基于模拟退火遗传算法的聚类分析*

武兆慧，张桂娟，刘希玉
（山东师范大学 信息管理学院，山东 济南 250014）

摘 要：将模拟退火遗传算法用于聚类分析，通过对聚类中心进行编码，定义适应度函数，选择、交叉、变异操作以及模拟退火算法的运用，给出了一种新的基于模拟退火遗传算法的聚类算法，实验结果显示该方法优于基本的遗传算法。

关键词：聚类；遗传算法；模拟退火算法；模拟退火遗传算法
中图法分类号：TP391 文献标识码：A 文章编号：1001-3695(2005)12-0024-03

Clustering Based on Simulated Annealing Genetic Algorithm

WU Zhao-hui, ZHANG Gui-juan, LIU Xi-yu
(School of Information & Management, Shandong Normal University, Jinan Shandong 250014, China)

Abstract: This paper combines the simulated annealing genetic algorithm with clustering. A new clustering algorithm based on simulated annealing genetic algorithm is presented by encoding the cluster centers, defining the fitness, selection, crossover, mutation and using simulated annealing algorithm. Experimental results demonstrate that this method is better than simple genetic algorithm.

Key words: Clustering; Genetic Algorithm; Simulated Annealing Algorithm; Simulated Annealing Genetic Algorithm

1 引言

聚类分析是依据样本间关联的度量标准将其自动分成几个群组，且使同一群组内的样本相似，而属于不同群组的样本相异的一组方法。一个聚类分析系统的输入是一组样本和一个度量两个样本间相似度（或相异度）的标准。聚类分析的输出是数据集的几个组（类），这些组构成一个分区或一个分区结构。

Metropolis等人于1953年提出了模拟退火算法，其基本思想是把某类优化问题的求解过程与统计热力学中的热平衡问题进行对比，固体退火过程的物理图像和统计性质是模拟退火算法的物理背景，Metropolis接受准则使算法跳离局部最优的“陷阱”，而冷却进度表的合理选择是算法应用的前提。固体退火是先将固体加热至熔化，然后徐徐冷却使之凝固成规整晶体的热力学过程。从统计物理学的观点看，随着温度的降低，物质的能量将逐渐趋近于一个较低的状态，并最终达到某种平衡。

遗传算法的主要思想是基于C. R. Darwin的生物进化论和G. Mendel的遗传学。遗传算法结合了Darwin的适者生存和随机交换理论，是一种自然进化系统的计算模型，也是一种通用的求解优化问题的适应性搜索方法。

遗传算法在运行早期个体差异较大，当采用经典的轮盘赌方式选择，后代产生的个数与父个体适应度大小成正比，因此在早期容易使个别好的个体的后代充斥整个种群，造成早熟（Premature）。在遗传算法后期，适应度趋向一致，优秀的个体在产生后代时，优势不明显，从而使整个种群进化停滞不前。因此对适应度适当地进行拉伸是必要的，这样在温度高时（遗传算法的前期），适应度相近的个体产生的后代概率相近；而当温度不断下降后，拉伸作用加强，使适应度相近的个体适应度差异放大，从而使得优秀的个体优势更明显。

本文将模拟退火算法与遗传算法相结合并进行了改进再用于聚类分析，由于模拟退火算法和遗传算法可以互相取长补短，因此有效地克服了传统遗传算法的早熟现象，同时根据聚类问题的具体情况设计遗传编码方式、适应度函数，使该算法更有效、更快速地收敛到全局最优解。

2 模拟退火遗传算法

2.1 模拟退火算法

模拟退火算法（Simulated Annealing Algorithm）于1983年成功地应用在组合优化的问题上。其思想是通过模拟高温物体退火过程的方法来找到优化问题的全局最优或近似全局最优解。首先产生一个初始解作为当前解，然后在当前解的邻域中，以概率P(T)选择一个非局部最优解，并令这个解再重复下去，从而保证不会陷入局部最优。开始时允许随着参数的调整，目标函数偶尔向增加的方向发展（对应于能量有时上升），以利于跳出局部极小区域。随着假想温度的降低（对应于物体的退火），系统活动性降低，最终以概率1稳定在全局最小区域。模拟退火算法描述如下：

（1）选S_0作为初始状态，令$S(0)=S_0$，同时设初始温度

收稿日期：2004-10-20；修返日期：2005-02-05
基金项目：国家自然科学基金资助项目（60374054）；山东省中青年科学家奖励基金资助项目（304065）；山东省科技攻关项目（012090101）

T;令 i=0。

(2)令 T=T_i,以 T和 S_i调用 Metropolis抽样算法,返回状态 S'作为本算法的当前解,S_i=S'。

(3)按照一定方式降温,即 T=T_{i+1},其中 T_{i+1}<T,i=i+1。

(4)检查终止条件,如果满足则转步骤(5),否则转步骤(2)。

(5)当前解 S_i为最优解,输出结果,停止。

Metropolis抽样算法描述如下:

(1)令 k=0时,当前解 S(0)=S,在温度 T下,进行以下各步操作。

(2)按某个规定的方式根据当前解 S(k)所处的状态 S产生一个近邻子集 N(S(k))⊂S,从 N(S(k))中随机得到一个新状态 S'作为下一个候选解,计算能量之差 ΔC'=C(S')-C(S(k))。

(3)如果 ΔC'<0,则接受 S'作为下一个当前解,否则,以概率 exp(-ΔC'/T)接受 S'作为下一个当前解。若 S'被接受,则令 S(k+1)=S',否则 S(k+1)=S(k)。

(4)k=k+1,检查算法是否满足终止条件,若满足,则转步骤(5),否则转步骤(2)。

(5)返回 S(k),结束。

2.2 模拟退火遗传算法

Paul L. Stoffa借鉴模拟退火思想提出了模拟退火遗传算法(Simulated Annealing Genetic Algorithm, SAGA)。

模拟退火遗传算法的基本思想是将遗传算法与模拟退火算法结合起来构成的一种混合优化算法。遗传算法的局部搜索能力较差,但把握搜索过程总体的能力较强;而模拟退火算法具有较强的局部搜索能力,并能使搜索过程避免陷入局部最优解,但模拟退火算法却对整个搜索空间的了解不多,不便于使搜索过程进入最有希望的搜索区域,从而使得模拟退火算法的运算效率不高。但如果将遗传算法和模拟退火算法相结合,互相取长补短,则有可能开发出性能优良的新的全局搜索算法[1]。

与基本遗传算法的总体运行过程相类似,模拟退火遗传算法也是从随机产生的初始解(初始群体)开始全局最优解的搜索过程,它先通过选择、交叉、变异等遗传操作来产生一组新的个体,然后再独立地对所产生的各个个体进行模拟退火过程,以其结果作为下一代群体中的个体。这个运行过程反复迭代地进行,直到满足某个终止条件。

模拟退火遗传算法描述如下:

(1)进化代数计数器初始化,t←0。

(2)随机产生初始群体 P(t)。

(3)评价群体 P(t)的适应度。

(4)个体交叉操作,P'(t)←Crossover[P(t)]。

(5)个体变异操作,P''(t)←Mutation[P'(t)]。

(6)个体模拟退火操作,P'''(t)←SimulatedAnnealing[P''(t)]。

(7)评价群体 P'''(t)适应度。

(8)个体选择、复制操作,P(t+1)←Reproduction[P(t)YP'''(t)]。

(9)终止条件判断。若不满足终止条件,则 t←t+1,转到步骤(4),继续进化过程;若满足终止条件,则输出当前最优个体,算法结束。

本文结合遗传算法对模拟退火算法进行改进:采用动态调节近邻子集的方法,利用以下公式确定近邻子集的容量 M的大小:M=λ×(f_max-f_avg)。其中,f_max为最大适应度值,f_avg为平均适应度值,λ为系数。这样在算法前期,f_max与 f_avg相差较大,所以此时取近邻子集的容量较大;在算法后期,f_avg与 f_max的差别越来越小,近邻子集的容量也随之变小。通过动态地调节近邻子集的大小可以使搜索在更有效的范围内进行。

3 基于模拟退火遗传算法的聚类算法

(1)编码方式。遗传聚类算法中,可以采用的染色体编码方式有两种:基于聚类中心的浮点编码和基于聚类划分的整数编码。由于聚类问题的样本数目一般都远大于其聚类数目,因此采用聚类中心的编码方式更有效。由于二进制编码在运算时要反复进行编码、译码操作,所以本文采用聚类中心的浮点编码方式,每条染色体由 k个聚类中心组成:C=c_1 c_2… c_k 对于 m维的样本向量,其染色体为长度是 k×m的浮点码串。

(2)适应度函数。遗传算法在进化搜索中以适应度函数为依据,利用种群中每个个体的适应度值进行搜索,因此适应度函数的选取至关重要,直接影响算法的收敛速度及最优解的寻找。在选定 k个聚类中心后,将每个样本向量 x_i=[x_i1, x_i2, …, x_im]^T(m为输入向量的维数)按下列欧氏距离归入中心为 c_i的类中:‖x_i-c_i‖=min_j‖x_i-c_j‖。定义目标函数 $J=\sum_{k=1}^{m}\sum_{x_k\in C_i}|x_k-c_i|$,其中 C_i是中心为 c_i的聚类块;定义适应度函数 $f=\frac{1}{1+J}$,这样目标函数 J的值越小,说明类内离散度和小,相应的适应度值就越大。

(3)交叉操作是把两个父个体的部分结构加以替换重组而生成新个体的操作。可以采用离散重组方法,即在个体之间交换变量的值,子个体的每个变量可按等概率随机地挑选父个体。如父个体 1:0.15 0.20 0.35,父个体 2:0.10 0.33 0.28,则重组之后的子个体可为子个体 1:0.10 0.33 0.35,子个体 2:0.15 0.33 0.35。

(4)变异操作体现了生物遗传的多样性。变异操作将每个基因位上的浮点数以一定的概率发生变异,发生变异的基因位被一个随机数代替。

(5)个体模拟退火操作是用适应度值作为模拟退火算法中的能量,当适应度值增加时,接受该解作为下一个当前解,否则以一定的概率接受该解。算法采用如下的适应度拉伸方法:$f_i=\frac{e^{f_i/T}}{\sum_{i=1}^{M}e^{f_i/T}}$,$T=T_0(0.99^{g-1})$,式中,f_i为第 i个个体的适应度,M为种群大小,g为遗传代数,T为温度,T_0为初始温度。温度 T由随着算法进程递减其值的控制参数担当。

(6)选择操作采用轮盘赌方法选择优良个体作为下一代。

(7)终止条件判断。兼顾算法的优化性能和优化速度,以最优指标连续 Q步保持不变为终止准则,即在算法初始化时置最优指标为群体中最优个体的综合指标值,同时置终止计数变量 d为 0,若每一次算法流程中最优指标发生改变就及时更新最优指标值并置 d为 0,否则令 d=d+1,当 d=Q时结束算法。

4 实验

实验是在一台内存为 256MB,中央处理器为赛扬 1.7GHz 的微机上进行的,所采用的软件为 Visual C++ 6.0. 本文将 SGA和 SAGA分别作用于两组随机产生的数据进行实验。其中,第一组数据由 50个二维平面上的点组成,这些点构成五个集合;第二组数据由 300个四维空间的点组成,聚类数目为 6,但彼此之间并没有明显的界限。本实验所采用的交叉概率为 0.8,变异概率为 0.02,种群规模为 50. 用这两种算法分别运行五次,计算它们的目标函数值 J,实验结果如表 1和表 2所示。

表 1 第一组数据实验结果

运行次数	SGA	SAGA
1	63.0819	63.0819
2	63.0819	63.0819
3	63.1205	63.0819
4	63.3213	63.0819
5	63.0819	63.0819

表 2 第二组数据实验结果

运行次数	SGA	SAGA
1	172.2511	171.2106
2	172.4621	171.2106
3	171.5006	171.2106
4	172.0615	171.2106
5	171.3290	171.2106

实验结果显示,本文提出的算法在解决聚类问题方面优于传统的遗传算法。用 SAGA作用于第一组数据可以发现每次实验都可以得到最优的目标函数值 J=63.0819,而采用 SGA并不是每次都能得到最优目标函数值;用该算法作用于第二组数据同样可以看出该算法优于传统的遗传算法,而且这种优势更加明显,也就是说当数据量较大时,SAGA的优越性更加明显。其主要原因是基本遗传算法在处理大规模数据时,容易收敛到局部最优解,而将遗传算法与模拟退火算法相结合形成一种混合算法后,可以有效地克服收敛到局部最优解的情况。

5 结束语

遗传算法作为一种优化算法特别适合于对象模型难以建立、搜索空间非常庞大的复杂问题的优化求解,它可以在对领域知识有较少了解的情况下解决问题。本文将模拟退火算法与遗传算法相结合并加以改进,然后用于聚类分析,利用模拟退火算法较强的局部搜索能力和遗传算法较强的全局搜索能力,有效、快速地解决了聚类问题。

参考文献:

[1] 史忠植. 知识发现[M]. 北京:清华大学出版社,2002.

[2] 王小平,曹立明. 遗传算法——理论、应用与软件实现[M]. 西安:西安交通大学出版社,2002.

[3] 傅景广,许刚,王裕国. 基于遗传算法的聚类分析[J]. 计算机工程,2004,30(4):122-124.

[4] Ali Kamrani, Wang Rong, Ricardo Gonzalez. A Genetic Algorithm Methodology for Data Mining and Intelligent Knowledge Acquisition[J]. Computers & Industrial Engineering, 2001, 40: 361-377.

[5] 高坚. 基于 C均值和免疫遗传算法的聚类分析[J]. 计算机工程,2003,29(12):65-66.

[6] Bandyopadhyay S, Maulik U. Genetic Clustering for Automatic Evolution of Clusters and Application to Image Classification[J]. Pattern Recognition, 2002, 35: 1197-1208.

[7] Richard J Roiger, Michael W Geatz. Data Mining a Tutorial-based Primer[M]. 北京:清华大学出版社,2003.

[8] 崔勇,吴建平,徐恪. 基于模拟退火的服务质量路由算法[J]. 软件学报,2003,14(5):877-884.

作者简介:

武兆慧(1981-),女,硕士研究生,主要研究方向为数据挖掘、遗传算法;张桂娟(1981-),女,硕士研究生,主要研究方向为进化计算;刘希玉,男,教授,博士生导师,博士,主要研究方向为进化计算、实体造型技术等。

(上接第 20页)

[50] Ha S H, Bae S M, Park S C. Web Mining for Distance Education[C]. Proc of IEEE International Conference on Management of Innovation and Technology (ICMIT), 2000. 219-715.

[51] Zaiane O R, Xin M, Han J. Discovering Web Access Patterns and Trends by Applying OLAP and Data Mining Technology on Web Logs[C]. International Conference on Management of Innovation and Technology (ICIT), 2002. 219-715.

[52] Ochi Y, Yano Y, Hayashi T, et al. JUPITER: A Kanji Learning Environment Focusing on a Learner's Browsing[C]. Proc of the 3rd Asia Pacific Conference on Computer Human Interaction, 1998. 446-451.

[53] McCalla G, Vassileva J, Bull S. Active Learner Modeling[C]. Proc of the 15th International Conference on Intelligent Tutoring Systems, 2000. 53-62.

[54] Tiffany Y Tang, Gordon McCalla. Student Modeling for a Web-based Learning Environment: A Data Mining Approach American Association for Artificial Intelligence[EB/OL]. http://www.aaai.org, 2004-08-03.

[55] 莫赞,冯珊,唐超. 智能教学系统的发展与前瞻[J]. 计算机工程与应用,2002,38(6):6-8.

[56] 范玉顺,曹军威. 多代理系统理论、方法与应用[M]. 北京:清华大学出版社,2002.

[57] 徐志伟,冯百明,李伟. 网格计算技术[M]. 北京:电子工业出版社,2004.

[58] Stewart Fraser, Steven Livingstone. C#XML入门经典[M]. 毛尧飞,崔伟. 北京:清华大学出版社,2003.

[59] Mark Graves. XML数据库设计[M]. 尹志军,等. 北京:机械工业出版社,2002.

[60] 马秀芳. 基于 LOM的资源检索 Agent系统研究[DB/OL]. http://www.swnu.edu.cn, 2004-12-20.

[61] 李建国,张小真. 计算机辅助教学[M]. 重庆:重庆大学出版社,1993.

[62] 邓志鸿,唐世渭,等. Ontology研究综述[J]. 北京大学学报(自然科学版),2002,3(5):16-20.

[63] 杜小勇,李曼,王大治. 语义 Web与本体研究综述[J]. 计算机应用,2004,24(10):14-16,20.

[64] Driscoll G O. The Essential Guide to Home Networking Technologies[M]. Prentice Hall PTR, 2001.

[65] Bray J. Bluetooth: Connect Without Cables[M]. Prentice Hall PTR, 2001.

作者简介:

李静(1980-),女,河南沈丘人,硕士研究生,主要研究方向为人工智能与计算机辅助教育;周竹荣,男,副教授,硕士生导师,博士,主要研究方向为人工智能与计算机辅助教育。