Journal of Computer Applications
计算机应用,2013,33(2):308 -310,315
ISSN 1001-9081
CODEN JYIIDU
2013-02-01
http://www.joca.cn

文章编号:1001 -9081(2013)02 -0308 -03
doi:10.3724/SP.J.1087.2013.00308

基于粘贴和 2-臂 DNA 模型的层次聚类算法

白　雪*,任晓玲,刘希玉
(山东师范大学 管理科学与工程学院,济南 250014)
(*通信作者电子邮箱 sdnubaixue@163.com)

摘　要:为了充分利用 DNA 分子在生物计算中的高度并行性和强大的存储能力,将 DNA 计算引入层次聚类实现对数据集的全局搜索。提出了粘贴模型与 2-臂 DNA 分子相结合的混合模型求解最近邻层次聚类的 DNA 算法。针对二维数据空间,算法首先基于最小生成树思想产生图的边的所有组合链;其次筛选含 $n-1$ 条边的链,基于边附着顶点,并选择包含全部顶点的复合链;再将复合链末尾连接相应边的权值片段,电泳出最短链;最后通过荧光分析法读解,得到最终的聚类结果。与已有文献同类算法对比表明,该算法在保持多项式操作时间下,更充分考虑连接边的长度,并将读解步骤数限定为常数步。

关键词:DNA 计算;层次聚类;最小生成树;粘贴模型;2-臂 DNA 分子
中图分类号: TP301.6　　**文献标志码**:A

Hierarchical clustering algorithm based on sticker and 2-armed DNA model

BAI Xue*, REN Xiaoling, LIU Xiyu
(*School of Management Science and Engineering, Shandong Normal University, Jinan Shandong* 250014, *China*)

Abstract: In order to take full advantage of the high parallelism and huge storage capacity of DNA molecules in biological computing, this paper introduced DNA computing into hierarchical clustering to do global research on data set. For realizing the nearest neighbor hierarchical clustering, an algorithm combining sticker model with 2-armed DNA molecules was put forward. Based on the idea of MST (Minimum Spanning Tree), the first thing to do was generating complex DNA strands of all combinations of edges and then screening those containing $n-1$ edges. Based on the edges, it is needed to set the corresponding vertex stickers and keep those strands covering all the vertices. After that, weight strands constructed by 2-armed molecules would be appended at the end of the complex strands and the shortest ones could be detected by gel electrophoresis. Finally, by fluorescence analysis the clustering result can be got. In computer simulation, this algorithm may take different lengths of edges into account instead of varying the polynomial time complexity and the number of steps to read final results is set as a constant.

Key words: DNA computing; hierarchical clustering; Minimum Spanning Tree (MST); sticker model; 2-armed DNA molecule

0　引言

作为 DNA 计算应用的主要模型,粘贴模型自 1996 年提出以来成功解决了多种 NP 问题[1-3],在求解装箱问题[4]、诱导路径问题等具有实际应用背景的数学问题时显示出一定的优越性,对其他相关领域的理论探索起到了一定程度的启发作用[5-6]。在粘贴模型中,存储链是一条含有多个不重叠子链的单链,粘贴链是与子链互补的 DNA 片段,常用的生物操作包括合并、分离、设置和清除。

2-臂 DNA 分子是 k-臂 DNA 分子的一种,k-臂 DNA 分子在自然界中[7]已经存在,例如大肠杆菌同源重组过程中有一种叫作 Holliday 的中间体结构。研究表明,在技术上能获得相当稳定的特殊设计的 2-臂 DNA 分子[8],被广泛应用于 DNA 计算研究中,多数情况下作为顶点块的基本构造单元使用[9]。2-臂 DNA 分子的 3′端为单螺旋延伸,由于两端都是黏性缺口链,使其可以根据需要逐步连接新的 DNA 分子。本文正是利用了 2-臂 DNA 分子这个特点,将其与粘贴模型结合使用,设计了解决层次聚类的 DNA 算法。

聚类问题是数据挖掘中的主要部分,其中层次聚类首先由 Kaufman 于 1990 年提出,包括凝聚层次聚类和分裂层次聚类两种。作为自底向上的分析方法,凝聚层次聚类首先将每个对象作为一个簇,通过相继合并相近的对象,直到所有对象合并为一个簇,或者达到终止条件为止,近年来在计算机网络中的 Web 服务[10]、文件数据集的处理[11]、神经网络的改进设计以及复杂网络社团发现中有着较为广泛的应用。

1　问题域转化

本文算法设计针对二维数据平面上的数据点进行聚类。凝聚层次聚类中最简单的是最近邻聚类算法,其样本点间的距离使用欧氏距离度量,类间距采用两类中相邻最近的两样本点的欧氏距离度量。算法首先将每个样本点视为一类,合并最近两样本点,随后通过计算类与类之间的距离不断合并并

收稿日期:2012-08-13;**修回日期**:2012-09-30。　**基金项目**:国家自然科学基金资助项目(61170038);山东省自然科学基金资助项目(ZR2011FM001);教育部人文社会科学研究项目(12YJA630152);山东省社会科学基金资助项目(11CGLJ22)。

作者简介:白雪(1989 -),女,北京人,硕士研究生,主要研究方向:DNA 计算、聚类、信息管理;　任晓玲(1988 -),女,山东威海人,硕士研究生,主要研究方向:DNA 计算、聚类;　刘希玉(1964 -),男,山东莱芜人,教授,博士研究生,主要研究方向:信息管理与电子商务、计算智能、计算机辅助创新设计。

最近的两个类,直到所有样本点合并为一个类为止[12]。

根据上述算法的思想将层次聚类问题转化为图论中的最小生成树问题:样本点对应图中的顶点,两点之间的欧氏距离对应图中两顶点间边的权值,从而得到一个赋权无向完全图。凝聚层次聚类问题就可以进一步转化为寻找图中的最小生成树问题。具体描述如下:

给定一个赋权无向完全图 $G=(V,E)$,其中顶点集为 $V=\{v_1,v_2,\cdots,v_n\}$,边集为 $E=\{e_1,e_2,\cdots,e_m\}$,边 e_i 的权值为 $w_i(w_i\geqslant 0,\ i=1,2,\cdots,m)$。若存在 $T=(V',E')(V'=V,E'\subseteq E)$ 为包含 n 个顶点,$n-1$ 条边的连通图,且为无循环图,使得 $w(T)=\sum_{e_i\in T}w_i$ 最小,则 T 为 G 的最小生成树。预先设定好一个阈值 τ,若边的权值大于该阈值则分割其连接的两顶点,若图中有 $k-1$ 条边被分割,则最终得到 k 个类的聚类结果。

2　DNA 编码设计

根据粘贴模型原理,由无向完全图 $G=(V,E)$ 设置一条长为 $((m+n+1)\times k)$ 位的存储链表示一个组合形式,其中前 m 位对应“边”的栈,接下来 n 位对应“顶点”的栈,最后一位是与2-臂分子黏性末端互补的寡聚核苷酸片段,用 A 表示。若生成树中包含边 $e_i=v_{i1}v_{i2}$,则存储复合物的第 i 个子链为“开”,同时第 $m+i_1$ 和 $m+i_2$ 个子链也设置为“开”,如图 1 所示。

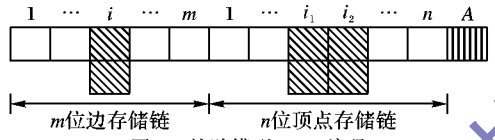
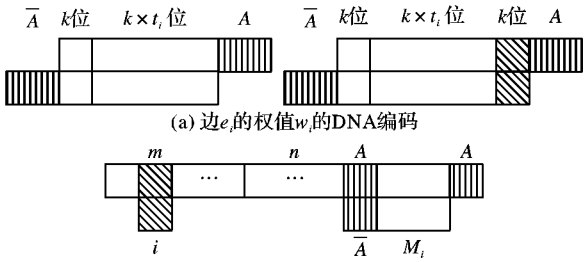


图 1　粘贴模型 DNA 编码

在构造粘贴模型的基础上,根据文献[13]将权值转化为整数,从而利用2-臂 DNA 分子 M_i 表示边 e_i 的整数权值 w_i。设定权值的阈值 τ,并将 w_i 按非递减顺序排序,令 t_i 为权值的序号。如图2(a)所示,M_i 包括以下四个部分:第一部分为 A 的补序列 $\overline{A}$;第四部分为寡聚核苷酸片段 A,在反应过程中起到连接作用;第二部分为边 e_i 对应的 DNA 片段;第三部分用长为 $k\times t_i$ 个碱基对表示 e_i 的权值。若 $w_i>\tau$,则在其后添加 k 位识别序列以作标记。若生成树中包含边 e_i,则将权值片段 M_i 连接到粘贴链的末端,以便求得代表最小生成树的 DNA 片段,如图2(b)所示。



(a) 边 e_i 的权值 w_i 的DNA编码

(b) 求解最小生成树问题的DNA编码

图 2　2-臂与混合模型 DNA 编码

3　算法实现

将最近邻层次聚类转化为最小生成树问题后,基于已成熟的生物操作实验,假设所有数据都进行正确的 DNA 编码,并且分子生物技术可以进行以下五种操作的组合,利用粘贴与 2-臂 DNA 模型求解该问题的基本算法及生物算法伪代码如下:

3.1　粘贴模型生物操作

Separate($T,T+,T-,i$):按照第 i 位的“开”、“关”情况将 T 中的链分在两个试管中,$T+$ 中盛放位置 i 为“开”DNA 链,$T-$ 则相反;

Merge(T_1,T_2):将试管 T_1 和 T_2 中的 DNA 分子合并到一起;

Detect(T):检验试管 T 中是否存在 DNA 分子;

Append(T,i):使用位置 i 对应的互补粘贴链将 T 中所有 DNA 链的位置 i 置为“开”;

Discard(T):清除试管 T 中的分子;

Connect(T,M):在 T 中所有分子的后面通过互补连接一段序列 M。

3.2　基本算法

1)生成图 G 的 $(n-1)$-边导出子图;

2)得到无向完全图 G 的生成树;

3)得到图 G 的最小生成树;

4)读出解序列,获得最终聚类结果。

3.3　生物算法

如下算法中,1) ~ 11)行从包含 m 条边所有可能组合的初始数据池中选出含有 $n-1$ 条边,即前 m 位中有 $n-1$ 位为“开”,其余位为“关”的存储复合物;12) ~ 17)行将第 $i(1\leqslant i\leqslant m)$ 位为“开”的链的第 $m+i_1$ 和 $m+i_2$ 位也设置为“开”,其中 $m+i_1$ 和 $m+i_2$ 表示第 i 条边连接的两个顶点;18) ~ 29)行筛选出第 $m+1$ 至 $m+n$ 位均为“开”的存储混合物,至此获得图 G 的 $(n-1)$-边导出子图,即图 G 的生成树。

30) ~ 34)行在存储复合物后连接权值片段,即若边 e_i 在生成树中,则将其权值对应的 2-臂 DNA 分子 M_i 连接至反应物末端;35)行将缓冲液与连接酶加入最终试管 T,使单双链混合结构全部成为双链结构;36)行使用凝胶电泳技术筛选出最轻的双链 DNA 分子,至此求得图 G 的最小生成树;37)行利用聚合酶链式反应(Polymerase Chain Reaction,PCR)扩增和 DNA 纳米金标识的探针识别出最小生成树包含的 $n-1$ 条边及其中大于阈值的边,使用荧光分析法读取颜色,若有 $k-1$ 个阈值片段的颜色,说明点集合被分为 k 个簇,同时通过不同的颜色得到 $n-1$ 条边的构成,进而还原成一棵树,得到所有顶点的聚类结果,具体流程如图 3 所示。

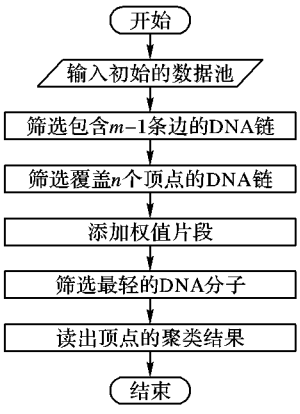


图 3　生物算法流程

```
1)  input:粘贴链前 m 位所有可能组合的链库 T0
2)  for i = 0 to m - 1
3)    for j = i down to 0
4)      Separate(Tj, T+, T-, i + 1)
                    //将试管 Tj 中的 DNA 分子按照第 i + 1 位的
                    //“开”“关”情况分别置于 T+ 和 T- 试管中
```

```
5)    Merge(T+, T_{j+1})
6)    Merge(T-, T_j)
7)  end for
8)  end for
9)  if (Detect(T_{n-1}) = true) then
        //如在试管 T_{n-1} 中存在满足筛选要求的 DNA 分子,
                                    //则执行下一步操作
10)   Merge(T_{n-1}, T)
11) end if
12) for i = 1 to m
13)   Separate(T, T+, T-, i)
14)   Append(T+, m + i_1)
        //将试管 T+ 中 DNA 链的第 m + i_1 位设置为"开"
15)   Append(T+, m + i_2)
16)   Merge(T+, T-, T)
17) end for
18) Discard(T_0)                          // 清空试管 T_0
19) Merge(T, T_0)
20) for i = 0 to n - 1
21)   for j = i down to 0
22)     Separate(T_j, T+, T-, m + i + 1)
23)     Merge(T+, T_{j+1})
24)     Merge(T-, T_j)
25)   end for
26) end for
27) if (Detect(T_n) = true) then
28)   Merge(T_n, T)
29) end if
30) for i = 1 to m
31)   Separate(T, T+, T-, i)
32)   Connect(T+, M_i)
                    //在试管 T+ 中第 i 位为"开"的 DNA 链后
                    //连接相应的权值片段 M_i
33)   Merge(T+, T-, T)
34) end for
35) Add buffer to T
36) Select the lightest from T
37) Read the sequences
```

4 算例

如图 4 所示，无向赋权完全图 $G=(V,E)$，其中 $V=\{v_1,v_2,v_3,v_4\}$，$E=\{e_1,e_2,e_3,e_4,e_5,e_6\}$，权值 $W=\{w_1,w_2,w_3,w_4,w_5,w_6\}=\{1,2,4,9,6,8\}$，阈值 $\tau=4$。

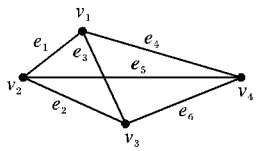


图 4 4 个顶点的无向简单图 G

首先对粘贴模型以及 2-臂分子结构 M_i 进行编码，其中 6 条边的编码如图 5(a)所示；其次通过筛选得到包含 3 条边和 4 个顶点的复合链，此时的 DNA 链所包含的边与顶点可以形成图 4 的生成树结构；为了方便筛选，使用缓冲液与连接酶将上步得到的混合结构全部变为双链，并通过凝胶电泳得到最短的双链 DNA 分子代表原图的最小生成树，其中未加入缓冲液之前的最小生成树结构如图 5(b)、(c)所示；最后通过读取图 5(b)、(c)的序列，发现边 e_1、e_2、e_5 为树的连接边，并且只有 e_5 部分的阈值链发光，从而可知原顶点集被聚为两类：顶点 v_1，v_2，v_3 为一类，v_4 单独作为一类。

e_1 -GGTG- e_4 -TTCC-
e_2 -GGCC- e_5 -CGCG-
e_3 -ACCA- e_6 -CTGT-

(a) 图G中6条边的编码

e_1	e_2	e_3	e_4	e_5	e_6	v_1	v_2	v_3	v_4	A
GGTG	GGCC	ACCA	TTCC	CGCG	CTGT					
CCAC	CCGG			GCGC						

(b) 筛选得到的混合链的前半段

	GGTG	GGGG	A	GGCC	GGGG	GGGG	A	CGCG	GGGG	GGGG	GGGG	GGGG	TTTT	A
	CCAC	CCCC		CCGG	CCCC	CCCC		GCGC	CCCC	CCCC	CCCC	CCCC	AAAA	

(c) 筛选得到的混合链的后半段,即连接在图(b)后的DNA链

图 5 图 G 中的边及最优解的 DNA 编码

5 结语

由于初始解空间生成边集的所有组合链，故其空间复杂度为指数级。像其他 DNA 算法一样，指数级的空间复杂度意味着当问题变量的规模达到某一值之后，所需的初始链会超过目前技术可操作的范围，所以该算法只能适用于一定的规模或者较小数据点的聚类情况。

尽管所需的初始解空间为指数级别，但该算法求解过程的复杂度为多项式级别，解的读取步骤为常数步。由于不同大小的金属纳米粒子能被单一波长的光激发而发出不同颜色的光，并且光稳定性高，不易降解[14]，所以可以用此方法标记边的探针，用来最后一次性检测出边集的构成。此外，上述算法在检测顶点链时采用粘贴模型中标准的迭代检测，但实际上可借用文献[15]中的半自动化手段经过一次操作就可以将顶点链部分没有完全形成双链的 DNA 分子删除，在最终池中得到满足条件的 DNA 结构。对于更大规模的二维聚类点集，可考虑首先使用网格方法处理成小规模的数据块再进行聚类。最后，该算法将 DNA 粘贴模型与 2-臂 DNA 分子模型相结合，充分利用了两种模型的结构特点，由于两种模型都是单双链结构，因此其操作手段灵活丰富并且简单易行，使该模型具有很大的应用弹性和可操作性。

参考文献：

[1] BRAICH R S, CHELYAPOV N, JOHNSON C, *et al*. Solution of a 20-variable 3-SAT problem on a DNA computer [J]. Science, 2002, 296(5567): 499 - 502.

[2] DAREHMIRAKI M , NEHI H M . Molecular solution to the 0-1 knapsack problem based on DNA computing [J]. Applied Mathematics and Computation, 2007, 187(2): 1033 - 1037.

[3] RAZZAZI M, ROAYAEI M. Using sticker model of DNA computing to solve domatic partition, kernel and induced path problems [J]. Information Sciences, 2011, 181(17): 3581 - 3600.

[4] ALONSO SANCHES C A, YOSHIHIRO SOMA N. A polynomial-time DNA computing solution for the bin-packing problem [J]. Applied Mathematics and Computation, 2009, 215(6): 2055 - 2062.

[5] SAKAKIBARA Y. Development of a bacteria computer: from in silico finite automata to in vitro and in vivo [C]// CiE'10: Proceedings of the Programs, Proofs, Process and 6th International Conference on Computability in Europe, LNCS 6158. Berlin: Springer-Verlag, 2010: 362 - 371.

(下转第 315 页)

表 1　三种算法的收敛次数和平均收敛值

函数	理论最优值	标准 BPSO 算法		EPSO 算法		带经验因子的 BPSO 算法	
		平均收敛值	最优值次数	平均收敛值	最优值次数	平均收敛值	最优值次数
f_1	78.6	78.144 5	9	78.422 4	121	78.426 8	137
f_2	157.28	156.36	2	156.98	128	157.18	889
f_3	3 095.93	3 848.33	1	3 882.36	207	3 882.02	14
f_4	55	51.102	1	53.971	192	53.971	241
f_5	80.7	80.7	667	80.5	356	80.7	989
f_6	25	24.951	6	24.951	141	24.989	163

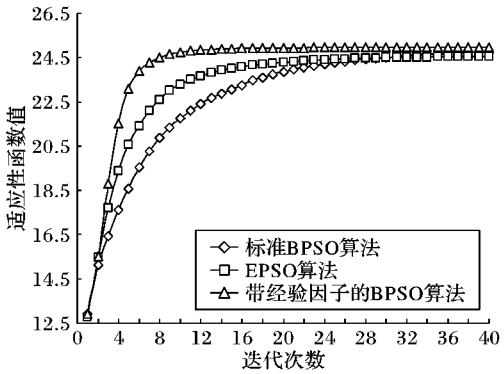


图 12　三种算法在 f_6 上的比较

4　结语

在标准二进制粒子群算优化法的基础上,充分利用了粒子群在寻优过程中产生的历史信息,提出一个带经验因子的二进制粒子群优化算法。该算法引入了记忆因子、历史遗忘系数、经验权重,并结合赏罚机制记录下各粒子每一维度取值的变化情况,引导粒子下次迭代的取值,从而防止算法过早收敛并提高算法的寻优效率。仿真实验结果表明,带经验因子的 BPSO 算法无论在收敛速度还是全局搜索能力上,都能达到更好的效果。

参考文献:

[1] KENNEY J, EBERHART R C. Particle swarm optimization [C]// Proceedings of the IEEE International Conference on Neural Networks. Piscataway: IEEE, 1995, 4: 1942 – 1948.

[2] 刘维亭,范洲远.基于混沌粒子群算法的无线传感器网络覆盖优化[J].计算机应用,2011,31(2):338 – 340.

[3] 汪楚娇,夏士雄,牛强. 免疫粒子群算法及其在矿井提升机故障诊断中的应用[J].电子学报,2010,38(2A):94 – 98.

[4] KENNEY J, EBERHART R C. A discrete binary version of the particle swarm algorithm [C]// Proceedings of the IEEE International Conference on Systems, Man and Cybernetics. Piscataway: IEEE, 1997, 5:4104 – 4108.

[5] TAŞGETIREN M F, LIANG Y-C. A binary particle swarm optimization algorithm for lot sizing problem[J]. Journal of Economic and Social Research, 2005, 5(2): 1 – 20.

[6] ANGHINOLFI D, PAOLUCCI M. A new discrete particle swarm optimization approach for the single-machine total weighted tardiness scheduling problem with sequence-dependent setup times [J]. European Journal of Operational Research, 2009, 193(1):73 – 85.

[7] PAN Q K, WANG L. No-idle permutation flow shop scheduling based on a hybrid discrete particle swarm optimization algorithm [J]. The International Journal of Advanced Manufacturing Technology, 2008, 39(7/8): 769 – 807.

[8] ONWUBOLU G C, BABU B V. New optimization techniques in engineering [M]. Berlin: Springer-Verlag, 2004.

[9] 王真. 基于离散粒子群的组合拍卖竞胜标确定问题研究[D].广州:中山大学,2009.

[10] 谭皓,王金岩,何亦征,等.一种基于子群杂交机制的粒子群算法求解旅行商问题[J].系统工程,2005,23(4):83 – 87.

[11] 周驰,高亮,高海兵.基于 PSO 的置换流水车间调度算法[J].电子学报,2006,34(11):2008 – 2011

[12] DORIGO M, BLUM C. Ant colony optimization theory: a survey [J]. Theoretical Computer Science, 2005, 344(2/3):243 – 278.

[13] de JONG K A. An analysis of the behaviour of a class of genetic adaptive systems [D]. Ann Arbor, MI: University of Michigan, 1975.

[14] YAP D F W, KOH S P, TIONG S K. Artificial immune system based remainder method for multimodal mathematical function optimization[J]. World Applied Sciences Journal, 2011, 14(10): 1507 – 1514.

[15] 陈恩修,刘希玉.一种简便高效的二元离散粒子群算法[J].控制与决策,2010,25(2):256 – 258.

(上接第 310 页)

[6] JIAO H Z, ZHONG Y F, ZHANG L P. Artificial DNA computing-based spectral encoding and matching algorithm for hyperspectral remote sensing data [J]. IEEE Transactions on Geoscience and Remote Sensing, 2012, 50(10): 4085 – 4104.

[7] PETRILLO M L, NEWTON C J, CUNNINGHAM R P, *et al*. The ligation and flexibility of four-arm DNA junction [J]. Biosystems, 1988, 27(9): 1337 – 1352.

[8] MA R I, KALLENBACH N R, SHEARDY R D, *et al*. Three-arm nucletic acid junctions are flexible [J]. Nucleic Acids Research, 1986, 14(24): 9725 – 9753.

[9] JONOSKA N, KARL S A, SAITO M. Three dimensional DNA structure in computing [J]. Biosystems, 1999, 52(1 – 3): 143 – 153.

[10] 刘兴伟,姚书怀.基于层次聚类的语义 Web 服务发现算法[J].计算机应用与软件,2007,24(7):173 – 178.

[11] ZHAO Y, KARYPIS G, FAYYAD U. Hierarchical clustering algorithms for document datasets [J]. Data Mining and Knowledge Discovery, 2005, 10(2): 141 – 168.

[12] 张鸿雁.基于 DNA 计算的聚类算法研究[D].济南:山东师范大学,2011.

[13] ABU BAKAR R B, WATADA J. A biologically inspired computing approach to solve cluster-based determination of logistic problem [J]. Biomedical Soft Computing and Human Sciences, 2008, 13(2):59 – 66.

[14] 孙伟,尤加宇,江宏,等.纳米粒子标记 DNA 探针的制备与检测应用[J].中国卫生检验杂志,2005,15(8):1008 – 1010.

[15] 殷志祥,石晓龙,徐涛,等.0-1 整数规划问题的半自动化 DNA 计算模型[J].生物信息学,2005,4(3):113 – 116.