU 模型相比于双向 LSTM 模型提前 28.33%～52.21%的收敛时间，因此从综合性能角度来说，本文模型优于其他的模型。并且在不同的数据集下本文模型均有较好的表现，同时验证了该模型具有一定适应性的优点。

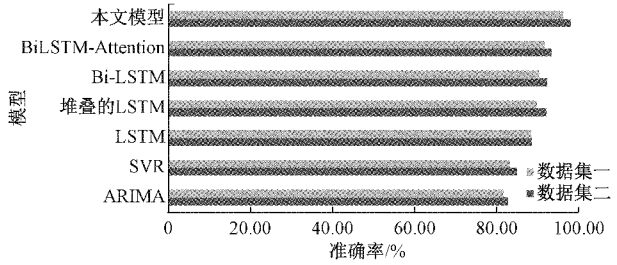

图 4　准确率对比图

3 结　论

内容资源流行度预测方法的提出是为了更好解决如何在有限地条件下合理地对内容进行缓存与调度,本文针对其存在的预测准确率不高,自适应能力较弱等问题提出了一种基于注意力机制的 Bi-GRU 流行度预测模型。该模型更好地挖掘了内容资源历史访问模式与其未来流行度之间的关系,充分利用了时序序列的相关性,提高了特征提取的效率以及质量。所以本文的模型可以更加准确地预测内容资源未来的流行度。在使用 Facebook 的真实数据集验证了本文的模型后,结果表明本文的模型确实优于其他已有的流行度预测模型。

下一步的研究将进一步拓展资源预测的可用信息维度,如引入边缘侧的用户群体行为作为辅助维度,或资源的内容属性信息(如内容相似资源流行度、资源兴趣类型等)。另外,如何基于流行度预测结果设计优化 CDN 资源数据文件的"推拉"、"更新"、"老化"等操作也是未来值得进一步研究的问题。

参考文献

[1] 孔令义. 面向 5G 的网络优化和重构[J]. 电信科学,2020,36(2):117-125.

[2] HASSINE N B, MARINCA D, MINET P, et al. Caching strategies based on popularityprediction in content delivery networks [C]. IEEE International Conference on Wireless & Mobile Computing, IEEE, 2016:1-8.

[3] 韩国栋,朱一戈,张帆. 一种基于认知的动态副本放置方法[J]. 计算机应用与软件,2013(1):83-87.

[4] 尹志鹏,侯方杰,王雷. 基于内容流行度的 ICN 缓存策略性能验证与分析[J]. 网络新媒体技术,2019,8(2):59-63.

[5] HASSINE N B, MILOCCO R, MINET P. ARMA based popularity prediction for caching in content delivery networks [C]. 2017 Wireless Days (WD), IEEE, 2017:113-120.

[6] 朱琛刚,程光,胡一非,等. 基于流行度预测的互联网+电视节目缓存调度算法[J]. 计算机研究与发展,2016,53(4):742-751.

[7] 徐冬,肖莹慧. 基于机器学习技术的网站用户行为预测[J]. 现代电子技术,2019,42(4):94-96,100.

[8] 方元武,何雪. 基于混合模型的内容资源流行度预测算法[J]. 微型电脑应用,2020,36(12):123-126.

[9] GAO S, MA J, CHEN Z. Modeling and predicting retweeting dynamics on microblogging platforms[P]. Web Search and DataMining, 2015:107-116.

[10] BAO P, SHEN H W, JIN X, et al. Modeling and predicting popularity dynamicsof microblogs using self-excited hawkes processes[J]. ACM, 2015:9-10.

[11] MISHRA S, RIZOIU M A, XIE L. Feature driven and point process approaches for popularity prediction [C]. Acm International, ACM, 2016:1069-1078.

[12] 陈亮,张俊池,王娜,等. 基于深度信念网络的在线视频热度预测[J]. 计算机工程与应用,2017,53(9):162-169,189.

[13] NIA Z M, KHAYYAMBASHI M R. Improving content popularity prediction with k-means clustering and deep-belief networks [J]. Multimedia Tools and Applications,2021, 80(11):1-20.

[14] QIANG C, XIN Q. Research on trend analysis and prediction algorithm based on time series[C]. 2019 3rd International Conference on Electronic Information Technology and Computer Engineering (EITCE), 2019:73-76.

[15] 宋旭鸣,沈逸飞,石远明. 基于深度学习的智能移动边缘网络缓存[J]. 中国科学院大学学报,2020,37(1):128-135.

[16] 鲍鹏,徐昊. 基于图注意力时空神经网络的在线内容流行度预测[J]. 模式识别与人工智能,2019,32(11):1014-1021.

[17] 武维,李泽平,杨华蔚,等. 融合内容特征和时序信息的深度注意力视频流行度预测模型[J]. 计算机应用,2021,41(7):1878-1884.

[18] 曾本冲,万旺根. 基于编解码器的组件式交通事故预测网络[J]. 电子测量技术,2021,44(6):90-95.

[19] 殷礼胜,孙双晨,魏帅康,等. 基于自适应 VMD-Attention-BiLSTM 的交通流组合预测模型[J]. 电子测量与仪器学报,2021,35(7):130-139.

[20] 梁宇佳,宋东峰. 基于 LSTM 和情感分析的股票预测[J]. 科技与创新,2021(21):126-127.

[21] 刘卓凡,郑庆庆,李俊,等. 基于注意力机制和 LSTM 的文本情感分析[J]. 信息与电脑(理论版),2021,33(18):63-65.

[22] 陈莉媛,毋涛. 融合主题模型与自注意力机制的短文本情感分析方法[J]. 国外电子测量技术,2021,40(11):18-23.

[23] 朱江,杜瑞,李建奇,等. 基于注意力机制的曲轴瓦盖上料机器人视觉定位和检测方法[J]. 仪器仪表学报,2021,42(5):140-150.

[24] TRZCINSKI T, ROKITA P. Predicting popularity of online videos using support vector regression[J]. IEEE Transactions on Multimedia, 2017, 19 (11): 2561-2570.

[25] CAO Q, SHEN H, CEN K, et al. DeepHawkes: Bridging the gap between predictionand understanding of information cascades[C]. The 2017 ACM. ACM,

2017:1149-1158.

[26] CHENG L, MA J, GUO X, et al. DeepCas: An end-to-end predictor of information cascades [J]. International World Wide Web Conferences Steering Committee, 2017, DOI: 10.1145/3038912.3052643.

[27] TAN J, LIU W, WANG T, et al. A high-accurate content popularity prediction computational modeling for mobile edge computing using matrix completion technology [J]. Transactions on Emerging Telecommunications Technologies, 2020(1):32, DOI: 10.1002/ett.3871.

[28] LIAO D, XU J, LI G, et al. Popularity prediction on online articles with deep fusion of temporal process and content features [J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2019, 33: 200-207.

作者简介

许阅，硕士研究生，主要研究方向为内容分发网络。

E-mail:446349037@qq.com

刘光杰(通信作者)，教授，博士，主要研究方向为网络与通信安全。

E-mail:gjicliu@gmail.com