

## 基于 DNA 计算的最大流算法在图聚类上的应用研究

张怀远, 向来生

(山东师范大学管理 科学与工程学院, 山东 济南 250014)

**摘 要:** 提出使用 DNA 计算解决图聚类问题, 提供了使用 DNA 两阶段法求最小切进行图分析的新思路. 在使用两阶段算法前, 首先根据一定的规则对给定图进行构造, 使其适合使用 DNA 两阶段算法. 在两阶段算法中, 使用 DNA 分子对图中顶点、边进行编码. 经过生化反应生成关于构造图从选定源节点到槽节点的所有路径, 再利用电子计算求出关于给定源节点和槽节点的最小切, 从而完成对图的划分, 然后迭代执行两阶段算法直到获得满意的聚类数目为止. 给出了算法的证明, 说明了算法的可行性.

**关键词:** DNA 计算; 图聚类; 最大流; 最小切

**中图分类号:** TP301.6

**文献标识码:** A

**文章编号:** 1000-7180(2013)04-0098-05

## The Study about Application on Graph Clustering Algorithm of Maximum Flow Based on DNA Computing

ZHANG Huai-yuan, XIANG Lai-sheng

(Management Science and Engineering College, Shandong Normal University, Jinan 250014, China)

**Abstract:** This article provides a new solution to problems about graph clustering by DNA computing, and also provides some new ideas on seeking for minimum cuts in the DNA two-stage method and afterwards future analyses a graph. Before carrying out the DNA two-stage method, some formation should be made on the given diagram to make it suitable for the premise of DNA two-stage method. Then when implementing the two-stage method, we shall encode the vertex and edge in the given graph first. After some biochemical reactions, it can generate all the routes from the source node to the sink node about the formatted graph. Then we could figure out the minimum cut of the given source node and the sink node through some computing, thus complete the division of the graph. Finally we can implement the two-stage methods iteratively until the satisfying cluster number is found. Besides, this article offers the testimony of the algorithm as well, which further illustrates its feasibility.

**Key words:** DNA computing; graph clustering; Maximum flow; minimum cut

### 1 引言

图聚类的任务就是把图中的各个顶点分成若干簇, 在这些簇的内部边的联系紧密, 在各个簇之间边的联系稀疏, 即簇内边的密度要大于两个簇之间边的密度. 图聚类的目的是使图的结构更加清晰. 图聚类的应用十分广泛, 涉及数据转换、生物和社会网络等.

随着图聚类研究的深入, 已经提出一些有特色的图聚类方法, 按其应用主要分为计算机图形学和社会学两大类. 在计算机图形学方面, 2000 年 Vempala 等人提出基于分割技术的图聚类方法, 并且引入图的谱信息, 在社会学领域, 主要是层次聚类方法.

本文将 DNA 计算技术引入图聚类, 提出了一种新的基于 DNA 计算和最大流技术的图聚类算

**收稿日期:** 2012-07-05; **修回日期:** 2012-08-06

**基金项目:** 国家自然科学基金(61170038); 山东省自然科学基金(ZR2011FM001); 软科学重大项目(2010RKMA2005); 教育部人文社会科学研究项目(2YJA630152)

法.  $s-t$  最大流问题被定义如下. 给定一个有向图  $G=(V,E)$ , 其中边的容量  $c(u,v) \in \mathbb{Z}^+$ , 两个顶点  $s, t \in V$ , 最大流就是找到能从源节点  $s$  到达槽节点  $t$  的通路, 而且必须遵循全部的容量限制.

Ford 和 Fulkerson<sup>[1-2]</sup> 提出的最大流最小切定理证明了网络的最大流问题等同于分割  $s$  和  $t$  的最小切问题. 许多多项式时间的算法可以解决最大流问题, 最大流问题的应用包括 VLSI 的设计, 调度问题, 图像的分割问题以及网络可靠性的测试.

## 2 无向图的聚类方法

### 2.1 构造适合于应用最大流算法的网络图

对于无向图  $D$  来说, 要应用 DNA 最大流算法首先要将其转换, 转换后的图称为  $D'$ , 转换的方法如下:

**步骤 1:** 首先把无向图  $D$  转换为有向图  $D_1$ , 把每一条容量为  $c(v,w)=x$  的边  $\{v,w\}$  替换为两条容量都为  $x$  的弧. 对于任意顶点对  $s,t$ , 在原图中一个切是最小  $s-t$  切, 当且仅当它在转换图中也是最小切.

**步骤 2:** 然后在图中插入一个人工槽节点  $t$ , 然后将其与图中所有的点连接, 连接边的方向指向槽节点  $t$ , 边的容量为  $k$  (对于  $k$  值的选取, 可以利用最小切的嵌套性质, 详见文献<sup>[3]</sup>), 这时产生的图的最小切等同于原图上的最小切, 证明见参考文献<sup>[4]</sup>. 对于边的容量  $c_{ij} = n_{ij}/m_{ij}$  ( $n_{ij}, m_{ij}$  互质), 令  $m$  为所有  $w_{ij}$  分母  $m_{ij}$  的最小公倍数, 则图  $D'=(V,E,mW)$  就是我们所要构造的图.

### 2.2 改进的 DNA 两阶段算法流程<sup>[5]</sup>

#### 第一阶段算法

**步骤 1:** 图  $D'$  中, 选定任一点作为源节点  $s$ , 删除  $s$  的入弧 (因为它的入弧一定不会成为增广路的一条弧). 令每条弧的长度为一个单位.

**步骤 2:** 对图  $D'$  中的顶点和弧进行等码长编码.

**步骤 3:** 随机产生图  $D'$  中从  $s$  到  $t$  所有路的 DNA 片段.

**步骤 4:** 仅保留图  $D'$  中从  $s$  到  $t$  所有路的 DNA 序列. 得到  $D'$  路集合  $H$ .

#### 第二阶段算法

介绍算法所需的基本理论:

**定理 1:** 令  $v_1, v_2 \in V$ ,  $S_1, S_2$  是在  $D'$  中关于槽节点  $t$  的社团, 那么  $S_1$  和  $S_2$  不相交或者一个是另一个的子集.

**定义 1:** 设容量网络  $D=(V,E,C)$  的容量集为

$C=\{c_{ij} \mid a_{ij} \in E(D), c_{ij} \in \mathbb{N}^+\}$ , 对  $D$  中给定的从  $x_1$  到  $x_n$  可行流  $f=\{f_{ij}\}$  构造集合  $E(f)=\{a_{ij} \text{ 和 } a_{ji} \mid a_{ij} \in E(D)\}$ ,  $C(\Delta, f)=\{c'_{ij} \text{ 和 } c'_{ji}\}$ , 其中:

$$c'_{ij} = \begin{cases} c_{ij} - f_{ij} & (c_{ij} - f_{ij} \geq \Delta \text{ 且 } \Delta c_{ij} \in c) \\ 0 & (c_{ij} - f_{ij} < \Delta \text{ 且 } \Delta c_{ij} \in c) \end{cases}$$

$$c'_{ji} = \begin{cases} f_{ij} & (f_{ij} \geq \Delta \text{ 且 } \Delta c_{ij} \in c) \\ 0 & (f_{ij} < \Delta \text{ 且 } \Delta c_{ij} \in c) \end{cases}$$

则称网络  $D(\Delta, f)=(V, E(f), C(\Delta, f))$  为网络  $D$  的  $\Delta$  松弛网络.

**命题 1** ① 设  $L_k$  为关于可行流  $f$  的一条增广路, 则沿  $L_k$  增广得到流  $f'=\{f'_{ij}\}$  为一个可行流, 且流量增加值为  $\theta$ . ② 若  $\forall a_j \in L_k$ , 使  $c'_j \neq 0$  ( $c'_j \in C(\Delta, f)$ ), 则  $L_k$  是  $D(\Delta, f)$ , 且  $L_k$  在  $D(\Delta, f)$  中的容量不低于  $\Delta$ . ③ 第一阶段算法中的路集  $H$  包含  $D(\Delta, f)$  所有从  $x_1$  到  $x_n$  路, 并且包含了流  $f=\{f_{ij}\}$  的所有增广路.

设第一阶段用 DNA 算法得到图  $D'$  路集为  $H=\{L(1), L(2), \dots, L(p)\}$ . 第二阶段求最小切的算法如下:

**步骤 1:** 令  $f_0=\{f_{ij}^{(0)} \mid f_{ij}^{(0)}=0\}$ ,  $\theta(0)=0$ , 赋初值:  $k=0, s=1, q=1$ , 按照定义 1 计算  $D(\Delta(1), f_0)$ , 其中  $\Delta(1)=(\frac{1}{2}) \cdot \min\{\sum_{a_{ij} \in E(D)} c_{ij}, \sum_{a_m \in E(D)} c_m\}$ ,  $t=1$ .

**步骤 2:** 若  $\exists a_j \in L(s)$ , 使  $c'_j=0$  ( $c'_j \in C(\Delta, f_k)$ ), 则转步骤 3. 否则转步骤 4.

**步骤 3:** 若  $s < p$ , 则令  $s=s+1, q=q+1$ , 则转步骤 2; 若  $s=p$  且  $q < p$ , 则令  $s=1, q=1$ , 转步骤 2; 若  $s=p$  且  $q=p$ , 则转步骤 5.

**步骤 4:** 计算  $L(s)^+=\{a_j \mid a_j \in L(s) \cap E(D)\}$ ;  $L(s)^-=L(s)-L(s)^+$ ;  $\theta(k+1)=\min_{a_{ij} \in L(s)}(c'_{ij})$ , 其中  $c'_{ij} \in C(\Delta(t), f_k)$ ,  $f_k=\{f_{ij}^{(k)}\}$  为第  $k$  次迭代的可行流, 于是由

$$f_{ij}^{(k+1)} = \begin{cases} f_{ij}^{(k)} + \theta(k+1) & (a_{ij} \in L(s)^+) \\ f_{ij}^{(k)} - \theta(k+1) & (a_{ij} \in L(s)^-) \\ f_{ij}^{(k)} & (a_{ij} \notin L(s)) \end{cases}$$

可得到新的可行流  $f_{k+1}=\{f_{ij}^{(k+1)}\}$ , 按照定义 1 计算  $D(\Delta(t), f_{k+1})$ , 令  $k=k+1, s=s+1$ , 转步骤 2.

**步骤 5:** 若  $\Delta(t) < 1$ , 则可行流  $f_k=\{f_{ij}^{(k)}\}$  为最大流, 转步骤 6. 否则, 取  $\Delta(t+1)=\Delta(t)/2$ , 计算  $D(\Delta(t+1), f_k)$ , 令  $t=t+1, s=1, q=1$ , 转步骤 2.

**步骤 6:** 记录下步骤 2 中提到的  $a_j$ , 则  $a_j$  就是切

集中的元素.

**步骤 7:** 求出切集中的所有元素, 则由切集分出来的顶点集合  $C$  就是源  $s$  所在的社团.

**步骤 8:** 假如在某一节点  $v$  和槽节点  $t$  之间的最小切产生了社团  $S$ , 那么在以后的聚类中可以不使用社团  $S$  中的顶点作为源节点, 因为根据定理它们不能产生更大的社团. 我们把它们标记到社团  $S$  内, 随后如果  $S$  变为一个更大的社团  $S'$  的一部分, 那么就把社团  $S$  中的全部节点标记为社团  $S'$  内的节点. 因此, 在社团  $C$  的外部选择一个节点作为源节点  $s$ , 重复第二阶段的算法直到所有的节点都被归入社团内部. 算法结束.

如果以上聚类结果中社团数目比期望的社团数目要多很多, 还可以采用层次聚类的方法对其进行进一步进行聚类.

**步骤 1:** 把各社团收缩为一个点, 把连接各个社团之间的边合并为一条, 边的容量为各边的容量之和.

**步骤 2:** 重复上述构造图  $D'$  的过程, 执行算法的第一和第二阶段, 这样原来的社团相互合并成为更大的社团. 社团的数目也随之减少.

### 3 有向图的聚类方法

#### 3.1 构造适合于应用最大流算法的网络图

对于有向图  $D$  来说, 要应用 DNA 最大流算法也要将其转换, 转换后的图称为  $D'$ , 转换的方法如下:

**步骤 1:** 对于有向图  $D$  中的每条弧增加一条与其方向相反的弧.

**步骤 2:** 然后在图中插入一个人工槽节点  $t$ , 然后将其与图中所有的点连接, 连接边的方向指向槽节点  $t$ , 边的容量为  $k$ , 其中  $1 < k < N_i/c$  ( $c$  表示原始切集(即不包含连接到槽节点  $t$  上的边的容量)的总容量,  $N_s$  表示源节点  $s$  所在的社团  $C$  中的节点个数,  $N_i$  表示社团  $V-C$  中节点的个数), 这时产生的图的最小切等同于原图上的最小切, 证明见参考文献<sup>[4]</sup>. 对于边的容量  $c_{ij} = n_{ij}/m_{ij}$  ( $n_{ij}, m_{ij}$  互质), 令  $m$  为所有  $w_{ij}$  分母  $m_{ij}$  的最小公倍数, 则图  $D' = (V, E, mW)$  就是所要构造的图.

#### 3.2 算法的流程

有向图的聚类方法和无向图类似, 只是在算法的第一阶段的步骤 1 稍作修改, 更改后结果如下: 选择任意节点为源节点, 删除源节点的入弧. 其他算法步骤与无向图的相同.

## 4 算法的正确性分析及生化实验操作

### 4.1 算法的正确性分析

下面命题说明第二阶段算法是正确的.

**命题 2** 第二阶段算法的步骤 5 得到的流  $f_k = \{f_{ij}^{(k)}\}$  是最大流问题的最优解.

**证明** 因为  $f_k = \{f_{ij}^{(k)}\}$  经过  $k$  条增广路增广零流  $f_0$  得到, 由命题 1 的 ①,  $f_k$  为可行流.

从算法步骤 2, 3 和命题 1 的 ②、③ 可知道, 对给定的  $\Delta$  在  $H$  上搜索出全部容量超过  $\Delta$  的增广路来增广流量. 只有当没有这样的增广路时, 才将  $\Delta$  的数值减少一半, 再继续在  $H$  上搜索. 因此当步骤 5 的条件达到时, 即  $\Delta(t) < 1$ , 在  $H$  上已经没有容量可以超过 1 的增广路了. 这一点由  $H$  的全体增广路性质可以保证.

由整流定理, 正整数容量网络  $D$  的最优解为整数, 因此增广路的容量为正整数, 这样, 当  $\Delta(t) < 1$  时实际上  $H$  在网络  $D$  上已经没有增广路, 从而, 已达到最大流. 而最大流对应的切集必然是最小切集, 所以, 步骤 6 中的切集是最小切集.

### 4.2 算法的生化实验操作

**步骤 1:** 构造顶点的 DNA 编码.

对  $V(D)$  中全部顶点  $x_i$  用不同的 DNA 编码与之对应, 如点  $x_i$  用  $p(i)q(i)$  ( $i = 1, 2, \dots, n$ ) 表示. 取  $|p(i)| = |q(i)| = l$ , 则  $|p(i)q(i)| = 2l$ . 注意: 用保护基对  $p(1)q(1)$  的 3' 端的羟基和  $p(n)q(n)$  的 5' 端的磷酸基进行活化处理, 使其失去化学活性.

**步骤 2:** 构造弧的编码.

对  $E(D)$  中每条弧  $a_{ij}$  用不同的 DNA 片段与之对应, 如弧  $a_{ij}$  用  $q'(i)p'(j)$  表示. 式中  $p'(i), q'(i)$  为  $p(i), q(i)$  补链.

**步骤 3:** 构造连接反应并固定在表面上.

将上面制作寡聚核苷酸  $p(i)q(i)$  (过量) 和 DNA 片段  $q'(i)p'(i)$  混合后放在试管中. 在将试管中的混合物缓慢升温至液相温度达  $94^\circ\text{C}$ , 充分反应后让试管进行缓慢冷却. 这样, 退火后的混合物包含起始于  $p(1)q(1)$  终止于  $p(n)q(n)$  的 DNA 片段和一些尚未反应的物质.

去掉 DNA 片段两端的保护基. 对反应物进行凝胶电泳, 按链长搜集 DNA 片段, 并去掉链长大于  $2l \times$  顶点数的 DNA 片段和小于  $4l$  的 DNA 片段. 将生物素分子连接到 DNA 片段  $p(1)q(1)$  的 3' 端, 再将 DNA 片段按照电泳收集的次序依次固定在经过修饰后的表面上, 并进行编号. 从而表面上按排列顺序

包含了图  $D$  起点为  $x_1$  终点为  $x_n$  长度不超过  $2l \times$  顶点数的所有链的序的 DNA 片段。

**步骤 4:**对表面 DNA 片段加热,解开双链,并冲洗掉表面上的所有补链。将表面 DNA 单链片段和具有荧光素的寡聚核苷酸  $p'(2)q'(2)$  均匀混合后,升温至  $94^{\circ}\text{C}$  再缓慢冷却,进行杂交反应。对反应后表面进行彻底清洗,除去未杂交的寡聚核苷酸片段。用激光共聚焦显微镜观察表面,对没有或只有 1 个亮点的 DNA 片段进行标记,并记录下不亮的 DNA 片段的位置。大于或等于两个亮点的 DNA 片段代表的链必定不是图  $D$  的路,在下一次实验中不用再观察和记录这些 DNA 片段。

**步骤 5:**对  $p(3)q(3)$  到  $p(n-1)q(n-1)$  重复进行步骤 4 的实验,并记录下这  $n-2$  次实验中均没有或只有 1 个亮点的 DNA 片段,这些片段代表的图  $D$  的链就是图  $D$  的从  $x_1$  到  $x_n$  的所有路。

**步骤 6:**对  $p(1)q(1)$  和  $p(n)q(n)$  进行实验,只记录下有一个亮点的 DNA 片段。

5 算例

下面以图 1 所示的简单无向图来说明两阶段算法的有效性和正确性。

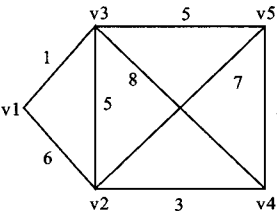


图 1 示例图

按照前面所述的方法构造适合于应用最大流算法的网络图,如图 2 所示。

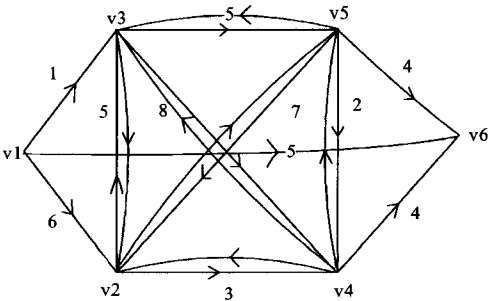


图 2 构造后的图

按照前面所述编码要求,对图 2 的顶点进行编码如表 1 所示。

表 1 顶点的编码

| Vertex | Oligonucleotide Encoding |
|--------|--------------------------|
| 1      | 3'-ACACTGTAGC-5'         |
| 2      | 3'-ACCAGCTGGT-5'         |
| 3      | 3'-AGAGCATCTC-5'         |
| 4      | 3'-TCAGTCGATC-5'         |
| 5      | 3'-ACAGACTGAC-5'         |
| 6      | 3'-TACGTCAGTC-5'         |

按照前面所述编码要求,对图 2 的边进行编码如表 2 所示。

表 2 边的编码

| edge            | Oligonucleotide Encoding |
|-----------------|--------------------------|
| e <sub>12</sub> | 5'-CATCGTGGTC-3'         |
| e <sub>13</sub> | 5'-CATCGTCTCG-3'         |
| e <sub>16</sub> | 5'-CATCGATGCA-3'         |
| e <sub>23</sub> | 5'-GACCATCTCG-3'         |
| e <sub>24</sub> | 5'-GACCAAGTCA-3'         |
| e <sub>25</sub> | 5'-GACCATGTCT-3'         |
| e <sub>32</sub> | 5'-TAGAGTGGTC-3'         |
| e <sub>34</sub> | 5'-TAGAGAGTCA-3'         |
| e <sub>35</sub> | 5'-TAGAGTGTCT-3'         |
| e <sub>42</sub> | 5'-GCTAGTGGTC-3'         |
| e <sub>43</sub> | 5'-GCTAGTCTCG-3'         |
| e <sub>45</sub> | 5'-GCTAGTGTCT-3'         |
| e <sub>46</sub> | 5'-GCTAGATGCA-3'         |
| e <sub>52</sub> | 5'-GACTGTGGTC-3'         |
| e <sub>53</sub> | 5'-GACTGTCTCG-3'         |
| e <sub>54</sub> | 5'-GACTGAGTCA-3'         |
| e <sub>56</sub> | 5'-GACTGATGCA-3'         |

反应后的一条 DNA 双链如图 3 所示,它表示的路径是 1→6。



图 3 DNA 双链

第一阶段算法结束后得到图 2 中从顶点 1 到顶点 6 的所有路径,然后便可执行第二阶段算法,得到如图 4 所示的聚类效果图。

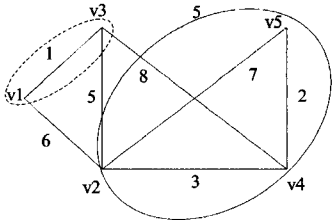


图 4 聚类效果图

## 6 结束语

本文提出的算法是应用在有向图上的,因为不管是对网络图进行聚类还是对现实中的公路运输图进行聚类,它们都是有方向的,这更符合实际,而且该算法采用 DNA 计算与电子计算想结合的算法,用 DNA 计算机完成算法中大量并行计算,然后将算法中操作步骤过于复杂的计算交给电子计算来完成. 计算速度更快,计算结果更精确. 可见, DNA 计算对于图聚类具有较大的应用以及理论价值. 文中提出的算法试验证明是可行的.

### 参考文献:

- [1] L R Ford Jr, D R Fulkerson. Maximal flow through a network[J]. Canadian J Math, 1956(8):399 - 404.
- [2] T H Cormen, C E Leiserson, R L Rivest. Introduc-

tion to algorithms[M]. 6th edition, MIT Press and McGraw-Hill Book Company, 1992.

- [3] G Gallo, M D Grigoriadis, R E Tarjan. A fast parametric maximum flow algorithm and applications[J]. SIAM Journal of Computing, 1989, 18(1):30-55.
- [4] G W Flake, R E Tarjan, K Tsioutsoulis. Graph clustering and minimum cut trees[J]. Internet Mathematics 2004, 1(1):385 - 408.
- [5] 周 康, 王子成, 许 进. 最大流的 DNA 两阶段算法[J]. 华中科技大学学报: 自然科学版, 2005, 33(8): 104-107.

### 作者简介:

张怀远 男, (1987—), 硕士. 研究方向为图聚类.

向来生 男, (1957—), 研究员, 博士. 研究方向为区域可持续发展.

(上接第 97 页)

- [3] Inyeol Lee, Gyudong Kim, Wonchan Kim. Exponential curvature-compensated BiCMOS bandgap references [J]. IEEE J Sol Sta Circ, 1994, 29(11):1396-1403.
- [4] TSIVIDIS Y P. Accurate analyses of temperature effects in  $I_C$ - $V_{BE}$  characteristics with application to bandgap reference sources[J]. IEEE J Sol Sta Circ, 1980(15):1076-1084.
- [5] 宋晶, 刘诗斌, 张勇, 等. 一种具有高稳定性的带隙基准电路[J]. 微电子学, 2007, 37(3):444-447.

- [6] Allen P E, Holberg D R. CMOS analog circuit design [M]. (中文版). 北京: 电子工业出版社, 2011.

### 作者简介:

盛健健 男, (1986—), 硕士. 研究方向为数模混合集成电路设计.

张国俊 男, (1960—), 教授. 研究方向为半导体功率器件、模拟集成电路、微细加工技术.