自适应小生境遗传算法在关联规则挖掘中的应用*

杨小影[1]，冯艳茹[1]，钱 娜[2]
（1.济源职业技术学院 计算机系，河南 济源 454650；
2. 济南大学 信息科学与工程学院，山东 济南 250022）

摘 要：传统的遗传算法存在早熟收敛和易于陷入局部搜索最优等缺陷；根据关联规则挖掘的要求和特点，提出一种应用于关联规则挖掘的自适应小生境遗传算法。

关键词：关联规则；自适应小生境遗传算法；选择；杂交
中图分类号：TP391.75 **文献标识码**：B

The application of adaptive niche genetic algorithm in association rules mining from data

YANG Xiao Ying[1], Feng Yan Ru[1], QIAN Na[2]
(1. Department of Computer, Jiyuan University of Vocation and Technology, Jiyuan 454650, China;
2. School of Information Science and Engineering, University of Jinan, Jinan 250022, China)

Abstract:Premature convergnce and weak local optimization are two key problems existing in the conventioal genetic algorithm.According to the requirement and character of association rules mining. The paper gives an adaptive niche genetic algorithm.

Key words: association rules mining; adaptive niche genetic algorithm; choose; hybrid

遗传算法（GA）是一种基于生物界适者生存理论的自适应搜索技术，其主要特点是群体搜索策略和群体中个体之间的信息交换，算法的搜索过程不依赖于目标函数的梯度信息[1-4]，目前它已经成功地应用于组合优化、自动控制等众多领域[5-6]。由于基本遗传算法所具有的特性，用它进行优化时的结果将使群体中的个体集中到目标函数值最大的一个峰值上，存在局部搜索能力不强，易陷入局部最优和早熟等缺陷，使得传统的GA在进行查询优化时效果不理想。在实际应用中有时希望最终搜索到的优化点不是只在一个峰值上，而是在多个峰值上都有分布，而且分布的多少与峰值的高低成正比。这就要求种群保持一定的个体多样性。这点在基于遗传的机器学习等问题中也尤为重要[2]。数据挖掘技术是机器学习、人工智能、数据系统等领域的研究方向。数据挖掘就是从大型数据库的大量原始数据中提取出人们感兴趣的、具有潜在应用价值的指示和信息。其中关联规则是最有用的信息之一，它用于发现大量数据项集合之间的关联[7]。本文提出一种自适应小生境遗传算法应用于关联规则挖掘技术。

1 关联规则的描述

令 $I=\{i_1,i_2,\ldots,i_d\}$是事务中所有项目的集合，而 $T=\{t_1,t_2,\ldots,t_n\}$是所有事务的集合。每个事务 t_i包含的项集都是 I的子集。在关联分析中，包含0个或多个项的集合被称为项集。关联规则(Association Rule)是形如 $X\rightarrow Y$的蕴涵表达式，其中 X和 Y是互不相交的项集。关联规则可以用它的支持度(support)和可信度(confidence)度量。支持度确定规则中给定数据集的频繁程度，而可信度确定 Y在包含 X的事务中出现的频繁程度。给定事务的集合 T，关联规则发现是指找出支持度大于等于minsup并且可信度大于minconf的所有规则，其中minsup和minconf是对应的支持度和可信度阈值[8]。研究表明，支持度阈值随着项集长度的增加而递减，因此用参考文献[9]针

*基金项目：山东省研究生教育创新计划(SDYY08032)

对支持度阈值设置惩罚函数可表示为：

$$\delta(l)=\frac{l}{(2^{l-1})^{\omega}}$$

其中 l 为相继长度，$\omega=(0,1]$。

2 自适应小生境遗传算法原理

2.1 小生境技术的生物学基础

在自然界，“物以类聚，人以群分”的小生境现象普遍存在，生物总是喜欢同自己形状、习性相似的生物在一起，并与同类交配繁衍后代，在生物学中，把某种特定环境及其在此环境中生存的组织称为小生境。小生境的形成在生物学上有着积极的意义，为新物种的形成提供了可能性[6]。

在具体的工程应用中，小生境技术演变为：将每一代个体划分为若干类，每个类中选出若干适应度较大的个体作为一个类的优秀代表组成一个种群，再在该种群与不同种群之间通过杂交、变异产生新一代个体群，同时采用预选择机制、排挤机制或共享机制完成选择操作。也就是说让个体在一个特定的生存环境中进化，形成多个小生境，最终达到小生境内的峰值，从而找到全局最优解。受此启发，近年来人们将小生境现象引入到遗传算法中，实践证明，这一技术对于改善遗传算法全局收敛性能具有良好的效果[10]。

2.2 自适应小生境遗传算法原理

为解决传统遗传算法种群多样性低的问题，自适应小生境遗传算法提出：首先将初时种群中的个体按适值排序，然后相似的若干个体进入一个小生境即子种群中独立进化。子种群的规模是随着大种群的多样性的变化而自适应变化的。设大种群的规模为 N，子种群规模为 K，则有：

$$K=\begin{cases} f(D) \\ 2 \end{cases}$$

其中，D 是大种群个体的方差，$f(D)$ 是关于 D 的一个函数，可根据问题的特征预先设置；σ 为一常数。

当大种群个体多样性降低时，D 就减小，当 D 小于某一阈值 σ 时，子种群规模 K 降低到最低限度 2。

在小生境技术中，插用（$\mu+\lambda$）选择机制，它被认为是集中流行进化算法的选择机制中选择率最高的一种。交叉操作采用均匀模板交叉算子。当交叉结束后，立即进入（$\mu+\lambda$）选择，以生成子种群的新一代个体。

新产生的个体进行随机变异，当变异的个体为子种群中的最佳个体时，应该对该最佳个体及其变异所得到的新个体进行（$1+l$）选择，以保证最优个体以概率 l 保留到下一代[11]。

算法描述如图 1 所示。

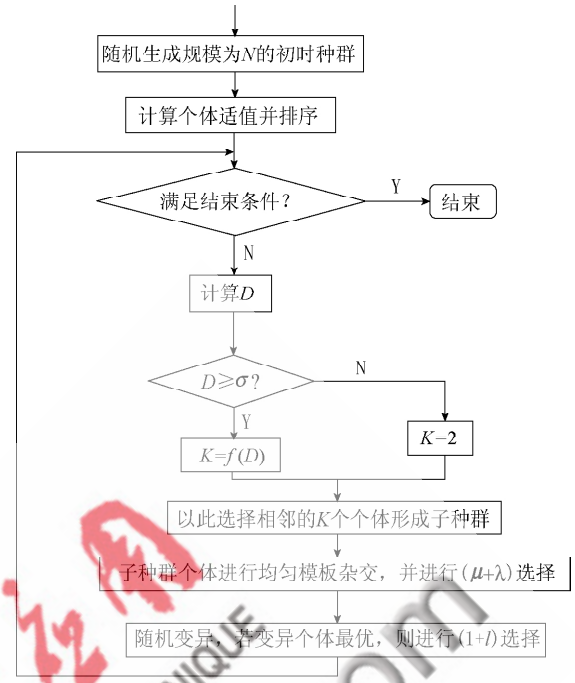

图 1 算法描述流程图

3 试验分析

3.1 数据库设计

采用某肿瘤医院的数据库进行试验。数据库中记录了从 1994 年～2003 年 2 600 多例肿瘤患者的病历，抽取出病历中的重要信息，构成数据表，如表 1 所示。

表1 数据表结构

字段名	类型	说明
name	Char	姓名
Num	Long	编号
Age	Int	年龄
Scx	Char	性别
Cac	Char	肿瘤种类
Tnm	Char	国际分期
Add	Char	住址
Zl	Char	诊疗计划
lx	char	治疗效果

3.2 数据库数字化

为了易于表示起见，将数据库中重要字段取值数字化。

肿瘤种类划分：肺癌；胃癌；乳腺癌；大肠癌；口腔癌；肝癌；宫颈癌；食管癌；其他。

诊疗计划划分：手术；放疗；化疗；生物免疫治疗；中医中药治疗。

治疗效果：治愈（5 年存活）；好转；恶化；死亡；自动出院。

国际分期划分：1（Ⅰ）；2（Ⅱa）；3（Ⅱb）；4（Ⅲ），5（Ⅳ）。

3.3 关联规则的提取

为挖掘数据库中蕴涵数字化属性间的关联规则，根

据以上数字化步骤，将4个属性分别划分为9、5、5、5个属性等级。设 X={肿瘤种类、国际分期}，Y={诊疗方案、治疗效果}，给定最小支持度和最小置信度都为0.02，表2列出部分有意义的所得到的优化语言值关联规则。

表2 部分关联规则

关联规则优化语言值	支持度	可信度
2 111	0.551 2	0.895 0
2 432	0.213 2	0.678 2
8 111	0.899 3	0.900 2
8 532	0.214 7	0.923 3
3 111	0.754 1	0.911 8
3 522	0.512 6	0.817 9

根据关联规则的特点和要求，提出了基于自适应小生境遗传算法的关联规则挖掘算法。试验显示，该方法快速有效。

参考文献

[1] RUDOLPH G.Convergence analysis of canonical genertic algorithme [J].IEEE Trans on Neural Network,1994,5(1)：96-101.

[2] 田盛丰.人工智能原理与应用[M].北京: 北京立功大学出版社，1993.

[3] FOGEL.An introduction to simulated evolutuionary ptionization[J]. IEEE Trans on Neural Network,1994,5(1): 3-14.

[4] 陈国良.遗传算法及其应用[M].北京: 人民邮电出版社，1996.

[5] SONG S K,GORLA N. Agenetic algorithm for vertical fragmentation and access path selection[J].The Computer Journal,2000,43 (1):81-92.

[6] JACK L B,NANDI A K. Genetic algorithms for feature selection in machine condition monitoring with vibration signals[J].IEEE Proceedings Vision,Image and Signal Processing，2000，47（3）：205 - 212.

[7] TAN Ping Ning ,STEINBACH M,KUMAR V.数据挖掘导论[M].北京：人民邮电出版社，2006.

[8] 潘舒，吴陈.基于遗传算法的关联规则挖掘[J].现代电子技术，2008，265（2）：90 - 92.

[9] 赵连朋，金喜子，孙亮，等.基于小生境遗传算法的关联规则挖掘方法[J].计算机工程，2008，34（10）：163-165.

[10] 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，2000.

[11] 郑宣耀，王芳.一种改进的小生境遗传算法[J].重庆邮电学院学报（自然科学学报），2005（2）.

（投稿时间：2008-12-06）

（上接第74页）

列可知，实施能量奖励策略后，多次实验的平均解要更接近指定的满意解，而且本文提出的能量奖励实施策略进一步提高了微正则退火算法的搜索能力。

图1是上述实验中4种算法最低运输成本的变化轨迹。可以看出微正则退火基本算法在最开始阶段的下降速度很快，但到后期不如其他算法。根据能量奖励策略的特点可知，它使得整个搜索过程更为平缓一些，前期目标函数值下降稍慢，这从一定程度上消弱了微正则退火算法本身的快速优化优势，但从图1可发现，施加了改进的能量奖励策略后，对上述问题有了明显的改善。

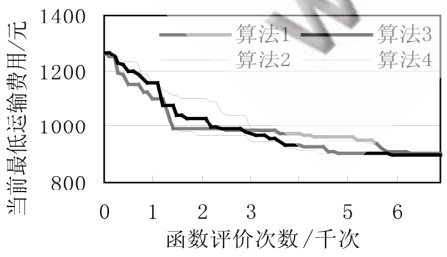

图1 最低运输费用的变化轨迹

微正则退火算法具有快速收敛特征，不少研究文献中都探讨了这种退火算法的工程应用价值。本文首先对笔者曾提出的能量奖励策略进行改进，根据能量差额比较的方法来决定是否启动奖励操作，并且尝试能量增幅比例 q 按线性方式调节。随后将新的改进算法应用车辆路径优化问题，并通过一个典型的单配送中心实例给出了初步比较。实验结果证明这种改进的微正则退火算法用于路径优化是十分有效的。本文的改进思路也有缺点，增加了算法代码的复杂程度，理论上会耗费更长的搜索时间，但在本实例上表现不明显，应该用更大规模的实例来检验，这将是下一步的工作内容。

参考文献

[1] PAOLO T, DANIELE V. The vehicle routing problem[M]. Philadelphia: Society for Industrial and Applied Mathematics,2002.

[2] 祝崇隽, 刘民, 吴澄. 供应链中车辆路径问题的研究进展及前景[J]. 计算机集成制造系统, 2001, 7 (11) :1-6.

[3] 崔雪丽, 马良, 范炳全. 车辆路径问题(VRP) 的蚂蚁搜索算法[J]. 系统工程学报, 2004, 19(4) :418-422.

[4] 胡大伟, 朱志强, 胡勇. 车辆路径问题的模拟退火算法[J]. 中国公路学报, 2006, 19(4) :123-126.

[5] 张波, 叶家玮, 胡郁葱. 模拟退火算法在路径优化问题中的应用[J]. 中国公路学报, 2004,17(1):79-81.

[6] CREYTZ M. Microcanonical Monte Carlo simulation[J]. Physical Review Letters, 1983, 50(19): 1411-1414.

[7] 李军, 谢秉磊, 郭耀煌. 非满载车辆调度问题的遗传算法[J]. 系统工程理论与实践, 2000, 20(3):235-239.

(收稿日期：2008-12-05)