第49卷第2期 中国科学技术大学学报 Vol. 49, No. 2
2019年2月 JOURNAL OF UNIVERSITY OF SCIENCE AND TECHNOLOGY OF CHINA Feb. 2019

文章编号:0253-2778(2019)02-0149-10

一种建模社交化点过程序列预测算法

江海洋[1],王 莉[2]

(1. 太原理工大学信息与计算机学院,山西晋中 030600;2. 太原理工大学大数据学院,山西晋中 030600)

摘要:根据序列数据预测下次事件类型和时间是一个值得研究的课题. 目前点过程强度函数算法仅从时间维度考虑背景知识和历史影响两个方面,没有从空间维度加入社交关系的影响. 针对该问题,提出基于时空深度网络的社交化点过程的序列预测算法(SPSP 算法). 该模型首先运用双LSTM(long short-term memory)分别建模强度函数的背景知识和历史影响;然后经过联合层将双LSTM 输出合并,生成事件类型和时间向量表征;最后在空间维度上加入社交关系网络影响,优化强度函数. 通过深度时空社交网络的多次训练,得到最优网络模型. 在新浪微博数据集上的实验验证算法的有效性,证明社交化点过程序列预测算法可高效准确预测出事件类型与时间.

关键词:序列预测;社交网络;点过程;LSTM;时空深度网络

中图分类号:TP18 **文献标识码**:A doi:10.3969/j.issn.0253-2778.2019.02.010

引用格式:江海洋,王莉. 一种建模社交化点过程序列预测算法[J]. 中国科学技术大学学报,2019,49(2):149-158.
JIANG Haiyang, WANG Li. A modeling socialization point process sequence prediction algorithm[J]. Journal of University of Science and Technology of China, 2019,49(2):149-158.

A modeling socialization point process sequence prediction algorithm

JIANG Haiyang, WANG Li[2]

(1. *College of Information and Computer, Taiyuan University Of Technology, Jinzhong* 030600, *China*;
2. *College of Big Data, Taiyuan University Of Technology, Jinzhong* 030600, *China*)

Abstract: Predicting the type and time of the next event according to the sequence data is a subject worth studying. At present, the point process intensity function only considers the background knowledge and historical influence from the time dimension, and has no influence on the social relations from the spatial dimension. Aiming at this problem, a sequence prediction algorithm (SPSP algorithm) is proposed based on the spatio-temporal deep network. In this model, firstly the background knowledge and historical influence of the intensity function are modeled with the dual LSTM (long short-term memory). Then the output of two LSTMs are combined by the union layer to generate the vector representation of event type and time. Finally, the influence of social networks on the spatial dimension is added to optimize the intensity function. Through multiple training of the Spatio-temporal deep neural network, the optimal network model is obtained. Sina weibo data sets are used to verify the validity of the algorithm, and it has been proven by experiments that the proposed algorithm can predict the event type and time efficiently accurately.

Key words: sequence prediction; social networking; point process; LSTM; deep time-space network

收稿日期:2018-09-21;**修回日期**:2018-12-04
基金项目:国家自然科学基金(61872260),山西省重点研发计划国际合作项目(201703D421013)资助.
作者简介:江海洋,女,1993年生,硕士生. 机器学习、序列预测. E-mail:2541529703@qq.com
通讯作者:王莉,博士/教授. E-mail:wangli@tyut.edu.cn

0 引言

序列推荐问题应用场景广泛,是一个值得研究的热点问题.很多领域都存在丰富的序列信息,比如电子商务平台的交易记录信息;社交网络平台的网络信息;设备故障的记录信息;十字路口交通事故的记录信息等.这些序列信息包含事件的时间、类型和属性.利用序列信息预测事件类型和时间,对捕获用户兴趣的动态变化、电子广告的精准投放、预防故障事故的发生等问题都具有重要意义.

根据序列生成方式的不同,将序列分为事件序列和时间序列.事件序列是随机异步生成的,时间序列是按照相同的时间间隔生成的.具体来讲事件序列指的是随机异步生成的时间戳序列,包含事件类型和事件发生的时间戳;时间序列指的是相同时间间隔内的信息记录序列,比如温度在每小时的记录信息.随机生成的时间戳使得事件序列和时间序列蕴含完全不同的信息.因为时间序列的时间仅起索引的作用,而事件序列中异步生成的时间戳携带丰富的动态信息,可以捕获系统的动态变化规律,因此本文同时利用两种序列做预测问题可以达到更高的预测效果.

序列预测模型的研究历程介绍如下:Aalen等[1]重点探索如何利用事件序列中时间戳信息研究序列的发展趋势,捕获系统的动态变化规律.Synder等[2]首先利用点过程建模事件序列,证实点过程中的强度函数是建模事件序列的强大数学工具.Zhou等[3-4]提出了先进的数学理论,使得点过程的强度函数得到优化.Xu等[5]根据先验知识为条件强度函数设计更加新颖的参数形式,进一步提高点过程建模效果.然而以上方法都是人为假设的强度函数,它将现实生活中的复杂情况简单化,得到的模型存在一定限制,不能准确地适用于现实应用场景中.针对这种情况,后来Zhou等[4]提出非参数形式来适应点过程的复杂性,但该方法仍具有选错模型的风险.随后Du等[6]提出recurrent marked temporal point process(RMTPP)模型,利用循环神经网络对事件序列进行点过程建模,忽略时间序列的特征,对强度函数的形式做部分参数化假设.Xiao等[7]提出了twin recurrent point process(TRPP)模型,用双LSTM对时间序列和事件序列分别建模,作为强度函数的背景知识和历史信息,解决了RMTPP忽略时间序列特征的问题.以上方法均是从时间维度上考虑建模强度函数,未考虑空间维度的影响.本文综合考虑空间和时间两方面的影响来建模强度函数,空间维度上具体考虑的是用户间的网络关系,时间维度上考虑的是强度函数的背景知识和历史影响,基于此提出社交化点过程序列预测算法(SPSP模型).

本文模型是对Xiao等[7]提出的TRPP模型的改进模型.具体过程介绍如下:首先在时间维度上利用双LSTM分别训练事件序列和时间序列,然后联合两个网络的表征作为强度函数的背景知识和历史特征生成预选的类型向量;最后在空间维度上根据深度时空社交网络的动态变化利用预选的类型向量更新强度函数的形式,优化训练整个模型,预测出最终事件类型和时间.

本文亮点介绍如下:

(Ⅰ)首次通过同时考虑时间维度和空间维度两方面建模点过程的强度函数,即将社交关系加入到点过程序列预测算法中,打开了建模强度函数的新思路;

(Ⅱ)SPSP模型具有可扩展性,可应用于多种实际场景.如微博预测、ATM故障预测、交通事故预测等.微博预测问题中社交关系网络为用户之间的好友关系网络;ATM的故障种类预测问题中,社交关系网络为ATM机的位置关系网络;十字路口事故预测问题中,社交关系网络为路口分布关联网络.本文主要从加入网络关系的角度改进点过程序列预测算法,从而提高模型的预测效果.

1 相关工作

SPSP模型的基础模块是循环神经网络,我们使用RNN建模两种序列数据,具体实验中采用是LSTM,因为它可以解决长期依赖性问题,并且效果较好.实际上也可以考虑其他变式RNN,如gated recurrent units(GRU).

(Ⅰ)RNN在序列数据上的应用:从应用场景上,RNN作为本文的构造模块,可应用于两种序列数据:时间序列和事件序列,两者可协同建模[7],具体介绍如下:

时间序列:指在相同时间间隔内记录相关数据统计信息的一种同步序列.在最近的时间窗口中时间序列可以及时捕获随时间变化的特征,因此RNN经常将时间序列作为输入做相关序列预测问题[8-10].例如,视频帧就可看作一种时间序列数据,

我们可以根据历史视频帧分析其内在联系，预测下一帧的内容. 最近 RNN 已经广泛应用于视频分析[11]和语音识别[12]等领域.

事件序列：指通过时间戳记录事件随机特征[13-14]的一种异步序列. 事件序列将随机生成的时间戳作为 RNN 的输入，使事件序列能更加高效地捕获事件间的长期依赖关系. 本文使用双 LSTM 对两种序列数据进行协同建模，既可以捕获同步信息的更新规律，又可以捕获突发信息的异步随机性，利用循环神经网络生成条件强度函数的非线性映射，避免了参数假设带来的限制.

(Ⅱ)点过程：一个时间序列或者事件序列的随机模型. 点过程分为事件点过程和物体点过程. 事件点过程：地震或者其他灾难事件；服务器的访问事件；工厂制造的不合格产品事件；三岔路发生的交通事故事件等. 带标记的点过程最初就是用来预测地震和余震的发生规律问题[15-17]；物体点过程：高速公路上汽车的位置；DNA 的基因等.

(Ⅲ)强度函数介绍：点过程作为建模序列模型的数学框架[1]，采用条件强度函数衡量点过程的动态变化性. 强度函数的定义为：在时间窗口 $[t+\mathrm{d}t)$ 中，$\lambda(t)$ 代表在历史事件 $H_t=\{z_i,t_i \mid t_i<t\}$ 发生的前提下新事件的发生概率，具体公式为

$$\lambda(t)=\lim_{\Delta t\to 0}\frac{E(N(t+\Delta t)-N(t)\mid H_t)}{\Delta t}=\frac{E(\mathrm{d}N(t)\mid H_t)}{\mathrm{d}t}$$

式中，$E(\mathrm{d}N(t)\mid H_t)$ 表示在历史 H_t 的基础上，在时间间隔 $[t+\mathrm{d}t)$ 内事件发生个数的期望值. 条件强度函数在点过程序列预测算法中扮演着关键性的作用，点过程建模效果随着强度函数参数化形式的不同而变化. 点过程的强度函数的变化过程包括：泊松过程[18]；加强版泊松过程[19-20]；霍克斯过程[21]；活性点过程[22]；自修正点过程[23]；RMTPP 模型；TRPP 模型.

强度函数的参数化形式均由两部分组成：背景知识和历史特征. 将以上方法总结如表 1 所示. 由表 1 可见，前 5 种强度函数是根据先验知识人为假设的参数模型，存在一定的限制，模型不能完全符合现实序列问题的复杂的动态变化. RMTPP 模型利用 LSTM 学习来自事件序列的历史特征的参数. 虽然忽略时间序列的背景特征，但是作为半参数化的强度函数，RMTPP 模型也取得了不错的效果；TRPP 模型使用双 LSTM 对时间序列和事件序列分别建模，作为强度函数的背景知识和历史信息，因此本文主要借鉴后两个模型的思想构造模型.

表 1　点过程强度函数

Tab. 1　Intensity function of point process

model	background knowledge	history influence
Poisson process	$\mu(t)$	0
Reinforced poisson process	0	$\gamma(t)\sum_{t_i<t}\delta(t_i<t)$
Hawkes process	$\mu(t)$	$\sum_{t_i<t}\gamma(t,t_i)$
Reactive point process	$\mu(t)$	$\sum_{t_i<t}\gamma_1(t,t_i)-\sum_{t_i<t}\gamma_2(t,t_i)$
Self-correcting process	0	$\exp(\mu t-\sum_{t_i<t}\gamma(t,t_i)$
RMTPP	$\exp(v^{t^T}\cdot h_j+w^t(t-t_j)+b_t)$	
TRPP	A nonlinear mapping by two LSTMs	

2　SPSP 模型

SPSP 模型的深度时空社交网络是一个同时考虑时间和空间动态变化的深度社交网络. 该模型的问题和目标介绍如下：

已知：网络的两种输入序列时间序列 $\{j_t\}_{t=1}^{T}$ 和事件序列 $\{z_i,t_i\}_{i=1}^{N}$.

目标：预测用户下次事件的类型 z 和时间 t.

2.1　时空深度社交网络

时空深度社交网络同时考虑空间和时间两个维度，建模用户与用户间的动态交互性而生成的深度动态网络，具体时空深度社交网络如图 1 所示. 本文的对比模型 TRPP 模型只是考虑时间维度上的信息，根据时间序列和事件序列的学习强度函数的背景知识和历史特征，但是忽略了空间维度上事件点间的相互影响. 由图 1 可知，对于相同时间范围内的

事件点之间存在着复杂的网络关系，随着时间的推移，事件点之间的网络关系也是动态变化的. 空间维度的网络关系和时间维度的序列演变是相互影响的. 空间维度上事件点 i 会受当前所在的网络关系 j 的影响，进而影响时间维度上事件点 i 下一刻的序列演变；反过来，每个时间维度事件点的动态变化也会影响空间维度网络关系 j 的形成；因此，时空两个维度相互关联，随时间动态演变. 如图 2 所示，对于用户 u_0 而言，在 time1 和 time2 时间段所处网络发生了动态变化，可能是 time2 时间段内用户 u_7 的出现，导致整个网络关系的动态变化，用户与好友之间的影响度也在动态变化(其中黑色圆点的大小代表影响的强弱). 同时考虑两个维度的影响，将能得到更加优化的模型.

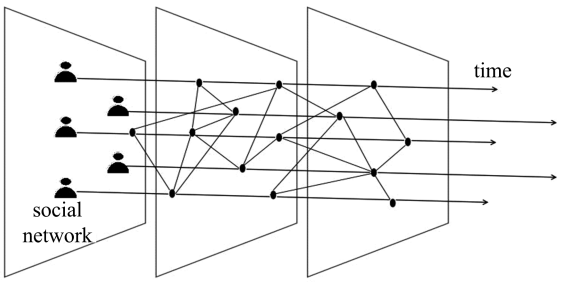

图 1 时空深度社交网络关系图

Fig. 1 Spatial and temporal depth social network diagram

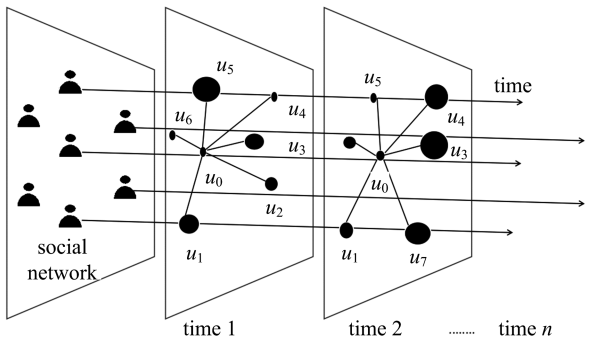

图 2 单用户时空网络动态图

Fig. 2 Dynamic graph of single user spatial and temporal network

该网络模型从时间和空间两个维度对点过程的强度函数进行建模，提高了强度函数的表达效果. 首先时间维度上利用双 LSTM 分别获取每个用户的时间序列表征和事件序列表征，联合训练所得的两个表征，作为用户的潜在表征，形成点过程强度函数的非线性映射；然后空间维度上，训练学习事件点之间的网络关系，通过多维时间点过程建模用户与用户的动态交互性；最后利用求得的用户间的交互性优化强度函数，提高综合模型的整体表达效果.

2.2 时间维度-变式 TRPP 模型

时间维度上我们借鉴 TRPP 模型[7]的思路，利用双 LSTM 分别建模事件序列和时间序列，联合两个 LSTM 的训练输出表征，通过 joint layer 形成对强度函数的非线性映射，得出预选事件类型和时间预测值，类型预选层用于空间维度上结合社交关系影响优化强度函数. 其中事件 embedding 层是全连接的，使用 tanh 激活函数并且输出一个 16 维的向量，进而通过事件序列 LSTM 层训练. 因为本文数据中类型数目较多，所以采用 embedding 层，使数据更加紧凑高效. 具体实验中的变式 TRPP 模型如图 3 所示.

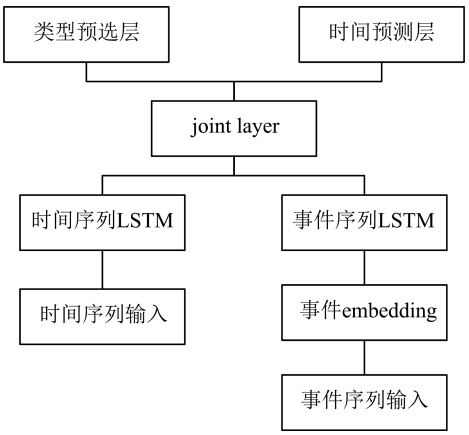

图 3 TRPP 模型

Fig. 3 TRPP model

2.3 空间维度-社交网络模型

社交网络模型主要利用类型表征的更新函数 $f_u(t)$ 和用户间的影响度概念.

定义 2.1 类型表征更新函数 $f_u(t)$. 对于每一个用户 u，使用时间维度上 TRPP 模型生成的低维向量 $f_u(t)$ 作为 t 时间用户 u 类型的初始表征. 用户的类型表征是随机变化的，因此这些表征向量随着时间动态变化，我们用更新函数 $f_u(t)$ 来建模表征的动态演化过程. 当用户发布一条新事件，函数就随之动态改变，函数的更新取决于整个神经网络，它集合了用户时间维度上的背景知识和历史影响，空间维度上社交关系影响. 具体函数形式为

$$f_u(t+1)=\sigma(W_1 y_t^j + W_2 y_t^z + \sum_{m=0}^{m} W_m f_m(t)) \tag{1}$$

式中，y_t^j 和 y_t^z 分别代表时间维度上的背景知识和历史影响，第 3 部分代表空间维度的社交影响. 根据用户影响度初始化社交影响中的参数 W_m. $m-1$ 代表用户 u 的关注用户数.

定义 2.2　用户影响度[24]. 本文用户影响度的概念采用文献[24]中的基于 PageRank 的微博用户影响度评估算法的思想. 一个偶像对其粉丝的吸引程度取决于 3 个因素：

(Ⅰ)用户 v 和其偶像 u 的兴趣相似度 SI_{vu}；

(Ⅱ)偶像 u 发帖活跃度 A_{vu}；

(Ⅲ)v 对偶像 u 所发内容的反馈程度 D_{vu}.

本文将用户 u 对其粉丝 v 的吸引程度表示为

$$F_{vu}=\begin{bmatrix}\mathrm{SI}_{vu}\\ A_{vu}\\ D_{vu}\end{bmatrix} \tag{2}$$

式中，$\mathrm{SI}_{vu}=\mathrm{ST}_{vu}+\mathrm{SF}_{vu}$. ST_{vu} 表示用户 v 和用户 u 之间的兴趣相似度，即利用余弦相似度计算用户 v 和用户 u 之间所发微博内容的文本相似度. SF_{vu} 表示用户 v 和用户 u 间的好友相似度，利用 Jaccard 相似度来计算. 具体公式为

$$\mathrm{ST}_{vu}=\frac{\sum_{k=1}^{h}w_k(t_v)\times w_k(t_u)}{\sqrt{\sum_{k=1}^{h}w_k(t_v)^2}\times\sqrt{\sum_{k=1}^{h}w_k(t_u)^2}},$$

$$w_k(t)=f_k(t)\times\log\frac{N}{n_k}.$$

式中，h 表示所有用户微博中关键字的总数，t 表示一用户所有文本内容，$w_k(t)$ 为关键字 k 在 t 中权重，N 为用户总数，n_k 代表事件内容中出现关键字 k 的用户总数. 即

$$\mathrm{SF}_{vu}=\frac{F(v)\cap F(u)}{F(v)\cup F(u)}.$$

式中，F 代表一个用户的邻居好友集合. 一般认为，一个偶像对其粉丝的影响程度与其自身的微博发布活跃度直接相关. 偶像活跃度越高，粉丝受影响越大. 偶像 u 相对其粉丝的事件发生活跃度为

$$A_{vu}=\frac{f(u)}{\sum_{k\in I(i)}f(k)}.$$

式中，A_{vu} 表示用户 v 相对于用户 u 的事件发生活跃度，$f(v)$ 表示用户 v 的事件发生频率，$I(u)$ 表示用户 u 所有偶像用户的集合，$f(k)$ 表示用户 u 的偶像用户 k 的微博发布频率.

用 D_{vu} 表示节点 u 对其邻居节点所发内容的反馈程度，公式为

$$D_{uv}=\frac{\lambda nr_{uv}+\mu nc_{uv}}{n_u}.$$

式中，n_u 表示用户 u 的事件内容总条数，nr_{uv} 表示用户 u 转发偶像 v 的总条数. nc_{uv} 表示 u 回复 v 内容的总条数，λ 和 μ 为调节系数. 本文取 $\lambda=0.50$，$\mu=1.00$.

社交媒体传播平台上，信息的传播往往受用户与用户的关系影响[25]，用户的好友发布/转发的内容将对用户产生一定的影响，因此采用此方法初始化参数 W_m .

利用用户之间影响度，将用户关注人的事件类型预测向量 $f_u(t)$ 与社交关系影响因子 F_i 和时间影响因子 T_i 相乘再求和即 $\sum_{i=1}^{m}T_iF_if_i(t)$，作为模型(1)中的第 3 部分的初始值(其中 F_i 表示用户 u 所关注的第 i 个用户的影响度).

社交网络中，只要有用户发布/转发新的微博，相应用户的 $f_u(t)$ 就会被更新，从而影响整个社交网络关系.

2.4　整体模型-SPSP 模型

本节主要结合前两节的变式 TRPP 模型和社交网络模型，综合介绍 SPSP 模型的整体框架. 首先介绍模型的基本模块 LSTM[26]的定义，本文使用双 LSTM 分别训练事件序列和时间序列. 循环神经网络的变式 LSTM 具体公式定义如下：

$$i_t=\mathrm{g}(W_ix_t+U_iy_{t-1}+V_ic_{t-1}+b_i),$$
$$f_t=g(W_fx_t+U_fy_{t-1}+V_fc_{t-1}+b_f),$$
$$c_t=f_tc_{t-1}+i_t\odot\tanh(W_cx_t+U_cy_{t-1}+b_c),$$
$$o_t=g(W_ox_t+U_oy_{t-1}+V_oc_t+b_o),$$
$$y_t=o_t\odot\tanh(c_t).$$

式中，i_t 表示输入门，f_t 表示遗忘门，o_t 表示输出门，c_t 表示细胞单元. o_t 和 y_t 共同表示输出门的计算公式，o_t 表示利用 g 函数确定单元状态的输出部分，最后 y_t 表示单元状态利用 tanh 处理之后和 o_t 相乘即为确定要输出的部分. LSTM 通过 3 个门决定细胞单元中记住什么遗忘什么. $\odot$ 代表数组元素依次相乘，函数 g 采用 sigmoid 激活函数. $\{x_t\}_{t=1}^{T}$ 表示输入序列，$\{y\}_{t=1}^{T}$ 表示生成隐含层状态. 以上公式可以简化为

$$(y_t,c_t)=LSTM(x_ty_{t-1}+c_{t-1}) \tag{3}$$

网络的输入序列由两部分组成：时间序列 $\{j_t\}_{t=1}^{T}$ 和事件序列 $\{z_i,t_i\}_{i=1}^{N}$. 两个序列分别用于学习背景知识和历史影响，运用双 LSTM 分别训练时间序列和事件序列表示为

$$(y_t^j,c_t^j)=\mathrm{LSTM}_j(j_t,y_{t-1}^j+c_{t-1}^j) \tag{4}$$
$$(y_t^z,c_t^z)=\mathrm{LSTM}_z(z_t,y_{t-1}^z+c_{t-1}^z) \tag{5}$$

将双 LSTM 的输出 y_t^j 和 y_t^z 通过 joint layer 利用 tanh 函数合并得到图 3 中的 joint layer 的方程为

$$e_t=\tanh(W_f[y_t^j,y_t^z]+b_f) \tag{6}$$

对 e_t 利用 softMax 函数得到

$$f_u(t)=\text{softMax}(W_u e_t+b_u) \tag{7}$$

式中，$f_u(t)$ 代表类型.

以上过程为时间维度上的变式 TRPP 模型的训练过程. 接下来介绍空间维度上的优化模型过程.

$$f_u(t+1)=\sigma(W_1 y_t^j+W_2 y_t^z+\sum_{m=0}^{m} T_m I_m f_m(t)) \tag{8}$$

$$s_t=W_s e_t+b_s \tag{9}$$

式(8)表示将得到的事件点的预选表征在空间维度上加入社交关系影响更新优化的 $f_u(t+1)$，然后经过分类损失层，通过时空深度网络的不断迭代训练得出最优事件点预测. 最后介绍式(9)即图 3 的时间预测层，时间预测采用和 TRPP 模型相同策略，均使用回归损失函数做误差损失，其中 s 代表时间.

本实验采用的损失函数是时间损失和类型损失之和，这样不仅保证了预测类型的准确率，还保证预测时间的准确率，损失函数的公式为

$$\sum_{l=1}^{N}(-w_u^l\log(u_t^l)-\log(f(s_t^l\mid y_{t-1}^l))) \tag{10}$$

式中，N 代表用户总数，事件点通过 l 来索引. s_t^l 是下次事件点的时间戳，y_{t-1}^l 代表的是时间点的历史信息. 对于时间损失函数部分采用高斯惩罚函数

$$f(s_t^l\mid y_{t-1}^l)=\frac{1}{\sqrt{2\pi\sigma}}\exp(\frac{-(s_t^l-\tilde{s_t^l})^2}{2\sigma^2}) \tag{11}$$

本实验采用的 RMSprop 梯度算法[27]对于训练神经网络模型具有很好的学习效果.

SPSP 模型框架流程如图 4 所示. 图 4 下部分代表时间维度上的变式 TPRR 模型，上部分的网络关系层是本文模型的主要创新，即空间维度. 由图 4 可知，本模型通过时间预测层和事件类型预测层分别输出下个事件点的预测类型和相应的时间戳信息. 对事件类型的预测采用平方损失函数，对事件时间的预测采用交叉熵损失函数.

SPSP 模型训练过程由时间维度和空间维度两部分组成，时间维度采用 TPRR 模型的思路，即通过双 LSTM 分别学习时间序列和事件序列建模强度函数的背景知识和历史影响，得出用户类型的初步预测作为空间维度 $f_u(t)$ 的类型输入. 然后空间维度上进一步通过综合社交关系影响和时间维度的表征，重新构造动态变化的事件类型 $f_u(t)$. 最后根据分类损失层和回归损失层经过多次迭代不断训练模型，得到最终预测的事件类型和事件时间戳.

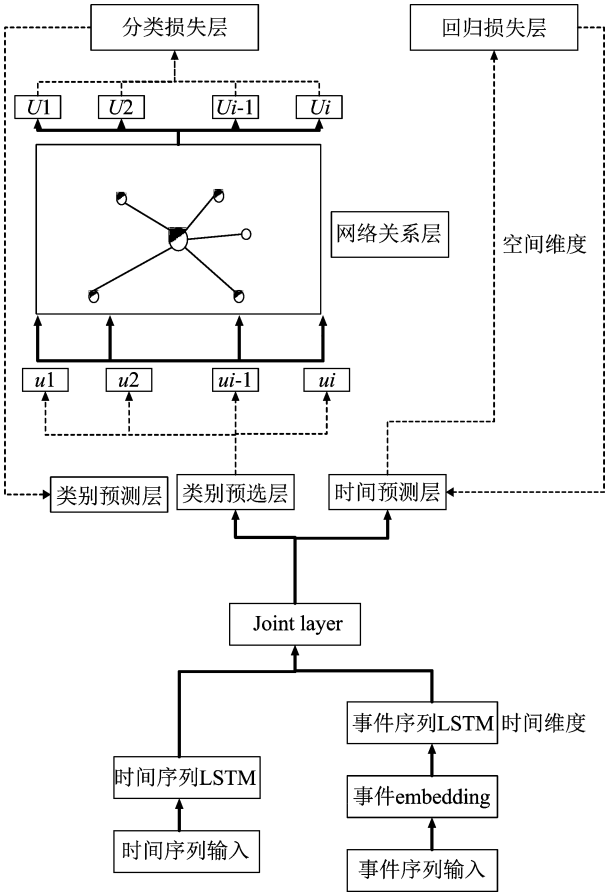

图 4 SPSP 模型框架图

Fig. 4 SPSP model diagram

3 实验及分析

本文将预测用户下次发布微博类型作为基于事件的点过程建模的例子，使用新浪微博数据集验证 SPSP 模型在实际应用中的效果. 对于点过程的建模，本实验不采用先验知识，即没有提前对模型作参数化假设. 实验数据中的社会层级、工作领域、年龄、性别、喜好各有不同. 在背景条件复杂多变的情况下对模型做参数化设计耗时耗力，而且不一定能符合实际情况，因此本实验用循环神经网络 LSTM 学习点过程强度函数的背景知识和历史影响.

3.1 研究问题和数据集描述

本文研究问题是根据用户发布或转发微博的历史事件序列和时间序列预测用户下一次发布微博的类型和时间戳.

数据集中微博用户通过关注-被关注彼此关联，数据集中包括 170 万用户数据，好友关系数据达 3 亿. 每个微博用户大约有 200 个粉丝. 本实验抓取 1 700 个用户的最近 3 年发布的微博作为训练数据. 表 2 列出了数据集的相关信息.

表 2　新浪微博数据

Tab. 2　Sina weibo data

数据集	用户数	关注关系数	原始微博数	转发微博数
新浪微博	1 700	308 489	3 000	23 755

3.2　用户模型

实验训练集数据和测试集数据没有交叉、独立分开. 训练数据由 1 500 个微博用户的 3 年的微博信息组成，测试数据由 200 个微博用户的微博信息组成，总共 1 700 个微博用户的 3 年的微博信息，覆盖个体用户和机构或组织用户. 个体用户又分为名人用户、普通用户，名人用户指的是娱乐明星、业内专家、学者等. 机构或组织用户包括：政府机构、公司企业等.

3.3　事件类型

本实验将所有用户发布的微博运用聚类算法分为 10 类：①闲趣美妆；②两性情感；③教育政治；④社会生活；⑤综艺影视；⑥社交媒体；⑦金融数码；⑧小说动漫；⑨饮食养生；⑩医疗保健. 其中聚类算法采用的是 word2vec 中的 K-均值聚类算法，它是一种基于原型的、中心的距离技术，具有简洁高效的优势. 图 5 给出了各事件类别数据统计图.

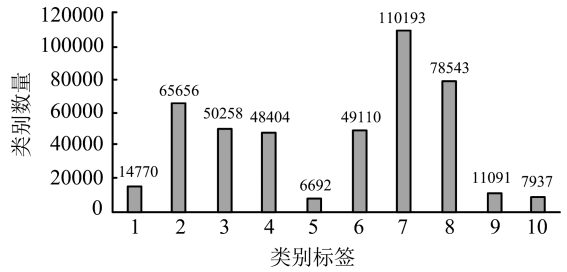

图 5　事件类别数据统计

Fig. 5　Event category data statistics

3.4　输入特征

两个 LSTM 的输入特征是：

（Ⅰ）时间序列 LSTM. 实验以一个月作为子窗口，在时间序列 LSTM 中提取的特征包括：①微博用户详细信息：用户 ID、年龄、位置、性别等；②事件类型统计信息，包括每月十类微博类型的出现次数，它们的出现频率被用作特性. 以上两类特征的连接作为每个子窗口的特征，即时间序列点.

（Ⅱ）事件序列 LSTM. 微博事件类型和每两条微博之间的时间间隔. 其中事件时间间隔统计图如图 6 所示.

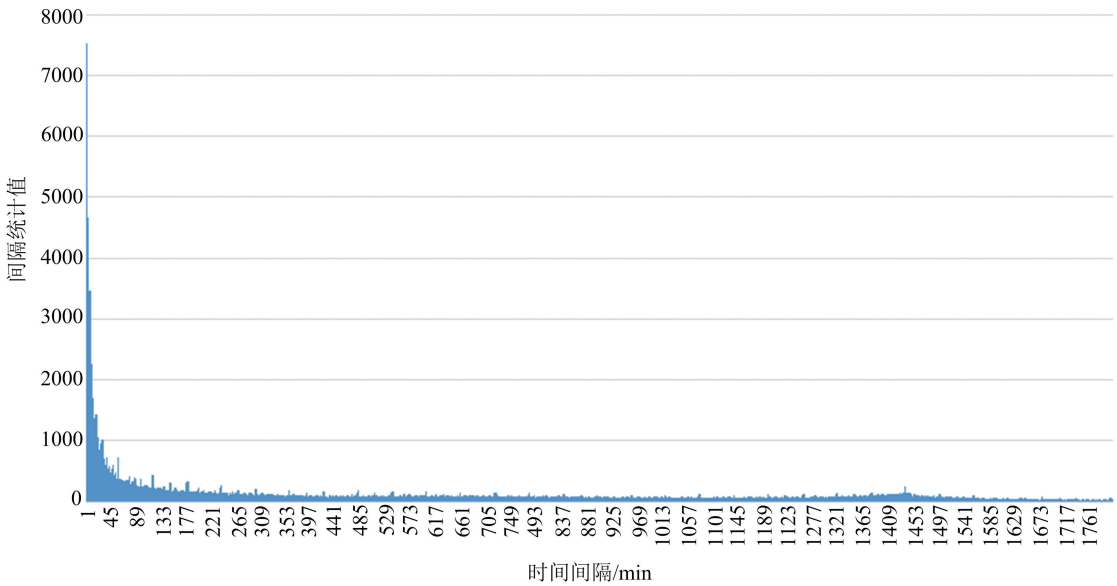

图 6　时间间隔统计图

Fig. 6　Graph of time interval statistics

3.5　模型设置

实验中时间序列 LSTM 的隐含层大小为 16，隐含层采用 sigmoid/tanh 激活函数. 事件序列 LSTM 的隐含层大小为 16，使用 tanh 激活函数. 实验中事件序列表征是由事件序列的类型和时间戳联合表征，其中事件类型采用 one-hot 表征.

对于时间序列 LSTM，实验设置每个子窗口的长度为一个月，子窗口的个数为 5. 实验参数设置为学习率{0. 1，0. 01，0. 001，0. 004}；训练批次{50；100；200}.

3.6　对比试验与评测标准

为了验证 SPSP 模型有效性，将 SPSP 模型和 7 种模型进行对比.

（Ⅰ）事件序列 RNN(ERNN)：删除时间序列的

输入，只训练事件序列. 这种方法与文献[7]采用的对比方法一致.

(Ⅱ) SERNN：社交化事件序列 ERNN 模型，即将第一种模型加入空间维度的社交网络关系影响，验证只在事件序列上加入空间维度的社交网络关系可否提高模型预测准确度.

(Ⅲ)时间序列 RNN(TRNN)：删除事件序列的输入，只训练时间序列. 这种设计与文献[9]采用 LSTM 视频分析的方法类似，即帧序列可以作为时间序列作为 LSTM 的输入.

(Ⅳ) STRNN：社交化时间序列 TRNN 模型，类似于模型(Ⅱ)的思想，验证只在时间序列上加入空间维度上的社交网络关系的影响可否提高模型预测准确度.

(Ⅴ) TRPP 模型：即强度函数 RNN，训练两种序列信息. 此模型即文献[7]提出的模型，也就是本文要改进的模型.

(Ⅵ) Logistics 模型：对事件时间戳预测采用逻辑回归；对类型预测采用逻辑回归分类算法.

(Ⅶ) RMTPP 模型：该模型由 Du 等[6]提出，通过 RNN 训练事件序列，对强度函数做部分参数化.

评估标准：实验采用序列预测算法的常用评估方法. 对事件类型的预测，采用准确率、召回率、$F1$ 值. 所有的这些评测方法是对每一种类型来计算的，然后计算平均值作为最终结果. 对于时间戳的预测采用准确率和 MAE 两种评测标准. 其中事件时间戳的准确率指的是预测时间和真实时间差值的绝对值在 1 小时以内记为预测准确.

3.7 结果和讨论

实验均表明 SPSP 模型具有良好的效果. 在新浪微博数据集上的事件类型和时间预测准确率实验结果如表 3 所示.

表 3 多种方法预测评估对比表

Tab. 3 Comparison table of multiple methods evaluation prediction

model	type accuracy	time accuracy	precision	recall	F_1 score	MAE
ERNN	0.560	0.64	0.558	0.533	0.545	4.52
SERNN	0.590	0.65	0.56	0.54	0.55	4.30
TRNN	0.553	0.564	0.556	0.512	0.533	4.48
STRNN	0.554	0.566	0.562	0.520	0.540	4.40
TRPP	0.610	0.73	0.616	0.614	0.615	4.20
Logistic	0.375	0.55	0.261	0.273	0.267	4.61
RMTPP	0.564	0.73	0.561	0.563	0.562	4.31
SPSP	0.690	0.89	0.676	0.680	0.678	4.00

为使结果更加形象，我们将实验结果绘制成柱状图，由图 7～12 可看到，SPSP 模型在各方面均较对比方法有所改进.

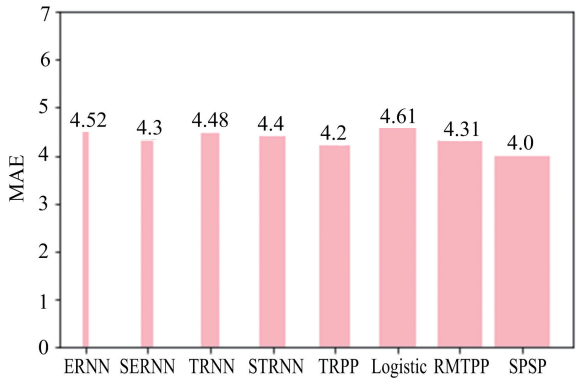

图 7 MAE 对比图

Fig. 7 MAE comparison

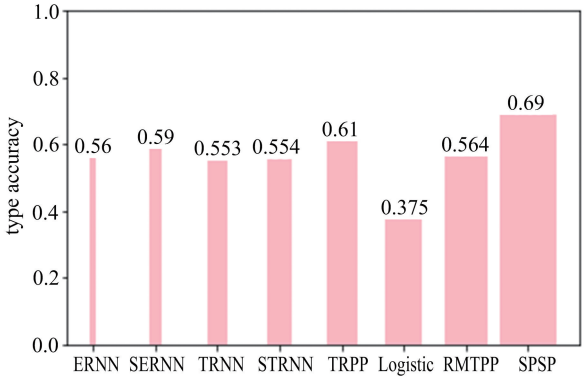

图 8 类型准确率对比图

Fig. 8 Type accuracy comparison

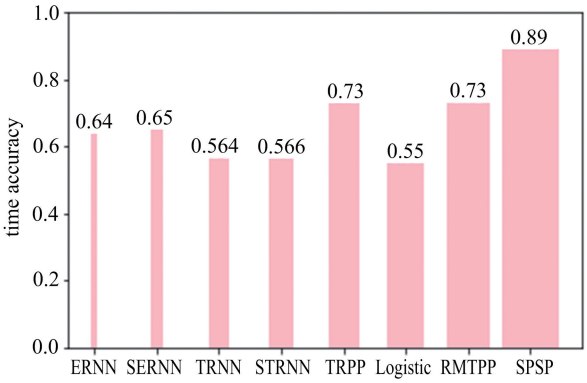

图 9 时间准确率对比图

Fig. 9 Time accuracy comparison

由图 9 可以直观地看到，SPSP 模型在时间准确率的预测方面具有较大的提升效果. 图 8 表明，SPSP 模型的类型准确率也有所提升. 这证明了本文 SPSP 模型的有效性.

基于实验结果，经过分析讨论得出以下结论：

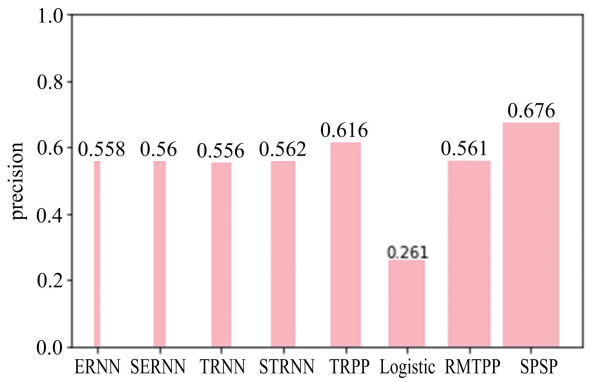

图 10　精确率对比图

Fig. 10　Precision comparison

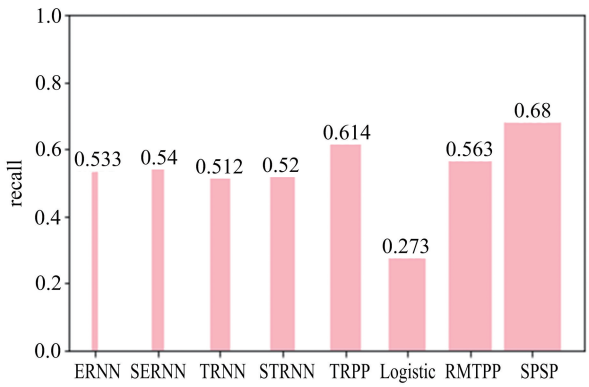

图 11　召回率对比图

Fig. 11　Recall rate comparison

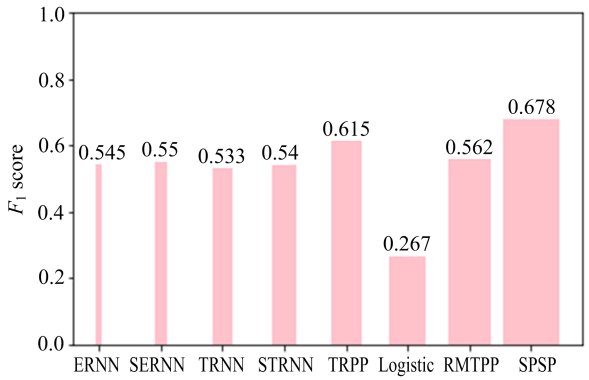

图 12　F_1 值对比图

Fig. 12　F_1 value comparison

(Ⅰ)由图 7～12 可知,本文的 SPSP 模型无论是在事件类型预测效果还是事件时间的预测效果上都比仅考虑时间维度的 TRPP 模型的实验效果好.

(Ⅱ)比较前 3 种方法效果可知,强度函数 RNN>事件序列 RNN>时间序列 RNN. 说明历史事件比背景知识的影响比重更大.

(Ⅲ)SPSP 模型相比经典方法 Logistic 模型和 RMTPP 模型具有更加优异的效果.

(Ⅳ)当事件序列 RNN 和时间序列 RNN 加入空间维度后,模型的预测效果也得到了相应的提高,并且社交化的事件序列 RNN 模型的效果的提高更加明显. 说明社交关系和历史影响更适宜结合,同时考虑两种信息能达到更好的序列预测效果.

(Ⅴ)由图 9 可知,SPSP 模型的时间准确率比其他对比方法有了很大的提高,从原来的 73%提高到了 89%.

(Ⅵ)本模型的优势在于提高准确性上没有考虑模型资源的占用情况. 对于计算效率,由于神经网络迭代训练,需要耗费较多时间,所以需要进一步改进计算效率的问题.

4　结论

霍克斯过程完全参数化建模强度函数,RMTPP 模型部分参数化强度函数,TRPP 模型利用双 LSTM 学习强度函数的背景知识和历史影响,避免人为参数化;然而以上方法都是从时间维度上纵向考虑,没有从空间维度上横向考虑. 本文提出的 SPSP 模型从纵向、横向两个维度建模点过程的强度函数,具体考虑了社交关系网络对于建模强度函数的影响,进一步改进了点过程强度函数的建模效果. 本文的模型首次同时利用时间、空间两个维度去建模点过程强度函数,为进一步改进该模型提出新的研究方向. 并且 SPSP 模型具有可扩展性,可应用于多种实际场景. 经过实验证明本模型在实际应用中具有良好的表现.

参考文献(References)

[1] AALEN O, BORGAN Ø, GJESSING H K. Survival and Event History Analysis: A Process Point of View[M]. Berlin: Springer Science & Business Media, 2008.

[2] SNYDER D L, MILLER M I. Random point processes in time and space [M]. Berlin: Springer Science &-Business Media, 2012.

[3] ZHOU K, ZHA H Y, SONG L. Learning social infectivity in sparse low-rank networks using multi-dimensional Hawkes processes [J]. Proceedings of Machine Learning Research, 2013, 31: 641-649.

[4] ZHOU K, ZHA H Y, SONG L. Learning triggering kernels for multi-dimensional Hawkes processes[C] // Proceedings of the 30th International Conference on Machine Learning. Atlanta, USA: ICML, 2013: 1301-1309.

[5] XU H T, WU W C, NEMAT S, et al. Patient flow

prediction via discriminative learning of mutually-correcting processes[J]. IEEE Transactions on Knowledge and Data Engineering, 2017, 29(1): 157-171.

[6] DU N, DAI H J, TRIVEDI R, et al. Recurrent marked temporal point processes: Embedding event history to vector[C]// Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Francisco, USA: ACM, 2016: 1555-1564.

[7] XIAO Shuai, YAN Junchi. Modeling the intensity function of point process via recurrent neural networks [C]// Thirty-First AAAI Conference on Artificial Intelligence. San Francisco, 2017: arXiv:1705.08982v1.

[8] CHEN P A, CHANG L C, CHANG F J. Reinforced recurrent neural networks for multistep-ahead flood forecasts[J]. Journal of Hydrology, 2013, 497: 71-79.

[9] CHUNG J, GULCEHRE C, CHO K, et al. Empirical evaluation of gated recurrent neural networks on sequence modeling[C]// Presented in NIPS 2014 Deep Learning and Representation Learning Workshop. 2014. In arXiv:1412.3555.

[10] BENGIO S, INYALS O, JAITLY N, et al. Scheduled sampling for sequence prediction with recurrent neural networks[C]// Proceedings of the 28th International Conference on Neural Information Processing Systems. Montreal, Canada: MIT, 2015: 1171-1179.

[11] JAIN A, SINGH A, KOPPULA H S, et al. Recurrent neural networks for driver activity anticipation via sensory-fusion architecture [C]// International Conference on Robotics and Automation (ICRA). Stockholm, Sweden: IEEE, 2016: 3118-3126.

[12] TRIPATHI S, LIPTON Z C, BELONGIE S, et al. Context matters: Refining object detection in video with recurrent neural networks [C] // York, UK: BMVC. 2016(9):19-22.

[13] ESTEBAN C, STAECK O, BAIER S, et al. Predicting clinical events by combining static and dynamic information using recurrent neural networks[C]// IEEE International Conference on Healthcare Informatics. Chicago, USA: IEEE, 2016: arXiv:1602.02685.

[14] CHE Z P, PURUSHOTHAM S, CHO K, et al. Recurrent neural networks for multivariate time series with missing values[J]. Scientific Reports, 2016, 8 (1): 6085-6097.

[15] HAWKES A G. Spectra of some self-exciting and mutually exciting point processes [J]. Biometrika, 1971, 58(1): 83-90.

[16] HAWKES A G, OAKES D. A cluster process representation of a self-exciting process[J]. Journal of Applied Probability, 1974, 11(3): 493-503.

[17] OGATA Y. Space-time point-process models for earthquake occurrences[J]. Annals of the Institute of Statistical Mathematics, 1998, 50(2): 379-402.

[18] KINGMAN J F C. Poisson processes[M]// Vol. 3, Oxford : Oxford university press, 1992.

[19] SHEN H W, WANG D S, SONG C M, et al. Modeling and predicting popularity dynamics via reinforced Poisson processes[C]// Québec, Canada: AAAI, 2014: .

[20] PEMANTLE R. A survey of random processes with reinforcement[J]. Probability Survey, 2007, 4(0): 1-79.

[21] HAWKES A G. Spectra of some self-exciting and mutually exciting point processes[J]. Biometrika, 1971.

[22] ERTEKIN S, RUDIN C, MCCORMICK T H. Reactive point processes: A new approach to predicting power failures in underground electrical systems[J]. The Annals of Applied Statistics, 2015, 9(1): 122-144.

[23] ISHAM V, WESTCOTT M. A self-correcting pint process[J]. Advances in Applied Probability, 1979, 37: 629-646.

[24] OU W, OU B Y, XIE Z F, et al. A PageRank-based algorithm to estimate microblog users' influences[J]. 计算机与现代化,2013,12: 1006-2475.

[25] MYERS S A, ZHU C G, LECKOVEC J. Information diffusion and external influence in networks [C]// Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining. Beijing: ACM, 2012: 33-41.

[26] GRAVES A. Generating sequences with recurrent neural networks [M]// Computer Science. 2013: arXiv:1308.0850.

[27] DAUPHIN Y N, DE VRIES H, CHUNG J, et al. RMSProp and equilibrated adaptive learning rates for non-convex optimization [J/E]. Machine Learning, 2015: arXiv:1502.04390.