计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.190700136

基于特征分类的链路预测方法综述

王 慧[1,2] 乐孜纯[3] 龚 轩[1] 武玉坤[1] 左 浩[1]
1 浙江工业大学计算机科学与技术学院 杭州 310023
2 江西理工大学应用科学学院 江西 赣州 341000
3 浙江工业大学理学院 杭州 310023

摘 要 复杂网络链路预测作为网络科学研究中一个重要的研究方向,受到了越来越多来自各个学科领域专家的关注,它可以利用现有的网络信息,如节点和边缘的特征,来预测未来可能形成的关系、网络中缺失的信息以及新的或正在消失的信息,识别虚假交互,评估网络演化机制,进行网络重构等。当前链路预测的文献主要来自工程学、计算机科学与物理学的专家,它们各自为政,缺少合作,结合多学科进行链路预测的综述论文少之又少。因此,文中从计算机科学和物理学的视角全面回顾、分析和讨论基于特征分类的链路预测算法的研究进展,介绍了该领域专家们提出的多种特征提取技术,首次把分层的思想引入链路预测算法分类中,将分类模型分为3层,即元数据层、特征分类层和特征抽取层。该分类模型包括"2个大块7个方面",即把常用的链路预测算法分为2个大块(特征提取方法和特征学习方法)和7个方面(基于相似性的方法、基于似然分析的方法、基于概率模型的方法、矩阵分解方法、基于随机游走的方法、基于神经网络的方法和基于自定义损失函数的方法)。该分类方法覆盖了各学科中许多经典的和最新的链路预测技术,包括当前最流行的图神经网络链路预测技术GNN(Graph Neural Network),GCN(Graph Convolutional Network),RNN(Recurrent Neural Network)和RL(Reinforcement Learning)。文中研究了这些算法的模型复杂性和预测性能的差异,并对当前链路预测技术未来所面临的挑战进行了讨论。

关键词:链路预测;复杂网络;机器学习;特征分类;图神经网络
中图法分类号 TP391

Review of Link Prediction Methods Based on Feature Classification

WANG Hui[1,2],LE Zi-chun[3],GONG Xuan[1],WU Yu-kun[1] and ZUO Hao[1]
1 College of Computer Science and Technology,Zhejiang University of Technology,Hangzhou 310023,China
2 College of Applied Science,Jiangxi University of Science and Technology,Ganzhou,Jiangxi 341000,China
3 College of Science,Zhejiang University of Technology,Hangzhou 310023,China

Abstract As an important research direction of network science research,link prediction of complex network has been more and more attention from experts from various disciplines,and it can make use of existing network information,such as the features of the nodes and edges,to predict possible future relationships,missing information in the network,and new or disappearing information,identify the false interactions,evaluate network evolution mechanism and conduct network reconfiguration,etc. Current articles on link prediction mainly come from the expert in engineering,computer science and physics. They are independent and lack cooperation,and there are few review articles combining multidisciplinary link prediction. The goal of this article is from the perspective of computer science and physics comprehensive to review,analyze,and discusse the research progress of link prediction algorithm based on feature classification,introduces a variety of feature extraction techniques in this field. This paper firstly introduces the idea of layering into the classification of link prediction algorithm,which divides the classification model into three layers,the metadata layer,features classification layer and feature extraction layer. The classification model includes two parts and seven aspects,that is,the prediction algorithm is divided into two parts,features extraction method and features learning method. The seven aspects are the similarity based methods,probabilistic methods,likelihood based methods,matrix factorization based method,random walk based method,neural network based method and custom loss function based method. This classification method covers many classical and latest link prediction techniques in various disciplines,including GNN,GCN,RNN and RL,which are the most popular link prediction techniques in graph neural networks. The differences of these algorithms in model

到稿日期:2019-07-19 返修日期:2019-09-22 本文已加入开放科学计划(OSID),请扫描上方二维码获取补充信息。
基金项目:浙江省国际合作"一带一路"专项(2015C04005)
This work was supported by the Special Funding of "the Belt and Road" International Cooperation of Zhejiang Province (2015C04005).
通信作者:王慧(540168713@qq.com)

complexity and prediction performance are studied，and the challenges of current link prediction technology in the future are discussed.

Keywords　Link prediction，Complex network，Machine learning，Feature classification，Graph neural network

1　引言

从 2000 年 Sarukkai[1] 首次在万维网中应用马尔可夫链进行链路预测和路径分析，以及 2007 年 Liben-Nowel 等研究了社交合作网络中的链路预测问题算起[2]，对链路预测的研究已经持续近 20 年。图 1 给出了用主题“链路预测”在 Web of Science 上进行搜索得到的 2000 年 1 月—2019 年 6 月发表的 SCI 论文数，达到了 32 858 篇，尤其在过去的 5 年中研究该问题的文献量有上升趋势。从图 2 可以发现，最近 5 年以“链路预测”为主题的 SCI 论文主要来自工程学、计算机科学、物理学和环境科学等，各学科独自为政，缺少合作，多学科交叉的论文和综述少之又少，如图 3 所示，发文总数为 32 858 篇，而综述论文只有 1 612 篇，不到 5%。

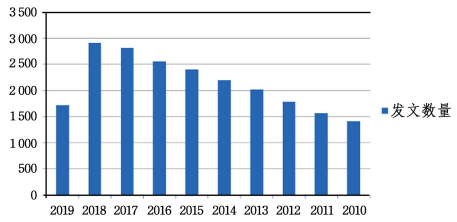


图 1　以“链路预测”为主题发表的 SCI 论文数

Fig. 1　Number of SCI papers published on link prediction

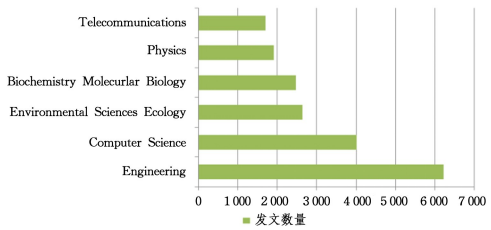


图 2　以“链路预测”为主题的 SCI 论文的研究领域

Fig. 2　Research area of SCI papers on link prediction

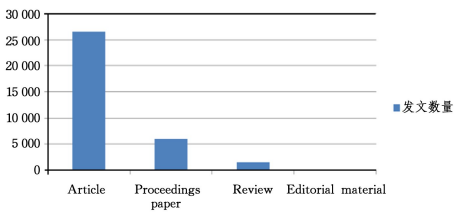


图 3　以“链路预测”为主题的文章的类型

Fig. 3　Type of article on link prediction

2010 年 Lv[3] 发表了首篇与链路预测研究相关的中文综述，该论文获得了 364 次引用，它比较了若干有代表性的链路预测方法，如基于相似性的链路预测、基于最大似然估计的链路预测和基于概率模型的链路预测。但是，该论文强调的是物理学方面的贡献，而不是计算机科学方面的研究成果。2011 年，Hasan 等[4] 综述了一些具有代表性的社交网络链路预测方法，根据网络的分类主要考虑了 3 种类型的模型：二分类模型、概率模型和线性代数模型。它涉及一些新的有代表性的链路预测工作，但适应范围有限，仅适合于社交网络。2016 年 Lv 等[5] 发表了一篇专题文章《网络链路预测：概念与前沿》，其作为一篇比较全面的综述，结合了物理学和计算机科学的最新研究成果，获得了较多的引用。但是由于当时图神经网络技术还不是很成熟，因此其缺少对图神经网络链路预测的论述。

为了填补现有研究工作的不足，本文试图全面、系统地对基于特征分类的链路预测方法进行综述，涵盖各个领域许多经典和最新的链路预测技术、链路预测问题和未来的发展方向。本文首先给出了链路预测语句的形式化定义，特别是从计算机科学和物理学的视角提出了一种新的链路预测分类方法；然后从 2 大块（特征提取和特征学习）7 个方面阐述了常用的链路预测技术：基于相似性的方法、基于似然分析的方法、基于概率模型的方法、矩阵分解方法、基于随机游走的方法、基于神经网络的方法和基于自定义损失函数的方法。

所有方法都提供了可输入的特征，使用监督、无监督、半监督和强化学习的算法，通过聚类或分类来解决链路预测问题。此外，在深入研究基于特征分类的链路预测技术的基础上，对当前链路预测技术存在的问题进行了分析和讨论，提出了未来链路预测问题所面临的挑战。

本文第 2 节介绍了链路预测的基本概念；第 3 节提出了一种新的链路预测分类方法，重点讨论了经典的、新兴的链路预测技术；第 4 节比较了各种链路预测方法的时间复杂度；第 5 节分析了其面临的挑战；最后总结全文。

2　链路预测的定义

2007 年 Liben-Nowel 等[2] 首次科学地定义了链路预测问题：给定一个信息网络 $G(V,E)$，V 代表顶点的集合，E 代表边的集合，$e=\langle u,v\rangle \in E$ 表示节点 u 和 v 的连边，链路预测的任务是指通过对已知网络结构，包括一些可能的节点信息的分析，来评估尚不相连的两个节点之间产生链接的可能性，进而实现预测。

我们可以把它看成是一个典型的二分类问题，令 x 和 y 为图 $G(V,E)$ 中的节点，$l^{(x,y)}$ 是节点对实例 (x,y) 的标签，在链路预测中，每对非连接的节点对应一个实例，包括类标签和描述这对节点的特征。因此，如果节点之间存在链接，那么标记这对节点为正链接，否则标记这对节点为负链接。节点 x 和 y 的标签定义如下：

$$l^{(x,y)}=\begin{cases}1, & \text{if } (x,y)\in E\\ -1, & \text{if } (x,y)\notin E\end{cases}$$

3　一种新的分类方法（分层模型）

当前有许多来自物理学、计算机科学、工程学等领域的专家提出了许多通用的、基本的链路预测方法，这些方法利用节

点的属性信息、拓扑信息以及社会理论来计算节点对之间的相似性。我们从计算机科学和物理学的视角，提出了一种新的链路预测分类技术，将当前的基于特征分类的链路预测模型分为3层，即元数据层、特征分类层和特征抽取层，下面分别对各层及其对应的链路预测方法进行系统的论述（见图4）。

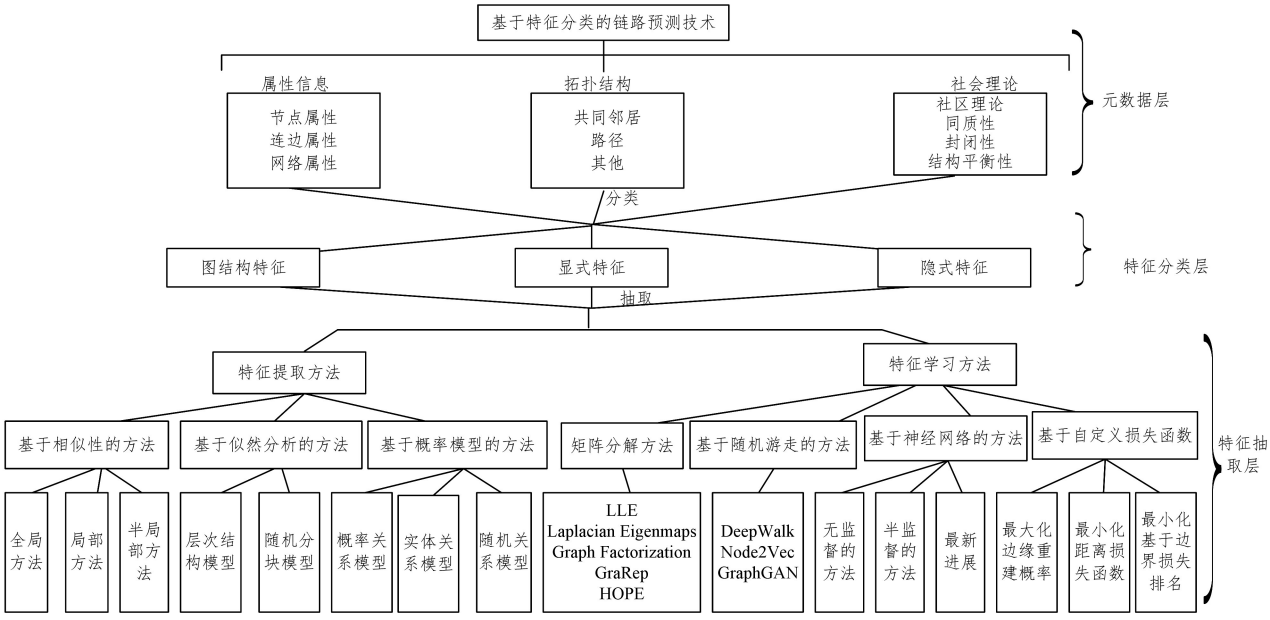


图4 基于特征分类的链路预测技术分层模型

Fig. 4 Hierarchical model of link prediction technology based on feature classification

3.1 第一层：元数据层

基于特征分类的链路预测算法采用3层模型，最底层为元数据层，该层包括3部分：属性信息、拓扑信息和社会理论。

(1)属性信息

属性信息包括节点属性、连边属性和网络属性。描述节点属性最常用的方法是采用文本形式的标签，如年龄、种族、教育、宗教、兴趣、地点和国籍，即用户倾向于与自己相似的人建立关系，两个节点的属性越相似，就越可能产生联系。连边属性包括亲缘关系和交易关系等。网络属性包括连通性、稳定性和鲁棒性等。

(2)2007年Liben-Nowell等[2]将网络的拓扑信息分为基于共同邻居和基于路径两大类，并分析了各种方法在社交网络中的预测效果，他们发现在只考虑节点邻居信息的若干方法中，AA参数[2]表现最好。

(3)社会理论

近年来，越来越多的研究采用古典社会理论，如社区理论、三元理论、封闭性、强与弱关系、同质性和结构平衡性，解决了社交网络分析和挖掘问题。与原先仅使用节点属性和拓扑结构度量不同，基于社会理论的预测指标可以通过捕获有用的附加信息来提高预测性能。

例如Valverde等[6]结合了拓扑结构和社区信息，考虑了用户的需求、兴趣和行为，来预测未来在Twitter的链接。结果表明，该方法具有较高的效率，提高了有向和非对称的大规模社交网络的链路预测性能。Liu等[7]提出了一种链路预测模型，该模型基于弱连接及共同邻居的3个节点中心性，即度、紧密度、中心性。每一个共同的邻居都扮演着不同的角色，根据它的中心性定义节点是否会连接。网络中，节点不仅喜欢链接到相似的节点，而且喜欢链接到中心节点。

在兴趣网络中[8]，同质性不仅被用来预测用户和其他感兴趣的服务之间的链接，还可以用来预测相同兴趣用户之间的链接。

3.2 第二层：特征分类层

该层把元数据层提供的特征分成三大类：图结构特征、显式特征和潜在特征。

(1)图形结构特征位于网络观察到的节点和边缘结构中，可以直接观察和计算。Common neighbors，Jaccard，Preferential attachment，Adamic-Adar，Resource allocation，Katz，PageRank，SimRank等都属于图的结构特征。除此之外，节点中心性（degree，closeness，betweenness，PageRank）等也属于图的结构特征。

(2)潜在特征是节点的潜在属性或表示形式，通常通过因子分解得到，它更倾向于关注全局的属性信息，对于链路预测来说这些信息是非常有用的。它的优势有3个：1)可以模拟现实世界中的社会现象；2)能发现实体之间潜在的结构或与其相关的未知属性；3)可以自动学习潜在特征，并对链路结构做出准确的预测。

事实证明，潜在特征一般具有较强的学习能力，从网络结构的表现形式来实现最先进的预测性能。例如，Cho等[9]提出了一种社交网络模型，其通过标签信息收集用户的位置。结合用户和额外的地理信息，可以分析用户的动向和友情。

(3)显式特征即能直接从元数据中提取的节点属性信息、连边信息和网络信息。

3.3 第三层：特征抽取层

特征抽取层为模型的第三层，该层的作用是对上一层中的图结构特征、显式特征和潜在特征进行抽取。本节针对该

层从 2 大块（特征提取和特征学习）7 个方面分析常用的链路预测技术。

3.3.1　特征提取方法

特征提取方法包括基于相似性的方法、基于概率模型的方法和基于似然分析的方法。

（1）基于相似性的方法

该方法是当前链路预测最通用的框架，其基于非常简单的思想来计算节点之间的相似性，如果两个节点更相似，则它们更有可能在未来产生连接。该类方法包括全局的方法、局部的方法和半局部的方法，如图 5 所示，它们使用节点的拓扑结构作为输入特征。

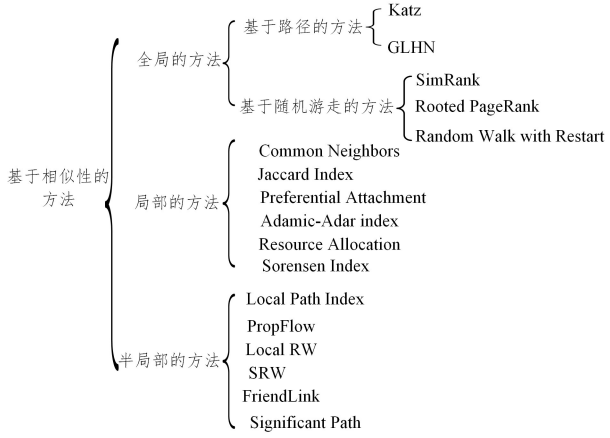


图 5　基于相似性的方法

Fig. 5　Similarity based approaches

本文中的一些标准符号如表 1 所列。

表 1　标准符号

Table 1　Standard symbols

Notations	Descriptions
G	A network
V	The set of nodes
E	The set of edges
$\|E\|$	number of edges
$\|V\|$	number of nodes
m	Number of node attributes
d	The dimension of node vector
$X\in R^{\|V\|\times m}$	The feature matrix
$\tau(x)$	Neighbor of node x
$\|\tau(x)\|$	Degree of neighbor x
A	Represents the adjacency matrix

1）全局方法

与基于局部相似性的方法不同，它利用网络的整体拓扑结构之间的相似性进行节点排序，虽然考虑整个网络的拓扑结构使得链路预测分析更加灵活，但也增加了时间复杂度。根据该方法的特点，使用时可将其分为基于路径的方法和基于随机游走的方法，前者包括 Katz 和 GLHN（Global-Leicht-Holme-Newman Index），后者包括 SimRank，Rooted PageRank，Random Walk with Restart 等。

Katz：直接对两顶点之间的所有路径求和，并且通过惩罚因子 β 来惩罚较长的路径，从而达到减少较长路径对节点间相似性的贡献。因此其可以看成是基于全局信息的指标，定义为[10]：

$$S^{KI}_{(x,y)}=\sum_{i=1}^{\infty}\beta^i\,|paths^{\langle I\rangle}_{xy}| \tag{1}$$

其中，$\beta>0$ 为可调参数，$|paths^{\langle I\rangle}_{xy}|$ 表示连接节点 x 和 y 的路径中长度为 I 的路径数。

GLHN：与 Katz 相似，如果这些节点之间的路径相似度总数高，则这些节点的相似性高[11]。

$$S^{GLHN}=\lambda(I-\beta A)^{-1} \tag{2}$$

对于更短的路径，可调参数 β 和 λ 被认为是更重要的，β 给出更小的值。

SimRank：是一种基于图的拓扑结构信息来衡量任意两个对象间相似程度的模型，该模型由 Glen 提出，其核心思想为：如果两个对象连接到相似的节点，那么这两个对象是相似的。在最坏的情况下，SimRank 的计算复杂度为 $O(n^4)$，其中 n 为顶点数，如此高的计算成本限制了它在大规模网络中的广泛应用[12]，SimRank 在计算相似度时不仅时间复杂度高，而且还存在“零相似度”问题。当 $x=y$ 时，$SimRank=1$。

$$S^{SR}_{(x,y)}=r\cdot\frac{\sum_{a\in\tau(x)}\sum_{b\in\tau(y)}S^{SR}_{(a,b)}}{|\tau(x)|\cdot|\tau(y)|} \tag{3}$$

Rooted PageRank（RPR）[13]：是 PageRank 算法的一个改进，通过搜索引擎来对搜索结果排序。图中，节点的排序与随机游走的概率成正比。此外，因子 ε 指算法访问节点邻居的可能性与重新开始的可能性的差值；$\boldsymbol{D}$ 是一个对角矩阵，$D_{i,i}=\sum_j A_{i,j}$，RPR 定义如下：

$$RPR=(1-\varepsilon)(I-\varepsilon\boldsymbol{D}^{-1}\boldsymbol{A})^{-1} \tag{4}$$

Random Walk with Restart（RWR）：直接应用了 PageRank 的算法思想，其假设随机游走粒子每走一步都以一定的概率返回初始位置[14]。

$$S^{RW}_{(x,y)}=p^{\rightarrow y}_x+p^{\rightarrow x}_y \tag{5}$$

其中，$p^{\rightarrow y}_x$ 为从节点 x 出发的粒子最终走到节点 y 的概率。RWR 指标已被应用于推荐系统的算法研究中。

2）局部方法

基于局部相似性的方法认为如果节点对具有公共的邻居或者其中某个节点有较高的度，则该方法在未来将可能产生连接。因为它们仅仅使用局部拓扑信息而不是整个网络结构，所以速度快于基于全局相似性的方法，特别是在动态网络上。局部相似性的方法包括 Common Neighbors，Jaccard Index，Preferential Attachment，Adamic-Adar index，Resource Allocation 和 Sorensen Index 等，下面将分别进行论述。

Common Neighbors（CN）：是最简单、最常用的度量方法，它表示节点 x 和节点 y 的相似性为它们共同邻居的数量[15]，即：

$$CN(x,y)=|\tau(x)\bigcap\tau(y)| \tag{6}$$

Jaccard Index（JC）：常用在信息检索中，假定节点对的值越高，它们共享的公共邻居相对于总邻居的比例就越高[16]，即：

$$JC(x,y)=\frac{|\tau(x)\bigcap\tau(y)|}{\tau(x)\bigcup\tau(y)} \tag{7}$$

Preferential Attachment（PA）：是一种偏好链接指标，一条新边连接到节点 x 的概率和 x 节点的度 $k(x)$ 成正比，因此

节点 x 和 y 连接的概率与两节点的度的乘积成正比[17]。

$$PA(x,y)=k_x\times k_y \tag{8}$$

Adamic-Adar(AA) index：AA 由 Adamic 和 Adar 提出，被用于计算两个网页之间的相似性[18]，此后被应用于社交网络。这个指标是在共同邻居算法的基础上赋予其权重，共同邻居越少，权重越大，即：

$$AA(x,y)=\sum_{z\in\tau(x)\cap\tau(y)}\frac{1}{\log|\tau(z)|} \tag{9}$$

Resource Allocation (RA)：类似于 AA，但惩罚的程度比 AA 更重，精确性上略胜 AA 指标，它和 AA 一样都是抑制度大的共同邻居的贡献，RA 和 AA 不仅使用直接的邻居，而且考虑了邻居的邻居[19]。

$$RA(x,y)=\sum_{z\in\tau(x)\cap\tau(y)}\frac{1}{|\tau(z)|} \tag{10}$$

Sorensen Index(SI)：主要用于生态学的数据研究[20]，除了考虑共同邻居的大小，也指出节点度越小，连接的可能性越大。

$$SI(x,y)=\frac{|\tau(x)\bigcap\tau(y)|}{|\tau(x)|+|\tau(y)|} \tag{11}$$

3)半局部方法

考虑到局部方法虽然利用较少的网络结构信息，具有较低的时间复杂度，但是预测效果较差，而全局方法利用了大部分的网络结构信息，具有较高的时间复杂度，但预测效果较好。为了平衡时间复杂度和效率，研究者提出了半局部方法。半局部方法包括 Local Path index, PropFlow, Local RW, FriendLink 和 Significant Path，下面将分别进行介绍。

Local Path index：在直接共同邻居的基础上考虑了长度为 3 的路径上的间接邻居的贡献[21]，并且通过 a 给间接邻居一定的惩罚。其中，$\boldsymbol{A}$ 为邻接矩阵；$a\in(-1,1)$ 为可调参数，给予长路径的惩罚因子，当 $a=0$ 时，$LP=CN$。

$$S^{LP}=\boldsymbol{A}^2+a\boldsymbol{A}^3 \tag{12}$$

PropFlow：通过一个本地路径计算节点之间的信息，以评估两个节点之间的关系[22]。PropFlow 值越大，未来连接的概率越大。

$$S^{PFP}_{(x,y)}=s^{PFP}_{(a,x)}\frac{\boldsymbol{\omega}_{xy}}{\sum_{k\in\tau(x)}\boldsymbol{\omega}_{xy}} \tag{13}$$

现在 PFP 已广泛应用在有向的、无向的、加权的、未加权的、稀疏的或稠密的网络。

Local RW(Local Random Walk, LRW)：虽然基于随机游走的算法具有良好的稀疏性，但对该算法来说，数据量仍然是一个具有挑战性的问题，因此 Lv 等提出了局部随机游走 LRW，它只考虑有限步数的随机游走过程。此方法的时间复杂度为 $O(N\langle K\rangle^n)$，比全局随机游走低，适合规模较大、较稀疏的网络[23]。$p_y^{\rightarrow x}(t)$ 为节点 y 出发的随机游走粒子在 t 时刻到达节点 x 的概率。

$$S^{LRW}_{(x,y)}(t)=\frac{|\tau(x)|}{2|E|}p_y^{\rightarrow x}(t)+\frac{|\tau(y)|}{2|E|}p_x^{\rightarrow y}(t) \tag{14}$$

SRW(有重叠的随机游走指标)：许多实际的网络都具有较大的集聚系数，网络更倾向于与较近的节点相连接。为了实现这一目标，Lv 等提出了有重叠的随机游走指标，SRW 的值等于不同时刻下的 LRW 的结果的叠加[23]。

$$S^{SRW}_{(x,y)}(t)=\sum_{l=1}^{t}S^{LRW}_{(x,y)}(l) \tag{15}$$

FriendLink 方法：把 x 和 y 之间的相似性定义为从 x 到 y 可指定长度的路径上的邻居节点(包括直接邻居和间接邻居)对这两个节点的相似性贡献。其中，$|V|$ 为图中节点的总数，l 为最大路径的长度，$|A^i_{x,y}|$ 表示连接两节点 x 和 y 之间长度为 i 的所有路径的集合[24]。

$$S^{FL}_{(x,y)}=\sum_{i=1}^{l}\frac{1}{i-1}\frac{|A^i_{x,y}|}{\prod_{j=2}^{i}(|V|-j)} \tag{16}$$

Significant Path 方法：它的核心思想包括两个方面：一方面是连接两个节点的不同长度的路径对其相似性贡献不同，另一方面是在连接两个节点的路径中，直接邻居和间接邻居的度对两个节点的相似性不同[25]。其中，$M(p)$ 是指路径 P 的中间节点的集合，α 和 β 为惩罚因子。

$$\sum_{p\in P^2_{uv}}\sum_{w\in M(p)}d_w^{\beta}+\alpha\sum_{p\in P^3_{uv}}\sum_{w\in M(p)}d_w^{\beta} \tag{17}$$

(2)基于似然分析的方法

Newman 提出的基于最大似然分析的方法，根据已知的网络结构的组织原则、目前已经观察到的链路计算网络的似然值以及网络似然最大化，通过详细规则和具体参数来计算每一对未连接的节点产生连边的可能性，因此在处理草原食物链这种层次结构的网络时具有较好的效果，但由于该算法在每次预测时都要更新样本，时间复杂度较高，实用性不强。

1)层次结构模型

Aaron 认为网络都是具有某种层次结构的，网络可以用族谱树表示，每个非叶子节点有一个概率值，则两个叶子节点连接的概率等于它们最近共同祖先节点的概率值[26]。其中，D 为族谱树，每个内部节点 r 都有一个概率值 P_r，$P_r\in[0,1]$。

$$L(D,\{pr\})=\prod_{r\in D}p_r^{E_r}(1-p_r)^{L_rR_r-E_r} \tag{18}$$

2)随机分块模型

当网络不能表示为层次结构时，可以采用另一种方法。Guimerar 等提出基于随机分块模型的算法，该方法假设节点位于某个社区中，节点可以被分为若干簇或小组，两个节点间连接的概率与它们的自身属性无关，仅取决于节点所在社区的关系。随机分块模型的效果优于层次结构模型，并且该方法还可以剔除网络的错误连边，如纠正 Yeast 网络中的错误连边[27]，但缺点是时间复杂度较高。

$$L(A|M)=\prod_{\alpha\leqslant\beta}Q_{\alpha\beta}^{l_{\alpha\beta}}(1-Q_{\alpha\beta})^{r_{\alpha\beta}-l_{\alpha\beta}} \tag{19}$$

3)基于概率模型的方法

概率模型的目的是观察网络抽象底层的结构关系，运用学习模型来预测缺失的链接。根据函数建立可调参数，通过可调参数来达到与目标网络数据最佳的匹配结果。概率模型作为数据挖掘领域的典型模型，不仅考虑了网络的拓扑结构信息，还考虑了节点的属性信息，预测结果比较理想，但由于概率模型在获取节点属性信息时时间复杂度较高，因此在实施上有些困难。

当前的链路预测有3种主要的概率模型：概率关系模型(Probabilistic Relational Model，PRM)、随机关系模型(Stochastic Relational Model，SRM)和实体关系模型。为了理解链路预测概率关系模型的基本概念和符号，必须熟悉关系代数符号、关系贝叶斯网络(Relational Bayesian Networks，RBN)和关系马尔可夫网络(Relational Markov Networks，RMN)[28]。

1)概率关系模型

PRM继承和综合了关系模型和概率模型的有关技术和思想，形成了一种适合网络的预测模型，该模型包括类和关系，每个实体都包含一些属性，即包括了某类实体内部的关系和实体之间的关系，每个属性值都限制在预定义的域中[29]。

该模型通过创建方法来划分，可以分为贝叶斯网络关系模型、马尔可夫网络关系模型和关系依赖网络模型(Relational Dependent Network，RDN)，这些模型的缺点是计算量很大。

2)实体关系模型

实体关系模型直接从现实世界中抽象出实体类型和实体间关系表示的数据模型。最重要的和广泛使用的实体关系模型是有向无环实体关系模型(Directed Acyclic Probabilistic Entity-Relationship model，DAPER)，一个DAPER由6个类组成：实体类、关系类、属性类、弧线类、局部概率分布类和限制条件类。

3)随机关系模型

随机关系模型的关键概念是“随机”的实体过程，它是由多个实体相互作用而产生的高斯过程(Gaussian Processes，GP)[29]。

假设链接 r 由潜在关系函数 t 来表示：$U\times V\rightarrow R$，其中，$r_{i,n}$是链接，$t_{i,n}$是潜在的值。θ_{Σ} 和 θ_{ω} 分别是GP核函数在 U 上的超参数和GP核函数在 V 上的超参数，边际似然为：

$$P(R_I|\theta)=\int \prod_{(i,n)} p(r_{i,n}\mid t_{i,n})p(t|\theta)\mathrm{d}t \tag{20}$$

3.3.2　特征学习方法

为了学习图结构特征、显式特征和隐式特征，近年来，各领域的专家提出了许多基于特征学习的链路预测方法，这些方法将节点嵌入到图并学习图的表示技术，可以将其看作一种映射的图降维技术，将图中的每一个节点都映射到一个低维向量空间，并且在此空间内保持原有图的结构信息或距离信息。

按时间先后顺序对这些方法进行分类。第一类是传统方法，包括LDA(Linera Discriminant Analysis)，PCA(Principle Component Analsysis)和MDS(MultiDimensional Scaling)等降维方法。第二类是在2000年左右提出的较为经典的方法，包括同态映射ISOMAP、局部线性镶嵌LLE和拉普拉斯特征分解LE。第三类是最近5年提出的新方法，可将其分成4类：基于矩阵因子分解的方法、基于随机游走的方法、基于神经网络的方法和基于自定义损失函数的方法。

(1)矩阵分解的方法

节点之间的链接可以用邻接矩阵表示，其中每一行和每一列表示不同的节点和布尔变量，用来表示节点对之间是否存在链接。Memon等[30]将链路预测作为矩阵补全问题，扩展了矩阵分解方法来求解链路预测问题。该模型可以通过双线性回归模型将图中的节点、链接的潜在特征和显式特征结合起来用矩阵分解方法解决。潜在特征也可以与其他链路预测模型相结合，直接优化了AUC值，解决了在所有链接中正链接占比很小、负链接占比很大的不平衡问题。

Aditya等[31]提出了一种监督的矩阵分解方法。该方法能够学习潜在的拓扑结构特征，将潜在特征、可选显式节点和边缘特性组合在一起，其性能比单独使用这两种特征更优。

Kunegis等[32]也提出了一种基于网络邻接矩阵的代数谱变换的方法，来预测网络中链路的存在性和链路的权值。相比其他几种机器学习的方法，该方法要学习的参数较少，但计算复杂度较高。

总之，当前所使用的矩阵分解方法多数采用的是图嵌入技术，通常图嵌入旨在低维空间中表示图，以保留尽可能多的图的属性信息[33]。不同的算法对节点(边、子结构、子图)的相似性以及如何在嵌入空间中保留这些特征，有不同的见解。下面介绍几种常用的矩阵分解方法：Graph Factorization，Locally Linear Embedding(LLE)，HOPE，Laplacian Eigenmaps和GraRep，以及它们如何量化图的属性并解决图嵌入问题。

1)Graph Factorization(图因子分解)

图因子分解方法使用矩阵的方式描述形成的网络，并通过矩阵分解得到每一个节点的嵌入。它对图的拉普拉斯特征映射法使用相同的损失函数，通过矩阵分解用低维向量表示每个顶点，并使用随机梯度下降法对其进行优化。其能够处理大型的网络，但只适用于无向网络[34]。它使用的最小化损失函数如下：

$$\Phi(Y,\lambda)=\frac{1}{2}\sum_{(i,j)\in E}(w_{ij}-\langle Y_i,Y_j\rangle)^2-\frac{\lambda}{2}\sum_i\|Y_i\|^2 \tag{21}$$

其中，λ 是正则化系数。

2)Locally Linear Embedding(LLE)

LLE假设每个节点都是在嵌入空间中相邻节点的线性组合，它能够使降维后的数据保持原有的拓扑结构特征。LLE算法可以归纳为三步。第一步寻找每个样本点的K个近邻点[35]。第二步根据每个样本点的近邻点计算出该样本点的局部重建权值矩阵。第三步由该样本点的局部重建权值矩阵和其近邻点计算出该样本点的输出值。LLE方法操作简单，且算法中的优化不涉及局部最小化，能解决非线性映射，但是当数据的维数过大、数量过多，以及稀疏矩阵过大时此方法不适用。

$$\Phi(y)=\sum_i|Y_i-\sum_j W_{ij}Y_j|^2 \tag{22}$$

3)HOPE

HOPE是一种高阶近似保留嵌入方法，用于解决有向图嵌入中的非对称传递性问题。这个属性将因子分解图的顶点映射到向量空间时，对于捕获节点的结构是至关重要的。将一个奇异值分解应用于邻接矩阵，利用奇异值来优化向量表示值[36]，其时间复杂度与图形大小呈线性关系，该方法常用奇异值分解来获得高效的嵌入。目标函数为：

$$\|\boldsymbol{S}-\boldsymbol{Y}_S\boldsymbol{Y}_t^{\mathrm{T}}\|_F^2 \tag{23}$$

其中，$\boldsymbol{S}$ 是相似度矩阵，$\boldsymbol{Y}_s$ 和 $\boldsymbol{Y}_t$ 表示嵌入向量。HOPE 通过最小化式(23)来保留更高阶的近似。

4)Laplacian Eigenmaps

该算法由 Belkin 等提出，首先使用 ε 或 K 最近的邻居构造图，然后使用节点对的权值最小化损失函数[37]。如果两个节点之间的权值越大，则表明这两个相似节点之间的距离越小，通过给相距较远的两个相似节点更多的惩罚，最终使得节点间有较高的相似度。最小化目标函数如下：

$$\Phi(Y)=\frac{1}{2}\sum_{i,j}|Y_i-Y_j|^2 w_{i,j}=\mathrm{tr}(y^{\mathrm{T}}LY) \tag{24}$$

其中，L 是图 G 的拉普拉斯算子。

5)GraRep

GraRep 是基于全局结构信息的图表征学习，它使用矩阵分解的方法来学习节点表示，该算法考虑了一种特别的关系矩阵，通过 SVD 分解对该关系矩阵进行降维，从而得到 K 步网络表示。其将图形邻接矩阵提升到不同的幂来利用不同尺度的节点共现信息，将奇异值应用于邻接矩阵的幂以获得节点的低维表示[38]。GraRap 的缺点是可扩展性差，在计算关系矩阵时计算效率较低。

(2)基于随机游走的方法

将深度优先(Depth First Search，DFS)和广度优先(Breadth First Search，BFS)搜索引入随机游走序列的生成过程，对于大型的图特别是社交网络图，研究节点的特征(节点中心性和节点相似性)的作用是非常明显的，当只能观察到一部分图或是图太大无法完整测量时，随机游走的效果较好[39]。本文总结了当前采用的图的随机游走嵌入技术，即 DeepWalk，Node2vec 和 GraphGAN。

1)DeepWalk

DeepWalk 是最近提出的一种社交网络嵌入方法，采用神经语言模型 SkipGram 进行图嵌入，把图的表示学习作为一种自然语言的处理问题[40]。其利用节点在网络上的随机游走路径来模仿文本生成过程，使用截断的随机游走从输入图中采样一组路径，提供一个节点序列，每个路径相当于句子，节点相当于单词，然后将 SkipGram 和 Hierarchiacal 应用于路径，对随机游走序列中每个局部窗口内的节点对进行概率建模，最大化随机游走序列的似然概率值，并使用随机梯度下降学习参数。DeepWalk 的目标函数如下：

$$L(s)=\frac{1}{|s|}\sum_{i=1}^{|s|}\sum_{i-t\leqslant j\leqslant i+t,j\neq i}\log Pr(v_j|v_i) \tag{25}$$

其中，

$$Pr(v_j|v_i)=\frac{\exp(v_j'\cdot v_i)}{\sum_{v\in V}\exp(v'\cdot v_i)} \tag{26}$$

2)Node2vec

Node2vec 的本质是进行文本的嵌入，该算法通过改变随机游走序列生成的方式来进一步扩展 DeepWalk 算法，它是一种学习网络中的节点连续特征的框架，它将节点映射到一个低维特征空间，以最大限度地保护节点的网络邻域[41]。其实质是通过节点的特征来预测周围的词。Aditya 等认为 Node2vec 的 AUC 值在链路预测方面优于 Common Neighbors，Jaccard's Coefficient，Adamic-Adar。其目标函数如下：

$$\max_f\sum_{u\in V}\log P(Ns(u)|f(u)) \tag{27}$$

其中，$P(Ns(u)|f(u))=\sum_{n_i\in Ns(u)}P(n_i|f(u))$，$V$ 是网络中节点的集合，f 将节点映射到特征空间。

3)GraphGAN

当前的大多数特征学习方法可以分为三大类。第一类为生成式模型，该模型旨在网络中学习一个潜在的连通性分布，前文介绍的 DeepWalk 和 Nodevec 属于此类。第二类为判别式模型，此模型试图学习一个判别器，利用此判别器来判断两个节点之间存在连边的概率，SDNE 和 PPNE 属于此类，SDNE 算法将邻接矩阵当作每一个节点的原始特征，然后用自动编码器监督式地获得稠密低维的节点表达；PPNE 算法通过正样本和负样本来监督地学习节点向量。第三类为生成式＋判别式结合模型，GraphGAN 属于此类。

GraphGAN[39]是一个在网络生成学习中将生成模型和判别模型加以结合的框架，它由两部分组成，即生成器和判别器，生成器的主要功能是生成与真实数据尽可能相似的数据，去“欺骗”判别器；判别器设计使用 Sigmoid 函数尽可能地将真实数据和生成的数据区分开。

(3)基于神经网络的方法

上文讨论的所有节点嵌入方法都是浅嵌入方法，其中编码器只是一个简单的嵌入查找，这导致算法存在一些缺点：1)编码器内的节点之间不共享任何参数，计算效率低；2)浅嵌入不能在编码期间利用节点属性。最近研究者提出了一些方法来解决这些问题，这些方法与浅层框架不同，它们使用深度神经图表示(DNGR)和结构深度网络嵌入(SDNE)来解决缺点 1)。DNGR 和 SDNE 直接使用深度神经网络将图结构合并到编码器算法中。这些方法的基本思想是使用自动编码器来压缩一个节点的本地邻域信息。

下面本文介绍其他基于深度神经网络的方法来解决上述浅嵌入的问题。将基于神经网络的方法分为无监督的方法、有监督的方法和最新的进展。其中，无监督的方法即图自动编码器 GAE，其主要由两部分组成：自编码器 AE 和变分自编码器 VAE。半监督的方法主要包括图神经网络 GNN 和图卷积网络 GCN。最新的方法包括 Graph RNNS 和 Graph RL。具体如图 6 所示。

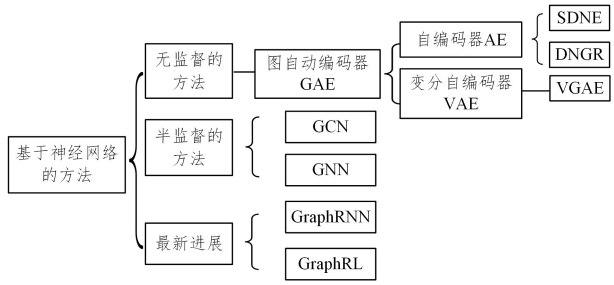


图 6 基于神经网络的方法

Fig. 6 Methods based on neural network

1)无监督的方法——图自动编码器 GAE

图自动编码器 GAE 由编码器和译码器组成，编码器把图形映射到低维空间存储图的结构信息。解码器学习解码嵌入的信息并输出标签。基于神经网络的自动编码器能够提取

出图的复杂结构特征，并能根据自动编码器隐藏层的通道数来进行降维，这不仅有利于测试不可见的顶点实现高的预测性能，而且有利于聚类节点，改善预测的精度。该方法优于基于因子分解的节点映射。它包括自编码器 AE 和变分自编码器(VAE)两种方法。其中，AE 包括 SDNE 和 DNGR，VAE 包括 VGAE。

Structured Deep Network Embedding(SDNE)：其本质就是基于图的自动编码器。Wang 等[42]提出使用深度自动编码器来保持一阶和二阶网络的邻近度。该方法采用高度非线性函数进行嵌入，它由监督和无监督两部分组成，使用了两种降维方法，即 L2-Reconstruction 和拉普拉斯特征映射。

DNGR：其使用了 L2-Reconstruction 方法进行降维，结合了随机游走和深度自编码器，该模型由 3 部分组成：随机游走、正点互信息(PPMI)计算和叠加去噪自编码器[43]。在输入图上使用随机游走模型生成共现概率矩阵，类似于 HOPE 中的相似矩阵。将该矩阵转化为 ***PPMI*** 矩阵，并将其输入到叠加去噪自动编码器中，得到节点的嵌入。输入 ***PPMI*** 矩阵保证了自动编码器模型能够捕获更高阶的近似度。此外，使用叠加去噪自动编码器有助于保证模型在图中存在噪声时的鲁棒性，以及捕获任务(如链路预测和节点分类)所需要的底层结构。

Variational Graph Auto-Encoders (VGAE)：其是变分自编码器 VAE 中的一种，它利用潜在的变量学习无向图的潜在表示，它由一个图形卷积网络 GCN 编码器和一个简单的内积译码器[44]组成，输入邻接矩阵，通过 GCN 来学习节点之间的高阶依赖关系。实验结果表明，在引文网络中使用图变分自编码器能取得较好的预测性能。

2)半监督的方法

图卷积网络 GCN：对于大型稀疏图来说，采用 SDNE 和 DNGR 方法，以每个节点的全局邻域作为输入，是一种计算代价高且不适用的方法。因此，图卷积网络 GCN 应运而生，其通过在图上定义卷积算子来解决这个问题。它设计了一种卷积神经网络作用于网络结构，并使用一种基于边的标签传播规则实现半监督的网络表示学习。该模型迭代地聚合了节点的邻域嵌入，并使用在前一次迭代中获得的嵌入及其嵌入的函数来获得新的嵌入[44]。通过局部邻域的聚合嵌入使其具有可扩展性，多次迭代允许学习嵌入一个节点来描述全局邻域。

图神经网络 GNN：其是一种在图域上操作的深度学习方法[45]。其遵循循环递归邻域聚合(或者消息传递)的模式，其中每个节点聚合其相邻节点的特征向量，以计算其新的特征向量。在 K 轮聚合迭代后，通过其转换的特征向量来表示该节点，该向量捕获节点的 K-hop 网络邻节点的结构信息，然后通过 pooling 来获得整个图结构的表征，例如对图中所有节点的表征向量求和。

GNN 统一了递归神经网络和马尔可夫链，在许多任务中达到了最佳的性能。未来，可以将 GNN 与 GCN 统一到一个通用框架中。

3)图神经网络的最新进展

图递归神经网络(Graph RNN)：You 等[46]针对图生成问题采用了 Graph RNN 方法。他们采用两个 RNN，一个用于生成新的节点，另一个用于以自回归的方式为新添加的节点生成边。结果表明，与传统的基于规则的图生成模型相比，这种分层 RNN 结构在具有可接受的时间复杂度的同时，能够有效地从输入图中学习特征，实现良好的预测性能。

我们也可以将 Graph RNN 与其他架构(如 GCN 或 GAE)结合使用。例如，MGCNN[47]将 CNN 与 RNN 进行耦合，逐步重构图，旨在解决图的稀疏性问题。Dynamic GCN[48]应用 LSTM 在动态网络中收集不同时间片段的 GCN 结果，其目的是获取空间和时间图信息。

图强化学习(GraphRL)：强化学习是深度学习的一个方面，已经被证实在 AI 领域非常有效。RL 被认为能够处理不可微的目标和约束，其有两种常用方法：1)GCPN[46]将图表示与 RL 相结合，利用 RL 生成目标导向的分子图，以处理不可导目标和约束，实验结果证明了 GCPN 在解决各种图生成问题和预测问题时的有效性。2)MolGAN[49]也采用了同样的思想，即使用 RL 生成分子图，其建议直接生成完整的图，而不是通过一系列的动作来生成图，这对小型网络很有效。

4)基于自定义损失函数的方法

基于自定义损失函数的方法使基于节点嵌入建立的边尽可能与输入图中的边相似，它包括最大化边重建概率、最小化基于距离的损失函数和最小化基于边距的损失排名。

最大化边重建概率：好的节点嵌入方法应该能够重建原始输入图中的边。其可以通过使用节点嵌入，最大化所有观察到的边的生成概率来实现。生成概率可以用一阶邻近度和二阶邻近度来衡量，例如 LINE 方法。

LINE(Large-scale Network Information Embedding)[50]，是一种非常成功的浅嵌入方法，它定义了两种损失函数，一种是一阶邻近关系，另一种是二阶邻近关系。

所谓一阶邻近关系指两个点之间是否直接相连，它使用基于 sigmoid 函数的解码器。二阶邻近关系指两个点的邻居的相似度，即两个节点虽然没有直接相连，但它们的邻居是完全重合的，因此可以认为这两个点具有较强的二阶邻近关系。基于这样一种邻近关系，LINE 定义了两个损失函数 O_1 和 O_2，然后基于这两个损失函数对一阶、二阶目标进行优化。虽然该算法具有很好的学习效果，但是它们并不能很好地捕捉到远距离节点之间的关系。

$$O_1 = -\sum_{(i,j)\in E} W_{i,j} \log p_1(v_i, v_j) \tag{28}$$

$$O_2 = \sum_{i\in V} \lambda_i d(\hat{p}_2(\cdot \mid v_i), p_2(\cdot \mid v_i)) \tag{29}$$

最小化基于距离的损失：可以基于节点嵌入来计算节点邻近度，或者通过观察到的边，凭经验计算节点邻近度。该方法最小化两种类型的邻近度之间的差异，使基于节点嵌入计算的节点邻近度尽可能接近基于观察到的边计算的节点邻近度[50]。

$$O_{\min}^{(1)} = \min -\sum_{e_{ij}\in E} A_{ij} \log p^{(1)}(v_i, v_j) \tag{30}$$

$$O_{\min}^{(2)} = \min -\sum_{e_{ij}\in E} A_{ij} \log p^{(2)}(v_j, v_i) \tag{31}$$

最小化基于边距的排名损失：在基于边距的排名损失优

化中，输入图的边指节点对之间的相关性。图中的一些节点通常与一组相关节点相关联。例如，在 cQA 站点中，一组答案标记与给定问题相关。节点的嵌入更类似于相关节点的嵌入，而不是任何其他不相关节点的嵌入，例如 TransX 方法[51]。

TransX 表示一系列方法，X 可以指代任何字母，这些方法都是基于 KG 嵌入的。KG 嵌入指将 KG 中的每个 entity 和 relation 都映射到一个低维连续空间中，并且保持原来的图结构信息。比较经典的方法有 TransE，TransH 和 TransR，统称为基于翻译的方法。

例如：TransE 基于实体和关系的分布式向量表示，将每个三元组实例(head，relation，tail)中的关系 relation 看作实体 head 到 tail 的翻译，即强制让 Head＋relation＝tail。换而言之，把 head 加 relation 翻译成 tail。

4 链路预测方法的时间复杂度分析

链路预测方法的复杂度分析如表 2 所列。

表 2 链路预测方法的复杂度分析

Table 2 Complexity analysis of link prediction method

Category	Algorithm	Complexity
基于相似性的方法	CN	$O(n^2)$
	Jaccard	$O(2n^2)$
	PA	$O(2n)$
	AA	$O(2n^2)$
	RA	$O(2n^2)$
	SI	$O(n^2)$
	SimRank	$O(n^4)$
	Local RW	$O(N\langle K\rangle^n)$
基于矩阵分解的方法	LLE	$O(\|E\|d^2)$
	Graph Factorization	$O(\|E\|d)$
	GraRep	$O(\|V\|^3)$
	Laplacian Eigenmaps	$O(\|E\|d^2)$
	HOPE	$O(\|E\|d^2)$
基于随机游走的方法	DeepWalk	$O(\|V\|d)$
	Node2Vec	$O(\|V\|d)$
	GraphGAN	$O(\|V\|\log\|V\|)$
基于神经网络的方法	SDNE	$O(\|V\|\|E\|)$
	DNGR	$O(\|V\|^2)$
	GCN	$O(\|E\|d^2)$
	GNN	$O(\|E\|d^2)$
基于自定义损失函数	LINE	$O(\|E\|d)$

5 面临的挑战

2000 年 Sarukkai 首次在万维网中应用马尔可夫链进行链路预测和路径分析，20 年间，来自计算机科学、物理学、生物学等多学科领域的专家根据学科特色提出了许多不同的研究成果。这些成果不仅从理论上推动了基于特征分类链路预测的发展，还解决了许多实际应用中的问题。目前已被提出的一些挑战性问题都得到了不同程度的推进，但是仍然存在一些问题有待解决。本节总结了基于特征分类的链路预测方法未来的 7 个有待解决的问题。

(1)不同类型的网络和图。现有的网络包括同构网络、异构网络(如 DBLP，DBpedia 和 Flickr)和多层网络，输入特征有文本和图像特征，由于输入数据的结构极其不同，现有方法无法处理所有类型的数据。下一个重要的方向是设计特定的深度学习模型来处理这些不同类型的图和文本。

(2)动态网络和动态图。大多数现有方法关注于静态网络。然而，许多真实网络本质上是动态的，其中节点、边及其特征可以随着时间的推移而改变。如何对动态网络和图的演化特征进行建模，并支持模型参数的增量更新，在很大程度上仍是未解决的问题。

(3)实际应用。首先预测结果必须具有可解释性，链路预测有许多实际应用，因此对预测结果作出解释对于决策问题至关重要。例如，在医学领域，在将计算机实验转化为临床应用方面，可解释性是必不可少的。其次，在社交网络中，对用户个人信息的保护是至关重要的，如何保证高精度又不侵犯用户隐私，也是不可回避的问题。

(4)组合性。如前所述，许多现有的架构可以一起工作，如何充分利用各模型的优势，将它们组合使用，也是当前基于特征分类的链路预测有待解决的问题。

(5)语义问题。由于存在不同类型的实体，如文本、图像或视频之间的相互依赖关系，实体非常复杂，因此使得现有的方法很难测量丰富的语义和顶点之间的接近度，未来需要探索更好的方法用于测量跨模式数据之间的接近性和它们与网络结构之间的相互作用。特别是在研究边缘语义方面，在有符号的网络中，有些研究通过直接建模链接的极坐标来学习节点的特征。如何充分编码网络结构和顶点属性，对于有符号网络的嵌入仍然是一个悬而未决的问题。

(6)可扩展性和并行化。在大数据时代，真实的图会有数百万个节点和边，例如社交网络或电子商务网络。因此，如何设计具有线性时间复杂度的可扩展模型，成为一个关键的问题。此外，由于图的节点和边是相互连接的，通常需要作为一个整体进行建模，因此，如何进行并行计算是另一个关键问题。

(7)跨学科性[52]。链路预测问题吸引了各领域专家的关注，多学科交叉既带来了机遇，也带来了挑战，领域知识用来解决特定的问题，但是交叉集成领域知识可能使得模型设计更加困难。

结束语 本文综述了许多经典的和最新的链路预测技术，包括当前最流行的图神经网络链路预测技术 GNN，GCN，RNN 和 RL。本文提出了一种新的分类方法，首次把分层设计思想引入算法分类中，提出了 3 层模型和 2 大类 7 个方面的划分方法，把基于特征分类的链路预测方法分为 3 层：元数据层、特征分类层和特征抽取层。其中特征抽取层包括“2 大块 7 个方面”，即把常用的链路预测算法分为 2 大块，即特征提取方法和特征学习方法，7 个方面为基于相似性的方法、基于似然分析的方法、基于概率模型的方法、基于矩阵分解方法、基于随机游走的方法、基于神经网络的方法和基于自定义损失函数的方法。本文还对这些方法的时间复杂度、优缺点进行了详细的对比，最后讨论了链路预测研究未来所面临的挑战和下一步有待解决的问题。

参 考 文 献

[1] RAMESH R S. Link prediction and path analysis using markov chains[J]. Computer Networks，2000，60(33)：377-386.

[2] LIBEN-NOWEL D，KLEINBERGN，JON K. The link-prediction problem forsocial networks[J]. Journal of The American Society For InformationScience and Technology，2007，58(7)：1019-1031.

[3] LV L L. Link prediction on complex networks [J]. Journal of University of Electronic Science and Technology of China，2010，39(5)：651-661.

[4] HASAN M A，ZAKI M. A survey of link prediction in social networks[J]. Social Network Data Analytics，2011，40(5)：243-275.

[5] LV L L，REN X L，ZHOU T. Network link prediction：concepts and frontiers [J]. Communication of China Computer Society，2016，12(4)：12-21.

[6] VALVERDE R J，LOPES A A. Exploiting behaviors of communities of twitter users for link prediction[J]. Soc. Netw. Anal. Min.，2013，20(3)：1063-1074.

[7] LIU H，HU Z，HADDADI H. Hidden link prediction based on node centrality and weak ties [J]. Europhys Lett.，2013，101(1)：65-75.

[8] YANG S H，LONG B，SMOLA A，et al. Like like alike：joint friendship and interest propagation in social networks[C]// Proceedings of the 20th International Conference on World Wide Web. ACM，2011：537-546.

[9] CHO E，MYERS S A，LESKOVEC J. Friendship and mobility：user movement in location-based social networks[C]// Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM. 2011：1082-1090.

[10] LEO K. A new status index derived from sociometric analysis [J]. Psychometrika，1953，50(18)：39-43.

[11] ELIZABETH A L，PETTER H，MARK E N. Vertex similarity-in networks[J]. Physical Review E，2006，73(2)：121-130.

[12] JEH G，WIDOM J. SimRank：a measure of structural-context similarity[C]// Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'02). ACM，2002：538-543.

[13] KLOUMANN I M，UGANDER J，KLEINBERG J. Block models and personalized PageRank[J]. The National Academy of Sciences，2017，114(1)：33-38.

[14] SHANG M S，LÜ L，ZENG W，et al. Relevance is more significant than correlation：Information filtering on sparse data[J]. Europhys Lett.，2009，88(6)：68008p1-68008p4.

[15] NEWMAN M E J. Clustering and preferential attachment in growing networks[J]. Phys. Rev. E.，2001，64(10)：251-264.

[16] SALTON G，MCGILLM. Introduction to Modern Information Retrieval[J]. McGraw-Hill，1983，32(6)：528-536.

[17] ADAMIC L A，ADAR E. Friend and neighbors on the web[J]. Soc Networks，2003，25(6)：211-230.

[18] LIU Y Y，ZHAO C L，WANG X J. The degree-related clustering coefficient and its application to link prediction[J]. Physica A，2016，454(15)：24-33.

[19] ZHOU T，LÜ L，ZHANG Y C. Predicting missing links via local information[J]. Eur. Phys. J. B.，2009，80(71)：623-630.

[20] RAVASZ E，SOMERA A L，MONGRU D，et al. Hierarchical organization of modularity in metabolic networks[J]. Science，2002，297(12)：1551-1558.

[21] LV L，JIN C H，ZHOU T. Similarity index based on local paths for link prediction of complex network[J]. Phys. Reve.，2009，80(4)：211-223.

[22] WANG P，XU B W，WU Y R. Link predictionin social networks：the state-of-the-art[J]. Science China Information Sciences，2015，58(1)：1-38.

[23] HU R J，TANG J X，YUAN Y N. Link prediction in complex networks based on the interactions among paths[J]. Physical A，2019，510(4)：52-67.

[24] ALEXIS P，PANAGIOTIS S，YANNIS M. Fast and accurate link prediction in social networking systems[J]. Systems and Software，2012，60(9)：2119-2132.

[25] ZHU X，TIAN H，CAI S，et al. Predicting missing links via significant paths[J]. EPL，2014，108(4)：18008-18014.

[26] AARON C，CRISTOPHER M，MARK E J. Hierarchicalstructure and theprediction of missing links in networks[J]. Nature，2008，453(71)：98-101.

[27] GUIMERÁ R，SALES P M. Missing and spurious interactions and the reconstruction of complex networks[J]. PNAS，2009，106(52)：22073-22078.

[28] BEN T，PIETER A，WONG M F. Relationalmarkov networks [J]. Introduction to Statistical Relational Learning，2007，40(13)：175-200.

[29] LV L Y，ZHOU T. Link prediction in complex networks：A survey[J]. Physica A：Statistical Mechanics and Its Applications，2011，390(6)：1150-1170.

[30] MENON A K，ELKAN C. Link prediction via matrix factorization in Machine Learning and Knowledge Discovery in Databases [C]// Springer. 2011：437-452.

[31] ADITYA K M，CHARLES E. Link Prediction via Matrix Factorization[C]// Proceedings of the 2011 European Conference on Machine Learning and Knowledge Discovery in Databases. ACM，2011：437-452.

[32] KUNEGIS J，LOMMATZSCH A. Learning spectral graph transformations for link prediction [C]// Proceedings of the 26th Annual International Conference on Machine Learning. ACM，2009：561-568.

[33] CHEN Z，ZHANG W. A Marginalized Denoising Method for Link Prediction in Relational Data [C]// Proceedings of the 2014 SIAM International Conference on Data Mining. ACM，2015：64-72.

[34] DAI L Y，KONG X Z，YUAN S S，et al. Integrative graph regularized matrix factorization for drug-pathway associations analy-

sis[J]. Computational Biology and Chemistry, 2019, 78(7): 474-480.

[35] REN Y W, ZHOU J L, WANG J. Quality-Relevant Fault Monitoring Based on Locally Linear Embedding Orthogonal Projection to Latent Structure[J]. Industrial & Engineering Chemistry Research, 2019, 58(3): 1262-1272.

[36] OU M D, CUI P, PEI J, et al. Asymmetrictransitivity preserving graph embedding[C]// Proceedings of the 22nd ACMSIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2017: 1105-1114.

[37] BELKIN M, NIYOGI P. Laplacian eigenmaps and spectral techniques for embedding and clustering[C]// Advances in Neural Information Processing Systems. 2002: 585-591.

[38] CAO S, LU W, XU Q. Grarep: Learning graph representations with global structural information[C]// Proceedings of the 24th ACM International on Conference on Information and Knowledge Management. ACM, 2015: 891-900.

[39] WANG H W, WANG J L, XIE X, et al. GraphGAN: Graph representation learningwith generative adversarial nets[J]. IEEE Transaction on Knowledge and Data Engineering, 2017, 30(22): 11-19.

[40] CAI L J, XU Y B, HE T Q, et al. A New Algorithm of DeepWalk Based On Probability[J]. Journal of Physics: Conference Series, 2019, 1069(1): 130-135.

[41] GROVER A, LESKOVEC J. Node2Vec: Scalable Feature learning for Network[J]. HHS Public Access, 2016, 14(8): 855-864.

[42] WANG D, CUI P, ZHU W. Structural deep network embedding [C]// Proceedings of the 22nd International Conference on Knowledge Discovery and Data Mining. ACM, 2016: 1225-1234.

[43] FRENSNO D, CARLOS W L, SARAI M C, et al. DNGR-1 in dendritic cells limits tissue damage by dampening neutrophil recruitment[J]. Science, 2019, 6412(362): 351-356.

[44] TOMAS N K, MAX W, Variational Graph Auto-Encoders[J]. Springer, 2016, 28(3): 61-63.

[45] LUCA F, MATHIAS N, MASSIMILIANO P. Learning Discrete Structures for Graph Neural Networks[J]. ICML. 2019, 30(11): 960-973.

[46] YOU J X, REX Y, XIANG R. Graphrnn: Generating realistic graphs with deepauto-regressive models[C]// International Conference on Machine Learning. ACM, 2018: 5694-5703.

[47] MONTI F, BRONSTEIN M, BRESSON X. Geometric matrix completionwith recurrent multi-graph neural networks[J]. Advances in NeuralInformation Processing Systems, 2017, 30(4): 3697-3707.

[48] BUI L Q, PHAN A V, NGUYEN Y L H. DGCNN: A convolutional neural network over large-scale labeled graphs[J]. Neural Networks: The Official Journal of the International Neural Network Society, 2017, 108(11): 533-543.

[49] CAO D N, KIPF T. Molgan: An implicit generative model forsmall molecular graphs[J]. ICML 2018 Workshop on Theoretical Foundations and Applications of Deep Generative Models, 2018, 97(30): 181-192.

[50] TANG J, QU M, WANG M, et al. Line: Large-scale information network embedding[C]// Ternational Conference on Word Wide Web. ACM, 2015: 1067-1077.

[51] KIM Y H, YOO J, KIM S J, et al. Comparison of Scattering Cross-Sections by MCNP5 and TRANSX/TWODANT Codes in Sodium-Cooled Fast Reactor[J]. Transactions of the American Nuclear Society, 2018, 106(1): 719-722.

[52] WANG H, LE Z C, GONG X, et al. Link predicton of complex network is analyzed from the perspective of informatics[J]. Journal of Chinese Computer Systems, 2020, 41(2): 316-326.



WANG Hui, born in 1983, Ph. D student, lecturer, is a member of China Computer Federation. Her main research interests include link prediction, deep learning, AI and big data.