文章编号:1671-8836(2005) S2-0019-03

基于遗传算法的图论聚类方法

张 爱 华
(武汉大学 数学与统计学院, 湖北 武汉 430072)

摘 要: 针对传统图论聚类算法对初始聚类中心的敏感性以及聚类结果与样本输入次序等问题,提出了基于遗传算法进行图论聚类分析的基本原理和实现方法.实验结果表明,遗传算法应用于图论聚类分析能够搜索到更为精确的聚类中心值,其结果明显好于传统图论聚类算法.

关 键 词: 遗传算法; 聚类分析; 图论; 最小生成树
中图分类号: TP 301.6 **文献标识码**:A

0 引 言

聚类就是按照一定的要求和规律对事物进行区分和分类的过程[1].在这一过程中没有任何关于类分的先验知识,没有教师指导,仅靠事物间的相似性作为类属划分的准则,因此属于无监督分类的范畴.传统的聚类方法如图论聚类方法都是直接利用样本进行聚类,没有进行相关的预处理,其有效性在很大程度上取决于样本的分布情况,而对于呈任意形状簇分布的情况则聚类效果较差.

为此,本文提出了一种基于遗传算法的图论聚类方法,旨在实现分布呈任意形状簇的样本聚类,该方法既克服了聚类有效性对于样本分布的依赖性,又增加了聚类的灵活性和可视化.遗传算法是一种全局搜索算法[2],而图论聚类算法本质上是一种局部搜索算法,收敛于部分最优,容易陷入局部极小值,对于聚类样本数量较大的情况尤为明显.本文提出的方法在性能上较传统图论聚类方法有一定改进,聚类更准确,收敛时间较快,仿真实验验证了其有效性和可行性.

1 图论聚类方法

图论聚类方法最早是由 Zahn 提出来的[3],又称作最大(小)支撑树聚类算法,后来经过改造从而可以实现模糊聚类分析.首先,简单回顾一下有关图论的基本概念[4].一个无向图 G 是由一组节点 X 和连接在节点上的边 E 构成的.

$$G = [X, E], X = \{x_1, x_2, \cdots, x_n\},$$
$$E = \{e_{ij} = (x_i, x_j) \mid x_i, x_j \in X\}$$

图 G 中一条长度为 K 的路径 P 是一系列连接的节点, $P = \langle x_1, x_2, \cdots, x_{K+1}\rangle$,其中对 $\forall i \in (0, K)$, $(x_i, x_{i+1}) \in E$;如果图 G 中没有一条非零长度的路径 $P = \langle x_1, x_2, \cdots, x_{k+1}\rangle$,且 $x_1 = x_{k+1}$,则称图 G 不包含环;图 G 的支撑树是指由连接所有接点的 $n-1$ 条边构成的无环图 $[X, T]$.显然,一个图 $[X, T]$ 中当且仅当任意两对节点之间只有一条路径时才是树.通常在一个图 G 中可以构造多个支撑树 $[X, T_i]$ $(i > 1)$,如果给图中每条边 e 赋以权值 $\omega(e)$,那么最小支撑树(MST)是指满足下列条件的支撑树:

$$\omega(\mathrm{MST}) = \min_i \{\sum_{e \in T_i} \omega(e)\}$$

对于一棵树 $[X, T]$,如果移去一条边 e,则生成两组连通的节点 $A \subset X$ 和 $\overline{A} = X - A$,定义 θ 为共环边:$\theta = \{e_{ij} \mid x_i \in A, x_j \in \overline{A}\}$.也就是说,$\theta$ 为图 $[X, G]$ 中连接两组节点 A 和 $\overline{A}$ 的一组边;森林是指不包含环的非连通图.其中的每一个连通的部分被称为一棵树.

在传统的图论聚类分析中,首先把待分类的对象 $X = \{x_1, x_2, \cdots, x_n\} \subset \mathbf{R}^p$ 看作一个全连接的无向图 $G = [X, E]$ 中的节点,然后给每一条边赋以权值,比如可以用任意两个节点 (x_i, x_j) 在特征空间的海明距离定义边 e_{ij} $(1 \leq i, j \leq n)$ 的权值为 $\omega(e_{ij}) = \| x_i - x_j \|$, $x_i, x_j \in X$,然后用以下步骤对该组对象进

收稿日期: 2005-09-05
作者简介: 张爱华(1978-),男,硕士生,现从事系统优化等方面研究. E-mail:aihua1682@sina.com

行聚类分析:

步 1 利用 Prim 算法[5]在图 $G=[X,E]$ 上构造最小支撑树 MST:

$$\mathrm{MST}=\{(A,T)\mid A=X,T=\{e_1,\cdots,e_{n-1}\}\},$$

$$\omega(\mathrm{MST})=\min\{\omega(\mathrm{Tree})\mid \mathrm{Tree}=(X,T')\};$$

步 2 给定一个阈值 γ,从 MST 中移去权值大于阈值的边,形成 X 上的森林 F:

$$F=\{(X,E')\mid E'=T-\{e'\mid \omega(e')>\gamma\}\};$$

步 3 获得包含在森林 F 中的所有树$\{(X_i,T_i)\mid i=1,2,\cdots,m\}$:

$$F=\bigcup_{i=1}^{m}(X_i,T_i),\text{其中}\bigcup_{i=1}^{m}X_i=X,\bigcup_{i=1}^{m}T_i=E';$$

步 4 每棵树(X_i,T_i)被称作一个聚类.

2 图论聚类的遗传算法

由于上述方法不适用于大数据量的情况,难以满足实时性要求较高的场合,从而不能保证聚类结果的最优性,针对这些缺陷,本文提出了基于遗传算法的图论聚类方法.

2.1 编码设计

设某无向图中共有 m 个顶点、n 条边,对图中的节点和边进行编号. 以图中的所有边为编码变量,各个编码变量的取值为 0 或 1,则用一个长度为 n 的二进制字符串表示该图的所有子图. 当字符串中某位上的字符值为 1 时,表示它所对应的一条边是构成子图的边;当字符值为 0 时,表示它所对应的边不是构成子图的边.

2.2 适应度函数设计

遗传算法在进化搜索过程中只要求目标函数是非负的最大值形式,而对目标函数的定义域、连续性、可微性等没有限制. 在具体应用中可根据问题的性质,采取一定的转化策略,将优化问题的目标函数映射成求非负的最大值形式,通常将转换后的目标函数称为"适应度函数".

设某一无向图中各边的权值为 $\omega_i(i=1,2,\cdots,n)$,对于求权最小的生成树问题,可定义其适应度函数为

$$F=F_0-\sum_{i=1}^{n}\omega_i x_i,\ x_i\in[0,1],i=1,2,\cdots,n$$

其中 F_0 是一个较大的正常数,保证个体的适应度 F 总为非负;ω_i 为第 i 条边的权值;x_i 为个体编码中第 i 位上的字符值(对应于图中的第 i 条边),为 0 或 1.

在遗传进化过程中,必须对每一个个体所代表的子图进行一次深度优先搜索,若能搜索到所有的顶点,则说明该个体是图的一个生成树,可用所定义的适应度函数表达式计算该个体的适应度值;否则令该个体的适应度值为零,使其在进化过程中的生存能力最低,逐渐从群体中淘汰出去.

2.3 遗传算子设计

根据生成树的特点可知,在长度为 n 的字符串中,必须有 $n-1$ 个字符位上的字符值为 1,才有可能是一个生成树,有小于或大于 $n-1$ 个字符值为 1 的个体必定不是一个生成树. 为了减少进化过程中不可行方案的产生,必须控制所产生的每个个体只有 $n-1$ 个字符值为 1,使其具备成为可行解的必要条件,然后通过连通性检验确定是否为一个生成树.

标准遗传算法主要通过交叉算子和变异算子产生新的子代个体. 对于最小生成树问题,交叉算子和变异算子容易破坏子体成为可行解的特性,即具有 $n-1$ 个字符值为 1,所产生的新个体往往不能构成有效的生成树,降低了遗传算法搜索最小树的能力. 为提高遗传算法搜索最小树的效率,结合生成树的图论特性,本算法放弃了标准遗传算法的交叉算子和变异算子,设计出新的单亲换位算子和逆转算子.

单亲换位算子对所选择个体的基因链上的任意一对基因进行交换,且所执行的基因交换次数以及被交换的基因位都是随机产生的. 个体 A 10101010 一次随机交换为个体 A' 11101000.

单亲换位算子能使任何一个母体通过有限次的基因换位生成一个新个体. 通过这种基因重组方式可以从一组群体出发,以较高的概率搜索到解空间的各个可行解,能有效避免遗传过程中无效个体的产生.

单亲逆转算子则是将母体基因链上的任意一段基因进行逆转,一次性地完成从一个母体突变为一个新子体. 与换位算子相比,逆转算子的执行速度较快,有助于将母体中未发生突变的有效基因段,直接遗传到子体中. 个体 A 10101010 基因段逆转成为个体 A' 10010110.

改进后的遗传算法具有一个突出特点:在子代群体的生成过程中,每个子体只有一个母体,通过对母体执行换位算子或逆转算子,产生出具有不同性状的新个体. 单亲换位算子和逆转算子不仅可以保证子代个体具有成为可行解的基本特性,而且可以提高对解空间的搜索能力.

2.4 对最小生成树进行聚类分析

选择某一个 α 值作截集,将 $T_{\max}$ 中小于 α 的边断开,使相连的各节点构成一类,当 α 由 1 下降到 0

时,所得的分类由细变粗,各节点所代表的分类对象逐渐归并,从而形成一个动态聚类谱系图.

3　仿真实验

为了验证上述算法的有效性,我们以某一数据库中的记录为例,采用传统图论聚类算法(C)和基于遗传算法的图论聚类算法(G)分别进行计算,计算结果如表 1.

由仿真结果表明,传统图论聚类算法因不能有效地处理局部极值问题,因此当初始聚类中心在整个样本空间不平衡时,它很难将这种不平衡纠正过来,从而导致聚类结果对初始聚类中心的选取有着很大的敏感性;而基于遗传算法的图论聚类算法因具有很好的处理局部极值能力,因此对初始聚类中心的选取以及样本的输入次序没有任何要求. 另一方面,从它们各自的收敛速度上来看,基于遗传算法的图论聚类算法的收敛速度较快,而且比用常规方法对不同初始聚类中心进行聚类来获取全局最优解要有效的多.

表 1　两种算法的结果比较

样本	(0,1)	(1,0)	(1,1)	(2,2)	(2,3)	(3,2)	(5,6)	(6,5)	(6,6)	(7,7)	(7,8)	(8,7)
C	1	2	2	3	3	3	4	4	4	4	4	4
G	1	1	1	2	2	2	3	3	3	4	4	4

C 和 G 聚类算法的目标函数分别为 11.938 和 5.483

4　总　结

以上研究了基于遗传算法的图论聚类方法及寻找聚类中心的效果,并与传统图论聚类方法结果进行了比较. 传统图论聚类方法的有效性依赖于样本的分布情况,若样本界限分明,则聚类效果好. 但是实际情况往往是样本分布呈任意形状簇,对于这类情形,已有的方法效果不佳. 本文提出的基于遗传算法的图论聚类方法,通过遗传算法最终得到全局最优解,且动态实现聚类. 仿真结果表明,该方法在性能上较传统图论聚类算法有一定改进,不依赖于样本特征空间的分布情况,具有更准确的聚类能力和较快的收敛速度.

参考文献:

[1] 黄凤岗, 宋克欧. 模式识别[M]. 哈尔滨: 哈尔滨工程大学出版社, 1998.

[2] 刘　勇,康立山,陈毓屏. 非数值并行算法(第二册,遗传算法)[M]. 北京:科学出版社,1997.

[3] Zahn C T. Graph-Theoretical Methods for Detecting and Describing Gestalt Clusters[J]. *IEEE Trans on Computers*,1971,**20**(1):68-86.

[4] Balakrishnan R, Ranganathan K. *A Textbook of Graph Theory* [M]. New York: Springer-Verlag, 1999.

[5] 朱求长. 运筹学及其应用[M]. 武汉:武汉大学出版社,1997.

Graph-Theoretical Clustering Method Based on Genetic Algorithms

ZHANG Ai-hua

(School of Mathematics and Statisics, Wuhan University, Wuhan 430072, Hubei, China)

Abstract: To solve the problem of sensitivity with the original clustering center and clustering results depended on the order of the input example in common graph-theoretical clustering algorithm, the basic rules and procedures of applying genetic algorithms to graph theoretical clustering analysis are studied. Computing results show that applying it to clustering can accurately locate the clustering centers and is superior to common graph-theoretical clustering algorithm.

Key words: Genetic Algorithms; cluster analysis; graph theoretical; minimum spanning tree