基于不确定图的层次聚类算法研究

李俊辉

(广西柳桂高速公路运营有限责任公司,南宁 530000)

[摘 要] 传统的基于图论的层次聚类算法都是对确定图进行分割,然而实际中,很多网络的图结构都是不确定的,边是以概率存在的。因此,本文提出了基于不确定有向加权图的层次聚类算法。该算法首先求出不确定图的可能强连通子图作为聚类中心,再对剩余节点进行层次聚类。算法主要考虑边的权重和存在概率。最后以一个例子说明算法的流程。

[关键词] 不确定图;有向加权图;聚类中心;层次聚类

doi: 10.3969/j.issn.1673-0194.2012.24.048

[中图分类号] TP311 **[文献标识码]** A **[文章编号]** 1673-0194(2012)24-0079-02

图的聚类算法是从节点开始的,一般的过程是:先是把图中的每一个节点都初始化成一个子图,图中有多少节点,就形成多少子图,设计一种子图间相似度或距离计算方法,定义子图之间的相似度或距离,以及凝聚过程的结束条件[1]。首先计算每个子图对之间的相似度,合并相似度最高,距离最近的子图成一个新的子图,再重新计算所有子图对之间的相似度或者距离数值,将此过程不断重复,直到达到过程结束条件,最终得出整个图的层次结构。聚类结果可以用树状图来表示,这种方式可以清晰看出各种不同需求下得到的划分结果。

主要的层次聚类算法包括 Girvan 和 Newman[2]提出的基于边的介数的聚类算法,Newman[3]为解决大规模网络提出的快速算法,以及 CURE 算法[4]和 Chameleon 算法[5]等,近年来,许多研究者改进了传统的层次聚类算法。例如,王小黎[6]等对相似度度量和时间复杂度进行改进,提出了基于相异度度量的凝聚聚类方法,采用曼哈坦距离作为节点相似性度量,并且证明了该方法的有效性。于慧娟[7]等提出了一种基于社团结构核心区域集的图聚类方法,社团结构核心区域集是满足某些限制条件的完全子图的集合。同时分析了聚类过程,通过实验表明了该方法能提高聚类的精度。然而目前的基于图的层次聚类算法大部分都是在改进算法的运算效率以及结果的精度,很少有考虑到图自身的结构。以往的研究都是基于图的结构是完全确定的,并没有考虑图的结构可能是变化的。诸如一些信息流网络,由于信息流并不是确定存在的,是以某种概率存在的。本文在此基础上提出了基于不确定图的层次聚类算法,并且最后以一个例证来说明算法的流程。该算法的思想是初步提出,算法的精度及有效性将会在以后的工作中呈现。

1 不确定图的相关定义

假设一个不确定有向加权图为 $G=(V,E,W,Pr)$,其中 $V=\{v_1,v_2,\cdots,v_n\}$ 是图中节点的集合,E 是网络中有向边的集合 $E=\{e_{11},e_{21},\cdots,e_{pq}\}$,$e_{pq}=(v_p,v_q)$ 表示 v_p 到 v_q 之间相连的边,m 表示网络中的边数,其中 $(v_p,v_q)\neq(v_q,v_p)$。$W=\{w_{11},w_{21},\cdots,w_{pq}\}$ 表示每条有向边的权值。$Pr=\{pr_{11},pr_{21},\cdots,pr_{pq}\}$ 表示边存在的概率,其中 $pr_{ij}\in(0,1]$。在有向图中,(v_p,v_i) 表示 v_p 的出集,(v_i,v_p) 表示 v_p 的入集。

一个不确定图可以派生出一组确定图 $G'=(V',E',W')$[8],此确定图为可能图,它满足 $V'=V,E'\in E,W'\in W$。假设不确定图不同边的概率分布是相互独立的,可能图的概率为 $Pr(G')=\prod_{e\in E'}Pr(e)\cdot\prod_{e\in E\setminus E'}(1-Pr(e))$。并且任何可能图有 $Pr(G')>0$,$\sum Pr(G')=1$。

定义 1:若 n 阶有向图 G 中至少有 $n(n-1)$ 条确定相连的边,则 G 为 n 阶确定强连通图,记为 $K_n(n\geqslant 1)$。

定义 2:若 n 阶有向图 G 中至少有 $n(n-1)$条以某个概率相连的边,则 G 为 n 阶可能强连通图。其确定强连通的概率可以定义为:$Pr(K_n)=\prod_{s,d\in V}q_{pr}(s,d)$

图中的强连通子图所在处,是网络最密集的地方,在这里的节点互相紧密连接,有时甚至有些完全子图互相重叠交错。因此,把不重叠作为聚类中心选择的限制条件之一。选择聚类中心的具体步骤如下:

(1) 求出网络中所有强连通子图作为子图集。

(2) 删除强连通概率不满足 θ 的子图。

(3) 删除子图集中重叠交错的子图。

(4) 重复步骤(2)、(3),最终获得所有聚点中心。

网络的聚类中心结果并不是唯一的,但对于稳定的聚类算法,初始聚类中心并不会使聚类结果差异太大。也就是说,不同的初始聚类节点在算法条件下具有等价性。

2 层次聚类算法

聚类中心算法仅用于初始节点的计算,获得初始聚类中心之后,计算剩余节点到各个聚类中心的聚集程度,将聚集程度最高的节点归入到相应的聚类中心,逐步归类,直到循环完所有的剩余节点,形成第二层虚拟节点或介节点。然后再重新计算介节点之间的权重和存在概率,根据聚合度合并介节点,形成图的第三层,如此循环,直到合并成一个最高层的介节点为止。

在层次聚类之前,需要先定义节点到聚类中心的凝聚程度,以便将节点归纳到凝聚程度最高的聚类中心。以往的研究大部分都将节点间的最小距离作为合并的衡量标准,然而本文中需要考虑边的权重以及存在概率,因此需要重新定义节点凝聚的衡量标准。

定义 3:给定图 $G=(V,E,W,Pr)$,其中 $V=\{v_i\},i=1,\cdots,n$,$E=\{e_{ij}\},i\neq j,e_{ij}\neq e_{ji},W=\{w_{ij}\},Pr=\{pr_{ij}\}$ 以及聚类中心的集合 $H=\{H_1,H_2,\cdots,H_k\},H_i\subset G$ 作为初始类,G 中任意一不属于 H 的节点 v 与 H_i 的凝聚程度为:

$$\mathrm{coh}(v_i,H_j)=\sum_{V_j\in H_j}(\beta w_{ij}pr_{ij}+w_{ji}pr_{ji})$$

[收稿日期] 2012-10-10

其中，w_{ij}，pr_{ij} 分别表示节点 v_i 到 v_j 的边的权重和存在概率，v_j 属于聚类中心 H_j。$\beta > 1$ 是节点出集的影响因子。

以上公式表示了节点与各个聚类中心的凝聚程度，这个程度是相对的。$\mathrm{coh}(v_j, H_j)$的值越大表示节点与该聚类中心的凝聚程度越高。如果节点到两个聚类中心的凝聚值是相同的，则随机选择一个聚类中心。决定凝聚程度的参数是边的权重和存在概率，如果两个节点之间没有边，则权重和存在概率都为 0。节点的出集和入集对凝聚程度具有不同的影响。例如在信息流网络中，一个节点调用其他节点的信息比被其他节点调用的耦合度会更高，因此在凝聚值计算中加入了参数 β。

通过上述过程，可以将初始节点聚类合并到聚类中心，形成了第二层介节点，介节点包含该类所有元节点的资源和功能。然而在第二层介节点中，需要重新计算介节点之间的权重和存在概率，以便形成第三层节点。计算方法如下：

定义 4：介节点 H_a 到 H_b 之间的出集或入集存在概率：

$$pr(H_a, H_b) = \max(pr(v_i, v_j))$$

其中，$v_i \in H_a, v_j \in H_b$。

由上式可知，出集或入集的计算是选取两个介节点中的子节点最大存在概率作为介节点的相邻存在概率。

定义 5：介节点 H_a 到 H_b 之间的出集或出集权重：

$$w(H_a, H_b) = \frac{\sum w(v_i, v_j)}{\sum e_{ij}}$$

其中，$v_i \in H_a, v_j \in H_b$。

入集权重的计算方法和出集权重的计算方法相同，都是取介节点中子节点出集或入集的权重的平均值。

权重和存在概率重新计算之后，再根据凝聚度的公式，合并凝聚度较高的介节点。两个介节点之间最多只有一个出度和一个入度，形成一个 2 阶的完全图，这种情况说明两个介节点的交互较多，合并成一个更高层级的介节点的可能性较大。这种计算方法不容易产生碎片和边缘类，比较符合信息系统的现实意义。选择凝聚度大的类逐渐合并，知道满足聚类数为止，得到最终的聚类结果。

3 算法例证

给定有向加全图 $G = (V, E, W, Pr)$，该图是有 15 个节点，42 条边构成的。如图 1 所示，图中每条边上的数值为(权重/存在概率)。

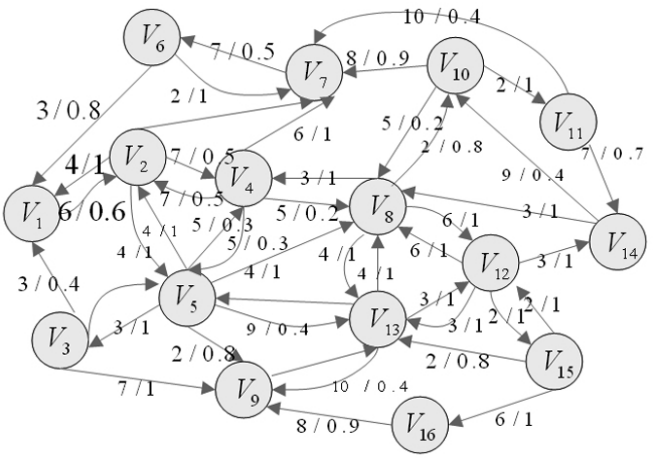

图 1 不确定有向加权图

根据层次聚类算法的步骤，首选求出图中的聚类中心，设置聚类中心每条边的存在概率阈值为 0.5。图 1 中有两个强连通图分别是 $G_1 = (v_2, v_4, v_5)$，$G_2 = (v_8, v_{12}, v_{13})$。然而在 G_1 中，$Pr(e_{45}) < 0.5$ 和 $Pr(e_{54}) < 0.5$ 不满足聚类中心的要求。因此，图中的聚类中心为 $H_1 = (v_8, v_{12}, v_{13})$。

获得聚类中心后，遍历剩余的所有节点。对某个节点而言，如果该节点的所有相邻节点中，$\mathrm{coh}(v_j, H_1)$的值最大，则将该节点并入 H_1 中，否则并入其他节点。设置 $\beta = 1.2$，计算每个节点的凝聚值，直到每个节点都与其他节点合并。有此可得，第二层的节点集合为 $v_{21} = (v_1, v_2, v_4)$，$v_{22} = (v_3, v_9)$，$v_{23} = (v_5, v_{11}, v_{14}, H_1)$，$v_{24} = (v_6, v_7, v_{10})$，$v_{25} = (v_{15}, v_{16})$

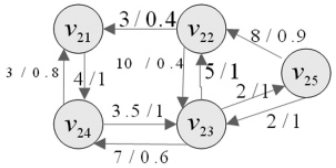

图 2 第二层图

根据图 2 进行聚类，可得第三层的节点集合：$v_{31} = (v_{21}, v_{24})$，$v_{32} = (v_{22}, v_{23}, v_{25})$。

第四层节点集合为：$v_{41} = (v_{31}, v_{32})$。至此，层次聚类算法结束。

4 总 结

传统的图聚类算法的研究都是基于结构确定图，然而，现实生活中，很多图的结构都是不确定的，边以某个概率存在。因此，本文提出了基于不确定图的层次聚类算法，该算法考虑边的存在概率，将边权重较大，且存在概率最高的节点聚合到一起。由于该算法初步提出，有很多方面未作研究，比如该如何验证聚类的结果等。下一步工作便是考虑算法的速率及精确度等。

主要参考文献

[1] 郭春艳. 基于连接度的图聚类方法研究[D]. 太原：山西大学，2008.

[2] M Girvan，ME J Newman. Community Structure in Social an Biological Networks [C] // Proceedings of the National Academy of Sciences of the United States of America，2002.

[3] M E J Newman. Fast Algorithm for Detecting Community Structure in Networks[J]. Phys Rev E，2004，69(6)：066-133.

[4] Sudipto Guha，Rajeev Rastogi，Kyuseok Shim. Cure：An efficient Clustering Algorithm for Large Databases [C] // Proceedings of the Fourth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining，1998.

[5] G Karypis，E H Han，V Kumar. Chameleon：A Hierarchical Clustering Algorithm Using Dynamic Modeling [J]. IEEE Transaction of Computer，1999，32(8)：68-75.

[6] 王小黎. 一种改进的图聚类的相异度度量方法 [J]. 计算机应用与软件，2011，28(5)：139-141.

[7] 于慧娟，崔军，毋晓志，等. 一种改进的凝聚图聚类方法[J]. 山西煤炭管理干部学院学报，2010(3).

[8] 袁野，王国仁. 面向不确定图的概率可达查询[J]. 计算机学报，2010，33(8).