基于遗传算法的关联规则挖掘

朱彦廷
(广西现代职业技术学院 计算机系,广西 河池 547000)

【摘 要】根据关联规则挖掘的要求,结合遗传算法的特点,提出了一种基于遗传算法的关联规则挖掘算法,在基本遗传操作选择、交叉、变异的基础上,引入了挑选操作,取消了交叉、变异概率,给出了详细的算法设计及描述,并通过实例证明了算法的性能。

【关键词】数据挖掘;关联规则;遗传算法
【中图分类号】TP311.13 【文献标识码】A 【文章编号】1673-1891(2010)03-0060-03

1 引言

数据挖掘(Data Mining)就是从大量的、不完全的、有噪声的数据中提取隐含的、有用的信息的过程。关联规则挖掘是其中的一个重要方面,是指发现数据库中项集之间的关联,在决策支持、医疗诊断、市场策略、优化调度等领域有广泛的应用价值。

在现有的关联规则挖掘的算法中,Agrawal等人提出的Apriori算法最为流行,很多算法都是对它的改进,如Savasere等人提出的基于划分的算法,Park等人提出的基于杂凑的算法。虽然进行了改进,但是Apriori算法计算量大,处理复杂数据库效率低的缺陷还是难以克服。在信息数量大、变化快的今天,人们迫切需要一种快速、有效的算法。

遗传算法(Genetic Algorithm)是美国Michigan大学教授J.Holland在1975年提出的,它是一种模拟生物进化过程的算法,与传统算法相比有高鲁棒性、全局搜索性、内在并行性等优点,为许多以前无法解决或难以解决的复杂问题提供了新的计算方法。遗传算法是一种基于群体的算法,能快速地进行搜索,适用于处理复杂数据库,还能避免停滞在局部最优解上,有望发现真正有用的规则。

基于遗传算法的关联规则挖掘是针对一个具体问题,随机产生一组规则,然后利用生物进化的原理进行优化,最后得到较为满意的规则,从而挖掘出隐含的知识。

2 关联规则的基本概念

关联规则是形如A⇒B的蕴含式,通常前件是合取式形式,每个项是一个特征属性的一个值,后件是类别属性的一个值。

衡量一条规则的好坏主要有3个指标:可信度、支持度、覆盖度。

设D是一个记录集,|D|为D的基数,r是一条规则,D_A是D中与规则r前件匹配的子集,D_B是与r后件匹配的子集,$D_A \cap D_B$是与r前件和后件均匹配的子集,可信度

$C(r)=|D_A \cap D_B|/|D_A|$,表示A⇒B的准确性。

支持度

$S(r)=|D_A|/|D|$(或$|D_A \cap D_B|/|D|$),表示规则前件(或规则)的普遍性。

覆盖度

$A(r)=|D_A \cap D_B)|/|D_B|$,表示B⇒A的准确性,和可信度侧重点相反,使用很少。

3 算法设计

3.1 编码

遗传算法的运算对象是表示个体的符号串。一般采用二进制编码,假定属性的取值是离散的(如果是连续的,可以转化为离散的),根据属性的取值范围,用一个适当长度的二进制串来表示,将所有属性的二进制串连接在一起,组成一个个体,表示一个关联规则。如研究氮肥浓度、磷肥浓度、海水密度对紫菜幼苗生长的影响,属性的取值及编码如表1所示,规定属性的排列顺序为:氮肥浓度、磷肥浓度、海水密度、(幼苗)长度,则个体00101000表示的意义是:氮肥浓度:3ppm,磷肥浓度:2ppm,海水密度:1.0239g/cm^3⇒长度:11.1~12.0cm。如果某特征属性编码全为*,表示该属性无论取何值,不影响规则的成立与否。表示特征属性的编码不能全为*,它表示的规则没有意义。

在随机产生个体时,应按属性产生,以氮肥浓度为例,随机产生一个0~4间的整数,如果为0,个体的0、1位是00;为1,个体的0、1位是01;为2,个体的0、1位是10;为3,个体的0、1位是11;为4,个体的0、1位是**。最后还要检查个体表示特征属性的编码是否全为*,如果全为*,重新产生个体。

3.2 适应度函数

遗传算法在进化过程中只利用种群中每个个

收稿日期:2010-06-27
作者简介:朱彦廷(1976-),男,辽宁建昌人,硕士,讲师,研究方向为遗传算法等。

体的适应度来进行搜索,因此适应度函数的设计至关重要。本文定义适应度函数F(r)=aS(r)+bC(r)+cA(r),这样比较全面,其中,a,b,c为常数且0≤a,b,c≤1,由用户根据需要设置,注意通常支持度(几乎不可能接近1)的均值比可信度(或覆盖度,可能为1)小些,如果a、b(或c)相等,实际可信度(或覆盖度)偏重。

表1　编码

属性	取值	编码	属性	取值	编码
氮肥浓度	3	00	海水密度(g/cm^3)	1.0229	00
(ppm)	5	01	长度(cm)	1.0234	01
	7	10		1.0239	10
	9	11		1.0244	11
磷肥浓度	0	00		11.1~12.0	00
(ppm)	1	01		12.1~13.0	01
	2	10		13.1~14.0	10
	3	11		14.1~15.0	11

3.3 遗传操作

3.3.1 选择

采用比例选择,用个体的适应度与群体中个体的适应度的总和的比值作为其被选择的概率。个体的适应度越高,被选择的概率越大。

3.3.2 交叉

采用单点交叉。本例要保持表示属性的二进制串完整,交叉起始位不是任何一位都可以,只能是0、2、4、6。先随机产生一个0~3间的整数,如果为0,交叉位是0,为1,交叉位是2,为2,交叉位是4,为3,交叉位是6。然后交换2个个体从交叉起始位开始的串,形成2个新个体,例如个体是1111111和0000000,交叉位是2,则交叉后的新个体为11000000和00111111。最后还要检查新个体表示特征属性的编码是否全为*,如果全为*(如11****00和**111100交叉,交叉位是0),随机产生1个个体,这样可以增强群体的多样性。

3.3.3 变异

采用基本位变异,随机产生变异位,然后对变异位作翻转操作,形成1个新个体。例如个体是11111111,变异位是3,则变异后的新个体为11011111。如果变异位编码为*,随机产生1个个体。

3.3.4 挑选

这是算法特有的也是最复杂的操作。

交叉、变异产生的新个体有可能不如原个体,为了避免丢失较好的个体,文献[1]保存原群体中较好的个体,用来替换新群体中较差的个体,但这仍有可能丢失较好的个体,因为可能新群体中较差的个体的适应度高于原群体中较好的个体的适应度,可是却被替换;也可能新群体中较好的个体的适应度低于原群体中较差的个体的适应度,可是却被保留。

适应度很高的个体很可能被选择多次,相同的这样个体交叉,得到同样的个体,经过若干代占据群体的大部分甚至全部,致使进化陷于停滞,因此文献[1]在得到新群体后,删除重复的个体,随机生成同样数量的个体补充,但这仍可能引入重复的个体,因为关联规则挖掘最后输出的是一个规则的集合,最后输出的实际规则数量将减少。

本文对于要计算适应度的新个体,先和原群体中的个体以及已计算适应度的新个体比较,如果重复,直接淘汰,如果不重复,计算适应度,最后从原群体和新群体中选出适应度最高的个体组成新群体。这将减少扫描数据库次数,避免丢失较好的个体,保证新群体中没有重复的个体。

由于从原群体和新群体中选出适应度最高的个体组成新群体,如果交叉、变异产生的新个体不如原个体,不会有什么不利影响,因此可以放心交叉、变异,无须设置交叉、变异概率(相当于均为1。本来这2个参数设置得过大、过小都将影响算法的性能[2],并且最好随着进化的进行而变化[3])。

3.4 算法描述

算法流程如下:

(1)初始化:输入记录集U,设置适应度函数的系数a、b、c、群体大小M、终止代数T,进化代数计数器t=0,随机生成M个个体作为初始群体;

(2)个体评价:扫描U,计算各个个体的适应度;

(3)选择;

(4)交叉;

(5)变异算;

(6)挑选,得到下一代群体;

(7)终止条件判断:如果t<T,则t=t+1,转到(2),如果t=T,则结束运算;

(8)设置可信度、支持度阈值,输出大于阈值的规则。

个体之间可能相互包含(如:00****00和0010**00)、矛盾(如:00****00和00****01)。文献[4]通过处理使群体中的个体不相互包含、矛盾,本文认为不做处理,由用户取舍更好,因为:对于相互包含的个体,00****00的支持度一定比0010**00高,但可信度则不一定。如果记录集如表2所示,那么00****00的可信度是100%,0010**00的可信

度是50%。如果记录集如表3所示，那么00****00的可信度是67%，0010**00的可信度是100%。因此，00****00不一定比0010**00更有价值；对于相互矛盾的个体，假设00****00的可信度为0.5，00****01的可信度为0.45，那么00****01也是有价值的。

表2 紫菜幼苗培育试验数据

序号	氮肥浓度（ppm）	磷肥浓度（ppm）	海水密度（g/cm³）	长度（cm）
1	3	2	1.0229	11.1
2	3	3	1.0229	11.9

表3 紫菜幼苗培育试验数据

序号	氮肥浓度（ppm）	磷肥浓度（ppm）	海水密度（g/cm³）	长度（cm）
1	3	2	1.0229	11.1
2	3	2	1.0229	11.9
3	3	3	1.0229	12.8

文献[1]预先设置支持度阈值、可信度阈值，以控制输出的规则数量，便于分析，本文认为阈值设置得较大，输出的规则过少甚至为零，设置得较小，输出的规则过多，改为最后设置支持度阈值、可信度阈值，可根据输出进行调整。

4 算法性能

关联规则挖掘最后输出的不是一个最优解，而是一个规则的集合，追求的是挖掘出适应度高、数量多的规则，因此衡量算法性能的一个重要指标是群体的平均适应度。采用紫菜幼苗培育试验数据，对比文献[1]算法和本文算法，参数设置如下：a=1，b=0.2，c=0，M=32，T=20，对于前者，交叉概率Pc=0.95，变异概率Pm=0.01，保存群体中较好个体的数量为8，运行程序，如图1所示，其中，f(t)是t代群体的平均适应度，最后得到的群体的平均适应度前者是0.218，后者是0.366。

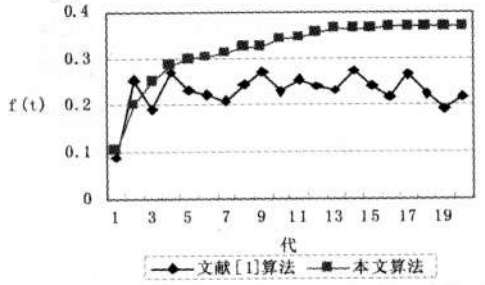

图1 平均适应度

前者只设置1次支持度阈值、可信度阈值；后者可设置多次，如设置支持度阈值为0.3，可信度阈值为0.25，输出8条规则，再设置支持度阈值为0.3，可信度阈值为0.3，输出4条规则，因数量较少，列举如下：

表4 扫描数据库次数

代数	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
文献[1]算法	32	32	32	32	32	32	32	32	32	32	32	32	32	32	32	32	32	32	32	32
本文算法	32	27	23	25	23	26	26	25	20	31	23	26	20	16	24	22	20	22	20	13

1101，S(r)=0.380，C(r)=0.421，F(r)=0.464

00****00，S(r)=0.360，C(r)=0.444，F(r)=0.449

01****01，S(r)=0.320，C(r)=0.438，F(r)=0.407

****0001，S(r)=0.300，C(r)=0.333，F(r)=0.367

扫描数据库次数如表4所示，文献[1]算法总计640次，本文算法总计464次，减少了28%。

这方面的算法不少，文献[1]算法是其中一个优秀的算法，和它相比，本文算法群体的平均适应度、稳定性都有较大提高，扫描数据库次数有较大减少（这意味着执行时间的减少），另外也比较简单，便于编程实现，其中一些做法（如从原群体和新群体中选出适应度最高的个体组成新群体，因此无须设置交叉、变异概率）也可以应用到一般遗传算法中去。

注释及参考文献：

[1]蒋志全，陈燕.基于遗传算法的关联规则挖掘模型[J].大连海事大学学报，2003，29(3)：97-100.

[2]施建强，刘晓平.基于遗传算法的数据挖掘技术的研究[J].电脑与信息技术，2003(1)：13.

[3]陈曦，林涛，唐贤瑛.遗传算法的参数设计与性能研究[J].计算机工程与设计，2004，25(8)：1310

[4]张志立，张鹏，齐德昱.一种基于遗传算法的知识规则挖掘算法[J].郑州大学学报(理学版)，2004，36(3)：19-20.

[5]高坚.基于免疫遗传算法的多维关联规则挖掘[J].计算机工程与应用，2003(32)：185-186.

[6]彭建.一种基于遗传算法的关联规则挖掘方法[J].计算技术与自动化，2005，24(2)：75-77.

[7]任颖，李华伟，吕红.遗传算法在关联规则挖掘中的应用[J].电脑知识与技术，2009，5(16)：4260-4261.

(下转67页)

Research on the Performance Optimization of ASP.NET Enterprise Website

YUE Fu-qiang

(Xichang College, Xichang, Sichuan 615013)

Abstract: The performance optimization of ASP.NET enterprise website is extremely important to an enterprise, but the work of the website performance optimization is various and will take a long-term. This paper introduces the ideas of the website overall performance optimization, and then discusses the methods and techniques of the website performance optimization from the website architecture, page optimization, business logic, data access, cache, IIS website configurations and so on.

Key words: ASP.NET; Enterprise website; Performance optimization; Cache

(上接62页)

Association Rule Discovering Algorithm Based on Genetic Algorithm

ZHU Yan-ting

(Department of Computer Science, Guangxi Modern Polytechnic College, Hechi, Guangxi 547000)

Abstract: According to the requirement of association rule data mining, an genetic algorithm based on association rule data mining algorithm was developed. In addition to three basic operators, selection, crossover and mutation, the new operators-pick was included. The designing idea and algorithms were presented in detail. We design the experiment to test the performance of the algorithms. It is proved that the efficiency of the algorithms is excellent.

Key words: Data mining; Association rule; Genetic algorithm