一种融合贝叶斯概率的社区结构发现方法研究

王 刚
(安康学院,陕西 安康 725000)

摘 要:社区结构通常具有动态、不对称、模糊的特性。为了更好地发现社区结构以及描述社区成员之间的关系,针对当前方法的一些不足,对利用贝叶斯概率来改进社区结构发现的方法进行研究。贝叶斯概率在描述成员之间动态、因果、模糊关系时具有优势,通过引入信息熵,提出了一种融合贝叶斯概率的社区发现方法。该方法首先计算成员之间的贝叶斯概率,研究贝叶斯关系网络构建方法,得出成员之间不对称贝叶斯概率矩阵;然后根据系统内信息的熵相对稳定的性质,把成员间贝叶斯概率作为信息熵的概率输入,计算出新成员加入后信息熵的变化值,根据熵值变化情况来确定成员是否属于社区,从而在发现社区结构的同时,也能描述社区成员之间的不对称、动态和模糊关系。实验结果证明了该方法的有效性。

关键词:社区发现;贝叶斯概率;信息熵;数据挖掘
中图分类号:TP301 **文献标识码**:A **文章编号**:1673-629X(2019)01-0110-04
doi:10.3969/j.issn.1673-629X.2019.01.023

Research on a Community Structure Detection Method Based on Bayesian Probability

WANG Gang
(Ankang University, Ankang 725000, China)

Abstracts: Community structure is commonly dynamical, fuzzy and asymmetric. In order to find the community structure and describe the relationship between members, for the shortcomings of current methods, we try to study a new method of community detection based on Bayesian probability which has advantages in describing dynamic, causal and fuzzy relations among members. By introducing information entropy, we propose a community discovery method integrating Bayesian probability. It first calculates the Bayesian probability among members, studies the Bayesian network construction method, and obtains the asymmetric Bayesian probability matrix between members. Then according to the nature of relatively stable information entropy in the system, the Bayesian probability between members as probability input of information entropy, the variable of information entropy after new members joining is calculated to determine whether members belong to the community. Thus, the asymmetrical, dynamic and fuzzy relationships among community members can be described while the community structure is discovered. Experiment shows that the method is effective.

Key words: community detection; Bayesian probability; information entropy; data mining

1 概 述

针对社区发现的研究是当前国内外研究的热点之一[1]。社会网络(social networks)是指社会中个人之间、组织之间或者个人与组织之间比较持久、稳定的社会关系模式,通常表示为节点的集合,同一社区节点集合之间的连接较多,如果节点集合之间连接较少,则认为属于不同的社区。研究网络中的社区有助于分析网络行为,对于个性化推荐具有重要的意义。

目前主要的社区发现算法有谱平分法[2-3]、KL 算法、层次聚类法、GN 算法等。谱平分法由于涉及到许多矩阵特征向量的计算,计算的时间复杂度可达 $O(n^3)$,在大数据环境中,它的效率成为瓶颈。有研究人员从社会学角度研究了社区结构,研究了人际之间的信任关系及提取[4],对相似度计算方法进行改进[5-6],采取角色连接轮廓方法从结构上进行划分,发现它们属于外围串类型。也有研究人员利用社会网络个体间的关系类别和个体间对应社会属性相似度引入关系模型来进一步量化团队成员个体间的关系强

收稿日期:2018-01-19 **修回日期**:2018-05-23 **网络出版时间**:2018-09-21
基金项目:国家自然科学基金(61152003);陕西省社科基金(2015M004)
作者简介:王 刚(1972-),男,博士,教授,研究方向为人工智能。

度[7]。此外,还有学者对从朋友关系、距离变化等角度发现社区结构进行研究[8-9]。上述研究能在一定条件和背景下发现成员之间的关系和结构,这些算法注重于网络节点的信息,如出度、入度[10],而对社区结构的动态、不对称、模糊的特性考虑不够。通常认为,结构上临近的节点,应该属于一个社区,而且关系是对称的,即:如果 a 信任 b,那么 b 就信任 a。而实际应用中,考虑到一些具体的关系[11],上述结论不一定有效,如应用 GN 算法,可能会把边"$a \to b$"作为两个社区的分界线,如果 a 和 b 本身属于事件非常重要的成员,把他们分割到两个社区就不合适。因此,结合社区结构的动态、不对称、模糊的特性,探讨一种新的方法很有必要。根据信息熵理论,随着网络节点的扩充,信息量增加,网络蕴含不确定信息的概率就会增加[12]。因此,一个社区内部,由于成员的增加,出现不确定性信息的概率应该增加,熵就增加,反之,则减少。一个社区内部,由于信息交流频繁,表现为一个稳定的综合体,不确定信息出现的概率不会剧烈增加或减少,这使得根据节点集合熵的变化来确定不同的社区成为可能。判断一个节点是否属于一个社区,可以通过判断节点加入社区后社区熵的变化来确定。

对于确定节点 a、b 之间关系的紧密程度,有许多方法,如计算余弦相似度、皮尔逊相似度系数等。通常的相似度计算方法不能很好地解决应用数据的动态模糊性,而贝叶斯概率在描述个体之间关系的动态和模糊性方面有一定的优势,因此文中提出了一种融合贝叶斯概率和熵的社区发现算法,并将其与余弦相似度计算方法进行比较。

2　信息熵及贝叶斯网络的建立

2.1　信息熵

1948 年,香农第一次将熵这一概念引入到信息论中,从此,熵这一概念被作为信息的度量,在自然科学和社会科学等领域应用广泛,并成为一些新学科的理论基础[13-14]。当一种信息出现概率更高的时候,表明它被传播得更广泛,或者说,被引用的程度更高。

一个信源发送出什么信息是不确定的,衡量它可以用其出现的概率来度量,信息熵概率公式表示为:

$$H(x) = -\sum_{x \in X} p(x)\log p(x) \tag{1}$$

其中,x 表示随机变量,与之相对应的是所有可能输出的集合,定义为符号集 $H(x)$;$p(x)$ 表示输出概率函数。

信息熵用来度量系统的有序化程度,一个系统越是有序,信息熵就越低,系统越是混乱,信息熵就越高。

2.2　贝叶斯概率与贝叶斯网络的生成

贝叶斯概率是由贝叶斯理论所提供的一种对概率的解释,它将概率定义为主体对一个命题的信任程度。贝叶斯网络通常表示为有向无环图,刻画一组变量的联合概率分布,每个变量在贝叶斯网络中表示为一个节点,网络弧表示断言"此变量在给定其直接前驱时,条件独立于其非后继",当 Y 到 X 存在一条有向的路径,称 X 是 Y 的后继。对每个变量有一个条件概率表,表示该变量在给定其立即前驱时的概率分布[15]。贝叶斯法则提供了计算假设概率的方法,提供了从先验概率 $p(h)$ 以及 $p(D)$ 和 $p(D \mid h)$ 计算后验概率 $p(h \mid D)$ 的方法。

贝叶斯公式表示为:

$$p(h_i \mid D) = \frac{p(D \mid h_i)p(h_i)}{\sum_{i=1}^{n} p(D \mid h_i)p(h_i)} \tag{2}$$

其中,$h_1, h_2, \cdots, h_n$ 为完备事件组,即 $\bigcup_{i=1}^{n} h_i = \Omega$,$h_j = \varphi$,$P(h_i) > 0$。

当前推荐系统通常采用关联规则挖掘方法,可能产生大量的序列模式。序列模式只是项集(Item set)的简单组合,依据支持度和可信度进行剪枝,而支持度和可信度依据项的个数计算得来,这样有些序列模式可能被认为毫无意义而被剪枝,如因果关系。由于刻画项之间深层关系存在的局限性,由于没有达到支持度阈值而被剪枝,而项之间的因果关系是非常重要的关系,不能简单凭借数量来判断,应该根据概率来判断是否剪枝。

贝叶斯概率在描述事件因果关系方面具有优势,所以采用贝叶斯条件概率来描述项之间的概率关系,依据概率的大小,寻找社区中的项集合,这样就能发现和挖掘项之间的因果关系,并用于推荐系统中,从而克服其他方法的局限性。

以商品推荐为例,利用算法 1 构建了用户贝叶斯关系网络图,它以用户为节点,节点之间的边表示用户之间的贝叶斯关联概率,形成的有向图如图 1 所示。

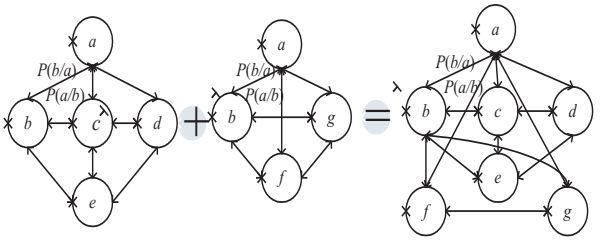

图 1　用户关系边有向图

算法 1:生成贝叶斯关系网络有向图。

Begin:

(1)一条记录包含多个商品,构成节点 $U = \{U_i\}$,$U_i = \{g_k\}$

(2)计算 $P(u_j/u_i) = \dfrac{P(u_i u_j)}{p(u_i)}$,得到边 $E(u_i, u_j)$ 的条件概率 $P(u_j/u_i)$

$$p(u_i) = \frac{\sum_{u_i}^{g_k}}{\sum g}, P(u_i u_j) = \frac{\sum n_{\text{umber}}(u_i, u_j)}{\sum g}, P(u_j/u_i) \neq P(u_i/u_j)$$

End

3　社区结构的发现

社区内各要素的内部及其相互间具有相对稳定的关系,个体成员之间因为互动而形成相对稳定的结构,由于蕴含有的共同信念和意愿,成员之间表现出强的趋一性。根据熵理论,由于系统的稳定和凝聚性,新的信息加入会对原系统的熵造成波动,如果新加入的节点与原系统能融入在一起,或接近度比较高,熵值的波动就较小,否则熵值的波动就较大。可以根据熵值的变化程度来决定新加入的节点是否被系统接纳。该算法随机确定一个节点为栈顶,建立一个栈,采用递归的思想,根据节点加入时熵值的变化,用阈值进行剪枝,超过阈值的节点不会加入到已有社区集合,满足条件时结束递归,输出产生的社区。

算法 2:基于信息熵的社区结构发现。

Begin

(1) $k = 0$

(2)建立节点堆栈 S

(3)确定栈顶 U_0, U_0 可以随机选取

(4)选取 U_0 的邻节点$\{U_k\}$

(5)计算 $\nabla H(H_0, H_k) = | H_0 - H_k |$, H_0 为系统现有的熵, H_k 为邻节点的熵

(6)如果 $\nabla H(H_0, H_k) < \varepsilon$,则该边入社区,把 U_k 压入堆栈 S,边 $E(H_0, H_k)$ 标记为True, $U_0 = U_k$, 获取栈顶的邻节点$\{U_j\}$, $U_k = U_j$, ε 为熵阈值,转步骤 4。否则如果 $\nabla H(H_0, H_k) \geqslant \varepsilon$, U_k 出栈,该边不加入社区,边 $E(H_0, H_k)$ 标记为 False

(7)输出标记为 True 的边 $E(H_0, H_k)$

(8) $k = k + 1$,选取标记为 False 的边节点 H_k ,作为栈顶,转步骤 3,直到所有节点标记完成或社区数满足要求

End

算法采用堆栈的形式遍历了关系网络图,属于一个社区的边进行了标记。算法可以发现不能划分到社区的孤立点,对应到事件集合中的某些离散值独立的事件。算法在实际运行过程中能够发现满足要求的社区。

4　实　验

实验对上述算法的正确性进行测试,同时展示发现的社区结构,并与其他算法进行比较,展示该算法的有效性。

实验选取图书馆图书借阅记录 10 条,每条记录包括借阅人员的学号,专业,借阅书籍类别。利用算法 1 计算出读者之间的关系矩阵,矩阵的值 U_{ij} 为有向边 $U_i \rightarrow U_j$ 之间的条件概率,可见 U_{ij} 不一定等于 U_{ji} ,如图 2 所示。

	U_1	U_2	U_3	U_4	U_5	U_6	U_7	U_8	U_9	U_{10}
U_1	1	0.2	0.4	0.1	0.4	0.5	0.3	0.1	0.2	0.5
U_2	0.75	1	0.5	0.25	0.75	0.5	0.75	0.75	0	0.75
U_3	0.57	0.28	1	0.14	0.57	0.57	0	0.28	0.14	0.14
U_4	0.2	0.2	0.2	1	0.2	0.4	0	0.2	0	0.4
U_5	0.57	0.43	0.57	0.14	1	0.57	0.43	0.28	0.14	0.57
U_6	0.63	0.25	0.5	0.25	0.5	1	0.25	0.13	0.25	0.5
U_7	0.38	0.38	0	0	0.38	0.25	1	0.13	0.38	0.38
U_8	0.25	0.75	0.5	0.25	0.5	0.25	0.25	1	0.5	0.25
U_9	0.66	0	0.17	0	0.17	0.33	0.5	0.33	1	0.17
U_{10}	0.63	0.38	0.13	0.25	0.5	0.5	0.38	0.25	0.13	1

图 2　用户关系矩阵

把信息熵的阈值设置为 0.5,发现的社区结构如图 3 所示。

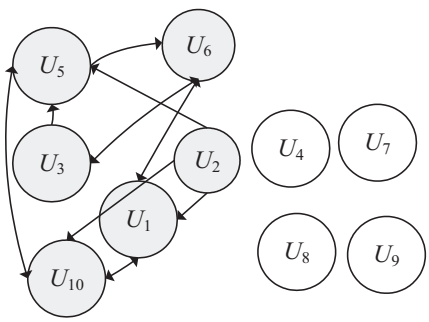

图 3　基于信息熵变化发现的社

由于不具有出入度的特征,所以不能根据出度、入度来发现社区结构。

相似度计算方法也是发现社区结构的一种常用方法。为便于对比研究,文中选用余弦相似度计算方法来发现社区结构。采用与方法 1 同样的数据,该方法采用余弦相似度计算出节点之间的相似度,并根据相似度大小发现社区结构,如图 4 所示。

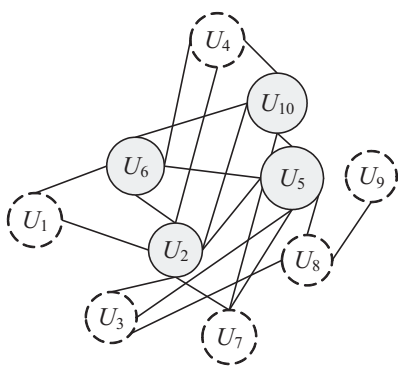

图 4　基于余弦相似度的社区结构发现

从两种方法发现的结果来看,两种结果存在一些共同点,也存在一些差异。两种方法都发现了 U_2, U_5, U_6, U_{10} 属于一个社区,方法 1 发现 U_1, U_3 与 U_5, U_6,

U_{10} 属于一个社区,不同之处在于,方法 1 发现的社区结构是有向图,方法 2 发现的结构是无向图。在实际生活中,有向图显然具有更好的科学性,描述的信息更准确。如图 3 中, U_5 到 U_6 有一条有向边,而 U_6 到 U_5 没有有向边。本实验中,如果 U_5 发生, U_6 发生的概率为 0.57,图 3 无法描述这些信息。所以方法 1 具有较大的优越性。方法 2 之所以没有发现 U_1, U_3,是因为彼此之间相似度较低,而方法 1 计算的是彼此之间的概率,所以被发现的可能性较大。

从两种方法对比来看,方法 1 能够发现更有价值的社区结果,同时对成员之间的关系进行描述,方法 2 不能做到这点。

5　结束语

提出的方法既能发现社区结构,也能发现社区内成员之间的动态模糊关系,这里用贝叶斯概率来描述成员之间的关系,采用了信息熵变化趋势来发掘社区成员。现实中,社区成员之间的关系有很多,如关联关系、因果关系、相反关系等,深入发掘这些关系并用于个性化推荐系统尤为重要。文中只用余弦相似度进行了对比,对该方法进行了改进,还需要与其他相似度计算方法进行对比。所用的样本记录远称不上大数据,在大数据环境下该方法表现如何有待进一步探索。根据熵的变化判断成员是否属于一个社区,熵阈值确定很重要,如何确定更好的阈值,使得发现的结果有更好的可用性,也是目前所有社区发现方法面临的问题。

参考文献:

[1] 丁元竹. 社区研究的理论与方法[M]. 北京:北京大学出版社,1995.

[2] 潘　磊. 若干社区发现算法研究[D]. 南京:南京大学,2014.

[3] 张　博. 有向网络的社区发现算法研究[D]. 成都:电子科技大学,2013.

[4] CHRISTAKISN A,FOWLER J H. Social network sensors for early detection of contagious outbreaks[J]. PLoS One,2010,5(9):e12948.

[5] 占文威,席景科,王志晓. 基于相似性模块度的层次聚合社区发现算法[J]. 系统仿真学报,2017,29(5):1028-1032.

[6] 窦炳琳,李澍淞,张世永. 基于结构的社会网络分析[J]. 计算机学报,2012,35(4):741-753.

[7] 谢晓芹,韩　帅,陈　敏,等. 基于社会网络的团队生成方法研究[J]. 计算机学报,2017,40(3):712-728.

[8] LEEK W R,LIM E P. Friendship maintenance and prediction in multiple social networks[C]//27th ACM conference on hypertext and social media. Halifax, NS, Canada: ACM, 2016:83-92.

[9] SHAO Junming,HAN Zhichao,YANG Qinli,et al. Community detection based on distance dynamics[C]//Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining. Sydney, NSW, Australia:ACM,2015:1075-1084.

[10] WHANG J J, GLEICH D F, DHILLON I S. Overlapping community detection using neighborhood-inflated seed expansion[J]. IEEE Transactions on Knowledge and Data Engineering,2016,28(5):1272-1284.

[11] 林友芳,王天宇,唐　锐,等. 一种有效的社会网络社区发现模型和算法[J]. 计算机研究与发展,2012,49(2):337-345.

[12] ZHANG Xiao,MEI Changlin,CHEN Degang,et al. Feature selection in mixed data:a method using a novel fuzzy rough set-based information entropy[J]. Pattern Recognition,2016,56(1):1-15.

[13] 王　刚,钟国祥. 基于信息熵的社区发现算法研究[J]. 计算机科学,2011,38(2):238-240.

[14] 彭长根,丁红发,朱义杰,等. 隐私保护的信息熵模型及其度量方法[J]. 软件学报,2016,27(8):1891-1903.

[15] MITCHELL T M. 机器学习[M]. 曾华军,张银奎,译. 北京:机械工业出版社,2003.