

基于基因表达式的太湖水质聚类及模型发现

鞠训光¹, 张三友², 徐德兰¹

(1 徐州工程学院 信电工程学院, 江苏 徐州 221018; 2 苏州大学 计算机科学与技术学院, 江苏 苏州 215006)

摘 要: 应用基因表达式编程研究太湖水质自动聚类及其数学模型发现, 解决了传统聚类算法需预先设定类别数目及不能对聚类建立数学模型的问题. 与实际数据比较, 表明聚类结果正确; 所发现的营养盐对水质影响的数学模型, 对监控、优化太湖水质具有建设性意义.

关键词: 基因表达式编程; 自动聚类; 模型发现; 太湖水质

中图分类号: TP391

文献标识码: A

文章编号: 1000-7180(2011)10-0039-05

Discovery Clustering and Model on Water Quality of Taihu Lake Based on Gene Expression

JU Xun-guang¹, ZHANG San-you², XU De-lan¹

(1 College of Information and Electrical Engineering, Xuzhou Institute of Technology, Xuzhou 221018, China;

2 School of Computer Science and Technology, Soochow University, Suzhou 215006, China)

Abstract: Mined clustering and mathematical model on Taihu water quality discover by applying Gene Expression Programming, Traditional Cluster Algorithm Preset the number of category to be solved and constructed mathematical model. The Experimental result compared to the actual data, the clustering is correct. The mathematical model of nutritive salt affecting water quality discovered that constructive significance is provided with onsupervisory control and optimizing Taihu water quality.

Key words: gene expression programming; auto-clustering; model discovery; Taihu Lake water quality

1 引言

聚类分析是数据挖掘的重要研究内容. 目前聚类技术主要可以分为以下几类: 基于划分的方法、基于层次的方法、基于密度的方法、基于网格的方法以及基于模型的方法等^[1]. 但从前人研究结果和实践看, 很多已有的算法需要预先设定聚类个数, 即在没有先验知识的情况下不能对数据进行自动聚类^[2], 也无法对聚类结果进行数学模型的发现.

而基因表达式编程 (Gene Expression Programming, GEP) 是 Candida Ferreira 在结合了遗传算法 (GA) 和遗传程序设计 (GP) 思想的基础上于 2001 年提出的一种基于基因型和表现型的自适应

演化算法^[3]. 由于 GEP 能通过简单的编码解决复杂的结构优化问题和模式发现问题^[4], 同时 GEP 较遗传算法和遗传程序设计具有易于表达、速度更快、求解能力更强等特点^[5].

本文就是利用基因表达式编程方法的优势, 尝试解决上述问题, 研究太湖水质、矿物质及沉淀物的动态自动聚类及模型的发现挖掘, 尝试发现矿物质及沉淀物对水质影响的内在规律性模式, 为安全、高效治理、管理和优化太湖水质提供有意义的决策建议.

2 基因表达式编程基本思想

2.1 GEP 开放阅读框和基因

GEP 的结构组织从开放阅读框^[6] (Open Read-

收稿日期: 2011-04-25; 修回日期: 2011-06-26

基金项目: 徐州市科技项目“徐州市食品(餐饮)安全风险监测预警系统构建”(XJ09095); 徐州工程院校基金项目“基于遗传算法的交互式二维下料系统研究”(XKY2010104)

ing Frames, ORFs)的角度来说更容易理解. 生物学上, 一个 ORF 序列, 或者基因编码的序列, 开始于一个“起始”密码子, 接着是些氨基酸密码子, 最后止于一个“终止”密码子.

比如, 考虑式(1)的代数表达式.

$$\sqrt{S(a) \times (c+d) - b} \quad (1)$$

它可以表示成如图 1 所示的表达式树.

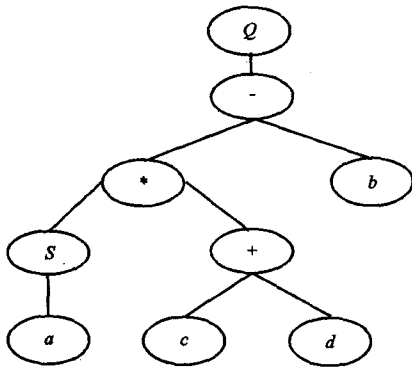


图 1 表达式树

此处“Q”代表开方函数, “S”代表正弦函数.

图 1 实际上代表了 GEP 个体的表现型, 根据表现型很容易推断出式(2)所示的基因型.

$$012345678: Q - * b S + a c d \quad (2)$$

该基因型是直接表达式树从左到右, 自上而下读出的结果. 式(2)就是一个 ORF, 从“Q”开始, 止于“d”, 称为 K-表达式(源自 Karva 语言)^[6].

2.2 GEP 的基因结构

GEP 的基因由头尾两部分构成. 头部由包含既代表函数又代表终点的符号构成, 而尾部仅仅含有终点. 对每个问题而言, 头的长度 h 是选定的, 而尾的长度 t 是 h 和 n 的函数, 其中 n 是所需变量数最多的函数的参数个数(也称为最大操作数), 例如, 对于开方或者三角函数, n 取 1, 对于加法等四则运算, n 取 2, t 的大小见式(3).

$$t = h(n-1) + 1 \quad (3)$$

采用这种头尾划分的好处在于, 一方面整个基因的结构在设定了函数集和头部长度之后就能够确定; 另一方面, 整个基因在式(3)的约束下, 一定能够保证结构上的正确性, 而不用担心会产生非法个体^[7].

对一棵表达式树, 按照从上到下, 从左到右的顺序遍历, 得到的序列即可称为 K-表达式.

因此, GEP 中, 基因型/表现型系统的根本特性: 句法封闭性是与之俱来的, 允许对基因型进行完全不受限制的操作, 并由此促成高效的进化.

2.3 GEP 适应度度量函数

C. Ferreira 提出了两种基于误差的适应度函数^[6]: 一种基于绝对误差, 见式(4); 另一种基于相对误差, 见式(5).

$$f_i = \sum_{j=1}^{C_i} (M - |C_{(i,j)} - T_j|) \quad (4)$$

$$f_i = \sum_{j=1}^{C_i} (M - |\frac{C_{(i,j)} - T_j}{T_j} \cdot 100|) \quad (5)$$

式中, M 是选择范围, $C_{(i,j)}$ 是染色体个体 i 对于适应度样本 j (来自 C_i 集合中) 的返回值. 而 T_j 是适应度样本 j 的目标值, 两个方程中的绝对值项都对应各自的误差, 该项称为精度 p , 因此, 对于一个完美适应的情况, $C_{(i,j)} = T_j$ 且 $f_i = f_{\max} = C_i M$.

2.4 GEP 算法流程图

具体的 GEP 算法流程图可参见文献[8].

3 GEP-Cluster 自动聚类算法

聚类的核心是搜索和评估, 基因表达式编程有很强的全局搜索能力, 在解决聚类问题方面有极大的潜力. 基于 GEP, 提出了 GEP-Cluster 聚类算法^[2], 要点如下:

(1) 在基因头中添加一些特殊的聚类运算符 $\cap$ 和 $\cup$ 来完成聚类的筛选和聚合;

(2) 从数据集中随机选若干点作为初始簇中心, 构成个体的染色体;

(3) 每代进化过程中, 交叉算子和变异算子引入新的数据点, 中心点可能漂移. 若干代进化后找到合适的簇中心, 从而找到最佳的聚类组合划分.

具体的步骤如下:

(1) 初始化种群;

(2) 按照个体所表达的聚类中心进行聚类;

(3) 计算个体的适应度;

(4) 保留最佳个体到下一代;

(5) 依照轮盘赌实施遗传操作;

(6) 按照交叉率 p_c 对个体进行遗传交叉;

(7) 按照变异率 p_m 对个体进行遗传变异;

(8) 若达到最大迭代次数则执行(9), 否则执行(2);

(9) 选择出最优个体;

(10) 对最优个体实施自动合并簇算法(AMCA);

(11) 输出聚类划分结果.

染色体编码: 由于聚类问题特殊性, 染色体采用整数编码, 由单基因构成. 每基因由基因头和基因尾

组成,其中基因头部包含聚类运算符“ $\cup$ ”和“ $\cap$ ”,基因尾部则由按照均匀分布从数据集中随机抽取无重复的数据点的序号构成。例如, $R_1 = \cup \cup \cap \cap \cup x_1 x_2 x_3 x_4 x_5 x_6$ 是一个合法的染色体,它描述了进化过程中某一代的聚类状态。其中, $x_i (i \in [1, n], n$ 为数据集中数据的总个数) 是数据集中每个数据对应的序号,且 x_i 是当前系统中某个簇 C_i 的中心。

若根节点是 $\cup$ 运算符,将左右子树分别放入两个不同的集合中,从而达到对簇的分割。若根节点是 $\cap$ 运算符,将左右子树中的数据放入同一个簇之中,计算其均值,并用均值作为簇的中心,从而达到对簇的合并,下列算法详细描述了解码过程:

输入:输入个体染色体经过编码后得到的聚类 ET。

输出:输出聚类 ET 对应的染色体所携带的簇中心。

(1)若 ET 空则执行(12)

(2)Switch (ET 根节点):

(3)Case ' $\cup$ ':

(4)剪枝左子树,将左子树放入集合 i 中

(5)剪枝右子树,将右子树放入集合 i 中

(6)FOR each SET i DO

(7)递归解码集合 i 中的每个子树

(8)Case ' $\cap$ ':

(9)递归解码左子树

(10)递归解码右子树

(11)簇中心序列 $\leftarrow$ 左右子树中叶子节点均值;

(12)返回染色体中携带的簇中心序列。

适应度计算:经过解码染色体信息得到各个可能的聚类簇的中心点坐标,然后对数据集中每个数据点根据其与其各个簇中心的欧氏距离,依次将其赋给最近的簇。重新计算聚类后的各个簇的中心坐标,并计算数据集中所有数据的平方误差总和 E 。 E 的定义式如(6):

$$E = \sum_{i=1}^k \sum_{p \in m_i} |p - m_i|^2, m_i = \frac{1}{n_i} \sum_{x_j \in c_i} x_j \quad (6)$$

式中, k 为数据集中的簇的个数, n_i 为簇 C_i 中数据点的个数。 E 值越小,则生成的簇会尽可能的紧凑和独立。由于在 GEP 中一般按照适应度由大到小选择个体,因此算法中取适应度函数 $f = 1/E$ 。

4 太湖水质聚类

根据 2004 年 5 月—9 月,太湖沿岸采样的(32 个点)数据^[9-10],其中磷、氮、硝态氮和叶绿素的含量聚类,32 个数据点聚类结果如图 2 所示。

其中每一类的具体数据显示如图 3 所示。

其中数据点 3、24、25、28、29、32 为一类,根据太

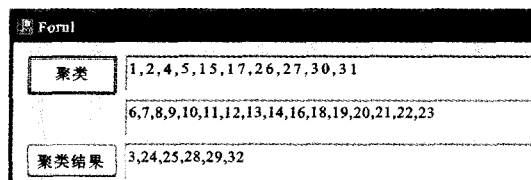


图 2 太湖 32 个数据点采样聚类结果

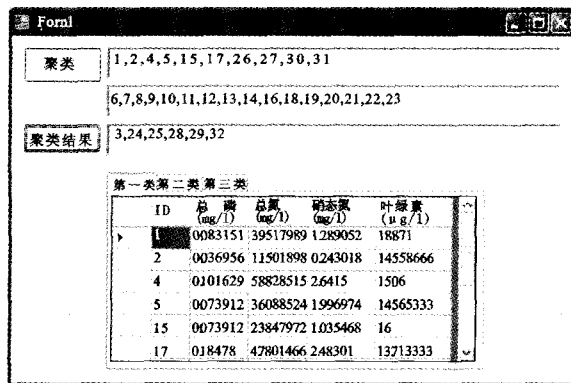


图 3 各类中磷、氮、硝态氮和叶绿素的含量聚类结果

湖沿岸采样的(32 个点)数据,可以看到这一类的叶绿素的含量比较高,所以这块区域波浪大、发现藻华一类。

数据点 6、7、8、9、10、11、12、13、14、16、18、19、21、22、23 为一类,根据采样的数据,可以看出氮、磷、硝态氮、叶绿素的含量比较适度。所以这块区域可能有养蟹场、水生植物多,水比较清。

数据点 1、2、4、5、15、17、26、27、30、31 归为一类。

根据聚类的结果分析三类的情况。

第一类:其他类。由于采样点的不准确或是其他客观原因,不做分析,将它归为一类。

第二类:养蟹场、水清、沉水植物多。水生植物能够有效的抑制底泥中氮的释放,是因为水生植物能够吸收底泥释放的氮;而植物进行光合作用,释放氧气,能够提供给水中的蟹、鱼等动物。水生植物也能有效的抑制水中的磷、硝态氮等含量。水生植物的生长也对水的净化起到了一定的作用。其中水生植物狐尾藻对氮的抑制作用最为明显。

第三类:波浪大,发现藻华。这一类中,叶绿素的含量比较多,氮、磷、硝态氮的含量也比较多。说明这区域水体富营养化。造成水体富营养化的原因有:由于城市工业废水和生活污水大量排入湖中,使营养物质在水体中富集,造成海域富营养化。其中氮、磷等含量增加。富营养化也造成藻类的生长,形成了“藻华”。

5 太湖水质模型或函数发现

在太湖水质模型函数发现中,我们用A、B、C、D来定义区间上的自变量的值,分别代表了影响太湖水质的主要因素磷的含量、氮的含量、硝态氮的含量和叶绿素的含量,y为因变量,代表均方差^[9].

GEP 进行模型发现主要需要以下五个步骤:

步骤 1:选择合适的适应度函数,对本问题,采用式(4)来计算适应度,对于一个完美适应的情况 $C_{(i,j)} = T_j$ 且 $f_i = f_{\max} = C_i M$. 采用绝对误差 100 作为选择范围,精度误差为 0.01,因此,最大适应度值 $f_{\max} = C_i M$.

步骤 2:选择终点集 T 和函数集 F,其中终点集由自变量和常数构成,自变量用 A、B、C、D 表示,常数用?表示. 终点集 $T = \{a, b, c, d, ?\}$;在这里选择 7 个函数,函数集 $F = \{+, -, *, /, Q, S, R\}$,其中 Q、S、R 分别代表开根号函数,正弦函数,余弦函数.

步骤 3:选择染色体组织结构,即头部和尾部的长度. 通过实验数据发现,当基因头部为 9,基因个数为 2 时,适应度达到最大为 991.781 287. 尾部长度为 $t = 10$.

步骤 4:选取连接函数. 在此问题中染色体后面附加一个和基因尾部等长的附加域 Dc,从结构上讲,Dc 在尾部的后面,和 t 长度相等,代表中间随机常数的符号构成. 在基因中生成了具有单独边界的,有自身字符集的另一个区域. 用于随机产生连接各个子表达式.

步骤 5:选择遗传算子集并决定其参数大小. 本文采用变异和重组算子来作为遗传算子,选取进化代数 50.

根据前面的聚类,以下为每类的函数发现:

第一类:其他类,具体的数据参见表 1.

其中上下边界值代表随机产生常数的上界和下界值,取[0,10]. 这个取值是根据 A、B、C、D 中最小值和最大值比较后得到,(如:A 中范围为 0.03 ~ 0.24;B 中范围为 0.75 ~ 12.05;C 中范围为 0.13 ~ 2.8;D 中范围为 5.02 ~ 90.44;B、C、D 范围值数据除以 10,最后得到一个最终的范围 0 ~ 10 为上下界值.

通过演化,在第 46 代发现最佳个体,最佳个体适应度为:990.601 823. 染色体:Q + * QDA? DQBAD?CBD?A?QRSASAC * DBAAAA???? A1R// - 1 - SQ2121221122, 函数表达式为 $y = (\sqrt{D * A + \sqrt{9.232\ 154\ 302}})$.

表 1 GEP 参数

运行代数——m_GepNum	50
种群大小——m_nGroupSize	50
下一个初始种群大小	10
一个染色体中基因的个数	2
下边界值——m_LowerBound	0
上边界值——m_UpperBound	10
头部长度——m_nHeadSize	9
基因重组率——m_PRecombination	0.7
基因变异率——m_PMutation	0.2

经过对上面得到的最佳染色体解码,得到两棵子树如下:对应的 sub-ET1、sub-ET2 如图 4 所示.

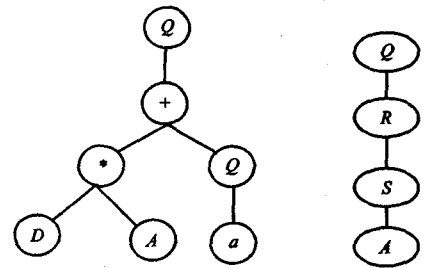


图 4 第一类中的 sub-ET1、sub-ET2

第一类的演化过程如图 5 所示.

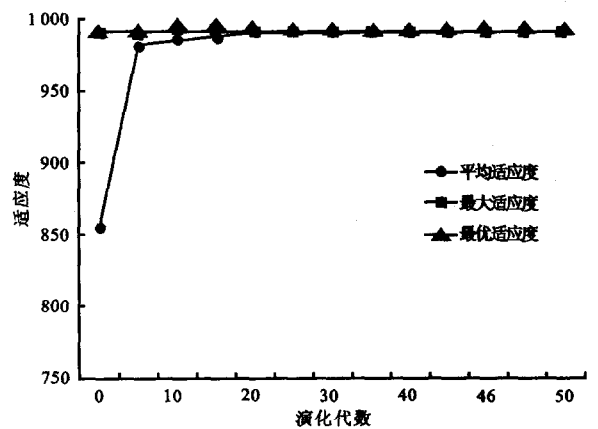


图 5 第一类演化示意图

第二类:养蟹场、水清、沉水植物多. 通过演化,在第 34 代发现最佳个体,最佳个体适应度为:1 593.898 439. 函数表达式为: $y = (9.992\ 980\ 743 / (D + \sqrt{C - C - A}))$.

第三类:波浪大,发现藻华. 通过演化,在第 43 代发现最佳个体,最佳个体适应度为:598.897 863. 函数表达式为: $y = (\cos((7.937\ 253\ 945 - C) / (B + C) * B) + C)$.

6 结束语

应用基因表达式编程对影响太湖水的主要因素:氮、磷、正磷酸盐、硝酸氮等营养盐进行自动聚类和公式发现研究,对监控、优化太湖水质具有一定的指导性意义。

参考文献:

- [1] Han J, Kamber M. 数据挖掘:概念与技术[M]. 范明, 孟小峰, 译. 北京:机械工业出版社, 2007.
- [2] 陈瑜, 唐常杰. 基于基因表达式编程的自动聚类方法[J]. 四川大学学报:工程科学版, 2007, 39(6): 107-112.
- [3] Candida Ferreira. Gene expression programming: a new adaptive algorithm for solving problems[J]. Complex Systems, 2001, 13(2): 87-129.
- [4] Ferreira C. Gene expression programming in problem solving[EB/OL]. [2001-10-24]. <http://www.gene-expression-programming.com/GEPBiblio.asp>.
- [5] 唐常杰, 张天庆, 左劼, 等. 基于基因表达式编程的知识

发现沿革、成果和发展方向[J]. 计算机应用, 2004, 24(10): 1-5.

- [6] Ferreira C. Gene expression programming: a new adaptive algorithm for solving problems[J]. Complex Systems, 2001, 13(2): 87-129.
- [7] 左劼. 基因表达式编程核心技术研究[D]. 四川: 四川大学, 2004: 36-42.
- [8] 颜雪松. 基于基因表达式程序设计的电路优化算法研究[J]. 微电子学与计算机, 2008, 25(1): 120-126.
- [9] 雷泽湘, 徐德兰. 太湖大型水生植物分布特征及其对湖泊营养盐的影响[J]. 农业环境科学学报, 2008, 27(2): 698-704.
- [10] 雷泽湘, 谢贻发, 徐德兰, 等. 大型水生植物对富营养化湖水净化效果的试验研究[J]. 安徽农业科学, 2006, 34(3): 553-554.

作者简介:

鞠训光 男, (1966—), 博士, 讲师. 研究方向为智能计算、数据挖掘.

张三友 硕士研究生. 研究方向为智能计算、图像处理.

(上接第 38 页)

再利用双线性插值算法恢复校正后图像. 实验表明, 在畸变不严重的情况下, 该方法取得很好的矫正效果.

对网格图和待矫正图像分别作挤压变形和球面化变形之后, 再矫正. 可以看到均有明显的矫正效果. 实验表明, 在畸变较严重的情况下, 应该将 MSE 的值适当降低, 以防止出现矫正过适, 即训练点矫正精度高, 而训练点之间出现过度扭曲的情况.

5 结束语

本文在总结畸变校正方法的基础上, 给出基于 BP 神经网络的畸变校正方法的具体介绍与实现. 通过实验结果表明, 基于 BP 神经网络畸变矫正方法能较全面地拟合畸变函数, 得到高精度的校正图像, 尤其是在畸变不是很严重的情况下, 而且易于实现, 能够有效地解决图像矫正在实际工作中的应用.

参考文献:

- [1] Zhang Z. A flexible new technique for camera calibration[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2000, 22(11): 1330-1334.
- [2] Park D B. Learning logic[C]// Center for computing research in economics and management sciences. Cambridge: MIT, 1985.
- [3] 朱铮涛, 黎绍发. 镜头畸变及其校正技术[J]. 光学技术, 2005(1): 136-141.
- [4] Bryson A E, Ho Y C. Applied optimal control[M]. New York: Blaisdell, 1969.
- [5] 史忠植. 神经网络[M]. 北京: 高等教育出版社, 2009.
- [6] 王昊京, 王建立, 王鸣浩, 等. 采用双线性插值收缩的图像修复方法[J]. 光学精密工程, 2010(5): 1234-1241.

作者简介:

金元 男, (1985—), 硕士研究生. 研究方向为图像处理.