DOI:10.19651/j.cnki.emt.2105839

基于编解码器的组件式交通事故预测网络*

曾本冲[1,2] 万旺根[1,2]
(1.上海大学 通信与信息工程学院 上海 200444; 2.上海大学 智慧城市研究院 上海 200444)

摘 要: 随着道路机动车数量的不断增多,交通事故已成为危害社会公共安全的主要因素之一,道路交通事故的预测也成为了研究热点。考虑到事故影响因素的错综复杂性和事故发生具有动态的时空变化性与数据稀疏性等问题,通过对多源数据的融合并按照时变和时不变数据进行特征提取,特别加入事故的文本描述特征提取上下文信息,同时采用负采样法平衡正负样本比例,最终提出了一种多特征组件组合训练的区域交通事故预测网络模型。在美国的 3 个具有不同事故稀疏性的城市数据集上进行了模型验证,实验结果表明该预测模型在各项评价指标上都优于对比的基础模型,各项指标提升约 2%~3%,可以看出该模型在一定程度上提升了预测性能,同时通过多特征组件的不同组合实验结果说明各项因素对事故发生具有影响性。

关键词: 交通事故预测;编解码网络;LSTM;注意力机制
中图分类号: TP391 **文献标识码**: A **国家标准学科分类代码**: 510.4030

Component traffic accident prediction network based on codec

Zeng Benchong[1,2] Wan Wanggen[1,2]
(1.School of Communication and Information Engineering, Shanghai University, Shanghai 200444, China;
2.Institute of Smart City, Shanghai University, Shanghai 200444, China)

Abstract: With the continuous increase in the number of road vehicles, traffic accidents have become one of the main factors that endanger social public safety, and the prediction of road traffic accidents has also become a research hotspot. Taking into account the intricacies of accident influencing factors and the dynamic spatio-temporal variability and data sparseness of accidents, the fusion of multi-source data and the feature extraction according to time-varying and time-invariant data, especially the text description of the accident. The feature extracts context information, and at the same time, the negative sampling method is used to balance the ratio of positive and negative samples. Finally, a regional traffic accident prediction network model trained by multi-feature component combination training is proposed. The model was validated on the data sets of three cities with different accident sparsity in the United States. The experimental results show that the prediction model is better than the basic model of comparison in various evaluation indicators, and each indicator is increased by about 2%~3%. It can be seen that the model has improved the prediction performance to a certain extent. At the same time, the experimental results of different combinations of multi-feature components show that various factors have an impact on the occurrence of accidents.

Keywords: traffic accident prediction; encoder decoder network; LSTM; attention mechanism

0 引 言

近年来,随着城市化进程的加快,人们的出行越来越便捷,然而这也带来了交通拥堵和交通事故等一系列社会问题,导致政府在交通管制方面压力巨大。根据世界卫生组织发布的《2018 年全球道路安全状况报告》,每年约有 135 万人死于交通事故[1]。为了减少交通事故造成的损失,在充分利用智能交通系统采集的海量多源异构数据的基础上,提前对区域交通事故发生可能性的准确评估可以指导驾驶员选择更安全的道路,同时辅助交通管理部门提前进行资源调配与决策。

交通事故分析与预测在过去几十年里一直是人们研究的热点。交通事故预测问题通常被表述为分类问题或回归问题。早期,传统的机器学习方法在该问题的研究应用比

收稿日期:2021-02-24
*基金项目:中国博士后科学基金(2020M681264)、上海市科委港澳台科技合作基金(18510760300)项目资助

较成熟和广泛。Hossain 等[2]应用贝叶斯信念网络建立实时公路碰撞预测模型。Lin 等[3]提出了一种基于频繁模式树(FP-tree)的方法来选择更可能导致交通事故的变量特征,然后利用 KNN 和贝叶斯网络对交通事故进行预测。Chen 等[4]使用逻辑回归模型和非负矩阵分解来预测事故。殷礼胜等[5]提出基于 EEMD-IPSO-LSSVM 的交通流组合模型完成交通预测任务。近些年来,一些学者更加关注深度学习方法在交通预测领域的应用。Zhang 等[6]提出 ST-ResNet 模型来预测城市各区域的人群流入和流出趋势,并实现项目的实际应用。程健等[7]基于人工智能算法建立实时交通流量预测。佟健颉等[8]等建立了一个深度残差网络模型用来进行短时交通流量预测。而在交通事故的预测上,Zheng 等[9]和 Najjar 等[10]使用卷积神经网络(CNN)预测交通事故风险图。Ren 等[11]和张志豪等[12]考虑交通事故风险的时间格局,利用交通流量、天气和空气数据,应用 LSTM 模型研究区域交通的时间周期性和趋势。Yuan 等[13]提出 Hetero-ConvLSTM 网络针对具有稀疏特性的交通事故数据集进行分析预测等。Zhu 等[14]结合行政区和出租车流划分研究区域,使用基于时空注意力机制的编解码结构网络对区域交通事故风险进行预测。Yu 等[15]融合多源交通事故数据集,包括车流、天气、POI 等信息,实现精确到路网层的事故预测。Zhou 等[16]提出了一种基于 GCN 网络的多任务时空细粒度交通事故预测框架,将预测的时间粒度提高到分钟级水平。基于深度学习的方法在大规模交通事故数据集的分析上性能总体优于传统机器学习方法。然而,以上方法也存在以下不足之处:1)未同时考虑时间的相关性和周期性、空间的依赖性问题,事故本身存在长期的历史和未来时空动态变化性;2)交通事故发生过程的文本描述隐藏了丰富的语义信息,现有的深度学习模型极少挖掘该特征;3)交通事故数据是极具代表性的稀疏性数据,深度学习模型在训练过程大都未做特殊处理,从而影响最终的预测精度。

为了解决以上的问题,本文提出了一种基于编解码结构的组件式区域交通事故预测网络(ED-CRTAN),旨在预测未来某个时间某个区域的交通事故发生风险。首先,基于大规模的美国交通事故数据集[17],本文以矩形法划分地理区域,以 15 min 作为时间间隔分别处理时变和时不变数据。然后,本文设计了一种组件式区域交通事故预测网络,通过加入时间注意力机制的编解码结构和对地理空间特征提取的嵌入层分别考虑交通事故的时间相关性和空间异质性。此外还采用负采样法解决正负样本数据不平衡的问题。最后,通过在真实的数据上进行对比实验,结果表明本文提出的模型训练性能更好。

1 数据集与问题描述

1.1 数据集描述

1)交通事故数据

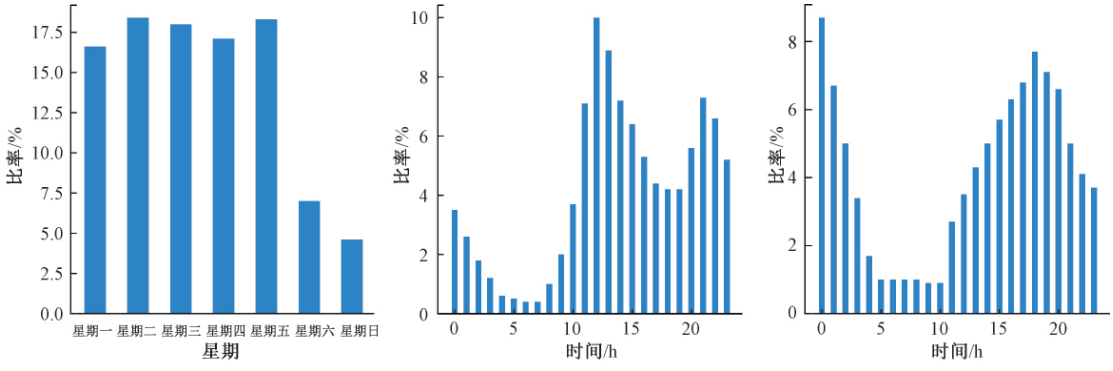
本文选取了该数据集中美国 3 个大型城市(亚特兰大、夏洛特、洛杉矶)从 2018 年 6 月~2018 年 8 月的道路交通事故记录,包含时间、经纬度、事故文本描述和 8 种交通状况类型等信息,如表 1 所示 3 个城市的数据量统计情况并计算出相应的事故率。交通状况类型体现了当前道路交通通行情况,如:车辆故障、交通限速和交通拥堵等。

表 1 交通事故和非事故数据统计情况

城市	事故	非事故	总计	事故率/%
亚特兰大	2 739	21 205	23 944	11.4
夏洛特	5 604	12 997	18 601	30.1
洛杉矶	8 303	84 685	92 988	0.09

2)日历数据

该数据描述了交通事故的日历特性,包括时段、工作日与周末、白天与黑夜信息。其中按照人们的正常生活习惯,将一天 24 h 分为 5 个时段:6:00~10:00、10:00~15:00、15:00~19:00、19:00~22:00、22:00~6:00。如图 1 所示,统计了美国交通事故在不同日历特性上发生的频率。



(a) 按星期统计事故率 (b) 按工作日的24 h统计事故率 (c) 按周末的24 h统计事故率

图 1 美国交通事故数据日历特性分布情况

3)天气数据

该数据收集了从 2018 年 6 月~2018 年 8 月的美国各城市天气数据记录,包含温度、湿度、气压、能见度、风速、降水量和特殊天气类型信息。其中特殊天气类型包括下

雨、下雪、大雾和冰雹天气。

4)POIs 数据

该数据从 Open Street Map(OSM) 上共收集了路网 15 种兴趣点标记信息，如：减速带、十字路口和停车标志等，其描述了路网结构的固有特性。

1.2　问题描述

定义 1(时空域)：描述一个城市的时空域为 $\boldsymbol{R}^m \times \boldsymbol{T}^n$。其中，空域 $\boldsymbol{R}^m$ 由 m 个子区域 r 组成，即 $\boldsymbol{R}^m = \{r_1, r_2, \cdots, r_m\}$，每个子区域 $r \in \boldsymbol{R}^m$ 为一个 $d \times d$ 的矩形区域，定义 $d = 5$ km；时域 $\boldsymbol{T}^n$ 由 n 个时间片 t 组成，即 $\boldsymbol{T}^n = \{t_1, t_2, \cdots, t_n\}$，每个时间片 $t = \Delta t \in \boldsymbol{T}^n$，定义 $\Delta t = 15$ min。

定义 2(交通事件)：描述某个子区域某个特定时间片内的交通事件为 $\boldsymbol{E}_{r_j}^{t_i} = \{e_1, e_2, \cdots\}$，其中 $r_j \in \boldsymbol{R}^m, t_i \in \boldsymbol{T}^n$。

定义 3(天气)：描述某个子区域某个特定时间片内的天气事件为 $\boldsymbol{W}_{r_j}^{t_i} = \{w_1, w_2, \cdots\}$，其中 $r_j \in \boldsymbol{R}^m, t_i \in \boldsymbol{T}^n$。

定义 4(兴趣点)：描述某个子区域的兴趣点标记为 $\boldsymbol{P}_{r_j} = \{p_1, p_2, \cdots\}$，其中 $r_j \in \boldsymbol{R}^m$。

问题求解：给定所有的历史交通事故序列，每个输入序列描述为 $\boldsymbol{X}_{r_j}^{t_i} = [\boldsymbol{E}_{r_j}^{t_i}, \boldsymbol{W}_{r_j}^{t_i}, \boldsymbol{P}_{r_j}]$，则对应的输出描述如式(1)所示。

$$\boldsymbol{Y}_{r_j}^{t_i} = \begin{cases} 1, & \text{发生} \\ 0, & \text{未发生} \end{cases} \tag{1}$$

使用最近 8 个时间片(前 2 h)的观测数据 $\boldsymbol{X}_{r_j}^{T} = \{\boldsymbol{X}_{r_j}^{t_{i-7}}, \boldsymbol{X}_{r_j}^{t_{i-6}}, \cdots, \boldsymbol{X}_{r_j}^{t_i}\}$ 来预测未来第 $\boldsymbol{T}+1$ 个时间片(15 min)内区域 r 的交通事故发生概率，如式(2)所示。

$$\hat{\boldsymbol{Y}}_{r_j}^{T+1} = f(\boldsymbol{X}_{r_j}^{T}, \boldsymbol{W}), r_j \in \boldsymbol{R}^m \tag{2}$$

通过最小化预测误差 ϑ，拟合以往的交通事故序列，训练出模型权值 $\boldsymbol{W}$，如式(3)所示。

$$\vartheta = \sum_{r_j} (\hat{\boldsymbol{Y}}_{r_j}^{T+1} - \boldsymbol{Y}_{r_j}^{T+1})^2 + \lambda \| \boldsymbol{W} \|^2, r_j \in \boldsymbol{R}^m \tag{3}$$

式中：$\boldsymbol{Y}_{r_j}^{T+1}$ 为区域 r_j 在第 $\boldsymbol{T}+1$ 个时间片发生交通事故的真实值；λ 为控制正则化重要程度的超参数。

2　研究方法

基于以上的多源融合数据源进行交通事故实时预测，本文提出了一种基于编解码结构的组件式区域交通事故预测网络框架，并命名为 ED-CRTAN。该模型如图 2 所示，本文将详细描述各组件的组成原理。

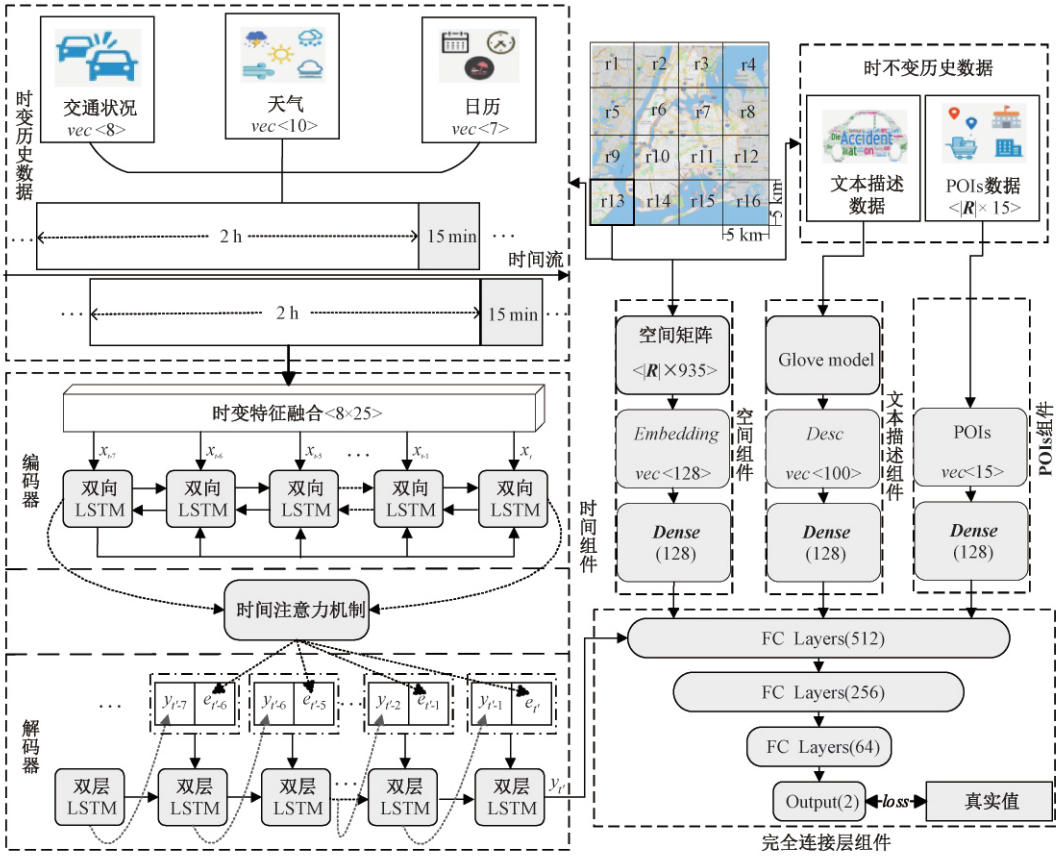


图 2　ED-CRTAN 网络框架

2.1　时间组件

在每个网格单元 $r_j \in \boldsymbol{R}^m$ 内，对时变特征建立基于编解码结构的时间组件：双向 LSTM 作为编码器，双层 LSTM 作为解码器，编码器和解码器之间加入时间注意力

机制模块来拟合历史时间对未来时间的影响程度。其中,双向 LSTM 的正向和反向隐藏层大小都为 64,则每个时间片的输出维度为 128 维;双层 LSTM 每一层的隐藏层大小都为 128,则每个时间片的输出维度仍为 128 维。双向 LSTM 和双层 LSTM 都是由若干个 LSTM 基本单元组成。

对于连接双向 LSTM 和双层 LSTM 的时间注意力模块,本文根据 Bahdanau 等[18]在神经网络机器翻译任务中提出的注意力机制原理,首先由式(4)计算编码器输出源序列中每个时间步长的隐藏状态 h_t 和解码器中输出目标序列中每一个时间步长的隐藏状态 $h_{t'-1}$ 的对齐分数。

$$e_{t'}^{t} = h_{t'-1} W_d h_t + b_d \tag{4}$$

式中:W_d 为参数矩阵;b_d 为偏置项。如式(5)所示,使用 softmax 函数对对齐分数 $e_{t'}$ 进行归一化,以得到检索注意力权值 $a_{t'}^{t}$。

$$a_{t'}^{t} = \frac{e^{e_{t'}^{t}}}{\sum_{t=1}^{T} e^{e_{t'}^{t}}} \tag{5}$$

由此,再根据式(6)计算编码器隐藏状态 h_t 与其对应的注意权值 $a_{t'}^{t}$ 的乘积线性和,从而得到解码器每个 LSTM 单元的上下文向量 $C_{t'}$。

$$C_{t'} = \sum_{t=1}^{T} a_{t'}^{t} h_t \tag{6}$$

2.2 空间组件

该组件实现对每个网格单元 $r_j \in R^m$ 的空间编码表示,包含对交通事故发生区域存在的空间异质性、交通特征以及环境影响因素等重要信息的空间编码。本文使用隐藏层大小为 128 的前馈神经网络训练编码空间特征,并使用 sigmoid 函数作为激活函数。

2.3 文本描述组件

该组件首先通过 Glove 模型实现对每个网格单元 $r_j \in R^m$ 中的历史交通事件的自然语言描述进行分词,得到 100 维的词向量矩阵。本文使用隐藏层大小为 128 的前馈神经网络训练编码空间特征,并使用 sigmoid 函数作为激活函数。

2.4 POIs 组件

该组件首先使用向量矩阵来表征道路路网结构的固有特性,通过使用隐藏层大小为 128 的前馈神经网络进行特征训练,并使用 sigmoid 函数作为激活函数。

2.5 完全连接层组件

该组件使用多个完全连接层实现对上述所有组件的输出进行最终预测,各层大小分别是 512、256、64 和 2。其中,本文使用 ReLU 函数作为前 3 层网络的激活函数,使用 softmax 函数作为最后一层网络的激活函数输出 0(事故不发生)或 1(事故发生)。同时,以本文 1.2 节给出的最小化预测误差计算公式作为 loss 损失函数。

3 结果分析

3.1 实验环境与超参数配置

电脑运行内存为 8 GB,CPU 为 Intel ® CoreTM i7-9750H CPU@2.60 GHZ×12,显卡为 GeForce GTX 1650/PCIe/SSE2,使用 python3.7 语言的 PyCharm 集成开发工具,使用 Keras 和 Tensorflow 神经网络框架搭建模型。

本文选择 70%的数据作为训练集进行模型训练,选择 10%的数据作为验证集来提升模型的泛化能力,选择剩下 20%的数据作为测试集来评价模型的性能。本文设置 epochs 大小为 100,batch 大小为 64,初始学习率为 0.01 的 Adam 来优化模型的训练过程。同时,本文使用 L2 正则化来避免模型过拟合,使用网格搜索进行超参数调整。

3.2 评价指标

本文采用精确率(Precision)、召回率(Recall)和 F1-Score三个指标综合评价各个模型的最终性能。其中,精确率描述了在被所有预测为正的样本中实际为正样本的概率,如式(7)所示。

$$Precision = \frac{TP}{TP + FP} \tag{7}$$

召回率描述了正确预测的正观测值与实际正观测值的比率,如式(8)所示。

$$Recall = \frac{TP}{TP + FN} \tag{8}$$

F1-Score 同时考虑了假阴性和假阳性,描述了精确率和召回率的加权平均值,如式(9)所示。

$$F1 - Score = \frac{2 * TP}{2 * TP + FN + FP} \tag{9}$$

其中,TP(true positives)表示被正确地划分为正例的个数;FP(false positives)表示被错误地划分为正例的个数;FN(false negatives)表示被错误地划分为负例的个数;TN(true negatives)表示被正确地划分为负例的个数。

3.3 实验对比结果

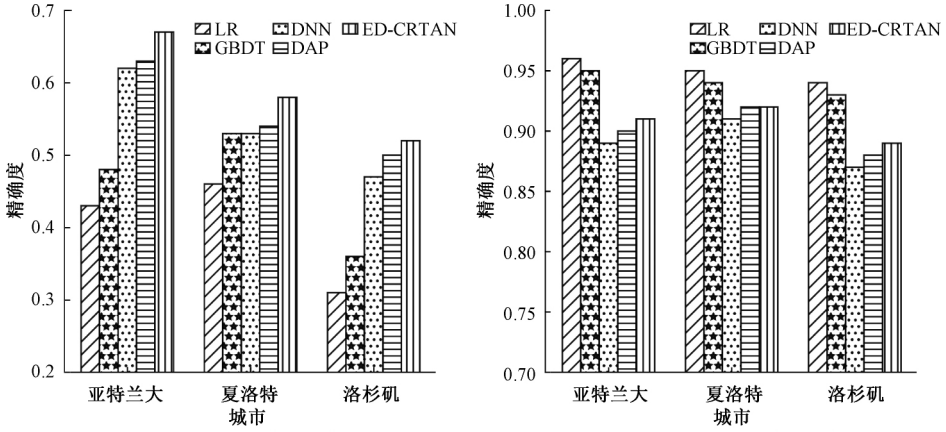
本文选择逻辑回归模型(LR)、梯度提升树模型(GBDT)、深度神经网络(DNN)和深度事故预测模型(DAP)[19]作为基础模型进行对比实验,并在 3 个美国大型城市数据上进行实验,选择精确率、召回率、F1-Score(F1)作为模型性能的评价指标,取多次实验计算各指标的事故级和非事故级加权平均值,如表 2 所示。

由表 2 可知,本文提出的 ED-CRTAN 模型在 3 个真实美国城市交通事故数据集上的性能优于 4 种基础模型,平均每个指标提升大约 2%。特别地,对于具有不同稀疏性的城市事故数据集,ED-CRTAN 模型同样具有实用性。例如,夏洛特城和洛杉矶城事故率分别为 30.1%和 0.09%,具有明显的事故数据稀疏特性,结果显示本文的 ED-CRTAN 模型相较于 2019 年提出的 DAP 模型在F1-Score上分别

表 2　ED-CRTAN 模型与基础模型的实验对比结果

模型	亚特兰大			夏洛特			洛杉矶		
	精确率	召回率	F1	精确率	召回率	F1	精确率	召回率	F1
LR	0.84	0.83	0.83	0.84	0.83	0.83	0.80	0.78	0.78
GBDT	0.84	0.84	0.84	0.85	0.84	0.84	0.80	0.78	0.78
DNN	0.84	0.83	0.83	0.83	0.81	0.81	0.79	0.76	0.77
DAP	0.84	0.84	0.84	0.84	0.81	0.82	0.80	0.77	0.78
ED-CRTAN	0.86	0.85	0.86	0.85	0.85	0.85	0.80	0.79	0.80

提升了 3%和 2%。为了更加突显在事故级数据稀疏的情况下模型性能提升情况，如图 3 所示，分别显示了各模型在 3 个城市的事故级和非事故级的预测平均精确率分布情况。



(a) 事故数据情况下不同模型的精确率对比　　(b) 非事故数据情况下不同模型的精确率对比

图 3　分别在事故级和非事故级下比较不同模型的精确率

由图 3(a)可以看出，在事故级的预测上，深度学习方法明显优于传统机器学习方法，前者能够充分挖掘稀疏性数据的潜在特征规律。特别地，本文提出的 ED-CRTAN 模型在稀疏事故级数据上的预测结果相比其它模型有显著提升。从图 3(b)可以看出，在海量的非事故级数据预测上，传统机器学习模型具有更好的分类效果。然而，在实际应用中更加关注事故级的预测，以便更好地预判某一特定时间和特定区域内事故发生情况。

4　结　　论

及时准确地交通事故预测对于城市交通和公共安全的治理至关重要。首先，本文基于海量的多源异构时空数据集，包括交通事故记录、天气记录和 POI 信息等，在数据融合的基础上完成对有意义的特征的提取。然后，本文提出了一种基于编解码结构的嵌入式组件交通事故预测框架 ED-CRTAN，包括时间组件、空间组件、文本描述组件、POIs 组件和完全连接层组件。最后，本文选择在真实的 3 个具有不同事故率的美国城市道路交通事故数据集上进行模型训练，并与 4 种基础模型进行了比较。实验表明，本文的模型性能优于 4 个现有模型。然而，本文提出的模型也存在一定的局限性。首先，在区域划分上还是采用的矩形法，只适用于区域预测，难以更加精确的定位到路网上，未来希望能够在路网层面实现预测。再者，事故受多种因素影响，当前数据集考虑的因素还比较少，未来希望扩充数据集，融合一些交通流、人文信息和经济信息等。

参考文献

[1] 张亚丽. 世界卫生组织发布《2018 年全球道路安全现状报告》[J]. 中华灾害救援医学，2019，7(2)：48.

[2] HOSSAIN M，MUROMACHI Y. A Bayesian network based framework for real-time crash prediction on the basic freeway segments of urban expressways[J]. Accident Analysis and Prevention，2012，45：373-381.

[3] LIN L，WANG Q，SADEK A W. A novel variable selection method based on frequent pattern tree for realtime traffic accident risk prediction[J]. Transportation Research Part C：Emerging Technologies，2015，55：444-459.

[4] CHEN Q，SONG X，FAN Z，et al. A context aware nonnegative matrix factorization framework for traffic accident risk estimation via heterogeneous data[C]. 2018 IEEE Conference on Multimedia Information Processing and Retrieval(MIPR)，2018：346-351.

[5] 殷礼胜，唐圣期，李胜，等. 基于 EEMD-IPSO-LSSVM

的交通流组合预测模型[J].电子测量与仪器学报,2019,33(12):126-133.

[6] ZHANG J,ZHENG Y,QI D.Deep spatio-temporal residual networks for citywide crowd flows prediction [C]. Proceedings of the 31st AAAI Conference on Artificial Intelligence(AAAI), 2017:1655-1661.

[7] 程健,毛忠峰.基于人工智能的实时交通流量预测[J].国外电子测量技术,2019,38(6):28-32.

[8] 佟健颉, 黎英, 王一旋.基于深度残差网络的短时交通流量预测[J].电子测量技术, 2019,42(18):90-94.

[9] ZHENG M, LI T, ZHU R, et al. Traffic accident's severity prediction: A deep learning approach based CNN network[J]. IEEE Access,2019,99:1-1.

[10] NAJJAR A, KANEKO S, MIYANAGA Y. Combining satellite imagery and open data to map road safety[C].AAAI,2017:4524-4530.

[11] REN H L, SONG Y, LIU J X, et al. A deep learning approach to the citywide traffic accident risk prediction[C]. Proceedings of the 21st International Conference on Intelligent Transportation Systems (ITS), 2018: 3346-3351.

[12] 张志豪,杨文忠,袁婷婷,等.基于 LSTM 神经网络模型的交通事故预测[J].计算机工程与应用,2019,55(14):249-253,259.

[13] YUAN Z,ZHOU X,YANG T.Hetero-ConvLSTM: A deep learning approach to traffic accident prediction on heterogeneous spatio-temporal data[C].The 24th ACM SIGKDD International Conference, 2018:984-992.

[14] ZHU L, LI T, DU S. TA-STAN: A deep spatial-temporal attention learning framework for regional traffic accident risk prediction[C]. 2019 International Joint Conference on Neural Networks(IJCNN),2019: 1-8.

[15] YU L, DU B, HU X, et al. Deep spatio-temporal graph convolutional network for traffic accident prediction[J]. Neurocomputing, 2021,423:135-147.

[16] ZHOU Z, WANG Y, XIE X, et al. RiskOracle: A minute-level citywide traffic accident forecasting framework[C]. The 34th AAAI Conference on Artificial Intelligence,2020,DOI:10. 1609/aaai.v34i01. 5480.

[17] SOBHAN M, MOHAMMAD H S, ARNAB N, et al. Short and longterm pattern discovery over large scale geospatiotemporal data[C]. Knowledge Discovery & Data Mining,2019:2905-2913.

[18] BAHDANAU D, CHO K, BENGIO Y. Neural machine translation by jointly learning to align and translate[C]. Proceedings of the 3rd International Conference on Learning Representations(ICLR),2015:1-12.

[19] MOOSAVI S,SAMAVATIAN M H,PARTHASARATHY S, et al. Accident risk prediction based on heterogeneous sparse data: New dataset and insights[C]. The 27th ACM SIGSPATIAL International Conference, 2019: 33-42.

作者简介

曾本冲,硕士,主要研究方向为时空大数据分析、交通事故分析和预测。

E-mail: 2550624530@qq.com

万旺根,教授,博士生导师,主要研究方向为计算机图形学、信号处理和数据挖掘。

E-mail: wanwg@staff.shu.edu.cn