基于模体演化的多因子动态链路预测方法

赵宇红 张晓炜
(内蒙古科技大学信息工程学院 内蒙古 包头 014010)

摘 要 为了提高动态网络链路预测准确率,从网络结构微观演化角度,提出基于模体演化的多因子动态链路预测方法(MFME)。在动态网络时间窗口划分优化的基础上,引入整合移动平均自回归模型构建预测模体演化的概率矩阵,综合考虑模体演化影响因子及模体演化概率,可获得任意节点间的连接边概率。在真实数据集的实验表明,所提方法能达到更好的链路预测效果。

关键词 动态链路预测 模体演化 时间窗口 整合移动平均自回归模型 模体演化影响因子

中图分类号 TP393 **文献标志码** A DOI:10.3969/j.issn.1000-386x.2022.03.038

MULTI FACTOR DYNAMIC LINK PREDICTION METHOD BASED ON MOTIF EVOLUTION

Zhao Yuhong Zhang Xiaowei
(*School of Information Engineering, Inner Mongolia University of Science and Technology, Baotou 014010, Inner Mongolia, China*)

Abstract To make the accuracy of dynamic link prediction higher, from the perspective of micro-evolution of network structure, a multi factor dynamic link predication method (MFME) is proposed. Based on the optimization of dynamic network time window size division, the integrated moving average autoregressive model was introduced to construct the probability matrix for predicting the motif evolution. Considering the influence factors and probability of motif evolution, the continuous edge probability between any nodes could be obtained. The experiments on real data set show that the proposed method can achieve better link prediction effect.

Keywords Dynamic link prediction Motif evolution Time windows Integrated moving average autoregressive model Motif influence index

0 引 言

链路预测[1]作为复杂网络的核心内容之一,可以从网络数据中挖掘有用的链路信息,即基于已存在的网络拓扑信息、节点属性和历史信息,预测未来尚未建立连接关系的节点对之间连接的可能性[2]。传统的链路预测方法主要是建立在静态网络[3]的基础上的,具有代表性的常用的方法有 Common Neighbor(CN)[4]、Adamic-Adar(AA)[5]和 Preferential Attachment Index(PA)[6]等,这些方法虽然能在静态网络中取到较好结果,但却忽略了网络的动态性特点。为了解决这一问题,动态网络的链路预测被提出,考虑了时间信息,可以根据网络历史时刻的拓扑结构来预测出该网络在未来的链接结构。

对未来可能产生的边进行预测的核心是对网络演化规律的把握,然而,现有的大多数动态链路预测方法都是直接在网络宏观的全局拓扑结构进行分析,并没有注意到网络中的微观结构对网络演化是如何影响的。模体(motif)是一种非常重要的网络微观结构[7],研究发现它的演化规律可以在动态网络分析中起到关键作用。文献[8]从网络的微观演化为切入点,通过分析网络中频繁出现的模式的演变情况来进行链路预测。文献[9]首次度量模体的转换概率并定义其转换概率矩阵进行链路预测。文献[10]改进了文献[9]的方法,换用三阶张量分解的方法计算模体转换概率矩阵,同时考虑模体结构指标。

但这些方法都没有对动态网络下时间窗口的准确

收稿日期:2020-01-06。国家自然科学基金项目(61562065)。**赵宇红**,副教授,主研领域:链路预测。**张晓炜**,硕士生。

选取做过多的探究,并且文献[10]的方法是针对无向网络的研究,而现实中的大多网络数据是有向的,同时研究中也没有考虑网络中重复连边的问题。针对这些问题,提出基于模体演化的多因子动态链路预测方法(MFME)。该方法在动态有向网络下从合适大小的时间窗口的划分入手,利用模体演化矩阵来记录网络中的模体演化情况,引入整合移动平均自回归模型有效地预测模体的演化概率。最后,通过结合模体演化影响因子来计算节点间产生连边分数来进行链路预测。通过真实数据集下与其他方法的比较,验证了本文方法能够达到更好的链路预测效果。

1　相关介绍

1.1　动态链路预测

假设存在一个有向网络 $G=(V,E)$,$V=\{v_1,v_2,\cdots,v_N\}$ 是该网络的节点集合,$E=\{e_1,e_2,\cdots,e_N\}$ 是网络中的节点间构成的带权有向边集合。对于一个动态有向网络 G,给出它的时间窗口序列 $1,2,\cdots,t$,在 t 个连续的时间窗口下 $G=(g_1,g_2,\cdots,g_t)$,其中 $g_t=(v_t,e_t)$,v_t 表示 t 时间窗口下网络的节点集合,e_t 表示 t 时间窗口下网络中的 v_t 个节点之间形成的带权有向边集合[11]。为了达到研究目的,本文根据已知动态网络 G 中 g_1 到 g_t 的演化情况,设计一种链路预测方法,预测时间窗口 g_{t+1} 中的任意有序节点间产生连边的分数值,该分数值越大,则表示两节点之间产生链路的概率越高。

1.2　模体及其演化理论

模体最早是在生物学的蛋白质网络里表示最基本的功能模块,引入到复杂网络中便可以表示为网络的基本子结构,其中由三个节点构成的结构为最基本的模体结构。有向网络内,共有 15 种三元模体[12]结构,如图 1 所示。图中标注的模体编号与结构一一对应,该编号将在本文中应用,且下文所述模体均为三元类型。

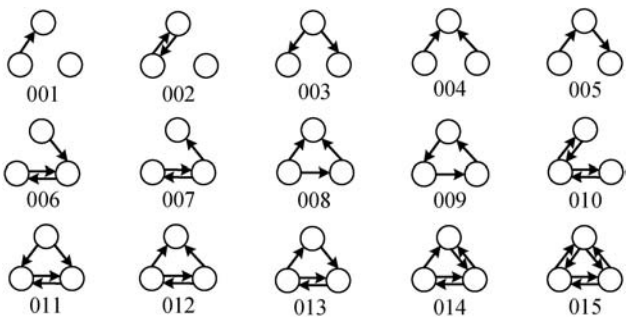

图 1　15 种三元模体结构

网络中不同模体间的演化规律的作用非常重要,本文正是利用各时序状态下不同模体间的演化规律来对动态网络的链路关系分析和预测。

定义 1　动态有向网络中,任意由三个点组成的节点集可以称为一个模体 m。m 为图 1 中的 15 种模体类型之一,用 mtf_i 来表示图 1 中第 i 类模体。动态网络的变化可以看作是大量模体间的演化过程,若在任意两个相邻的时间窗口下,某三个节点组成的模体由模体类型 i 转化成模体类型 j,那么这个过程可表示为:

$$mtf_i \rightarrow mtf_j \quad 1 \leqslant i,j \leqslant 15 \tag{1}$$

网络中各模体的演化即衍生了动态网络的演化。本文用 15×15 的模体演化矩阵 MEM(Motif Evolution Matrix)来表示动态网络相邻时间窗口上各模体间的演化规律。矩阵的行标分别对应前一时间窗口的 15 种不同类型的模体,而列标分别对应后一时间窗口的 15 种模体。

两相邻时间窗口对应矩阵中的每个元素都表示行标对应的模体转换到列标对应的模体的概率。将 15×15 种模体之间的历史转移概率看作特定的时间序列,使用 $MEM_{t,i,j}$ 表示在 t 时刻 mtf_i 转换为 mtf_j 的概率。

在时间窗口 $1\sim t$ 之间,任意两类模体 mtf_i、mtf_j 之间均存在演化概率时间序列 $MEM_{1,i,j}$,$MEM_{2,i,j}$,$\cdots$,$MEM_{t-1,i,j}$。通过上述的演化序列,如果 $MEM_{t,i,j}$ 的值能被准确预测,则会有利于动态链路预测的研究。

1.3　整合移动平均自回归模型

自回归移动平均模型(Autoregressive Moving Average Model)是一种时间序列预测方法。该模型的基本原理是:预测指标随时间形成的数据序列被视为随机序列,这些随机变量的相关性反映了原始数据在时间上的延续性。一方面,随机变量受某些因素的影响,另一方面,也有其自身的变化规律[13]。若一时间序列 $Y_t=\{Y_1,Y_2,\cdots,Y_T\}$ 满足自回归移动平均模型,可表示为:

$$\Phi(B)\nabla^d Y_t=\Theta(B)\varepsilon_t \tag{2}$$

式中:$\Phi(B)=1-\varphi_1B-\cdots-\varphi_pB^p$,$B^p$ 为延迟算子,与 Y_t 作用后得到 Y_{t-p},φ_p 为自回归系数;$\nabla^d=(1-B)^d$,d 为差分次数;$\Theta(B)=1-\theta_1B-\cdots-\theta_qB^q$,$\theta_q$ 为移动平滑系数;ε_t 为零均值白噪声序列。式(2)可简写为:

$$\nabla^d Y_t=\frac{\Theta(B)}{\Phi(B)}\varepsilon_t \tag{3}$$

将模型中心化后,可将公式简写为:

$$Y_t=\phi_0+\phi_1Y_{t-1}+\cdots+\phi_pY_{t-p}+\varepsilon_t+\theta_1\varepsilon_{t-1}-\cdots-\theta_q\varepsilon_{t-q} \tag{4}$$

整合移动平均自回归模型是自回归移动平均模型的改进,它不仅有自回归移动平均模型的优点,而且还

针对性地加入了非稳定序列的预测,并且它完成预测不需要借助其他外生变量。对于非数值型时间序列的处理,本文采用面向属性的预测思路,即预测每种不同类型属性在不同取值上的数量分布,可将非数值型预测转化为数值型预测。预测序列的稳定性对模型预测结果影响很大,数据稳定表示数据序列是没有趋势,没有周期性的;若数据是不稳定的,则无法捕捉到规律,难以预测。

1.4 模体演化影响因子

1.4.1 链接权重

现实中的动态网络往往存在重复的连边,把任意两节点关系的强度用其链接重复次数表示。假如模体内的某条边重复出现,则认为这条边代表的关系也就越强。若网络中两节点出现的重复连边 $e_1,e_2,\cdots,e_n$,用最初出现的边 e_1 作为这条链接的代表,把出现的次数作为 e_1 的权重 w,用 w 表示这两点间联系的强弱程度,链接关系的强弱将对连边的演化有着重要的影响。

1.4.2 闭包三原组

三元闭包是一种非常直观和自然关系的描述。例如,如果两个人 B 和 C 存在一个共同朋友 A,则这两个人在未来成为朋友的可能性就会提高。共同的朋友 A 直接导致他们彼此见面的几率增加,并且关系链形成的过程中,B 和 C 都和 A 是朋友,这为他们提供了陌生人所缺乏的基本信任[14]。这种三节点之间的关联形成三元闭包,而闭包也呈现不同的结构,且不同结构的闭包也会影响链接的演化结果。

1.5 实验数据集

本文使用两个真实的数据集进行链路预测实验。安然(Enron)网络[15]是 2000 年至 2002 年间 150 位安然员工之间发送的 60 多万封电子邮件的公开数据库,该数据集网络的连边中带有时间戳注释;Facebook-wosn-wall[16]是某一用户发给其他用户的一小部分帖子的定向网络,网络中的节点是 Facebook 用户,每个有向边代表一个帖子,并且连边具有时间戳。这两个真实网络中的节点关系都是有向的,并且这些节点关系都包含时间信息,因此都是动态的有向网络。数据具体参数如表 1 所示。

表 1 实验数据集参数表

参数	Enron	Facebook-wosn-wall
节点数	36 692	45 813
总边数	183 831	876 993
时间跨度/天	760/天	1 561/天

2 算法思路与设计

本文的研究重点是对动态链路预测的多影响因子的获取,主要分为两部分,一是如何对动态网络进行合适大小的时间窗口划分;二是怎样考虑模体演化过程中链接权重和闭包三元组对连边形成的影响。首先,确定两个动态网络中以时间窗口为共同变量且呈现趋势相反的函数,利用两函数的差值最小化来找到合适的窗口大小;其次,利用预测模型得到模体演化预测矩阵 MEPM(Motif Evolution Prediction Matrix)后,结合两个模体演化影响因子计算出相应的模体影响指数,综合预测矩阵得出任意两节点的产生连边的概率。

2.1 合适的时间窗口划分

对于动态网络预测研究,时间窗口的划分对其预测结果有较大的影响,因此选取适当的时间窗口划分方法非常必要。本文采用 SOTS(Segment of Time Series)方法对动态网络划分时间窗口。

在给定窗口大小为 ω 和相应的网络 $g(\omega)$ 的情况下,f 代表动态网络时间快照的不同统计信息,S_ω 表示其对应的时间序列信息:

$$S_\omega(g)=[f(G_1),f(G_2),\cdots,f(G_t),\cdots,f(G_T)] \tag{5}$$

为了保障时间窗口中保留较完善的网络快照信息,采用如下方法:

(1) 方差:$V(S_\omega)$ 表示 S_ω 的方差,可以看作是一种时间序列 S_ω 的噪声度量方法,公式如下:

$$V(S_\omega)=\frac{1}{T}\sum_{i=1}^{T}[f_\omega(G_i)-\varepsilon(S_\omega)]^2 \tag{6}$$

$$\varepsilon(S_\omega)=\frac{1}{T}\sum_{i=1}^{T}f_\omega(G_i) \tag{7}$$

式中:$\varepsilon(S_\omega)$ 为 S_ω 的均值。随着时间窗口 ω 的变化,方差的值也会发生改变。如果方差值较大的话,说明 S_ω 随着时间发生了剧烈的变化,这样会产生信息冗余,产生很多噪声;反之,会使得 S_ω 过于平滑,丢失很多有用信息。

(2) 时序压缩比:d 表示 $S_\omega(g)$ 的序列长度,c_d 表示利用数据压缩算法压缩后的 $S_\omega(g)$ 的序列长度,则 $S_\omega(g)$ 的时序压缩比 $R(S_\omega)$ 表示如下:

$$R(S_\omega)=\frac{d}{c_d} \tag{8}$$

$R(S_\omega)$ 是一种对 S_ω 的信息编码方式,一个较小的时序压缩比值表示 S_ω 的信息中存在大量的噪声;反之,则表示混乱度低,信息含量多。

由上边的描述可以看出,方差和压缩比随着窗口大小 ω 的变化呈现着相反的趋势,方差和时序压缩比

的最小差值也就是最优的时间窗口划分值。SOTS 时间窗口划分方法步骤如算法 1 所示。

算法 1　SOTS 时间窗口划分方法

输入:1. 带有时间标签的有向网络

2. 该网络下窗口划分的最大可能值 $\omega_{\max} \geqslant 1$

输出:合适的窗口大小 ω

```
for ω = 1 to ω_max do
    根据图序列计算时间序列 S_ω:[f(G_1),f(G_2),…,f(G_t),…,f(G_T)];
    计算对应窗口下的方差 V(S_ω)和时序压缩比 R(S_ω);
    if V(S_ω) - R(S_ω) = 0 then
        输出 ω;
    end if
end for
```

2.2　构建模体演化概率预测矩阵

以优化的时间窗口对动态网络进行等值窗口切分,对历史演化信息统计,则可以得到相邻时间窗口下不同模体类型的转换概率。在两个数据集下对所获取的模体演化信息进行分析,图 2 和图 3 分别是 Enron 数据集和 Facebook 数据集所截取的前 15 个时间窗口的不同模体下的演化情况,可以看出模体在演化过程中会出现一定的趋势性和周期性,即演化过程中的不稳定性。例如,图 2 中 008 号模体到 011 号模体的演化过程,图 3 中 005 号模体到 006 号模体的演化过程等,这样的不稳定性对预测模型的训练和预测结果都会有不好的影响。

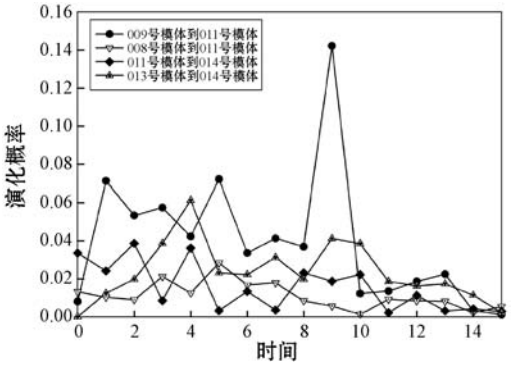

图 2　Enron 数据集下不同模体间演化概率

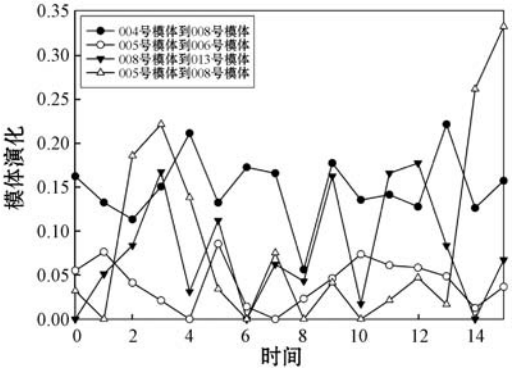

图 3　Facebook 数据集下不同模体间演化概率

本文基于 TTM prediction[9] 算法将每两个相邻时间窗口下的模体演化概率分别记录到对应的模体演化矩阵 MEM 当中,最后利用整合移动平均自回归的时序预测模型计算两个相邻时间窗口下的模体间演化概率并得到预测矩阵。

在划分好的相邻时间窗口下,若某两个类型模特间的演化概率的时间序列为 $MEM_{1,i,j}, MEM_{2,i,j}, \cdots, MEM_{t,i,j}$,表示为 $Y_t = (y_1, y_2, \cdots, y_t)$,使用整合移动平均自回归模型可表示为式(2)的形式。

(1) 用单位根检验法来判断数据的平稳性。当数据不平稳时,对其进行差分处理,差分阶数的选取从 1 逐渐增加,直至满足校验。检验方法如下:

假设序列经过 d 阶差分后平稳,可以设:

$$\Phi(B) = \prod_{i=1}^{p}(1 - \lambda_i B) \quad |\lambda_i| < 1, i = 1,2,\cdots,p \tag{9}$$

$$\Phi(B)\nabla^d = [\prod_{i=1}^{p}(1 - \lambda_i B)](1 - B)^d \tag{10}$$

由式(10)知,模型可以得出 $p + q$ 个根。当 $d \neq 0$ 时,可判断序列在模型下不平稳。

(2) 确定模型最佳的 p、q 值。根据所选数据的自相关函数 ACF 和偏自相关函数 PACF 的拖尾性和截尾性来确定 p 和 q。本文采用 AIC 标准,即利用使 AIC 达到最小值的自回归移动平均模型进行拟合。AIC 的标准函数如下:

$$AIC = n \ln L + 2(p + q + 1) \tag{11}$$

式中:L 为似然函数。选择使 AIC 达到最小的 p、q 值为最佳 p、q 值。

(3) 估计预测模型中参数的值。本文选用最小二乘法来估计参数值。方法如下:

$$\hat{\partial} = (\varphi_1, \varphi_2, \cdots, \varphi_p, \theta_1, \theta_2, \cdots, \theta_q)' \tag{12}$$

$$f_t(Y, \hat{\partial}) = \varphi_1 Y_{t-1} + \varphi_2 Y_{t-2} + \cdots + \varphi_p Y_{t-p} - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - \cdots - \theta_q \varepsilon_{t-q} \tag{13}$$

式中:φ_i 为自回归系数;θ_i 为移动平均系数;ε_i 为零均值白噪声序列。则残差项为:

$$\sigma_t = Y_t - f(Y; \hat{\partial}) \tag{14}$$

若 $\sum_{t=1}^{n} \sigma_t^2$ 最小,则 $\hat{\partial}$ 为最小二乘参数估计值。

至此,确立适合本文的模型及其参数。利用得到的模型预测出 $T-1$ 到 T 时间窗口下任意两类模体间的演化概率,从而得到预测矩阵 MEPM。整个过程可用图 4 表示。

图 4 模体演化矩阵及预测矩阵构建过程

每相邻的两个时间窗口下的模体演化信息用一个 MEM 记录,并把前 $T-1$ 个 MEM 作为历史序列,用整合移动平均自回归模型学习并预测出最后的 MEPM。

2.3 基于模体演化的多因子链路预测方法

网络中的一个节点对往往可能属于多个模体之中。如图 5 所示,一个简单的有向网络中存在节点对 (v_m,v_n),该节点对可以属于模体 (v_m,v_n,v_2)、(v_m,v_n,v_5),也可以属于 (v_m,v_n,v_4) 等。因此,当得到模体预测矩阵时,并不能完全直接由其得出某节点对之间的连边概率。但是,每个包含该节点对的模体的状态都为连边可能性的预测提供了参考,并且对该节点对影响更大的模体对预测结果往往也起到积极的作用。

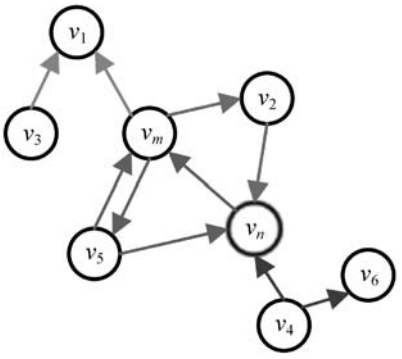

图 5 简单动态有向网络结构示例图

2.3.1 模体演化影响因子

影响模体的演化主要集中表现为两个方面,一方面是单个模体中各节点间的连边次数,连接频率越高,则说明该模体的影响度越大;另一方面是模体的闭合次数,在模体演化的过程中,如果某一模体形成的闭合次数越多,说明该模体内节点间关系越紧密,对节点对连边的预测也越重要。此外,针对有向网络而言模体的闭合类型也不一样,比如模体 (v_m,v_n,v_2) 是单层闭合,而模体 (v_m,v_n,v_5) 则存在一条双层闭合。

基于上述分析,对于任一个模体 (v_m,v_n,v_j),本文定义模体影响指数来综合描述模体演化影响因子即连边次数和闭包闭合次数对链路预测的影响程度:

$$I_{(v_m,v_n,v_j)} = \delta\sum_{i=1}^{Max_i}\sqrt{W_{e_i}} + (1-\delta)\sum_{t=1}^{T}\theta\cdot syn_t \quad e_i \in link(v_m,v_n,v_j) \tag{15}$$

式中:$I_{(v_m,v_n,v_j)}$ 表示该模体的模体影响指数;W_{v_i} 表示模体内任两节点间的连边次数;$link(v_m,v_n,v_j)$ 表示当前模体内所有存在的连边;syn_t 表示 t 时刻模体是否闭合,闭合为 1,不闭合为 0;δ 表示连边次数和闭合次数的影响占比。对于一个有向网的模体闭合状态,考虑 θ 为模体闭合程度,取值如式(16)所示。

$$\theta=\begin{cases}0.25 & \text{模体为全单层闭合}\\ 0.5 & \text{存在一组双层闭合}\\ 0.75 & \text{存在两组双层闭合}\\ 1 & \text{模体为全双层闭合}\end{cases} \tag{16}$$

根据得到的 MEPM 和模体影响指数,可以计算出预测的 T 时刻中任意节点对 (v_m,v_n) 间的连边概率:

$$P_{(v_m,v_n)} = \sum_{j=1}^{Max_j} I_{(v_m,v_n,v_j)}\cdot MEPM(mtf_{T-1},mtf_T) \quad v_j \in neig(v_m,v_n) \tag{17}$$

式中:mtf_{T-1} 为 $T-1$ 时刻 (v_m,v_n,v_j) 的模体名;mtf_T 为 T 时刻 (v_m,v_n,v_j) 的模体名;$neig(v_m,v_n)$ 为节点 v_m、v_n 的邻居节点集。

2.3.2 算法描述

在上述分析基础上,本文提出基于模体演化的多因子链路预测方法 MFME,算法如下:

算法 2 基于模体演化的多因子链路预测方法

输入:1 到 T 时刻的时间窗口

输出:T 时刻各节点对的连边概率

```
初始化:计算 1 ~ T-1 相邻时间窗口的模体演化矩阵;
计算 T-1 到 T 时间窗口的模体演化预测矩阵 MEPM;
for each(v_m, v_n) do
    for each(v_m, v_n) ∈ (v_m, v_n, v_j) ;        //v_j ∈ neig(v_m, v_n)
        用式(15)计算模体影响指数;
    end
    用式(17)计算节点对的连边概率;
end
```

3 实验与结果分析

3.1 实验设计

本文研究采用的实验环境是 AMD Ryzen 5 2600X,16 GB 内存,系统采用 Windows 10,实验程序采用 MATLAB 编写,版本为 R2019a。

本文所提的 MFME 方法将与 TTM[9]、TCM[10]、TS[18] 进行对比实验。其中,TTM 是最先利用模体的转换概率信息进行链路预测的方法,并得到了较好结果;TCM 是在 TTM 的基础上用三阶张量分解的方法计算模体转换概率矩阵,效果优于 TTM 方法;而 TS 方法是一种经典的动态链路预测方法。AUC[19] 可以从整体上衡量结果的精确度,对比实验采用 AUC 指标对预测结果的准确性进行验证。

3.2　实验结果及分析

下面先利用 SOTS 算法对选定的两个真实数据集进行时间窗口的划分。图 6 为 Enron 数据集在 SOTS 算法下的方差和时序压缩比的计算结果,可以看到随着时间窗口 ω 的增大,方差是总体呈下降趋势的;反之时序压缩比呈持续增长的趋势,并且很明显可以看到在 $\omega=6$ 附近时两者的值出现交点,因此我们将选取 $\omega=6$ 为 Enron 数据集的时间窗口划分大小。

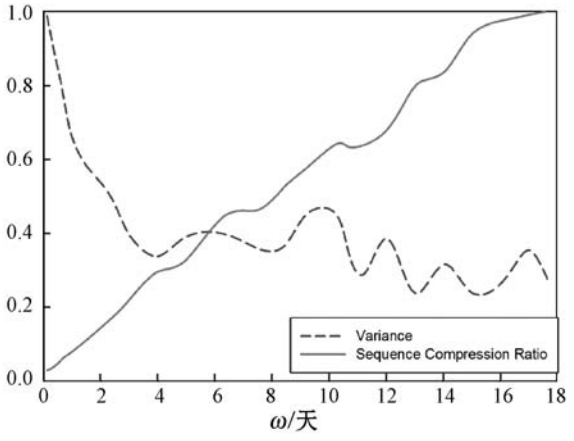

图 6　SOTS 算法在 Enron 下的划分情况

由于 Facebook-wosn-wall 数据集的网络规模和时间跨度都更大,因此在利用 SOTS 算法进行划分的时候,所呈现的曲线也有很大的不同。如图 7 所示,该数据集下的方差曲线呈快速下降的趋势,而时序压缩比曲线在 $\omega=20$ 附近才开始有明显的增长趋势,并且两条曲线在 $\omega=30$ 附近出现交点,因此,本文将选取 $\omega=30$ 为 Facebook-wosn-wall 数据集的时间窗口划分大小。

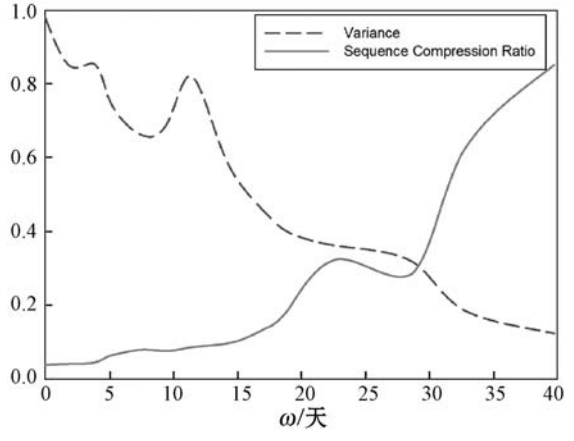

图 7　SOTS 算法在 Facebook-wosn-wall 下的划分情况

为了验证算法的有效性并能从中体现出时间窗口对实验结果产生的影响,分别对所用的两个数据集进行了多组实验。图 8 是 Enron 数据集在不同时间窗口下各算法的 AUC 值曲线,可以看出各算法在不同时间窗口下的 AUC 值有着较为明显的变化趋势,并且时间窗口在 5 ~8 的范围里各方法的 AUC 值都达到了最大值;图 9 是 Facebook-wosn-wall 数据集在不同时间窗口下各算法的 AUC 值变化曲线,可以看到曲线随着窗口大小的增长同样有明显的变化趋势,而且在窗口大小为 28 ~35 的区间里四种方法的 AUC 值也是达到一个最高的水平。

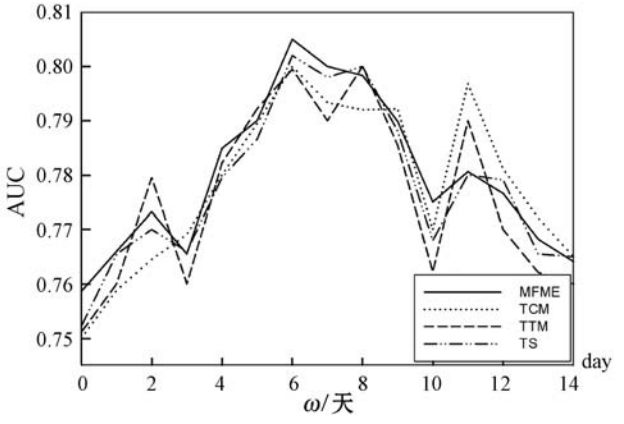

图 8　Enron 数据集不同时间窗口下各算法 AUC 对比

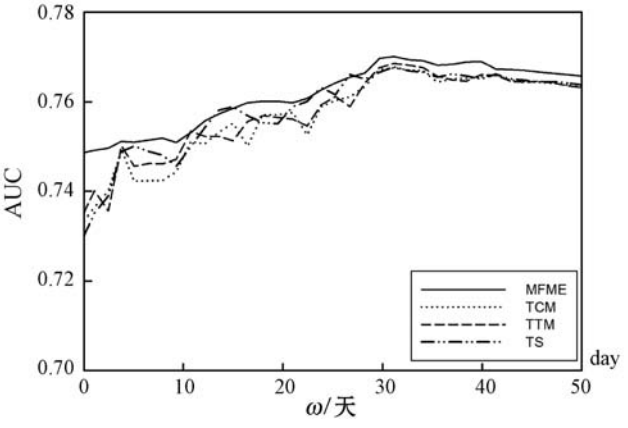

图 9　Facebook-wosn-wall 数据集不同时间窗口下各算法 AUC 对比

实验表明在 SOTS 的时间窗口划分结果下进行实验,各预测算法都能在选取的窗口大小中得到更为精准的预测结果,说明 SOTS 算法对动态网络时间窗口划分的有效性;同时利用两个时序相关函数的计算,也能更为快速地确定划分结果,这也体现了 SOTS 算法的高效性。

基于优化的时间窗口的划分,分别在两个数据集上进行了上述 4 种链路预测方法的对比实验。实验中,考虑到稀疏网络中模体的闭合次数较少,因此将连边次数的影响占比 δ 设为 0.7,则模体的闭合次数影响占比为 0.3。实验结果如图 10 所示,TS 算法作为一种经典的时序链路预测方法还是有着较高的准确性的,它在 Enron 数据集上的表现明显比 TTM 和 TCM 算法更好,但在 Facebook-wosn-wall 数据集上的表现较差,由于 Facebook-wosn-wall 数据集规模相对较大,模体演化的规律性更容易体现出来并被模型所学习,三种基于模体转换分析的算法都有着相似的表现。

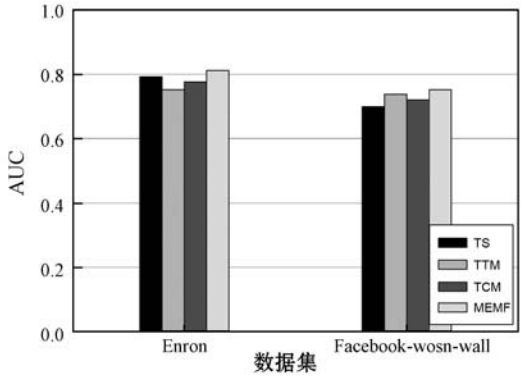

图 10 不同数据集下各算法的 AUC 值对比

虽然 TTM 和 TCM 都对模特转换进行了分析,但 TTM 缺乏对不同模体转换概率随时间变化的规律的进一步挖掘,而 TCM 缺乏对动态网络时间窗口划分准确合理的处理和对真实网络有向性的考虑。与其他三种算法相比,MFME 方法利用 SOTS 算法快速得到了更为准确合理的时间窗口的划分,对模体演化规律分析后选用合理的时序预测方法,同时充分考虑了模体的两个影响因子对模体演化的影响,使得 MFME 算法在两个数据集下的预测效果都有较好表现。

3.3 算法复杂度分析

本文所提的算法可分为三个主要部分,一是时间窗口大小的确定,二是模体演化概率矩阵的构建,三是模体影响指数的计算。其中,SOTS 算法的时间复杂度为 $O(n)$,模体演化概率矩阵计算时间复杂度为:如果采用传统的对网络进行遍历的方式计算则时间复杂度为 $O(tn^3)$,模体影响指数计算时间复杂度为 $O(nt)$。其中,n 为网络节点数,b 为网络最大度,t 为时间窗口。

4 结 语

链路预测是动态网络演化研究的重要内容,合适的时间窗口划分是对动态网络演化研究的必要且关键的一步。将模体这种微观结构的演化规律应用到动态网络的链路预测之中,可以更加深入、准确地挖掘节点间的关联变化从而取得较好的链路预测效果。现实中的大多数动态网络都是有向网络,本文基于动态有向网络首先研究时间窗口的划分,基于优化的时间窗口利用模体演化矩阵来记录网络中不同时间序列模体演化情况,引入整合移动平均自回归模型来预测模体的演化概率。最后,综合考虑模体演化影响因子获得连边影响指数并结合模体演化概率来计算节点间产生连边的分数。仿真实验表明,本文所提方法在时序链路预测中可获得更精确的预测值。在以后的研究中,欲更加深刻挖掘模体演化的周期性和趋势性等规律对动态有向网络链路预测的影响。

参 考 文 献

[1] 张月霞,冯译萱. 链路预测的方法与发展综述[J]. 测控技术,2019,38(2):8 - 12.

[2] Pujari M, Kanawati R. Link prediction in complex networks [M]//Advanced Methods for Complex Network Analysis, 2016:1196 - 1236.

[3] 刘继嘉. 基于相似性演化的动态网络链路预测算法研究[D]. 合肥:中国科学技术大学,2018.

[4] Yang Y, Zhang J, Zhu X, et al. Link prediction based on the tie connection strength of common neighbor[J]. International Journal of Modern Physics C(IJMPC),2019,30(11): 1 - 16.

[5] Rossi R A, Rao A, Kim S, et al. Higher-order ranking and link prediction: from closing triangles to closing higher-order motifs[EB]. arXiv:1906.05059, 2019.

[6] Chowdhury G G. Introduction to modern information retrieval [M]. Facet publishing, 2010.

[7] 杜凡,刘群. 有向动态网络中基于模体演化的链路预测方法[J]. 计算机应用研究,2019,36(5):167 - 171,179.

[8] Bringmann B, Berlingerio M, Bonchi F, et al. Learning and predicting the evolution of social networks[J]. IEEE Intelligent Systems, 2010, 25(4): 26 - 35.

[9] Juszczyszyn K, Musial K, Budka M. Link prediction based on subgraph evolution in dynamic social networks[C]//2011 IEEE 3rd International Conference on Privacy, Security, Risk and Trust; 2011 IEEE 3rd International Conference on Social Computing. IEEE, 2011:27 - 34.

[10] 王守辉,于洪涛,黄瑞阳,等. 基于模体演化的时序链路预测方法[J]. 自动化学报,2016,42(5):735 - 745.

[11] 刘书新,刘群,杜凡. 基于模体演化与社区一致性的时序链路预测方法[J]. 计算机应用研究,2019(12):3674 - 3678.

[12] Ortmann M, Brandes U. Efficient orbit-aware triad and quad census in directed and undirected graphs[J]. Applied Network Science, 2017, 2(1):13.

[13] 郑洋洋,白艳萍,续婷. 基于 SARIMA-SVR 组合模型的空气质量指数预测[J]. 河北工业科技,2019(6):436 - 441.

[14] 刘幼迟. 弱关系优势的分析逻辑:绝对论与相对论的比较[J]. 社会发展研究,2018,5(4):154 - 172,245.

[15] Enron email dataset[DS/OL]. [2020 - 01 - 06]. http://www.isi.edu/adibi/Enron/Enron.htm.

[16] WOSN 2009 data sets[DS/OL]. [2020 - 01 - 06]. http://socialnetworks.mpi-sws.org/data-wosn2009.html.

[17] 刘若愚,刘立波. 基于 ARIMA 模型的游客人数分析与预测[J]. 电脑与电信,2019(1):1 - 4.

[18] Huang Z, Lin D K J. The time-series link prediction problem with applications in communication surveillance[J]. Informs Journal on Computing, 2009, 21(2):286 - 303.

[19] Carter J V, Pan J, Rai S N, et al. ROC-ing along: Evaluation and interpretation of receiver operating characteristic curves[J]. Surgery, 2016, 159(6): 1638 - 1645.