

面向强稀疏性移动社交网络的链路预测深度学习方法

何亚迪, 刘林峰

(南京邮电大学计算机学院, 江苏 南京 210023)

摘要: 链路预测是利用深度学习技术分析网络数据, 挖掘网络中潜在的节点关系, 通常应用于网络安全、信息挖掘等领域。通过预测网络中节点间的链路, 可以识别社交工程攻击、欺诈行为和隐私泄露风险。但移动社交网络的拓扑结构随时间变化, 链路稀疏, 影响预测准确性。为了解决移动社交网络中链路预测的强稀疏性问题, 提出基于深度学习的预测方法, 即面向强稀疏性移动社交网络的链路预测深度学习方法 (deep learning-based method for mobile social networks with strong sparsity for link prediction, DLMSS-LP)。该方法综合运用了图自编码器 (graph auto-encoder, GAE)、特征矩阵聚合技术以及多层长短期记忆网络 (long short-term memory, LSTM), 旨在降低了模型的学习成本, 更有效地处理高维和非线性的网络结构, 并且捕捉移动社交网络中的时序动态变化, 进而增强模型对现有链路生成可能性的预测能力。对比其他方法在 AUC (area under curve) 和 ER (error rate) 指标上有明显提升, 体现了模型对不确定链路预测的高准确率和强鲁棒性。

关键词: 链路预测; 移动社交网络; 强稀疏性; 深度学习

文献标志码: A

DOI: 10.11959/j.issn.2096-109x.2024044

Deep learning-based method for mobile social networks with strong sparsity for link prediction

HE Yadi, LIU Linfeng

School of Computer Science, Nanjing University of Posts and Telecommunications, Nanjing 210023, China

Abstract: Link prediction, the process of uncovering potential relationships between nodes in a network through the use of deep learning techniques, is commonly applied in fields such as network security and information mining. It has been utilized to identify social engineering attacks, fraudulent activities, and privacy breach risks by predicting links between nodes within a network. However, the topology of mobile social networks is subject to change over time, and the sparsity of links affects the accuracy of predictions. To address the issue of strong sparsity in link prediction for mobile social networks, a deep learning-based prediction method named DLMSSLP (deep learning-

收稿日期: 2024-01-21; 修回日期: 2024-04-09

通信作者: 刘林峰, liulf@njupt.edu.cn

基金项目: 国家自然科学基金 (62272237)

Foundation Item: The National Natural Science Foundation of China (62272237)

引用格式: 何亚迪, 刘林峰. 面向强稀疏性移动社交网络的链路预测深度学习方法[J]. 网络与信息安全学报, 2024, 10(3): 117-129.

Citation Format: HE Y D, LIU L F. Deep learning-based method for mobile social networks with strong sparsity for link prediction[J]. Chinese Journal of Network and Information Security, 2024, 10(3): 117-129.

based method for mobile social networks with strong sparsity for link prediction) was developed. This method was designed to employ a combination of a Graph Auto-Encoder (GAE), feature matrix aggregation, and multi-layer long short-term memory networks (LSTM). It aimed to reduce the learning cost of the model, process high-dimensional and nonlinear network structures more effectively, and capture the temporal dynamics within mobile social networks, thereby enhancing the model's predictive capability for the generation of existing links. When compared to other methods, DLMSSLP demonstrated significant improvements in AUC and ER metrics, showcasing the model's high accuracy and robustness in predicting uncertain links.

Keywords: link prediction, mobile social networks, strong sparsity, deep learning

0 引言

随着网络信息技术的发展, 新型网络如互联网、物联网和移动社交网络 (mobile social network, MSN)^[1]的出现。这些网络可以分为静态网络和动态网络两类。在现实世界中, 绝大多数网络是动态的, 意味着网络中的节点和链路会随时间的推移而出现或消失。在动态网络中分析未来会出现的链路显得尤为重要, 因为链路能够揭示节点之间的潜在时空关系, 关系的识别对于网络信息安全具有重要意义。异常检测^[2]: 动态网络中未来会出现链路的预测可以帮助识别异常行为或潜在的安全威胁。例如, 在移动社交网络中, 异常的连接模式可能存在恶意活动, 如账户被盗用或社交工程攻击。安全防护策略的优化^[3]: 通过分析动态网络的潜在链路, 可以更好地理解网络的演变趋势和潜在风险点, 从而设计出更有效的安全防护策略。数据泄露风险评估^[4]: 在物联网和企业网络中, 链路预测可以揭示设备或系统间不寻常的交互, 有助于提前识别数据泄露的风险。另外链路预测还有入侵检测^[5]、网络流量分析^[6]和社交工程攻击预防^[7]等应用。

动态网络中的链路预测在网络信息安全领域具有广泛的应用前景, 可以显著提升对网络行为的洞察力和对安全威胁的响应能力。

对于链路预测^[8], 如何提高预测精度仍然是主要关注的问题。预测精度较高的链路预测方法, 可以更好地分析动态网络的变化, 提取出更加难以察觉的信息。例如, 在分子网络^[9]中, 酵母和蛋白质之间 80% 的相互作用关系还很难确定, 如果能采用适当的链路预测方法来预测相互效应, 则可以极大地降低在分子网络上进行生物实验的成本。

单分散神经网络是典型的动态网络, 其中节点之间的关系代表了人类、组织、城市以及国家之间的通信。由于这些关系的时变特征, MSN 是高度动态和复杂的。在 MSN 中, 链路预测技术可以预测人与人之间的关系或手持设备间的通信关系。具体来说, 这些关系揭示了人的社交角色、习惯、爱好和社交倾向, 通过这些关系可以提供便利和个性化的服务。例如, 通过分析高峰时段拥挤路段的未来人口密度 (由人与人之间预测的关系获得), 出租车司机可以更容易地接到乘客; 在社交媒体上, 通过匹配个性或爱好, 可以更容易地向人推荐潜在的朋友; 在线购物平台上, 通过探索客户的浏览历史, 可以更准确地推荐商品。总之, 对 MSN 进行准确的链路预测可以提高社会资源的消费效率, 方便人们生活。

由于网络中节点和链路发生变化, 动态网络中的链路预测问题比静态网络中的问题要复杂得多。在动态网络 (如 MSN) 中, 需要提取时变特征来预测未来会出现的链路, MSN 中的链路预测如图 1 所示。

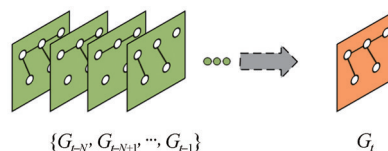


图1 MSN中的链路预测
Figure 1 Link prediction in MSN

近年来, 一些深度神经网络被应用于预测动态网络未来会出现的链路。例如, 递归神经网络 (recurrent neural network, RNN)^[10]和一些它的变体已经被用来学习网络的时序序列, 并从大量的网络中提取出时序特征。RNN 通过不断学习历史网络拓扑结构, 大大提高了预测精度。由于以高维网络和非线性网络为输入, 网络的数据量

越大,可以得到的预测结果越准确。然而,RNN通常会将网络的时序序列按时间顺序排列。随着网络的增加,RNN及其变体也面临着梯度爆炸和消失的问题^[11]。

MSN的突出特点是强稀疏性^[12],即现有链路数量比节点少得多。节点越多的MSN,MSN稀疏性越强。然而,传统的深度学习模型不能很好地处理MSN强稀疏性,因为有以下两个事实。

1) 由于强稀疏性,模型训练效果不理想,使深度学习模型难以收敛,特别是在训练样本数量非常小的情况下。

2) 以one-hot形式编码后的输入特征维度更高,因此需要的深度学习模型更大。

基于上述考虑,提出面向强稀疏性移动社交网络的链路预测尝试学习方法(deep learning-based method for mobile social networks with strong sparsity for link prediction, DLMSS-LP)。基于相遇矩阵和度矩阵,DLMSS-LP可以有效地处理高维和非线性网络。DLMSS-LP是由图自编码器(graph auto-encoder, GAE)^[13]、特征矩阵聚合和多层长短期记忆(long short-term memory, LSTM)^[14]组成的。GAE利用相遇矩阵作为节点的特征矩阵,降低了模型学习的成本,提高了现有链路在预测中的生成可能性。图卷积网络(graph convolutional network, GCN)^[15]的度矩阵和输出矩阵的融合可以消除MSN强稀疏性对深度神经网络的负面影响。此外,多层LSTM可以学习MSN时序序列的变化,从而提高预测精度。

1 相关工作

链路预测的目标是根据历史链路(或历史网络拓扑结构)推断出最有可能出现或消失的链路。在早期,链路预测技术侧重应用于静态网络或拓扑结构变化较少的网络,并根据节点相似性预测未来会出现的链路。目前,随着深度学习模型的发展,一些深度学习模型已被应用于动态网络的链路预测。

1.1 静态网络的链路预测方法

Zhang等^[16]探讨了用于链接预测的启发式学习范式,并且提出了新颖的递减启发式理论。该理论在单一框架内统一了广泛的启发式方法,并

证明所有这些启发式都可以从局部子网络中很好地近似。实验结果表明,局部子网络保留了与链接存在相关的丰富信息。基于递减理论,使用图神经网络(graph neural network, GNN)从局部子网络中学习启发式的新方法。这种方法具有很好的泛化性,能够在其他类似的问题上表现良好。

在社交网络分析中,通常链路预测是根据节点相似性来预测节点之间未来可能出现的链路,而传统的相似度算法将相似度视为标量值,导致一些信息丢失。Su等^[17]提出新颖的节点相似度的计算方法,该方法将节点之间的信息投射到多个维度上来测量节点相似度。定义全局相似度、局部相似度和元相似度,来综合计算节点相似性,并测量了不同场景中节点相似性计算结果的差异。与一些先进的方法相比,该相似度方法提高了推荐系统中的预测准确性。

此外,Rafiee等^[18]提出了CNDP(common neighbors degree penalization),其节点相似性根据网络的拓扑特性和节点的特征来确定。在CNDP中,引入了新的链路预测度量,将网络的聚类系数(clustering coefficient)作为网络特征。此外,除每对节点公共的邻接节点之外,CNDP还考虑了邻接节点的邻接节点,这使CNDP比其他类似的链接预测方法取得了更好的预测效果。在现实世界的网络上的评估实验结果表明,CNDP预测结果更准确,CNDP的复杂度也更低。

这些方法大多在静态网络或网络拓扑结构变化很小的网络中表现良好,当网络的拓扑结构随时间变化很大或频繁变化时,链路预测的准确性就会出现大幅度地下降。

1.2 动态网络的链路预测方法

基于深度学习的动态网络的链路预测方法通常是通过建立一些深度学习模型来学习网络时序序列的演化。例如,Li等^[19]提出了统一框架来通过对节点相似性进行编码并计算邻接节点之间的相似度作为链路预测的特征,称为网络嵌入的判别性距离度量学习(discriminative distance metric learning for network embedding, DDNE)。DDNE通过学习网络的非线性变换,将网络中的节点相似度一致的节点对投影到相同的低维空间中。

DDNE 在 3 个社交网络分析任务（网络重构、链路预测和推荐系统）上对 4 个真实世界数据集进行了广泛的实验，其准确度与基准方法的准确度相比具有一定的竞争性和优越性。Qu 等^[20]将 GNN 和时序神经网络结合起来，提出了时间依赖图神经网络（temporal dependent graph neural network, TDGNN）。TDGNN 是简单而有效的动态网络，表示深度学习框架，将网络时间信息融入 GNN 中，并引入新的时间聚合器（temporal aggregator, TDAgg）来聚合邻接节点的特征和边的时间信息，以获得目标节点的特征。TDGNN 在 6 个真实世界动态网络数据集上进行的连续时间链路预测取得了很好的效果。

此外，Chen 等^[21]受 LSTM 在处理时间序列方面出色表现的启发，提出了新的编码器—长短期记忆网络—解码器（encoder-LSTM-decoder, E-LSTM-D）深度学习模型，用于动态链路预测。该模型能够处理长时序的链路预测问题，并适用于不同规模的网络。E-LSTM-D 能够在统一的框架内自动学习网络结构的时间特征，预测网络中之前从未出现过的链路。

对于基于深度学习模型的链路预测方法，需要使用大量的网络时序序列进行模型训练，然后应用训练后的模型进行链路预测。然而，MSN 的强稀疏性严重影响了传统深度学习模型的训练效果，导致预测精度下降，这是本文主要关注的问题。

2 深度学习预测方法

本节首先提供 MSN 中链路预测的相关定义，然后介绍本文提出的 DLMSS-LP 的细节，最后分析 DLMSS-LP 的抗噪声性和空间复杂性。

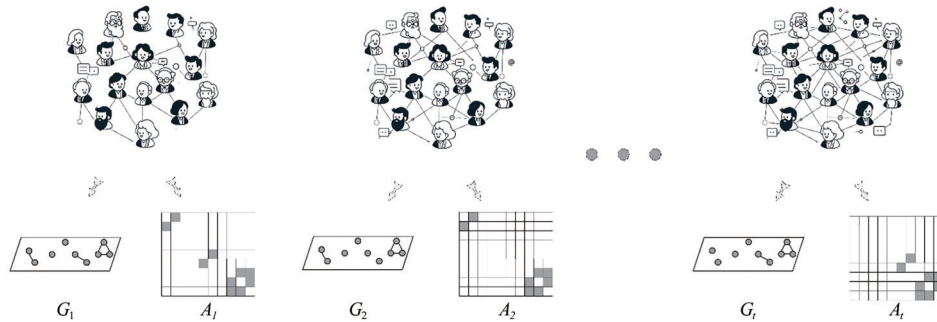


图 2 会议人员之间的链路映射到邻接矩阵
Figure 2 Mapping meeting members and their links into adjacency matrices

2.1 问题定义

2.1.1 移动社交网络

本文使用邻接矩阵的形式来表示 MSN 的快照。将会议人员之间的链路映射到邻接矩阵中，如图 2 所示，会议的成员是节点，成员之间的交流构成链路。会议在每一个时刻的快照可以用一个邻接矩阵表示。一组 MSN 时序序列由一组快照（图） $\{G_1, \dots, G_k, \dots, G_t\}$ 表示。其中 $G_k = (V, E_k)$ 表示第 k 张快照， V 表示节点集合， $E_k \subseteq V \times V$ 表示第 k 张快照中的链路集合。 A_k 表示快照 G_k 的邻接矩阵。如果在第 k 张快照中节点 v_i 和节点 v_j 之间存在链路，则为 $a_{k,i,j} = 1$ ；否则为 $a_{k,i,j} = 0$ 。

2.1.2 度矩阵序列

度矩阵是对角矩阵，其中对角线上的元素表示节点的度。具体地说，在无向图中节点 v_i 的度等于连接到的链路数；在有向图中，节点 v_i 的度等于入度与出度的和。提供度矩阵可以帮助深度学习模型学习网络时序数据。然而，单个度矩阵不能反映 MSN 的时变特征，如果使用度矩阵 D_{t-1} 来预测度矩阵 D_t （在当前快照中），深度学习模型无法完全学习网络拓扑结构的时序变化。因此将度矩阵序列 $\{D_1, \dots, D_{t-1}\}$ 作为深度学习模型的特征来预测未来会出现链路。

度矩阵序列的构造如图 3 所示。

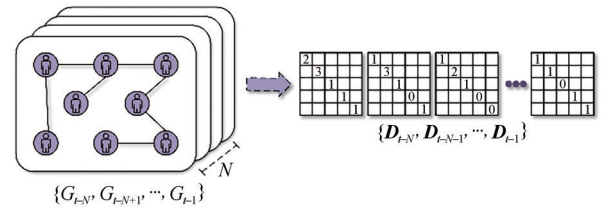


图 3 度矩阵序列的构造
Figure 3 Construction of a degree matrix sequence

2.1.3 强稀疏性

G_k 的稀疏度的计算表达式如下。

$$\text{Sparsity}(A_k) = \frac{\sqrt{n^2 - \sum |x_i|}}{\sqrt{(\sum x_i^2) \cdot n^2 - 1}} \quad (1)$$

其中, n 表示 MSN 中的节点数, x_i 表示第 i 个邻接矩阵 A_k 中的元素。稀疏度的值在区间 $[0, 1]$, 并且快照中的链路越少则稀疏度越大, 当稀疏度的值大于 0.9 时, 则定义该快照有强稀疏性。

2.1.4 链路预测方法

邻接矩阵序列为 $A = \{A_{t-N}, A_{t-N+1}, \dots, A_{t-1}\}$ 。考虑 MSN 的时变特性, 深度学习模型首先通过输入邻接矩阵序列 A 来学习 MSN 的网络拓扑结构的变化, 然后预测第 t 次快照中的邻接矩阵 A_t 。

将存在的链路称为正样本, 不存在的链路称为负样本。由于 MSN 的强稀疏性, 正样本通常很少, 这大幅度降低了链路预测的准确性。本文提出的 DLMSS-LP 可以平衡正样本和负样本的数量。DLMSS-LP 认为最近出现过的链路在不久的将来还出现的可能性会更大。基于这个前提, 引入相遇矩阵, 每个元素在相遇矩阵中表示两个节点之间链路出现的次数。此外, DLMSS-LP 将度矩阵和编码器的输出矩阵融合以增强未来链路出现的可能性, 通过引入相遇矩阵和度矩阵, 并调整中正样本和负样本的比例, 加强正样本对模型的训练的影响, 使 DLMSS-LP 的输出 (链路预测结果) 倾向于生成正样本 (链路), 从而提高链路预测的预测精度。

2.2 DLMSS-LP 模型

强稀疏性可能导致在模型训练过程中正样本和负样本之间的不平衡, 严重影响了预测准确

性。为此, 特别针对具有强稀疏性的 MSN 设计深度学习模型 DLMSS-LP, 如图 4 所示。DLMSS-LP 主要由 GAE 模块、矩阵聚合模块和 LSTM 模块组成。此外, 将相遇矩阵作为特征矩阵, 并采用随机负采样的方法生成不存在的链接 (负样本), 以此来平衡实际存在的链接 (正样本)。这种方法主要通过 3 种方式缓解数据不平衡的影响。首先, 通过增加负样本数量来平衡正负样本的比例, 避免学习算法过度偏向于多数类; 其次, 提高计算效率, 因为在极端稀疏的网络中使用所有可能的负样本进行训练是不现实的; 最后, 随机负采样通过引入更广泛的负面情况来增加模型的泛化能力, 从而提高未来链路生成的可能性, 减少训练时间。

GAE 由编码器和解码器构成, 分别负责提取相遇矩阵的特征, 以及输出预测结果的邻接矩阵。LSTM 模块位于编码器和解码器之间以提取编码器输出的时变特征。

2.2.1 编码器

编码器的主要功能是提取相遇矩阵的特征并降低特征的维度, 从而减少 LSTM 模块学习时序特征所需要的时间。然而, 编码器并不能很好地平衡正样本和负样本之间数量, 降低强稀疏性对预测精度的影响。所以 DLMSS-LP 通过将相遇矩阵作为编码器的输入, 使链路 (未来可能存在的) 在预测结果的邻接矩阵中对应的值接近于 1, 来尽可能地降低强稀疏性对预测精度的影响。

GAE 使用 GCN 作为编码器, GCN 的架构如图 5 所示。编码器的执行步骤分别表示为式 (2)、式 (3) 和式 (4)。

$$\tilde{A}_t = A_t + I_N \quad (2)$$

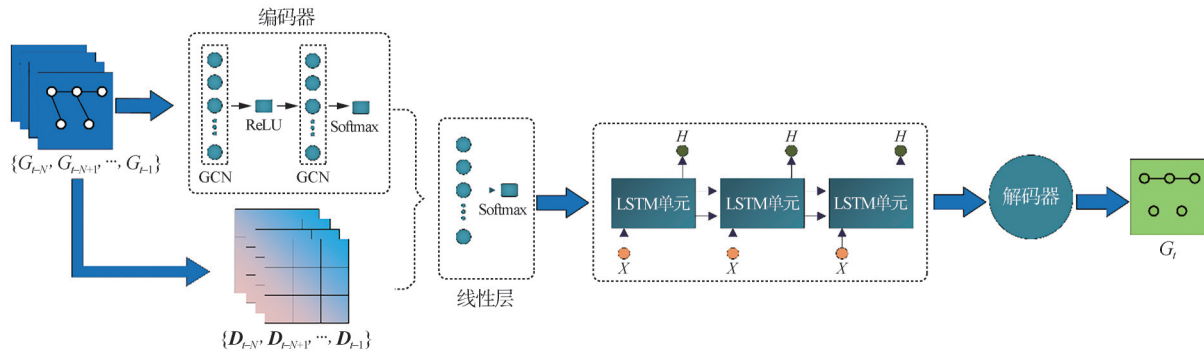


图4 DLMSS-LP 架构
Figure 4 Architecture of DLMSS-LP

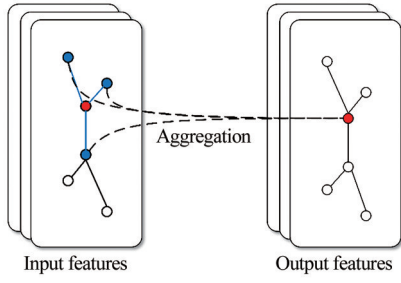


图 5 GCN 架构
Figure 5 Architecture of GCN

其中, A_t 表示第 t 张快照中的邻接矩阵, I_N 表示维数为 $N \times N$ 的单位矩阵, $\tilde{A}_t$ 表示 A_t 的过渡形式。

$$\dot{A}_t = \tilde{D}_t^{-\frac{1}{2}} \cdot \tilde{A}_t \cdot \tilde{D}_t^{-\frac{1}{2}} \quad (3)$$

其中, $\tilde{D}_t$ 表示 $\dot{A}_t$ 的度矩阵。和 $\dot{A}_t$ 是 $\tilde{A}_t$ 的过渡形式。

$$Z_t = \text{Softmax}(\dot{A}_t \cdot \text{ReLU}(\tilde{A}_t \cdot W^{(0)}) \cdot W^{(1)}) \quad (4)$$

其中, X_t 表示第 t 张快照的特征矩阵, $W^{(0)}$ 表示第 1 层 GCN 的权重, $W^{(1)}$ 表示第 2 层 GCN 的权重, Z_t 表示 GCN 的输出。DLMSS-LP 选择 ReLU 和 Softmax 作为第 1 层 GCN 的激活函数与第 2 层 GCN 的激活函数。

2.2.2 聚合模块

聚合模块使用一个线性层和 Softmax 函数对编码器输出的度矩阵和特征矩阵进行聚合:

$$\hat{Z}_t = \text{Softmax}(\hat{W}_t \cdot (D_t \cdot Z_t \cdot D_t^{-1}) + \hat{b}_t) \quad (5)$$

其中, $\hat{W}_t$ 和 $\hat{b}_t$ 表示的线性层的权重和偏置。 D_t 表示第 t 张快照的度矩阵, D_t^{-1} 表示 D_t 的逆矩阵, Z_t 表示第 t 张快照的编码器输出, 并且 $\hat{Z}_t$ 表示聚合后的特征矩阵。此外将 Softmax 函数作为激活函数处理 $\hat{Z}_t$ 。

2.2.3 LSTM 模块

LSTM 模块是 RNN 类型, 用于解决普通

RNN 中的消失梯度问题。LSTM 的架构如图 6 所示。LSTM 单元由遗忘门、输入门和输出门组成。其中, 遗忘门是基于 h_{t-2} (前 $t-2$ 张快照序列的隐藏状态) 和 x_{t-1} (第 $t-1$ 张快照的输入)。在图 6 中, f_{t-1} 表示第 $t-1$ 张快照的遗忘门输出, i_{t-1} 表示第 $t-1$ 张快照的输入门输出, o_{t-1} 表示第 $t-1$ 张快照的输出门输出, $\tilde{C}_{t-1}$ 表示前 $t-1$ 张快照序列的 LSTM 单元状态, y_{t-1} 表示前 $t-1$ 张快照序列的 LSTM 单元输出。 $\{x_{t-N}, x_{t-N+1}, \dots, x_{t-1}\}$ 表示 $\{G_{t-N}, G_{t-N+1}, \dots, G_{t-1}\}$ 经过降维处理后的快照序列。

遗忘门决定了哪些信息应该从 LSTM 单元状态中丢弃。输入门决定了哪些信息应该存储在 LSTM 单元状态, 将从输入门保留的 f_{t-1} 添加到 LSTM 单元状态 C_{t-1} 中。输出门根据 LSTM 单元状态决定哪些信息应该输出, 并且根据 C_{t-1} 和 h_{t-2} 更新 h_{t-1} 。

遗忘门将 h_{t-2} 与 x_{t-1} 相结合起来, 并输出向量 f_{t-1} , 其中 f_{t-1} 值在区间 $[0, 1]$ 内。当 f_{t-1} 某个元素值等于 1 时, 将 C_{t-2} 对应位置上的元素输入到 C_{t-1} 中。相反, 当 f_{t-1} 对应某个元素值等于 0 时, 则表示应该丢弃在 C_{t-2} 对应位置上的元素。此过程可表示为:

$$f_{t-1} = \sigma(W_f \cdot [h_{t-2}, x_{t-1}] + b_f) \quad (6)$$

其中, h_{t-1} 表示前 $t-1$ 张快照序列的隐藏状态, W_f 为遗忘门的权值, b_f 为遗忘门的偏置, σ 为遗忘门的激活函数 Sigmoid。

之后, x_{t-1} 和 h_{t-2} 被转移到输入门中。输入门会过滤来自 x_{t-1} 和 h_{t-2} 的信息, 过滤后的信息用于更新 C_{t-1} 。此过程可表示为:

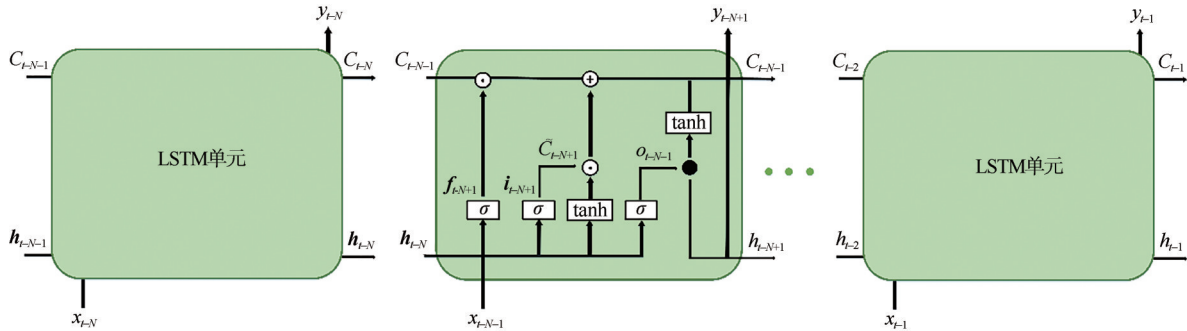


图 6 LSTM 架构
Figure 6 Architecture of LSTM

$$\mathbf{i}_{t-1} = \sigma(\mathbf{W}_i \cdot [\mathbf{h}_{t-2}, \mathbf{x}_{t-1}] + \mathbf{b}_i) \quad (7)$$

$$\tilde{\mathbf{C}}_t = \tanh(\mathbf{W}_c \cdot [\mathbf{h}_{t-2}, \mathbf{x}_{t-1}] + \mathbf{b}_c) \quad (8)$$

其中, $\mathbf{i}_{t-1}$ 表示第 $t-1$ 张快照中输入门的输出, 而 $\tilde{\mathbf{C}}_{t-1}$ 是基于 $\mathbf{x}_{t-1}$ 和 $\mathbf{h}_{t-2}$ 得到的。当 $\mathbf{i}_{t-1}$ 某个元素值等于 1 时, 将 $\tilde{\mathbf{C}}_{t-1}$ 对应位置上的元素输入 $\mathbf{C}_{t-1}$ 中。 $\mathbf{W}_i$ 和 $\mathbf{W}_c$ 分别是 $\mathbf{i}_{t-1}$ 和 $\tilde{\mathbf{C}}_{t-1}$ 对应的权重。同样, $\mathbf{b}_i$ 和 $\mathbf{b}_c$ 分别是与 $\mathbf{i}_{t-1}$ 和 $\tilde{\mathbf{C}}_{t-1}$ 对应的偏置。

然后, 遗忘门和输入门根据 $\tilde{\mathbf{C}}_{t-1}$ 和 $\mathbf{C}_{t-2}$ 更新单元状态 $\mathbf{C}_{t-1}$:

$$\mathbf{C}_{t-1} = \mathbf{f}_{t-1} \cdot \mathbf{C}_{t-2} + \mathbf{i}_{t-1} \cdot \tilde{\mathbf{C}}_{t-1} \quad (9)$$

最后, 基于 $\mathbf{C}_{t-1}$ 和 $\mathbf{o}_{t-1}$ 计算隐藏状态 $\mathbf{h}_{t-1}$:

$$\mathbf{h}_{t-1} = \mathbf{o}_{t-1} \cdot \tanh(\mathbf{C}_{t-1}) \quad (10)$$

其中, $\mathbf{o}_{t-1}$ 表示中第 $t-1$ 张快照的输出门的输出, $\mathbf{o}_{t-1} = \sigma(\mathbf{W}_o \cdot [\mathbf{h}_{t-2}, \mathbf{x}_{t-1}] + \mathbf{b}_o)$, 其中 $\mathbf{W}_o$ 和 $\mathbf{b}_o$ 表示输出门的权重和偏置。而门控循环单元 (GRU, gated recurrent unit) [22] 简化了 LSTM 的模型结构, 它只有两个门 (更新门和重置门), 没有独立的细胞状态, 并且它将遗忘门和输入门合并为一个单一的更新门。所以 LSTM 由于其更复杂的内部结构, 理论上能够捕捉更长的序列依赖关系。

2.2.4 隐藏层单元数

隐藏层单元数直接影响深度学习模型的学习能力。如果单元数量太少, 深度学习模型的学习能力就会不足; 如果单元数量太大, 深度学习模型的复杂性就会大幅增加, 从而导致过拟合问题和不必要的计算。因此, 适当的隐藏层单元数量可以提高预测精度和训练效率。假设隐藏层的维数为 $p \times q$, 隐藏层单元数可由式 (11) 计算。

$$N_{\text{cell}} = \sqrt{p \cdot q + \alpha} \quad (11)$$

其中, α 是整数, 并且值在区间 $[0, 10]$ 。

2.2.5 解码器

解码器的功能是处理 LSTM 模块输出的特征矩阵, 然后得到最终预测网络的邻接矩阵 (用 $\hat{\mathbf{A}}_t$ 表示):

$$\hat{\mathbf{A}}_t = \mathbf{o}_t \cdot \mathbf{o}_t^T \quad (12)$$

其中, $\mathbf{o}_t$ 表示中第 t 张快照的 LSTM 模块输出, $\mathbf{o}_t^T$ 表示 $\mathbf{o}_t$ 的转置。

2.2.6 梯度算法

梯度算法决定了影响 DLMSS-LP 的收敛速

度。DLMSS-LP 使用 Adadelta 优化器 [23], 与随机梯度下降 (stochastic gradient descent, SGD) 相比, 随机梯度递减 Adadelta 优化器不需要手动调整学习速率, 能自动根据噪声、模型架构以及参数的变化进行调整。

2.2.7 损失函数

DLMSS-LP 取均方误差 (mean squared error, MSE) 作为损失函数, 其表达式为:

$$\text{MES} = \frac{1}{n} \sum_{i=1}^t |\mathbf{y}_{\text{true}, i} - \mathbf{y}_{\text{pred}, i}|^2 \quad (13)$$

其中, $\mathbf{y}_{\text{true}, i}$ 和 $\mathbf{y}_{\text{pred}, i}$ 分别表示与第 t 张快照的标签值和对应的 DLMSS-LP 输出值。

2.3 训练过程

邻接矩阵 $\mathbf{A}_t$ 的元素值为 0 或 1, 因此需要对 $\mathbf{A}_t$ 的元素处理。如果在 $a_{t, i, j} > 0.5$, 设定 $a_{t, i, j} = 1$, 表明在节点 v_i 和节点 v_j 之间存在链路, 否则 $a_{t, i, j} = 0$ 。当损失函数的值非常小, 并且在连续的几个训练时期中保持不变时, 则认为模型已经收敛。

2.4 抗噪性分析

噪声是由数据冗余、数据异常和不确定的链路引起的。数据冗余和数据异常易于消除, 以降低噪声。不确定的链路指的是过去存在的、将来会不存在的链路。深度学习模型不能完全学习不确定的链路, 因此很难预测不确定的链路。为了保持数据集的完整性, 本文不从数据集中排除不确定的链路, 从而尽可能地保持数据集的完整性和真实性。

2.5 空间复杂度

2.5.1 GAE 模块的空间复杂度

GAE 模块包括编码器和解码器。编码器处理相遇矩阵。编码器由两层 GCN 构成, 并且每个 GCN 层由 128 个细胞组成。解码器对 LSTM 的输出矩阵进行处理, 因为其没有学习能力, 所以不消耗任何额外的空间。因为相遇矩阵的维数是 $n \times n$, GAE 的空间复杂度为 $O(128n)$ 。

2.5.2 特征矩阵融合模块的空间复杂度

特征矩阵聚合模块主要由线性层组成。假设 GCN 的输出矩阵的维度是 $n \times m$, 其中 m 表示特征的尺寸。因此, 特征矩阵聚合模块的空间复杂性特征矩阵聚合模块的复杂度为 $O(n \cdot m)$ 。

2.5.3 LSTM的空间复杂度

LSTM主要包括3部分,即遗忘门、输入门和输出门。因为输入门中有两个单元格门,所以输入门(或输出门)的权重数是遗忘门的两倍,并且DLMSS-LP中有3个LSTM层。因此,LSTM参数的总数写成: $12 \cdot \text{hidden_size} \cdot (\text{input_size} + \text{bias} + \text{output_size})$ 。

由于LSTM模块的输出 y 等于 h_t 。所以隐含状态不需要额外的空间消耗。假设LSTM模块的输入矩阵的维度为 $n \times m$,则LSTM模块的空间复杂性可表示为 $O(12(n \cdot m + n + n^2))$ 。因此,DLMSS-LP的空间复杂度为 $O(140n + 13 \cdot n \cdot m + 12 \cdot n^2)$ 。

3 实验

基于3个具有强稀疏性的KONECT数据集,将DLMSS-LP与4种基准方法(Node2vec^[24]、DDNE、E-LSTM-D和E-GRU-D^[25])在预测精度方面进行比较。

3.1 数据集

本实验的3个数据集中,节点代表人,链路代表人之间的通信(每个链路持续时间超过20 s)。

HyperText数据集记录了ACM超文本会议中的参会人员面对面交流的社交网络,Haggle数据集记录了由所携带的无线设备收集到的人与人之间的通信,Infectious数据集记录了在都柏林科学画廊举办的“*Infectious: stay away*”展览期间,参观者之间的交流,这些数据集信息见表1。

表1 数据集信息
Table 1 Details of three datasets

数据集	HyperText	Haggle	Infectious
节点数量	113	274	410
链路数量	20 818	28 202	17 298
总时间/天	3	4	9
平均稀疏度	0.910	0.989	0.996

为了保持链路的时间顺序,首先根据其时间戳对链路进行排序。然后,根据时间戳、头节点和尾节点删除相同的链路,以降低这些数据集的冗余性。为了处理数据异常,删除没有头节点或尾节点的不完整链路。最后,以固定的时间间隔(20 s)将链路划分为不同的组,每一组代表MSN的一个快照。

将得到的快照转换为邻接矩阵。如果某些参与者(节点)在快照中相遇,则将邻接矩阵中对应的元素值设置为1,否则,将元素值设置为0。此外,将75%的数据作为训练集,其余的作为测试集。

3.2 基准方法

4种基准方法(Node2vec、DDNE、E-LSTM-D和E-GRU-D)的实现细节如下。

1) Node2vec是将宽度优先搜索与深度优先搜索相结合的图嵌入方法。Node2vec通常用于静态网络的链路预测。基于嵌入层,Node2vec通过嵌入层可以将节点映射到低维特征空间,以降低计算成本。

2) DDNE类似于自动编码器,使用两个GRU层作为编码器和解码器。编码器用于提取MSN的特征和降低输入矩阵的维数,而解码器则恢复输出矩阵的维数以得到预测结果的邻接矩阵。

3) E-LSTM-D结合了自动编码器和LSTM模块。自动编码器的功能是提取输入数据的时序特征,降低特征矩阵维度,并且输出预测结果,从而使E-LSTM-D降低学习成本。LSTM模块的功能是提取时间序列时序特征。

4) E-GRU-D将自动编码器与GRU模块相结合,以降低训练成本。在E-GRU-D中,GRU模块保留了门控循环单元,减少更新门中的参数数量,以缩短训练时间。

根据数据集的规模和隐藏层单元数的计算公式,将编码器中的单元数设置为128个。由于解码器中的单元数与输出矩阵和输入矩阵的维数有关,所以解码器中的单元数等于数据集中的节点总数。隐藏层中的单元数见表2。

表2 隐藏层中的单元数
Table 2 Number of cells in hidden layers

数据集	HyperText	Haggle	Infectious
编码器	128	128	128
LSTM模块	129	196	274
解码器	113	274	410

3.3 评价指标

用于测量预测精度的指标有AUC(area under curve)^[26]和ER(error rate)。AUC引入了混淆矩阵(confusion matrix)的概念,它可以衡量

链路预测的整体性能。ER表示预测错误的链路数占链路总数的比例。

AUC为ROC曲线下的面积。如果两个节点之间存在链路，则该链路为正样本，否则为负样本。

具体来说，AUC的计算方法是：

$$AUC = \frac{\sum \left(r_i - \frac{n_0(n_0 + 1)}{2} \right)}{n_0 \cdot n_1} \quad (14)$$

其中， n_0 和 n_1 分别表示正样本和负样本的个数。正样本排名的例子见表3， r_i 表示排名表中第 i 个正样本的节点相似度值的排名。AUC值越大，说明深度学习模型的预测精度越高。

表3 正样本排名的例子
Table 3 Example of ranks of positive samples

	+	-	-	-	+	+	-	+	+
i	1				2	3		4	5
r_i	1				6	7		9	10

ER的计算表达式为：

$$ER = \frac{\sum_{j=0}^n \sum_{i=0}^n |a_{t,i,j} - a'_{i,j}|}{\eta^2} \quad (15)$$

其中， $a_{t,i,j}$ 表示第 t 张快照的邻接矩阵的元素值（第 i 行和第 j 列）， $a'_{i,j}$ 表示预测结果的邻接矩阵的元素值（第 i 行和第 j 行列），并且所有邻接矩阵的维数为 $n \times n$ 。

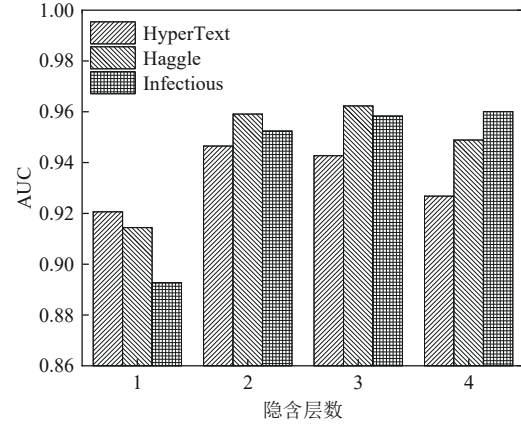
3.4 超参数的影响

LSTM模块和GAE模块的超参数会影响DLMSS-LP的预测精度。

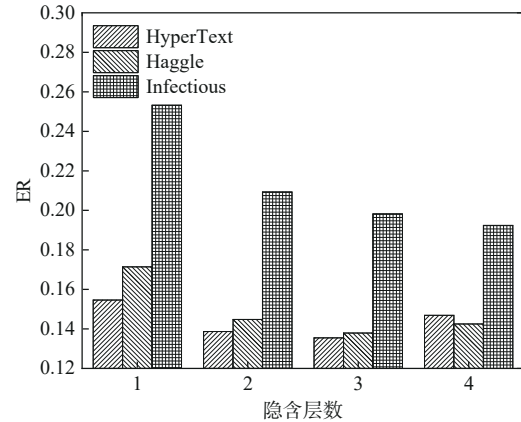
LSTM模块的超参数包括隐藏层数和隐藏层中的单元数。LSTM隐藏层中的单元数量由式（11）计算。当单元数不变时，隐藏层数对DLMSS-LP预测精度的影响如图7所示。

当隐藏层的数量被设置为1时，在3个数据集上进行实验，DLMSS-LP测试结果的AUC值很小，而ER值相对较大，因为DLMSS-LP模型架构简单，使DLMSS-LP的学习能力不足。随着隐含层数的增加，DLMSS-LP的预测精度不断提高。但是，当隐藏层数设置为3和4时，DLMSS-LP对HyperText数据集的预测精度略有降低，而对Haggle数据集的预测精度先提高，然后降低。此外，在Infectious数据集上，随着隐含层数的增加，DLMSS-LP的预测精度逐渐提高。以上结果

表明，隐藏层越多，权值也随之增加，进而DLMSS-LP可以更好地学习特征。但是DLMSS-LP复杂度的增加也会导致过拟合，因此应该根据数据集的规模大小和特征设置适当数量的隐藏层。



(a) 隐藏层数和AUC值



(b) 隐藏层数和ER值

图7 隐藏层数对DLMSS-LP预测精度的影响

Figure 7 Impacts of the number of hidden layers on the prediction accuracy of DLMSS-LP

GAE模块的超参数主要是GCN层的数量。GCN层的数量对DLMSS-LP的预测精度的影响如图8所示。当GCN层的数量设置为1或2，DLMSS-LP的AUC值较高，且DLMSS-LP的ER值相对较低。当GCN层数变大时，DLMSS-LP的预测精度显著降低，因为GCN的主要功能是聚合来自邻接节点的信息。随着GCN层数的增加，更多邻接节点的信息聚合到中心节点，当有足够多的GCN层时，就可以从所有节点的信息中聚合节点的特征。同时，随着深度学习模型复杂度的增加，很容易导致过拟合，从而降低了预测精度。根据实验结果，DLMSS-LP选择两层的GCN。

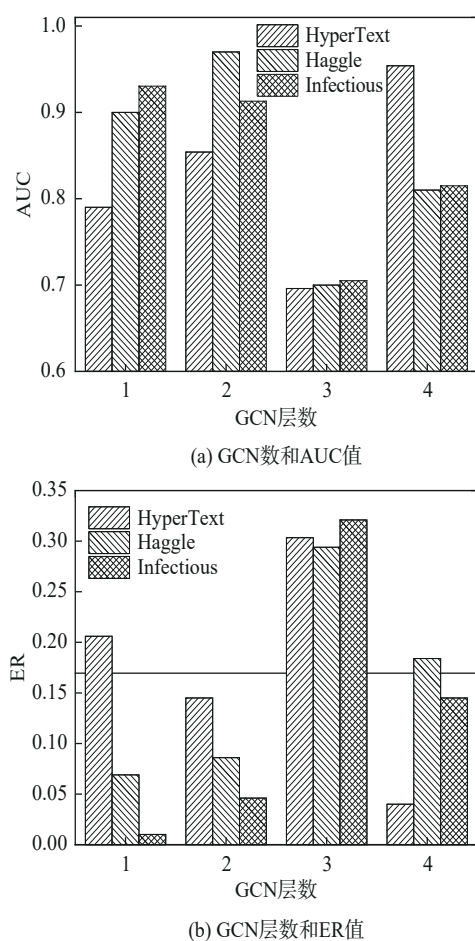


图8 GCN层数对DLMSS-LP预测精度的影响
Figure 8 Impacts of GCN layers on the prediction accuracy of DLMSS-LP

3.5 GCN中的聚合方案

GCN中的聚合方案负责聚合来自邻接节点的特征。选择不同的聚合方案（加权、平均值和最大值），聚合方案对DLMSS-LP预测精度的影响如图9所示。当采用加权作为聚合方案时，DLMSS-LP的AUC值最高，ER值最低。

由于MSN的稀疏性，聚合后的节点的特征值很小，并且采用最大值和求平均值的聚合方案会缩小不同节点之间特征值的差距，从而导致很难进行链路预测（很难判断一些链路是否存在）。当采用加权作为聚合方案时，节点的特征值会增大，不同节点的特征值之间的差距也相应增大，进而提高DLMSS-LP的预测精度。因此，在DLMSS-LP中采用加权的聚合方案。

3.6 特征矩阵聚合模块中层数的影响

特征矩阵聚合模块中的层数对DLMSS-LP预

测精度的影响如图10所示。随着特征矩阵聚合模块中层数的增加，DLMSS-LP的AUC值和ER值基本保持相同。此外，线性层越多，训练所需的资源 and 时间就越多，因此在DLMSS-LP中，特征矩阵聚合模块中的层数被设置为1。

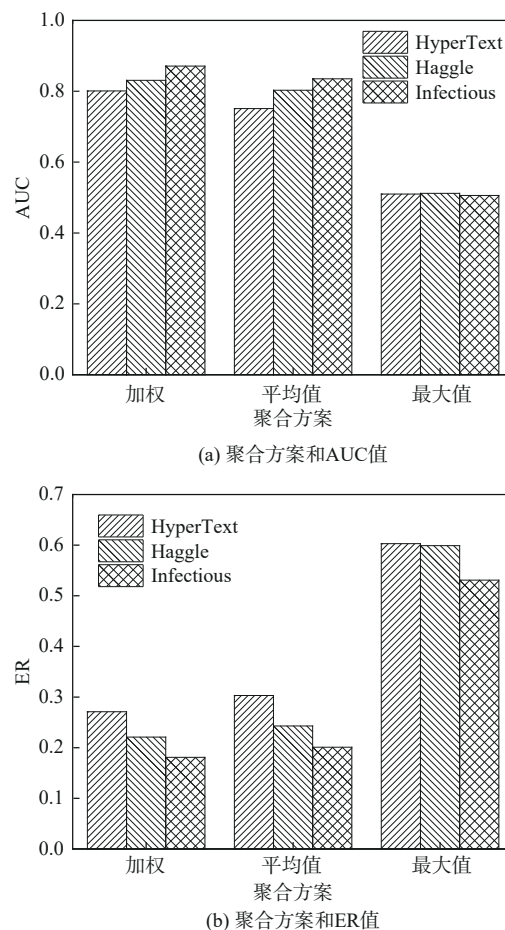


图9 聚合方案对DLMSS-LP预测精度的影响
Figure 9 Impacts of different aggregation schemes on the prediction accuracy of DLMSS-LP

3.7 不确定链路的影响

不同的数据集上的AUC值和不同的数据集上的ER值分别如图11和图12所示，实线表示训练轮次与AUC值（或ER值）之间的关系，红色虚线表示AUC值（或ER值）的平均值。

AUC值和ER值的曲线随着训练轮次的增加而明显波动。这一现象是由于这些数据集中的不确定链路使DLMSS-LP不能完全学习MSN中节点的时序特征。因此，HyperText数据集的不确定链路数大于Infectious数据集和Haggle数据集。

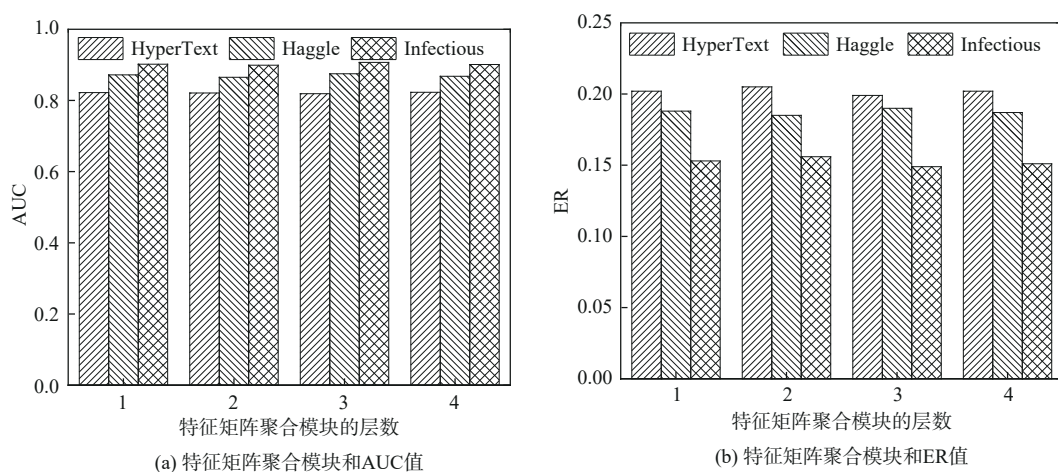


图10 特征矩阵聚合模块中层数对DLMSS-LP预测精度的影响
Figure 10 Impacts of number of layers in feature matrix aggregation module

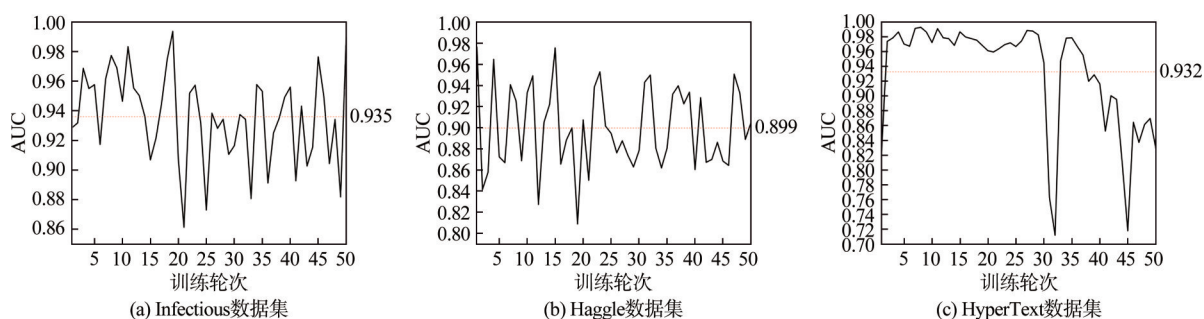


图11 不同的数据集上的AUC值
Figure 11 AUC values on different datasets

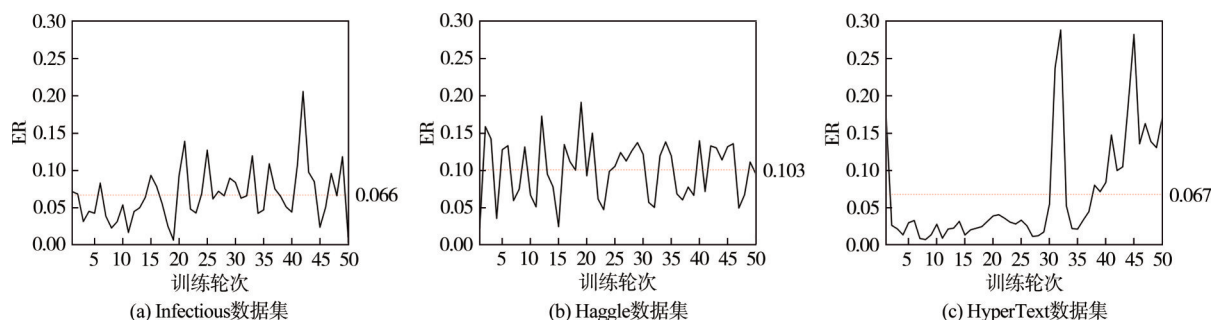


图12 不同的数据集上的ER值
Figure 12 ER values on different datasets

3.8 不同方法之间的比较

本节将DLMSS-LP与4种方法的AUC值和ER值进行了比较,实验结果见表4和表5。

在表4和表5中,Node2vec的预测精度远低于其他包含自编码器结构的方法。实验结果表明,Node2vec不适用于具有强稀疏性的MSN的链路预测。

与DDNE、E-LSTM-D和E-GRU-D相比,DLMSS-LP在所有3个数据集上的ER值都达到

了最佳结果,且DLMSS-LP的AUC值接近这些方法的最佳结果。

MSN中链路较少,稀疏性较强,且AUC受MSN中链路数量的影响,因此DLMSS-LP的AUC值在这些方法中不是最好的。此外,网络的强稀疏性,使在训练数据中链路个数过少,从而导致模型会更倾向于学习无链路的情况,所以会使基线方法出现零矩阵的情况。因此DLMSS-LP、DDNE、E-LSTM-D和E-GRU-D模型获得的

AUC 值差异不明显。

此外, ER 值是由预测错误的链路数计算的, 因此用 ER 度量比 AUC 度量来衡量链路预测的能力更重要。以相遇矩阵和度矩阵为特征矩阵, 采用随机负抽样方法, 通过 DLMSS-LP 可以尽可能地消除强稀疏性对预测精度的负面影响, 从而通过 DLMSS-LP 得到最小的 ER 值。

表 4 不同方法的 AUC 值
Table 4 The AUC values of different methods

方法	数据集		
	HyperText	Haggle	Infectious
Node2vec	0.510	0.592	0.537
DDNE	0.935	0.901	0.897
E-LSTM-D	0.958	0.946	0.941
E-GRU-D	0.952	0.960	0.936
DLMSS-LP	0.899	0.932	0.935

表 5 不同方法的 ER 值
Table 5 The ER values of different methods

方法	数据集		
	HyperText	Haggle	Infectious
Node2vec	0.971	0.951	0.981
DDNE	0.379	0.334	0.455
E-LSTM-D	0.160	0.152	0.238
E-GRU-D	0.181	0.147	0.235
DLMSS-LP	0.103	0.067	0.066

4 结束语

本文提出了基于深度学习的移动社交网络强稀疏性链路预测方法 (DLMSS-LP)。使用相遇矩阵和度矩阵来提高生成链路的可能性, 并消除强稀疏性对链路精度的负面影响。此外, 采用 GAE 模块和特征矩阵聚合模块的架构有助于降低非线性网络的维数, 降低 LSTM 模块提取时变特征的计算成本。大量实验结果表明, 对于 MSN 的链路预测, DLMSS-LP 在 AUC 值和 ER 值方面都表现较好, 说明 DLMSS-LP 具有较高的预测精度。在未来, 研究将集中在由严重噪声引起的 DLMSS-LP 性能波动问题, 例如, 不确定链路的数量非常大的问题。

参考文献:

- [1] QIU T, CHEN B C, SANGAIAH A K, et al. A survey of mobile social networks: applications, social characteristics, and challenges[J]. IEEE Systems Journal, 2018, 12: 3932-3947.
- [2] HUANG Z, ZENG D D. A link prediction approach to anomalous email detection[C]//Proceedings of the 2006 IEEE International Conference on Systems, Man and Cybernetics. Piscataway: IEEE Press, 2006: 1131-1136.
- [3] LIM M, ABDULLAH A, JHANJHI N Z. Performance optimization of criminal network hidden link prediction model with deep reinforcement learning[J]. Journal of King Saud University - Computer and Information Sciences, 2021, 33(10): 1202-1210.
- [4] ELMRABIT N, YANG S H, YANG L L, et al. Insider threat risk prediction based on Bayesian network[J]. Computers & Security, 2020, 96: 101908.
- [5] WANG. B. IDS: an intelligent data semantics system for communication link prediction in SLoT[C]//Proceedings of the 17th International Conference on Mobility, Sensing and Networking (MSN). Piscataway: IEEE Press, 2021.
- [6] ZHANI M, ELBIAZE H, KAMOUN F. Analysis and prediction of real network traffic[J]. J Networks, 2009, 4: 855-865.
- [7] SYAFITRI. W. Social engineering attacks prevention: a systematic literature review[J]. IEEE Access, 2022(10): 39325-39343.
- [8] WANG L, REN J, XU B, et al. MODEL: motif-based deep feature learning for link prediction[J]. IEEE Transactions on Computational Social Systems, 2020, 7(2): 503-516.
- [9] HAWES J S, THEIS H J, HEINIG M. Inferring interaction networks from multi-omics data[J]. Frontiers in Genetics, 2019, 10.
- [10] PASCANU. R, MIKOLOV. T, and BENGIO. Y. On the difficulty of training recurrent neural networks[J]. JMLR. org, 2012, 52: 337-345.
- [11] HIREMATH A, SAYYED I, IJARI K, et al. Optimizing recurrent neural network using neuromodulation[C]//Proceedings of the 2021 12th International Conference on Computing Communication and Networking Technologies (ICCCNT). Piscataway: IEEE Press, 2021: 1-7.
- [12] SHO T. Empirical analysis of the relation between community structure and cascading retweet diffusion[J]. Proceedings of the International AAAI Conference on Web and Social Media, 2019, 13: 493-504.
- [13] VAIBHAV, HUANG P Y, FREDERKING R. RWR-GAE: random walk regularization for graph auto encoders[EB]. arXiv Preprint arXiv: 1908.04003, 2019.
- [14] CHEN W, YANG L Q, ZHA B, et al. Deep learning reservoir porosity prediction based on multilayer long short-term memory network[J]. Geophysics, 2020, 85(4): WA213-WA225.
- [15] ZHANG S, TONG H H, XU J J, et al. Graph convolutional networks: a comprehensive review[J]. Computational Social Networks, 2019, 6(1): 11.
- [16] ZHANG M H, CHEN Y X. Link prediction based on graph neural networks[EB]. arXiv Preprint arXiv: 1802.09691, 2018.

- [17] SU Z, ZHENG X L, AI J, et al. Link prediction in recommender systems based on vector similarity[J]. Physica A: Statistical Mechanics and Its Applications, 2020, 560: 125154.
- [18] RAFIEE S, SALAVATI C, ABDOLLAHPOURI A. CNBP: link prediction based on common neighbors degree penalization[J]. Physica A: Statistical Mechanics and Its Applications, 2020, 539: 122950.
- [19] LI X X, LI Y X, SHANG Y M, et al. DDNE: discriminative distance metric learning for network embedding[M]. Lecture Notes in Computer Science. Cham: Springer International Publishing, 2020: 568-581.
- [20] QU L, ZHU H S, DUAN Q Q, et al. Continuous-time link prediction via temporal dependent graph neural network[C]//Proceedings of the Web Conference 2020. New York: ACM, 2020: 3026-3032.
- [21] CHEN J Y, ZHANG J, XU X H, et al. E-LSTM-D: a deep learning framework for dynamic network link prediction[J]. IEEE Transactions on Systems, Man, and Cybernetics: Systems, 2021, 51(6): 3699-3712.
- [22] YIN Y T, WU Y J, YANG X B, et al. SE-GRU: structure embedded gated recurrent unit neural networks for temporal link prediction[J]. IEEE Transactions on Network Science and Engineering, 2022, 9(4): 2495-2509.
- [23] GILL K S, SHARMA A, ANAND V, et al. Brain tumor detection using VGG19 model on adadelta and SGD Optimizer[C]//Proceedings of the 2022 6th International Conference on Electronics, Communication and Aerospace Technology. Piscataway: IEEE Press, 2022: 1407-1412.
- [24] GROVER A, LESKOVEC J. Node2vec: scalable feature learning for networks[C]//Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM, 2016: 855-864.
- [25] LIU L, YU Z, ZHU H. A link prediction method based on gated recurrent units for mobile social network[J]. Journal of Computer Research and Development, 2023, 60(3): 707-716.
- [26] WELLECK S J. Efficient AUC optimization for information ranking applications[M]. Lecture Notes in Computer Science. Cham: Springer International Publishing, 2016: 159-170.

[作者简介]



何亚迪（2000—），男，江西上饶人，南京邮电大学硕士生，主要研究方向为深度学习的时序分析、车联网。



刘林峰（1981—），男，江苏丹阳人，博士，南京邮电大学教授、硕士生导师，主要研究方向为移动计算、车联网、机器学习方法。