一种改进的动态遗传 Apriori 挖掘算法

詹 芹[a], 张幼明[b]
(九江学院 a. 信息科学与技术学院; b. 机械与材料工程学院, 江西 九江 332005)

摘 要: 在经典关联规则算法 Apriori 的基础上,提出了一种改进的动态遗传 Apriori 挖掘算法。通过动态遗传 Apriori 挖掘算法对学生成绩管理数据库中的课程进行分析,找出各课程之间的隐藏关系,得到一些合理、可靠的课程关联规则,从而根据这些规则进行课程的合理设置。实验结果表明,该算法能高效地解决数据挖掘问题。

关键词: 关联规则; 数据挖掘; 遗传算法; Apriori 算法
中图分类号: TP391 **文献标志码**: A **文章编号**: 1001-3695(2010)08-2929-02
doi:10.3969/j.issn.1001-3695.2010.08.031

Improved dynamic genetic Apriori mining algorithm

ZHAN Qin[a], ZHANG You-ming[b]
(*a. School of Information Science & Technology, b. School of Mechanical & Materials Engineering, Jiujiang University, Jiujiang Jiangxi* 332005, *China*)

Abstract: This paper proposed a dynamic genetic Apriori algorithm based on traditional Apriori algorithm. Analyzing student's achievement utilizing dynamic genetic Apriori algorithm to administer to educational management database, relationship concealing finding out between each curriculum, it was rightful to obtain some dependable association rules, moreover on the basis of these regulations carries on the rightful installation of courses. Experiment results demonstrate that this method can solve data mining effectively.

Key words: association rule; data mining; genetic algorithm(GA); Apriori algorithm

随着数据和信息的爆炸式增长,人类分析数据和从中提取有用信息的能力远远不能满足实际需要,因此数据挖掘技术应运而生。数据挖掘是从数据库中抽取隐含的、以前未知的、具有潜在应用价值的信息过程[1]。关联规则挖掘是数据挖掘中非常重要的内容。近年来,人们研究了许多挖掘算法。Apriori 算法[2]是关联规则挖掘的经典算法,但计算复杂度高,不能满足对大规模数据库的实时挖掘要求;Park 等人[3]提出了 DHP 算法,该算法通过引入 hash 技术来提高生成频繁 2 项集的效率从而提高整个算法运行效率;FP-growth 算法[4]创建了一种关系紧密的树结构,有助于候选项目集的产生,但需要遍历整个数据库,计算量大,对大规模数据库而言,效率非常低;Brin 等人[5]提出了动态项集计数算法;文献[6]将遗传算法应用到关联规则挖掘中;文献[7]提出一种关联规则的矩阵算法;文献[8]设计一种基于事务压缩和项目压缩的算法,这些算法都在不同方面对关联规则算法尤其是经典 Apriori 算法进行了优化。

遗传算法(GA)是一种基于群体的进化算法,具有很强的随机性、鲁棒性和隐含并行性,能快速、有效地进行全局优化搜索,是处理大规模数据项目集的有效方法[9,10]。本文将遗传算法策略引入 Apriori 算法,提出了一种基于自适应策略的动态遗传 Apriori 挖掘算法(dynamic genetic Apriori algorithm, DGAA),并成功应用到某高校学生成绩管理数据库中。

1 问题的定义

关联规则是数据项之间存在的规则,是在同一事件中出现的不同项之间的相关性[11]。为了便于分析问题,特作以下形式化定义:

定义 1 数据项集 $I=\{i_1,i_2,\cdots,i_n\}$。其中 $i_k(1\leqslant k\leqslant n)$ 是数据项。

定义 2 事务数据集 $T=\{T_1,T_2,\cdots,T_m\}$。其中,$T_k(1\leqslant k\leqslant m)$ 是事务数据集 T 的数据项,也是数据项集 I 中的数据项,并且 $T\subseteq I$。

定义 3 关联规则 $X\Rightarrow Y$。关联规则描述的是数据项集 X 中的项目出现时,数据项集 Y 中的项目也跟着出现的规律。$X\subset I, Y\subset I, X\cap Y=\varnothing$。

定义 4 支持度 $\sup(X\Rightarrow Y)$。

$$\sup(X\Rightarrow Y)=P(X\cup Y) \tag{1}$$

支持度反映了该规则 $X\Rightarrow Y$ 在 T 中所占的比例,说明了 $X\Rightarrow Y$ 在事务集 T 中出现的普遍程度。

定义 5 可信度 $\mathrm{con}(X\Rightarrow Y)$

$$\mathrm{con}(X\Rightarrow Y)=P(Y|X) \tag{2}$$

可信度 $\mathrm{con}(X\Rightarrow Y)$ 说明 $X\Rightarrow Y$ 成立的必然程度,表明由规则的前提导致结论的可信程度。

2 DGAA 算法在关联规则挖掘中的应用

本文设计的 DGAA 算法模型定义如下:

定义 6 DGAA 算法模型。$\mathrm{DGAA}=(En,F,S,C,M)$。其中:$En$ 表示染色体编码;F 代表适应度函数;选择算子、交叉算子和变异算子分别用 S、C 和 M 表示。

收稿日期: 2010-01-04; **修回日期**: 2010-06-14

作者简介:詹芹(1977-),女,湖北仙桃人,讲师,主要研究方向为数据挖掘(zhanqin2008@tom.com);张幼明(1976-),男,江西鄱阳人,讲师,硕士,主要研究方向为数字控制、嵌入式.

DGAA 算法的描述如下：

a)数据预处理。对原始数据进行处理，把数值型数据转换为由项集组成的事务数据库，并把关系数据库中的数值属性离散化，以便进行染色体编码。

DGAA 算法的染色体长度为 5，由四部分组成：第 1 位数字表示学期(1，2，…，8 分别代表大一上学期，大一下学期，…，大四下学期)；第 2 和 3 位数字一起表示课程代码(01 表示计算机导论，02 表示 C 语言，23 表示计算机系统结构)；第 4 位数字表示选修或重修(0 表示选修，1 表示重修)；第 5 位数字表示课程分数等级(1 表示等级为优秀，分数在 90～100；2 表示等级为良好，分数在 80～89；3 表示等级为中等，分数在 70～79；4 表示等级为及格，分数在 60～69；5 表示等级为不及格，分数在 60 分以下)。例如，染色体串 41304 表示的含义是大二下学期数字电子技术课程选修的分数在 60～69，等级为及格。

b)编码。根据染色体编码方法，把数据库中的每条记录一次性全部编码，作为初始种群。

c)评估。计算个体的支持度、可信度和适应度值。适应度函数 $F(X\Rightarrow Y)=(S(X\Rightarrow Y)+C(X\Rightarrow Y))/2$，$F(X\Rightarrow Y)$ 表明在各种规则的竞争中，只有支持度和可信度都高才有可能生存下来。

d)遗传操作。执行选择操作，保留支持度和可信度分别大于最小支持度和最小可信度的个体；然后按照传统遗传算法的方法执行两点交叉和高斯变异操作，最终产生频繁 k 项集的集合 L_k。

e)连接。如果两个项集 L_{k-1} 前面的 L_{k-2} 相同，而最后一项不同，则将这样的两个 L_{k-1} 进行连接后得到候选 k-项集的集合 C_k。

f)剪枝。对候选 k-项集 C_k 进行剪枝，从 C_k 中删除所有不包含 C_{k-1} 的事务，根据用户给定的最小支持度生成 L_{k+1}。

g)如果 L_k 为空，算法停止；否则，$k=k+1$，返回到 c)。

h)关联规则提取。从包含项数最多的频繁项集的集合 $L_{\max}$ 开始依次递减直到 L_2 为止执行循环操作。在每次循环中，对 $L_k(2\leqslant k\leqslant \max)$ 的每个元素 $l_k\sim L_{k-1}$ 中找子集 l_{k-1}，如果找到子集，并且 $\sup(l_k)/\sup(l_{k-1})\geqslant \text{mincon}$，则输出该规则。

3　实验

目前大多数高校的教学计划制定工作一般都是由一些富有多年本专业教学经验的专家来完成。这种主要凭借主观经验的制定方法因缺少客观事实根据多少会产生一些偏差。通过关联规则在学生成绩管理系统中的应用，可以挖掘出一些隐含其中的课程间相关联系，为高校的教学计划的编排及修订提供参考[12,13]。这里以某高校计算机科学与技术专业的学生成绩管理数据库的数据为例，挖掘课程之间的相关联系。数据项集 I 是由学生的 32 门课程成绩组成的集合。事务数据集 T 为学生成绩库记录的集合，其中每一个记录 T_k 是 I 中一组属性的集合。为了方便实验，只考虑学生成绩库中课程的成绩，其余的属性忽略。

在应用算法挖掘关联规则之前，根据第 2 章描述的编码方法对原始数据进行处理，把数值型数据转换为由项集组成的事务数据库。实验中用到的主要参数设置如下：变异概率 $P_m=0.15$，交叉概率 $P_c=0.8$，最小支持度 minsup = 0.7，最小可信度 mincon = 0.1。表 1 为应用 DGAA 算法生成的部分关联规则。

表 1　DGAA 算法生成的部分关联规则

编号	规则	可信度	支持度	含义
1	20203⇒51002	10	78	如果大一下学期 C 语言课程选修成绩为中等，那么大三上学期 C++课程选修成绩很可能为良好
2	20212⇒51004	8.6	82	如果大一下学期 C 语言课程重修成绩为良好，那么大三上学期 C++课程选修成绩很可能为及格
3	20203⇒31002	6.8	86	如果大一下学期 C 语言课程选修成绩为中等，那么大二上学期 C++课程选修成绩很可能为良好
4	31203⇒41302	14	92	如果大二上学期模拟电子技术课程选修成绩为中等，那么大二下学期数字电子技术课程成绩很可能为良好
5	41304⇒51403	8	88	如果大二下学期数字电子技术课程选修成绩为中等，那么大三上学期计算机组成原理课程成绩很可能为中等

为了测试本文提出的 DGAA 算法性能，将 DGAA 与经典的 Apriori 算法和文献[6]的 GA 算法进行了对比实验。图 1～3 分别显示的是三种算法在关联规则数目、精确率和查全率方面的比较情况。从表 1 的实验结果可以看出，应用本文设计的 DGAA 挖掘算法，能有效地挖掘课程之间的关联关系，对高校的课程内容设置和时间先后安排具有一定的指导意义。图 1～3的实验结果反映出 DGAA 算法的性能明显优于 Apriori 和 GA 算法，将 GA 思想融入到 GA 算法中，显著提高了算法的运行效率，大大缩短了扫描执行时间。

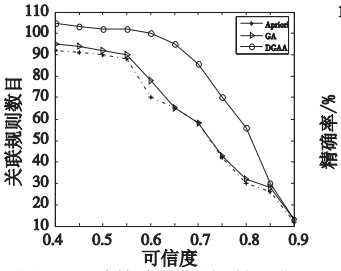

图1　三种算法挖掘关联规则数目的比较

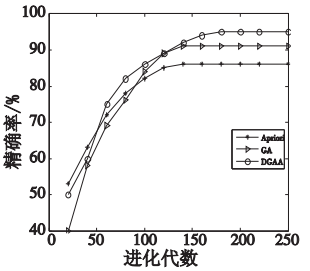

图2 三种算法的精确率比较

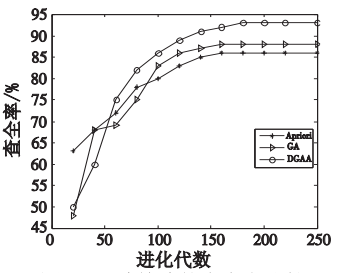

图3　三种算法的查全率比较

4　结束语

遗传算法在解决大空间、多峰值、非线性、全局优化等复杂度较高的问题时具有独特的优势。本文利用动态遗传 Apriori 挖掘算法构造了数据挖掘中关联规则挖掘的模型，通过支持度和可信度的对比，发现隐藏在数据库中的规律。通过实验显示该算法可以有效、快速地解决关联规则挖掘问题。

参考文献：

[1] 惠晓滨，张凤鸣，虞健飞. 一种基于栈变换的高效关联规则算法[J]. 计算机研究与发展，2003，40(2)：330-335.

[2] EKELIN C，JONSSON J. Solving embedded system scheduling problem using constraint programming[J]. IEEE Real-Time Systems Symposium，2000，24(2)：27-30.

（下转第 2935 页）

a)初始化算法的相关参数,如设置最大进化代数 N_{max},种群规模 m_0等。

b)根据 3.1 节中的方法,对 JSS 问题进行编码,并构建初始种群。

c)根据 3.3 节中的方法,对上述群体进行分裂、水平选择操作。

d)根据 3.4 节中的方法,对上述群体进行变异和垂直选择操作。

e)根据 3.5 节中的方法,对上述群体进行交叉操作。

f)若 $N < N_{max}$的,则 $N = N + 1$,转步骤 c);若 $N = N_{max}$则终止算法。

g)输出最终解。

在 HDEA 的每次迭代中,个体的进化操作包括分裂、水平选择、变异、垂直选择和交叉五个环节,选择算子中个体适应度函数计算公式见 3.2 节。DEA 的优势在于个体自身的进化,使得个体在解空间的局部区域内进行深度搜索,而遗传算法的交叉算子,使个体能够跳出局部极值,有效地解决了早熟收敛的问题。

4　案例分析

本文给出一个 3×3 的模糊 JSS 问题的例子,每个工序在相应机器上的加工时间如表 1 所示。HDEA 的参数设置为:$m_0 = 50$,$N_{max} = 300$,$p_c = 0.5$,采用 VC#2005 编程,运行 10 次,在 190 代内都可以收敛并得到最优目标值,其中 7 次得到最优值。最小完工时间为(16,18,22,24),最优 $E[C_{max}] = 20.0$。在机器 M_1上,各操作的执行顺序为 $O_{12}\rightarrow O_{22}\rightarrow O_{21}$;在机器 M_2上,操作执行顺序为 $O_{11}\rightarrow O_{23}$;在机器 M_3上,操作执行顺序为 $O_{13}\rightarrow O_{32}\rightarrow O_{31}$。

表 1　3×3 JSS 问题的模糊加工时间表

		M_1	M_2	M_3
J_1	O_{11}	(1.6,1.8,2.2,2.4)	(4.8,5.4,6.6,7.2)	(11.2,12,15.4,16.8)
	O_{21}	(4.8,5.4,6.6,7.2)	(3.2,3.6,4.4,4.8)	(9.6,10.8,13.2,14.4)
	O_{31}	(9.6,10.8,13.2,14.4)	(1.6,1.8,2.2,2.4)	(6.4,7.2,8.8,9.6)
J_2	O_{12}	(1.6,1.8,2.2,2.4)	(8.0,9.0,11.0,12.0)	(4.8,5.4,6.6,7.2)
	O_{22}	(3.2,3.6,4.4,4.8)	(6.4,7.2,8.8,9.6)	(14.4,16.2,19.8,21.6)
	O_{32}	(14.4,16.2,19.8,21.6)	(1.6,1.8,2.2,2.4)	(3.2,3.6,4.4,4.8)
J_3	O_{13}	(1.6,1.8,2.2,2.4)	(14.4,16.2,19.8,21.6)	(4.8,5.4,6.6,7.2)
	O_{23}	(8.0,9.0,11.0,12.0)	(3.2,3.6,4.4,4.8)	(6.4,7.2,8.8,9.6)

本文还选择了具有更大规模的 8×8 的 JSS 实例[1],将其中的操作时间 t 由确定值调整为梯形模糊数($\delta\times t$,$\theta\times t$,$\psi\times t$,$\varphi\times t$),令 $\delta = 0.85$,$\theta = 0.95$,$\psi = 1.05$,$\varphi = 1.15$,从而构造出 8×8模糊 JSS 问题。分别采用 HDEA 和 GA 进行求解,单点交叉概率 0.85,换位变异概率 0.09。为便于比较,设置 HDEA 与 GA 种群规模和最大进化代数相同,分别运行 20 次。实验表明,与标准 GA 相比,DEA-GA 具有较快的收敛速度,可以在 290 代内获得具有理想制造跨度的调度方案,同时不存在 GA 容易早熟的问题,如表 2 所示。

表 2　HDEA、GA 求解 8×8 JSS 问题性能比较

算法	最差 $E[C_{max}]$	最优 C_{max}	最优 $E[C_{max}]$	最大收敛代数	最优值比率
HDEA	27.3	(18.7,20.9,23.2,25.3)	22.1	290	68.3%
GA	28.1	(18.6,21.3,23.7,25.5)	23.2	310	57.1%

5　结束语

本文对具有模糊加工时间的 JSS 问题进行了研究,用梯形模糊数来表征时间参数,并给出了相应的目标函数。在 JSS 问题求解方面,本文采用集成 GA 和 DEA 的混合进化算法,取得了较理想的结果。DEA 是单亲进化算法,其分裂算子和变异算子使个体的进化过程具备良好的连续性,适合局部搜索,而 GA 是基于种群的优化算法。本文的 HDEA 算法将单亲繁殖和种群交叉的优势结合起来,有效提高了算法的性能。仿真实验表明,与遗传算法相比,HDEA 有更好的全局性和鲁棒性,尤其在求解较大规模问题时,HDEA 收敛性能的优势更为明显。笔者还将对此作进一步的研究。

参考文献:

[1] CACEM I, HAMMADI S. Approach by localization and multiobjective evolutionary optimization for flexible job-shop scheduling problems [J]. IEEE Trans on Systems, Man, and Cybernetics, Part C: Applications and Reviews, 2002, 32(1): 1-13.

[2] CHEN Hao-xun, LHLOW J, LEHMANN C. A genetic algorithm for flexible job shop scheduling[C]//International Conference on Robitics & Automation. 1999.

[3] 余文,李人厚. 一种有效的双向进化算法[J]. 小型微型计算机系统,2003,24(3):527-530.

[4] 牛群,顾幸生. 基于 DNA 进化算法的 Flow shop 生产调度问题[J]. 上海大学学报,2004,10(S):88-92.

[5] 牛群,顾幸生. 基于 DNA 进化算法的车间作业调度问题研究[J]. 控制与决策,2005,20(10):1157-1160.

[6] LIU Bao-ding, LIU Yian-kui. Expected value of fuzzy variable and fuzzy expected value model[J]. IEEE Trans on Fuzzy Systems, 2002,10(4):445-450.

(上接第 2930 页)

[3] PARK J S, CHEN M S. An effective hash based algorithm for mining association rules[C]//Proc of International Conference on Management of Data. New York: ACM Press, 1995:175-186.

[4] STAKOVIE J. Misconceptions about real-time computing: a serious problem for next generation system [J]. IEEE Computer, 1998, 21(10):10-19.

[5] BRIN S, MOTWANI R. Dynamic item set counting and implication rules for market basked data[C]//Proc of ACM SIGMOD International Conference on Management of Data. New York: ACM Press, 1997:255-264.

[6] 赵方方,刘万军,陈芳元. 遗传算法在关联规则挖掘中的应用研究[J]. 沈阳理工大学学报,2006,25(4):51-54.

[7] 曾万聃,周绪波,戴勃. 关联规则挖掘的矩阵算法[J]. 计算机工程,2006,32(2):45-47.

[8] 彭仪普,熊拥军. 关联规则挖掘 AorionTrid 算法优化研究[J]. 计算机工程,2006,32(5):55-57.

[9] 刘勇国,李学明,张伟. 基于遗传算法的特征子集选择[J]. 计算机工程,2003,29(6):67-70.

[10] 许国艳,史宇清. 遗传算法在关联规则挖掘中的应用[J]. 计算机工程,2002,23(7):122-124.

[11] 张宗平. 一种更新关联规则的方法[J]. 计算机工程, 2008,34(1): 64-65, 68.

[12] 陈熔. 数据挖掘技术在课程相关性中的应用研究[J]. 西昌学院学报:自然科学版,2007,21(2):67-69.

[13] 刘红梅. 关联规则在学生成绩分析中的应用[J]. 长江大学学报:自然科学版,2008,12(5):357-359.