基于遗传算法的关联规则数据挖掘

肖冬荣[a]，　杨　磊[b]

（南京信息工程大学 a. 电子与信息工程学院；b. 信息与控制学院，江苏 南京 210044）

【摘　要】关联规则是数据挖掘的重要手段，它基于支持度、置信度等对规则进行筛选，生成有用的规则，由于根据实际情况有时会产生虚假规则，所以兴趣度也自然被引入。遗传算法是自动化技术、专家系统等经常采用的算法。通过改进的遗传算法进行关联规则数据挖掘并进行了实例应用。遗传算法能较好地得出发生交通事故原因与结果的关联规则，提高数据挖掘的效率。

【关键词】数据挖掘；关联规则；遗传算法

【中图分类号】TP312　　【文献标识码】A　　【文章编号】1002-0802(2010)01-0205-03

Association Rule Data Mining Based on Genetic Algorithm

XIAO Dong-rong[a]，　YANG Lei[b]

（a. College of Electronic & Information Engineer; b. College of Information & Control,
Nanjing University of Information Science & Technology, Nanjing Jiangsu 210044, China）

【Abstract】The association rule is the important means for data mining. It is mostly based on the degrees of support and confidence for the choice of useful rule. Due to the actual situation, it may sometimes produce the false rule, so the interest degree should also be introduced. The genetic algorithm is usually used in automation field, expert system, and so on. The association rule data mining is implemented via the modified genetic algorithm and is put into practical application. The genetic algorithm could fairly achieve the association rule of the accident cause and result and improve the efficiency of data mining.

【Key words】data mining; association rule; genetic algorithm

0 引言

关联规则挖掘是找出满足给定支持度和可信度的有用规则[1]。Apriori 算法和它的一些改进算法是通过对数据库进行多次遍历来产生频繁项集，这些算法是非常耗时间和空间的[2]。而基于遗传算法的关联规则挖掘是通过选择、交叉、变异等对种群进行遗传操作，再对其进行筛选得出有用的规则。

1 兴趣度的引入

任何一条规则需要满足支持度和可信度的要求。但由于根据实际情况有时会产生虚假规则，所以兴趣度也自然被引入。先看下面一个喝茶与喝可乐的例子。

用图 1 分析规则{喝茶}=>{喝可乐}。由于该规则支持度(15%)和可信度(75%)都比较高，所以可得到喜欢喝茶的人也喜欢喝可乐。但是在所有人中不管是不是喝茶，喝可乐的人都占总人数的 80%，而既喝茶又喝可乐的人却只占 75%。也就是说一个人如果喝茶，则他喝可乐的可能性从 80%降到 75%，因此该规则可信度虽然高，却产生误导[3]。

	喝可乐	不喝可乐	
喝茶	30	10	40
不喝茶	130	30	160
	160	40	200

图 1 喝茶与喝可乐人数对照

支持度的缺点在于许多潜在的有意义的模式由于包含了支持度小的项而被删去，而可信度的缺点在于忽略了规则类别属性项集的支持度，而有可能产生误导，产生虚假规则，所以引入兴趣度，我们可以记作 inte(X=>Y)=P(X∪Y)/P(X)P(Y)=supp(X∪Y) /supp(X)supp(Y)。兴趣度越大于 1，用户对此规则就越感兴趣。兴趣度是用于过滤常规的，实用价值不高的规则，是用户对挖掘出来的知识的简洁性、确定性、实用性等的综合度量。

收稿日期：2009-02-17。
作者简介：肖冬荣（1944-），男，教授，博士生导师，主要研究方向为系统科学与控制工程；杨　磊(1984-)，男，硕士研究生，主要研究方向为计算机集成制造系统、数据库技术。

2 遗传算法的实现

2.1 编码

由于十进制编码准确性高，波动性小，本文采用十进制编码。假设所有特征属性和类别属性是离散的，并且取值范围已知，对每个属性的取值用一个十进制数表示，每个十进制数代表一个基因，将一个事务的所有属性的十进制数连接起来，形成十进制串，这就是染色体。每个染色体就是一条关联规则。编码[4]时字段顺序要保持不变。

2.2 适应度函数

遗传算法中适应度函数用来评价群体中的个体的优良程度，本文采用支持度来筛选规则，适应度函数为：fit(R)=S^*-Smin，S^*为经过遗传操作形成一条新规则的支持度，Smin 为用户给定的支持度阀值，若 fit(R)>0，保留此规则进入下一代，若 fit(R)<0，此规则将在遗传中被淘汰。

2.3 选择、交叉与变异

选择的本质是优胜劣汰，由适应度大小来决定。本文只要适应度大于零的规则就保留下来。如若为了体现好的个体的竞争力，更好的地实现优胜劣汰，也可以将适应度大的个体多复制一份放入种群中。

交叉算子是产生新个体的主要方法，将群体中 M 个个体随机组成 $M/2$ 对，对随机选择的交叉点按交叉概率进行交叉，由于效率高，模式保存较好，采用单点交叉。

变异算子是保持种群多样性，防止早熟的重要手段。采用基本位变异，对随机选择的变异位，每个个体按变异概率进行变异，变异个体的变异基因座上的基因值用基因座上的其它等位基因来替换，替换的基因值是其对应字段取值空间编码范围的任意值。

为了防止优良基因结构遭到破坏，可以适当地动态改变交叉概率 Pc、变异概率 Pm。设 Pc、Pm 分别有两个值 $Pc1$、$Pc2$、$Pm1$、$Pm2$（$Pc1<Pc2$，$Pm1<Pm2$），如果交叉两个体较小的那个个体适应度大于平均适应度，则 Pc 取较小值，$Pc=Pc1$；如果变异个体适应度大于平均适应度，则 Pm 取较小值，$Pm=Pm1$；相反地，如果交叉两个体较小的那个个体适应度小于平均适应度，则 Pc 取较大值，$Pc=Pc2$；如果变异个体适应度小于平均适应度，则 Pm 取较大值，$Pm=Pm2$。

交叉算子、变异算子配合，共同完成对搜索空间的全局搜索和局部搜索，从而完成寻优过程。

2.4 算法的流程图

算法的流程图如图 2 所示。

3 规则提取

根据遗传算法所形成的规则只能表明这些属性是关联的，但它们之间如何关联是不知道的，所以特征属性组合推导出类别属性组合必须满足可信度和兴趣度的要求，满足用户给定可信度和兴趣度的规则才输出。

步骤：一条规则，

If　　C>Cmin

　Then　计算此规则的 I

　If　　I>Imin

　Then　输出此规则

Else　　结束提取，转到下一条规则

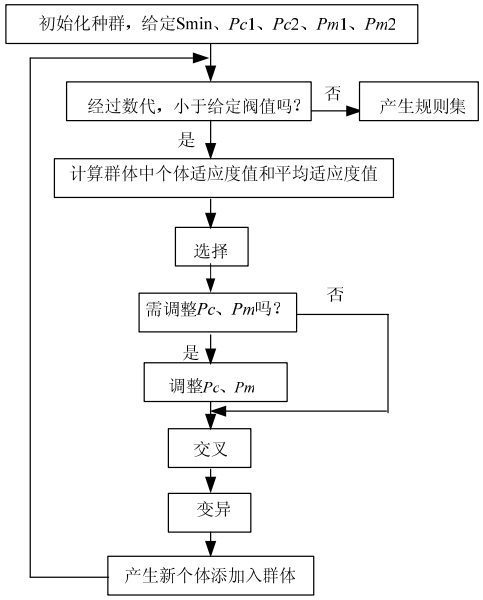

图 2 算法的流程

4 应用实例

随着城市汽车数量的增加，交通事故时有发生。各种因素导致了不同交通事故，它们之间存在着内在联系。

对数据库中的事故记录进行处理，保留对事故的产生有影响的字段，去除无用字段或影响甚微的字段，例如司机姓名、性别等字段。事故的诱因很多，如图 3 所示。

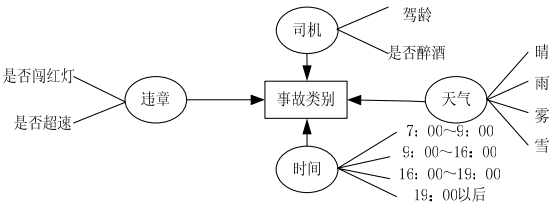

图 3 事故发生因素与事故类别关联

上图有 7 个属性，编码串长度设为 7，属性如图 4 所示。

司机		违章				
驾龄	是否醉酒	是否闯红灯	是否超速	时间	天气	事故类别

图 4　事故属性

对上图各个属性值进行十进制编码，注意编码时，字段顺序需保持不变，编码如下：

驾龄字段属性值分别为 2 年以下、2～10 年、10 年以上，对应码值分别为 1、2、3，同理醉酒为 1，没醉酒为 2；闯红灯为 1，没闯红灯为 2；超速为 1，没超速为 2；时间字段属性值分别为 7：00～9：00、9：00～16：00、16：00～19：00、

19：00 以后，码值分别为 1、2、3、4；天气字段属性值分别为晴、雨、雾、雪，码值分别为 1、2、3、4；事故类别字段属性值分别为严重、中等、轻微，码值分别为 1、2、3。

由于产生的规则包括上述全部字段，若出现某个或某几个字段与该规则无关的情况下，可以用 0 表示，也就是说，上述每一个字段属性值都有一个 0 值码，用户不需要此属性与其它属性的关联[5]。例如，码串“0111031”表明“醉酒、闯红灯、超速以及雾天的情况下，造成严重交通事故”，这与司机驾龄以及事故发生时间没有关系或关系甚微可忽略，并且此规则还覆盖了诸如“1111331”、“2111231”等码串。

用遗传算法进行关联规则数据挖掘可得出诸如晚上司机醉酒超速驾驶容易造成中等交通事故；晚高峰时司机驾龄在 2 年以下，超速驾驶容易造成中等交通事故；晚上司机醉酒闯红灯超速驾驶容易造成严重交通事故；雾天早高峰时司机驾龄在 2～10 年，超速驾驶容易造成轻微交通事故；雨天司机超速驾驶容易造成中等交通事故等规则[6]。

根据以上分析，我们发现了事故发生的因素与事故的类别有着内在联系，这为交管部门采取措施控制交通事故的发生或降低事故的重大程度提供了参考和决策支持。

5 结语

遗传算法能较好地得出发生交通事故原因与结果的关联规则。就关联规则而言，只考虑支持度、可信度，有时会出现虚假规则，而兴趣度可真正挖掘出用户感兴趣的规则。就遗传算法来说，从适应度函数、编码、选择、交叉、变异有一定改进，但由于在计算支持度时，一个规则要和其它所有规则比较后才能计算出该规则的支持度，因此该算法在计算规则支持度时要花费较多时间。

参考文献

[1] 李岚. 数据挖掘技术在电子商务中的应用[J].通信技术,2007,40(08):74-76.

[2] 陈安,陈宁,周龙骧,等.数据挖掘技术及应用[M].北京:科学出版社,2006.

[3] 周皓峰,朱扬勇,施伯乐.一个基于兴趣度的关联规则采掘算法[J].计算机研究与发展,2002,39(04):450-457.

[4] 李敏强,寇纪淞,林丹,等.遗传算法的基本理论与应用[M].北京:科学出版社,2002.

[5] 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(02):75-77.

[6] 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.

（上接第 204 页）

主机板上插一块功能板卡就可以实现，而且主机的双总线设计，大大提高了测试台的性能，在功能方面也提高了测试台的可扩展性[7]。

4 结语

在成功地利用 FPGA 作为整个硬件电路的控制逻辑上，所设计的测试台有效地提高了系统的可靠性、安全性、可扩展性，具有功能可扩展性。作者在总结了以往电路设计经验的基础上，提出了电路系统的可靠性设计方法，另外，FPGA 程序设计采用自顶向下，逐步求精的结构设计方法及采取的抗干扰措施，都证实了可以大大提高测试台运行的可靠性。测试台通过与变换器联试，能够对变换器的各项性能进行检测，满足了系统提出的各种要求。并通过实验证明了设计的实用性，已成功应用于某飞行器测试系统。文中创新点在于利用 FPGA 产生控制信号来对数据选择器进行控制，从而产生了一种效率更高、更准确的陀螺脉冲信号。并且还实现了 2 路不带电和六路 28V 带电的指令，同时拥有了传统的数字量变换测试平台和指令量变换测试平台的功能。

参考文献

[1] 李邦复.遥测系统(下)[M].北京:宇航出版社,1994.

[2] 高谦. 基于FPGA 的高性能MFCC特征参数提取[J]. 通信技术,2008,41(06):153-154.

[3] 任勇峰.飞航导弹遥测匹配装置自动监测系统研究[D].华北工学院,2000:18-20.

[4] 宋吉江.光电隔离器的工作原理及应用[J].微电子技术,2001,9(05):55-57.

[5] James Haigh. Bringing Together Drivers and Fieldbus Technology Control and Instrumentation[C].USA:[s.n.],2000:58-63.

[6] 宁楠,鲍慧,宋文妙,等.一种基于FPGA 的纠错编译码器的设计与实现[J].通信技术,2008，41(08):95-97.

[7] 杨永辉.脉冲信号对载波提取锁相环的干扰分析[J].通信技术,2007,40(05):8-10.