DOI：10.11830/ISSN.1000-5013.201702064

采用基因表达式编程的自适应层次聚类方法

姜代红[1,2]，尹洪胜[2]，张三友[2]

（1. 徐州工程学院 信电工程学院，江苏 徐州 221008；
2. 中国矿业大学 信息与电气工程学院，江苏 徐州 221008）

摘要： 针对层次聚类算法高维度数据计算复杂度较高、抗干扰性较差、误差较大等不足，在结合基因表达式编程(GEP)非线性演化优越性能的基础上，提出一种基于 GEP 计算模型的层次聚类算法(GEPHCA)，寻找经过基因遗传进化适应度最高的聚类中心.通过试验对比验证可知：基于基因表达式编程的自适应层次聚类方法在实际应用中是有效的，不仅能够实现自动聚类，而且和一般的聚类方法进行比较，具有自适应迭代、速度较快、稳定高效等优点.

关键词： 基因表达式编程；层次聚类；自适应方法；选择算子

中图分类号： TP 301 **文献标志码：** A **文章编号：** 1000-5013(2018)03-0435-04

Self-Adaptive Hierarchical Clustering Algorithm Using Gene Expression Programming

JIANG Daihong[1,2], YIN Hongsheng[2], ZHANG Sanyou[2]

(1. School of Information and Electronic Engineering, Xuzhou Institute of Technology, Xuzhou 221008, China;
2. School of Information and Electrical Engineering, China University of Mining and Technology, Xuzhou 221008, China)

Abstract: Aiming at the disadvantages of the hierarchical clustering algorithm has toward the high-dimension data in the respects of high computational complexity, poor anti-interference and large error. Based on the superior performance of nonlinear evolution of gene expression programming (GEP), a kind of gene expression programmed hierarchical clustering algorithm (GEPHCA) is proposed to discover the most suitable cluster centers through gene genetic evolutionary adaptation. Through the experimental verification, the adaptive hierarchical clustering method based on gene expression programming is effective in practical application. It not only can realize automatic clustering, but also have the advantages of adaptive iteration operating speed, faster operating speed, stable and efficient compared with the general clustering method.

Keywords: gene expression programming; hierarchical clustering; self-adaptive method; selection operation

聚类算法是研究大数据分析和数据挖掘的一个热点，近年来，专家学者对聚类算法进行了研究[1-2].目前，聚类算法主要分为以下几种：基于划分的聚类算法[3-5]，如 k-means，k-prototypes，CLARANS 算法等；基于层次的聚类算法[6-7]，如 CURE，BIRCH 算法等；基于密度聚类算法[8]，如 DBSCAN，OPTICS 算法等；基于网格的聚类算法[9-10]，如 Wave Cluster，CLIQUE 算法等；基于神经网络的聚类算法[11]，

收稿日期： 2017-03-29

通信作者： 姜代红(1969-)，女，教授，博士，主要从事智能计算嵌入式技术的研究. E-mail：daihongjiang@163.com.

基金项目： 国家自然科学基金资助项目(61379100)；国家星火计划项目(2015GA690085)；江苏省徐州市省科技计划项目(KC16SQ178)

如SOM算法等;基于统计学的聚类算法[12],如CobWeb,Bayesian算法等.各种聚类算法各有优劣,其中,层次聚类方法由于算法简单、普适性强、并具有高可伸缩性的优点,因此,在聚类分析中应用非常广泛.基因表达式编程(GEP)计算模型于2001年首次由Candida提出,融合了遗传算法和遗传程序设计的优点,编码简单,较其他遗传演化算法易于表达、速度更快、求解能力更强.层次聚类方法结合基于基因表达式编程的演化遗传优化,应用到聚类分析计算模型中,具有以下3个优点.1)能够通过迭代演化自动寻找最佳聚类中心;2)在保证聚类效果的前提下,有效减少聚类时间;3)能够解决诸多聚类实际问题,在应用方面具有良好的自适应性.本文提出采用基因表达式编程的自适应层次聚类方法.

1 层次聚类算法的基本原理

层次聚类算法通过将数据组织为若干组并形成树进行聚类,因此,也称为树聚类算法.设集合 $X\in\mathbf{R}^{n\times d}$,$n$ 为样本数.首先,通过计算样本之间的相似性构成一个相似性矩阵 $\boldsymbol{R}=(r_{i,j})_{n\times n}$;然后,根据该样本集的样本之间的相似性矩阵形成层次结构,对每个分层节点进行二叉划分,生成从1到 n 的树形聚类分析结果,从而构成样本集 X 的系统树图 $H=\{H_1,H_2,\cdots,H_q\}$.对于 $q\leqslant n$,$C_j\in H_j$ 且 $1<l<m<q$,在 $C_i\subset C_j$ 或 $C_i\cap C_j=\phi$ 的条件下,满足全部 $j\neq i$.设初始聚类结果 $C=\{C_1,C_2,\cdots,C_k\}$ 及聚类中心 $s=\{s_1,s_2,\cdots,s_k\}$,根据距离合并簇,递归进行层次聚类.定义 $d(\boldsymbol{x}_i,\boldsymbol{x}_j)$ 为簇 i,j 之间的对称距离,即 $d_{i,j}=d_{j,i}$.i 的特征向量为 $\boldsymbol{x}_i$,用中心向量表示簇.层次聚类方法通过迭代进行数据分析,具体有如下6个步骤.

步骤1 for each$(i,j\in C_2)$,$d_{i,j}=d(s_i,s_j)$;

步骤2 $(i^*,j^*)=\arg\min(d_{i,j}),(i,j)\in C_2$;

步骤3 Set $C_m\leftarrow C_i^*\bigcup C_j^*$;

步骤4 $C\leftarrow\{C-\{i^*\}-\{j^*\}\}\bigcup\{m\}$;

步骤5 中心调整 $s_m=\dfrac{1}{\|C_m\|}\sum_{v\in C_m}\boldsymbol{x}_i$;

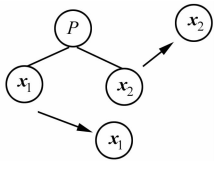
步骤6 重复步骤1~5,直到所有对象聚类到一个单一的聚类中为止.

2 基于基因表达式编程的层次聚类算法

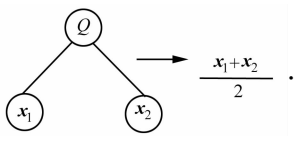
基于GEP,在对层次聚类算法进行非线性迭代优化的基础上,提出了基于GEP计算模型的层次聚类算法(GEPHCA).该算法能够通过基因遗传自适应演化寻找适应度最高的聚类中心.

2.1 基因编码

基因表达式编程编码由基因头和基因尾两部分组成.头部由非终结符构成,而尾部由终结符构成.为使基因染色体表达树达到自动聚类的目标,设计头部函数集 $F=\{P,Q\}$.P 为聚类中心取节点值,如



;Q 为聚类中心取平均值,如



.基因尾部终结符集合 $T=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\boldsymbol{x}_3,\cdots,\boldsymbol{x}_n\}$,其中,$\boldsymbol{x}_i$ 为 C_t 的聚类中心,如基因 $G=PQP\boldsymbol{x}_1\boldsymbol{x}_2\boldsymbol{x}_3\boldsymbol{x}_4$,翻译成的聚类树,如图1所示.

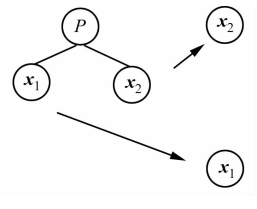


图1 基因 G 翻译成的聚类树

Fig. 1 Clustering tree from gene G

2.2 适应度函数

在层次聚类方法基础上,依据适应度函数解码染色体可得到各可能的簇中心.适应度函数为

$$f=1\Big/\sum_{r=1}^{p}\sum_{i=1}^{m_r}\|X_i^r-C_r\|^2. \tag{1}$$

式(1)中:聚类中心 $C_r=\dfrac{1}{m_r}\sum_{i=1}^{m_r}X_i^{(r)}$,$i=1,2,\cdots,m_r$,$r=1,2,\cdots,P$;$\sum_{r=1}^{p}m_r=N$,$m_r$ 为第 r 类样本的数目,X_i^r 为数据 $\boldsymbol{x}_i$ 隶属第 r 类,N 为样本的数目,$P(2\leqslant P\leqslant N-1)$ 为输出的聚类

中心个数.

2.3　遗传算子

1）选择算子.使用的轮盘赌选择方法进行选择操作，并引入精英策略进行择优选择，直接选择复制上一代种群中的最佳个体到下一代，算法描述如下.

```
nextEvolutionStep()//精英选择策略
    {
        generationNumber++;//种群代数加 1
        List<T> elites=new ArrayList<T>();
        for(0 到最佳个体数)
        {
            elites.add(上一代种群中的最佳个体);
        }
        选择种群中的个体;
        对选中的个体进行遗传修饰;
        返回下一代种群
    }
```

2）基因重组和基因变异基因头采用函数符替换法进行变异，即 $P\rightarrow Q$，$Q\rightarrow P$，基因尾部采用随机选取一个数据的方式进行无重复替换变异.

2.4　自动合并聚类算法

经过基因遗传自适应演化寻找适应度最高的聚类中心后，通过自动合并聚类算法对簇进行整合优化，自动归并可详见文献[11].

2.5　基于基因表达式编程的层次聚类算法

基于基因表达式编程(GEP)的层次聚类算法流程，如图 2 所示.具体有如下 11 个主要步骤.

步骤 1　输入样本数据集和算法参数，如最大迭代代数、种群大小、头部的长度、基因重组率和基因变异率等.

步骤 2　按聚类基因编码方式随机初始化种群.

步骤 3　按照适应度函数计算种群适应度值.

步骤 4　判断是否已经迭代到最大演化代数条件，假如达到，则根据自动合并聚类算法合并聚类，输出聚类中心和聚类结果 C；否则，转向步骤 5.

步骤 5　对种群中个体尾部的簇中心进行聚类.

步骤 6　按照层次聚类合并簇算法，更新聚类结果 C.

步骤 7　根据步骤 5 更新聚类中心.

步骤 8　精英保留策略保留最佳个体到下一代.

步骤 9　按输入的基因重组率进行单点和两点基因重组.

步骤 10　按输入的基因变异率进行变异.

步骤 11　产生新种群，转向步骤 3，重新计算新种群适应度值.

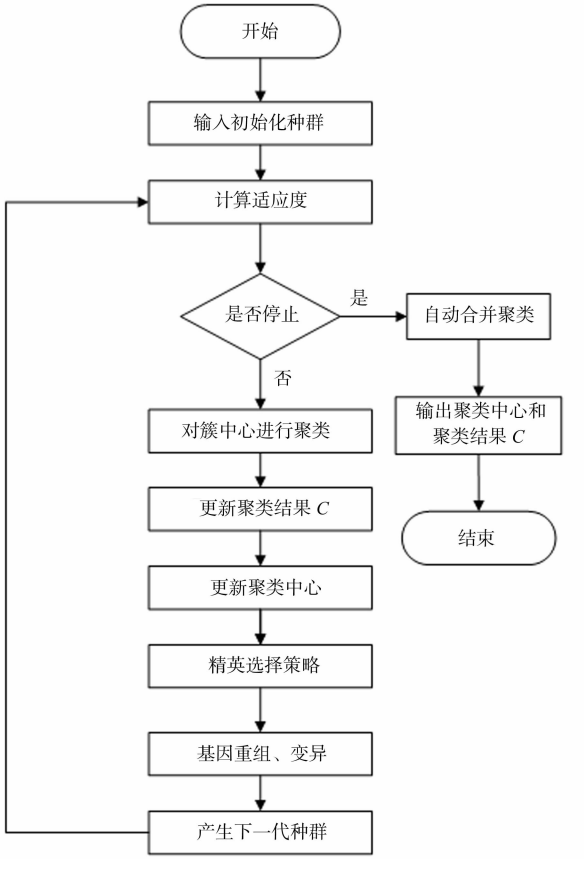


图 2　基于 GEP 的层次聚类算法流程图

Fig. 2　Flow diagram of hierarchical clustering algorithm based on GEP

3　实验性能与对比分析

基于 GEP 的层次聚类算法配置：种群大小为 100；基因头部长度为 8；变异率为 1.0；基因尾部长度

为9;基因数目为3;染色体长度为51;远程安装服务(RIS)转移率为0.5;RIS长度为1,2,3;函数集为"+,-,*,/";变量集 $F=\{P,Q\}$;连接函数为"+";适应度函数为式(1).实验环境:操作系统为Windows 7操作系统;CPU为酷睿i5双核,2.70 GHz,2.00 G内存;开发环境和运行环境为Visual C^{++}.利用MATLAB编程对参考文献[12]的合并聚类算法和文中GEPHCA算法两种聚类效果进行了比较,结果如图3所示.由图3可知:GEPHCA算法适应度能够达到700,而文献[12]算法最高达650就收敛了.这说明GEPHCA算法比文献[12]算法聚类效果要好.因此,GEPHCA算法是一种稳定高效的自适应聚类方法,聚类结果更有参考意义.

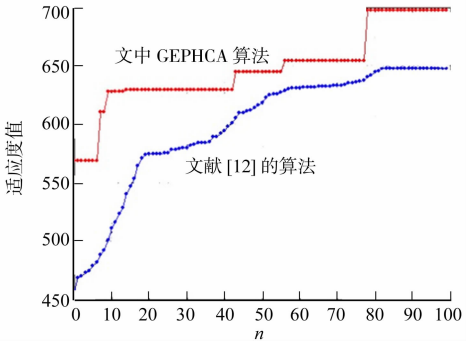


图3 两种聚类算法的聚类比较

Fig. 3 Comparison between two kinds of clustering algorithm

4 结束语

基于基因表达式编程算法,结合层次聚类,提出了基于GEP演化的GEPHCA.GEPHCA具有GEP聚类不需要任何先验知识的条件下进行自动聚类分析且聚类精度较高的优点,提高了聚类收敛速度.最后,通过仿真实验验证了GEPHCA算法的稳定性与有效性.

参考文献:

[1] 金建国.聚类方法综述[J].计算机科学,2014,41(b11):288-293.

[2] 王千,王成,冯振元,等.K-means聚类算法研究综述[J].电子设计工程,2012,20(7):21-24.DOI:10.3969/j.issn.1674-6236.2012.07.008.

[3] JI Jinchao,PANG Wei,ZHENG Yanlin,*et al*.A novel cluster center initialization method for the k-prototypes algorithms using centrality and distance[J].Applied Mathematics and Information Sciences,2015,9(6):2930-2933.

[4] 段明秀.QPSO优化的改进CLARANS聚类算法[J].Computer Engineering and Applications,2013,49(9):61-64.DOI:10.3778/j.issn.1002-8331.1109-0356.

[5] 李松,崔环宇,张丽平,等.基于CURE聚类算法的静态R树构建方法[J].计算机科学,2015,42(10):193-197.DOI:10.11896/j.issn.1002-137X.2015.10.039.

[6] 韦相.基于密度的改进BIRCH聚类算法[J].计算机工程与应用,2013(10):66-69.DOI:10.3778/j.issn.1002-8331.1112-0567.

[7] 马超,侯天诚,徐瑾辉,等.泛函深度神经网络及其在金融时间序列预测中的应用[J].徐州工程学院学报(自然科学版),2017,32(2):46-53.DOI:10.15873/j.cnki.jxit.000154.

[8] CHEN Lin,YU Ting,CHIRKOVA R.Wavecluster with differential privacy[C]// ACM International on Conference on Information and Knowledge Management.[S.l.]:ACM,2015:1011-1020.

[9] 杨善红,梁金明,李静雯.基于网格密度影响因子的多密度聚类算法[J].计算机应用研究,2015,32(3):743-747.DOI:10.3969/j.issn.1001-3695.2015.03.024.

[10] SABAR N,AYOB M,KENDALLl G,*et al*.Automatic design of a hyper-heuristic framework with gene expression programming for combinatorial optimization problems[J].IEEE Transactions on Evolutionary Computation,2015,19(3):309-325.DOI:10.1109/TEVC.2014.2319051.

[11] 姜代红.基于基因表达式编程编程的ISODATA模糊聚类算法[J].计算机应用,2011,31(12):3252-3254.DOI:10.3724/SP.J.1087.2011.03252.

[12] 刘贝贝,马儒宁,丁军娣.基于密度的统计合并聚类算法[J].智能系统学报,2015,10(5):712-721.DOI:10.11992/tis.201410028.

(**责任编辑**:陈志贤 **英文审校**:吴逢铁)