一种基于改进型遗传算法的关联规则提取算法及其应用*

王礼刚[1]，左源瑞[1]，李盛瑜[2]

（1. 西南大学 计算机与信息科学学院，重庆 400715；2. 重庆工商大学 计算机科学与信息工程学院，重庆 400067）

摘　要：对关联规则的数据挖掘和遗传算法进行了概述，阐述了关联规则数据挖掘的现实意义，提出了一种采用改进型遗传算法的关联规则提取方法，并给出了具体的算法，最后结合一个具体实例进行了应用。

关键词：数据挖掘；关联规则；遗传算法

中图分类号：TP181；TP301.6　　文献标识码：A　　文章编号：1672-6693(2006)02-0042-04

An Algorithm for Mining Association Rules Based on Improved Genetic Algorithm and its Application

WANG Li-gang[1]，*ZUO Yuan-rui*[1]，*LI Sheng-yu*[2]

(1. College of Computer and Information Science, Southwest Normal University, Chongqing 400715;
2. College of Computer Science and Information Engineering,
Chongqing Technology and Business University, Chongqing 400067, China)

Abstract: This paper summarized date mining of association rules and genetic algorithm, clarified the realistic significance of data mining of association rules, and put forward an algorithm for mining association rules based on improved genetic algorithm. Finally a case is provided for applying this algorithm.

Key words: data mining; association rule; genetic algorithm

数据挖掘是20世纪90年代中期兴起的一项新技术，它是知识发现(KDD)过程中的关键步骤。对信息社会中数据和数据库的爆炸式增长，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。当前，很多成功的企业正在应用数据挖掘进行关联规则的提取来帮助它们更好地制定决策，利用功能强大的数据挖掘技术，可以把数据转化为有用的信息以帮助制定决策，从而在市场竞争中获得优势地位。

关联规则的挖掘是数据挖掘领域的一个重要内容，它代表数据库中一组对象之间某种关联关系的规则。最早提出的Apriori算法可以说是关联规则挖掘的经典算法，但是这个算法仍然存在很多弊端，本文提出了一种基于改进型遗传算法的关联规则提取算法，大大提高了算法的效率。

1　关联规则的数据挖掘

1.1　数据挖掘的定义及相关技术

所谓数据挖掘(Data Mining，简称DM)，就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的过程[1]。也有一些文献把数据挖掘称为知识抽取(Knowledge Extraction)、数据考古学(Data Archaeology)、数据捕捞(Data Dredging)等等[2]。数据挖掘目前是人工智能(AI)领域的一大研究热点，它也是一个新兴的边沿学科，汇集了来自机器学习、数据库、模式识别、统计学、人工智能、管理信息系统等多个学科的研究成果，吸引了各方面的专家学者投身此领域的研究和开发工作，而且被许多工商界

* 收稿日期：2006-01-16

作者简介：王礼刚(1976-)，男，重庆铜梁人，硕士研究生，研究方向为人工智能、数据库。

人士看作是能带来巨大回报的重要工具[3]。

数据挖掘常用的技术有关联规则、决策树、粗糙集、神经网络、遗传算法、概率统计、信息论、聚类规则及各种技术的融合等。本文主要研究改进型遗传算法在关联规则数据挖掘中的应用,提出了一种基于改进型遗传算法的关联规则提取算法。

1.2 关联规则的形式化定义

设 T 是事务数据,即 $T = \{i_1, i_2, i_3, \cdots, i_m\}$,其中 $i_i(1 \leqslant i \leqslant r_n)$ 是每个事务的数据,这些数据称为数据项。I 是 T 中所有数据项(物品)的集合,即 $I = \{i_1, i_2, i_3, \cdots, i_n\}$,$i_j(1 \leqslant j \leqslant n)$ 是 T 中的一个数据项。每个事务中含有 I 的一个子集。关联规则是一种蕴含关系 $X \Rightarrow Y$,其中 $X \subset I, Y \subset I, X \cap Y = \varnothing$,$X$ 称作是前提,Y 称作是结果。有两个因子与这条规则有关 如果事务数据库中有 $s\%$ 的事务包含 $X \cup Y$,那么就说关联规则 $X \Rightarrow Y$ 的支持度(support)为 s。如果事务数据库里包含 X 的事务中有 $c\%$ 的事务同时也包含 Y,那么就说关联规则 $X \Rightarrow Y$ 的可信度(Confidence)为 c。其中支持度表示 $X \Rightarrow Y$ 在 T 事务数据中出现的普遍程度,可信度说明 $X \Rightarrow Y$ 成立的必然程度,如果支持度和可信度都超过各自的阈值,则 $X \Rightarrow Y$ 可以看成是 T 中的一个有意义的关联规则[4,5]。

关联规则的挖掘是寻找在同一事件中出现的不同项的相关性,通过它可以发现大量数据中数据项集之间存在的关联,更普遍地,关联规则可表示为

$$A_1 \wedge A_2 \wedge A_3 \wedge \cdots \wedge A_n \Rightarrow B_1 \wedge B_2 \wedge \cdots \wedge B_n$$

1.3 关联规则的挖掘算法

关联规则挖掘问题可以划分成两个子问题[6]。

(1)发现所有支持度不少于最小支持度的项目集,即频集,也称为大项集,这个子问题是最重要的,开销最大,因此各种算法主要致力于提高发现频集的效率。

(2)使用频集来产生关联规则。对于每一个频集 L,发现它所有的非空子集,对于任何一个非空子集 X,可能存在关联规则 $X \Rightarrow (L - X)$,计算这个关联规则的可信度 $support(L)/support(X)$,如果最小可信度 $\leqslant support(L)/support(X)$,那么输出关联规则 $X \Rightarrow (L - X)$。

关联规则的挖掘是数据挖掘的重要内容之一,关联规则的挖掘国际上目前已有较多的研究,如国际上较有影响的挖掘方法 Apriori[7] 及其相关改进算法,FP-Growth[8] 算法等。最早提出的关联规则挖掘算法是 Apriori 算法。但是 Apriori 算法需要多次遍历数据库,I/O 开销大。

2 基于遗传算法的关联规则挖掘

遗传算法(Genetic Algorithms,简称 GA)是一种借鉴生物界自然选择和自然遗传机制,模拟自然进化过程搜索最优解的方法,通过自然选择、遗传、变异等作用机制,实现各个体适应性的提高,由于其解决问题以混沌、随机和非线性为典型特征,因而为其它科学技术无法解决或难以解决的问题提供了新的计算模型[9]。习惯上将 J. H. Holland 提出的遗传算法称为简单遗传算法(Simple Genetic Algorithm,简称 SGA)。通常,简单 GA 算法中的交叉操作是随机取两个染色体进行交叉操作(单点交叉、多点交叉、树交叉等),即在以高适应度模式的祖先的“家族”中取一点,可见这种取法存在一定的片面性,并且简单遗传算法在任何情况下都是收敛的,即不能搜索到全局最优解[10],而改进的遗传算法则能保证全局最优解。本文解决问题所采用的是在简单遗算法基础上的改进型遗传算法。

2.1 改进型遗传算子

针对关联规则挖掘的特殊性以及传统的基本遗传算法的缺陷,本文对基本遗传算法进行了改进,结合自适应调整遗传算法的控制参数的思想,提出了一种改进的自适应遗传算法。该算法不权能加快遗传进化速度,而且还能增强算法的全局收敛性能,从而得到满意的全局最优值。本文对基本遗传算法的主要改进如下。

(1)选择算子。简单遗传算法采用赌轮方式选择交配组,即根据个体的适应度与平均适应度的比例来确定该个体的复制比例。本文用一种基于种群的按个体适应度大小排序的选择算法来代替赌轮选择方法。其算法描述如下:

```
fitsort()      /* 定义一个函数 fitsort() */
    {
    将种群中的个体按适应度大小进行排序;
    }
  do while 种群还没有扫描完
    {
     排在前面的个体复制两份;
     中间的复制一份;
     后面的不复制;
```

}

(2)交叉算子。在随机选择出父本和母本以后,按照交叉方法(单点,多点)进行 n 次交叉,产生 $2n$ 个个体,再从这 $2n$ 个个体中挑选出最优的两个个体加入新的种群中。这样既保存了父本和母本的基因,又在进化的过程中大大地提高了种群中个体的平均性能。

(3)变异算子。不采用固定的变异概率 P_m,这里采用一种可变变异概率的方法,这一方法的算法描述为

if(个体的适应度 < 平均适应度) then P_m 取值很小或为零;

else P_m 取值相对很大

这样就使得种群中好的基因不被破坏,既有利于不良基因的去除,又有利于新基因的引入,从而可以很大程度地提高遗传算法的性能。

2.2 关联规则的遗传算法编码改进

在本文中,结合遗传算子、关联规则挖掘的需要,采用了实数数组的编码方法,这种编码方法具有精度高、便于空间搜索的优点,实现起来也比较简便。在利用实数数组的方法进行编码过程中,实数数组的元素个数与事务数据库中的字段的个数相对应,实数数组的元素值则代表了字段的属性值。本文用一个元组为 N 的数组来表示如上所示的事务数据库的个体编码,$A[1]$ 表示字段 1,$A[2]$ 表示字段 2,……,$A[N]$ 表示字段 N;将属性值用数值型的值表示,例如用数值 1 表示属性值 1,数值 2 表示属性值 2,……,数值 E 表示属性值 E,这样就可以用数组 $A[N]$ 的元素值来表示相对应的字段的属性值了。另外再用 0 值表示此属性与其它的属性无关联。

2.3 适应度函数的构造

采用关联规则的支持度来定义它的适应度函数。可以这样来筛选规则,先用支持度来筛选规则,然后在满足最小支持度的规则中确定它的关联程度和关联性。因此规则的适应度可以简便地定义为

$$fitnes(R_i) = \frac{S^*}{S} = \begin{cases} p & \text{当 } S^* > S \\ q & \text{当 } S^* < S \end{cases}$$

式中 S^* 为经过遗传操作所形成的一条新规则的支持度,S 为用户给定的支持度的阈值。当 R_i 为符合要求的规则时,它的适应度函数值应大于 1,否则适应度函数值将小于 1,这条规则在下一代遗传中就会被淘汰。

2.4 基于改进型遗传算法的关联规则挖掘算法

Step1 随机生成一个初始种群 $P = \{A_1, A_2, \cdots, A_n\}$(即种群中有 N 个个体);

Step2 在个体种群中,对所有的个体按其适应度大小进行排序,然后计算个体的支持度 S 和可信度 C;

Step3 选择。根据适应度的大小按一定比例进行个体复制(排在前面的个体复制两份,中间的复制一份,后面的不复制),并计算保留下来的个体数 M;

Step4 如果 $M < N$,则随机生成($N - M$)个个体,否则跳过 Step4;

Step5 变异。根据个体的适应度与平均适应度的比较,确定个体变异的概率;

Step6 从复制组中随机选择两个个体,对这两个个体进行多次交叉,从所得的结果中选择一个最优个体存入新种群;

Step7 若满足结束条件,则停止,否则跳转到 Step2,直至找到所有符合条件的规则;

Step8 进行规则的提取。

3 应用实例——智能型教学评测管理系统

首先,根据学院 2005 级学生的资料建立了一个数据库(包括以下信息:姓名、学号、学期、性别、籍贯、民族、入学类别、学生干部、英语成绩(四或六级)、政治面貌、出生年月、学习成绩、奖惩情况),并将相应的字符型数据转换成数值型数据。根据前面提出的算法,在这个学生资料数据库中发现部分关联规则如下:

(1) <4200> ⇒ <0001> (5% support,99% confidence),即 <民族:汉,入学类别:统招> ⇒ <英语成绩:四级>。这条规则意味着汉族的统招生英语四级大部分都过了;

(2) <4> ⇒ <200001> (55% support,90% confidence),即 <民族:汉> ⇒ <入学类别:统招,英语成绩:四级>。这条规则意味着汉族的学生大部分为统招生,通过了英语四级,粗看起来这条规则和上一条规则差不多,但是实际上,它们的可信度是不同的;

(3) <001> ⇒ <4000> (6% support,89% confidence),即 <学生干部:学习委员> ⇒ <学习成绩:80~89>。这条规则意味着担任学习委员的学生,他们的学习成绩在 80 分到 89 分之间的可信度为 89%,支持度为 6%,也就是说,大部分学习委员的

学习成绩为良好；

(4) < 0004 > ⇒ < 100 > (6% support, 100% cinfidence),即 < 学习成绩:80 ~ 89 > < 奖惩:奖励 >。这条规则意味着学习成绩为 80 ~ 89 分的学生百分之百都会获得获励；

……

通过获取以上信息,在学院的教学管理工作方面可以从中得到一些启示,并针对一些问题作出改进,比如在以后的教学工作中加强对少数民族学生的教学的针对性等。

还可以通过类似的方法,对不同学生的不同的资料信息建立相应的数据库,并在此基础上进行关联规则的挖掘,发现有用的知识,并把这些知识应用到学生的教学管理上去,针对出现的不同问题,对原来的教学制度和计划做出相应的调整,以更加有益于学生的培养和教育。这些说明本文所提出的内容具有一定的应用价值。

4 结束语

遗传算法在数据挖掘技术中占有很重要的地位,这是由它本身的特点和优点所决定的。遗传算法具有十分顽强的鲁棒性,其在解决大空间、多峰值、非线性、全局优化等复杂度高的问题时具有独特的优势。遗传算法可以单独用于数据仓库中关联规则的挖掘,还可以和其它的数据挖掘技术相结合,本文将改进后的遗传算法应用到关联规则的数据挖掘中去,并结合实例提出了一种基于这种改进型遗传算法的关联规则提取算法。基于遗传算法的关联规则挖掘技术还可以应用在销售分析、金融信贷风险分析、物流货源分析等其他领域,具有较好的研究和应用价值。

参考文献：

[1] FRAWLEY W J,PIATETSKY S G,MATHEUS C G. Knowledge Discovery in Database:An Overview[A]. PIATETSKY S G,FRAWLEY W J. Knowledge Discovey in Database[C]. Massachusetts:AAA/MIT Press,1991. 1-27.

[2] HAN J. Data Mining Techniques[R]. Canada:Simon Fraser University,1996.

[3] HAND D,MANNILA H,SMYTH P. 数据挖掘原理[M]. 张银奎,廖丽 译. 北京:机械工业出版社,2003.

[4] 张云涛,龚玲. 数据挖掘原理与技术[M]. 北京:电子工业出版社,2004.

[5] 刘同明. 数据挖掘技术及其应用[M]. 北京:国防工业出版社,2001.

[6] AGRAWAL R,IMIEINSKIT,SWAMI A. Mining Association Rules Between Set of Item in Large Databases[A]. In SIGMOD-93[C]. Washington:DC,1993. 207-216

[7] AGRAWAL R,SRIKANT R. Fast Algorithms for Mining Association Rules[A]. BOCCA J B,JARKE M ,ZANIOLO C. VLDB94[C]. Chile:Morgan Kaufmann,1994. 487-499.

[8] HAN J,PEI J,YIN Y. Mining Frequent Patterns Without Candidate Generation[A]. In SIGMOD00[C]. Dallas:TX,2000. 1-12.

[9] 陈国良,王煦法,庄镇泉,等. 遗传算法及其应用[M]. 北京:人民邮电出版社,1996.

[10] 杨大地,张春涛. 均匀两点交叉遗传算法[J]. 重庆师范大学学报(自然科学版),2004,21(1):1-3.

(责任编辑　游中胜)