隐马尔可夫模型的异质网络链接预测方法研究

钱榕[1,2]，许建婷[2]，张克君[1,2]，董宏宇[1]，邢方远[1]
（1. 北京电子科技学院网络空间安全系，北京 100070；2. 西安电子科技大学计算机科学与技术学院，陕西 西安 710071）

摘 要：为了解决异质网络的结构信息和语义信息挖掘不全面的问题，针对异质网络的链接预测，提出了将基于元路径的分析方式与隐马尔可夫模型相结合的链接预测方法。考虑到聚簇可以有效地捕获异质网络的结构信息，将 k-means 算法进行改进得到基于距离均方差最小的初始聚簇中心方法，并将其应用到隐马尔可夫模型（HMM）中，设计了基于聚簇的一阶隐马尔可夫模型（C-HMM$^{(1)}$）的链接预测方法，同时提出基于聚簇的二阶隐马尔可夫模型（C-HMM$^{(2)}$）的异质网络的链接预测方法。进一步考虑数据的特征信息，提出了将最大熵模型和二阶隐马尔可夫模型相结合的链接预测方法 ME-HMM。实验结果表明，ME-HMM 比 C-HMM 方法的链接预测精确度更高，且 ME-HMM 因充分考虑到数据的特征信息比 C-HMM 的性能更加优异。

关键词：异质网络；链接预测；隐马尔可夫模型；聚簇；最大熵
中图分类号：TP181
文献标志码：A
DOI: 10.11959/j.issn.1000−436x.2022095

Research on HMM based link prediction method in heterogeneous network

QIAN Rong[1,2], XU Jianting[2], ZHANG Kejun[1,2], DONG Hongyu[1], XING Fangyuan[1]
1. Department of Cyberspace Security, Beijing Electronic Science and Technology Institute, Beijing 100070, China
2. College of Computer Science and Technology, Xidian University, Xi'an 710071, China

Abstract: In order to solve the problem that incomplete mining of structural information and semantic information in heterogeneous networks, a link prediction method combining meta-path-based analysis and hidden Markov model was proposed for link prediction of heterogeneous network. Considering that clustering could effectively capture the structural information of heterogeneous network, the k-means algorithm was improved to obtain the initial clustering center method based on the minimum distance mean square error, and it was applied to the hidden Markov model, first-order cluster hidden markov model (C-HMM$^{(1)}$) link prediction method, and a link prediction method for heterogeneous network with second-order cluster hidden Markov model (C-HMM$^{(2)}$) were designed. Further, considering the feature information of the data, a link prediction method called ME-HMM that combined the maximum entropy model and the second-order Markov model was proposed. The experimental results show that the ME-HMM has higher link prediction accuracy than the C-HMM, and the ME-HMM method has better performance than the C-HMM method because it fully considers the feature information of the data.

Keywords: heterogeneous network, link prediction, hidden Markov model, clustering, maximum entropy

0 引言

随着网络信息技术的迅速发展，人类社会进入复杂网络时代，人们的生产和生活越来越依赖于以Internet、WWW、通信网络、社会关系网络、经济网络等为代表的复杂网络系统的安全可靠和有效运行。异质网络从拓扑结构上看拥有不同类型的顶点和不同类型的边，更加贴近于客观世界。因此人

收稿日期：2022−01−10；**修回日期：**2022−04−07
基金项目：国家重点研发计划基金资助项目（No.2018YFB1004101）
Foundation Item: The National Key Research and Development Program of China (No.2018YFB1004101)

们对异质网络进行研究具有现实意义和挑战性。

链接预测是复杂网络研究的一个重要方向，旨在根据已知的网络结构和信息，发现和还原网络中丢失的信息，或者预测节点之间未来可能存在的关系[1]。在现实生活中，链接预测已经有很多成功的应用，比如推荐系统，可以给新用户推荐兴趣相近的朋友，或给用户推荐其可能感兴趣的视频；再如疫情传播动力学中，可以还原一些丢失的传播路径。随着异质网络研究的深入，元路径思想给异质网络预测链接的研究提供了新的思路。基于元路径的异质网络链接预测能够考虑不同节点类型及节点之间的关系，更好地提取网络中不同的语义信息，从而大大提高网络链接预测方法结果的精确度。近年来，更多的研究学者将元路径融入异质网络的研究中，如将元路径权重融合到异质网络的表征学习中[2]，将元路径与图神经网络结合[3]采样异质邻居节点，使用元路径高效提取异质网络的语义信息[4-5]等。因此，将元路径与隐马尔可夫模型（HMM, hidden Markov model）结合并应用于异质网络的预测有着十分重要的理论意义与应用前景。

本文的主要研究工作如下。

1) 提出基于聚簇的一阶隐马尔可夫模型的链接预测方法，即 C-HMM$^{(1)}$。将 HMM 应用于链接预测中，并将数据簇的方法应用于 HMM。对 k-means 算法在确定初始聚簇中心时进行改进，得到基于距离均方差最小的方法，使簇中心同时满足到其他簇中心之间的距离之和最大以及与其他所有的簇中心之间距离的均方差最小。

2) 提出基于聚簇的二阶隐马尔可夫模型的链接预测方法，即 C-HMM$^{(2)}$，并分析了 C-HMM$^{(2)}$在链接预测上的有效性。C-HMM$^{(2)}$在考虑模型当前状态的基础上，又合理地考虑了概率和模型历史状态之间的关联性，提高了链接预测的准确率。

3) 提出基于最大熵（ME, maximum entropy）的 C-HMM$^{(2)}$的链接预测方法，即 ME-HMM。通过最大熵模型，在链接预测中加入训练数据的特征信息；同时通过 C-HMM$^{(2)}$考虑状态转移概率和观测值输出概率和模型历史状态之间的关联性，进一步提高了链接预测的准确率。

4) 对本文提出的方法与已有的链接预测方法进行实验对比分析。实验结果表明，本文提出的方法优于已有的链接预测方法。

本文方法的研究框架如图 1 所示，其中 C-HMM$^{(1)}$与 C-HMM$^{(2)}$统称为 C-HMM 方法。

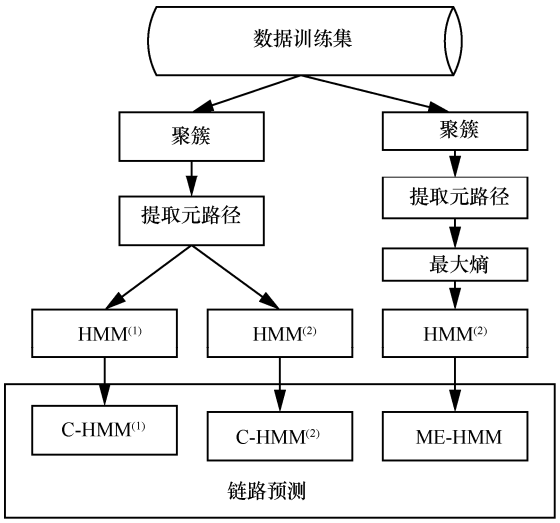


图 1　本文方法的研究框架

1　相关工作

基于异质网络中的链路预测问题引起了学者们的广泛关注，目前基于元路径的链接预测研究十分火热。

1.1　异质网络

异质网络[6]是一个带有对象类型映射函数 $\phi:\nu\to A$ 和链接类型映射函数 $\psi:\varepsilon\to R$ 且 $|A|+|R|>2$ 的有向图 $G=(\nu,\varepsilon)$，其中 ν 和 ε 分别表示网络的节点集合和链接的边集合，A 和 R 分别表示网络中的节点类型集合和链接类型集合。本文选用的 DBLP（digital bibliography & library project）数据集中选取了论文、作者、会议 3 种类型节点，以及作者与论文之间的撰写关系、论文与会议之间的发表关系等。

1.2　元路径

元路径[7]是在异质网络模式上链接两类节点的路径。元路径 P 定义为 $A_1\xrightarrow{R_1}A_2\xrightarrow{R_2}\cdots\xrightarrow{R_l}A_{l+1}$ 的形式，它表示节点 A_1 和 A_{l+1} 之间的复合关系 $R=R_1\circ R_2\circ\cdots\circ R_l$，其中，$A_i$ 为节点类型，R_i 为关系类型，$\circ$ 为关系间的复合算子。在 DBLP 数据集中，作者与作者合作撰写论文关系的元路径可以表示为“作者(A)$\xrightarrow{\text{撰写}}$论文(P)$\xrightarrow{\text{被撰写}}$作者(A)”，如果作者与论文、论文与作者之间没有任何其他的链接关系，则该元路径可以简化为“APA”，语义信息为与指定作者合作的作者。图 2 为异质网络 DBLP 的常见元路径。

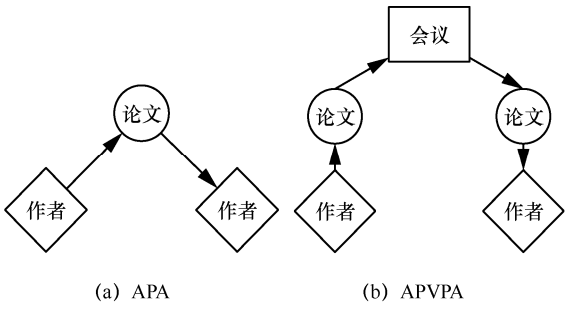


图 2 异质网络 DBLP 的常见元路径

1.3 链路预测技术

链路预测技术是研究异质网络中重要的研究方向之一。郭振宏等[8]提出对不同元路径综合网络的拓扑特征获得异质和同质数据，再通过逻辑回归作为链路预测模型。韩忠明等[9]提出基于动态网络表示的链接预测（DNRLP, dynamic network representation based link prediction）模型，基于连接强度的随机游走算法模拟网络的动态信息扩散，对新时刻下得到的节点表示通过度量相似度得到预测结果。刘大伟等[10]提出将局部共同邻居集合根据全局最短路径信息进行建模的链路预测算法。董鑫等[11]提出基于 Boosting 的异质信息网络链路预测方法，通过选择样本择优的方式和增设阈值分别对 Boosting 算法的训练速度和防止过拟合方面进行改进，并通过集成学习的思想改进链路预测性能。赵妍等[12]提出挖掘有效元路径来生成带节点属性的子图，用子图代表被预测链路，用图核方法计算子图相似性，再训练支持向量机得到预测结果。孙诚等[13]提出基于共同邻居（CN, common neighbor）、路径和随机游走的 8 种常用链路预测指标的线性组合作为度量指标，并找到较好的优化参数，提出相应的神经网络模型。黄立威等[14]分析了异质信息网络不同元路径上不同类型节点和关系的不同语义，并通过不同元路径上节点之间的连接概率进行链接预测。

本文选用易于扩展、可用于大型数据集且最常用的基于相似性指标的 CN 方法[15-16]、能够获得大量节点之间的上下文信息的重启随机游走（RWR, restart in random walks）方法[17]、单一的 HMM 方法[18]，以及近些年学者提出的针对不同类型对象间来凝结关系而重构异质网络的 BRLinks 方法[19]、通过将异质网络划分多通道再对其进行图卷积网络学习网络节点的向量表示的 MDGCN（multichannel deep graph convolutional network）方法[20]、将异质网络关系预测视为 PU 学习问题的 PURP（positive and unlabeled relationship link prediction）方法[21]，在此只简单介绍 CN 和 RWR 方法。

1) CN 方法

CN 是最常见的相似性指标之一。CN 指的是 2 个节点之间的公共邻居节点。公共邻居节点数目越多，这 2 个节点就越有可能产生链接。若节点 x 的邻居节点集合是 $\Gamma(x)$，那么节点 x 和节点 y 的相似性指标 CN 为

$$S_{xy}^{\mathrm{CN}} = \left|\Gamma(x) \cap \Gamma(\mathrm{y})\right| \tag{1}$$

2) RWR 方法

RWR 是在随机游走方法的基础上改进得到的，其主要原理是从图中的某一节点开始出发，随机选择该节点相邻节点的一个或返回开始的节点。RWR 方法有一个重启概率 α，而 $1-\alpha$ 则表示移动到相邻节点的概率，经过多次迭代达到稳定状态之后结束。RWR 的定义式为

$$\boldsymbol{r}_t = \alpha W \boldsymbol{r}_t + (1-\alpha)\boldsymbol{e}_t \tag{2}$$

其中，$\boldsymbol{r}_t = [r_{i,j}]$ 是 $n \times 1$ 的得分向量，$r_{i,j}$ 是节点 j 到节点 i 的相关度得分；$\boldsymbol{e}_t$ 是 $n \times 1$ 的初始向量，第 i 个元素为 1，其他为 0。

2 基于聚簇隐马尔可夫的异质网络链接预测

2.1 隐马尔可夫模型

HMM 常用来描述一个含有隐含未知参数的马尔可夫过程。在 HMM 中，状态是不直接可见的，可以通过观测序列的随机过程表现出来。HMM 可以用 S、O、$\boldsymbol{\pi}$、$\boldsymbol{A}$、$\boldsymbol{B}$ 这 5 个元素表示。其中，S 为隐含状态，O 为可观测状态，$\boldsymbol{\pi}$ 为初始状态概率矩阵，$\boldsymbol{A}$ 为隐含状态转移概率矩阵，$\boldsymbol{B}$ 为观测状态转移概率矩阵。HMM$^{(1)}$如图 3 所示。HMM 在研究中被广泛应用，如用 HMM 进行食品安全风险评估[22]、进行人脸特征标注与识别[23]等。

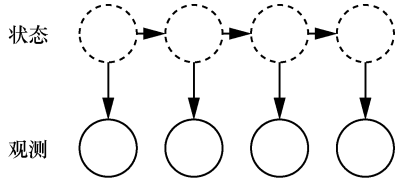


图 3 HMM$^{(1)}$

2.2 基于 C-HMM$^{(1)}$的链接预测

1) 数据聚簇

由 k-means 算法的基本思想[24]可知，选到合理

节点作为初始簇中心的可能性比较小，算法的复杂性也相对增加，不一定能得到理想的聚簇结果。本文了解到有学者通过聚簇分析理论提出了层次和密度聚簇分析方法的航迹关联算法[25]，提高了目标数量多且相互位置较近时航迹关联的准确性；以及为了改进空间聚簇算法的效率提出了距离代价函数的概念，利用距离代价最小准则，设计了一个新的 k 值优化算法[26]，对空间聚簇算法 k-means 算法和 k-中心法进行改进等。因此本文也尝试通过提高聚簇的性能，提出一种基于距离均方差最小的初始聚簇中心方法，思想如下。

一般用对象之间的距离表示相似度[27]，首先根据数据集中节点的相似性矩阵（计算式如式(3)所示），得到具有最大距离的 2 个节点，作为 2 个聚簇的初始簇中心。

$$\begin{pmatrix} d_{(1,1)} & \cdots & d_{(1,j)} & \cdots & d_{(1,N)} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ d_{(i,1)} & \cdots & d_{(i,1)} & \cdots & d_{(i,1)} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ d_{(N,1)} & \cdots & d_{(N,1)} & \cdots & d_{(N,1)} \end{pmatrix} \tag{3}$$

其中，$d_{(i,j)}$ 表示节点 i 和节点 j 之间的距离，其计算式为

$$d_{(i,j)} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} p_{Iij}\log\frac{p_{Iij}}{p_{Jij}} + \sum_{i=1}^{n}\sum_{j=1}^{n} p_{Jij}\log\frac{p_{Jij}}{p_{Iij}}}{2n} \tag{4}$$

其中，p_{Iij} 表示第 I 个节点从状态 s_i 到状态 s_j 的转移概率。2 个节点之间的相似性程度越高，其距离值 d 越小。其余聚簇中心的确定方式如下：假设需要聚簇的总个数为 k，已经得到的初始簇中心的个数为 n，那么第 $n+1$ 个待确定的簇中心与其他已得到的簇中心的距离为

$$E = \sum_{i=1}^{n}\frac{d_{i,n+1}}{n} \tag{5}$$

其中，$d_{i,n+1}$ 表示第 $n+1$ 个簇中心与第 i 个簇中心之间的距离。同时，第 $n+1$ 个待确定的簇中心节点需满足该簇中心到其他所有簇中心之间的距离之和最大，且和其他所有簇中心之间距离的均方差最小，也就是

$$\begin{cases} \max\left[\sum_{i=1}^{n} d_{i,n+1}\right] \\ \min\left[\sqrt{\sum_{i=1}^{n}(d_{i,n+1}-E)^2}\right] \end{cases} \tag{6}$$

常用平方误差作为准则函数，即

$$E = \sum_{i=1}^{n}\sum_{p\in C_i}|p-m_i|^2 \tag{7}$$

其中，C_i 表示第 i 个聚簇，p 表示 C_i 中的一个数据节点($p\in C_i$)，m_i 表示 C_i 的簇中心。

采用改进 k-means 进行聚簇，聚簇算法如算法 1 所示。

算法 1　聚簇算法

输入　n 个已标记的数据训练集，聚簇数目 k

输出　满足函数收敛的 k 个聚簇

① 用最大似然估计（MLE, maximum likelihood estimate）中的统计公式计算已标记的数据训练集的状态转移概率，计算训练集的初始状态概率，训练 HMM；

② 用式(4)计算节点之间的距离，构造节点相似性矩阵；

③ 利用初始簇中心算法，初始化 k 个聚簇中心；

④ 将数据训练集中的节点分配给与簇中心距离最近的簇；

⑤ 计算簇的平均值，更新各簇的簇中心；

⑥ 若准则函数（式(7)）未收敛，或仍有节点需要进行重新分配，则重复执行步骤④；否则结束算法，输出聚簇结果。

2) 基于聚簇一阶隐马尔可夫模型的链接预测算法

采用 DBLP 数据集引入元路径的思想，抽取节点之间的关系，预测某个作者与某个会议是否存在链接关系或是否将会发生链接关系。C-HMM(1)算法主要分为以下 5 个步骤。

① 在保留语义信息的基础上，去除冗杂数据，提取作者、论文、会议信息。

② 提取异质网络中的作者信息，采用改进的 k-means 进行聚簇分析。

③ 挖掘异质网络中节点之间的相关性，提取各聚簇对应的元路径。

④ 对每个聚簇分别进行训练，得到状态转移

概率矩阵。

⑤ 对预测节点所在的聚簇使用 Viterbi 算法，选择概率最大的结果作为链接预测的最终结果。

C-HMM$^{(1)}$算法的流程如图 4 所示。

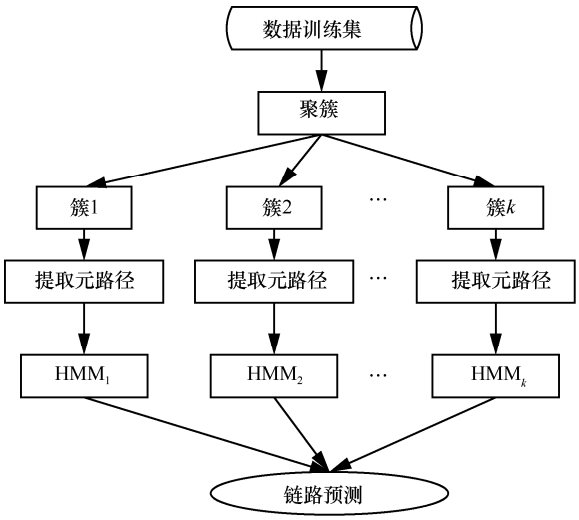


图 4 C-HMM$^{(1)}$算法的流程

将数据聚簇，从而得到多个簇，分别训练 HMM 中的 3 个参数 $\boldsymbol{\pi}$、$\boldsymbol{A}$、$\boldsymbol{B}$。运用 MLE 算法训练数据集并对 HMM 的参数进行学习，其参数估计式为

$$\gamma_t=\frac{\alpha_t(i)\beta_t(i)}{\sum_{j=1}^{N}\alpha_t(j)\beta_t(j)} \tag{8}$$

$$\varepsilon_t(i,j)=\frac{\alpha_t(i)a_{ij}b_j(o_{t+1})\beta_{t+1}(j)}{\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_t(i)a_{ij}b_j(o_{t+1})\beta_{t+1}(j)} \tag{9}$$

$$a_{ij}=\frac{\sum_{t=1}^{T-1}\varepsilon_t(i,j)}{\sum_{t=1}^{T-1}\gamma_t(i)} \tag{10}$$

$$b_j(k)=\frac{\sum_{t=1}^{T}\gamma_t(j)\delta(o_t,v_k)}{\sum_{t=1}^{T}\gamma_t(i)},\delta(o_t,v_k)\begin{cases}1,o_t=v_k\\0,o_t\neq v_k\end{cases} \tag{11}$$

从数据集训练出 HMM 的参数后，使用 Viterbi 算法来预测待预测节点的元路径。将 δ_t 定义为在时刻 t 且状态为 i 的所有路径 $(i_1,i_2,\cdots,i_t)$ 中的概率最大值，其计算方法为

$$\varepsilon_t=\max_{i_1,i_2,\cdots,i_{t-1}}P\left(i_t=i,i_{t-1},\cdots,i_1,o_t,\cdots,o_1\mid\lambda\right) \tag{12}$$

递推式为

$$\delta_{t+1}=\max_{1\leqslant j\leqslant N}\left[\delta_t(j)a_{ij}\right]b_i(o_{t+1}) \tag{13}$$

定义在时刻 t 且状态为 i 的所有路径中概率最大路径的第 $t-1$ 个节点的计算式为

$$\psi_t=\arg\max_{1\leqslant j\leqslant N}\left[\delta_{t-1}(j)a_{ij}\right],i=1,2,\cdots,N \tag{14}$$

2.3 基于 C-HMM$^{(2)}$的链接预测

HMM$^{(2)}$如图 5 所示。HMM$^{(2)}$增加了观测值输出的约束条件，相应地，链接预测的准确率也有较大的提高。

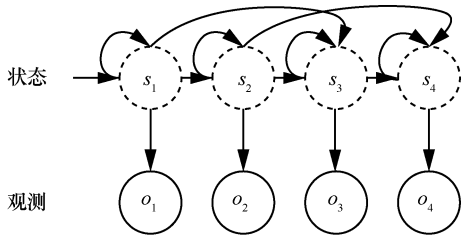


图 5 HMM$^{(2)}$

HMM$^{(2)}$ 可以被定义为七元组 $(S,O,\boldsymbol{\pi},\boldsymbol{A}_1,\boldsymbol{A}_2,\boldsymbol{B}_1,\boldsymbol{B}_2)$，七元组中 S、O、$\boldsymbol{\pi}$ 和 C-HMM$^{(1)}$中的含义一样，其他需要重点说明的介绍如下。

隐含状态转移概率矩阵 $\boldsymbol{A}_1$ 和 $\boldsymbol{A}_2$，$\boldsymbol{A}_1=(a_{ij})_{N\times N}$，$\boldsymbol{A}_2=(a_{ijk})_{N\times N\times N}$，其中

$$a_{ij}=P\left(q_t=s_j\mid q_{t-1}=s_i\right),1\leqslant i,j\leqslant N$$

$$a_{ijk}=P\left(q_{t+1}=s_k\mid q_t=s_j,q_{t-1}=s_i\right),1\leqslant i,j,k\leqslant N$$

观测状态转移概率矩阵 $\boldsymbol{B}_1$ 和 $\boldsymbol{B}_2$，$\boldsymbol{B}_1=\left[b_j(o_t)\right]_{N\times N}$，$\boldsymbol{B}_2=\left[b_{ij}(o_t)\right]_{N\times N\times N}$，其中

$$b_j(o_t)=P\left(o_t=o_l\mid q_t=s_j\right),1\leqslant j\leqslant N,1\leqslant l\leqslant M$$

$$b_{ij}(o_t)=P\left(o_t=o_l\mid q_t=s_j,q_{t-1}=s_i\right),$$
$$1\leqslant i,j\leqslant N,1\leqslant l\leqslant M$$

HMM$^{(2)}$与 HMM$^{(1)}$一样，在链接预测中用来解决学习问题和解码问题。

C-HMM$^{(2)}$算法利用数据集中的训练样本，首先用 C-HMM$^{(2)}$中的 MLE 算法学习 HMM$^{(2)}$的参数，然后用 C-HMM$^{(2)}$中的 Viterbi 算法获取链接预测的最大概率的状态序列。基于 C-HMM$^{(2)}$的链接预测与基于 C-HMM$^{(1)}$的链接预测类似：收集处理异质网络数据，对其聚簇；提取元路径；应用 C-HMM$^{(2)}$。

1) C-HMM$^{(2)}$的 MLE 算法模型

初始状态概率计算式为

$$\pi_i = \frac{\text{Init}(i)}{\sum_{j=1}^{N}\text{Init}(j)}, 1 \leqslant i \leqslant N \tag{15}$$

其中，$\text{Init}(i)$ 表示在数据集训练样本中初始状态 s_i 的序列数目；$\sum_{j=1}^{N}\text{Init}(j)$ 表示在数据集训练样本中所有状态的序列数目之和。

隐含状态转移概率计算式为

$$a_{ij} = \frac{c_{ij}}{\sum_{k=1}^{N} c_{ik}}, 1 \leqslant i, j \leqslant N \tag{16}$$

$$a_{ijk} = \frac{c_{ijk}}{\sum_{u=1}^{N} c_{iju}}, 1 \leqslant i, j, k \leqslant N \tag{17}$$

其中，c_{ij} 表示数据集训练样本中，在时刻 t 状态为 s_i、时刻 $t+1$ 状态转换为 s_j 的次数；c_{ijk} 表示数据集训练样本中，在时刻 $t-1$ 状态为 s_i、时刻 t 状态为 s_j、时刻 $t+1$ 转换为状态 s_k 的次数；$\sum_{u=1}^{N} c_{iju}$ 表示在数据集训练样本中，在时刻 $t-1$ 状态为 s_i、时刻 t 状态为 s_j，时刻 $t+1$ 转换到所有状态的次数之和。

观测状态转移概率计算式为

$$b_j(o_k) = \frac{E_j(o_k)}{\sum_{i=1}^{M} E_j(o_i)}, 1 \leqslant i \leqslant N \tag{18}$$

$$b_{ij}(o_k) = \frac{E_{ij}(o_k)}{\sum_{u=1}^{M} E_{ij}(o_u)}, 1 \leqslant i, j \leqslant N, 1 \leqslant k \leqslant M \tag{19}$$

其中，$E_j(o_k)$ 表示数据集训练样本中，在时刻 t 状态为 s_j 时可观测状态为 o_k 的次数；$E_{ij}(o_k)$ 表示数据集训练样本中，在时刻 $t-1$ 状态为 s_i、时刻 t 状态为 s_j 时可观测状态为 o_k 的次数；$\sum_{u=1}^{M} E_{ij}(o_u)$ 表示数据集训练样本中，在时刻 $t-1$ 状态为 s_i、时刻 t 状态为 s_j 时所有观测状态为 o_k 的次数之和。

接着用C-HMM$^{(2)}$的Viterbi算法解决解码问题：利用递归调用在给定条件下求最佳的状态序列 $Q^* = (q_1^*, q_2^*, \cdots, q_T^*)$。定义 $\delta_t(i,j)$ 是在时刻 t 路径为 $q_1, q_2, \cdots, q_t\ (q_{t-1} = s_i, q_t = s_j)$ 且观测状态序列为 $o_1, o_2, \cdots, o_t$ 的最大概率，其计算式为

$$\delta_t(i,j) = \max_{q_1, \cdots, q_{t-2}} P(q_1, q_2, \cdots, q_{t-1} = s_i, q_t = s_j, o_1, o_2, \cdots, o_t | \lambda), 1 \leqslant i, j \leqslant N, 2 \leqslant t \leqslant T \tag{20}$$

其中，$\varphi_{t+1}(j,k)$ 是记录节点的数组。在时刻 $t+1$ 时，有

$$\delta_{t+1}(j,k) = \max\left[\delta_t(i,j)a_{ijk}\right] b_{ij}(o_{t+1}), \quad 1 \leqslant i, j \leqslant N, 2 \leqslant t \leqslant T-1 \tag{21}$$

C-HMM$^{(2)}$的 Viterbi 算法流程如算法 2 所示。

算法 2　C-HMM$^{(2)}$的 Viterbi 算法

输入　观测值序列 $O=(o_1, o_2, \cdots, o_T)$，HMM$^{(2)}$ = $(\boldsymbol{\pi}, \boldsymbol{A}_1, \boldsymbol{A}_2, \boldsymbol{B}_1, \boldsymbol{B}_2)$

输出　$P(Q|O, \boldsymbol{\pi})$ 的最佳状态序列 $Q^* = (q_1^*, q_2^*, \cdots, q_T^*)$

1) 初始化方法

$\delta_2(i,j) = \pi_i a_{ij} b_i(O_1) b_{ij}(O_2), 1 \leqslant i, j \leqslant N$

$\varphi_2(i,j) = 0, 1 \leqslant i, j \leqslant N$

2) 递归过程

where$(1 \leqslant j, k \leqslant N, 2 \leqslant t \leqslant T-1)$

$\delta_{t+1}(j,k) = \max\left[\delta_t(i,j)a_{ijk}\right] b_{ij}(o_{t+1})$;

$\varphi_{t+1}(j,k) = \arg\max_{1 \leqslant i \leqslant N}\left[\delta_t(i,j)a_{ijk}\right], 1 \leqslant i, j \leqslant N$

3) end

$P^* = \max_{1 \leqslant i, j \leqslant N}\left[\delta_T(i,j)\right]$

$q_{T-1}^*, q_T^* = \arg\max_{1 \leqslant i, j \leqslant N}\left[\delta_T(i,j)\right]$

4) 最佳状态序列为

$q_{t-1}^* = \varphi_{t+1}(q_t^*, q_{t-1}^*), t = T-1, T-2, \cdots, 2$

3　基于最大熵隐马尔可夫模型的异质网络链接预测

最大熵原理的主要思想是在只掌握知识的部分信息且是未知知识时，应该选择对已知的知识熵最大的概率分布。为了进一步提高链接预测性和准确率，在最大熵模型中加入了特征信息对链接预测中状态转移的影响，提出了 ME-HMM 方法，如图 6 所示。ME-HMM 方法处理流程大致概括为：对异质网络数据集聚簇并提取元路径；进行最大熵处理提取特征；应用 C-HMM$^{(2)}$。

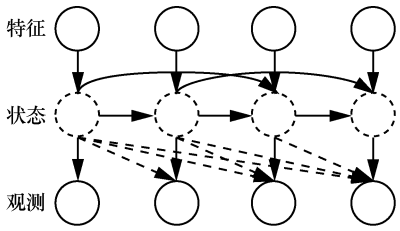


图 6 ME-HMM 方法

3.1 特征−状态转移概率矩阵的训练

首先在数据集中选择合适的特征信息，考虑这些特征信息对链接预测模型状态转移的影响，将其加入链接预测模型的训练中，训练特征−状态转移概率矩阵。

1) 提取特征信息。根据本文采用的数据集，选取一些特征信息，如是否包含人名、是否都是大写字母、是否以“.”结尾、是否以“a”开头、是否以“v”开头。

有些特征可能是多个数据的特征，假设特征集是 $T=\{T_1,T_2,\cdots,T_L\}$，其长度是 L；状态集是 S，那么为了表示其是否具有某些特征信息，定义以下二值函数。

如果特征信息 l 只影响一个状态 s_j，则应用

$$f_{l,j}(o_t,s_t)=\begin{cases}1, o_k\text{具有特征信息}l,\text{且}s_t=s_j,o_l=o_k\\0,\text{其他}\end{cases}$$

如果特征信息 l 同时影响状态 s_j 和状态 s_i，则应用

$$f_{l,i,j}\left(o_t,s_t,s_{t-1}\right)=\begin{cases}1, o_k\text{具有特征信息}l,\text{且}s_t=s_j,s_{t-1}=s_i\\0,\text{其他}\end{cases}$$

2) 计算特征−状态转移概率矩阵。当特征信息 l 只影响状态 s_j 时，特征−状态转移概率矩阵为 $\boldsymbol{M}=\{M_{i,j}\}$，其中 $M_{i,j}$ 是从状态 i 到状态 j 的转移概率，且满足

$$\sum_j M_{i,j}=1, 1\leqslant l\leqslant N_{\mathrm{F}}, 1\leqslant j\leqslant N_{\mathrm{S}}$$

其中，N_{F} 表示特征的个数，N_{S} 表示状态的个数。

当特征信息 l 同时影响状态 s_j 和 s_i 时，特征−状态转移概率矩阵为 $\boldsymbol{M}=\left\{M_{l,i,j}\right\}$，其中 $M_{l,i,j}$ 是从状态 l 到状态 i、状态 j 的状态转移概率，且满足

$$\sum_j M_{l,i,j}=1, 1\leqslant l\leqslant N_{\mathrm{F}}, 1\leqslant i,j\leqslant N_{\mathrm{S}}$$

训练特征−状态转移概率矩阵的步骤如下。

① 计算数据集中每个特征−状态的平均值。假设可观测状态的长度是 m_s，第 l 个特征信息、第 j 个状态平均值的计算式为

$$F_{l,j}=\frac{1}{m_s}\sum_{r=1}^{m_s} f_{l,j}(o_t,s_t)_r$$

第 l 个特征信息、第 i 个状态、第 j 个状态平均值的计算式为

$$F_{l,i,j}=\frac{1}{m_s}\sum_{r=1}^{m_s} f_{l,i,j}\left(o_t,s_t,s_{t-1}\right)_r$$

② 根据 GIS 参数估计算法，得出特征−状态概率矩阵。

3.2 结合最大熵的状态转移概率

在最大熵的 C-HMM[(2)]中，时刻 $t+1$ 的状态 s_k 的转移概率由时刻 t 的状态 s_j、时刻 $t-1$ 的状态 s_i，以及在时刻 t 得到的特征−状态转移概率共同决定。

当第 l 个特征信息仅影响状态 s_k 时，状态 s_k 的状态转移概率的计算式为

$$P\left(s_{t+1}=s_k\left|s_t,s_{t-1},o_{t+1}\right.\right)=\frac{1}{\gamma}\left(\lambda\alpha_{i,j,k}+(1-\lambda)\sum_l\left(M_{l,k}f_{l,k}(o_{t+1},s_{t+1})\right)\right)$$

$$\gamma=\sum_k\left(\lambda\alpha_{i,j,k}+(1-\lambda)\sum_l\left(M_{l,k}f_{l,k}\left(o_{t+1},s_{t+1}\right)\right)\right)$$

其中，$\alpha_{i,j,k}$ 是时刻 $t-1$ 状态为 s_i、时刻 t 状态为 s_j 时，时刻 $t+1$ 状态是 s_k 的状态转移概率，γ 是归一化常数。

当第 l 个特征信息同时影响 s_k 和 s_i 时，状态 s_k 的状态转移概率的计算式为

$$P\left(s_{t+1}=s_k\left|s_t,s_{t-1},o_{t+1}\right.\right)=\frac{1}{\gamma}\left(\lambda\alpha_{i,j,k}+(1-\lambda)\sum_l\left(M_{l,j,k}f_{l,j,k}\left(o_{t+1},s_{t+1},s_t\right)\right)\right)$$

$$\gamma=\sum_k\left(\lambda\alpha_{i,j,k}+(1-\lambda)\sum_l\left(M_{l,j,k}f_{l,j,k}(o_{t+1},s_{t+1},s_t)\right)\right)$$

其中，λ 是调节特征−状态转移概率和状态转移概率矩阵重要性的权重。

4 实验结果与分析

4.1 数据预处理

本文所采用的 DBLP 数据集为 XML 格式文件且文件较大，在此选择 SAX 解析方式来处理文件。

DBLP 数据集按年份列出了 60 000 多名作者的科研成果，包括 72 902 篇论文和 464 个会议。其中，每条数据<article>中包含<author>、<title>、<page>、<year>、<volume>、<journal>、<ee>、<url>等信息。

数据预处理首先需要提取数据中所收录的论文，每一条记录中都包含论文的作者、标题、会议等信息。论文中涉及的作者为合作关系，同时将论文的标题信息加入相关作者的属性信息中。另外，在当前数据库中检索当前作者相关的其他论文，保证作者的文本属性完整。本文只提取论文中前 3 位作者与该论文的关系。DBLP 数据预处理流程如图 7 所示。

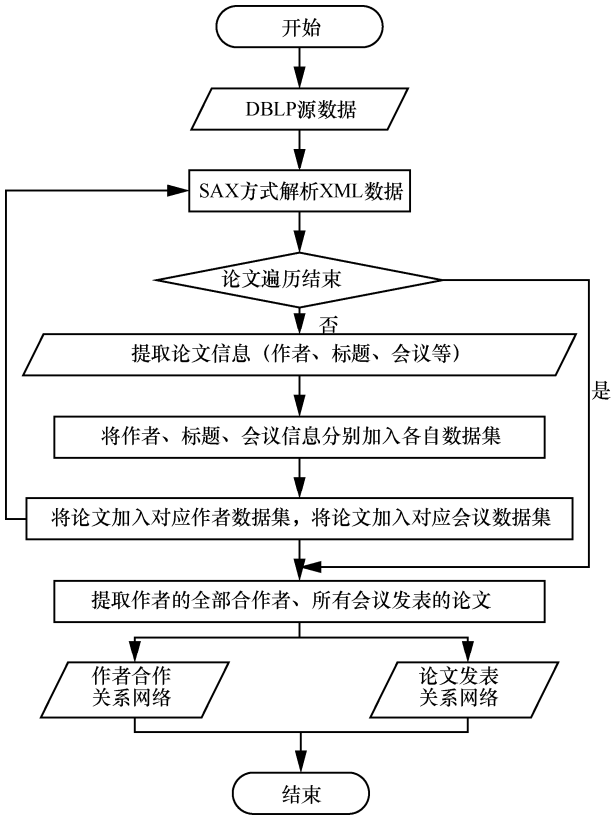


图 7　DBLP 数据预处理流程

本文在 Intel i7 10750H 6 核 12 线程 CPU 主频为 2.6 GHz 的机器上运行程序，随机抽取 90%的数据集作为训练集，剩下的 10%作为测试集。

4.2　抽取元路径

本文通过随机游走模型生成元路径来表示网络结构和节点的上下文关系，并采用 skip-gram 模型生成异质网络的节点表示，提取出待预测的 2 个节点之间的元路径作为候选集。同时通过调查大量基于元路径的研究和工作发现，异质网络中最常用的也是最有效的元路径方案是 APVPA，表示的语义信息是两位作者在同一会议发表论文的关系。考虑到本文的实验环境，本文抽取的元路径长度是 100，每个节点的游走次数是 500。部分抽取的元路径如图 8 所示。

aMirenI.Bagüés vICSNW aAlbertoDamiano vSSDBM aDennisShasha vDMDW aCarstenSapia
vDawak aGiuseppePsaila vVLDB aHectorGar-Molina vJCDL aIchiroFujinaga vJCDIL
aNingHu vJASIST aHemalataIyer vJASIST aNaoyukiTokuda vJASIST aShukYingho
vJASIST aBlumaC.Persitz vJASIS aJohnR.0ttensmann vJASIS aBryceAllen vJSIS
aBellaHassweinberg vJASIS aB.S.Manjunath vADL aSang-gooLee vDOLAP aCarlMedsker
vSIGM0DRecord aFranKEliassen vIEEEDataENG.Bull.aViswanathPoosala vVLDB
aManuelBarrenaGarcia vSSTD aPanoskalnis vSSTD aMaxJ.Egebhofer vACM-GIS
aDavidj.Russomanno vJ.Intell.Inf.Syst.aShyamalaDoraisamy vSIGIRForum
aChristopherStokoe vTREC aYvesRasolofo vTREC ajamesP.Callan vSIGIRForum
aTheop.vanderweide vNLDB aw.Mustapha-Elhadi vNLDB avéroniquePlihon vCAiSE
aH.Heijes vCAiSE aFadip.Deek vCAiSE aIngevandeWeerd vCAiSE aHenderiKAlexProper
vDataKnowl.Eng.aAmedeoCappelli vDataKnowl.Eng.aNobuyoshiMiyazaki vD00D
aHandjürgen0hlbach vERWorkshops aAgnarRenplen vERWorkshops aClaudTäubner
vICDEWorkshops aAngeloChianese vDEXAWorkshops aCyrillabbé vDEXA
aKunihikoKaneko vDEXAWorkshop aAntonioConte vDEXAWorkshop aF ionnMurtagh
vSIGM0DRecord alomsooWong vDBPL ajianwenSu vIEEETrans.Knowl.DataEng.
aNicalaLeone vFMLD0 aHolgerRiedel vEC00PWorkshopon0bject-0rientedDatabases
aVanjaosifovski vWWW aDarrenMundy vWWW aHieoshiNakagawa vDEXAWorkshops
aKhedijaAeour viCDMWorkshops aChenxi vICDMWorkshops aGangChen vWAIM
aYanqiuzhang vIDEAS akeirB.Davis vIDEAS aALberto0.Mendelzon vCASC0N
aYann-GaëlGéhéneuc vCASCON aAndyY.Mao vCASAON aDouglasR.Bloch vCASON

图 8　部分抽取的元路径

4.3　实验评价指标

采用计算测试集的精确度（Precision）、召回率（Recall）、F 值（F-Measure）3 种指标作为衡量本文提出的链接预测模型的指标。本文将链接预测的结果分为有链接和无链接 2 种，将有链接的数据定义为正例，无链接的数据定义为负例。针对以上 2 种不同情况，分别计算 Precision 和 Recall。总体的 Precision 和 Recall 分别如式(22)和式(23)所示。

$$\text{Precision}=\frac{\text{Precision}_{\text{E}}+\text{Precision}_{\text{N}}}{2} \tag{22}$$

$$\text{Recall}=\frac{\text{Recall}_{\text{E}}+\text{Recall}_{\text{N}}}{2} \tag{23}$$

其中，$\text{Precision}_{\text{E}}$ 和 $\text{Precision}_{\text{N}}$ 分别为正例和反例的准确率，Recall_{E} 和 Recall_{N} 分别为正例和反例的召回率。精确率表示预测正确的链接数目占预测为有链接（无链接）的数目的比值；召回率又称查全率，表示预测正确的链接数目占实际存在链接（无链接）的数目的比值。采用 Precision 和 Recall 这 2 个指标的调和平均值 F-Measure 作为评估标准，其计算式为

$$\text{F-Measure}=\frac{\left(\beta^2+1\right)\text{Precision}\,\text{Recall}}{\beta^2\,\text{Precision}+\text{Recall}} \tag{24}$$

其中，β 是衡量精确度和召回率重要性的权值，本文将 β 的值设为常数 1。

4.4 实验结果及分析

本文选择 CN、RWR、HMM、BRLinks、MDGCN、PURP 方法来进行对比实验。首先确定聚簇的数目，以达到最佳的链接预测性能。通过多次改变聚簇数目进行实验测试，得到链接预测的总精确度与聚簇数目的关系，如图 9 所示。

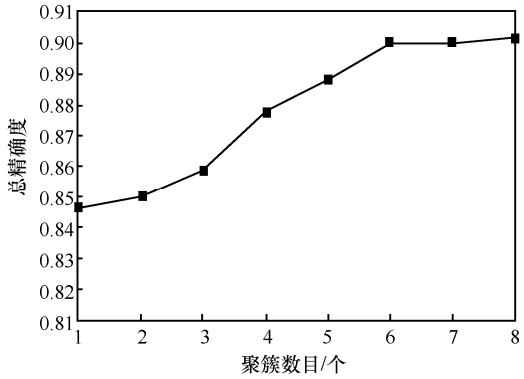


图 9 链接预测的总精确度与聚簇数目的关系

从图 9 中可以看出，链接预测的总精确度随着聚簇数目的增加而提高，但当聚簇数目为 6 时，精确度不再有明显的提高。其原因一方面是有些聚簇训练得到的 HMM 的链接预测准确率大致相同；另一方面是对于同一条元路径来说，不同模型的链接预测结果可能相同；除此之外，聚簇数目多的簇中可能包含的训练数据不多，通过这些簇得到的链接预测模型很少作为最终的链接预测结果。因此，这些新增簇对 HMM 链接预测的结果影响不大。同时，本节又对不同聚簇数目下的程序运行时间做了测试，如图 10 所示。

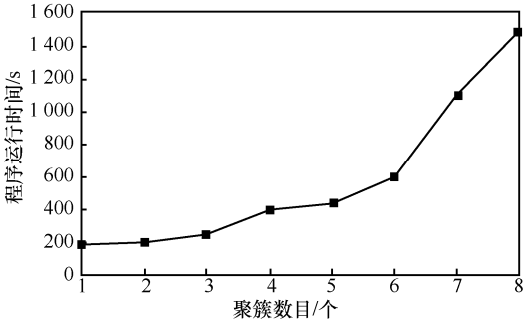


图 10 不同聚簇数目下的程序运行时间

从图 10 中可以看出，聚簇数目越多，程序运行时间也就更多。图 9 和图 10 的实验数据表明，本文实验选择聚簇数目为 6 的效果是最佳的，因此以下对比实验均采用聚簇数目为 6。

1) C-HMM$^{(1)}$和 HMM 链接预测的实验对比

选择最佳聚簇数目 6 进行 C-HMM$^{(1)}$和 HMM 这 2 种方法的实验。C-HMM$^{(1)}$和 HMM 链接预测总精确度比较如图 11 所示。

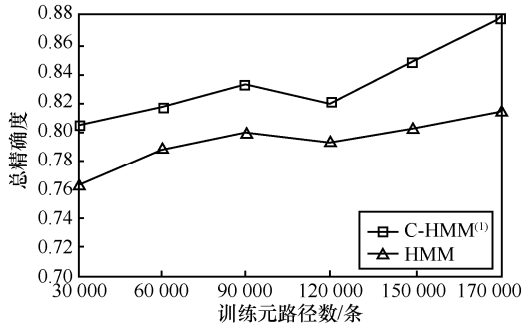


图 11 C-HMM$^{(1)}$和 HMM 链接预测总精确度比较

从图 11 中可以看出，C-HMM$^{(1)}$总精确度比单独使用 HMM 精确度高。当训练元路径数为 170 000 时，预测类型节点链接预测精确度比较如表 1 所示，进一步说明聚簇能更有效地捕捉异质网络的结构信息，因此 C-HMM$^{(1)}$有更好的链接预测性能。

表 1 C-HMM$^{(1)}$和 HMM 对各状态链接预测精确度比较

待预测节点	C-HMM$^{(1)}$	HMM
作者	0.836 246	0.823 226
标题	0.858 673	0.853 853
会议	0.887 625	0.835 826

2) C-HMM$^{(1)}$和 C-HMM$^{(2)}$链接预测的实验对比

训练集从 30 000 条开始，不断增加到 170 000 条，C-HMM$^{(1)}$和 C-HMM$^{(2)}$ 链接预测总精确度比较如图 12 所示。

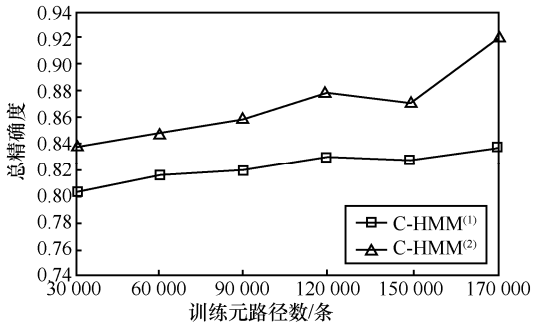


图 12 C-HMM$^{(1)}$和 C-HMM$^{(2)}$链接预测总精确度比较

从图 12 中可以看出，选取不同数量的元路径，C-HMM$^{(2)}$链接预测的总精确度都高于 C-HMM$^{(1)}$。随着训练数据的增加，C-HMM$^{(2)}$的总精确度始终较高，

这是因为随着训练数据的增加，C-HMM$^{(2)}$更加优化，识别错误的能力更强。当训练数据从 120 000 条增加到 150 000 条时，2 种方法精确度提高不大，这可能是因为此时增加的训练集和测试集中数据的匹配度较低。

以 150 000 条元路径作为训练集为例，C-HMM$^{(1)}$和 C-HMM$^{(2)}$链接预测的总精确度如表 2 所示。

表 2　C-HMM$^{(1)}$和 C-HMM$^{(2)}$链接预测的总精确度

待预测节点	C-HMM$^{(1)}$	C-HMM$^{(2)}$
作者	0.823 226	0.836 246
标题	0.853 853	0.858 673
会议	0.835 826	0.887 625

从表 2 中可以看出，基于 C-HMM$^{(2)}$的链接预测精确度高于 C-HMM$^{(1)}$的链接预测精确度，更进一步论证了 C-HMM$^{(2)}$的算法性能高于 C-HMM$^{(1)}$。

3) ME-CHMM 链接预测的实验对比

选取相关的特征信息，建立相关的特征匹配字典进行匹配。在判断是否包含人名特征信息时，使用从网上下载的外国人名进行匹配；对于是否都是大写字母、是否以“a”“v”开头、是否以“.”结尾的特征信息，不需要建立特征字典，在程序中可以通过逻辑判断的方式进行匹配。在实验中，进行交叉测试，取平均值，不断调整特征−状态转移概率和状态转移概率的相对重要程度 λ 的值。当 λ 取 0.6、训练集从 30 000 条增加到 170 000 条时，部分链接预测方法总精确度的比较如图 13 所示。

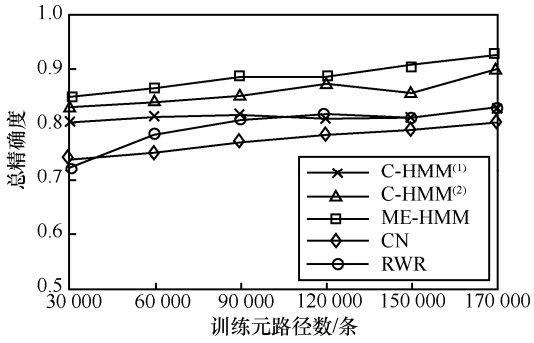


图 13　部分链接预测方法总精确度的比较

从图 13 中可以看出，随着训练数据的增加，各方法的总精确度都有不同程度的提升，其中 ME-HMM 的链接预测总精确度高于其他模型，尤其是高于 CN、RWR 和基准方法。C-HMM$^{(1)}$和 C-HMM$^{(2)}$随着训练数据的增加，链接预测的总精确度有所下降，这是由于新增加的训练集与测试集匹配度不高，但是在 ME-HMM 模型的预测中，这种情况并没有出现，这表明在 HMM$^{(2)}$中加入数据特征信息的最大熵模型对链接预测更加有效。多种链接预测方法实验比较如表 3 所示。

表 3　多种链接预测方法实验比较

链接预测方法	Precision	Recall	F-Measure
CN	0.827	0.836	0.843
RWR	0.853	0.862	0.876
HMM	0.829	—	—
BRLinks	—	—	0.873
MDGCN	0.830	—	—
PURP	—	—	0.805
C-HMM$^{(1)}$	0.853	0.883	0.903
C-HMM$^{(2)}$	0.921	0.876	0.922
ME-HMM	0.943	0.923	0.957

从表 3 中可以看出，在链接预测精确度上，C-HMM$^{(2)}$比 C-HMM$^{(1)}$更高；在召回率方面，C-HMM$^{(2)}$没有优势，但 C-HMM$^{(2)}$总的 F-Measure 比 C-HMM$^{(1)}$高，这说明 C-HMM$^{(2)}$的链接预测性能更高。ME-HMM 在进一步提高链接预测精确度的同时，也提高了链接预测的召回率。HMM、MDGCN 的精确度分别为 0.829、0.830，而本文提出的方法的精确度都在 0.853 以上；BRLinks、PURP 的 F-Measure 的值分别为 0.873、0.805，而本文提出的方法的 F-Measure 都在 0.903 以上，通过对比可以发现，本文提出的方法的 F-Measure 至少提高了 3 个百分点。由以上实验结果可知，本文提出的方法性能较为优异，其中 ME-HMM 方法的性能最好。

5　结束语

在复杂网络中，如何更加精确地对异质网络进行链接预测一直是人们研究的热点之一，据此本文提出了通过改进 k-means 算法实现基于聚簇的隐马尔可夫模型的异质网络的链接预测，并通过分析在 C-HMM$^{(2)}$中状态转移概率、观测值输出概率和模型历史状态之间的关系，可知链接预测的准确率有较大的提高。在此基础上，提出 ME-HMM，将数据的特征信息加入链接预测中。实验表明，ME-HMM 在本文中的链接预测中具有最优的性能，最大程度地提高了链接预测的准确率。但本文使用的是静态数据，没有考虑到数据的实时动态问题。但在实际生活中，数据每时每刻都在快速增长，对增量数据集的研究更符合现实网络的特征。因此，在后续的

工作中，可以更深一步地对异质网络链接预测的有关课题进行研究。

参考文献：

[1] LEE J B, ADORNA H. Link prediction in a modified heterogeneous bibliographic network[C]//Proceedings of 2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining. Piscataway: IEEE Press, 2012: 442-449.

[2] 蒋宗礼, 陈浩强, 张津丽. 基于融合元路径权重的异质网络表征学习[J]. 计算机系统应用, 2019, 28(12): 28-36.
JIANG Z L, CHEN H Q, ZHANG J L. Heterogeneous network representation learning based on fusion meta-path weights[J]. Computer Systems & Applications, 2019, 28(12): 28-36.

[3] MA Y, CHENG G Q, LIANG X X, et al. Heterogeneous graph neural networks based on meta-path[C]//Proceeding of 2020 3rd International Conference on Algorithms, Computing and Artificial Intelligence. [S.l.:s.n.], 2020: 95-98.

[4] 朱恺. 异构社交网络中基于元路径的链接预测研究[D]. 南京: 南京大学, 2020.
ZHU K. Research on meta path-based link prediction in heterogeneous social networks[D]. Nanjing: Nanjing University, 2020.

[5] 郑玉艳. 异质信息网络的语义元路径分析方法研究[D]. 北京: 北京邮电大学, 2019.
ZHENG Y Y. Research on semantic meta path analysis method of heterogeneous information networks[D]. Beijing: Beijing University of Posts and Telecommunications, 2019.

[6] 孙艺洲, 韩家炜. 异构信息网络挖掘: 原理和方法[M]. 段磊, 朱敏, 唐常杰, 等, 译. 北京: 机械工业出版社, 2017.
SUN Y Z, HAN J W. Mining heterogeneous information networks: principles and methodologies[M]. Translated by DUAN L, ZHU M, TANG C J, et al. Beijing: China Machine Press, 2017.

[7] SUN Y, HAN J, YAN X, et al. PathSim: meta path based top-k similarity search in heterogeneous information networks[J]. Proceedings of the VLDB Endowment, 2011, 4(11): 992-1003.

[8] 郭振宏, 李海峰. 异质信息网络中演员合作关系的链路预测[J]. 计算机工程, 2017, 43(1): 219-225.
GUO Z H, LI H F. Link prediction of actor cooperation relationship in heterogeneous information network[J]. Computer Engineering, 2017, 43(1): 219-225.

[9] 韩忠明, 李胜男, 郑晨烨, 等. 基于动态网络表示的链接预测[J]. 物理学报, 2020, 69(16): 332-345.
HAN Z M, LI S N, ZHENG C Y, et al. Link prediction model based on dynamic network representation[J]. Acta Physica Sinica, 2020, 69(16): 332-345.

[10] 刘大伟, 吕元娜, 余智华. 一种改进的复杂网络链路预测算法[J]. 小型微型计算机系统, 2016, 37(5): 1071-1074.
LIU D W, LV Y N, YU Z H. An improved link prediction algorithm for complex networks[J]. Journal of Chinese Computer Systems, 2016, 37(5): 1071-1074.

[11] 董鑫. 基于 Boosting 的异质信息网络链路预测方法研究[D]. 黑龙江: 哈尔滨工程大学, 2017.
DONG X. Research on boosting based method of link prediction in heterogeneous information network[D]. Heilongjiang: Harbin Engineering University, 2017.

[12] 赵妍, 赵书良, 马秋微. 基于图核的异质信息网络链路预测方法[J]. 计算机应用研究, 2021, 38(10): 3125-3130.
ZHAO Y, ZHAO S L, MA Q W. Graph kernel based link prediction in heterogeneous information network[J]. Application Research of Computers, 2021, 38(10): 3125-3130.

[13] 孙诚, 王志海. 社会网络中基于神经网络的链路预测方法[J]. 数学建模及其应用, 2017, 6(4): 10-17.
SUN C, WANG Z H. The link prediction algorithms based on neural networks in social networks[J]. Mathematical Modeling and Its Applications, 2017, 6(4): 10-17.

[14] 黄立威, 李德毅, 马于涛, 等. 一种基于元路径的异质信息网络链路预测模型[J]. 计算机学报, 2014, 37(4): 848-858.
HUANG L W, LI D Y, MA Y T, et al. A meta path-based link prediction model for heterogeneous information networks[J]. Chinese Journal of Computers, 2014, 37(4): 848-858.

[15] 王凯, 刘树新, 丁洪涛, 等. 基于共同邻居有效性的复杂网络链路预测算法[J]. 电子科技大学学报, 2019, 48(3): 432-439.
WANG K, LIU S X, YU H T, et al. Predicting missing links of complex network via effective common neighbors[J]. Journal of University of Electronic Science and Technology of China, 2019, 48(3): 432-439.

[16] 汤永新, 齐敬英. 基于共同邻居的小度节点有利链路预测算法[J]. 现代电子技术, 2021, 44(5): 37-40.
TANG Y X, QI J Y. Algorithm of predicting missing links of small promoted index via common neighbors[J]. Modern Electronics Technique, 2021, 44(5): 37-40.

[17] JIN W, JUNG J, KANG U. Supervised and extended restart in random walks for ranking and link prediction in networks[J]. PLoS One, 2019, 14(3): e0213857.

[18] DONG S L, WU Z G, SHI P, et al. Quantized control of Markov jump nonlinear systems based on fuzzy hidden Markov model[J]. IEEE Transactions on Cybernetics, 2019, 49(7): 2420-2430.

[19] 杨妮亚. 异质网络中基于元路径的链路预测方法的研究[D]. 长春: 吉林大学, 2018.
YANG N Y. Meta path-based link prediction research for heterogeneous information networks[D]. Changchun: Jilin University, 2018.

[20] 赵宇红, 吴昊. 基于图表示深度学习的异质网络链路预测研究[J]. 小型微型计算机系统, (2021-12-13)[2022-01-10].
ZHAO Y H, WU H. Link prediction in heterogeneous networks based on deep learning of graph representation[J]. Small Microcomputer Systems, (2021-12-13)[2022-01-10].

[21] 彭高婧. 基于 PU 学习的链接预测方法研究[D]. 南京: 南京邮电大学, 2018.
PENG G J. Research on link prediction method based on PU learning[D]. Nanjing: Nanjing University of Posts and Telecommunications, 2018.

[22] 崔仕颖. 基于隐马尔可夫模型的食品安全风险评估方法研究及应用[D]. 北京: 北京化工大学, 2020.
CUI S Y. Research and application of food safety risk assessment al-

gorithm based on hidden Markov model[D]. Beijing: Beijing University of Chemical Technology, 2020.

[23] 安晓宁, 王智文, 张灿龙, 等. 基于隐马尔可夫模型的人脸特征标注和识别[J]. 广西科技大学学报, 2020, 31(2): 118-125.
AN X N, WANG Z W, ZHANG C L, et al. Face feature labeling and recognition based on hidden Markov model[J]. Journal of Guangxi University of Science and Technology, 2020, 31(2): 118-125.

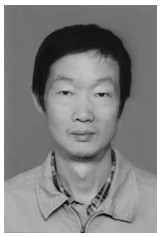
[24] MACQUEEN J. Some methods for classification and analysis of multivariate observations [C]//Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability. California: University of California Press, 1967: 281-297.

[25] 丛蓉, 王秀坤, 李进军, 等. 基于层次和密度聚类分析的航迹关联算法[J]. 系统仿真学报, 2005, 17(4): 841-843.
CONG R, WANG X K, LI J J, et al. Plot-track association algorithm based on hierarchical and density clustering analysis[J]. Acta Simulata Systematica Sinica, 2005, 17(4): 841-843.

[26] 李永森, 杨善林, 马溪骏, 等. 空间聚类算法中的 K 值优化问题研究[J]. 系统仿真学报, 2006, 18(3): 573-576.
LI Y S, YANG S L, MA X J, et al. Optimization study on K value of spatial clustering[J]. Journal of System Simulation, 2006, 18(3): 573-576.

[27] KAUFMAN L, ROUSSEEUW P J. Finding groups in data: an introduction to cluster analysis[M]. Saarland: DBLP, 2009.

[作者简介]



钱榕（1970− ），男，福建福州人，博士，北京电子科技学院副教授、硕士生导师，主要研究方向为复杂网络、数据挖掘、云计算安全等。

许建婷（1997− ），女，河北石家庄人，西安电子科技大学硕士生，主要研究方向为复杂网络、数据挖掘等。

张克君（1972− ），男，山东临沂人，博士，北京电子科技学院教授、博士生导师，主要研究方向为智能计算、信息安全等。

董宏宇（1994− ），女，内蒙古乌兰察布人，北京电子科技学院硕士生，主要研究方向为复杂网络、数据挖掘等。

邢方远（1998− ），男，辽宁宽甸人，北京电子科技学院硕士生，主要研究方向为复杂网络、数据挖掘等。

《通信学报》第十届编辑委员会