文章编号:1001 - 9081(2011)12 - 3252 - 03
doi:10.3724/SP. J. 1087. 2011. 03252

基于基因表达式编程的 ISODATA 模糊聚类算法

姜代红[1,2]
(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)
(jdh@ xzit. edu. cn)

摘 要:针对 ISODATA 算法需要人为给定分类数,对初始聚类中心较为敏感,没有显示出自动聚类效果等不足,结合基因表达式编程(GEP)嵌套构成迭代自组织模糊聚类进行优化计算。该方法不仅能在不需要先验知识的条件下对数据进行自动聚类,而且充分利用了 GEP 算法的全局寻优能力及 ISODATA 算法的软性分类特性,提高了算法的收敛速度和聚类精度。通过仿真验证及对比分析,运用到地理信息系统(GIS)物流选址实际问题中,得到了理想聚类效果。

关键词:基因表达式编程;模糊 ISODATA;聚类;地理信息系统
中图分类号: TP311. 13 **文献标志码**:A

Fuzzy ISODATA clustering algorithm based on gene expression programming

JIANG Dai-hong[1,2]
(1. *School of Information and Electronic Engineering, Xuzhou Institute of Technology, Xuzhou Jiangsu* 221008, *China*;
2. *School of Information and Electrical Engineering, China University of Mining and Technology, Xuzhou Jiangsu* 221008, *China*)

Abstract: Concerning the defects of the artificial setting of the categories number, the sensitiveness to initial cluster centers and the lack of automatic clustering effects on the ISODATA algorithm, in combination with Gene Expression Programming (GEP) a nested iterative self-organizing fuzzy clustering was formed up. This paper presented a new algorithm: fuzzy ISODATA clustering algorithm based on GEP. This algorithm not only conducted automatic clustering under the condition of no prior knowledge, but also fully used the capability of global optimization of GEP algorithm and soft classification features of ISODATA, which resulted in the increase of the convergence speed and the clustering accuracy. It is verified by simulation and comparative analysis of the practical problems in GIS logistics location.

Key words: Gene Expression Programming (GEP); fuzzy ISODATA; clustering; Geographics Information System (GIS)

0 引言

聚类分析是数据挖掘中的一种重要技术,目前已成为数据挖掘中的一个热点领域。聚类需要解决的问题是将已给定的若干无标记的模式聚集起来使之成为有意义的类[1]。选取和设计更加有效的聚类算法,对于研究数据挖掘,在海量数据中发现有效而实用的模式,具有科学的参考价值。

基于模糊划分的迭代自组织数据分析技术——模糊 ISODATA (Interactive Self-Organizing Data Analystic Techniques)算法[2]是解决若干样本聚类问题实用的方法之一。作为一种无监督迭代自组织聚类划分算法,与传统分类方法的根本区别是,它是一种软性分类,而传统聚类划分是一种非此即彼的硬性划分,因此具有更强的科学性和实用性。基因表达式编程(Gene Expression Programming, GEP)是 Ferreira 于 2001 年提出的[3-4],是一种全新的模拟生物进化算法,借鉴了生物进化遗传的基因表达规律提出的知识发现新技术。将 GEP 用于聚类分析中,具有在不需要先验知识的条件下对数据进行自动簇的划分及合并的优点,将它与模糊 ISODATA 算法结合,具有以下三个优点:1)能够在不需要任何先验知识的情况下自动确定最佳聚类中心;2)算法的收敛速度和寻优精度都得到了很大提高;3)是一种能够知道聚类中心信息的软划分,从而知道各类样本隶属于某类的隶属度,使得聚类结果更具参考价值。本文将该方法用于地理信息系统(Geographics Information System, GIS)物流选址中,可对数据信息实体化和综合化,从而给决策者提供了更为科学有效的参考与建议。

1 模糊 ISODATA 算法基本原理

模糊 ISODATA 是用隶属度确定每个数据点属于某个聚类的程度的一种聚类算法。1973 年,Bezdek 利用模糊集合的概念提出了该算法,作为早期硬 C 均值聚类(Hard C-Means clustering, HCM)方法的一种改进[5]。模糊 ISODATA 把 n 个向量 $\boldsymbol{x}_i(i = 1,2,\cdots,n)$ 分为 c 个模糊组,并求每组的聚类中心,这种分类结果对应的分类矩阵,就是一个模糊分类矩阵。设模糊分类矩阵为 $\boldsymbol{U}$:

$$\boldsymbol{U} = \begin{bmatrix} u_{11} & u_{12} & \cdots & u_{1n} \\ u_{21} & u_{22} & \cdots & u_{2n} \\ \vdots & \vdots & & \vdots \\ u_{c1} & u_{c2} & \cdots & u_{cn} \end{bmatrix} \tag{1}$$

其中:n 表示聚类样本个数;c 表示分类数;u_{ij} 表示第 j 个样本属于第 i 个分类的隶属度,且满足

$$\sum_{i=1}^{c} u_{ij} = 1;\ j = 1,\cdots,n \tag{2}$$

模糊 ISODATA 的目标函数为

$$J(U,c_1,\cdots,c_c) = \sum_{i=1}^{c} J_i = \sum_{i=1}^{c}\sum_{j}^{n} u_{ij}^m d_{ij}^2 \tag{3}$$

收稿日期:2011 - 06 - 24;**修回日期**:2011 - 08 - 11。
基金项目:江苏省高校自然科学研究计划项目(10KJD520008);"青蓝工程"资助项目;徐州市科技计划项目(XX10A021)。
作者简介:姜代红(1969 -),女,湖南郴州人,副教授,博士研究生,主要研究方向:嵌入式技术、数据库。

式(3)中 u_{ij} 介于0,1 间,为模糊组 i 的聚类中心,$d_{ij} = \| c_i - x_j \|$ 为第 i 个聚类中心与第 j 个数据点间的欧几里得距离;$m \in [1,\infty)$ 是一个加权指数,Bezdek 证明当 $m > 1$ 的条件下,一定可以算出最佳划分,参数 $m = 2$ 时最优[4]。一个分类 U_0 若是最佳分类,应使 $J(U_0,c_1,\cdots,c_c)$ 最小,构造如下新的目标函数,可求得使式(3)达到最小值的必要条件:

$$\bar{J}(U,c_1,\cdots,c_c,\lambda_1,\cdots,\lambda_n) = J(U,c_1,\cdots,c_c) + \sum_{j=1}^{n}\lambda_j\left(\sum_{i=1}^{c}u_{ij}-1\right) = \sum_{i=1}^{c}\sum_{j}^{n}u_{ij}^{m}d_{ij}^{2} + \sum_{j=1}^{n}\lambda_j\left(\sum_{i=1}^{c}u_{ij}-1\right) \tag{4}$$

式(4)中 $\lambda_j(j = 1,2,\cdots,n)$ 是式(1)中的 n 个约束式的拉格朗日乘子。对所有输入参量求导,使式(4)达到最小的必要条件为:

$$c_i = \frac{\sum_{j=1}^{n}u_{ij}^{m}x_j}{\sum_{j=1}^{n}u_{ij}^{m}} \tag{5}$$

$$u_{ij} = \frac{1}{\sum_{k=1}^{c}\left(\frac{d_{ij}}{d_{kj}}\right)^{2/(m-1)}} \tag{6}$$

2　基于 GEP 的模糊迭代自组织算法

本文应用基因表达式编程(GEP),结合 ISODATA 模糊聚类算法,提出了基于基因表达式的模糊迭代自组织(Fuzzy Interactive Self-Organizing Data Analystic, FISODATA-GEP)算法。该算法通过遗传演化,不断优化种群结构,自动寻找最佳聚类中心,从而得到最优解。

2.1　基因编码

GEP 的基因由一头一尾两部分构成。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[8]。对每个问题而言,头的长度 h 是选定的,而尾的长度 t 满足式(7)[6]:

$$t = h(n-1)+1 \tag{7}$$

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),本文中 $n = 2$,同时为使得聚类表达式树(Expression Trees,ET)具有自动聚类功能,设计头部函数集 $F = \{P,Q\}$,其中 P 函数将每个子表达式树中的元素构成一个簇的中心点,Q 函数计算叶子节点均值作为一个簇的中心点,其功能示例图如图 1 所示;终结符集合 $T = \{x_1,x_2,x_3,\cdots,x_n\}$,其中 x_i 代表当前系统中某个聚类 C_i 的中心,如染色体 $G = PQPx_1x_2x_3x_4$,编码对应的聚类表达式树[4]如图 2 所示。

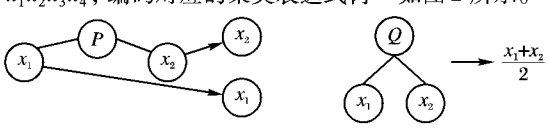

图 1　函数功能示例图

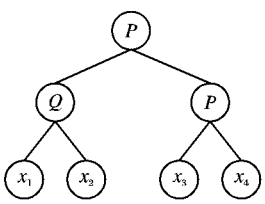

图 2　染色体 G 对应的表达式树

2.2　适应度函数

解码染色体可得到各可能的簇中心,再根据每个数据点与各个簇中心的距离,依次将其赋给最近的簇,然后重新计算聚类后各个簇的中心,本文在模糊 ISODATA 聚类方法基础上,取适应度函数 $f = 1/J(\boldsymbol{U},c_1,\cdots,c_c)$,$J(\boldsymbol{U},c_1,\cdots,c_c)$ 的值越小,表明生成的簇越紧密和独立。

2.3　遗传算子

1)选择算子。采用被广泛使用的轮盘赌选择策略[7],同时用精英保留策略来保持进化中出现的优良特性。适应度越好的个体被复制到下一代的可能性越大,复制过程中根据轮盘赌原则的结果确定被复制的次数,并且种群大小保持不变。

2)重组算子和变异算子。在 GEP 中有三种重组算子:单点重组、两点重组和基因重组。本文采用单点重组和双点重组方式,基因头部变异法则为函数符 $P\rightarrow Q$,$Q\rightarrow P$,基因尾部的变异采用从样本集中随机选取一个数据的方式进行无重复替换,即基因尾部变异后产生的新基因码与变异前的旧基因码不相同。

2.4　自动合并聚类算法

经过 GEP 进化计算找到一个最佳的聚类划分之后,可以对聚类结果进行进一步的整理操作,进一步合并其中的某些簇。其步骤见文献[9]。

2.5　FISODATA-GEP 算法步骤

FISODATA-GEP 算法的形式化语言描述如下。

```
输入:输入 FISODATA-GEP 样本数据集,算法参数参见表 1。
输出:聚类中心和模糊分类矩阵 U。
Procedure FISODATA-GEP
  {
  Initialize 初始化
  聚类样本集、种群大小 m_nGroupSize
  基因重组率 m_PRecombination
  基因变异率 m_PMutation
  运行代数 m_GEPNum
  种群大小 m_nGroupSize;
  g = 0;                          /*g 为进化代数计数器 */
  初始化种群 P(0);
  while (g < m_GEPNum){
    for (i = 0;i < m_nGroupSize;i++){
      将染色体翻译成表达式树;
      计算第 g 代种群各个体的适应度;
      }
    for (i = 0;i < m_nGroupSize;i++){
      对簇中心进行聚类并更新聚类中心和模糊分类矩阵 U;
      精英保留策略保留最佳个体;
      对 P(t) 代种群进行选择、重组、变异操作;
      }
    产生下一代初始种群;
    }
  自动合并聚类;
  输出;
  }
```

该算法的时间复杂度主要耗费在更新模糊矩阵和 GEP 演化过程。模糊 ISODATA 算法的时间复杂度是 $O(s \times k \times d)$,其中 s 为样本数,k 为类别数,d 为样本维数。设 m 为 GEP 中 k-表达式字符串的长度,k 为总算法的运行代数,则 FISODATA-GEP 算法的时间复杂度为 $O(k*m + s \times k \times d)$。

在理论上讲进行 GEP 演化操作的样本容量越大,聚类的误差越小,而 GEP 算法又可以对全局进行最优化处理,因此 FISODATA-GEP 算法在可接受的时间代价下,在聚类精度和收敛速度上都能显著提高,该算法利用模糊 ISODATA 良好的非监督学习动态聚类能力,结合基因表达式对聚类进行自动寻优,避免陷入局部最优解,在性能上相比于传统的模糊

ISODATA 聚类算法获得较大提高。

3 实验性能与对比分析

FISODATA-GEP 算法融合了基因表达式编程和模糊 ISODATA 聚类算法[10]，只要基因的尾部长度固定，则经过自动聚类演化并合并后的簇的个数小于等于基因的尾部长度，能够自动提供给决策者最佳的选址方案，而且聚类结果中生成隶属度矩阵能从侧面反映出每个样本划分在某类中的归属程度，具有很强的实用性，将其运用于 GIS 物流选址[11]中，使得聚类结果更加具有参考价值。本文即采用在 GIS 物流选址问题中进行实验性能测试，并与参考文献[12]中的聚类算法进行对比分析。

3.1 实验环境与数据

本文实验的硬件环境是：Pentium4 CPU(1.80 GHz)，内存容量 1 GB，软件环境为：Microsoft Windows XP 操作系统，VC++6.0和 Matlab 编程实现。本文中考虑到 GIS 物流选址中影响物流选址的因素有：人口(population)、地段价格(price)、劳动力成本(labor)、距离(distance)、竞争对手数目(opponent)等五个因素。实验数据采用文献[10]中的物流选址样本集和数据标准化方法，使用本文 FISODATA-GEP 算法对其进行聚类，并与 *K*-means 算法、模糊 ISODATA 算法及参考文献[11]中的基于 GEP 的 *K* 均值自动聚类算法(*K*-means Auto-Clustering Algorithm, GEPKMCA)进行对比分析。

3.2 实验参数

本文中 FISODATA-GEP 算法输入参数见表 1，基因头部长度 $h = 8$，根据式(7)可得基因尾部长度 $t = 9$，即随机初始化聚类为 9 类，经过 FISODATA-GEP 演化，输出聚类可以自动合并。

表 1 FISODATA-GEP 算法输入参数设置

参数名	参数值
ISODATA 加权指数 *m*	2
运行代数 *m_GEPNum*	100
种群大小 *m_nGroupSize*	50
下一个初始种群大小 *m_nNextGroupSize*	10
头部的长度 *m_nHeadSize*	8
基因重组率 *m_PRecombination*	0.72
基因变异率 *m_PMutation*	0.09

3.3 实验结果对比分析

由表 2 的实验结果对比中可以看出，FISODATA-GEP 算法相比前面几种算法，由于充分利用了 GEP 的全局寻优能力和模糊 ISODATA 的软分类特性，使得聚类结果的精度有显著提高。由于 *K*-means 和模糊 ISODATA 算法给定的初始聚类数是固定的，所以得到的聚类个数都为 9，而 GEPKMCA 和 FISODATA-GEP 由于引进了 GEP 自动合并聚类功能，在输出聚类中心的同时聚类结果进行了进一步的整理，合并了某些簇，因此最终得到的聚类个数比初始化时的聚类个数要少，减少了计算量。在运行时间方面，*K*-means 和模糊 ISODATA 都是直接基于相似性度量的非监督学习聚类，未引入 GEP 演化迭代，而引入 GEP 算法用于聚类分析中，经过不断的演化迭代，在不需要先验知识的条件下对数据进行自动簇的划分及合并，运行时间可控，得到的聚类精度更高，智能化程度更高。

四种算法聚类过程比较见图 3，从图中可以清晰地看出 FISODATA-GEP 的优势所在，该算法是一种稳定自动聚类算法，不仅提高了聚类效率，而且提高了聚类精度，使得聚类结果更具有参考价值。

表 2 实验结果对比

算法	聚类精度/%	运行时间/ms	聚类数
K-means	68.5	80.5	9
模糊 ISODATA	70.2	92.0	9
GEPKMCA[10]	78.8	112.0	7
FISODATA-GEP	83.6	120.5	6

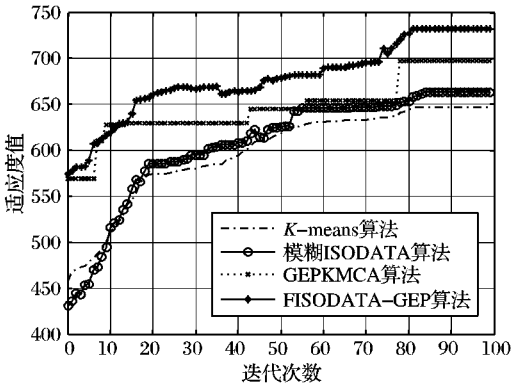

图 3 四种聚类算法的聚类比较过程

4 结语

本文结合基因表达式编程和模糊 ISODATA 算法，提出了 FISODATA-GEP 算法，并介绍了该算法的流程及步骤，最后通过将其运用到 GIS 物流选址中并与同类算法进行了结果对比分析，从结果可知 FISODATA-GEP 算法的稳定性与高效性，聚类结果具有很高的参考价值。但是，将 GEP 算法应用于聚类分析中，仍然还有许多问题需要研究与完善，如改进基因编码结构和适应度函数以进一步提高聚类精度，设定阈值来减少迭代次数，提高收敛速度等，这是下一步研究的主要内容。

参考文献：

[1] HAN J W, 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, 2005.

[2] 钱夕元, 邵志清. 模糊 ISODATA 聚类分析算法的实现及其应用研究[J]. 计算机工程与应用, 2004, 40(15): 70 - 71.

[3] 杨柳, 何锫, 潘小海. 基于堆栈解码的元胞基因表达式编程算法[J]. 计算机应用, 2009, 29(12): 3280 - 3282.

[4] FERREIRA C. Gene expression programming: A new adaptive algorithm for solving problems[J]. Complex Systems, 2001, 13(2): 16 - 30.

[5] BEZDEK J C. Pattern recognition with fuzzy objective function algorithms[M]. New York: Plenum, 1981.

[6] FERREIRA C. Gene expression programming in problem solving [EB/OL]. [2011 - 03 - 20]. http://www.gene-expression-programming.com/webpapers/ferreira-WSC6.pdf.

[7] FERREIRA C. Gene expression programming [M]. Portugal: Angra do Heroismo, 2002.

[8] MICHALEWICZ Z. 演化程序: 遗传算法和数据编码的结合[M]. 周家驹, 译. 北京: 科学出版社, 1999.

[9] 陈瑜, 唐常杰, 叶尚玉, 等. 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.

[10] 关庆, 邓赵红, 王士同. 改进的模糊 C-均值聚类算法[J]. 计算机工程与应用, 2011, 47(10): 27 - 29.

[11] 毛克彪, 谭志豪. 空间数据挖掘与 GIS 集成及应用研究[J]. 测绘与空间地理信息, 2004, 27(2): 14 - 17.

[12] 姜代红, 张三友. 基于基因表达式编程的 *K* 均值自动聚类算法[J]. 计算机仿真, 2010, 27(12): 216 - 219.