融合时序知识图谱的路段级交通事故风险预测

唐伟文[1,2]　郭晟楠[1,2]　陈　炜[1,2]　林友芳[1,2]　万怀宇[1,2]

一、论文目的

准确预测未来交通事故发生的风险,可以大幅减少人员伤亡、降低交通事故造成的经济损失,对市民出行、政府管控都具有重要意义。围绕这一课题,路段级的交通事故风险预测尤为重要,具有实际应用价值。 因为，准确的路段级的事故风险预测，可以及时为交通参与者提供有针对性的行车建议和预警信息,保障出行安全。 例如,当预测某个路段的交通事故风险在未来一小时会激增时,可以对该路段上正在通行的车辆广播安全行车准则,并对计划通行该路段的车辆发送预警信息。

实现准确的路段级交通事故风险预测面临的三个问题:

1)交通事故风险会受到多源因素的动态影响。具体而言，交通事故风险会同时受到交通状态和已发生的交通事故的影响。

2)交通事故的发生存在复杂的时空相关性。 在时间维度上,路段上发生的历史事故会非线性地影响未来事故的发生. 在空间维度上,事故的发生会受到相邻路段和周围兴趣点的影响,而且影响程度是随时间动态变化的。

3)由于交通事故的发生是一个小概率事件,因此经统计得到的全路网的路段级事故风险中会出现大量零值,导致模型预测值收敛于零值,即出现零膨胀的问题。而收敛于零的预测值对于实际应用是没有意义的。

二、时序知识图谱构建

时序知识图谱嵌入模块首先基于训练集数据构建一个交通事故时序知识图谱,然后设计可以融入秒级时间戳信息的交通事故时序知识图谱历时嵌入模型(DETA),建模多源影响因素之间的动态、高阶相关性,最后通过预训练的方式,得到多源影响因素的深层嵌入表示。

1.构建交通事故时序知识图谱：基于时序知识图谱的定义,给出交通事故时序知识图谱的定义

$$\mathcal{G}=\{(u,r,v,t)\mid u\in\mathcal{E},v\in\mathcal{E},r\in\mathcal{R},t\in\mathcal{T},\ \varphi(u)\in\mathcal{A},\varphi(v)\in\mathcal{A},\phi(r)\in\mathcal{B}\}.$$

A 有 6 种实体类型，包括:事故等级与原因实体、交通状态等级实体、出租车服务区域实体、兴趣点实体、天气实体和路段实体。 B 有 9 种关系类型,包括:天气状态关系、交通状态关系、区域相邻关系、路段下游关系、路段交叉关系、区域包含兴趣点关系、区域包含路段关系、路段邻近兴趣点关系和事故发生关系。

二、时序知识图谱嵌入

下图为交通事故时序知识图谱的一个实例，图中不同的节点表示不同实体,不同的边表示不同的关系，边是有向的。

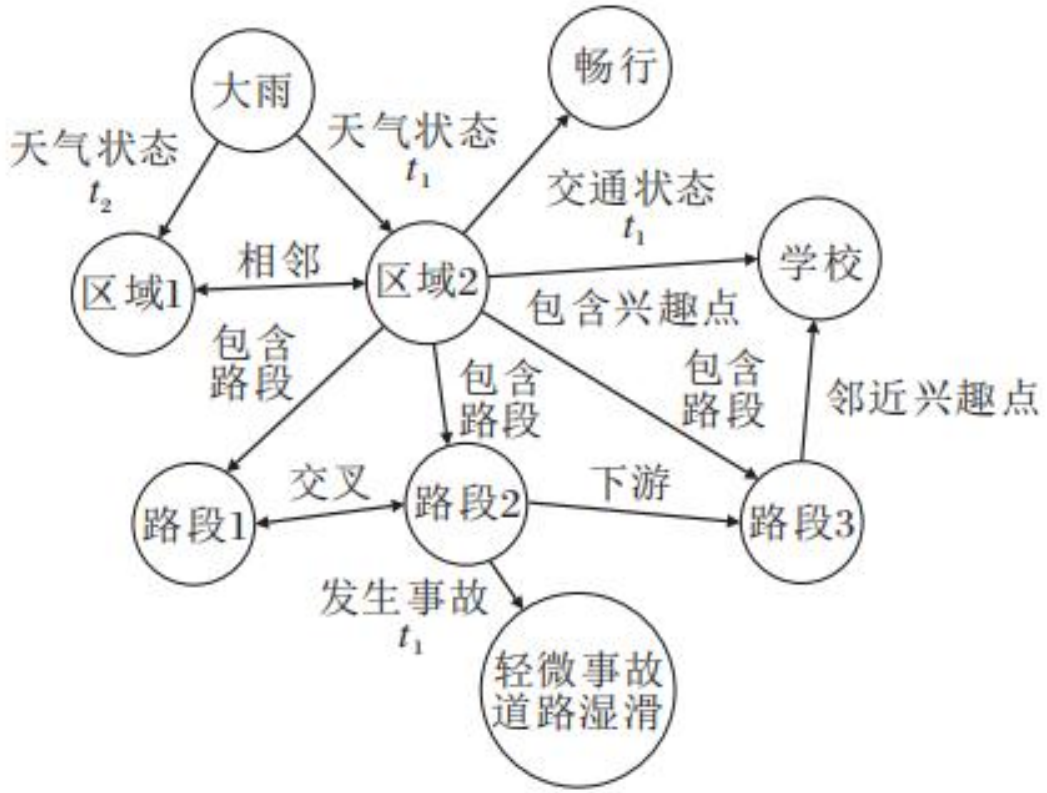

图 2　交通事故时序知识图谱实例

三、时序知识图谱历时嵌入模型

为了获得秒级时间粒度的历时实体嵌入，本文设计可以处理秒级时间戳的交通事故时序知识图谱历时嵌入模(DETA)。 模型由三部分组成:关系嵌入方法、实体嵌入方法和得分函数。

1)关系嵌入方法. 关系 $r \in R$嵌入表示Zr的定义如下:

$$z_r[i] = \{l_r[i], t_r[i]\}, 1 \leqslant i \leqslant d, \qquad (1)$$

其中，$L_r \in R_d$表示正向学习关系r的可学习向量，$t_r \in R_d$表示逆反关系r的可学习向量，d表示关系嵌入维度。

2)实体嵌入方法.实体$v \in \varsigma$在时刻t作为头实体的历时嵌入表示Z^t_v定义如下:

$$\begin{cases} \sum_{j \in P_t} a_v^j[i] \sin(w_v^j[i] j + b_v^j[i]), & 1 \leqslant i \leqslant \gamma d \\ m_v[i], & \gamma d < i \leqslant d \end{cases} \qquad (2)$$

三、时序知识图谱历时嵌入模型

3)得分函数. 给定任意事实四元组 $f = (u,r,v,t)$,得分函数会给出事实发生的概率:

$$\psi(u,r,v,t) = \frac{1}{2}\sum_{i=1}^{d}\left((z_u^t l_r)\overleftarrow{z}_v^t + (z_v^t t_r)\overleftarrow{z}_u^t\right), (3)$$

得分函数是正向事实四元组与逆向事实四元组语义匹配得分的平均值,头实体历时嵌入经过关系嵌入的线性映射后得到的嵌入与尾实体历时嵌入越相近，语义匹配得分越高。

四、实验

1.实验数据集：基于第三方库 OSMnx 和纽约市公开城市数据(https: / / opendata. cityofnewyork. us)，本文构建两个真实的路段级交通事故风险数据集:布鲁克林数据集和曼哈顿数据集,具体时间跨度为 2019 年 1 月1 日至2019 年12 月31 日. 两个数据集的统计情况如表 1 所示.

表 1　数据集统计信息

Table 1　Statistics of datasets

信息	布鲁克林	曼哈顿
交通事故数	41167	28287
出租车订单数	6259063	80184108
兴趣点数	5587	5866
天气/h	8760	8760
路段数	958	475

四、实验

2.实验结果:

表 3　各模型在布鲁克林数据集上的预测性能

Table 3　Prediction performance of different models on Brooklyn dataset

模型	RMSE	RMSE*	Recall /%	Recall* /%	MAP /%	MAP* /%
HA	0.2781	0.3107	19.69	19.37	27.61	32.57
MLP	*0.2640*	0.2955	25.81	23.61	36.20	41.92
GRU	0.2649	0.2964	23.08	21.43	37.69	42.74
ConvLSTM	*0.2640*	*0.2949*	26.94	23.53	37.90	43.61
SDCAE	0.2648	0.2960	24.96	23.42	29.73	35.05
GSNet	0.3057	0.3541	*27.02*	*24.38*	*38.84*	42.87
AGCRN	0.2656	0.2961	24.72	23.73	37.19	*44.14*
STGN-TKG	**0.2633**	**0.2946**	**28.58**	**24.41**	**39.11**	**45.34**

表 4　各模型在曼哈顿数据集上预测性能

Table 4　Prediction performance of different models on Manhattan dataset

模型	RMSE	RMSE*	Recall /%	Recall* /%	MAP /%	MAP* /%
HA	0.2752	0.3039	28.40	28.60	34.36	38.72
MLP	0.2621	0.2899	35.66	37.27	46.00	51.68
GRU	0.2631	0.2908	37.53	41.25	49.70	54.66
ConvLSTM	*0.2619*	**0.2887**	*43.07*	*44.36*	50.40	56.20
SDCAE	**0.2616**	*0.2888*	38.95	41.07	*52.22*	*57.79*
GSNet	0.2934	0.3263	40.01	41.95	49.60	56.86
AGCRN	0.2698	0.2968	38.42	41.91	48.91	53.49
STGN-TKG	*0.2619*	*0.2888*	**43.61**	**45.80**	**53.86**	**60.81**