动态多模网络中演化社区发现算法改进

胡 昊,张小燕,苏 勇
(江苏科技大学 计算机科学与工程学院,江苏 镇江 212003)

摘 要:在动态多模式网络中发现社区可以帮助人们了解网络的结构属性,解决数据不足和不平衡问题,并且可以协助解决市场营销和发现重要参与者的问题。一般来说,网络和它的社区结构是不均匀进化的。通过使用时态信息来分析多模网络,分析时态正则化架构和它的收敛属性。提出的算法可以解释为一个迭代的潜在语义分析过程,允许扩展到处理带有参与者属性和模内联系的网络。

关键词:数据挖掘;社区发现;社区演化;多模网络;动态网络
中图分类号:TP391 **文献标识码**:A **文章编号**:1674-7720(2011)24-0072-04

Identifying evolving groups in dynamic multi-mode networks

Hu Hao, Zhang Xiaoyan, Su Yong
(School of Computer Science and Engineering, Jiangsu University of Science and Technology, Zhenjiang 212003, China)

Abstract: Identifying communities in a multi-mode network can help understand the structural properties of the network, address the data shortage and unbalanced problems, and assist tasks like targeted marketing and finding influential actors within or between groups. In general, a network and its group structure often evolve unevenly. The paper tried to address this problem by employing the temporal information to analyze a multi-mode network. A temporally-regularized framework and its convergence property were carefully studied. It showed that the algorithm can be interpreted as an iterative latent semantic analysis process, which allows for extensions to handle networks with actor attributes and within-mode interactions.

Key words: data mining; community detection; community evolution; multi-mode networks; dynamic networks

当今网络拥有海量数据,要从海量数据中得到有用的信息是很困难的,因此网络分析[1]和建模[2]受到越来越多的关注。目前很多研究工作都只涉及一种模式的网络,即网络中只存在一种类型的参与者(点),参与者之间只存在同种类型的关系(联系)。但是,最近迅猛发展的 Web 数据挖掘涉及到了不止一种类型的参与者,这些参与者之间的关系也不再仅限于一种。这种类型的网络称为多模网络[3]。

在多模网络中,不同模中点的进化是不相同的。对于具有动态关系的异构实体,发现演化社区有很多的好处:(1)能够清晰地了解迥异模式之间的联系和长期演化模式;(2)可以形象化具有多种实体和多种关系的复杂网络;(3)有助于在多种领域中做决策;(4)在早期如果发现不良的演化样式,也可以发出事件警告。

在动态多模网络中发现社区演化还是很困难的,原因有二:(1)不同的模式之间的演化是有关联的;(2)不同模式具有独特的演化样式。本文采用谱聚类架构,提出一种发现动态多模网络中演化社区的一般方法。一个动态多模网络会包含一系列的网络快照,利用这些快照可以找出社区是如何演化的。在这个模型下,加入正则项反映时态变化[4],可以将有联系模式的聚类结果和相邻时间戳作为一个模式下的社区更新的属性,是一个将动态多模网络分析和常规的基于属性的数据挖掘联系起来的新方法。

1 问题阐述

给出含有 m 种类型元素 $X_1, X_2, \cdots, X_m$ 的 m 模网络,找出每一模中的潜在社区是如何演化的[5]。在架构中,通过一系列的网络快照只关注离散时间戳,这个方法在正则项网络分析中得到广泛应用。表 1 所示为下文中所涉符号及其表示的内容。

1.1 使用网络序列发现社区

在动态多模网络中,有多种多样的网络快照。不考虑时态效应的目标函数 F_1 可以写为:

$$F_1: \min \sum_{t=1}^{T} \sum_{1 \leqslant i<j \leqslant m} w_a^{(i,j)} \left\| \boldsymbol{R}_{i,j}^{t} - \boldsymbol{C}^{(i,t)} \boldsymbol{A}_{i,j}^{t} (\boldsymbol{C}^{(j,t)})^{\mathrm{T}} \right\|_F^2 \quad (1)$$

表 1 符号及其表示的内容

符号	表示的内容
m	模式的数量
n_i	模式 i 中参与者的数量
X_i	模式中的实体
$\boldsymbol{R}_{i,j}^{t}$	时刻 t 下模式 i 和 j 的联系
k_i	模式 i 中潜在社区的数量
$\boldsymbol{C}^{(i,t)}$	时刻 t 下模式 i 的社区指示符矩阵
$\boldsymbol{A}_{i,j}^{t}$	在模式 i 和模式 j 之间的组联系密度
$w_a^{(i,j)}$	模式 i 和模式 j 之间联系的权重
$w_b^{(i)}$	时态正则化的权重

$$s.t.(\boldsymbol{C}^{(i,t)})^{\mathrm{T}}\boldsymbol{C}^{(i,t)}=I_{k_i}\quad i=1,\cdots,m,t=1,\cdots,T \tag{2}$$

有如下定理：

定理 1：如果 $\boldsymbol{C}^{(i,t)},1\leqslant i\leqslant m,1\leqslant t\leqslant T$ 是方程 F_1 的有效解，那么 $\widetilde{\boldsymbol{C}}^{(i,t)}$ 也是具有相同目标值的有效解，其中 $\widetilde{\boldsymbol{C}}^{(i,t)}$ 定义如下：

$$\widetilde{\boldsymbol{C}}^{(i,t)}=\boldsymbol{C}^{(i,t)}\boldsymbol{Q}^{(i,t)}$$

$$s.t.(\boldsymbol{Q}^{(i,t)})^{\mathrm{T}}\boldsymbol{Q}^{(i,t)}=\boldsymbol{Q}^{(i,t)}(\boldsymbol{Q}^{(i,t)})^{\mathrm{T}}=I_{k_i}$$

$$\boldsymbol{Q}^{(i,t)}\in\boldsymbol{R}^{k_i\times k_i}$$

1.2 使用时态正则化发现社区

公式 F_1 并没有关注连续时间戳之间的关系。可以将 F_1 归结为每一个快照单独地进行聚类。实际生活中的社区演化是非常缓慢的，为了得到平滑的社区演化，将增加时态正则项 Ω，它可以迫使聚类序列通过不同的时间戳时变得平滑。

$$\Omega=\frac{1}{2}\sum_{t=2}^{T}\left|\left|\boldsymbol{C}^{(i,t)}(\boldsymbol{C}^{(j,t)})^{\mathrm{T}}-\boldsymbol{C}^{(i,t-1)}(\boldsymbol{C}^{(j,t-1)})^{\mathrm{T}}\right|\right|_F^2 \tag{3}$$

式(3)中，“1/2”只是为了下面求导方便做的计数。实际上，这里进行一阶马尔科夫假设，要求当前的聚类与前一个时间戳的聚类很相似。

$$\Omega=\sum_{t=2}^{T}\left|\left|\boldsymbol{C}^{(i,t)}-\boldsymbol{C}^{(i,t-1)}\right|\right|_F^2 \tag{4}$$

正如定理 1 指出的，$\boldsymbol{C}^{(i,t)}$ 在正交变换下是相等的，因此，可以在式(4)中直接比较 $\boldsymbol{C}^{(i,t)}$ 和 $\boldsymbol{C}^{(i,t-1)}$，而不必得出它们在不同时间戳下聚类指示符的不同。相比较而言，式(3)中正则项与正交变换无关，因此应该得到相邻时间戳下社区结构的不同。因为正则化，发现演化社区的问题就可以阐述成：

$$F_2:\min\sum_{t=1}^{T}\sum_{i<j}w_a^{(i,j)}\left|\left|R_{i,j}^{t}-\boldsymbol{C}^{(i,t)}\boldsymbol{A}_{i,j}^{t}(\boldsymbol{C}^{(j,t)})^{\mathrm{T}}\right|\right|_F^2+$$

$$\frac{1}{2}\sum_{i=1}^{m}w_b^{(i)}\sum_{t=2}^{T}\left|\left|\boldsymbol{C}^{(i,t)}(\boldsymbol{C}^{(i,t)})^{\mathrm{T}}-\boldsymbol{C}^{(i,t-1)}(\boldsymbol{C}^{(i,t-1)})^{\mathrm{T}}\right|\right|_F^2 \tag{5}$$

$$s.t.(\boldsymbol{C}^{(i,t)})^{\mathrm{T}}\boldsymbol{C}^{(i,t)}=I_{k_i}\quad i=1,\cdots,m,t=1,\cdots,T \tag{6}$$

其中，$w_b^{(i)}$ 是联系的块体模型近似和时态正则化的平衡因子。因为演化非常缓慢，所以这里只是寻找到和联系矩阵一致的社区结构[6]，而不是查找和先前时间戳中相差很大的社区。

2 时态正则化多模聚类

为了在动态多模网络中得到演化社区，通过使用迭代算法对 F_2 求解。对于 $\boldsymbol{A}_{i,j}^{t}$ 和 $\boldsymbol{C}^{(j,t)}$，如果其他变量不变，存在一个闭合形式解。为了便于理解和扩展，将在属性视图中解释这个算法。

2.1 A 的估计

定理 2：对于给定的 $C^{(i,t)}$，最佳的社区作用矩阵 $\boldsymbol{A}_{i,j}^{t}$ 可以使用下式得到：

$$A_{i,j}^{t}=(\boldsymbol{C}^{(i,t)})^{\mathrm{T}}R_{i,j}^{t}\boldsymbol{C}^{(j,t)}$$

2.2 C 的计算

给出最佳的 $\boldsymbol{A}_{i,j}^{t}$，如下验证：

$$\left|\left|R_{i,j}^{t}-\boldsymbol{C}^{(i,t)}\boldsymbol{A}_{i,j}^{t}(\boldsymbol{C}^{(j,t)})^{\mathrm{T}}\right|\right|_F^2$$

$$=\left|\left|R_{i,j}^{t}\right|\right|_F^2-\left|\left|(\boldsymbol{C}^{(i,t)})^{\mathrm{T}}R_{i,j}^{t}\boldsymbol{C}^{(j,t)}\right|\right|_F^2 \tag{7}$$

同时：

$$\frac{1}{2}\left|\left|\boldsymbol{C}^{(i,t)}(\boldsymbol{C}^{(i,t)})^{\mathrm{T}}-\boldsymbol{C}^{(i,t-1)}(\boldsymbol{C}^{(i,t-1)})^{\mathrm{T}}\right|\right|_F^2$$

$$=\frac{1}{2}tr[\boldsymbol{C}^{(i,t)}(\boldsymbol{C}^{(i,t)})^{\mathrm{T}}+\boldsymbol{C}^{(i,t-1)}(\boldsymbol{C}^{(i,t-1)})^{\mathrm{T}}-$$

$$2\boldsymbol{C}^{(i,t)}(\boldsymbol{C}^{(i,t)})^{\mathrm{T}}\boldsymbol{C}^{(i,t-1)}(\boldsymbol{C}^{(i,t-1)})^{\mathrm{T}}]$$

$$=k_i-\left|\left|(\boldsymbol{C}^{(i,t)})^{\mathrm{T}}\boldsymbol{C}^{(i,t-1)}\right|\right|_F^2 \tag{8}$$

因为式(7)中的 $\left|\left|\boldsymbol{R}_{i,j}^{t}\right|\right|_F^2$ 和式(8)中的 k_i 是不变的，所以可以将 F_2 转化为下面的目标函数：

$$F_3=\max\sum_{t=1}^{T}\sum_{1\leqslant i<j\leqslant m}w_a^{(i,j)}\left|\left|(\boldsymbol{C}^{(i,t)})^{\mathrm{T}}\boldsymbol{R}_{i,j}^{t}\boldsymbol{C}^{(j,t)}\right|\right|_F^2+$$

$$w_b^{(i)}\sum_{t=2}^{T}\sum_{i=1}^{m}\left|\left|(\boldsymbol{C}^{(i,t)})^{\mathrm{T}}\boldsymbol{C}^{(i,t-1)}\right|\right|_F^2 \tag{9}$$

注意，$\boldsymbol{C}^{(i,t)}$、$\boldsymbol{C}^{(j,t)}$ 以及 $\boldsymbol{C}^{(i,t-1)}$ 都是有联系的。一般来说，这个函数没有解析的闭合形式解。但是如果有给定的 $\boldsymbol{C}^{(j,t)}$ 和 $\boldsymbol{C}^{(i,t\pm1)}$，则可以根据定理 3 直接得到最佳的 $\boldsymbol{C}^{(i,t)}$。

定理 3：如果给定 $\boldsymbol{C}^{(j,t)}$ 和 $\boldsymbol{C}^{(i,t\pm1)}$，那么 $\boldsymbol{C}^{(i,t)}$ 就可以作为矩阵 $\boldsymbol{P}_i^t$ 的顶左奇异向量求解，$\boldsymbol{P}_i^t$ 通过以下矩阵在列方向的级联得到。

$$\boldsymbol{P}_i^t=[\{\sqrt{w_a^{(i,j)}}\}_{j\neq i},\sqrt{w_b^{(i)}}\,\boldsymbol{C}^{(i,t\pm1)}] \tag{10}$$

2.3 属性视图中的算法

借助轮换寻优思想解决式(9)中的 F_3 问题。即求解 $\boldsymbol{C}^{(i,t)}$ 时，固定其他变量的值。这个过程一直迭代，直到函数收敛。收敛之后，$\{\boldsymbol{C}^{(i,t)}\}$ 就是近似的社区指示符矩阵。通常使用后处理视图恢复社区中不相邻部分，即对社区

指示符采用 k-均值聚类。综上所述，时态正则化多模聚类算法如下：

输入：$R, k_i, w_a^{(i,j)}, w_b^{(i)}$

输出：$idx^{(i,t)}, \boldsymbol{C}^{(i,t)}, \boldsymbol{A}_{i,j}^{t}$

产生初始聚类指示符矩阵 $\boldsymbol{C}^{(i,t)}$

迭代

For $t=1:T, i=1:m$

适当地增加或减小 $\boldsymbol{C}^{(i,t\pm1)}$

根据定理 3 计算 $\boldsymbol{P}_i^t$（或者 M_i^t）

计算 $\boldsymbol{P}_i^t$ 的奇异值分解（或者 M_i^t 的特征向量）

根据顶左奇异（特征）向量更新 $\boldsymbol{C}^{(i,t)}$

直到 F_3 的目标值小于 ε

在属性视图中解释这个算法，更新 $\boldsymbol{C}^{(i,t)}$ 的每一步都和潜在的语义分析过程一致。根据定理 3，社区指示符 $\boldsymbol{C}_i^t$ 和矩阵 $\boldsymbol{P}_i^t$ 的左奇异向量是一致的，在式(10)中也做了定义。如果将 $\boldsymbol{P}_i^t$ 作为正规实例-属性矩阵，则找出社区指示符和进行潜在的语义分析 LSA(Latent Semantic Analysis)同样重要。图 1 指出了整个算法的过程。

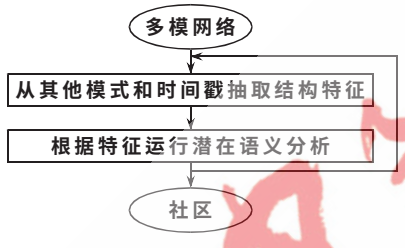

图 1 树形视图中的算法：迭代的 LSA

3 动态数据上的实验

实验中进行下面三种方法的比较：静态聚类、在线聚类和本文提出的时态正则化多模聚类。静态聚类是一种基线方法，它不关心任何的时态规则化。静态聚类通过对式(1)中的 F_1 求解对每一个网络快照单独进行聚类。

因为真实的数据不包含验证信息，即在不同时间戳下的社区联系，因此，使用合成数据验证提出算法的有效性。

3.1 实验设置

构建一个三模动态网络，模分别有 2、3、4 个社区和 50、100、200 个元素。每两个模型之间都可以发生联系，为了产生迭代，条件设置为：(1)为每个元素决定潜在的社区；(2)实体之间基于团体同一性产生的关系符合伯努利分布。

为了模拟演化，在不同的时间戳下按照如下规则发生联系：

(1)在每个时间戳下，参与者的部分(20%~50%)要转化为与先前时间戳下的组完全不同的组中，这是为了模拟社区演化。

(2) 两个组之间发生联系的概率是高斯分布的样本，主要和先前时间戳下的概率有关。这是为了模拟联系的改变。

(3)在每个时间戳下添加噪音。例如噪音强度是 0.2，则联系矩阵中大概有 20%的词条被随机设为 0，另外 20%的词条被随机设为 1。噪音和潜在的组结构是独立的，因此，噪音强度越高，在联系中的社区结构越不容易被发现。

为了研究不同聚类方法的属性，本文在不同的噪音强度下产生数据。对于每一个噪音强度构建 100 个合成网络，为简单起见，将式(5)中的权重 $w_a^{(i,j)}$ 和 $w_b^{(i)}$ 都设为 1。如果算法目标函数的相对变化小于 10^{-6}，算法结束。采用归一化互信息 NMI(Normalized Mutual Information)评价聚类的效果。π^a 和 π^b 表示社区两个不同的部分，则 NMI 可表示为：

$$\mathrm{NMI}=\frac{\sum_{h=1}^{k^{(a)}}\sum_{l=1}^{k^{(b)}}\log\left(\frac{n\times n_{h,l}}{n_h^{(a)}\times n_l^{(b)}}\right)}{\sqrt{\left(\sum_{h=1}^{k^{(a)}}n_h^{(a)}\log\frac{n_h^{(a)}}{n}\right)\left(\sum_{l=1}^{k^{(b)}}n_l^{(b)}\log\frac{n_l^{(b)}}{n}\right)}}$$

其中，n 表示数据实例的总数，$k^{(a)}$ 和 $k^{(b)}$ 分别代表社区 π^a 和 π^b 部分的数量，$n_h^{(a)}$、$n_l^{(b)}$、$n_{h,l}$ 分别表示第 h 个社区 π^a 部分的数量、第 l 个社区 π^b 部分的数量和第 h 个社区 π^a 部分及 π^b 部分的总数量。NMI 是介于 0 和 1 之间的一个量，等于 1 时表示两个部分相当。

3.2 实验结果

表 2 列出了超过 100 次运行的结果取平均值，其中，粗体表示结果中最好的。从表中数据可见，聚类效果随着噪音的加强变得越来越坏。总体来看，规则化聚类比在线聚类的效果好，在线聚类比静态聚类的效果好。从结果中也注意到，当噪音强度比较大(即噪音强度为 0.55)时，社区结构的平滑性遭到了破坏，也因此时态正则化聚类效果比静态聚类还差。

表 2 不同噪音强度下各种聚类比较

噪音强度	静态聚类	在线聚类	正则化聚类
0.05	0.925 1	0.926 7	**0.927 3**
0.10	0.899 8	0.899 7	**0.910 2**
0.15	0.891 1	0.894 5	**0.902 9**
0.20	0.856 9	0.857 4	**0.869 5**
0.25	0.787 5	0.787 6	**0.790 4**
0.30	0.710 9	0.713 5	**0.724 2**
0.35	0.588 8	0.591 2	**0.603 9**
0.40	0.471 9	0.476 0	**0.488 1**
0.45	0.317 7	0.320 9	**0.321 9**
0.50	0.199 1	**0.200 5**	0.199 2
0.55	0.120 0	**0.121 5**	0.113 5

图 2 显示平均计算时间。噪音越大，计算时间越长。静态聚类需要的时间是最短的，在线聚类的时间相对较长，时态正则化聚类的时间是最长的，特别是当噪音强度非常大时，时间变得不可接受。在这种情况下，时态平滑性已经被损害，算法需要更多的迭代找到最优解。

为了显示参数调整的效果，选择中等噪音强度的数据集，使用在线聚类和正则化聚类，时态权重 w_b 从 0.01~1 000 进行调整，w_a 固定为 1。如图 3 所示，时态权重过大反而得到不好的效果，即时态正则化处于首要地

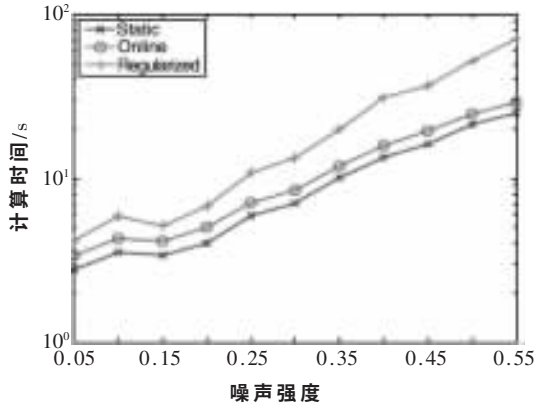

图 2 不同方法的计算时间

位。大部分时间,时态规则化有利于聚类考虑时态信息,时态权重在 0.01~100 的范围内体现的尤为明显。

在实际应用中,异构参与者之间的互相作用形成了多模网络。正是在这样的网络中,不同模的参与者构成社区并慢慢演化。本文提出了时态正则化多模聚类算法在动态多模网络中发现演化社区。这个算法可以理解为迭代的 LSA 过程,在不同模和时间戳下的属性构成社区矩阵。基于这种属性视图,提出的算法也能扩展到处理带有属性的网络、模内联系以及休眠点和活跃点。实验结果证明该算法能够根据一系列的快照找到更精确的社区结构和社区演化。

参考文献

[1] NEWMAN M.The structure and function of complex networks[J].SIAM Review,2003,45(2):167-256.

[2] CHAKRABARTI D,FALOUTSOS C.Graph mining:laws,generators,and algorithms[J].ACM Comput.Surv.,2006,38(1):65-78.

[3] WASSERMAN S,FAUST K.Social network analysis:methods and applications[M].Cambridge University Press,1994.

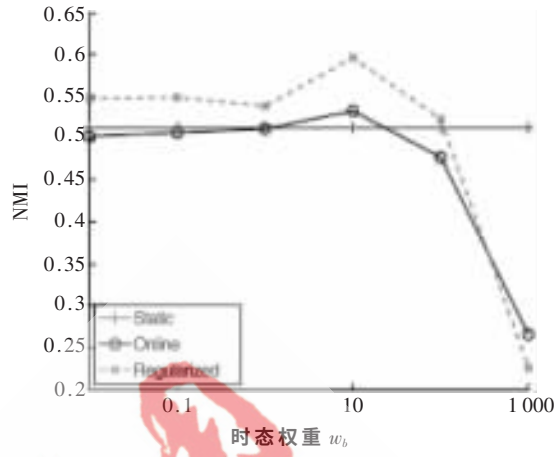

图 3 时态权重的性能敏感度

[4] BAUMES J,GOLDBERG M,WALLACE W,et al.Discovering hidden groups in communication networks[C].In 2nd NSF/NIJ Symposium on intelligence and Security Informatics,2004.

[5] LONG B,ZHANG Z M,WU X,et al.Spectral clustering for multi-type relational data[C].In ICML'06:Proceedings of the 23rd international conference on Machine learning.ACM,2006:585-592.

[6] 王林,戴冠中.基于复杂网络中社区结构的论坛热点主题发现[J].计算机工程,2008,34(11):214-21.

(收稿日期:2011-08-29)

作者简介:

胡昊,男,1985年生,硕士研究生,主要研究方向:数据挖掘。

张小燕,女,1987年生,硕士研究生,主要研究方向:数据挖掘。

苏勇,男,1958年生,硕士研究生导师,主要研究方向:知识发现与数据挖掘。