南京大学

研究生毕业论文
(申请硕士学位)

论文题目 基于深度学习与点过程的事件序列预测算法研究

作者姓名 王言

学科、专业名称 计算机科学与技术

研究方向 人工智能

指导教师 申富饶 教授

2022 年 6 月 2 日

学　　号：MG1937025

论文答辩日期：2022 年 5 月 17 日

指 导 教 师：　　（签字）

Event Sequence Prediction Algorithm Based on Deep Learning and Point Process

by

Yan Wang

Supervised by

Professor Furao Shen

A dissertation submitted to
the graduate school of Nanjing University
in partial fulfilment of the requirements for the degree of
MASTER
in
Computer Science and Technology

School of Artificial Intelligence
Nanjing University

May 17, 2022

南京大学研究生毕业论文中文摘要首页用纸

毕业论文题目：基于深度学习与点过程的事件序列预测算法研究

计算机科学与技术 专业 2019 级硕士生姓名：王言

指导教师(姓名、职称)：申富饶 教授

摘　要

随着互联网和大数据技术的发展，针对事件序列的分析在现实世界中被广泛运用，例如社交网络分析，用户行为建模以及疾病的早期干预。由于序列中的事件间存在潜在的相关性，对这种相关性进行挖掘和利用具有重要的社会价值，事件序列预测则是其中重要的研究问题。点过程算法是建模事件序列预测问题的重要方式，随着数据规模的不断增大，基于深度学习的点过程算法逐渐成为主流方法。然而，在许多真实场景下，由于存在事件间关系复杂、数据分布具有偏向性等问题，现有事件序列预测模型仍面临挑战。

本文关注事件序列预测问题，从表示学习的角度出发，研究如何优化模型以提升其处理复杂事件数据的能力。本文分析现有模型面临的局限性，聚焦其中两个主要问题进行研究，分别是的复杂事件关系建模和历史序列中的长尾分布问题。针对上述问题，本文提出两种基于深度学习的点过程模型，并将其应用在实际系统中。本文的主要贡献如下：

针对事件间复杂关系建模的问题，本文提出一种基于渐进生成图的时序点过程模型。该模型重新定义事件关系模型框架，首先将事件间的复杂关系视为隐变量，之后设计一种渐进生成图模型学习历史事件间的关系，该模型由简单的前置图和复杂的多维关系图构成，可以渐进式的学习事件间关系并具有复杂关系推理的能力。并且我们针对不同历史事件构建关系图结构，而非事件类型，从而实现动态关系建模。我们引入多维图对事件关系进行定义，使得模型可以表示事件间不同类型的影响关系。实验结果验证此模型在处理标记类型较多的事件序列预测问题上具有显著优势。

针对事件序列中的长尾分布问题，本文提出一种基于双重平衡软标签的辅助训练模型。我们将事件序列预测问题中的长尾分布问题定义为离散标记空间和连续时序空间中的长尾问题。为了缓解长尾分布对于模型特征表示学习的影响，我们引入解耦学习框架，并设计软标签生成方式对此进行优化。具体而言，我们提出一种辅助模型，设计三种辅助模型与原模型之间的的信息传输通道，使其为解耦学习的两个阶段分别生成对应软标签。针对时序空间中长尾分布问题，我们设计一种连续空间上的软标签，并且使用标签分布平滑方法进行代价敏感学习。实验结果验证我们的模型在处理真实数据上具有一定优势。

最后我们将所提出的两种模型应用于真实场景下，即图书推荐管理系统中，对用户行为进行分析建模并为其提供图书推荐服务，验证本文所提出模型的实用性。本文对所提出的方法进行总结分析，归纳并介绍了在我们工作的基础上需要进一步探索的研究方向。

关键词： 事件序列；点过程；深度学习；表示学习

南京大学研究生毕业论文英文摘要首页用纸

THESIS: Event Sequence Prediction Algorithm Based on Deep Learning and Point Process

SPECIALIZATION: Computer Science and Technology

POSTGRADUATE: Yan Wang

MENTOR: Professor Furao Shen

ABSTRACT

With the development of Internet and big data technology, the analysis of event sequence data is widely used in the real world, such as social network analysis, user behavior modeling, and early intervention of diseases. Due to the potential correlation between events, mining this kind of correlation has important social value, and event sequence prediction is an important research direction in this field. The point process is an important method for modeling event sequence prediction problems. As the size of the dataset continues to increase, the point process based on deep learning has become the main method gradually. However, in many practical applications, the existing event sequence prediction models still have difficulties due to the complex relationship between events and the biased data distribution.

This paper focuses on the problem of event sequence prediction. From the perspective of representation learning, we explore how to optimize the model and improve its ability to deal with complex event sequence data. This paper analyzes the limitations of existing models and focuses on two main problems: modeling complex relationship between events and dealing with long-tailed distribution in historical sequence. To solve the above problems, we propose two point process models based on deep learning and apply them to our actual systems. The main contributions of this paper are as follows:

Firstly, in order to model the complex relationship between events, we propose a Temporal Point Process model based on Progressive Generative Graph. PGG-TPP redefines the model framework. The complex relationship between events is regarded

as implicit variable. Then, a progressive generative graph model is designed to learn the relationship between historical events, which is composed of a simple front graph and a complex multi-view relation graph, so that the model can gradually learn the relationship between events and has the ability of complex relationship reasoning. We define event relationships as multi-view graphs, so that the model can represent different types of influence relationships between events. Moreover, we build relation graph for different historical events, rather than event types, so as to realize dynamic relationship modeling. The experimental results demonstrate that our model has significant advantages in dealing with event sequence prediction with more types of markers.

Secondly, for the long-tailed distribution in event sequence data, we propose an auxiliary training model based on Dual-Balanced Soft Labels. We define the imbalance problem in event sequence prediction as long-tailed distribution in discrete marker space and continuous temporal space. In order to alleviate the impact of long-tailed distribution on feature representation learning, we use decoupled learning framework and design a soft label generation method for optimization. We propose an auxiliary model, and design three information transmission channels between the auxiliary models and the original model, so as to generate corresponding soft labels for the two stages of decoupled learning. For long-tailed distributions in temporal series space, we design a type of soft label on continuous space, and use the Label Distribution Smoothing for cost-sensitive learning. The experimental results demonstrate that our model has certain advantages in dealing with real data.

Finally, we apply our two models to the book management system we built, model user behavior and provide book recommendation service to verify the practicability of the model proposed in this paper. After that, we summarize the proposed methods and introduce several directions which can be further explored on the basis of our work.

KEYWORDS: Event sequence, point process, deep learning, representation learning

目　　录

插图清单

附表清单

第一章　绪论

1.1　研究背景及意义

时间序列分析（Time Series Analysis）是机器学习领域重要的研究方向之一，被广泛应用于气候建模，数字金融，通信工程等领域。在我们的日常生活中，时间序列相关数据每时每刻都在被记录，研究人员通过分析这些时序数据，发现其中的潜在规律，可以对动力系统进行显式建模，获取有效系统信息并对其进行利用。事件序列预测（Event Sequence Prediction）是时间序列分析的一个子领域。与采样时间间隔为恒定值的一般时间序列数据不同，事件序列（Event Sequence）是在连续时间内上，对离散发生事件进行采样所得到的序列数据，其在时间域和空间域上通常是非均匀分布的，即事件之间的间隔通常是非恒定的，具有一定的随机性。事件序列中每个采样节点表示一个自然发生的事件，记录该事件发生的时间和类型，例如类型可能是事件发生的地点信息。序列中的事件并不是独立发生的，当前时刻事件是否发生会受到历史事件的影响。因此事件序列预测任务要求我们对历史事件序列进行分析，建模事件之间的影响关系，并对未来事件发生的时间和类型进行准确的预测，为我们做出决策和计划提供辅助信息。

随着数字设备的普及和互联网时代的到来，各种现实世界中产生的行为被记录，形成事件序列数据。这些数据涉及我们生活的方方面面，例如社交软件用户在平台上的发帖、评论和分享；智能家居系统中户主与各种连接设备的交互过程；教育系统中每个学生参与的课程种类和考试情况，这些用户行为都可以被视为随时间推移的事件流数据，等待被进一步的分析和挖掘。事件序列的相关研究被应用在社交网络分析[1-4]、精准医疗[5-7]、犯罪行为建模[8-9]、金融分析[10-14] 等领域，在我们的实际生活中发挥了重要作用。因此，如何构建一个预测准确、具有一定泛化能力和鲁棒性的事件序列预测模型成为一个重要的课题。

近年来，随着设备计算能力的提升和高质量数据集的构造，深度学习技术（Deep Learning）在计算机视觉（Computer Vision)[15-18]，自然语言处理（Natural

Language Processing)[19-22]，强化学习（Reinforcement Learning)[23-26] 等领域取得显著的成功。深度学习可以学习高阶非线形的数据表示，对其进行自动特征提取，弱化了研究人员对于手工特征工程的需求；可以构造端到端的学习模型，直接建模样本空间和标签空间之间的复杂映射，简化了模型的部署流程，并且具有应用场景所需的泛化能力。得益于这些优点，深度学习模型有能力处理日趋复杂的真实场景数据，基于深度学习的事件序列预测模型已经被亚马逊、阿里巴巴、华为等公司用于实际业务场景中。我们列举事件序列预测技术在真实场景下的实际应用。

随着医疗水平的提升，疾病的早期发现可以最大程度控制其损害，因此进行有效的疾病早期预警可以帮助患者尽早开展干预，尽可能降低风险。随着医院信息化管理水平的提升，电子病历逐渐普及，使得针对不同患者的个性化疾病早期预警成为可能。电子病历存储着有关患者每次在医院进行检查、诊断和药物治疗的相关信息，包括具体的就诊情况和时间戳，这些记录可以视为事件序列。针对患者历史诊断数据进行分析，可以实现对部分疫病的有效预警[27]。基于机器学习的事件序列预测模型可以有效地解决此类问题[28]，提高早期预警的准确度，挽回疾病对患者家庭和社会造成的各种损失。

对企业而言，进行有效的设备异常检测，尽早发现设备故障进行维护，是保证企业生存发展、降本增收的重要方式。规范运行的企业设备通常会存档系统日志，这些系统日志都可以自然地表示为连续时间上的离散事件，检测此类数据中的异常可以为企业提供巨大的工业价值。例如，互联网公司所使用的服务器会输出系统日志，日志中的异常条目可能对应于未被注意到的服务器故障。由于此类数据规模庞大，手动检查此类事件序列数据通常是低效而不可行的；并且由于所使用软件更新频繁，人工制定的规则实效性较差。事件序列预测模型可以针对历史数据进行挖掘，并自动检测异常事件序列进行设备预警，辅助设备正常运行[29-30]。

在亚马逊、淘宝等电子商务平台上，针对用户的商品推荐发挥着至关重要的作用。随着时间的推移，用户的偏好和商品的受欢迎程度都是动态变化的，如何捕捉这种动态变化是许多电商平台面临的挑战。用户在电商平台的行为序列，例如在平台上浏览或购买商品，是典型的事件序列数据。基于对话的推荐系统要求算法分析用户行为数据，是推荐系统的重要分支，主要通过针对用户与平

台交互的顺序依赖性进行建模，得出用户的实时兴趣，为用户建议可能感兴趣的商品[31-33]。事件序列预测模型是构建基于对话推荐系统的有效方式之一，可以为用户提供有效的个性化推荐，提升平台收益。

在人工智能技术高速发展的背景下，本文旨在研究基于深度学习和点过程的事件序列预测模型，以实现对于用户行为数据的高效利用，本文的方法已经应用在实际系统之中对用户行为进行有效分析。

1.2 研究现状及挑战

早在二十世纪七十年代，对离散的事件序列的相关研究就已经开始吸引人们的注意，Hawkes 等人在论文[34-35] 中对离散的事件序列预测问题进行规范化，并提出一种建模未来事件发生概率的算法，由于事件流数据在我们身边大量生成，此研究引起各领域科学家的广泛关注。例如 Ogata 等人在论文[36] 中将相关模型引入地震学进行分析，之后相关模型成为地震序列分析的重要工具[37-39]；Hasbrouck 等人[40] 利用此处理离散事件数据的模型对高频金融数据进行分析，至今为止，相关方法及其变体仍是高频计量经济学中常用的模型[41-43]。随着事件序列预测基础理论的不断发展，相关研究还被广泛应用于生态学[44-45]、疫情传播[46-47]、犯罪预测[8-9] 等领域，为其提供有力的理论保障。

时至今日，事件序列预测已经成为时间序列分析领域重要研究方向之一。以预测目标为分类依据，常见的时间序列分析任务主要分为四类，分别是时序预测、时序分类、时序聚类以及异常检测，本文所关注的事件序列预测任务属于时序预测领域。事件序列预测要求我们根据观察到的历史事件序列，预测未来时刻会发生哪种类型的事件。基于此目标，良好的事件序列模型应该研究序列中不同事件之间的影响关系，以及事件再次发生的动力学机制，根据历史事件预测未来事件的动态规律，从而我们可以针对模型的预测结果设计有效的控制措施进行干预，引导事件的动态走向我们期望的结果。序列建模（Sequence Modeling）是实现时序预测任务中特征学习的重要方法。对事件序列进行建模，学习输入事件序列数据的特征表示，对于预测未来事件至关重要。以序列建模为基础，现有针对事件预测的主流建模方法主要分为四类，分别为基于传统机器学习的建模方法、基于深度学习的建模方法、基于对抗训练的建模方法以及基于强化学

习的建模方法，我们在本节中分别介绍这四种建模方式。

为了建模一系列离散发生的事件，我们需要思考如下问题：事件的发生率是否随时间而变化？历史事件会影响其他事件的触发概率吗？所触发的事件在时间域和空间域上是如何分布的？为了回答上述问题，在介绍四种具体的建模方法之前，我们先简单介绍时序点过程（Temporal Point Process），其为事件序列分析的重要工具。时序点过程是该领域经典的数学工具[48]，对上述问题进行合理假设，认为事件的发生概率会随着时间的推移而变化，并且相关的历史事件会影响未来时刻发生事件的概率。在此基础上对事件的概率分布进行建模，将连续时间上的事件序列描述为一个随机过程，使得模型不仅可以表示事件流上离散发生的事件，同时可以对采样点之间没有发生事件的时间戳进行表示，实现对于连续时间轴上完整事件流的建模，从而对未来事件聚合的间隔长度进行选择。作为经典数学工具，时序点过程具有较为完善的理论基础[49]。

基于传统机器学习的事件序列模型是该领域早期的方法，在过去的几十年间被广泛研究。基于传统机器学习的方法主要分为两个方向，分别为基于马尔可夫模型的方法和基于时序点过程的方法。基于马尔可夫的方法[50-51] 通常引入变阶马尔可夫模型（Varying-order Markov Models），将此问题抽象为离散化的时间序列预测任务来处理，基于观察到的数据进行分析，捕获序列中高阶和低阶马尔可夫依赖性。基于所观察的历史事件序列，变阶马尔可夫模型学习事件序列中的概率有限状态自动机，事件类型的预测由状态转换过程中下一步最可能的演变成状态给出。但是基于马尔科夫模型的方法具有明显缺陷，由于状态空间的大小随着马尔可夫模型中假设的时间步数呈指数增长，因此马尔可夫模型无法处理历史事件的长期依赖性。与此相比，基于时序点过程的传统机器学习方法[34-35] 是早期处理事件序列的更通用框架，利用构造的条件强度函数（Conditional Intensity Function），以历史事件为条件构建针对未来事件的随机过程模型，可以建模事件序列中的长期依赖。但是基于传统机器学习的点过程模型由于需要定义显式参数化的强度函数，使得模型表达能力受限，可能出现欠拟合的问题，难以捕捉事件之间复杂的影响关系。

为了改善传统机器方法的局限性，基于深度学习的事件序列模型逐渐引起研究人员的关注，推动该领域研究从显式参数化模型过渡到隐参数化模型。相比传统机器学习方法中显式参数化的条件强度函数，基于深度学习的时序点过程

模型利用神经网络对条件函数进行建模，可以拟合事件间更加复杂的相关关系，改善了模型需要手动预设强度函数的局限性。基于深度学习的事件序列预测模型，其整体流程主要包含三个步骤，分别为对于单个事件的编码、对于历史事件序列的编码、以及对于未来事件的预测。对于历史事件的编码，Du 等人在论文[52] 中首先利用循环神经网络（Recurrent Neural Network）编码历史事件序列，提出循环标记时间点过程（Recurrent Marked Temporal Point Processes）模型，之后循环神经网络逐渐成为用于历史事件序列编码表示的主要方法[53-55]。随着注意力机制（Attention Mechanism）在深度领域中大放异彩，基于注意力机制的历史序列编码方法也应运而生，并展现了良好性能[56-57]，在一定程度上提高深度学习模型的可解释性。目前为止，对于未来事件的预测，构造神经网络化的条件强度函数仍是主要方式[52,56,58]，但是免强度函数的方法被逐渐关注，例如 Takahiro 等人在论文[59] 中跨过强度函数，直接学习条件强度函数的积分，从而构造一个更加灵活的完全神经化的点过程模型（Fully Neural Network Point Processes）。

对于事件序列分析，虽然上述基于判别式的深度学习模型可以得到相对较好的预测结果，但是相对而言缺乏对于事件序列数据的完整理解。而生成模型直接对联合概率密度分布进行建模，可以使得模型学习到更多事件序列自身的信息。如果事件序列模型可以生成接近真实的事件序列，则我们可以相信该模型相对准确地捕获了序列内部的潜在变化过程。对抗生成网络（Generative Adversarial Networks，GAN）已被证明是一种很有前途的生成模型，其具有广泛的理论基础和实验验证[60-62]。Xiao 等人在文章[63] 中为时序点过程引入改进后的 Wasserstein 对抗生成网络（Wasserstein GAN），利用点过程之间的 Wasserstein 距离来训练生成模型，建模离散事件的分布。随后，基于对抗训练的时序点过程模型逐渐发展起来[64-65]。

基于强化学习（Reinforcement Learning）的事件序列预测模型，将事件序列分析问题视为一个序列决策问题，使用强化学习框架研究事件的动态建模，该类方法将已经给定历史序列的下一个事件时间的条件分布视为模型要学习的策略。在基于循环神经网络的事件序列预测模型中，由于训练和推理之间的差异性会导致误差生成，这些误差会沿着预测的序列快速累积，而基于强化学习的方法可以缓解上述问题，会自然会考虑策略的长期累积奖励或损失，从而避免在上述学习和推理过程之间的不匹配问题[66-68]。例如 Li 等人在文章[66] 中通过

逆强化学习（Inverse Reinforcement Learning）将问题转化为专家点过程和学习者点过程之间差异的最小化问题，从而学习未知的奖励函数。

随着人们对于深度学习技术探索逐渐深入，事件序列相关模型的精度进一步提升，但是现有的事件序列预测相关研究重点在于建模未来事件的条件分布，对于历史事件序列的特征编码常使用现有时间序列建模相关方法[52,57]，如何从事件序列数据中学习更加有效的特征表示仍是该领域的主要难题。针对历史事件序列进行特征表示和学习，最关键的挑战是合理建模不同事件之间的影响关系，也是事件序列预测问题的主要目标之一。现有方法常使用循环神经网络和注意力机制对事件序列进行特征编码，对问题进行简化，认为不同事件是时间流上的顺序结构，而忽略了事件之间更复杂的因果结构。合理建模不同事件之间的影响关系不仅可以帮助提升预测准确率，同时对于我们理解用户行为，对未来事件进行更有效的干预有重要意义。其次由于事件序列数据往往采样自实际应用场景，因此用户行为序列会具有一定的偏向性，使得数据具有一定的不平衡特征，这种不平衡性使得事件序列呈现长尾分布（Long-Tailed Distributions），即少数类型的事件拥有充足的样本，而大部分事件仅有少量样本。这种不平衡性导致模型难以对样本数据较少的类型进行充分的特征学习，也会影响模型对于不同事件的特征编码能力，如何处理具有长尾分布的真实事件序列也是特征表示学习的重要挑战。

1.3 研究内容与贡献

本文在点过程建模的基础上，研究基于深度学习的预测相关算法，以针对历史事件序列的特征表示学习为切入点，主要关注其中的两个挑战，即如何学习历史序列中事件之间的相关结构信息，以及如何缓解历史序列数据中的长尾分布对于特征表示学习的影响。针对上述挑战，本文提出两种针对性的模型来进行处理，我们分别在仿真数据和真实数据上进行实验，验证算法的有效性。最后，我们将本文所提出的模型应用于实际的图书管理系统中。本文的主要研究成果如下：

1. 针对事件间复杂关系建模的问题，本文分析现有事件序列模型进行关系建模的局限性，创新性地提出一种基于渐进生成图的事件序列预测模型。该模

型利用动态图神经网络对历史序列中的不同事件间关系进行建模，同时利用前置图和多维关系图学习事件间不同复杂度的关系图，使得模型由简单到复杂地学习事件间多维关系图。考虑到事件间关系并非静态，我们对关系的动态变化进行建模。为了能够更灵活的捕捉事件间关系，我们将事件间关系结构视为隐变量进行自动推理；并且将事件间关系用多维图进行描述，使得模型具有表示事件间不同类型影响关系的能力。

2. 针对历史序列数据中的长尾分布问题，本文深入分析事件序列预测中长尾分布问题的具体表现，首次将该问题描述为标记空间上的长尾分布和时序空间上的长尾分布，并且引入解耦学习框架来进行处理。同时针对解耦学习框架的局限性，我们提出一种基于双平衡软标签的辅助训练模型，使用软标签来提高解耦学习的整体性能。具体来说，我们设计了一个专用的辅助网络来分别为事件类型预测器和时间间隔预测器的两个不同的训练阶段生成辅助软标签。针对时间间隔预测器，辅助网络使用生成模型预测时间间隔的分布，作为连续空间上的软标签。为了使得主网络和辅助网络间进行更高效的特征传递，我们引入了特征级蒸馏方法，并通过多尺度特征融合改进了一般特征的学习。

3. 我们本文将所提出的两个事件序列预测模型应用于实际系统中。在我们所搭建的图书管理系统中，我们建模用户浏览图书的行为序列进行分析，并将用户行为特征编码到推荐系统中，实现图书推荐的功能，充分说明了所提出的模型具有较强的实际应用价值。

1.4 本文的组织结构

本文围绕基于深度学习和点过程的事件序列预测模型进行分析和讨论，主要关注其中针对事件的特征表示学习，并聚焦两个核心问题。针对事件间关系建模的挑战，本文提出一种基于渐近生成图的模型，利用前置图和多维关系图对事件间影响关系进行动态建模。针对历史序列中的长尾分布挑战，本文首先定义事件序列的长尾分布，通过设计一种基于双平衡软标签的辅助训练模型，缓解长尾问题对于模型进行特征学习的影响。最后将算法应用于实际的图书管理系统中，证明其实用性。本文一共分为六章，第一章为绪论，主要从实际应用角度出发介绍事件序列预测问题的研究背景和意义，将研究现状总结为四个方向

分别进行介绍，并指出该领域的研究挑战。第二章为相关工作，主要介绍事件序列预测问题和点过程的概念及问题定义，介绍该领域相关研究，并重点介绍现有基于深度学习的事件序列预测方法。第三章对基于渐近生成图的事件序列预测模型进行介绍，此章介绍该模型的研究动机和关键技术，然后从模型结构设计和训练方法的角度介绍该模型，并介绍和分析相关实验结果。第四章对基于双平衡软标签的辅助训练模型进行介绍，分别在模拟数据和真实数据上进行实验和分析，证明模型的有效性。第五章为上述两种方法在图书管理系统中的应用。第六章对全文所有工作进行总结，并对事件序列预测领域的未来研究方向进行展望。

第二章　相关工作

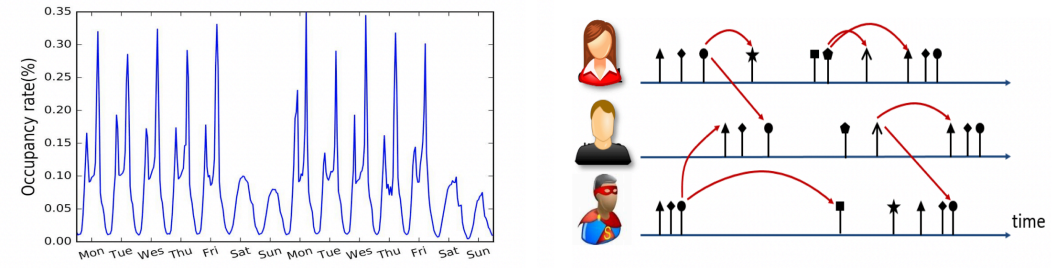
事件序列预测是时间序列分析领域重要的研究方向之一，早在二十世纪七十年代研究人员就已经开始针对该领域的进行探索研究，并获得了诸多成果。早期针对该问题的算法以随机过程相关理论为基础，使用基于传统机器学习的相关模型。随着人工神经网络和深度学习领域的快速发展，该领域目前以基于深度学习的算法为主流。本章将对事件序列预测问题进行形式化的定义，并介绍相关领域的经典算法及近期具有代表性的算法研究工作。

2.1　事件序列预测问题及点过程定义

时间序列数据是机器学习领域中一种常见的数据类型，定义为将同一统计指标的观察数值按其所发生时间的先后顺序排列而成的序列数据。事件序列数据作为时间序列数据的一种，也是按照离散事件发生的先后顺序被记录的，具有自然的时间顺序性。但是与规则采样的一般时间序列不同，事件序列数据是不规则的，同一序列上相邻事件的发生可能任意时间，换言之，同一序列上的相邻事件发生时间的间隔是不规则的，具有一定的随机性；但是事件之间仍然具有一定的相关性，这种隐式的相关性具有统计分析的价值。事件序列数据与一般时间序列数据对比如图2-1所示。

给定时间区间 $[0,T]$，事件序列数据是在这段时间间隔内发生的具有先后顺序的离散事件的序列。所记录的事件按照标记信息的有无一般分为两种，即有标记信息的事件和无标记信息的事件，无标记信息的事件序列通常是为了便于理论分析而针对该问题的一种简化，由于现实世界中事件通常具有额外的信息，因此对于有标记信息的事件的分析更具有价值，本文主要关注有标记信息的事件序列。一个具有标记信息的事件 e_i 被可以定义为 $[t_i, m_i]$，其中 $t_i \in [0,T]$，表示序列中第 i 个事件发生的时间；$m_i \in \{\emptyset, 1, \ldots, K\}$ 为所记录的第 i 个事件的标记信息，即事件 e_i 所对应的的事件类型，事件类型所对应的标记空间一般为离散

的空间。时间间隔 $[0,T]$ 上的事件序列可以被定义为 $X=\{(t_1,m_1),\ldots,(t_i,m_i)\}$，表示记录在连续时间 $[0,T]$ 上发生的离散标记空间 M 内的事件。其中 $M=\{\emptyset,1,\ldots,K\}$，若 $m_i=\emptyset$ 我们认为该时刻 t_i 没有观测到对应事件发生；仅当 $m_i\neq\emptyset$ 时，该时刻有观测事件发生，为 K 种类型中的一种。同一事件序列上的不同事件需要按照发生的先后顺序进行排序，即事件发生时间的集合 $\{t_1,\ldots,t_i\}$ 需要满足 $0<t_1<\cdots<t_i\leq T$。



(a) 一般时间序列数据 (b) 事件序列数据

图 2-1: 事件序列数据与一般时间序列数据对比

本文主要关注针对事件序列的预测问题。事件序列预测问题是指我们利用输入的事件序列 X 来预测未来将要发生事件 e_{i+1}。为了方便表示，我们定义 τ 为事件发生时间的间隔，第 i 个事件与上一事件的时间间隔为 τ_i，即 $\tau_i=t_i-t_{i-1}$，τ 也被称为连续事件间的持续时间。同时为了区别特定时刻对应的历史序列和输入的整个序列，我们定义 His_t 为时刻 t 之前的历史序列，即 $His_t=\{(t_j,m_j),t_j<t\}$，利用此定义我们同样可以表示事件 e_i 对应的历史序列为 His_{t_i}。

时序点过程的基本思路是将事件预测问题视为连续时间上针对事件的计数过程，一般定义 $N(t)$ 表示到时间 t 为止事件的累积发生数量。为了处理事件序列预测问题，我们需要判断时刻 t 上有对应事件发生的概率，如果未来时刻 t 有事件发生的概率较大，则我们可以输出该时刻作为预测结果。为了实现上述功能，时序点过程定义条件强度函数（Conditional Intensity Function）为 $\lambda^*(t)$，用于建模下一事件发生时间的随机模型。具体而言，条件强度函数表示在给定历史序列 H_t 的情况下，在时间窗口 $[t,t+dt)$ 内有新事件发生的概率，在 $\lambda^*(t)$ 的基础上，我们可以得到对于该问题的如下形式化表示：

$$\lambda^*(t)dt=\mathbb{P}\left(\text{event in }[t+dt]\mid H_t\right) \tag{2-1}$$

即可以定义为：

$$\lambda^*(t)dt = \mathbb{E}\left(dN(t) \mid H_t\right), \tag{2-2}$$

其中 $\mathbb{E}$ 用以表示期望，dt 用以表示时刻 t 之后一段极小的时间间隔，因为此段间隔极小，我们进行合理假设 $dN(t) \in \{0,1\}$。从而在 $N(t)$ 定义的基础上，我们利用 $\mathbb{E}\left(dN(t) \mid H_t\right)$ 表示在给定历史序列情况下，窗口 $[t, t+dt)$ 内事件累积发生数量的期望，$*$ 用以表示条件函数依赖于历史序列。

但是强度函数仅能表示某一时刻窗口是否有事件发生，无法直接对未来事件的发生时间进行表示。为了直接计算未来事件的发生时刻，我们定义条件密度函数（Conditional Density Function）为 $f^*(t)$，表示时刻 t 将有下一未来事件发生。在条件密度函数的基础上，利用链式法则，我们可以得到历史时间序列的联合密度函数。此联合密度函数 $f^*(t_1, t_2, \ldots, t_n)$ 表示已有观测序列的联合似然。之后建立条件密度函数和条件强度函数之间的关联，进而计算此联合密度函数。在此基础上，我们利用条件强度函数可以对对数似然函数进行表示：

$$\log f\left(t_1, t_2, \ldots, t_n\right) = \sum_{j=1}^{n} \log \lambda^*\left(t_j\right) - \int_{t_0}^{t_n} \lambda^*(\tau) d\tau \tag{2-3}$$

此为时序点过程模型的优化函数。在模型学习的过程中，传统的时序点过程模型通常给定预先定义的参数化条件强度函数 $\lambda_\theta^*(t)$ 和一系列观测到的事件序列作为训练数据，我们通过最大化似然函数来优化模型参数 θ。我们可以发现条件强度函数可以直接影响模型的表达能力，因此为条件强度函数 $\lambda_\theta^*(t)$ 选择一个合适的参数形式对于模型至关重要。

2.2 基于传统机器学习的点过程算法

基于传统机器学习的相关算法是处理事件序列预测问题的早期方法，主要包括基于马尔可夫性质的相关算法和基于传统点过程的相关算法，本小节主要介绍基于传统机器学习的点过程算法。如上述介绍，基于点过程算法的核心是设计合适的条件强度函数，本文将介绍传统机器算法中两种基础点过程算法，即泊松过程和自激励过程，以及两个算法的相关拓展工作。

2.2.1 泊松过程

泊松过程（Poisson Process）模型是点传统过程模型中一类相对基础的模型，其基本假设是事件之间相互独立，即条件强度值与历史的事件序列无关。齐次泊松过程为泊松过程中的最简单的模型，为了简化问题，其认为在整个序列中，条件强度的值为不随时间变化的固定值，即有：

$$\lambda(t) = \gamma_0 \geq 0 \tag{2-4}$$

在模型中 γ_0 为需要学习的参数。但是齐次泊松过程的表达能力非常有限，当事件序列本身的动态特征变化明显时，模型难以捕捉这种动态变化。为了提升泊松过程的表达能力，其改进工作非齐次泊松过程被提出。在同样的事件相互独立假设下，非齐次泊松过程建模条件强度函数为时间相关的函数，即条件强度值会随时间而变换：

$$\lambda(t) = p(t) \geq 0 \tag{2-5}$$

其中 $p(x)$ 可以帮助模型捕捉序列中的动态特征。在非齐次泊松过程的基础上，Cox 等人[69] 探索泊松过程，提出著名的重随机泊松过程（Doubly Stochastic Poisson Process），改善条件强度函数使得其不仅捕捉序列本身的动态特征，并且可以捕捉环境因素等协变量的特征。但是此方法依然没有考虑序列内事件之间的相关关系，当事件序列较为复杂时无法取得较好的预测效果。

2.2.2 自激励过程

现实世界中相邻事件往往存在相关性，针对不同事件的独立性假设是一个较强的假设，具有此假设的泊松过程使用场景有限。为了改善泊松过程事件独立性假设过强的问题，建模事件之间的相关性，Hawkes 等人在论文[35] 中提出一种自激励过程（Self-exciting process）模型，后续也被称为霍克斯过程[70]（Hawkes Process）。该模型是时序点过程领域中建模事件间关系的极具影响力的工作。基础的自激励过程假设事件具有相关性，并且假设这种相关性为正向的，同一序列上过去发生的事件将对条件强度函数产生积极作用，即自激励过程认为若历史发生相关事件，则未来发生该事件的概率将增加。这种历史事件对强度函数

的积极的影响是可以叠加的，并且这种影响随时间的进行而衰减。自激励过程的条件强度函数可以表示为：

$$\lambda^*(t) = \gamma_0 + \alpha \sum_{t_j<t} \gamma\left(t - t_j\right) \tag{2-6}$$

自激励过程的强度函数分为两个部分，第一部分为基础强度 $\gamma_0 \geq 0$，其为独立于历史事件并且不随时间变换的强度值。第二部分为历史事件对于当前时刻的影响关系，其中 $\gamma(t - t_j) > 0$ 是捕捉事件间依赖性的核函数，并且可以描述历史事件对当前时刻影响随时间的衰减过程，常用的核函数是指数衰减函数，即 $\gamma(t - t_j) = \exp\left(-\beta\left(t - t_j\right)\right)$，需要注意的是自激励过程使用加性模型统计所有历史事件的影响。从条件强度函数的形式我们认为，当历史近期有事件发生时，当前时刻事件发生的强度值将快速提高，并且随时间流逝，当前时刻发生事件的强度值逐渐回归基础强度。

上述为基础的自激励过程模型，在其基础上，有许多拓展工作。例如 Bowsher 等人在论文[71] 中考虑的模型外生非平稳性，将历史事件的无关强度 γ_0 拓展为时间相关的函数 $\gamma_0(t)$，即：

$$\lambda^*(t) = \gamma_0(t) + \alpha \sum_{t_j<t} \gamma\left(t - t_j\right) \tag{2-7}$$

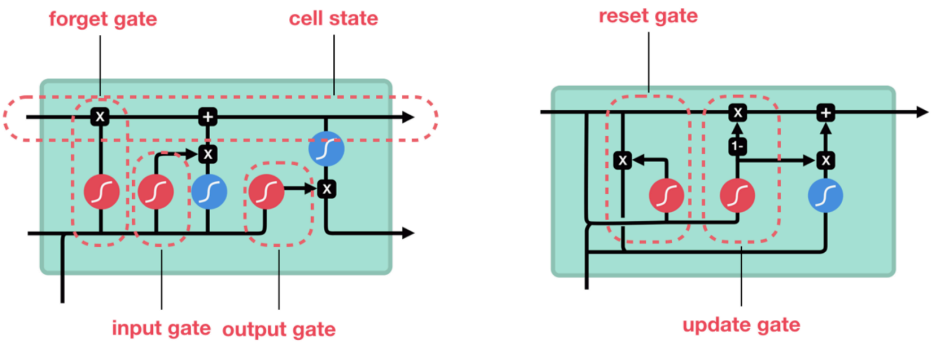
从而提升模型的表达能力，使得模型可以对非平稳系统进行表示。该模型被用于解决金融场景下的由季节性引起的动态变化[72]。相比模型的外生非平稳性，自激励过程的内生非平稳性同样被研究，即建模由于事件间内生相互作用引起的非平稳性[73-75]。基础的自激励过程虽然考虑累积所有的历史时间影响，但是仅使用简单的加性模型进行建模，这种针对事件强度的加性模型在实际应用场景下缺乏合理性，因此研究人员考虑自激励过程中历史事件影响的非线形叠加[76-77]。

自激励过程相比泊松过程，建模了序列中不同事件的相关性，提升了模型的表达能力。但是自激励过程中历史事件对于当前事件的积极性假设依然限制了其自身的表达能力，从我们日常生活中可以感受到，在一些场景下历史事件可以对我们当前的决定产生消极的影响，例如我们需要使用笔记本电脑，如果我们过去时刻已经在购物平台上购买了该设备，那么在较长的一段时间内我们

无需购买相同类型的设备。为了描述这种非积极的影响关系，Ertekin 等人在文章[78] 中提出一种响应点过程模型（Reactive Point Process），通过在基础模型上补充一个描述自抑制关系的核函数 $\kappa(t-t_j)$，来表示部分历史事件对于未来事件的抑制作用，响应点过程模型可以视为自激励过程的拓展。

2.3 基于深度学习的序列建模算法

序列建模是时序分析领域进行序列数据特征学习、处理预测任务的重要工具，如今基于深度学习的序列建模算法已经成为其中的主流方法，也是处理事件序列预测任务的基础，在本节中我们将介绍常用的基于深度学习的序列建模方法。



(a) LSTM (b) GRU

图 2-2: 循环神经网络结构示意图

深度学习的迅速崛起得益于人工神经网络领域的不断发展。逐步完善的参数优化方法使得深度学习被不同研究领域应用，其中就就包括时间序列分析领域。早期的多层感知机模型虽然可以实现序列建模，但是其由于神经元本身没有固有顺序，难以检测时序数据中的随时间变化的动态模式，并且无法处理变长的序列，为了解决上述问题，循环神经网络被提出[79]。循环神经网络的核心思想是在结构单元中定义一个内部的存储空间，该存储空间可以被循环神经网络用于记录历史状态，存储空间内储存历史序列的紧凑特征表示，该记忆状态处理每个时间步所输入当前的观测值，进行参数更新。由于循环神经的每个时间步使用相同的结构单元，因此可以处理变长的序列数据，并且网络的权重可以跨时间共享，使其有能力捕捉序列内部的动态变化。在第 i 个时间步，循环神

经网络的核心可以表示为:

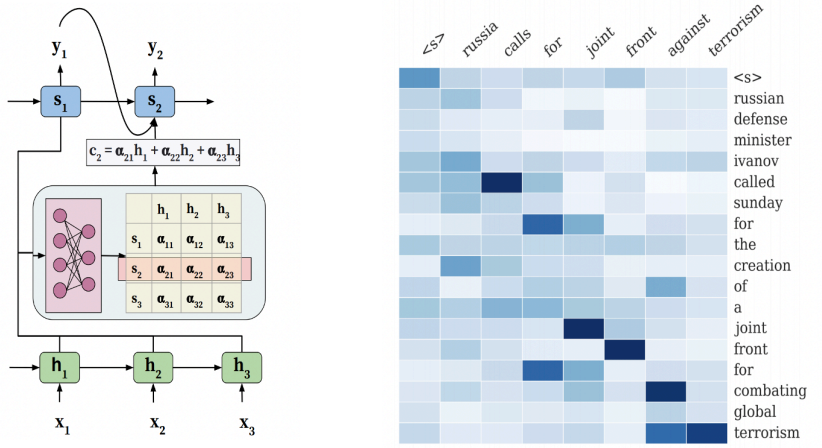
$$h_t = \tanh\left(W_{hh}h_{t-1} + W_{xh}x_t\right) \tag{2-8}$$

其中 x_t 为第 i 个时间步的序列输入，h_i 为网络的第 i 个时间步隐层输出，即为上述存储空间所存储的包含历史信息的特征，$tanh$ 为循环神经网络常用的激活单元，W_{hh} 和 W_{xh} 分别为网络针对历史状态和序列输入进行特征学习的参数。但是基础的循环神经网络模型在处理较长序列的问题时，由于梯度消失等问题的出现而难以正常训练，并且会随着序列的增长而逐渐失去早期序列信息，因此研究人员通过引入门结构来优化模型，典型的工作为长短期记忆网络[80]（Long Short-Term Memory networks，LSTMs），在长短期记忆网络中，三种类型门结构被用于对记忆方式进行控制，实现去除或者增加信息到存储空间的操作，使得模型可以处理更长的输入序列。门控循环单元[81]（Gated Recurrent Unit，GRU）是在长短期记忆网络的基础上改良而来的模型，它将遗忘门和输入门合成了一个更新门，并且混合了网络的单元状态和隐藏状态。除了门控结构，针对循环神经网络的其他方向的改良也被不断发掘[82-84]。

另一类常用的基于深度学习的序列建模是时序卷积网络（Temporal convolutional network, TCN）。循环神经网络已经被证明是进行时序建模的有效工具，但是其依然存在因自身机制导致的缺点。因为不同时间步共享网络单元，因此后面的时间步必须等待之前时间步的操作完成，这种固有的顺序性使得循环神经难以进行并行化计算，网络训练和推理的效率较低；并且由于网络对临近序列更敏感，虽然长短期记忆网络等结构对其进行改良，但是灾难性遗忘问题依然存在。为了缓解循环神经网络的上述问题，时序卷积网络被提出[85]。时序卷积网络包含三种核心结构，包括因果卷积 (Causal Convolution)，膨胀卷积 (Dilated Convolution) 和残差链接 (Residual Connections)，其中因果卷积可以消除一般卷积结构应用于序列数据时的导致的信息泄露问题。上述结构使得时序卷积网络相比循环循环神经网络具有更优秀的计算并行性和更稳定的梯度更新。并且由于膨胀卷积的引入，时序卷积网络可以更加灵活的调整自身的感受野。Bai 等人在论文[86] 中提出改良后的网络网络（Trellis Networks），设计不同时序卷积层之间的权重共享机制，并且证明网格网络与截断循环网络之间的等价性，从理论

角度分析时序卷积和循环神经网络两种序列建模之间的相关性。

近年来，基于注意力机制的大规模预训练模型[19,21]在自然语言处理领域大放异彩，考虑注意力机制对于序列化数据的天然优势，研究人员开始探索将注意力机制应用于序列建模的可能性。注意力机制的提出借鉴人类大脑学习庞大输入数据的方式，主要选择关键信息进行记忆学习，而在一定程度上忽视非关键信息，通过信息筛选来提升学习的效果。注意力机制借鉴此学习方式，允许模型去着重关注过去时间内的重要信息，基本结构如图2-3所示。注意力机制的实现本质上是一种针对给定查询选择对应相关信息的过程，由于弱化了感受野的概念，使得注意力机制可以在很长一段历史序列中进行检索，因此理论上可以根本解决时间序列问题中的长期依赖遗忘的问题。以 Informer[87] 为代表的基于注意力机制的时序建模方法被证明具有优秀的性能，并且实验证明在处理长序列输入数据时候，其相比循环神经网络和时序卷积有更明显的优势。



(a) 注意力机制 (b) 注意力权重可视化

图 2-3: 注意力机制示意图

2.4 基于深度学习的点过程算法

随着设备计算能力的提升和各领域高质量数据集的构造，近年来深度学习技术被广泛应用。在上一小节中我们介绍了基于深度的序列建模相关算法，以此为基础，在本小节中我们将介绍如何利用基于深度学习的点过程方法处理事件序列预测任务。我们首先为读者介绍基于深度学习的点过程模型在处理事件序列数据时的整体流程，之后我们分别介绍基于循环神经网络和基于 Attention

的点过程算法。

2.4.1 深度学习点过程基础算法流程

基于深度学习的相关算法是目前用于处理事件序列预测问题的主流方法，在介绍具体的方法之前，我们先总结基于深度学习的点过程方法解决该问题的整体流程。我们可以将其总结为三个主要步骤，分别为对单个历史事件的特征编码、对整个历史事件序列的特征编码、以及利用历史序列的特征预测未来事件的发生。下面我们将分别对三个步骤所涉及的相关方法进行介绍。

基于深度学习的点过程算法第一步是对单个历史事件进行特征编码。该部分的输入是经过数据预处理后的事件序列，输出是历史序列上，每个事件对应的特征编码。每个单独的事件 e_i 由两部分组成，分别为事件的发生时间 t_i 和该事件所对应的标记信息 m_i。由于时间 t_i 处于连续的时序空间而标记 m_i 一般处于离散的类型空间，所以一般对两部分使用不同的方法进行单独编码，并进行特征融合得到事件 e_i 对应的特征表示 $\mathbf{c}_i$。对于事件时间的编码，我们一般对历史事件序列进行差分操作，利用事件发生的相对时间来代替绝对时间，即使用间隔时间 τ_i 来表示，然后利用嵌入层对其进行编码得到 $\mathbf{c}_i^{time}$。对于事件标记信息的编码，标记信息满足 $m_i \in \{\emptyset, 1, \ldots, K\}$，由于我们仅对发生的事件进行编码，因此我们不考虑标记信息为空的情况。我们首先利用独热编码（One-Hot Encoding），然后同样使用嵌入层进行编码得到 $\mathbf{c}_i^{mark}$。对于 $\mathbf{c}_i^{time}$ 和 $\mathbf{c}_i^{mark}$ 进行特征融合，基础的方案是将两部分特征投影到相同维度的特征空间相加，从而得到每个单独事件 e_i 的特征编码 $\mathbf{c}_i$。

第二步是对整个历史事件序列进行特征编码。该部分的输入为历史事件序列所对应的单个特征编码的序列，输出为整个历史事件序列的特征 $\mathbf{h}_i$。当前时刻 t 所对应的历史事件特征序列为 $(\mathbf{c}_1, \ldots, \mathbf{c}_{i-1})$。为了对相关历史序列进行充分编码，此部分可以被抽象为序列特征表示任务，相关方法通过引入时序建模相关算法实现特征编码。此部分常使用的序列建模算法为基于循环神经网络的方法[52-54] 和基于注意力机制的模型[56-57]。一般的循环神经网络和注意力机制模型处理的是时间轴上等间隔离散点的特征，由于点过程对事件序列预测问题建模方式的特殊性，要求模型具有对连续时间点进行特征表示的能力，因此一般的循环神经网络和注意力机制模型无法被直接使用。相关方法通常对现有序列建

模方法进行改良，使其满足点过程建模对于历史序列编码模型的要求。

第三步是利用历史序列的特征预测未来事件的发生情况。若历史序列包含 $i-1$ 个事件，则此部分的输入为前 $i-1$ 个事件所对应的历史序列编码 $\mathbf{h}_i$，输出为对未来事件发生的预测，包括所预测将要发生事件的具体发生时刻 $\hat{t}_i$ 和对应的标记信息 $\hat{m}_i$，即对历史嵌入信息建模条件分布 $P_i^*(t_i, m_i \mid \mathbf{h}_i)$。由于我们一般预测的是未来事件距离上一历史事件的时间间隔，因此该条件分布可以表示为 $P_i^*(\tau_i, m_i \mid \mathbf{h}_i)$。基础的算法一般针对预测结果 $\hat{t}_i$ 和 $\hat{m}_i$ 使用条件独立的建模方式，即认为:

$$P_i^*(\tau_i, m_i \mid \mathbf{h}_i) = P_i^{time}(\tau_i \mid \mathbf{h}_i) \times P_i^{mark}(m_i \mid \mathbf{h}_i) \tag{2-9}$$

其中 P_i^{time} 和 P_i^{mark} 分别为未来事件发生时刻和标记信息所对应的条件分布。实验表明，对时间和标记信息进行条件独立假设的模型可以取得不错的预测效果，但是需要承认的是这种建模方法限制了预测模型的表达能力。因此在独立假设的基础上，研究人员提出两种建模未来事件的时间和标记信息的方法，分别为标记信息以时间为条件的模型，以及时间以标记信息为条件的模型。通过建立未来事件时间和标记信息的相关性，模型性能得以进一步提升。

2.4.2 基于循环神经网络的点过程算法

循环神经网络是时序分析领域最常使用的深度学习模型，引入门结构的长短期记忆网络和门控循环单元模型缓解了基础循环神经网络的存在的难以建立长期依赖的问题，因此被广泛使用。将循环神经网络引入事件预测任务是非常自然的思路，当前在基于深度学习的事件序列预测模型中，基于循环神经网络的模型仍是主流的方法，本小节我们将介绍其中最具有代表性的工作。

Du 等人在论文[52] 中首次使用循环神经网络建模时序点过程问题。由于传统时序点过程算法在建模的过程需要对强度函数进行特定假设，存在假设与现实问题不匹配的问题，并且固定形式的强度函数限制了模型的表达能力。针对此问题 Du 等人提出循环标记时序点过程模型（RMTPP）。RMTPP 模型通过引入循环神经网络结构学习历史事件序列的表示，将条件强度函数视为历史事件的非线形方程，使得时序点过程模型更具有通用性。在隐藏层中，RMTPP 使用

循环神经网络如下：

$$\mathbf{h}_j = \max\left\{\mathbf{W}^m\mathbf{m}_j + \mathbf{W}^t\mathbf{t}_j + \mathbf{W}^h\mathbf{h}_{j-1} + \mathbf{b}_h, 0\right\} \tag{2-10}$$

在得到历史序列的特征表示 $\mathbf{h}_j$ 之后，RMTPP 定义条件强度函数如下：

$$\lambda^*(t) = \exp(\underbrace{\mathbf{v}^{t^\top}\cdot\mathbf{h}_j}_{\text{past}} + \underbrace{w^t\left(t-t_j\right)}_{\text{current}} + \underbrace{b^t}_{\text{base}}), \tag{2-11}$$

条件强度分为三个部分，分别为过去历史序列对未来事件的影响、当前事件对于未来事件的影响、以及基础强度。条件强度函数建模未来事件的发生时间，对于其发生的类型的预测，RMTPP 使用多项分布模型进行建模：

$$P\left(y_{j+1}=k \mid \mathbf{h}_j\right) = \frac{\exp\left(\mathbf{V}^y_{k,:}\mathbf{h}_j + b^y_k\right)}{\sum_{k=1}^{K}\exp\left(\mathbf{V}^y_{k,:}\mathbf{h}_j + b^y_k\right)}, \tag{2-12}$$

其中 $\mathbf{V}^y_{k,:}$ 为模型对于历史序列特征进行特征学习的参数，RMTPP 的目标函数预测未来事件时间的对数似然函数和预测未来事件标记信息的交叉熵。RMTPP 继承了循环神经网络和时序点过程模型两者优势，可以在没有任何针对潜在动态时序的先验知识的情况下预测未来事件的发生时间和标记信息。

虽然 RMTPP 引入循环神经网络进行特征学习，但其简化了模型，将历史事件和当前事件对未来事件的影响进行分离，并没有让循环神经网络对连续时间上的采样点进行特征表示。针对此问题，Mei 等人在论文[53] 中提出一种神经霍克斯模型（NHP），通过设计一种连续时间长短期记忆网络（CT-LSTM），实现对于时间轴上的连续采样点的特征表示，并将其体现在条件强度函数上。具体而言，相比长短期记忆网络，CT-LSTM 在事件发生之后的连续时间内，使得内部的存储单元可以向稳态值进行指数衰减。CT-LSTM 中内部更新过程表示如下：

$$\mathbf{c}_{i+1} = \mathbf{f}_{i+1}\odot\mathbf{c}\left(t_i\right) + \mathbf{i}_{i+1}\odot\mathbf{z}_{i+1} \tag{2-13}$$

$$\bar{\mathbf{c}}_{i+1} = \bar{\mathbf{f}}_{i+1}\odot\bar{\mathbf{c}}_i + \bar{\mathbf{I}}_{i+1}\odot\mathbf{z}_{i+1} \tag{2-14}$$

$$\delta_{i+1} = f\left(\mathbf{W}_{\mathrm{d}}\mathbf{k}_i + \mathbf{U}_{\mathrm{d}}\mathbf{h}\left(t_i\right) + \mathbf{d}_{\mathrm{d}}\right) \tag{2-15}$$

其中 f_i, i_t, o_t 分别表示三种门结构，即遗忘门，输入门和输出门，δ_i 描述了内部的衰减过程。在此基础上，连续内核单元 $c(t)$ 更新如下：

$$\mathbf{c}(t) \stackrel{\text{def}}{=} \bar{\mathbf{c}}_{i+1} + (\mathbf{c}_{i+1} - \bar{\mathbf{c}}_{i+1}) \exp\left(-\delta_{i+1}\left(t - t_i\right)\right) \text{ for } t \in (t_i, t_{i+1}] \tag{2-16}$$

基于 CT-LSTM 的连续内核单元 $c(t)$，NHP 模型中连续时间上的特征表示 $\mathbf{h}(t)$ 可以被表示为：

$$\mathbf{h}(t) = \mathbf{o}_i \odot (2\sigma(2\mathbf{c}(t)) - 1) \text{ for } t \in (t_{i-1}, t_i] \tag{2-17}$$

针对不同类型的事件，NHP 分别建立不同的条件强度函数。相比传统的霍克斯过程，NHP 可以自动学习历史事件对于不同类型的影响，包括促进和抑制效应，即历史事件可能降低未来事件的强度；并且可以帮助模型捕获事件的固有惯性，使得事件强度可以在连续事件之间非单调地波动。

RMTPP 和 NHP 是基于循环神经网络处理事件序列预测问题的两种非常具有代表性的模型。除此之外，研究人员也分别从提高模型训练效率、进行更充分的信息提取等角度改善基于循环神经网络的点过程模型。例如针对现有模型训练效率低的问题，Mei 等人在论文[58] 中利用噪声对比估计方法优化积分计算，降低模型的计算开销；为了探索利用更充分的信息，Xiao 等人在论文[88] 中提出循环点过程模型（RPPNs）利用两个循环神经网络编码离散事件序列和额外的连续输入特征，同时建模内生强度和外生强度。

2.4.3 基于注意力机制的点过程算法

基于注意力机制的模型已经被证明在时间序列预测任务上具有良好的性能性能[87]。基于注意力机制的点过程算法逐渐在事件序列预测领域引起研究研究人员的关注，本小节中我们将介绍基于注意力机制的事件序列预测模型。

Wang 等人在论文中[89] 中为时序点过程模型引入注意力机制，提出基于集联注意力机制的循环神经网络（CYAN-RNN）用于处理事件序列预测问题。在 CYAN-RNN 模型中，注意力层是建立在循环神经网络编码之上的。对于输入的事件序列数据，CYAN-RNN 首先利用循环神经网络进行作为编码器，然后在编码器之上引入注意力层，以帮助循环神经网络进行记忆捕捉。CYAN-RNN 中注

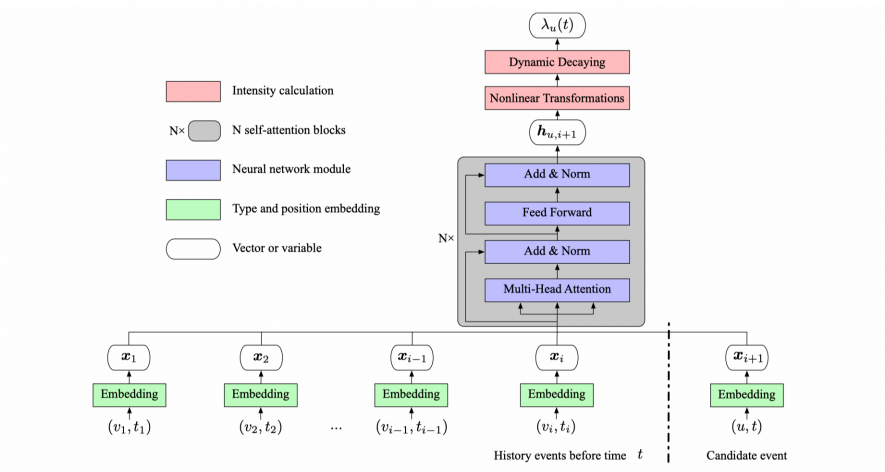


图 2-4: SAHP 模型结构示意图

意力权重计算过程为：

$$\alpha_{k,i} = \frac{\exp\left(e_{k,i}\right)}{\sum_{j=1}^{k} \exp\left(e_{k,j}\right)} \tag{2-18}$$

其中 $\alpha_{k,i}$ 表示事件 e_k 和 e_i 之间的相关系数，事件之间相关性的计算方式为：

$$e_{k,i} = a\left(T_{k-1}, h_i\right) = v^T \tanh\left(W^{(a)} T_{k-1} + U^{(a)} h_i\right) \tag{2-19}$$

其中 h_i 和 T_i 分别表示事件 e_i 在编码器和解码器中的特征表示。

CYAN-RNN 模型虽然引入注意力机制，但是其主要为循环神经网络提供辅助编码。Zhu 等人在论文[90] 中提出深度注意力点过程（Deep Attention Point Process，DAPP）模型。相比 CYAN-RNN 模型，DAPP 对于事件序列的特征编码不再依靠循环神经网络，仅依靠注意力机制完成。由于 DAPP 仅使用注意力机制进行特征编码，因此无法像 CYAN-RNN 一样利用循环神经网络进行特征提取后的编码信息作为事件相似度评判依据，为此 DAPP 建立注意力机制如下：

$$\widetilde{\nu}^{(k)}\left(x, x'\right) = \frac{\nu^{(k)}\left(x, x'\right)}{\sum_{t_i < t} \nu^{(k)}\left(x, x_i\right)}, k = 1, \ldots, K \tag{2-20}$$

其中 $\nu^{(k)}$ 表示 DAPP 模型的评分函数，$\nu^{(k)}$ 由傅里叶核嵌入进行定义：

$$\nu^{(k)}\left(x, x'\right) := \mathbb{E}\left[\phi_{\omega}^{(k)}(x) \cdot \phi_{\omega}^{(k)}\left(x'\right)\right] \tag{2-21}$$

其中 $\phi_{\omega}^{(k)}(x)$ 为注意力机制中基于傅立叶特征和线形特征嵌入的特征编码模块。

DAPP 在傅立叶空间中直接量化在序列中一个事件被另一事件触发的可能性，并将此作为两个事件相关性的依据。

DAPP 模型仅依靠注意力机制实现针对事件序列预测问题的建模过程，但是由于不同事件间的相关性度量是在傅立叶空间内完成的，因此模型的计算开销较大。Zhang 等人于论文[56] 中提出基于自注意力机制（Self-Attentive）的霍克斯过程（Self-Attentive Hawkes Process，SAHP），结构如图2-4所示。借助于 Self-Attentive 结构，相比 DAPP 模型，SAHP 可以将注意力机制进行高效的实现。但是由于丢弃了循环结构，而 Self-Attentive 本身不能对输入事件进行时序信息的编码，因此为了强化历史序列中不同事件在时序空间上的相对关系，SAHP 模型提出时移位置编码方法对事件的时间进行特征编码。SAHP 模型的强度函数定义如下：

$$\lambda_u(t) = \text{softplus}(\mu_{u,i+1} + (\eta_{u,i+1} - \mu_{u,i+1}) \exp(-\gamma_{u,i+1}(t - t_i))) \tag{2-22}$$

$\mu_{u,i}, \eta_{u,i}$ 为建立在历史序列特征 h_i 之上的非线形变换，softplus 函数是 ReLU 函数的平滑近似。类似 SAHP，Zuo 等人在论文[57] 中提出基于 Transformer 的霍克斯过程。仅利用注意力机制构建的时序点过程模型被证明可以实现超越基于循环神经网络的相关方法，因为其继承注意力机制的优势，相比循环神经网络，其可以帮助当前事件捕捉事件序列中更早之前的事件的相关性，这种建立长期依赖的能力是良好的事件序列模型所需要的。

2.5 本章小结

在本章中，我们对事件序列建模算法进行了梳理和介绍。本章首先对事件序列预测任务进行问题定义，然后介绍了时序点过程模型的建模流程。之后分别介绍了基于传统机器学习的点过程算法和基于深度学习的点过程算法。同时为了更方便的介绍基于深度学习的相关算法，我们梳理现有的基于深度学习的序列建模算法。基于传统机器学习的点过程模型通常会设计形式固定的条件强度函数，而基于深度学习的点过程模型则通过引入人工神经网络模型，消除了模型对于显式参数化条件强度函数的需求，使得模型对更复杂的事件序列数据进行建模。

第三章　基于动态关系建模的事件预测模型

在本章中，我们提出一种基于渐进生成图的时序点过程（Progressive Generative Graph-based Temporal Point Process，PGG-TPP）模型用于解决事件序列预测问题。该模型旨在通过对不同事件间关系进行动态建模，利用历史事件序列推理出未来事件信息，从而提升事件序列预测系统的整体性能，并为其提供一种可供参考的推理依据。

3.1　研究动机

事件序列预测任务的直接目标是更准确地预测未来事件发生的时间及其标记信息，但是不可否认的是，对同一序列中的事件进行合理的关系推理对于事件序列预测十分重要。事件间合理的关系推理可以帮助模型进行对应的因果关系发现，不仅可以利用历史序列对未来做出更精准的预测，而且可以帮助研究者加深对于事件序列内部相关关系的理解，为历史事件动态复发的机制提供更合理的解释。对于事件间关系的正确理解是利用事件序列模型进行行为决策的关键，其可以为系统推理提供一定可解释性能力，在此基础上我们可以更充分利用所预测的未来事件，设计特定场景下的干预和控制措施，以引导事件序列达到我们所期望的结果。我们可以认为对于未来事件的准确预测和对于事件间关系的合理建模是事件序列预测任务的两个核心问题。

我们分析事件序列中的关系推理任务，认为若想合理实现该目标则需要处理好其中三个关键问题：（1）历史序列中不同事件间的关系在现实世界中是相对复杂的，难以被显式地定义，可以被认为是隐式的，即使是具有丰富经验的相关领域人员也难以对其准确描述。(2) 不同事件间影响关系种类并非单一，例如历史事件可能对未来事件产生促进、抑制等不同影响，以自激励过程为代表的

模型为了简化模型，做出事件间关系仅有正向激励的假设，这种假设不仅限制模型的表达能力，而且在一定程度上破坏了模型的可解释性。（3）事件间的关系是随时间的推移动态变化的，即使是两种相同类型的事件，由于所发生时间不同，受到其他事件的影响，也会有不同的相关关系。以消费者购物为例，对于快餐和冰汽水，在夏季购买快餐可能对购买冰汽水有正向激励作用，但是在冬季这种关系会变化为负向。解决好上述三个问题对于事件关系的建模至关重要。

基于深度学习的算法已经成为处理事件序列预测问题的主流方法。现有基于深度学习的点过程模型通常利用循环神经网络对历史事件序列进行特征编码[52-54]，这类方法忽略了对于事件间关系的直接建模，仅在现有序列建模模型的基础上进行改良，使其适用于事件序列相关问题。由于循环神经网络本身的黑盒性，人们即使得到模型针对未来事件的预测结果，也难以建立对事件间相关性的理解。近几年来逐步发展的基于注意力机制的时序点过程模型[56-57]，利用评分函数构建事件之间的权重系数，这种权重系数一般是通过评价事件之间的相关性来构建的，并非对于事件间因果关系的直接描述。

近年来，涌现越来越多的工作研究事件间关系推理问题。现有直接对事件序列中不同事件间关系进行建模的模型主要分为两类，分别为基于因果推断（Causal Inference）的方法和基于图神经网络（Graph Neural Networks，GNN）的方法。第一类模型为基于因果推断的方法[91-93]，通常建立历史事件之间的格兰杰因果关系（Granger Causality）来定义事件之间的相互影响，但是此类方法通常会对模型做出强有力的假设，并且由于格兰杰因果关系的限制，模型难以学习事件之间不同类型的影响关系。因此基于因果推断的模型往往难以适用于复杂的真实场景。第二类模型为基于图神经网络的模型[94-96]，通常建立事件之间的相关图，在相关图上使用图神经网络完成事件间信息的传递。例如 Wu 等人在论文[94] 中构建图偏置时序点过程（Graph Biased Temporal Point Process）模型，将节点之间的直接影响与作为偏差项的间接历史影响分开测量；Xue 等人在论文[96] 中构造图正则点过程（Graph Regularized Point Process）模型，将数据原有的社会结构作为约束来学习社交网络中用户之间的影响模式。但是基于图神经网络的方法一般建立事件之间的静态图，其关注的重点在于信息在静态图上如何进行高效的传递，这类静态的图结构难以适用动态变换的真实场景。

针对现有模型的缺陷，我们总结并思考解决事件关系推理任务中三个关键

问题的方式，并在本章中提出的 PGG-TPP 模型，其可以自动学习所观测到的历史序列中事件间的影响关系图，并且这个过程为动态的。

3.2 基于渐进生成图的时序点过程

现有的基于深度学习的事件序列预测主要使用循环神经网络和注意力机制对历史序列进行建模，而忽视了事件序列预测的核心问题，即对事件间关系的合理推理。已有处理事件间关系建模的方法以基于因果推断的方法和基于图神经网络的方法为主，这些方法存在引入过强假设、对事件间关系描述能力不足的问题，并且现有方法对于事件间关系大多采用静态建模的方法，无法捕捉事件间关系随时间的动态变化。

在本节中，我们针对现有模型，提出一种基于动态关系建模的事件预测模型，即 PGG-TPP 模型。我们首先概括性地介绍模型的整体结构，之后分别对模型的编码器结构和解码器结构进行详细介绍，并介绍 PGG-TPP 模型在训练时候的学习算法。

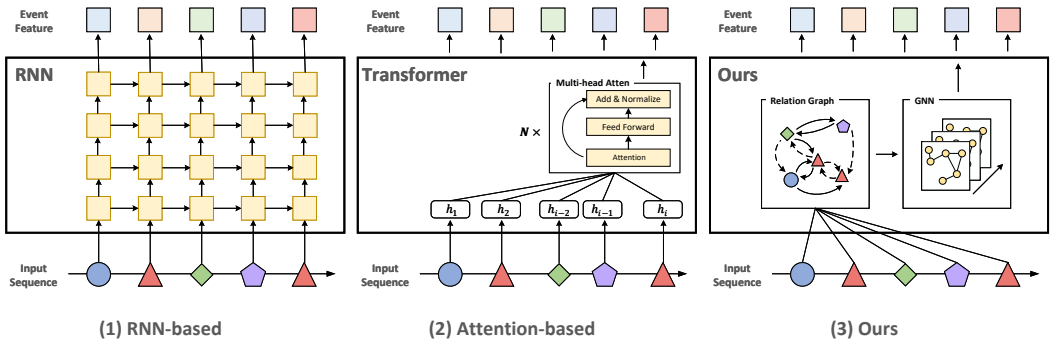


图 3-1: 事件序列预测模型中不同类型神经网络对事件间相关性的建模方式

3.2.1 模型构架

如3.1所述，对于历史序列中事件间的建模过程，主要存在三个挑战，分别为（1）现实世界中事件间关系复杂，难以被显式定义；（2）事件间影响关系种类并非单一；（3）事件间的关系是随时间的推移动态变化。我们重新思考处理三种挑战的方法，提出 PGG-TPP 模型。针对事件间关系难以被显式定义的问题，我们将事件间的相关性抽象为隐状态空间内的图结构，把整体模型定义为一个

隐变量模型（Latent Variable Model）。我们通过引入变分自编码器模型，实现对现有隐变量的模型的求解和优化。利用变分自编码器结构，我们得以实现对于事件间关系的推理过程，并生成事件间的关系图。同时，我们设计一种渐进式的学习方法，由易到难的学习事件间的关系推理过程。为了表示事件间不同种类的影响关系，我们将关系图定义为一种多维图（Multi-view Graph）结构，使得 PGG-TPP 相比现有基于图的点过程模型针对事件间影响关系具有更丰富的表达能力。得益于变分自编码器这类生成模型的灵活性，我们可以在每个时间窗口内进行当前序列内事件间的关系推理，构建事件间的动态关系，使模型不再受到静态相关关系的限制，可以更灵活的捕捉序列的动态特征。在推理出事件间多维关系图的基础上，我们利用图神经网络进行事件间的信息传递，实现对事件特征更充分的表示。由于模型对于关系图的有效推理，我们仅使用基础的图神经网络即可实现优异的性能。我们所提出的 PGG-TPP 模型与现有基于循环神经网络和注意力机制的模型对比如图3-1所示。通过对比我们可以发现，相比其他两种算法，PGG-TPP 模型对事件间相关性进行更显式的建模，为模型预测结果提供的更具可解释性的参考依据。

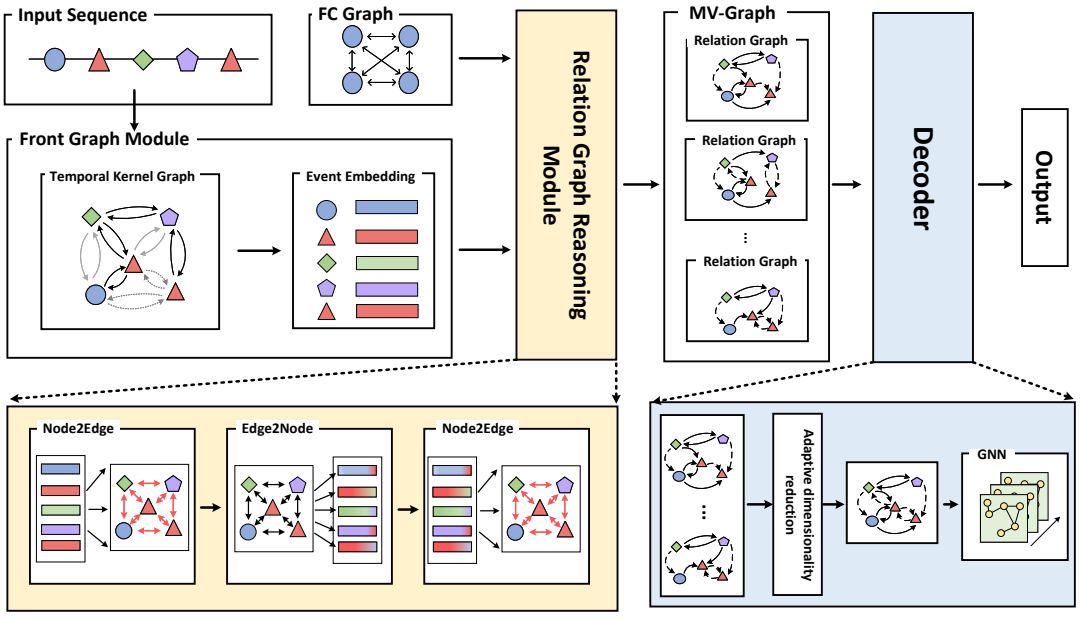


图 3-2: PGG-TPP 模型结构示意图

PGG-TPP 模型的具体结构如图3-2所示，模型整体分为两个部分，分别为编码器结构 Enc 和解码器结构 Dec。PGG-TPP 利用编码器学习当前输入事件序列的多维关系图（Multi-view Relation Graph），解码器利用此多维关系图结构和图

神经网络完成历史序列的特征编码，并根据历史序列特征实现对于未来事件的预测。模型整体的计算流程如下：

$$Graph_{mvr}, h_{enc} = Enc(\{(t_j, m_j), t_j < t_i\}) \tag{3-1}$$

$$(\hat{t}_{i+1}, \hat{m}_{1+i}) = Dec(Graph_{mvr}, h_{enc}) \tag{3-2}$$

其中 h_{enc} 表示编码器阶段学习到的特征表示，$Graph_{mvr}$ 表示上述多维关系图。为了使得图结构的学习过程更合理，我们提出一种渐进生成图（Progressive Generative Graph）的方式，用于编码器阶段的学习。同时为了在多维图上使用图神经网络，我们在解码器阶段提出一种针对多维图结构的自适应降维方法。我们分别在3.2.2小节和3.2.3小节对编码器和解码器进行详细的介绍。

3.2.2 编码器结构

PGG-TPP 的编码器实现对于输入历史序列中不同事件进行关系建模，为解码器阶段提供可靠的关系图结构。由于事件间的相关性较为复杂，若直接学习事件间关系图则难以取得较好的推理效果。为了使得编码器更充分的对事件间关系进行推理，我们设计一种基于渐进图的学习方式。我们参考课程学习（Curriculum Learning）的思路，模仿人类在学习困难知识时候的过程，先让模型学习一个较为简洁的图结构，我们称为前置图（Front Graph），然后在此基础上，再让模型对完整的关系图进行推理。我们将这个由简单图结构到复杂图结构的过程称为渐进图学习模型。与现有课程学习在数据层面的操作不同，渐进图学习模型是建立在模型结构层面的学习策略。为了实现上述过程，PGG-TPP 的编码器由两模块来完成对于图结构的逐步学习，分别为前置图学习模块（Front Graph Module，FGM）和关系图推理模块（Relation Graph Reasoning Module，RGRM），我们将分别对两个模块展开介绍。

对于前置图学习模块，我们希望编码器学习一个具有较为简单结构信息的图网络，利用此图模型，编码器可以初步得到具有图结构信息编码的历史事件特征。虽然这种结构信息相比完整的事件关系信息是简单的，但是也更容易进行学习，因此是模型进行渐进式学习的重要过程。在此阶段，我们设计一种时序核编码图（Temporal Kernel Graph，TKG）利用历史序列在时域上的相关信息构

造图结构。具体而言，前置图学习模块的输入为历史序列 $\{(t_1, m_1), \ldots, (t_n, m_n)\}$，对于历史序列上的两个事件 (t_i, m_i) 和 (t_j, m_j)，我们计算两者发生的时间差分 $(t_i - t_j)$，并使用核函数对时间差分进行编码 $K(t_i - t_j)$，我们利用编码后的信息 $d_{ij} = K(t_i - t_j)$ 作为图结构上两个事件之间边的权重。核函数可以被任意指定，在本文中，我们使用高斯核函数进行构造：

$$K\left(t_i, t_j\right) = \exp\left(-\frac{\left\|t_i - t_j\right\|^2}{2\sigma^2}\right) \tag{3-3}$$

利用核函数，我们可以构造出时序核特征图 Φ^{TKG}，并得到表示结构信息的加权临界矩阵 A^{TKG}。在时序核特征图的基础上，我们利用图卷积网络[97] 对事件进行特征编码：

$$Enc^{FGM}(C_i) = \sigma(\tilde{\Phi}_i^{TKG} C_i W_i^{FGM}) \tag{3-4}$$

其中 C_i 为事件原始特征的序列，其中 $C_i = \{c_1, \ldots, c_i\}$。对于事件的时序信息，我们利用嵌入层 Emb_{ori}^{time} 对时间间隔进行编码；对于事件的标记信息，我们使用独热编码进行初步的处理，然后使用嵌入层 Emb_{ori}^{mark} 对独热码特征进行表示。对于单个事件，特征表示过程如下：

$$c_i^{time} = Emb_{ori}^{time}(t_i - t_{i-1}) \tag{3-5}$$

$$c_i^{mark} = Emb_{ori}^{mark}(One - hot(m_i)) \tag{3-6}$$

$$c_i = fusion(c_i^{time}, c_i^{mark}) \tag{3-7}$$

此处 $fusion$ 的方式我们使用加性模型。在得到事件的初步编码后，经过前置图学习模块 Enc^{FGM}，我们得到具有图结构信息的事件特征编码 H^{FGM}。

对于 PGG-TPP 的关系图推理模块，其输入为前置图学习模块学习得到的事件特征编码 H^{FGM}，以及事件间没有先验信息的全连接图 Φ^{FC} 结构。由于关系图推理模块的输入包括事件间的全连接图，输出为体现事件间影响程度的关系图，因此关系图推理模块的关系推理过程可以视为对于无先验信息的 Φ^{FC} 的自动化剪枝过程，经过剪枝，事件间有实际相关性的边得以被保留。我们定义 HNd_i 表示推理过程中第 i 个节点的信息编码，对应为历史序列上的事件 e_i，满足 $HNd_i \in \mathbb{R}^{N_{HNd}}$；定义 $HEg_{(i,j)}$ 表示推理过程中第 i 个节点和第 j 个节点之间

连通边的信息编码，对应历史序列上的事件 e_i 和 e_j，满足 $HEg_{(i,j)} \in \mathbb{R}^{N_{HEg}}$。在此基础上，模块进行关系图推理过程可以表示如下：

$$HNd_i^{1p} = Emd^{bef}\left(H_j^{FGM}\right) \tag{3-8}$$

$$HEg_{(i,j)}^{1p} = Rea_{nd\to eg}^{one-pass}\left(\left[HNd_i^{1p}, HNd_j^{1p}\right]\right) \tag{3-9}$$

$$HNd_i^{2p} = Rea_{eg\to nd}^{one-pass}\left(\sum_{i\neq j} HEg_{(i,j)}^{1p}\right) \tag{3-10}$$

$$HEg_{(i,j)}^{2p} = Rea_{nd\to eg}^{two-pass}\left(\left[HNd_i^{2p}, HNd_j^{2p}\right]\right) \tag{3-11}$$

其中 Emd^{bef} 表示推理前的节点编码模块，$Rea_{nd\to eg}$ 表示由节点到边的推理过程，$Rea_{eg\to nd}$ 表示由边到节点的推理过程。我们的推理过程分为两个步骤，第一个步骤中，我们首先将相邻的两个节点的特征向量 HNd_i^{1p} 进行合并，并通过推理子模块 $Rea_{nd\to eg}^{one-pass}$ 学习两个节点间边的特征表示，然后我们聚合每个节点所连接的边特征，并通过 $Rea_{eg\to nd}^{one-pass}$ 再次得到节点的特征编码，此时的节点编码包含事件间的关系特征。在第二个步骤中，我们通过推理子模块 $Rea_{nd\to eg}^{two-pass}$ 重复上次操作，得到每条边的特征表示 $HEg_{(i,j)}^{2p}$。通过两个步骤的推理过程，模块模块完成了图上的信息传递与聚合。若仅进行单个步骤，则关系图中的边特征由于无法聚合多跳节点信息而不能被充分表示，因此两个步骤的推理是必要的。在此基础上，我们进行多维关系图的推理结果的输出：

$$q_\phi\left(\Phi^{MVRG} \mid HEg^{two-pass}\right) = \text{softmax}\left(Enc^{RG}(HEg^{2p})\right) \tag{3-12}$$

其中 Phi^{MVRG} 表示是事件间多种关系的多维关系图。至此，模块输出 Φ^{MVRG} 作为 PGG-TPP 的编码器阶段对于历史序列事件间关系的推理结果，以帮助解码器完成未来事件的预测。

3.2.3 解码器结构

PGG-TPP 的解码器实现对于未来事件发生事件和其对应标记信息的预测。在解码器阶段，PGG-TPP 将利用编码器 Enc 推理得到的事件间多维关系图 Φ^{MVRG}，以及编码器阶段对于历史事件的特征 C_i，利用图神经网络实现事件

特征在关系图上的信息传递，完成事件特征的重编码。需要注意的是，为了表征事件间不同种类的影响关系，PGG-TPP 模型所学习的关系图是多维的，因此无法直接使用图神经网络进行推理。为了处理此问题，我们提出一种自适应降维模块（Adaptive dimensionality Reduction Module，ADRM），将多维关系图降维为单张图结构。PGG-TPP 的解码过程利用三个模块实现，分别为自适应降维模块、图推理模块（Graph Reasoning Module，GRM）以及针对未来事件的预测模块，我们将对三个模块进行详细地介绍。

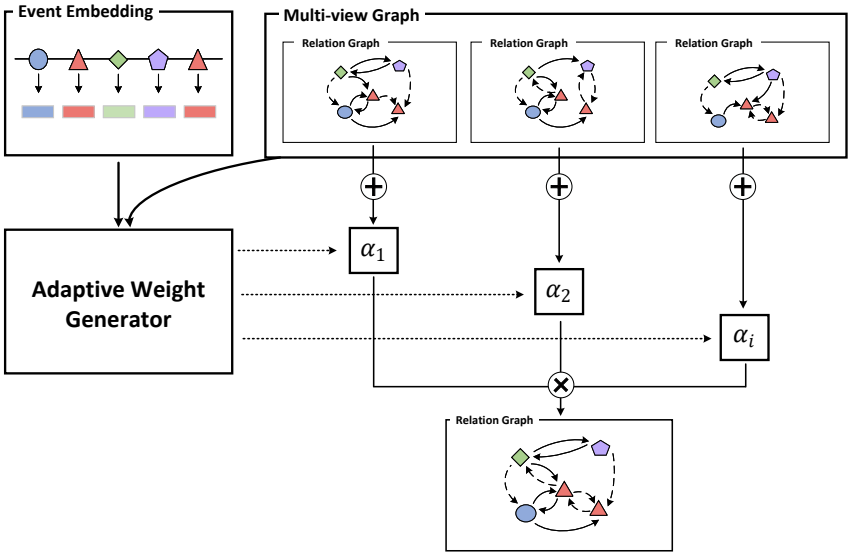


图 3-3: 自适应降维模块结构示意图

解码器 Dec 的输入为 PGG-TPP 模型在编码器阶段推理得到的事件间多维关系图。我们希望使用图神经网络进行关系图上的推理。由于关系图 Φ^{MVRG} 是能表征事件间多种关系的多维图，我们定义 PGG-TPP 模型可以表征 N^{RT} 种不同的影响关系，因此 Φ^{MVRG} 包含 N^{RT} 张单维图（Single-dimensional Graph），即 $\Phi^{MVRG} = \{\Phi_1^{MVRG}, \ldots, \Phi_{RT}^{MVRG}\}$，每张单维图 Φ_i^{MVRG} 对应一种影响关系。由于不同的单维关系图之间并非条件独立，例如一个事件如果抑制另一个事件的发生，则其正向的相关性会受到影响。因此直接使用处理单维图的推理模型并非合理的选择。现有处理多维图的方法主要分为两种，以 mGCN[98] 为代表的一类模型通过建立多维图中不同维度节点之间的联系，实现多维图上的信息传递；以 Multi-GCN 为[99] 的另一类模型利用子空间分析（Subspace Analysis）和流量形学习（Manifold Learning）等方法对多维图进行合并以实现降维的目的。然后现有处理多维图的方法多针对无边权重信息的多维图进行分析，我们参考第二种

思路，针对加权多维度图设计一种自适应降维模块。此模块输入当前事件节点的特征表示，以及多维图的结构信息，利用自适应权重生成器（Adaptive Weight Generator）为每张单维关系图生成对应的权重，通过对多个不同单维关系图进行加权，得到最终表示事件间影响情况的关系图。自适应降维模块的整体结构如图3-3所示。对于 N^{RT} 张单维关系图，我们分别生成其对应的自适应权重 α_i，计算过程如下：

$$\alpha_i = \begin{cases} \frac{1}{N^{RT}+\exp\left(-GW\left(\Phi_i^{MVRG},C\right)\right)}, & \text{if } 0 < i < N^{RT} \\ 1 - \sum_{j=1}^{N^{RT}-1} \alpha_j, & \text{if } i = N^{RT} \end{cases} \tag{3-13}$$

其中权重编码器 GW 实现自适应权重的生成。通过降维，模型计算如下：

$$\phi^{SDRG} = \sum_{i=1}^{N^{RT}} \alpha_i \phi_i^{MVRG} \tag{3-14}$$

其中 ϕ^{SDRG} 为经过降维的单位关系图。我们在 ϕ^{SDRG} 的基础上实现历史序列中事件间的信息传递。

解码器的图推理模块实现事件特征编码在关系图上的信息传递，我们通过引入图神经网络完成此功能。为了证明 PGG-TPP 模型推理事件间相互影响得到的关系图结构 ϕ^{SDRG} 对于事件序列预测任务的有效性，我们仅使用基础的 GCN[97] 进行后续的推理。在图推理模块，我们使用两层 GCN 完成推理过程，计算过程如下：

$$HS_i^1 = GCN^1(C_i) = \sigma(\tilde{\Phi}_i^{SDRG} C_i W_1^{GRM}) \tag{3-15}$$

$$HS_i^2 = GCN^2(HS_i^1) = \sigma(\tilde{\Phi}_i^{SDRG} HS_i^1 W_2^{GRM}) \tag{3-16}$$

HS_i^1 为图推理模块的隐层特征，HS_i^2 为该模块推理得到的历史事件序列的输出特征。

利用历史序列的特征编码，我们构建对于未来事件的条件分布。在 PGG-TPP 模型中，考虑图神经网络的使用，为了提升计算效率，我们使用免强度函数的建模方法。我们引入 FullyNN-TPP[59] 模型的思路，跨过强度函数直接对未来事件

的间隔分布进行建模，预测器对于未来事件的预测过程如下:

$$\hat{\tau}_{i+1} = Pred^{time}(HS_i^2) \tag{3-17}$$

$$\hat{m}_{i+1} = Pred^{mark}(HS_i^2) \tag{3-18}$$

其中 $\hat{\tau}_{i+1}$ 为模型预测的未来事件发生的事件间隔，发生时间可以计算为 $\hat{t}_{i+1} = t_i + \hat{\tau}_{i+1}$；$\hat{m}_{i+1}$ 为模型预测的未来事件所对应的标记信息。对于预测器 $Pred^{mark}$ 和 $Pred^{time}$，本文使用多层感知机模型实现。$\hat{\tau}_{i+1}$ 和 $\hat{m}_{i+1}$ 作为预测结果是 PGG-TPP 模型的最终输出。

3.2.4 学习算法

对于给定的训练样本 X，我们假设训练样本规模为 N_{train} 的模型预测未来事件信息为 $(\hat{\tau}_i, \hat{m}_i)$。由于 PGG-TPP 模型通过免强度函数的方式进行建模，目标函数被定义如下:

$$\text{loss} = \sum_{i=1}^{N_{train}} \left(l_{mark}\left(m_i, \widehat{m}_i\right) + l_{time}\left(\tau_i, \hat{\tau}_i\right) + \text{KL}\left[q_\phi(X_i \mid \Phi^{MVRG}) \| p_\theta(X_i)\right]\right) \tag{3-19}$$

PGG-TPP 模型的目标函数遵循证据下界（Evidence Lower Bound，ELBO）的形式，分为两部分。目标函数的第一部分为模型的重构误差，包含 l_{mark} 和 l_{time}，对于 l_{time} 我们使用均方误差，对于 l_{mark} 我们使用交叉熵损失函数；目标函数的第二部分为 KL 散度（Kullback-Leibler divergence），可以理解为对于分布 q_ϕ 的正则化项。我们使用反向传播方法优化网络参数，对于图中离散边值进行采样的过程是无法进行梯度传播的，因此我们使用 Gumbel 重参数化[100]（Gumbel Reparametrization）技巧，使得模型可以被正常训练。

3.3 实验与分析

我们提出一种基于动态关系建模的事件序列预测模型 PGG-TPP，在本节中，我们分别在仿真数据集和真实场景数据中进行充分的实验，通过和现有方法的比较，以及模型自身消融实验，证明 PGG-TPP 模型的有效性。在3.3.1小节中我

们将会介绍针对 PGG-TPP 实验的基本设置，在3.3.2小节中，我们将从预测精度和模型收敛性两个角度将 PGG-TPP 与现有模型进行充分的对比，在3.3.3小节中，我们将介绍针对模型本身的消融实验结果并分析，并在3.3.4中针对模型重要超参数进行敏感性分析。

3.3.1 实验设置

在本小节中，我们将介绍针对 PGG-TPP 模型进行实验的相关设置。本小节将分别对实验所使用时间序列预测问题相关的数据集、评价指标和对比方法进行介绍，并详细介绍 PGG-TPP 模型在训练中的参数细节。

1. 数据集

对于 PGG-TPP 模型的相关实验，我们使用两个仿真数据集（synthetic dataset）和三个真实场景下事件序列数据集，接下来我们分别介绍相关数据集。

a. Synthetic: Synthetic 数据集是事件序列预测领域被广泛使用的仿真数据集。Synthetic 数据集根据多元霍克斯过程（Multivariate Hawkes Process）生成样本，每个事件具有离散空间内的标记信息，多元霍克斯过程被广泛用于模拟社交网络中用户行为的生成过程[101]。我们参考 Wu 等人在论文[102] 中所提出的构建仿真数据集的方法。我们建立了相互关联的 U 个霍克斯过程，每个霍克斯过程对应一个节点。本文在第二章中对上述霍克斯过程进行了详细描述。在本实验中，我们分别使用定义 $U = 10$ 和 $U = 100$ 的两组仿真数据，分别为 Synthetic-10 和 Synthetic-100。

b. ATM: ATM 数据集[54] 是真实场景下采集到的数据集，是由银行的 1554 台自动柜员机（ATM）生成数据所组成的实际数据集，由位于北美的一家全球性银行提供。此数据集记录该银行自动柜员机收集的事件日志信息，包括故障通知单（TIKT）和错误报告，其中错误报告的类型包括：打印机错误、互联网数据中心断连、自动提款机模块错误、打印机监视器错误、通讯部分错误以及其它错误。事件序列预测模型将对其历史故障序列进行分析。

c. IPTV: IPTV 数据集[103] 是中国电信公司所提供的数据，记录电信网络电视（Internet Protocol Television）系统中用户的观看行为的序列，包括来自多个用户的电视节目观看事件的日志。日志信息包含每个观看记录的开始和结束时间戳，并记录用户所观看电视节目的名称和对应的的类别，包含 25 个类别的 9000

个电视节目。IPTV 数据集共采集 2012 年中 11 个月的数据，记录包含 2967 个用户的行为序列。事件序列预测模型一般对用户所观看节目进行节目类别粒度的分析。

d. Weeplace: Weeplace 数据集是 Cheng 等人[104] 所采集的 Twitter 用户所发布的兴趣点（Point of Interest，POI）数据。Foursquare 网站的用户一般通过 Twitter 软件来发布他们的 POI 信息。每个 POI 数据包含该用户所分享所在地理位置的描述信息，包括纬度和经度、以及地区相关的类别标签。Weeplace 数据集针对采集到的数据进行整理，筛选出 12422 个用户所发出的 46194 个 POI 信息。事件序列预测模型将对每个用户序列进行预测分析。

表 3-1: 数据集统计信息

Dataset	Train Events	Test Events	Mark Types
Synthetic-10	350000	70000	10
Synthetic-100	350000	70000	100
ATM	370000	182000	7
IPTV	731000	243000	25
Weeplace	98000	31000	8

为了方便统计，我们在表3-1中汇总上述仿真数据集和真实数据集的相关信息。对于仿真数据集，我们按照 5:1:1 的比例划分训练集、验证集和测试集；对于真实场景数据集，我们按照约 6：1：2 的比例划分训练集、验证集和测试集。对比实验的预测精度等相关比较仅在测试集上进行。

2. 评价指标

由于本节所提出的 PGG-TPP 模型是基于免强度函数的深度点过程模型，因此常用的负对数似然评价指标并不适用。我们使用事件序列预测任务另外两种常用的评价指标来评估 PGG-TPP 模型模型的性能。

对于评估未来事件发生时间的预测精度的评估，我们使用均方误差（Root Mean Squard Error，RMSE）进行判断：

$$RMSE(y_i, \hat{y}_i) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2} \tag{3-20}$$

对于模型预测事件标记信息的评估，我们使用预测准确率（Accuracy，ACC）作为评价指标。

3. 对比方法

我们将所提出的 PGG-TPP 模型分别与现有的五种最先进的（state-of-the-art）基于深度学习的事件序列预测模型进行对比，包含三种基于循环神经网络的点过程模型，分别为 RMTPP[52]，Intensity-RNN[54],NHP[53]; 以及两种基于注意力机制的点过程模型 SAHP[56]，THP[57]。我们对上述对比模型进行简要的介绍：

a. RMTPP: RMTPP 模型是 Du 等人在论文[52] 所提出的基于循环神经网络的深度点过程模型。其将时序点过程建模中的条件强度函数视为历史事件序列非线性函数，并使用循环神经网络对其进行参数化。RMTPP 模型将历史事件利用神经网络嵌入到一个紧凑的向量表示中，并自动从历史事件中学习相关影响的表示。

b. Intensity-RNN: Intensity-RNN 模型是 Xiao 等人在论文中[54] 提出的时序点过程模型。考虑外生影响，Intensity-RNN 模型同时对离散事件序列和连续时间序列数据进行特征学习，对于两部分输入，各使用一个单独的循环神经网络去进行特征提取。

c. NHP: NHP 模型是 Mei 等人在论文[53] 中提出的模型。NHP 模型提出连续时间内的 LSTM 网络来建模条件强度函数以解决传统霍克斯过程中的缺陷，NHP 模型使得过去事件对未来事件的综合影响可以是超加性（superadditive）、次加性（sub-additive）甚至是减性的，同时它还具有处理缺失数据的能力。

d. THP: THP 模型是 Zuo 等人于论文[57] 中提出的基于注意力机制的深度点过程模型。THP 模型通过在传统霍克斯过程中引入 Transformer 模块改善其缺陷，利用注意力机制直接对事件之间的相关性进行建模。该模型可以捕获事件序列中短期和长期依赖关系，并同时提升计算效率。

e. SAHP: SAHP 模型是 Zhang 等人在论文[56] 中提出的深度学习模型，其也是基于自注意力机制的点过程模型，并且针对原始 Transformer 的位置编码忽略事件间的时间间隔的问题，设计一种时移位置编码，提升对于事件发生时间的编码能力。

4. 训练细节

在本小节中，我们介绍模型训练的具体细节。对于 PGG-TPP 模型，我们设定

模型隐藏维度为 32，使用 dropout 技巧，设置 dropout 参数为 0.1。对于 PGG-TPP 模型中编码器的推理模型 $Rea_{nd\to eg}$ 和 $Rea_{eg\to nd}$，我们使用多层感知机模型进行建模；对于编码器的时序核特征图，我们使用高斯核函数。在训练的过程中，我们设定训练批量大小（Batch Size）大小为 1024，同时使用 Adam[105] 优化器进行模型参数优化，学习率设定范围为 $\{1e^{-3}, 1e^{-4}, 1e^{-5}\}$，同时使用学习率衰减方法，衰减系数设定为 $1e-8$，模型每次训练迭代 200 个 epoch。我们使用 Facebook 公司开发的开源深度学习构架 Pytorch 实现整个模型，所有的实验在 Ubuntu 16.04 操作系统下完成，训练过程使用 Nvidia GeForce RTX 2080Ti 显卡加速。为了对比实验的公平性，对于所有的模型，我们进行相同的数据预处理。

3.3.2 对比实验

在本节中，我们将介绍所提出的 PGG-TPP 模型与现有最先进事件序列预测模型的对比结果。我们在3.3.1所介绍的五个数据集上进行对比。我们分别从对未来事件发生时间预测精度和标记信息预测精度两个角度进行对比分析。

表3-2给出 PGG-TPP 模型与对比模型在事件预测精度上的对比结果。从表3-2的对比结果可以发现，我们所提出的基于动态事件关系建模的 PGG-TPP 算法，在五个事件序列预测数据集上的预测精度相比其他模型均具有一定的优势，这种精度的优势同时体现在对未来事件时间的预测和标记信息的预测。证明如果对事件间关系进行有效推理，仅使用简单的图神经网络模型即可以超越基于循环神经网络的模型和基于注意力机制的模型。同时我们可以发现，相比 ATM 数据集和 Weeplace 数据集，PGG-TPP 模型在 IPTV 数据集上对比现模型，体现出更明显的优势。我们认为其中一个主要的原因是 IPTV 数据集中事件标记更加丰富，这种标记信息的复杂因素加剧了事件序列预测任务的难度。在这种情况下，由于 PGG-TPP 使用基于关系推理的方法进行特征建模，使其更加具有建模事件间复杂影响因素的能力，因此更善于处理标记信息种类更多的场景。经过对比，我们同样可以发现，在仿真数据集 Synthetic-10 和 Synthetic-100 上，基于循环神经网络的深度点过程模型相比基于注意力的模型具有优势，我们认为主要的原因与仿真数据的生成方式有关。相比真实数据集，仿真数据集中每个事件序列的长度更短；注意力机制本身弱化了对于深度模型的结构假设，为模型带来灵活性的同时，使得模型更适用于长序列的建模，因此基于循环神经

表 3-2: PGG-TPP 模型与对比模型在不同数据集上的测试结果

Dataset	Model Type	Model	ACC	RMSE
Synthetic-10	RNN	RMTPP	32.44%	5.432
	RNN	Intensity-RNN	33.34%	4.343
	RNN	NHP	**33.63%**	4.476
	Attention	THP	33.25%	4.922
	Attention	SAHP	31.10%	6.067
	RelationGraph	PGG-TPP	33.38%	**4.405**
Synthetic-100	RNN	RMTPP	29.79%	5.314
	RNN	Intensity-RNN	29.91%	5.102
	RNN	NHP	30.17%	4.885
	Attention	THP	28.73%	**4.821**
	Attention	SAHP	27.02%	6.832
	RelationGraph	PGG-TPP	**30.67%**	4.934
ATM	RNN	RMTPP	76.70%	6.221
	RNN	Intensity-RNN	76.14%	3.302
	RNN	NHP	73.67%	7.031
	Attention	THP	70.71%	3.820
	Attention	SAHP	67.20%	4.591
	RelationGraph	PGG-TPP	**80.11%**	**3.108**
IPTV	RNN	RMTPP	56.67%	22.574
	RNN	Intensity-RNN	58.22%	20.218
	RNN	NHP	50.06%	18.812
	Attention	THP	72.10%	12.780
	Attention	SAHP	71.83%	13.211
	RelationGraph	PGG-TPP	**74.31%**	**11.131**
Weeplace	RNN	RMTPP	21.97%	7.320
	RNN	Intensity-RNN	23.72%	7.011
	RNN	NHP	25.17%	**6.219**
	Attention	THP	29.10%	6.695
	Attention	SAHP	28.65%	6.889
	RelationGraph	PGG-TPP	**30.38%**	6.445

网络的模型在仿真数据集上更具有优势。

3.3.3 消融实验

通过3.3.2小节的对比实验，我们已经证明 PGG-TPP 模型相比现有的其他模型具有更优秀的性能。在本节中，对 PGG-TPP 模型进行进一步分析，我们通过

对不同模块进行消融实验，证明模型各个结构的有效性。我们分别对 PGG-TPP 模型中的渐进图学习模型、多维度关系图、自适应融合模块进行相关实验。我们主要在 ATM 数据集上进行本节的消融实验。

1. 渐进图学习

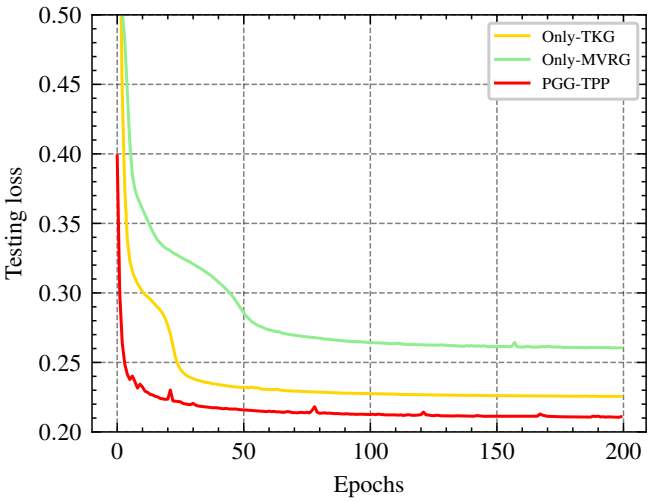


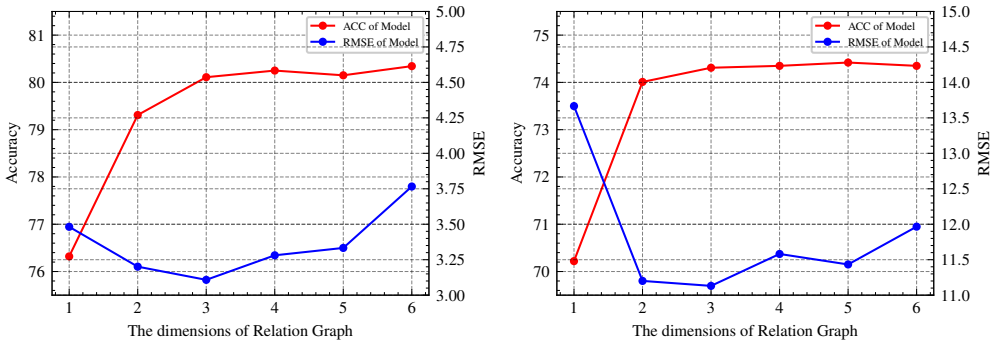
图 3-4: 不同学习方式下模型收敛性分析

我们首先探究渐进图学习模型的有效性。我们设计如下对比实验，即我们分别对比 PGG-TPP 和仅使用时序核特征图的模型，以及直接学习多维关系图而不使用时序核特征图的模型。实验结果如表3-3所示。通过对比我们可以发现，相比其他两种学习方式，顺序学习两种图结构时，PGG-TPP 具有最好的性能。这证明这种从简单到复杂的学习策略可以帮助模型更好的学习事件间的影响关系。

表 3-3: 不同学习策略下模型的性能对比（以 ACC 和 RMSE 度量）

Model	ACC	RMSE
Only-TKG	78.52%	3.472
Only-MVRG	77.84%	3.550
Ours	**80.11%**	**3.108**

在模型预测精度的对比之上，我们进一步对比模型的收敛情况。模型收敛实验在 ATM 数据集上进行，图3-4展示我们所提出的 PGG-TPP 模型相比其他模型的收敛情况对比。通过对比我们发现，PGG-TPP 相比仅使用时序核特征图和直接学习多维关系图的模型，具有更快的收敛速度。若不进行以时序核特征图为基础的前置图学习过程，则会明显影响模型的收敛速度，从而证明渐进图学



(a) ATM (b) IPTV

图 3-5: 关系图结构使用不同维度下模型的性能对比（以 ACC 和 RMSE 度量）

习方式的有效性。

2. 多维度关系图

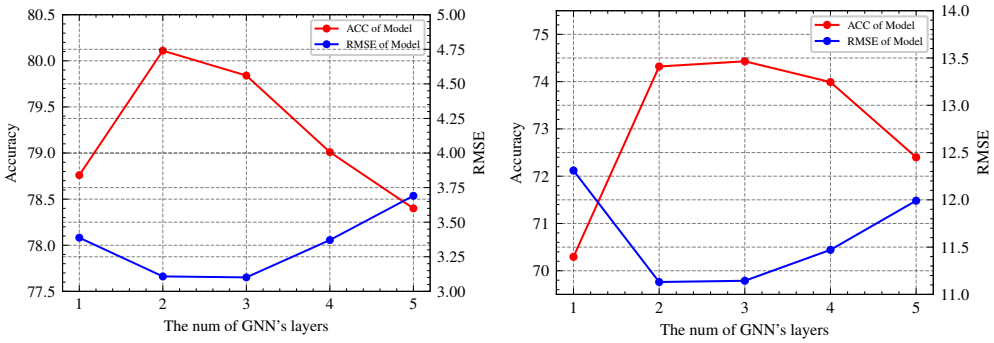
之后我们探索多维关系图对于 PGG-TPP 模型的贡献。我们分别对比多维图维度为 $N_r \in \{1,2,3,4,5,6\}$ 时，模型的整体性能。当 $N_r = 1$ 时，PGG-TPP 模型退化为直接学习单维的事件间关系图。实验结果如图3-5所示，我们可以发现，当 $N_r \in \{2,3,4,5,6\}$ 时候，相比 $N_r = 1$ 情况，模型具有明显优势，证明对多种关系进行建模对比提升模型预测精度是必要的。但是我们也发现，模型设置多维图的维度并非越多越好，当维度达到一定程度时候，模型精度提升有限，但是模型计算代价提升。在 ATM 数据集和 IPTV 数据集上，我们发现当 $N_r = 3$ 时是合适的学习策略，在保证预测精度的同时，没有过多增加模型复杂度，在一定程度上实现精度和模型复杂度的平衡；并且当 $N_r > 3$ 时，模型对于未来事件发生时间的预测精度将会明显下降。

3. 自适应融合模块

我们对于自适应模块进行消融实验，我们设计如下对比实验，针对多维图的降维，我们使用两种现有的基本策略，分别为通过直接相加和相乘来进行多维关系图内不同关系图的融合，实验结果如表3-4所示。通过对比我们发现，相比基础的加法模型和乘法模型，自适应融合模块可以帮助模型对多维关系图进行更好的结构降维。并且我们可以发现，使用乘法模型的降维方法相比使用加法模型的降维方法具有明显的劣势，证明其并非合理的特征降维方案。

表 3-4: 不同融合策略下模型的性能对比（以 ACC 和 RMSE 度量）

Model	ACC	RMSE
Add	79.52%	3.244
multi	73.95%	4.239
Ours	**80.11%**	**3.108**



(a) ATM (b) IPTV

图 3-6: 不同图神经网络层数下模型的性能对比（以 ACC 和 RMSE 度量）

3.3.4 敏感度分析

在对模型进行对比实验和消融实验之后，我们证明 PGG-TPP 模型中各个模块的有效性，在此基础上，我们研究模型中重要超参数对于模型的影响程度，对模型敏感度进行分析，此相关实验帮助我们更深层次理解 PGG-TPP 模型。我们分别探索图推理模块中图神经网络层数的选择、编码器结构中关系推理的步骤数目选择、以及时序核函数中不同核函数的优劣。

1. 图神经网络层数的选择

我们首先探索解码器中图推理模块所使用图神经网络的层数 N_g 对于 PGG-TPP 模型整体性能的影响。实验结果如图3-6所示，我们可以发现，当 $N_g = 2$ 时模型使用两层图神经网络即可以达到一定水平。当 $N_g > 3$ 后，随着层数的加深，模型对于未来事件标记信息和发生时间的预测性能均有下降趋势，我们认为其中一个重要原因是由于图神经网络模型在训练过程中产生的过度平滑问题。但是 PGG-TPP 没有针对此问题使用更复杂的图神经网络进行推理，而是选择使用简单的图神经网络，在此基础上模型即可以实现超越现有先进模型的效果，也在一定程度上证明建模事件间有效的关系结构对于事件预测任务的重要性。

2. 关系推理的步骤数目

其次我们探索编码器模块中，关系推理步骤 N_s 对于模型的影响。在 PGG-TPP 模型中，对于事件间关系推理，我们使用两个阶段的推理过程，每个推理过程包含一组节点到边的推理和边到节点的推理过程（即"Node->Edge" 和"Edge->Node"）。我们分别使用不同数目处理阶段对于模型性能的影响，实验结果如表3-5所示。由实验结果表明，随着推理步骤的增加，模型的性能趋于饱和。考虑模型推理效率和模型复杂度，我们定义 $N_s = 2$。

表 3-5: 不同关系推理步骤下模型的性能对比（以 ACC 和 RMSE 度量）

Model	ACC	RMSE
$N_s = 1$	77.33%	3.394
$N_s = 2$	80.11%	3.108
$N_s = 3$	80.08%	**3.096**
$N_s = 4$	**80.13**%	3.135

3. 核函数的对比

我们在前置图学习阶段引入时序核特征图作为先导图进行学习，在 PGG-TPP 模型中，我们使用高斯核函数进行特征编码，在本小节中，我们列举其他常用的核函数，使得 PGG-TPP 模型可以针对不同的数据集选择合适的核函数。我们引入核方法中常见的核函数，包括线形核函数（Linear Kernel）、多项式核函数（polynomial kernel）、以及拉普拉斯核函数（Laplacian kernel），其中核函数的具体定义如表3-6所示。

表 3-6: 不同类型核函数定义

Model	formula
Linear Kernel	$K\left(x_i, x_j\right) = x_i^T, x_j$
Polynomial kernel	$K\left(x_i, x_j\right) = \left(x_i^T, x_j\right)^d, d >= 1$
Gaussian kernel	$K\left(x_i, x_j\right) = \exp\left(-\frac{\left\|x_i - x_j\right\|^2}{2\sigma^2}\right), \quad \sigma > 0$
Laplacian kernel	$K\left(x_i, x_j\right) = \exp\left(-\frac{\left\|x_i - x_j\right\|}{\sigma}\right), \quad \sigma > 0$

我们展示四种核函数在 ATM 数据集上的对比实验结果，如表3-7所示，供读者参考。通过对比我们发现，在 ATM 数据集上，使用高斯核函数和拉普拉斯核函数模型可以取得较好的预测性能。

表 3-7: 不同核函数下模型的性能对比（以 ACC 和 RMSE 度量）

Model	ACC	RMSE
Linear Kernel	77.21%	3.580
polynomial kernel	78.38%	3.444
Gaussian kernel	**80.11%**	**3.108**
Laplacian kernel	79.86%	3.135

3.4 本章小结

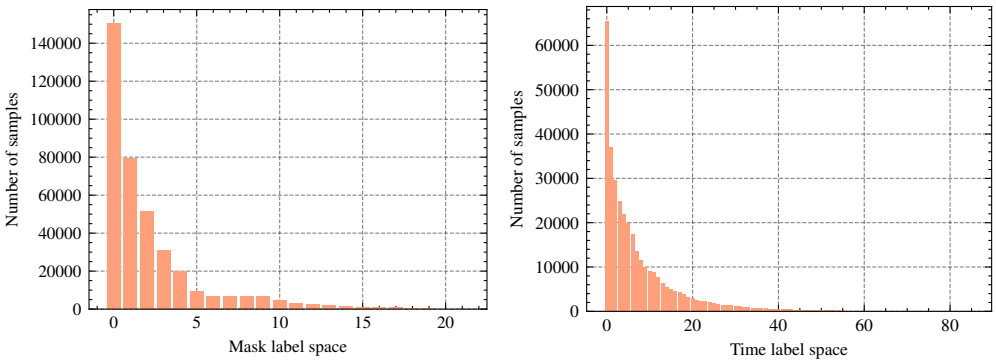
在本章中，针对事件序列预测问题，我们提出一种基于渐进生成图的深度点过程模型，即 PGG-TPP 模型，该模型可以实现对于历史序列中不同事件间关系的动态建模。针对真实场景下事件间关系复杂难以被显式定义的问题，PGG-TPP 模型引入变分自编码器结构，将事件间关系图定义为隐变量，利用隐变量模型推理事件间相关关系，同时为了使得模型可以学习复杂的关系网络，PGG-TPP 模型使用一种渐进图学习模型，并设计一种时序核特征图作为前置图进行学习；针对事件间关系非单一的问题，PGG-TPP 将事件间关系图定义为一种多维图结构，同时设计一种针对多维关系图的自适应降维模块，方便图神经网络对关系图进行处理；由于生成式模型的引入，使得模型可以对于历史序列间关系进行动态建模。我们在五个数据集上和现有先进模型进行对比，证明 PGG-TPP 模型的有效性，同时我们设计消融实验证明模型各模块的有效性，最后我们进一步对模型进行敏感度分析，研究重要超参数对于模型的影响。

第四章　基于软标签的事件序列预测辅助训练方法

在上一章中，针对事件序列预测问题，我们从模型结构的角度出发，设计一种基于动态建模事件间关系的 PGG-TPP 模型。虽然 PGG-TPP 模型可以通过推理历史序列中事件间的相关关系，提升模型预测精度。但是对部分真实场景下事件序列数据不平衡的问题，PGG-TPP 难以有效解决。针对此问题，在本章中，我们从模型目标函数和训练方式出发，提出一种基于双重平衡软标签的辅助训练方法（Dual-balanced soft labels for Auxiliary training，DBSL-Aux）。DBSL-Aux 模型通过设计事件时间域和标记信息域的软标签，结合深度解耦学习，缓解数据不平衡问题对于事件序列预测模型特征表示学习的影响，帮助模型提升预测性能。

4.1　研究动机

事件序列预测问题的核心任务是建模事件间的相关关系，预测未来事件的发生信息。事件序列预测模型若想较好地完成上述两个核心任务，针对不同的历史事件学习一种良好的特征表示是基础。例如在对未来事件进行预测的过程中，模型建立对于历史序列特征的条件概率分布，如果相关模型无法对历史事件及序列进行有效的特征表示，那么即使所构建的条件概率分布模型是合适的，模型整体也无法取得较好的预测结果。但是针对真实事件序列数据进行表示学习（Representation Learning）面临诸多挑战[106]，相关挑战可以被总结为结构保持（Structure-preserving），内容包含（Content-preserving），数据稀疏（Data sparsity）以及可伸缩性等问题。由于基于深度学习的相关方法可以学习原始输入数据的高阶非线形的特征，因此可以对数据进行自动特征表示和提取。得益于其优秀的表示学习能力，基于深度学习的时序点过程模型目前成为事件序列预测领域



(a) 信息标记空间中的长尾分布 (b) 时序空间中的长尾分布

图 4-1: 事件序列数据中的长尾分布问题

的主流方法。

基于深度学习的表示学习相关方法不断发展的一个重要原因是相关研究人员在不同领域构建了丰富的数据集，这些人工构造的数据集通常在收集时考虑类平衡的问题，以确保数据集中不同类别的数据之间数量差距较小。然而在现实世界中，不同类别的样本数量通常是不平衡的。少数类别占据大量样本，被称为头部类别（head categories）；而大多数类别只能占据少量样本，也称为尾部类别（tail categories），这种数据不平衡问题被称为长尾分布（long-tailed distribution）问题。相比自然数据集，由于用户行为本身具有较强的偏好性，这种长尾分布问题在事件序列数据中更为突出。对于事件序列相关数据，这种不平衡现象不仅体现在离散的标记信息空间，同样体现在连续的时序空间，图4-1展示 StackOverow 数据集中两种长尾分布现象。数据的不平衡加剧表示学习中数据稀疏的挑战，若我们在具有长尾分布的数据集上直接使用通用训练策略来训练模型，会导致模型性能显著下降。

现有基于深度学习的点过程模型，通过引入深度学习相关特征学习网络，例如长短期记忆网络[52-53]、注意力机制[56-57]、时序卷积[85] 等模型，提升点过程模型对于历史序列中不同事件的特征表示能力，但是并没有针对数据本身的长尾分布问题进行有效的处理，导致在模型训练的过程中，会产生对于头部类别所对应事件的偏向性，从而降低模型对于尾部类别所对应事件的预测精度。

然而现有处理深度学习中不平衡问题的相关方法，大多是针对图像识别、自然语言处理领域的数据进行设计的，例如以焦点损失函数[107]（Focal loss）为代表的代价敏感学习方法（Cost-sensitive based methods），和以过采样（Over-sampling）

策略为代表的重采样方法（Re-sampling based methods）。这些方法虽然可以有效处理数据不平衡问题对于深度学习模型表示学习过程的影响，但是现有深度不平衡学习相关方法仅能处理图像数据、自然语言数据等离散类别空间内的长尾分布问题，无法同时处理事件序列数据中连续时间域内的长尾分布问题。同时处理离散标记空间和连续时序空间中的长尾分布问题是处理事件序列预测中长尾分布问题的关键。并且现有深度不平衡模型，例如深度解耦学习模型，大多使用硬标签（Hard Label）来监督模型学习过程，即在样本对应的标签中仅有正确类别被标记为 1。我们认为使用硬标签并不是学习特征表示的最佳方式，在事件序列预测模型的学习阶段使用硬标签会增加模型对于头部类别的过度自信并加剧头部类别的模型偏差。

针对上述挑战，我们重新思考基于软标签（Soft Lable）的监督方法对于事件序列预测任务中长尾分布问题的作用，并提出一种基于双重平衡软标签的模型辅助训练方法，即 DBSL-Aux 模型。该模型可以同时处理事件序列数据中离散信息标记空间和连续时序空间中的长尾分布问题，帮助模型提升对于未来事件的预测性能。

4.2 基于双平衡软标签的辅助训练模型

基于深度学习的事件序列模型虽然使用了深度神经网络模型提升整体的特征学习能力，但是由于事件序列数据本身存在长尾分布问题，阻碍深度学习进行特征表示学习的效果。现有模型并没有针对这种长尾分布问题进行有效的处理。为了处理上述问题，在本节中我们提出一种基于双平衡软标签的辅助训练方法。我们将分别从模型整体结构、表示学习阶段、预测器学习阶段、以及辅助网络结构等角度介绍该模型。

4.2.1 模型结构

对于事件序列预测任务，现有基于深度学习的方法无法有效处理数据内部的长尾分布问题，这会影响现有模型对于事件序列数据的特征表示能力。对于事件序列，长尾分布问题不仅存在于离散的标记信息空间，同时存在于连续的时序空间中，现有处理长尾问题的方法大多关注训练数据在离散标签空间中的

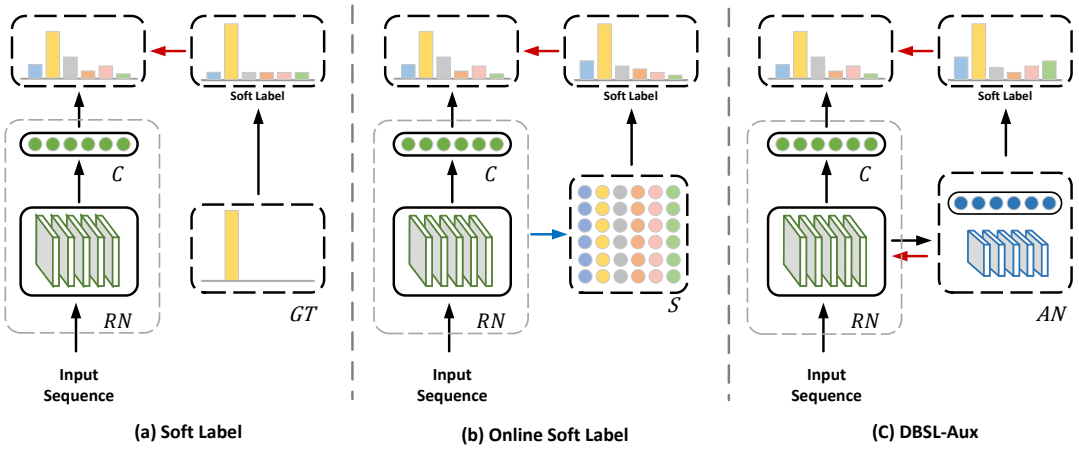


图 4-2: DBSL-Aux 算法与现有软标签生成算法的对比

不平衡分布问题，无法同时对离散空间和连续空间中的长尾分布模型进行处理。并且深度学习中现有处理长尾分布问题的方法大多使用基于硬标签的方法对模型进行监督，我们认为相比硬标签，使用软标签有益于提升模型的最终预测性能。

为了处理上述问题，我们提出一种基于双平衡软标签的辅助训练模型，即 DBSL-Aux 模型。针对事件序列在离散空间和连续空间均存在长尾分布的问题，我们整体通过引入深度解耦学习构架[108]，将模型训练过程中的特征学习阶段和预测器学习阶段相分离，分阶段学习通用的特征提取器和无偏的预测器，帮助模型缓解长尾分布问题。对于连续空间中的不平衡问题，在预测器学习阶段我们引入标签分布平滑（Label Distribution Smoothing，LDS）方法对时序空间的分布进行建模，在此基础上使用代价敏感学习方法进行模型训练。在解耦学习框架的基础上，我们使用软标签生成方法对其进行改进，提升模型在表示学习阶段和预测学习阶段的整体效果，具体而言，软标签可以帮助模型在特征表示学习阶段学习更紧凑的特征表示，并且可以帮助模型在预测器学习阶段降低模型的过度自信问题，提升模型对于尾部类别的预测效果。为了实现上述目标，我们需要在离散标记空间和连续时序空间为模型生成对应的软标签。由于解耦学习将模型两个学习阶段分离，现有的标签平滑相关方法无法被使用。我们设计一种辅助模型，分别为模型两个阶段的学习生成合适的软标签。对于标记空间，我们直接利用辅助模型推理离散空间的标签分布；但是对于时序空间，由于其为连续空间，现有的软标签方法不能被有效定义，针对此问题，我们设计一种时

序空间内生成模型，直接对间隔时序分布进行建模，为基础网络提供基于间隔分布的软标签。我们所提出的基于软标签的训练方法，与现有软标签生成模型：标签平滑[109] 和在线标签平滑[110] 的对比如图4-2所示。同时，由于我们的辅助网络仅在训练过程中为模型提供软标签作为监督信息，而不改变基础网络结构，因此不会影响模型在推理阶段的速度。

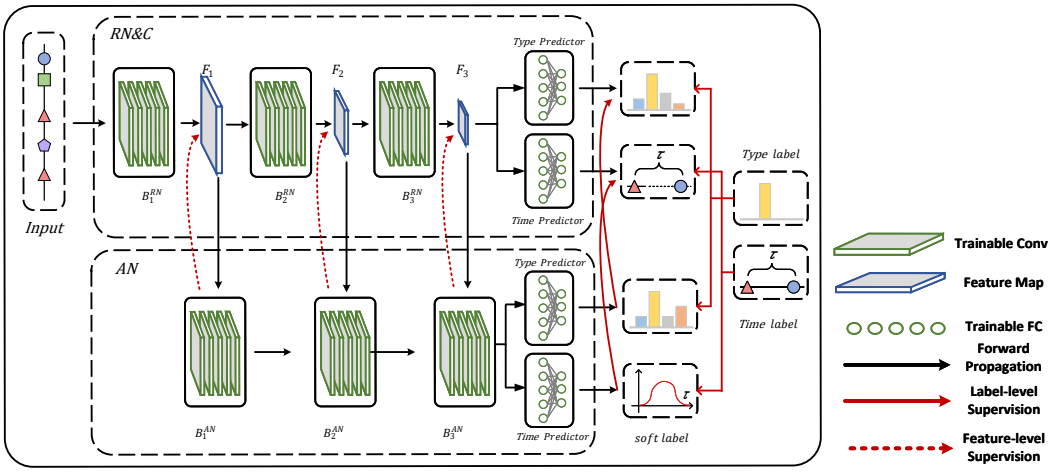


图 4-3: DBSL-Aux 模型结构示意图

DBSL-Aux 模型的整体结构如图4-3所示。模型的整体结构分为两个部分，分别为基础网络和辅助网络。其中基础网络为模型在训练和推理阶段所使用的网络，为模型的主体；辅助网络仅在模型的训练阶段为基础网络提供辅助监督信号，而不参与模型的推理阶段。在模型的表示学习阶段，DBSL-Aux 使用基于实例采样的训练方法；在预测器学习阶段，DBSL-Aux 使用代价敏感学习方法进行训练。为了使辅助网络对于基础网络进行更充分的监督，我们将其对于基础网络的监督分为两种，分别为标签粒度的监督和特征粒度的监督。同时为了使得事件特征在辅助网络中被充分学习，我们设计一种复杂的双向融合结构（complex two-way fusion structure），通过多尺度信息融合的方式使得辅助网络进行更充分的事件表征。我们分别在4.2.2和4.2.3中对两个阶段的学习进行详细介绍，并在4.2.4小节对所设计的辅助网络结构进行介绍。

4.2.2 表示学习阶段

若使用 DBSL-Aux 方法对模型进行训练，模型的整体结构分为两部分，分别为基础网络（Basic Network，BN）和辅助网络（Auxiliary Network，AN）。如4.2.1小

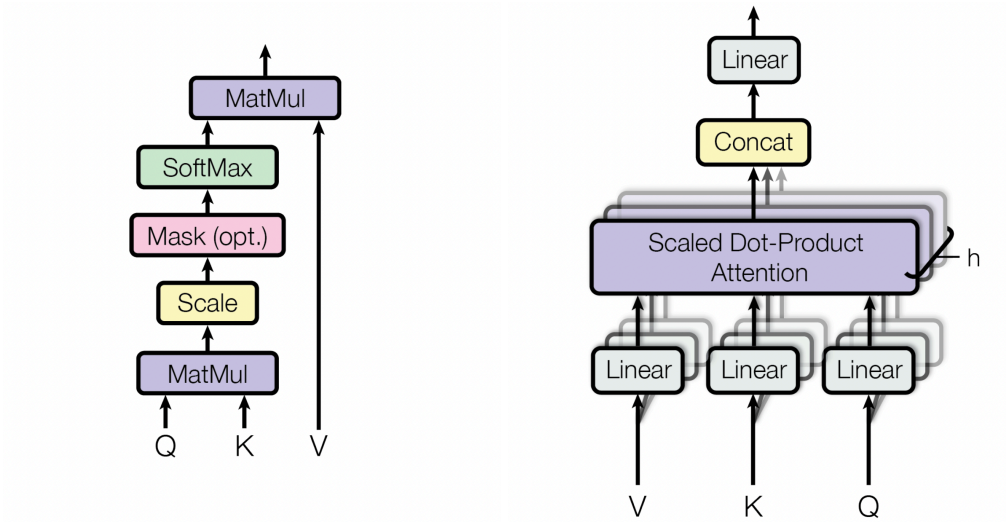
节所描述，基础网络是 DBSL-Aux 模型对未来事件进行预测的主要结构；辅助网络则是在深度解耦学习框架下为基础网络两个阶段的学习提供辅助训练用的软标签。由于两个阶段的学习方式不同，所以辅助网络在两个阶段具有生成不同特征的软标签。其中基础网络是基于时序点过程模型构建的，包括对历史事件序列进行特征提取的特征提取网络和构建历史序列与未来事件相关性的连续时间条件强度函数（Continuous Time Conditional Intensity），在此基础上，为了实现更优的性能，我们使用额外的预测器对未来事件信息进行预测，包括对于未来事件的时间预测器（Future Time predictor，FTP）和标记信息预测器（Future Mark predictor，FMP），其中时间预测器直接对未来事件所发生的事件间隔 τ_i 进行预测。辅助网络则包括特征提取模块和软标签生成器，为了保证辅助网络与基础网络的适应性，辅助网络的输入并非原始的事件序列，而且基础网络所提供的中间层特征。辅助网络利用标记软标签生成器（Mark Softlabel generator，MSG）和时序软标签生成器（Time Softlabel generator，TSG）为基础网络中所对应的两个预测器分别生成辅助训练用的软标签。

在特征学习阶段我们采用基于实例平衡采样的方法对基础网络和辅助网络同时进行训练。其中训练样本所对应的真实标签通过独热码的方式对两部分网络进行监督。在训练过程中，辅助网络生成针对样本标记信息和时间间隔的软标签分布，作为辅助监督信号在标签粒度（label-level）对基础网络产生监督作用。同时为了帮助原网络进行更充分的特征表示学习，辅助网络利用中间层特征为基础网络提供特征粒度 (feature-level) 的监督。我们认为两个粒度的监督可以有效帮助基础网络提升对于数据中长尾分布问题的处理能力。

对于基础网络和辅助网络的特征提取模块，为了方便模型构建当前事件和历史事件间有效的长期依赖，我们选择使用自注意力机制作为基础构架，结构如图4-4所示。对于原始输入，我们使用与3.2.2小节通过的方法提取事件的特征，得到事件原始特征的序列 $C_i = \{c_1, \ldots, c_i\}$，对于特征序列 C_i，自注意力结构描述如下：

$$SelfAtten(C_i) = \text{Softmax}\left(Q_i K_i^T / \sqrt{M_k} V_i\right) \tag{4-1}$$

$$Q_i = C_i W^Q, K_i = C_i W^K, V_i = C_i W^V \tag{4-2}$$



(a) Scaled Dot-Product Attention (b) Multi-Head Attention

图 4-4: Self-Attention 模型结构示意图

其中 Q_i, C_i, V_i 分别为输入特征序列 C_i 经过变换得到的特征矩阵，代表注意力机制中的 query、key 和 value 矩阵。在得到自注意力机制的编码 $S_i = SelfAtten(C_i)$ 之后，我们利用两层前馈神经网络（Feed forward Neural Networks）得到历史序列的隐层表示：

$$HE^b = \text{ReLU}\left(S_i W_1 + b_1\right) W_2 + b_2 \tag{4-3}$$

上述结构被定义为一个自注意力机制模块（Self-Atten Block），我们同时使用 B 个串联的自注意力机制模块，得到事件序列特征的最终表示 HE^B，在获得历史事件序列的特征表示之后，我们使用连续条件强度函数来推断事件在下一刻发生的概率，其中强度函数定义为：

$$\lambda\left(t \mid HE_t^B\right) = \sum_{k=1}^{K} \lambda_k\left(t \mid HE_t^B\right) \tag{4-4}$$

不同类型的标记信息分别对应单独的条件强度函数，单个条件强度函数 λ_k 可以被定义为：

$$\lambda_k\left(t \mid HE^B\right) = f_k(\underbrace{b_k}_{\text{base}} + \underbrace{\alpha_k \frac{t - t_j}{t_j}}_{\text{current}} + \underbrace{\mathbf{w}_k^\top HE^B}_{\text{history}}) \tag{4-5}$$

条件强度函数被分为三部分，分别表示基础强度、当前时间、历史事件序列的影响。在条件强度函数之外，为了实现更好的预测性能，我们使用时间预测器和标记信息预测器对未来事件的发生 (τ_{i+1}, m_{i+1}) 进行预测：

$$\hat{\tau}_{i+1} = \text{Pred}^{time}\left(HE_i^B\right) \tag{4-6}$$

$$\hat{m}_{i+1} = \text{Pred}^{mark}\left(HE_i^B\right) \tag{4-7}$$

其中 $(\hat{\tau}_{i+1}, \hat{m}_{i+1})$ 为基础网络的输出结果，也是模型在训练和推理阶段对于未来事件所发生信息的预测。

由于基础网络使用 B 个自注意力模块，因此处理最终的输出 $(\hat{\tau}_{i+1}, \hat{m}_{i+1})$ 时，我们同样可以得到 B 个对于历史序列的中间层特征表示 $SF = \{F_i^1, F_i^2, \ldots F_i^B\}$，辅助网络利用基础网络的中间层特征作为输入，进行辅助标签的生成。对于事件的标记空间，可以利用 MSG 直接生成离散的分布 $s^m = (s_1^m, s_2^m, \ldots s_K^m)$。但是对于时序空间，由于其为连续空间，现有方法没有针对连续空间定义软标签。针对时序空间，我们定义 TSG 为一个生成模型，利用历史序列特征 HE^B，我们预测时序事件的连续分布 s^t：

$$s^t = TSG((\mu^t, \sigma^t) \mid HE^B) \tag{4-8}$$

其中 (μ^t, σ^t) 为分布的均值和标准差，此处我们使用高斯分布作为生成模型的先验分布，针对不同的场景，我们定义更合适的先验分布。

为了实现辅助网络对于基础网络更充分的监督，我们引入一种特征粒度的监督方案。对于输入的中间层特征序列 SF，辅助网络可以得到对应的特征序列 $SFA = \{FA_i^1, FA_i^2, ...FA_i^B\}$，由于经过辅助网络的特征学习，我们认为相比 SF，SFA 更具表达能力，因此使用 SFA 对基础网络特征 SF 进行监督，监督方式为：

$$l_{a2r}^{feа}(FA, F) = \sum_{j=1}^{D} \left\| \frac{\text{Tr}\left(F_j\right)}{\left\|\text{Tr}\left(F_j\right)\right\|_2} - \frac{\text{Tr}\left(FA_j\right)}{\left\|\text{Tr}\left(FA_j\right)\right\|_2} \right\|_2 \tag{4-9}$$

其中 Tr 为基于注意力机制的映射[111]。

在表示学习阶段，DBSL-Aux 的模型目标函数可以构造如下：

$$loss = l_{gt2b}(X) + \beta_{gt2a} l_{gt2a}(X) + \beta_{a2b} l_{a2b}(X) \tag{4-10}$$

其中 l_{gt2b} 表示样本真实标签对于基础网络的监督，l_{gt2a} 表示真实标签对于辅助网络的监督，而 l_{a2b} 表示辅助网络对于基础网络的辅助监督。三种函数的详细定义如下：

$$l_{gt2b}(X) = \sum_{i=1}^{N} -\ell\left(x_i\right) + l^m_{gt2b}(m_i, \hat{m}_i) + l^t_{gt2b}(\tau_i, \hat{\tau}_i) \tag{4-11}$$

$$l_{gt2a}(X) = \sum_{i=1}^{N} l^m_{gt2a}(m_i, s^m_i) + l^t_{gt2a}(\tau_i, s^t_i) \tag{4-12}$$

$$l_{a2b}(X) = \sum_{i=1}^{N} l^m_{a2b}(\hat{m}_i, s^m_i) + l^t_{a2b}(\hat{\tau}_i, s^t_i) + l^{fea}_{a2r}(FA, F) \tag{4-13}$$

其中 $\ell(X)$ 为条件强度函数所对应的对数似然函数。lm, lt 为针对标记信息和预测时间的损失函数。我们使用交叉熵函数构造真实标签和两个网络之间标记空间中的损失函数。对于时序空间，由于辅助网络中 TSG 生成连续空间中的间隔时间分布，因此我们使用对数似然函数对其进行监督，而使用均方误差作为真实标签对于基础模型预测 $\hat{\tau}_i$ 的监督。在训练的过程中，我们使用 $\beta_{gt2a}, \beta_{a2b}$ 作为损失函数间权重的调节系数。

4.2.3 预测器学习阶段

对于 DBSL-Aux 训练方法，在表示学习阶段，我们对整个网络的模型参数进行更新，目的是通过更充足的样本使得模型中的特征提取模块具有更通用的特征表示能力；由于更充足的样本通常具有不平衡的特征，在事件序列数据中体现为长尾分布现象，这种长尾分布现象会影响预测器对于未来事件的预测性能。因此 DBSL-Aux 在完成表示学习之后，会再次进行预测器的学习。在预测器学习阶段，我们通过使用重平衡的训练对模型基础网络的预测器进行重新训练，在此过程中，辅助网络同样生成对应的软标签作为基础网络中预测器的辅助监督方法。

我们首先介绍预测器学习阶段中的重平衡训练方法。基础网络的预测器分

为两部分，分别为时间预测器 FTP 和标记预测器 MTP，由于基础网络的输出 (τ_i, m_i) 中，未来事件标记 m_i 对应离散空间，未来事件间隔 τ_i 对应连续空间，因此需要使用不同的重平衡策略。对于标记预测器的重平衡学习策略处理事件序列离散标记空间中的长尾分布问题，由于事件序列数据中不同事件间本身具有顺序性质，因此通过采样同样难以得到分布平衡的数据，因此我们使用基于代价敏感学习的策略，通过重新分配对于不同类别的权重系统，提升模型对于尾部类别的关注程度。我们使用类平衡的 softmax 交叉熵损失[112]（class-balanced softmax cross-entropy loss，CBCE），CBCE 计算方式如下：

$$l_{CBCE}(\mathbf{z}, y) = -\frac{1-\gamma}{1-\gamma^{n_y}} \log\left(\frac{\exp\left(z_y\right)}{\sum_{j=1}^{C} \exp\left(z_j\right)}\right) \tag{4-14}$$

其中 $\mathbf{z}$ 为模型的输出预测分布，y 为样本所对应的真实标签，n_y 为类别 y 所对应的训练样本的数目，γ 为调节系统。CBCE 损失函数通过不同类别的样本数量，调节模型对于不同类别数据的关注程度，使得模型可以对于样本较少的尾部标记类别进行更有效的推断。对于事件预测器的重平衡学习策略处理事件序列连续时序空间中的长尾分布问题，由于时序空间是连续的，并且连续空间中的间隔值之间的距离是有物理意义的，因此我们无法直接进行重平衡的处理，为了处理该问题，我们引入标签分布平滑方法，利用核密度估计的方法对现有标签分布进行平滑处理，得到平滑之后的样本标签分布，在此基础上使用代价敏感学习方法，分布平滑方法方法计算如下：

$$\tilde{p}\left(\tau'\right) \triangleq \int_{\tau} k\left(\tau, \tau'\right) p(\tau) d\tau \tag{4-15}$$

其中 $p(\tau)$ 为训练样本中事件间间隔的真实分布，这个分布是不平衡的，我们利用核函数 k 进行密度估计，得到平滑后的间隔分布 $\tilde{\tau}$，并在此基础上使用代价敏感学习方法。

由于预测器学习阶段中，基础网络的特征学习模块参数已经不再更新，因此辅助网络对于基础网络的辅助监督仅包含标签粒度的监督，即：

$$l_{a2b}(X) = \sum_{i=1}^{N} l_{a2b}^{m}(\hat{m}_i, s_i^t) + l_{a2b}^{t}(\hat{\tau}_i, \hat{s}_i^m) \tag{4-16}$$

此阶段训练过程中的目标函数为:

$$loss = l_{gt2b}(X) + \beta_{gt2a} l_{gt2a}(X) + \beta_{a2b} l_{a2b}(X) \tag{4-17}$$

在预测器学习阶段，通过使用辅助网络生成的软标签对于基础网络进行辅助监督，使得基础网络所学习的目标更加平滑，能够缓解预测器对于头部类的偏向性，这种缓解效果同时作用于模型对于未来事件时间和标记信息的预测。

4.2.4 辅助网络结构

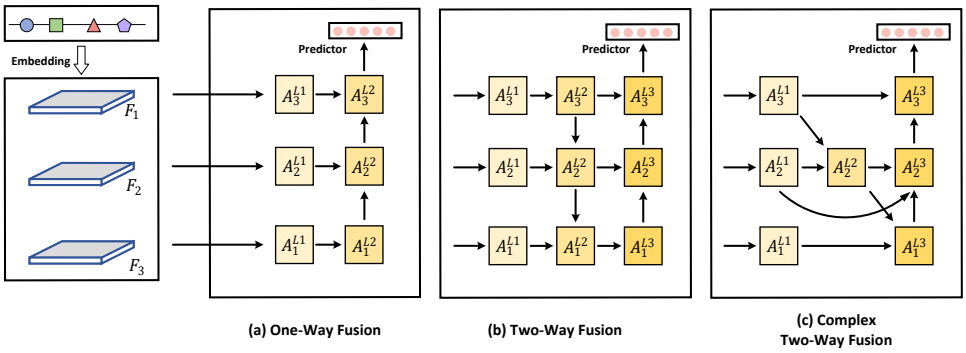


图 4-5: 辅助网络与原网络间三种信息传输路径的结构示意图

辅助网络是 DBSL-Aux 模型中为基础网络生成双重平衡软标签的结构，所提供的用于辅助训练的软标签结合深度解耦学习框架可以使得基于深度学习的时序点过程模型有效处理事件序列数据中的长尾分布问题。为了保证辅助网络与基础网络的适应性，辅助网络的输入并非原始事件序列数据，而是基础网络中 B 个自注意力机制模型的中间层输出 $SF = \{F_i^1, F_i^2, \ldots F_i^B\}$，如何针对特征序列 SF 有效处理是辅助网络所生成软标签是否有效的关键。为了使得辅助网络可以更充分地利用基础网络的特征信息，我们使用多尺度特征融合（Multi-scale feature fusion）的方法，混合使用不同深度的特征。为了使得辅助网络可以针对不同复杂程度的数据进行有效处理，我们设计其与原网络间的三种特征传递路径，分别为单向融合结构（Single-way fusion structure）、双向融合结构（Two-way fusion structure）、以及交叉双向融合结构（Cross two-way fusion structure），三种传递路径复杂度依次增加。三种特征传递路径的结构如图4-5所示。接下来我们分别对三种特征传递路径进行介绍。

a. 单向融合结构：单向融合是三种方案中模型复杂度最低的模型，仅包含一条由浅层特征到深层特征的融合路径，也被称为自底向上的融合路径（bottom-up path）。单向融合结构包含两组特征提取模块，在本章所描述的 DBSL-Aux 模型中使用自注意力模型，因此两组特征提取模块为 $Atten^{L1}$ 和 $Atten^{L2}$，两组特征模块分别包含 B 个自注意力模块，即 $Atten_{L1} = \{Atten_1^{L1}, \ldots, Atten_B^{L1}\}$。对于每一层的输入特征 SF_i，单向融合结构首先使用 $Atten^{L1}$ 初步处理，然后使用 $Atten^{L2}$ 形成自底向上的融合路径。$Atten^{L2}$ 的输出特征用于辅助网络中两种软标签生成器。

b. 双向融合结构：双向融合结构是在单向融合结构的基础上构造的，双向融合结构在其基础上增加从深层特征到浅层特征的路径，也被称为自顶向下的路径（top-down path），此时双向网络已经同时具有自顶向下和自底向上的特征传递路径，因此为双向融合结构。为了实现上述特征传递路径，在单线融合结构的基础上，我们增加一组特征提取模块，即 $Atten^{L3}$，此时网络具有三个自注意力模块，即 $Atten^{L1}$，$Atten^{L2}$ 和 $Atten^{L3}$。我们依旧使用 $Atten^{L1}$ 对输入特征进行初步处理，然后使用 $Atten^{L2}$ 构造自顶向下的特征传递路径，并使用 $Atten^{L3}$ 构造自低向上的路径，利用最后一层 $Atten_B^{L3}$ 的输出作为事件序列在辅助网络中的最终特征表示。

c. 交叉融合结构：对于不同深度的输入特征，双向融合结构实现了特征双向传递路径的初步方案，交叉双向融合机构在此基础上进行进一步改善。交叉融合结构同时使用三组特征提取模块，即 $Atten^{L1}$、$Atten^{L2}$ 和 $Atten^{L3}$，相比双向融合结构，该模型对特征传递路径进行改进，引入交叉尺度连接（Cross-Scale Connections）和加权跳跃连接（weighted skip connection）。对于中间的特征提取模块 $Atten^{L2}$，我们不再使用对应 $Atten^{L1}$ 的输出作为输入，而是使用交叉尺度连接实现自顶向下的融合，其具体操作如下：

$$Atten_i^{L2} = \text{Atten}\left(w_{i,1}^{A2} \cdot Atten_{L1}^{i} + w_{i,2}^{A2} \cdot \text{ Resize }\left(Atten_{i+1}^{L2}\right)\right) \tag{4-18}$$

对于 $Atten^{L3}$，我们使用加权跳跃连接结构，实现自底向上的特征传递：

$$Atten_i^{L3} = \text{Atten}\left(w_{i,1}^{A3} \cdot Atten_{L1}^{i} + w_{i,2}^{A2} \cdot Atten_{L2}^{i} + w_{i,3}^{A3} \cdot \text{ Resize }\left(Atten_{i-1}^{L3}\right)\right). \tag{4-19}$$

辅助网络同时为基础网络提供标签粒度和特征粒度的监督信息，因此针对输入特征序列构造有效的特征传递路径不仅影响辅助网络本身，而且可以帮助基础网络生成更具表达能力的中间特征。三种特征传递路径的实验对比结果在4.3.3中进行介绍和分析。

4.3 实验与分析

在本章中，针对事件序列任务，我们提出一种基于双重平衡软标签的辅助训练模型，即 DBSL-Aux 模型。在本节中，我们通过此模型和现有模型在真实数据集中的实验对比，证明所提出的模型可以有效处理事件序列数据中的长尾分布问题。我们在4.3.1中介绍本次实验的基本设置，在4.3.2中介绍针对 DBSL-Aux 模型中不同模块的消融实验，并在4.3.3中进行模型的敏感性分析。

4.3.1 实验设置

在本小节中，我们将介绍针对 DBSL-Aux 模型的相关实验设置，包括本次实验所使用的数据集，评价指标和对比方法，之后介绍 DBSL-Aux 模型训练过程中的具体参数。

1. 数据集

对于 DBSL-Aux 模型，我们将在三个真实场景数据数据集下开展实验和分析。我们将对所使用的数据集进行简单的介绍。

a. Retweets: Retweets 数据集是 Zhao 等人在论文[113] 中所采集整理的数据集。该数据集记录了相关用户于 2011 年 10 月 7 日至 11 月 7 日期间，在 Twitter 上转发新推文的相关情况，对于每次转发，采集相关推文相关信息，包括原始推文的时间、转发时间、转发者的关注人数，并根据粉丝数的多少将转发者分为“较多关注”、“中等关注”、“较多关注”三类。此数据集整理了 166076 条推文的相关转发数据。

b. StackOverow: StackOverow 数据集是对网络大型在线问答社区 StackOverFlow 相关交互数据的积累整理。具体来说，该网站用徽章奖励用户，如果用户进行了一定数量的高质量回答，社区将向用户发放奖励徽章，以促进参与社区活动，同一徽章可以多次奖励同一用户。StackOverow 数据集收集该网站两年内的数据，并将每个用户的奖励历史作为一个序列处理。序列中的每个项目都表示获得了特定种类的徽章。徽章的获得表示了用户不同类别的活跃方式，相关模型需要通过获得的徽章对用户行为数据进行分析。

c. MIMIC-II: MIMIC-II 数据集是通过电子病历（Electrical Medical Records）所采集的数据，记录医院中病人的医疗诊断相关数据。该数据集由位于美国波士顿的柏斯以色列狄肯尼斯医学中心（Beth Israel Deaconess Medical Center）和麻省理工学院联合收集，记录了去该医学中心就诊的 53423 名患者七年间 (2001-2008) 的就诊数据，该数据集中的每个事件包含对应的时间戳和就诊结果，每个患者的诊断结果被视为一个单独的事件序列。

表 4-1: 数据集统计信息

Dataset	Train Events	Test Events	Mark Types	IF
Retweets	1629000	543000	3	10.71
StackOverow	360000	120000	22	600.87
MIMIC-II	1812	604	75	314.50

我们将所使用数据集的情况统计在表4-1中。我们利用不平衡因子（Imbalance Factor，IF）描述不同数据集内标记空间中的长尾分布情况，其被定义为该数据集中最大类别对应的训练样本数除以最小类对应的训练样本数，不平衡因子值越大则表示该数据集中的长尾分布问题更严重。对于真实场景数据集，我们按照约 6:1:2 的比例划分训练集、验证集和测试集。

2. 评价指标

由于本文所提出的 DBSL-Aux 模型主要关注事件序列数据中的分布不平衡问题，对于事件标记信息预测准确的的评估指标，除了准确率，我们使用宏 F1 分数（Macro F1 Score，Macro-F1），被定义为：

$$\text{Macro-F1} = \frac{1}{n}\sum_{k=0}^{n} F1_k \tag{4-20}$$

其中 $F1_k$ 为第 k 类所对应的 F1 Score。

$$F1_k = \frac{2 \times \text{ precision }_k \times \text{ recall }_k}{\text{precision }_k + \text{ recall }_k} \tag{4-21}$$

其中 $precision$ 和 $recall$ 分别对应精度率和召回率。对于事件发生时间准确的评估，我们同时使用均方误差作为评判指标。

3. 对比方法

对于 DBSL-Aux 模型，我们将其与现有五种先进的基于深度学习的事件序列点过程模型进行比较，除了两种基于循环神经网络的点过程模型 RMTPP[52]，NHP[53]，两种基于注意力机制的点过程模型 THP[57]，SHAP[56]，为了比较更多类别的事件序列预测模型，我们同时对比一种免强度函数的完全神经化的时序点过程，即 FullyNN-TPP 模型。FullyNN-TPP 由 Takahiro 等人在论文[59] 中提出，为了克服使用数值逼近法对积分进行求值所带来的拟合精度的降低和计算资源的消耗，FullyNN-TPP 模型使用前馈神经网络直接对条件强度函数的积分进行建模，而不是直接对条件强度函数本身进行建模。

4. 训练细节

在本小节中，我们将介绍本章所提出模型的具体训练细节。对于 DBSL-Aux 模型，我们所使用的自注意力模块数量 B 为 4，同时我们使用多头注意力机制，使用多头数目 n_{head} 为 2。对于隐层特征维度 M_{hidden}，针对不同数据集，我们分别定义为 $M_{hidden} = \{256, 1024\}$，对于自注意力机制中 key 和 value 的维度 M_{key} 和 M_{value}，我们定义为 $\{64, 128\}$。在模型的训练过程中，我们设定训练批量大小为 $8, 16, 128$，训练过程同样使用 Adam[105] 算法，学习率设置为 $\{1e^{-3}, 1e^{-4}, 1e^{-5}\}$。训练过程中使用 Dropout 技术，Dropout 参数设置为 0.1，同时使用 L2 正则化方法，正则化系数设置为 $\{1e^{-4}, 1e^{-5}\}$。模型每轮训练会进行 100 个 epoch。模型的实现基于 Pytorch 深度学习框架，所有实验使用 Nvidia RTX 2080Ti 显卡进行加速。

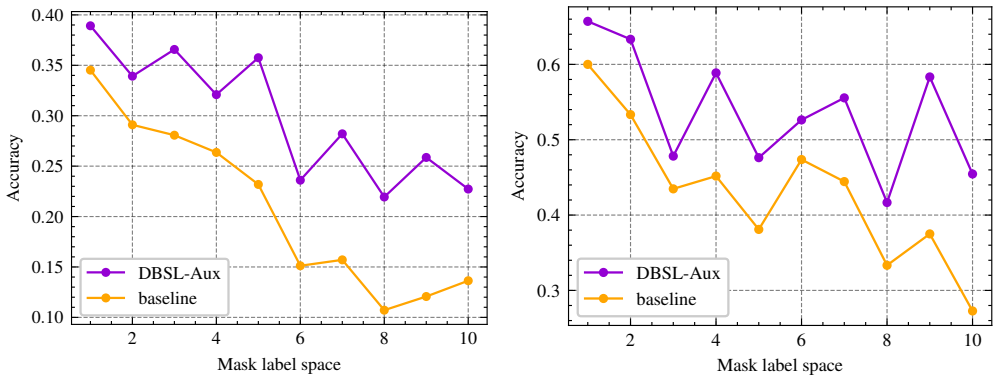
4.3.2 对比实验

在本节中，我们将介绍所提出的 DBSL-Aux 模型与现有先进事件序列预测模型的对比情况，我们利用离散空间和连续空间中的评判指标对模型进行评价。

表 4-2: DBSL-Aux 模型与对比模型在不同数据集上的测试结果

Dataset	Model Type	Model	ACC	Macro-F1	RMSE
Retweets	RNN	RMTPP	52.12%	0.4320	37.23
	RNN	NHP	54.97%	0.4872	35.21
	MLP	FullyNN-TPP	56.03%	0.5144	35.76
	Attention	THP	58.83%	0.5379	**33.24**
	Attention	SAHP	55.28%	0.5021	34.45
	Attention	DBSL-Aux	**59.27%**	**0.5414**	33.81
StackOverow	RNN	RMTPP	43.10%	0.2428	7.455
	RNN	NHP	44.21%	0.2547	6.084
	MLP	FullyNN-TPP	44.17%	0.2503	6.342
	Attention	THP	45.77%	0.2813	4.284
	Attention	SAHP	43.19%	0.2441	**3.807**
	Attention	DBSL-Aux	**46.28%**	**0.3006**	4.162
MIMIC-II	RNN	RMTPP	81.10%	0.4826	6.71
	RNN	NHP	83.66%	0.5275	3.38
	MLP	FullyNN-TPP	82.46%	0.4597	4.34
	Attention	THP	85.12%	0.5405	1.250
	Attention	SAHP	83.93%	0.5054	2.123
	Attention	DBSL-Aux	**85.75%**	**0.5535**	**1.192**

表4-2中展示了 DBSL-Aux 与五种对比模型在三个数据集上的实验结果。本章中所介绍的 DBSL-Aux 模型特征提取模块采用基于注意力机制的方法，同样可以将其拓展到其他深度学习结构中。从表4-2中的实验结果可以发现，使用 DBSL-Aux 训练的模型在三个真实数据集上均取得了优异的性能。通过 ACC 指标的比较，我们可以发现 DBSL-Aux 相比其他模型有所提升，通过对 Macro-F1 的对比，我们发现 DBSL-Aux 相比其它模型提升更明显，因为 Macro-F1 指标更可以反应模型对于不平衡数据的预测性能。我们同时发现，相比 Retweets 数据集，DBSL-Aux 在 StackOverow 数据集上相比其他模型提升更加显著，我们认为其中的一个重要原因是 StackOverow 数据集中的事件标记类别更多，当事件标记类别增多时，长尾分布问题对模型学习的影响更大，因此对于长尾分布进行针对性处理的 DBSL-Aux 具有更好的预测性能。我们同时发现 DBSL-Aux 在



(a) StackOverow (b) MIMIC-II

图 4-6: DBSL-Aux 模型对于尾部类别识别的影响（以 ACC 度量）

MIMIC-II 数据集上相比现有模型提升较小，我们认为其中一个原因是 MIMIC-II 虽然具有更多的标记类别数量，但是其数据集中训练样本数据整体较少，所以限制了以注意力机制为基础特征提取模块的 DBSL-Aux 模型的预测性能。在此基础上，为了进一步说明所提出模型可以改善长尾分布对于事件序列的影响，我们统计不同方法预测数据集中尾部类别时的识别准确率，结果如图4-6所示。通过对比可以发现，DBSL-Aux 可以有效提升模型对于尾部类别的预测准确率。

4.3.3 消融实验

在4.3.2小节中，我们通过与现有模型的对比实验，证明所提出的 DBSL-Aux 具有优秀的性能。在本小节中，我们利用消融实验对 DBSL-Aux 模型进行进一步的分析。我们设计相关实验，探究 DBSL-Aux 模型中各部分模块的有效性。我们分别探索所提出的软标签生成方法，以及辅助网络对于基础网络多种监督方法的有效性，同时我们对比4.2.4小节中所提出的辅助网络的三种特征信息传递方法。

1. 软标签生成方法

我们在所提出的 DBSL-Aux 模型中探索软标签监督方法对事件序列数据中长尾分布问题处理能力，并且提出一种基于辅助网络生成软标签的方法。为了证明我们所提出的软标签生成方法的有效性，我们分别将 DBSL-Aux 模型与现有基于软标签方法进行对比，对比方法包括标签平滑方法和在线标签平滑方法，实验结果如表4-3所示。通过对比实验可以证明，我们提出的基于辅助网络生成软标签的方法是有效的。我们同时可以发现，在解耦学习的框架下，直接使用

基础的标签平滑方法反而降低模型的预测精度，我们认为一个可能的原因是基础的标签平滑方法使用均匀分布作为进行分配的先验分布，在解耦学习框架下，这种均匀分布的先验假设会破坏预测器阶段的重平衡学习过程，因此动态的软标签生成方法对于解耦学习是更加有效的。

表 4-3: 不同软标签生成策略的性能对比（以 ACC，Macro-F1 和 RMSE 度量）

Model	ACC	Macro-F1	RMSE
baseline	45.22%	0.2768	4.540
Label Smoothing	44.93%	0.2702	4.451
OLS	45.30%	0.2759	4.201
Ours	**46.28**%	**0.3006**	**4.162**

2. 多种辅助监督方法

在 DBSL-Aux 中，我们通过辅助网络对基础网络在训练阶段提供辅助监督信息，包括对于基础网络的预测器的标签粒度的监督，以及特征提取网络的特征粒度的监督，其中标签粒度的监督分为对于事件时间预测的时序软标签，和对于事件标记信息的标记软标签。为了研究这三种辅助监督信号的有效性，我们设计如下消融实验，分别使用仅使用时序软标签、标记软标签和特征粒度监督对基础网络进行辅助训练。实验结果如表4-4所示，通过对比我们可以发现使用时序软标签或标记软标签可以分别提升模型对应部分的预测精度；若仅使用特征粒度的辅助监督，对模型的精度提升并不明显，但是当特征粒度的辅助监督配合时序软标签和标记软标签同时进行作用时，可以提升模型的性能。

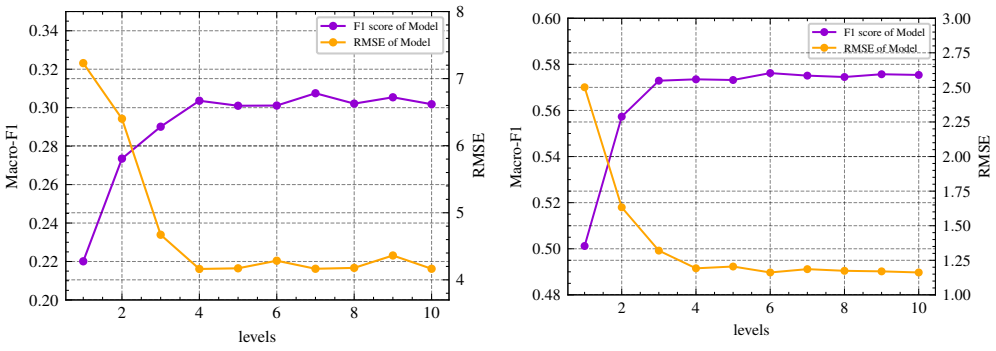
表 4-4: 不同辅助监督方法下模型的性能对比（以 ACC，Macro-F1 和 RMSE 度量）

Model	ACC	Macro-F1	RMSE
baseline	45.22%	0.2768	4.540
baseline + soft-mark-label	45.98%	0.2955	4.522
baseline + soft-time-label	44.86%	0.2751	4.173
baseline + feature-level	45.25%	0.2773	4.536
Ours	**46.28**%	**0.3006**	**4.162**

3. 信息传递路径

表 4-5: 不同类型信息传递路径下模型的性能对比（以 ACC，Macro-F1 和 RMSE 度量）

Model	ACC	Macro-F1	RMSE
Single-way fusion structure	45.72%	0.2908	4.243
Two-way fusion structure	46.04%	0.2942	4.256
Cross two-way fusion structure	**46.28%**	**0.3006**	**4.162**



(a) StackOverow (b) MIMIC-II

图 4-7: 不同特征提取模块数量下模型的性能对比（以 Macro-F1 和 RMSE 度量）

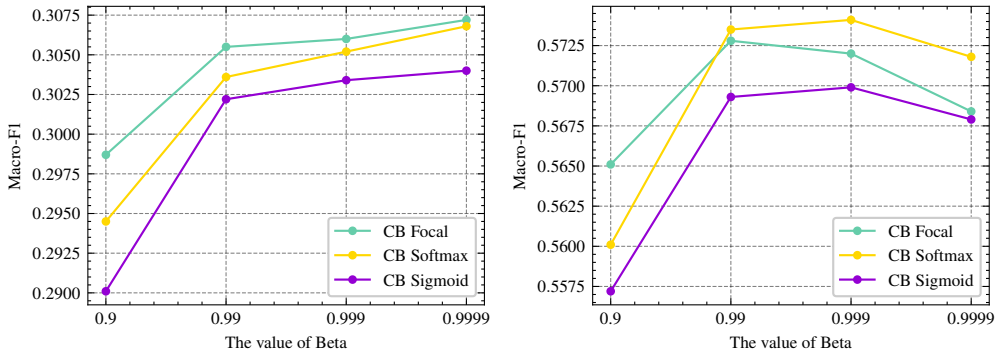
在4.2.4小节中，为了使得辅助网络对输入的特征序列进行更充分地利用，我们设计三种特征信息的传递方法，分别为单向传播结构、双向传播结构、以及交叉双向传播结构。在本小节中我们对比三种方案，实验结果如表4-5所示。通过对比可以发现，相比单向传播结构和双向传播结构，交叉双向传播结构可以取得更好的预测性能，证明使用更充分的特征传递方案可以帮助模型对输入特征进行更有效的学习和利用，提升模型对于未来事件的预测精度。

4.3.4 敏感性分析

在对 DBSL-Aux 模型进行对比实验和消融实验之后，为了进一步加深对于模型的理解，我们对 DBSL-Aux 模型进行敏感性分析，包括对于模型所使用特征提取模块数量 B 的分析和 CBCE 中权重参数 γ 的相关实验。

1. 特征提取模块数量

我们首先研究特征提取模块规模对于 DBSL-Aux 模型的影响。我们使用不同特征提取模块数量 B，并记录模型预测性能的变化，实验结果如图4-7所示。通过实验我们发现随着 B 的提升，模型预测性能逐渐提升并趋于饱和，可见使用更深的特征提取网络可以提升模型的预测性能。当满足 $B \leqslant 4$ 时，这种性能



(a) StackOverow (b) MIMIC-II

图 4-8: γ 不同取值下模型的性能对比（以 Macro-F1 度量）

的提升相对明显；但是当 $B > 4$ 时，模型的性能逐渐稳定，并且由于网络参数的增加，模型的训练和推理效率显著降低，我们认为 $B = 4$ 是保持模型性能和推理效率平衡的合适选择。

2. CBCE 中权重参数

随后我们研究预测器学习阶段所使用 CBCE 方法中，不同权重系统 γ 对于模型性能的影响。我们定义 $\gamma \in \{0.9, 0.99, 0.999, 0.9999\}$，为了说明方法的有效性，我们选择三种常用的损失函数并展示实验结果，包括 softmax cross-entropy loss，sigmoid cross-entropy loss 以及 focal loss，实验结果如图4-8所示。通过在 StackOverow 和 MIMIC-II 数据集上的对比实验，我们认为 $\gamma = 0.99$ 是相对稳定的参数值。

4.4 本章小结

在本章中，我们针对事件序列数据中事件空间和标记空间中的长尾分布问题，提出一种基于双重平衡软标签的深度点过程方法，即 DBSL-Aux 模型，该模型可以有效处理事件序列预测问题中历史序列间的长尾分布问题。针对事件序列数据在连续空间和离散空间中的长尾分布模型，DBSL-Aux 引入深度解耦学习模型，分解特征表示学习和预测器学习，使得模型在学习通用特征表示的同时，可以学习无偏向的预测器网络。同时 DBSL-Aux 模型通过软标签监督方法进一步提升解耦学习处理不平衡数据的能力，通过辅助网络为解耦学习的两个阶段生成对应的软标签，包括时序间隔软标签和标记软标签。我们在三个真实场景数据集上证明 DBSL-Aux 模型的有效性，同时对模型进行进一步分析，包

括相关消融实验和敏感性分析。

第五章　事件序列预测在图书管理系统中的应用

为了证明本文所提出的两种基于深度学习的事件序列预测模型的实用性，我们搭建了一个智慧教育平台，并将所提出的事件序列预测模型应用于其中的图书管理推荐系统中。在本章中我们将对图书管理系统，以及相关算法在其中的应用方式进行介绍，包括系统的需求分析、软件构架以及算法流程。

5.1　相关背景

随着人工智能技术的快速发展，以深度学习、知识图谱为代表的相关人工智能算法的使用逐渐成为社会未来发展的趋势。随着信息化时代的到来以及"AI+"概念的提出，通过人工智能领域的相关技术与工业界金融界等各个行业的融合，全面提升社会的生产力，为用户提供更舒适化、个性化的服务体验，成为社会的迫切需要。人工智能技术与教育行业的融合是“AI+”概念的重要体现。使用人工智能技术的相关系统已经被应用于教育行业，人工智能相关技术可以提高教学环节中的效率，简化管理任务，随着人工智能 + 教育解决方案的逐渐成熟，相关技术能够帮助教育行业填补针对学生学习和教学方面的许多空白，例如差异化和个性化学习服务、远程在线学习方式、以及自动化教学管理任务。其中差异化学习服务的一个重要体现在于，人工智能技术可以为学生提供自助且个性化的课外辅导和支持，拓展学生对于感兴趣领域的知识积累。针对此目的，我们搭建智慧教育平台，希望利用人工智能相关技术为学生提供更加智能便捷的课内外辅导服务。考虑到学生在查阅课外资料的过程中，可能缺乏相应的信息检索能力，难以获取有效的图书信息，该智慧教育平台中一个重要的功能是构建丰富的课外图书资料库，为学生提供图书介绍相关的查阅服务，帮助学生寻找感兴趣领域的优质图书，拓展学生视野，提升课外知识储备。

为了实现上述作用，我们的图书管理系统不仅希望为学生提供其所查询图书的信息，同时希望根据学生所查阅书籍的历史信息，发掘学生兴趣所在，并

为其提供感兴趣领域的优秀书籍，帮助学生发现更多优质图书。为了实现上述功能，我们引入推荐系统领域相关技术。推荐系统可以根据用户与平台的交互记录，自动发掘用户的兴趣点所在，学习到用户对不同信息或者商品的偏爱程度，并为用户提供所感兴趣的信息。由于信息化时代的到来，网络上的信息呈现爆炸式增长的趋势，推荐系统可以从大量信息中根据用户兴趣进行筛选，进而为用户提供个性化的服务。推荐系统是深度学习技术应用最广泛的领域之一，相关技术被应用于各个行业，例如电影行业、音乐行业，以及新闻媒体。现有推荐模型主要分为基于协同过滤的模型、基于内容的推荐模型以及一些混合推荐系统。目前基于深度学习的推荐模型已经逐渐成为推荐系统领域的主流方法，相关模型可以捕捉用户历史数据中复杂的交互关系，为用户带来高质量的推荐服务。

现有的推荐系统相关方法主要通过对用户与平台产生的用户-项目交互(user-item interactions）数据以及用户本身的信息进行学习，这种方法存在两种弊端。首先，许多用户在平台进行浏览的过程中并没有登陆账号，或者使用游客状态进行浏览，我们无法预先知道用户的相关信息，但是在用户允许的情况下可以通过Cookie 技术获取用户与平台的匿名交互信息，一般的推荐模型无法对交互历史较短且无身份特征的状态进行推荐；其次现有推荐系统模型主要是利用已知数据学习用户的静态特征，但是用户对于信息和商品的偏好性往往是动态的，随着时间的推移而改变。上述问题可以被描述为基于会话的推荐系统（Session-based Recommendation)。我们希望所设计的图书管理系统不仅为登陆用户进行图书推荐，也希望系统可以为以游客身份进行浏览的用户提供相关的推荐服务，因此我们将该问题抽象为一个事件序列预测问题，利用本文所提出的两种深度点过程模型，即 PGG-TPP 模型和 DBSL-Aux 模型，为此图书推荐场景下的基于会话的推荐问题提供一种解决方案。我们将在5.2.3小节详细的介绍我们所设计的算法解决方案，并在5.3小节进行系统的展示。

5.2 系统设计

在本小节中，我们对所搭建的智慧教育平台进行介绍，由于我们所提出的事件序列预测模型是用于其中的图书管理系统，所以我们将主要关注此图书管

理系统。我们分别介绍其系统需求、软件架构以及算法流程的设计方案。

5.2.1 系统需求

图书管理系统的主要目的是为用户提供需要查阅的图书的相关信息，并且根据用户查阅过程中所产生的交互信息为用户提供感兴趣图书的推荐。同时考虑到有一部分用户会以游客身份登陆系统，我们希望系统可以为此类用户提供同样的图书推荐服务。为了实现上述目的，我们图书管理系统的核心需求归纳为四个方面：

1. 账号注册和管理等基本功能：由于我们希望构造的是一个智慧教育平台中的图书管理系统，因此该系统应该具有一个用户系统的基本功能，包括用户账号的注册、登陆、验证、注销等功能，并且记录账号的基本信息。同时为了使得针对注册用户的推荐系统可以正常工作，管理系统的后台应该记录用户在平台上的交互记录，包括搜索记录和浏览记录，以便后期推荐模型可以利用交互历史构建用户个人画像。

2. 图书数据库构建及检索：图书管理系统的基本功能是为用户提供所查找图书的基本信息，因此我们需要构建并且维护一个针对图书信息的数据库，包括图书的名称、作者、出版社、出版事件、页数、类别、定价、国际标准书号（International Standard Book Number，ISBN）以及基本内容介绍，数据库支持后期图书信息的补充维护。同时我们需要为用户提供相关的搜索功能，包含图书名称搜索、关键词搜索等功能，并显示用户的搜索历史记录。

3. 针对注册用户的图书推荐：我们希望图书管理系统可以构建基础推荐模型，为已经登陆的用户构建用户个人画像，学习到用户本身的阅读兴趣，并根据用户特征为用户进行个性化的图书推荐服务，为用户提供固定数目个感兴趣的优质图书。并且由于用户与平台的交互数据是在不断生成的，因此推荐模型需要具有模型更新的能力。

4. 针对游客用户的图书推荐：在为登陆用户提供推荐服务的基础上，我们希望模型同样可以为没有登陆的游客用户提供图书推荐服务，图书管理系统需要构建辅助推荐模型实现该功能。辅助推荐模型无法获取游客用户的基本信息，仅能通过该用户当前与平台进行交互的数据进行推荐。辅助推荐模型同样需要具有模型更新的能力。

5.2.2 软件架构

本文的软件整体由两部分实现，分别为系统前端和系统后端，采用浏览器/服务器构架模型。系统前端作为与用户进行交互的界面，使得用户可以完成注册、登陆等基础操作，进行相关图书的搜索，并且可以在讨论区与同学进行交流互动。后端则实现相关功能对应的系统逻辑，并且对数据库进行维护。整体构架如图5-1所示。

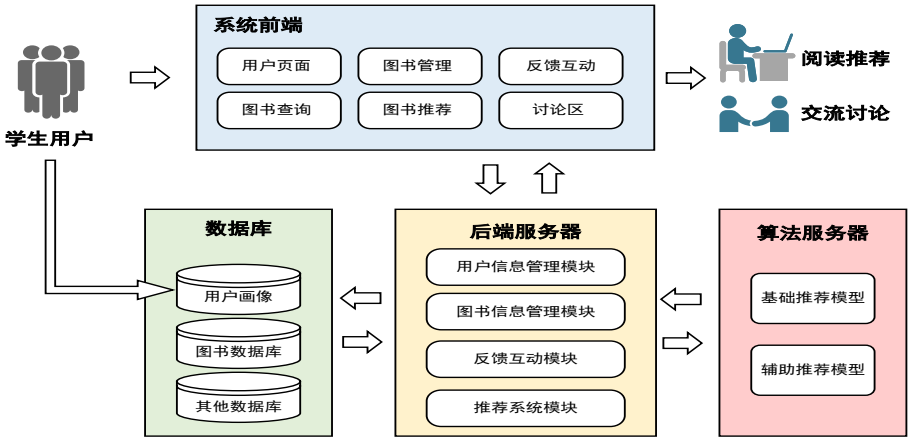


图 5-1: 系统整体架构图

对于系统前端，我们使用 javascript 语言，并采用 Bootstrap 框架实现；对于系统服务端，我们采用 Django 框架实现，并使用 MySQL 数据库进行数据存储，所存储数据包括用户的相关信息、所建模的用户画像、图书信息数据库等。针对系统服务端的算法服务部分，由于算法本身需要 GPU 进行加速，考虑部署的灵活性，我们在满足计算条件的算法服务器上进行部署。我们使用 Flask 框架在算法服务器上实现推荐算法的 REST-API 服务，算法部分利用 pytorch 构架实现。我们将推荐算法部署到具有单张 Nvidia 2080Ti 显卡的 GPU 服务器上，后端服务向其发送请求即可得到对应的图书推荐结果。

5.2.3 算法设计

图书管理的系统的算法模型包括两部分，分别为针对登陆用户提供图书推荐服务的基础推荐模型，以及针对未登陆的游客用户提供图书服务辅助推荐模型。工业级的推荐系统一般分为分为两个主要阶段，分别为召回阶段和精排阶段，这种多阶段的推荐算法是为了解决从海量数据集中进行信息筛选的计算效

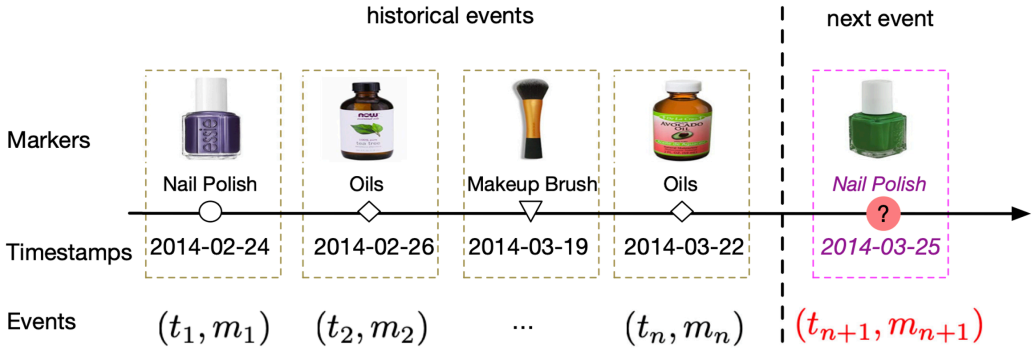


图 5-2: 利用事件序列预测模型实现基于会话的推荐

率和实时性的问题，在召回阶段，推荐系统一般会使用较为简单的模型，例如因子分解机（Factorization Machine，FM）等模型，利用较为简单的召回模型从大规模候选集中进行快速的筛选，保证整体的时间效率；在精排阶段，推荐系统则会使用较为复杂的模型和特征，在精排阶段完成精准的推荐功能，作为用户最终的推荐结果。但是由于我们所搭建的图书管理系统的数据规模相比工业界场景较小，仅具有图书类的数据，因此在管理系统推荐功能的实现过程中，我们不再使用多阶段的推荐，而是忽略召回阶段，仅使用精排阶段进行处理。

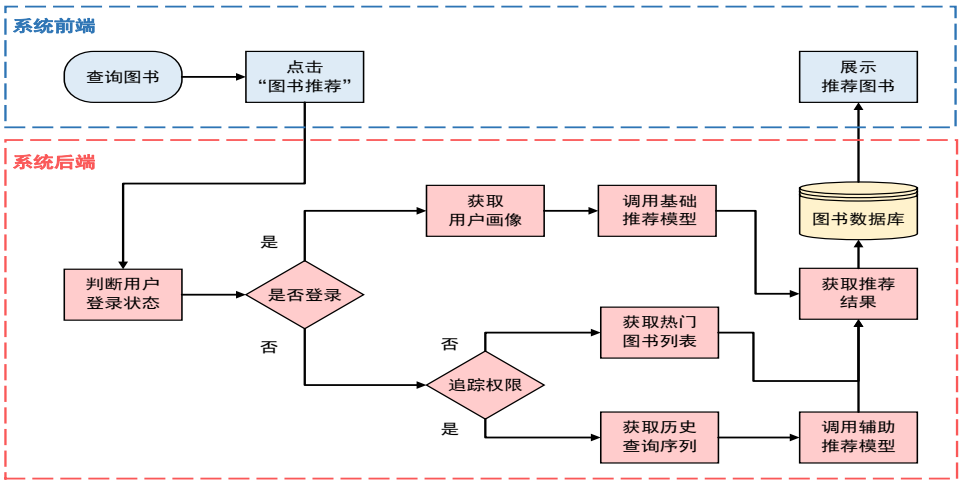
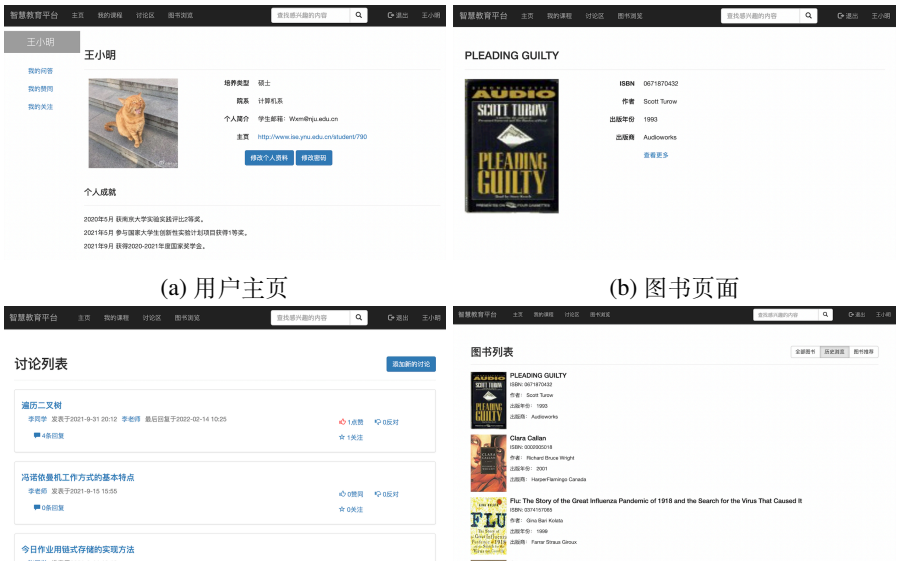


图 5-3: 系统图书推荐算法流程图

对于基础推荐模型，我们选择使用 Zhou 等人在论文[114] 中所提出的深度兴趣网络（Deep Interest Network，DIN）。深度兴趣网络是基于注意力机制的深度推荐模型，使用 *Embedding&MLP* 范式对稀疏特征进行编码，通过使用激活单



(a) 用户主页 (b) 图书页面

(c) 用户讨论页面 (d) 用户浏览页面

图 5-4: 图书管理系统基础功能展示

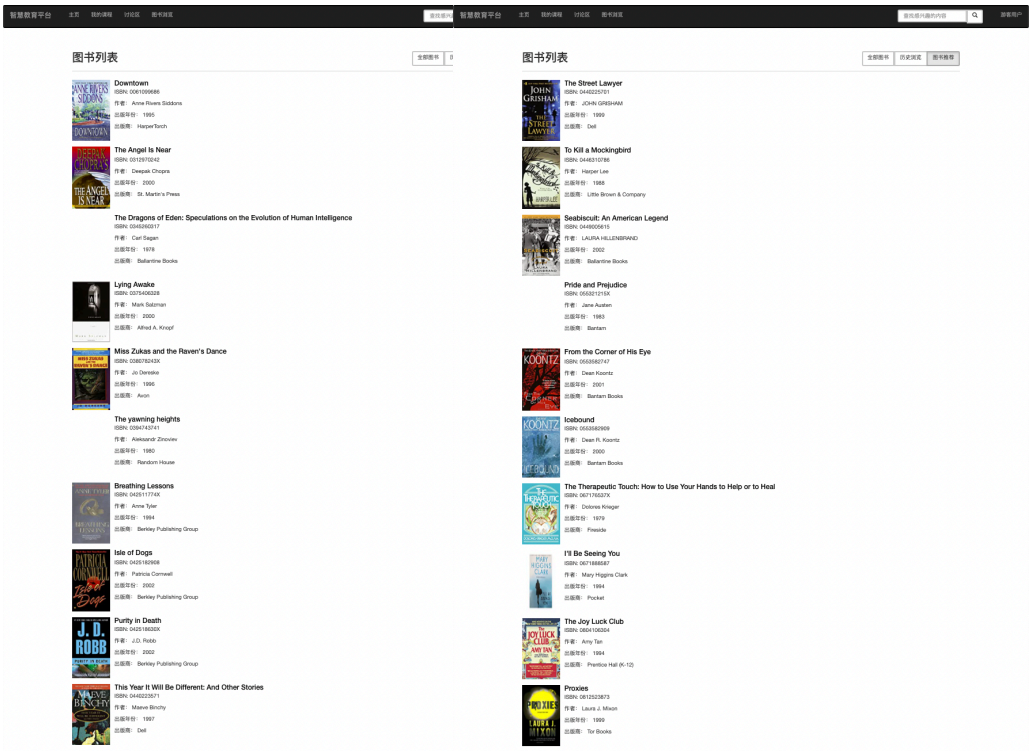
元（Activation Unit）描述用户对于项目的感兴趣程度：

$$v_U(A) = \sum_{j=1}^{H} a\left(e_j, v_A\right) e_j = \sum_{j=1}^{H} w_j e_j \tag{5-1}$$

其中 e_j 为所对应的嵌入特征，$a(\cdot)$ 是为馈神经网络。深度兴趣网络利用注意力机制处理用户兴趣的多样性和局部聚焦特征（Local Activation）。

对于辅助推荐模型，由于其仅能针对游客用户的无额外信息的交互事件进行推荐，因此该推荐过程有别于一般的推荐任务，可以视为基于会话的推荐任务，我们可以将其抽象为一个事件序列预测问题，其中一次事件表示一次用户与平台的交互，而事件的信息标记为此次交互的类型。图5-2展示了如何利用事件预测模型进行基于会话的推荐任务。我们利用游客用户在图书管理系统上针对感兴趣图书的历史浏览记录，预测用户下一次将要浏览的图书，作为针对该用户的推荐结果。我们利用本文所提出的两种基于深度学习的点过程模型作为辅助推荐模型。

系统后端中算法处理流程如图5-3所示。若用户点击“图书推荐”按钮，系统后端会对用户的登录状态进行判断，并根据不同的登录状态和浏览器历史追踪权限的情况，为用户进行相关推荐。



(a) 针对登录用户推荐　　(b) 针对游客用户推荐

图 5-5: 图书管理系统推荐功能展示

5.3 效果展示

图书管理系统基础功能如图5-4所示。在进入主页后，点击界面右上角的“登录”按钮，通过输入用户名和密码，可以进行用户登录，在图5-4(a)所示的用户主页中会显示用户的基本信息，用户可以根据需要进行个人资料的修改和密码的修改。在系统中，用户可以在搜索框中查询感兴趣的图书，点击图书后可以进入图5-4(b)所示的图书页面，浏览图书的相关信息。系统会记录记录用户的历史浏览记录，单击“历史浏览”按钮，如图5-4(d)所示，系统会显示用户近期浏览的图书。

系统的推荐功能如图5-5所示。在单击“图书推荐”按钮之后，图书管理系统可以分别为已经登陆的用户和没有登陆的游客用户进行图书推荐。图5-5(a)展示了系统为已登陆用户进行推荐服务的效果；图5-5(b)展示了系统为未登陆的游客用户进行推荐服务的效果。可以发现，在用户未登陆的情况下，图书管理系统依然可以根据该用户近期的浏览序列为其进行图书推荐。

5.4 本章小结

在本章中，为了证明本文所提出两种事件序列模型的有效性，我们搭建了一个智慧教育平台，并将 PGG-TPP 和 DBLS-Aux 模型用于图书管理系统，实现为未登陆用户提供推荐图书服务的功能。我们构建一个完整的图书管理系统，该系统可以同时为已登陆用户和未登陆用户进行图书推荐服务，充分说明本文所提出的两种深度点过程模型具有实际落地的价值。

第六章　总结与展望

事件序列预测问题是机器学习领域重要的研究方向，本文围绕事件序列预测问题进行分析和研究，将处理该问题的核心总结为事件间合理的关系推理和对未来事件的准确预测。现有的研究工作通常更关注预测的准确性，而忽视了对事件间关系进行学习并推理；并且在进行历史事件的表示学习过程中，现有模型没有处理事件序列数据中长尾分布问题，该问题将限制模型进行特征表达的能力。针对上述问题，本文提出两种基于深度学习的点过程算法，设计对应模型处理事件间关系推理和数据中两种长尾分布问题。并将所提出的算法用于一个应用系统中，证明所提出的方法的使用性。

从事件间关系推理的角度出发，本文提出一种基于渐进生成图的深度点过程模型，即 PGG-TPP 模型。PGG-TPP 模型将事件间关系推理抽象为隐变量模型，模型推理事件间的合理的相关关系。并且将关系图定义为多维图，使得模型具有表达事件间多种影响关系的能力。由于 PGG-TPP 模型直接针对历史事件间关系进行建模，因此可以在每个时间步对窗口内事件进行动态的关系推理。PGG-TPP 模型同时设计一种渐进生成图的方法，使得模型依次学习前置图和多维关系图，从简到难的的学习历史序列间的关系结构。为了证明所学习到的关系结构的有效性，该模型在推理阶段仅引入基础的图神经网络。经过实验验证，该模型在仿真数据集和真实采集数据集上均优于现有先进模型。

从优化模型对事件进行特征表示的角度出发，本文提出一种基于双重平衡软标签的辅助训练模型，即 DBSL-Aux 模型，主要处理事件序列数据中的长尾分布问题对表示学习的消极影响。我们分析事件序列数据特征，将事件序列数据中的长尾分布问题总结为离散标记空间中的长尾分布和连续时序空间中的长尾分布。为了处理该问题，我们整体引入深度解耦学习的思想，分离模型的特征学习过程和预测器学习过程。在此基础上，我们设计辅助网络生成软标签对基础网络进行监督，进一步优化模型特征学习效果和预测器偏向性问题。DBSL-Aux 模型通过引入标签分布平衡方法，使得连续空间中的代价敏感学习成为可能，并且定义连续空间上的软标签。通过在多个数据集上的实验，证明 DBSL-Aux 模

型可以有效缓解长尾分布模型对表示学习的影响，提升模型预测精度。

为了验证本文所提出两种算法的实用性，我们将上述两种基于深度学习的点过程模型应用于所搭建的智慧教育平台的图书管理系统。在图书管理系统中，我们将针对匿名用户图书推荐服务定义为基于对话的推荐问题，并通过事件序列预测模型进行处理。本文所提出的两种算法在系统中可以实现良好的推荐效果，证明模型的实用性。

在本文的研究基础上，可以从如下方向进行思考并开展进一步的研究工作。针对事件间关系推理问题，本文主要关注影响关系的推理生成过程，因此只使用了基础的图神经网络进行之后的节点信息传递，但是如何利用所生成的关系图结构进行更好的信息传递依然需要进一步探索；此外，在处理事件序列中的长尾分布问题中，本文在预测器学习阶段仅使用代价敏感学习的方式，如何针对事件序列数据进行有效的重采样方法也值得思考。最后，如何将事件序列预测问题应用到更多实际场景中，发挥其在预测决策任务中的更大作用，也具有实际的研究价值。

致　　谢

逝者如斯夫，不舍昼夜。转眼间，于南京大学三年的硕士生活即将画上句号，忆昔抚今，往昔之事历历在目，已为追忆。回首三年时光，有感慨、有感伤、有激动、有怀念。我喜欢这片校园，在南京大学生活的日子里，我曾因泥泞而步履蹒跚，也曾因星光而心弛神往，不曾改变的是心中的热情。回首三年时光，我收获了相伴我一生的宝藏。

我想感谢我的导师申富饶教授。申老师为人谦逊，做研究态度严谨，老师对于科研和生活的态度深深影响着我。在科研方面，申老师尊重我们的兴趣方向，并且为我们提供细致的指导，通过每周与老师的个人讨论，我总能有所收获；申老师教导我从解决问题的角度思考，帮助我在科研道路上走入正轨。在生活方面，申老师是我们的良师益友，老师总会设身处地为我们着想，在我遇到困难的时候老师的鼓励让我重拾信心，十分感谢申老师在这三年的生活中对我的关心和照顾。我还想感谢赵健老师，老师教授我们许多做科研的好方法，并且花费时间对我们的论文细致批改，并且提出建设性的意见。

我想感谢 RINC 实验室的同门们，尤其感谢实验室的师兄和师姐们，当我在科研上遇到困难的时候，是你们无私的为我提供帮助和指导，并且给我许多中肯的建议；与你们在生活上的交流同样让我受益匪浅。

我想感谢我的室友张永顺和张玉鹏，很庆幸研究生三年是与你们在同一个屋檐下度过。从一开始的拘谨到现在的亲密无间，是你们让我的研究生生活多了许多色彩与乐趣，感谢你们对我的真诚帮助，让我感受温暖。我想特别感谢高妍同学，是你一直以来的支持和鼓励让我可以坚持自我；是你和我一起面对挑战，不断前进。

最后，我要感谢我的父母，给予我最无私的爱和关怀，你们是我坚实的港湾，在我逆境之中给予鼓励，迷茫之中对我开导，让我可以不断向前，披荆斩棘。

参考文献

[1] FARAJTABAR M, DU N, RODRIGUEZ M G, et al. Shaping social activity by incentivizing users[C]//Advances in neural information processing systems: volume 27. : NIH Public Access, 2014.

[2] ZHAO Q, ERDOGDU M A, HE H Y, et al. Seismic: A self-exciting point process model for predicting tweet popularity[C]//Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining. 2015: 1513-1522.

[3] FARAJTABAR M, YE X, HARATI S, et al. Multistage campaigning in social networks[C]//Advances in Neural Information Processing Systems: volume 29. 2016: 4718-4726.

[4] LUKASIK M, COHN T, BONTCHEVA K. Point process modelling of rumour dynamics in social media[C]//Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing. 2015: 518-523.

[5] WANG Y, XIE B, DU N, et al. Isotonic hawkes processes[C]//International conference on machine learning. : PMLR, 2016: 2226-2234.

[6] TAGLIAZUCCHI E, BALENZUELA P, FRAIMAN D, et al. Point process analysis of large-scale brain fmri dynamics[J/OL]. arXiv preprint arXiv:1107.4572, 2011. https://arxiv.org/abs/1107.4572.

[7] XU H, WU W, NEMATI S, et al. Patient flow prediction via discriminative learning of mutually-correcting processes[J]. IEEE transactions on Knowledge and Data Engineering, 2016, 29(1):157-171.

[8] LOEFFLER C, FLAXMAN S. Is gun violence contagious a spatiotemporal test [J]. Journal of quantitative criminology, 2018, 34(4):999-1017.

[9] ZAMMIT-MANGION A, DEWAR M, KADIRKAMANATHAN V, et al. Point process modelling of the afghan war diary[C]//Proceedings of the National Academy of Sciences: volume 109. : National Acad Sciences, 2012: 12414-12419.

[10] BACRY E, MUZY J F. Hawkes model for price and trades high-frequency dynamics[J]. Quantitative Finance, 2014, 14(7):1147-1166.

[11] BACRY E, DELATTRE S, HOFFMANN M, et al. Modelling microstructure noise with mutually exciting point processes[J]. Quantitative finance, 2013, 13 (1):65-77.

[12] TOKE I M, POMPONIO F. Modelling trades-through in a limit order book using hawkes processes[J]. Economics, 2012, 6(1).

[13] LALLOUACHE M, CHALLET D. Statistically significant fits of hawkes processes to financial data[J]. Available at SSRN, 2014.

[14] FAUTH A, TUDOR C A. Modeling first line of an order book with multivariate marked point processes[J/OL]. arXiv preprint arXiv:1211.4157, 2012. https://arxiv.org/abs/1211.4157.

[15] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. Imagenet classification with deep convolutional neural networks[C]//Advances in neural information processing systems: volume 25. 2012: 1097-1105.

[16] 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. 2016: 770-778.

[17] REN S, HE K, GIRSHICK R, et al. Faster r-cnn: Towards real-time object detection with region proposal networks[C]//Advances in neural information processing systems: volume 28. 2015: 91-99.

[18] HE K, GKIOXARI G, DOLLÁR P, et al. Mask r-cnn[C]//Proceedings of the IEEE international conference on computer vision. 2017: 2961-2969.

[19] DEVLIN J, CHANG M W, LEE K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding[J/OL]. arXiv preprint arXiv:1810.04805, 2018. https://arxiv.org/abs/1810.04805.

[20] PETERS M E, NEUMANN M, IYYER M, et al. Deep contextualized word representations[C]//Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. : Association for Computational Linguistics, 2018: 2227-2237.

[21] BROWN T B, MANN B, RYDER N, et al. Language models are few-shot learners[C]//Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual. 2020.

[22] LE Q, MIKOLOV T. Distributed representations of sentences and documents [C]//International conference on machine learning. : PMLR, 2014: 1188-1196.

[23] SILVER D, HUANG A, MADDISON C J, et al. Mastering the game of go with deep neural networks and tree search[J]. nature, 2016, 529(7587):484-489.

[24] MNIH V, KAVUKCUOGLU K, SILVER D, et al. Human-level control through deep reinforcement learning[J]. nature, 2015, 518(7540):529-533.

[25] MNIH V, BADIA A P, MIRZA M, et al. Asynchronous methods for deep reinforcement learning[C]//International conference on machine learning. : PMLR, 2016: 1928-1937.

[26] VAN HASSELT H, GUEZ A, SILVER D. Deep reinforcement learning with double q-learning[C]//Proceedings of the AAAI conference on artificial intelligence: volume 30. 2016.

[27] KOMOROWSKI M. Clinical management of sepsis can be improved by artificial intelligence: yes[J]. Intensive care medicine, 2020, 46(2):375-377.

[28] ENGUEHARD J, BUSBRIDGE D, BOZSON A, et al. Neural temporal point processes for modelling electronic health records[C]//Machine Learning for Health. : PMLR, 2020: 85-113.

[29] ERTEKIN Ş, 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.

[30] ZHU S, YUCHI H S, XIE Y. Adversarial anomaly detection for marked spatio-temporal streaming data[C]//ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing. : IEEE, 2020: 8921-8925.

[31] CHEN T, WONG R C. Handling information loss of graph neural networks for session-based recommendation[C]//The 26th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. : ACM, 2020: 1172-1180.

[32] MI F, FALTINGS B. Memory augmented neural model for incremental session-based recommendation[C]//Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence. 2020: 2169-2176.

[33] YU F, ZHU Y, LIU Q, et al. TAGNN: target attentive graph neural networks for session-based recommendation[C]//Proceedings of the 43rd International ACM

SIGIR conference on research and development in Information Retrieval. : ACM, 2020: 1921-1924.

[34] HAWKES A G. Point spectra of some mutually exciting point processes[J]. Journal of the Royal Statistical Society: Series B (Methodological), 1971, 33(3): 438-443.

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

[36] OGATA Y. The asymptotic behaviour of maximum likelihood estimators for stationary point processes[J]. Annals of the Institute of Statistical Mathematics, 1978, 30(2):243-261.

[37] ADAMOPOULOS L. Cluster models for earthquakes: Regional comparisons [J]. Journal of the International Association for Mathematical Geology, 1976, 8 (4):463-475.

[38] OGATA Y. Statistical models for earthquake occurrences and residual analysis for point processes[J]. Journal of the American Statistical association, 1988, 83 (401):9-27.

[39] CLEMENTS R A, SCHOENBERG F P, SCHORLEMMER D. Residual analysis methods for space-time point processes with applications to earthquake forecast models in california[J]. The Annals of applied statistics, 2011:2549-2571.

[40] HASBROUCK J. Measuring the information content of stock trades[J]. The Journal of Finance, 1991, 46(1):179-207.

[41] ENGLE R F, RUSSELL J R. Autoregressive conditional duration: a new model for irregularly spaced transaction data[J]. Econometrica, 1998:1127-1162.

[42] BAUWENS L, HAUTSCH N. Modelling financial high frequency data using point processes[J]. Handbook of financial time series, 2009:953-979.

[43] BACRY E, DELATTRE S, HOFFMANN M, et al. Some limit theorems for hawkes processes and application to financial statistics[J]. Stochastic Processes and their Applications, 2013, 123(7):2475-2499.

[44] RENNER I W, WARTON D I. Equivalence of maxent and poisson point process models for species distribution modeling in ecology[J]. Biometrics, 2013, 69 (1):274-281.

[45] WARTON D I, SHEPHERD L C. Poisson point process models solve the" pseudo-absence problem" for presence-only data in ecology[J]. The Annals of Applied Statistics, 2010:1383-1402.

[46] MEYER S, ELIAS J, HÖHLE M. A space–time conditional intensity model for invasive meningococcal disease occurrence[J]. Biometrics, 2012, 68(2): 607-616.

[47] HÖHLE M. Infectious disease modelling[J]. 2016.

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

[49] DALEY D J, VERE-JONES D. An introduction to the theory of point processes: volume i: elementary theory and methods[M]. : Springer, 2003.

[50] BEGLEITER R, EL-YANIV R, YONA G. On prediction using variable order markov models[J]. Journal of Artificial Intelligence Research, 2004, 22:385-421.

[51] JANSSEN J, LIMNIOS N. Semi-markov models and applications[M]. : Springer Science & Business Media, 2013.

[52] DU N, DAI H, 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. 2016: 1555-1564.

[53] MEI H, EISNER J M. The neural hawkes process: A neurally self-modulating multivariate point process[C]//Advances in neural information processing systems: volume 30. 2017.

[54] XIAO S, YAN J, YANG X, et al. Modeling the intensity function of point process via recurrent neural networks[C]//Proceedings of the AAAI Conference on Artificial Intelligence: volume 31. 2017.

[55] CHOI E, BAHADORI M T, SCHUETZ A, et al. Retain: Interpretable predictive model in healthcare using reverse time attention mechanism[J]. Curran Associates Inc., 2016.

[56] ZHANG Q, LIPANI A, KIRNAP O, et al. Self-attentive hawkes process[C]// International Conference on Machine Learning. : PMLR, 2020: 11183-11193.

[57] ZUO S, JIANG H, LI Z, et al. Transformer hawkes process[C]//International Conference on Machine Learning. : PMLR, 2020: 11692-11702.

[58] MEI H, WAN T, EISNER J. Noise-contrastive estimation for multivariate point processes[C]//Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems. 2020.

[59] OMI T, UEDA N, AIHARA K. Fully neural network based model for general temporal point processes[C]//Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems. 2019: 2120-2129.

[60] GOODFELLOW I, POUGET-ABADIE J, MIRZA M, et al. Generative adversarial nets[C]//Advances in neural information processing systems: volume 27. 2014.

[61] BROCK A, DONAHUE J, SIMONYAN K. Large scale GAN training for high fidelity natural image synthesis[C]//7th International Conference on Learning Representations. : OpenReview.net, 2019.

[62] RADFORD A, METZ L, CHINTALA S. Unsupervised representation learning with deep convolutional generative adversarial networks[C]//4th International Conference on Learning Representations. 2016.

[63] XIAO S, FARAJTABAR M, YE X, et al. Wasserstein learning of deep generative point process models[C]//Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems. 2017: 3247-3257.

[64] WU Q, YANG C, ZHANG H, et al. Adversarial training model unifying feature driven and point process perspectives for event popularity prediction[C]// Proceedings of the 27th ACM International Conference on Information and Knowledge Management. 2018: 517-526.

[65] XIAO S, XU H, YAN J, et al. Learning conditional generative models for temporal point processes[C]//Proceedings of the AAAI Conference on Artificial Intelligence: volume 32. 2018.

[66] LI S, XIAO S, ZHU S, et al. Learning temporal point processes via reinforcement learning[C]//Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada. 2018: 10804-10814.

[67] UPADHYAY U, DE A, RODRIGUEZ M G. Deep reinforcement learning of marked temporal point processes[C]//Advances in Neural Information Processing

Systems 31: Annual Conference on Neural Information Processing Systems. 2018: 3172-3182.

[68] CHANG C H, MAI M, GOLDENBERG A. Dynamic measurement scheduling for event forecasting using deep rl[C]//International Conference on Machine Learning. : PMLR, 2019: 951-960.

[69] COX D R. Some statistical methods connected with series of events[J]. Journal of the Royal Statistical Society: Series B (Methodological), 1955, 17(2):129-157.

[70] LINIGER T J. Multivariate hawkes processes[D]. : ETH Zurich, 2009.

[71] BOWSHER C G. Modelling security market events in continuous time: Intensity based, multivariate point process models[J]. Journal of Econometrics, 2005, 141 (2):876-912.

[72] LUC B, NIKOLAUS H. Modelling financial high frequency data using point processes[J]. Discussion Papers (ECON - Département des Sciences Economiques), 2006.

[73] BRÉMAUD P, MASSOULIÉ L. Hawkes branching point processes without ancestors[J]. Journal of applied probability, 2001, 38(1):122-135.

[74] JAISSON T, ROSENBAUM M. Limit theorems for nearly unstable hawkes processes[J]. The annals of applied probability, 2015, 25(2):600-631.

[75] MASTROMATTEO I, BACRY E, MUZY J F. Linear processes in high dimensions: Phase space and critical properties[J]. Physical Review E, 2015, 91(4): 042142.

[76] SORNETTE D, OUILLON G. Multifractal scaling of thermally activated rupture processes[J]. Physical Review Letters, 2005, 94(3):038501.

[77] BAN Z, ROUEFF F, ABERGEL F. Ergodicity and scaling limit of a constrained multivariate hawkes process[J]. Post-Print, 2014.

[78] ERTEKIN Ş, 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.

[79] BENGIO Y, COURVILLE A, VINCENT P. Representation learning: A review and new perspectives[J]. IEEE transactions on pattern analysis and machine intelligence, 2013, 35(8):1798-1828.

[80] HOCHREITER S, SCHMIDHUBER J. Long short-term memory[J]. Neural computation, 1997, 9(8):1735-1780.

[81] CHO K, VAN MERRIËNBOER B, GULCEHRE C, et al. Learning phrase representations using rnn encoder-decoder for statistical machine translation[J/OL]. arXiv preprint arXiv:1406.1078, 2014. https://arxiv.org/abs/1406.1078.

[82] KOUTNÍK J, GREFF K, GOMEZ F J, et al. A clockwork RNN[C]//JMLR Workshop and Conference Proceedings: volume 32 Proceedings of the 31th International Conference on Machine Learning. : JMLR.org, 2014: 1863-1871.

[83] VAN DEN OORD A, KALCHBRENNER N, KAVUKCUOGLU K. Pixel recurrent neural networks[C]//JMLR Workshop and Conference Proceedings: volume 48 Proceedings of the 33nd International Conference on Machine Learning. : JMLR.org, 2016: 1747-1756.

[84] HU H, QI G. State-frequency memory recurrent neural networks[C]// Proceedings of Machine Learning Research: volume 70 Proceedings of the 34th International Conference on Machine Learning. : PMLR, 2017: 1568-1577.

[85] BAI S, KOLTER J Z, KOLTUN V. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling[J/OL]. arXiv preprint arXiv:1803.01271, 2018. https://arxiv.org/abs/1803.01271.

[86] BAI S, KOLTER J Z, KOLTUN V. Trellis networks for sequence modeling[C]// 7th International Conference on Learning Representations. : OpenReview.net, 2019.

[87] ZHOU H, ZHANG S, PENG J, et al. Informer: Beyond efficient transformer for long sequence time-series forecasting[C]//Thirty-Fifth AAAI Conference on Artificial Intelligence. : AAAI Press, 2021: 11106-11115.

[88] XIAO S, YAN J, FARAJTABAR M, et al. Learning time series associated event sequences with recurrent point process networks[J]. IEEE transactions on neural networks and learning systems, 2019, 30(10):3124-3136.

[89] WANG Y, SHEN H, LIU S, et al. Cascade dynamics modeling with attention-based recurrent neural network.[C]//IJCAI. 2017: 2985-2991.

[90] ZHU S, ZHANG M, DING R, et al. Deep fourier kernel for self-attentive point processes[C]//International Conference on Artificial Intelligence and Statistics. : PMLR, 2021: 856-864.

[91] XU H, FARAJTABAR M, ZHA H. Learning granger causality for hawkes processes[C]//International Conference on Machine Learning. : PMLR, 2016: 1717-1726.

[92] ACHAB M, BACRY E, GAIFFAS S, et al. Uncovering causality from multivariate hawkes integrated cumulants[C]//International Conference on Machine Learning. : PMLR, 2017: 1-10.

[93] EICHLER M, DAHLHAUS R, DUECK J. Graphical modeling for multivariate hawkes processes with nonparametric link functions[J]. Journal of Time Series Analysis, 2017, 38(2):225-242.

[94] WU W, LIU H, ZHANG X, et al. Modeling event propagation via graph biased temporal point process[J]. IEEE Transactions on Neural Networks and Learning Systems, 2020.

[95] SHANG J, SUN M. Geometric hawkes processes with graph convolutional recurrent neural networks[C]//Proceedings of the AAAI Conference on Artificial Intelligence: volume 33. 2019: 4878-4885.

[96] XUE S, SHI X, HAO H, et al. A graph regularized point process model for event propagation sequence[C]//2021 International Joint Conference on Neural Networks (IJCNN). : IEEE, 2021: 1-7.

[97] KIPF T N, WELLING M. Semi-supervised classification with graph convolutional networks[C]//5th International Conference on Learning Representations. : OpenReview.net, 2017.

[98] MA Y, WANG S, AGGARWAL C C, et al. Multi-dimensional graph convolutional networks[C]//Proceedings of the 2019 SIAM International Conference on Data Mining. : SIAM, 2019: 657-665.

[99] KHAN M R, BLUMENSTOCK J E. Multi-gcn: Graph convolutional networks for multi-view networks, with applications to global poverty[C]//Proceedings of the AAAI Conference on Artificial Intelligence: volume 33. 2019: 606-613.

[100] JANG E, GU S, POOLE B. Categorical reparameterization with gumbel-softmax [C]//5th International Conference on Learning Representations. : OpenReview.net, 2017.

[101] LI L, ZHA H. Dyadic event attribution in social networks with mixtures of hawkes processes[C]//Proceedings of the 22nd ACM international conference on Information & Knowledge Management. 2013: 1667-1672.

[102] WU W, LIU H, ZHANG X, et al. Modeling event propagation via graph biased temporal point process[J]. IEEE Transactions on Neural Networks and Learning Systems, 2020.

[103] LUO D, XU H, ZHA H, et al. You are what you watch and when you watch: Inferring household structures from iptv viewing data[J]. IEEE Transactions on Broadcasting, 2014, 60(1):61-72.

[104] CHENG Z, CAVERLEE J, KAMATH K Y, et al. Toward traffic-driven location-based web search[C]//Proceedings of the 20th ACM international conference on Information and knowledge management. 2011: 805-814.

[105] KINGMA D P, BA J. Adam: A method for stochastic optimization[C]//3rd International Conference on Learning Representations. 2015.

[106] ZHANG D, YIN J, ZHU X, et al. Network representation learning: A survey[J]. IEEE transactions on Big Data, 2018, 6(1):3-28.

[107] LIN T Y, GOYAL P, GIRSHICK R, et al. Focal loss for dense object detection [C]//Proceedings of the IEEE international conference on computer vision. 2017: 2980-2988.

[108] KANG B, XIE S, ROHRBACH M, et al. Decoupling representation and classifier for long-tailed recognition[C]//8th International Conference on Learning Representations. : OpenReview.net, 2020.

[109] SZEGEDY C, VANHOUCKE V, IOFFE S, et al. Rethinking the inception architecture for computer vision[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 2818-2826.

[110] ZHANG C, JIANG P, HOU Q, et al. Delving deep into label smoothing[J]. IEEE Trans. Image Process., 2021, 30:5984-5996.

[111] ZAGORUYKO S, KOMODAKIS N. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer[C]// 5th International Conference on Learning Representations. : OpenReview.net, 2017.

[112] CUI Y, JIA M, LIN T, et al. Class-balanced loss based on effective number of samples[C]//IEEE Conference on Computer Vision and Pattern Recognition. : Computer Vision Foundation / IEEE, 2019: 9268-9277.

[113] ZHAO Q, ERDOGDU M A, HE H Y, et al. SEISMIC: A self-exciting point process model for predicting tweet popularity[C]//Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. : ACM, 2015: 1513-1522.

[114] ZHOU G, ZHU X, SONG C, et al. Deep interest network for click-through rate prediction[C]//Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. : ACM, 2018: 1059-1068.

简历与科研成果

基本信息

王言，男，汉族，1997 年 6 月出生，河北省邢台人。

教育背景

2019 年 9 月 — 2022 年 6 月 南京大学人工智能学院 硕士

2015 年 9 月 — 2019 年 6 月 西北工业大学计算机学院 本科

攻读硕士学位期间完成的学术成果

1. Hongyan Hao, Yan Wang, Jian Zhao, Furao Shen, "Temporal Convolutional Attention based Network For Sequence Modeling" in *arXiv preprint arXiv*: 2002.12530, 2020.
2. 葛轶洲，刘恒，王言，徐百乐，周青，申富饶. 小样本困境下的深度学习图像识别综述. 软件学报. 21;33(1):193-210. Apr 2021.

攻读硕士学位期间完成的专利成果

1. 申富饶，王言，赵健. 一种基于时序卷积和关系建模的事件序列预测方法。专利申请号：202210305672.8
2. 周青，王言，葛轶洲，徐百乐，张歆，申富饶. 一种基于多尺度信息融合的增量式水声信号识别方法。专利申请号：202010673065.8

攻读硕士学位期间参与的科研课题

1. 国家自然科学基金 ‘‘基于深度感知增量式联想记忆神经网络的信息融合系统研究’’（课题年限 2019 年 1 月—2022 年 12 月），负责时间序列预测相关问题的研究。

学位论文出版授权书

本人完全同意《中国优秀博硕士学位论文全文数据库出版章程》（以下简称“章程”），愿意将本人的学位论文提交“中国学术期刊（光盘版）电子杂志社”在《中国博士学位论文全文数据库》、《中国优秀硕士学位论文全文数据库》中全文发表。《中国博士学位论文全文数据库》、《中国优秀硕士学位论文全文数据库》可以以电子、网络及其他数字媒体形式公开出版，并同意编入《中国知识资源总库》，在《中国博硕士学位论文评价数据库》中使用和在互联网上传播，同意按“章程”规定享受相关权益。

作者签名：_____

_____年_____月_____日

论文题名	基于深度学习与点过程的事件序列预测算法研究				
研究生学号	MG1937025	所在院系	人工智能学院	学位年度	2019
论文级别	☑ 硕士　☐ 硕士专业学位 ☐ 博士　☐ 博士专业学位　（请在方框内画勾）				
作者 Email	yanwang@smail.nju.edu.cn				
导师姓名	申富饶 教授				

论文涉密情况：

☐ 不保密

☑ 保密，保密期：_____年_____月_____日 至 _____年_____月_____日

注：请将该授权书填写后装订在学位论文最后一页（南大封面）。