广西科学院学报
Journal of Guangxi Academy of Sciences

2010,26(2):162~166
Vol. 26,No. 2 May 2010

一种基于遗传算法的分类规则挖掘算法*

An Algorithm of Classification Rules Based on Genetic Algorithms

黄宏本[1],李余琪[1],覃伟良[2]

HUANG Hong-ben[1],LI Yu-qi[1],QIN Wei-liang[2]

(1. 梧州学院计算机科学系,广西梧州 543002;2. 长洲中学,广西梧州 543006)

(1. Department of Computer Science, Wuzhou University, Wuzhou, Guangxi, 543002, China; 2. Changzhou School, Wuzhou, Guangxi, 543006, China)

摘要:对遗传算法应用于分类规则挖掘问题进行研究,提出一种基于遗传算法和 Apriori 算法的混合分类规则挖掘算法,该算法的具体方案包括分类规则编码、适应度函数定义以及对进化后的规则的约简方法,最后通过实例仿真证明了该算法的有效性。

关键词:分类规则 遗传算法 数据挖掘

中图法分类号:TP311.2 **文献标识码**:A **文章编号**:1002-7378(2010)02-0162-05

Abstract: The classification rules mining using genetic algorithms is studied. Then a new classification rules mining algorithm based on genetic algorithms and apriori algorithm, include the encoding of classification rules, definition of fitness function and a reduction method for the evolutional rules is introduced. Finally, the effiency of this algorithm by simulation is proved.

Key words: classification rules, genetic algorithms, data mining

遗传算法(GA)[1]是由 J. H. Holland 在1975年提出的一种基于模拟生物进化的学习方法,广泛应用于工业工程优化等领域。分类是数据挖掘的一个重要概念,广泛应用于为决策提供支持。分类的算法主要有决策树分类、贝叶斯分类、基于规则的分类、反向传播法分类、关联分类等,其中基于决策树的主要分类算法包括 J. Ross Quinlan 开发的 ID3算法[2]以及后继算法 C4.5算法[3]、L. Breiman 等[4]开发的分类与回归树(CART)算法,这3种算法都采用了贪心算法构造决策树,缺点是决策树容易过分拟合、规模过大、产生的规则长度过长等[5]。

针对上述这些缺点,本文提出一种基于遗传算法和 Apriori 算法[6]的混合分类挖掘算法——AGA 算法,给出了算法的具体方案,包括分类规则编码、适应度函数定义等,并通过实例证明该算法的有效性。

1 AGA 算法流程

AGA 算法流程如图1所示。AGA 算法中遗传算法流程如图2所示。

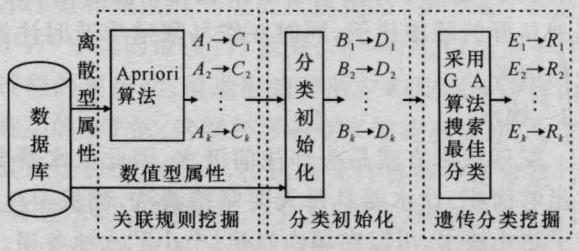

图1 AGA 算法流程

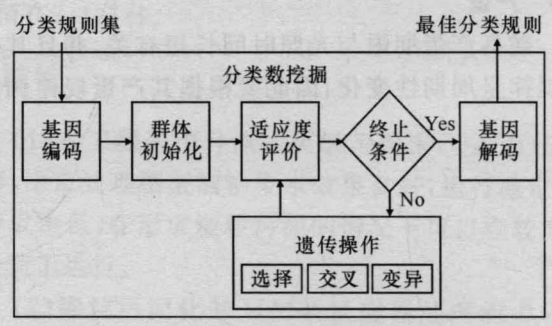

图2 AGA 算法中遗传算法流程

收稿日期:2010-04-12

作者简介:黄宏本(1977-),男,硕士研究生,讲师,主要从事数据挖掘研究。

*广西教育厅科研项目(200708MS056),梧州学院科研项目(2007C006)资助。

2 AGA 算法中的遗传算法的设计与实现

2.1 分类规则编码

分类规则为一逻辑公式,如 IF (Mtermrate= A or B) AND (Coursedif >1) THEN Pass,其中 Mtermrate 和 Coursedif 为特征属性,分别表示学生的期中成绩的等级和课程难易程度;Pass 为类别属性,表示通过期末考试。我们将该规则简化表示为 Mtermrate (A B) Coursedif (>1)=>Pass,现用二进制串来表示分类规则,首先对数值型特征属性进行离散化处理,Coursedif 属于数值型特征属性,其取值范围:0~2,将其分为4个区0<Coursedif≤0.5,0.5<Coursedif≤1,1<Coursedif≤1.5,1.5<Coursedif≤2,分别用4个类别(easy,midd,diff,verydiff)来表示;特征属性 Mtermrate 的取值域{A,B,C,D};类别属性 Pass 取值域为{pass,fail}。如果一个离散型特征属性有 k 种可能的取值,则在二进制串中为其分配 k 位,每一位与一特定的取值对应,取0表示析取式中没有该取值,取1相反。对类别属性,因只有2个类别,则分别用0和1表示。则分类规则可以表示成如下的染色体(二进制串)形式:

Mtermrate (A B) Coursedif (easy,midd)=>pass,

1100 1100 1

反过来,染色体0011001101对应规则为:Mtermrate (C D) Coursedif (diff, verydiff)=>fail。

2.2 适应度函数的定义

用遗传算法实现分类规则挖掘的过程是使"好的规则"得以生存,并作为父代繁殖出更好的规则,直至找到最优规则集为止。所谓"好的规则"是指规则与数据集中的各实例匹配程度高,适应度函数应能反映出规则对数据集的匹配程度。在适应度函数中,应综合考虑规则与实例的特征属性部分、类别属性部分及特征属性+类别属性(实例的全部)的匹配情况,只考虑其中的一种或两种情况是片面的,不能反映出最优规则的全部性质。为此作如下定义。

定义1 A 为数据集 R(由 n 条记录构成)的属性集合,由特征属性 C 和类别属性 D 构成(C,D 均为离散型),$A=C\cup D$ 且 $C\cap D=\Phi$。

定义 2 R 中与规则 r 特征部分成功匹配的数据集记为 R_C,与 r 类别部分成功匹配的数据集记为 R_D,与特征和类别均成功匹配的数据集记为 $R_C\cap R_D$。

定义 3 规则的支持度 $S(r)=|R_C|/n$。

定义 4 规则的置信度 $A(r)=|R_C\cap R_D|/|R_C|$。

定义 5 规则的覆盖度 $C(r)=|R_C\cap R_D|/|R_D|$。

以上定义的规则的支持度、置信度、覆盖度可以从不同的角度表明规则的性质,规则的支持度越大,说明规则在数据集空间所占的比例越大,规则的普遍意义越好;规则的置信度表示由特征(条件)推出类别(结论)的正确程度,当置信度为1时,规则恒真,此时只要条件为真,结论恒为真;规则的覆盖度表示结论包含于条件的正确程度,当覆盖度为1时,规则是完备的,此时条件为真是结论为真的必要条件。我们从支持度、可信度、覆盖度3个方面来综合评价一条规则的"好坏",在进化的过程中,让越"好"的规则获得越高的适应度值,使其在选择竞争中获得更大的生存和交叉的机会。

定义 6 适应度函数为:$F(r)=aS(r)+bA(r)+cC(r)$,其中:r 为规则变量;a,b,c 为常数且 $0\leqslant a,b,c\leqslant 1$;$S(r)$ 为规则支持度;$A(r)$ 为规则的置信度;$C(r)$ 为规则的覆盖度,a,b,c 的值由用户根据需要调整,从而对规则评价的偏重方面可以发生变化,使进化沿着用户期望的方向进行。

2.3 遗传操作的设计

遗传操作的设计主要包括选择、交叉和变异操作的确定。

2.3.1 选择操作的确定

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。这样就存在两个问题:(1)在进化的初期,有可能产生个别适应度特别高的个体,这些个体的适应度远大于种群的平均适应度。按照赌轮选择方式,它们能够复制出很多后代,所带的基因在几代之间就能占满整个种群。这就会使种群因基因单一而无法继续进化,从而使搜索过程陷于局优解。(2)在进化的后期,种群中的个体的适应度已经进化得相差无几,这时候采用赌轮选择方式就会使每一个个体都获得复制一份的机会,体现不出好的个体的竞争力,无法实现遗传算法的优胜劣汰的原则。

鉴于此,我们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。其过程描述如下:

```
fitsort( )
    {将种群中的个体按适应度大小进行排序;
    }
while 种群还没有扫描完 do
```

{排在前面的个体复制两份；
中间的复制一份；
后面的不复制；
}

2.3.2 交叉操作的确定

在解决过早收敛问题时，通常习惯于采用限制优良个体的竞争力(高适应度个体的复制份数)的方法。这样无疑会降低算法的进化速度，增大算法的时间复杂度，降低算法的性能。由于种群的基因多样性可以减小陷入局优解的可能，而加快种群进化速度又可以提高算法的整体性能。为了解决这一对矛盾，尝试一种在不破坏种群的基因多样性前提下加快种群的进化速度的方法，这一方法描述如下：在随机选择出父本和母本以后，按照交叉方法(单点，多点，一致交叉)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。

在此交叉操作采用单点交叉，随机产生交叉位，交换两个父代个体交叉位前后的串，形成两个新的个体. 例如对于父代个体1“1111111”和父代个体2“0000000”，随机产生的交叉点是3，则交叉后产生的子代个体1和子代个体2分别为“1110000”、“0001111”。

2.3.3 变异操作的确定

在遗传算法中，如采用固定的变异概率 P_m，则当 P_m 取值很小时，变异算子对群体不会产生影响，不利于新的基因的引入；当 P_m 取值很大时，有可能破坏群体中的优良基因，使得算法收敛速度变慢甚至不收敛。在这里，我们提出一种可变变异概率的方法，这一方法描述如下：

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，又有利于新的基因的引入，从而可以很大程度地提高遗传算法的性能。此时的变异操作方法是随机产生变异位，对变异位作翻转操作，例如对于个体“1111111”，产生的变异位为3；对第3个字符作翻转操作后个体变异为“1101111”。

2.4 分类规则约简

规则约简的目的是使用最少的特征属性集合和最少的分类规则数来描述对象的分类知识，达到对对象进行有效分类的目的。

执行遗传算法得到的最优规则集是针对适应度而言的，在这个规则集中会存在大量冗余规则，如规则 Mtermrate (A) Coursedif (easy，midd)＝＞pass 和规则 Mtermrate (B) Coursedif (easy，midd)＝＞pass 就是两条冗余规则，可以合并为一条规则，即 Mtermrate (A B) Coursedif (easy，midd)＝＞pass。合并后规则的语义是：学生期中成绩为 A 或 B 且课程难易程度为 easy 或 midd 则期末成绩一定能通过。另外提出一种规则约简方法，其基本思想是：如果分类知识中的两条分类规则中仅有一个特征属性的描述不同，则将这两个描述归纳为一个更广义的描述。这里所指的特征属性的不同描述包含以下两种情况：第一种情形是指两条规则中该属性的描述值完全不同，如 Mtermrate (A) 和 Mtermrate (B)；第二种情形是指某条规则中该属性的描述包含另一条规则中该属性的描述，如 Mtermrate (A) 和 Mtermrate (A B C)。因此在对规则进行归纳合并时应根据不同情形予以对待，对于第一种情形，可将该属性的不同描述值合并为一析取式，如对 Mtermrate (A) 和 Mtermrate (B) 合并后得析取式为：Mtermrate (A B)；对于第二种情形，可判定两条规则的包含与被包含关系，将被包含的那条规则删去，而保留包含对方的那条规则，如上例中将含 Mtermrate (A) 的规则删去，而保留含 Mtermrate (A B C) 的规则。对归纳合并后的规则集，还需检查其中每一条规则的每一个属性的描述值是否包含了该属性的全部取值，若是，则说明该属性不论取何值该规则都成立，因此需将该属性对应的合取项从该规则中删除以简化规则，否则对该属性不作任何处理。如对规则 Mtermrate (A B C) Coursedif (easy)＝＞pass，经过该项处理后，规则被进一步简化为 Coursedif (easy) ＝＞pass，其相应的语义也更为简洁：课程难易程度非常容易学生肯定可以通过的。简约过程描述如下。

输入：某一规则集；

输出：约简后的规则集。

Step1：建立数组 Rule [n] 以表示该规则集；

Step2：将规则集中的规则进行 $n-1$ 遍两两比较合并

For $i=1$ to $n-1$ Do

For $j=i+1$ to n Do

将 Rule [i] 与其后的 Rule [j] 进行比较，若两条规则中仅有一个特征属性的描述不同，则根据不同情形进行合并，合并后的规则存放在 Rule [j] 中，只要发生过一次规则的合并，就将 Rule [i] 置为空；

Step3:For $i = 1$ to n Do

检查 Rule [i] 条规则,若该规则的某个属性的描述值包含了该属性的全部取值,则从该规则中删除该属性对应的合取项,得到约简后的规则集。

3 AGA 算法应用实例分析

3.1 AGA 算法中的变量设计

通过采用一所大学的学生信息的数据集对基于 AGA 算法的挖掘结果和基于简单遗传算法的挖掘性能进行测试。模型系统中的变量如表1所示。

表1 模型中的变量

特征属性	对应的变量名	数据类型
X_1:院系代码	Depart-code	离散型
X_2:性别	Gender	离散型
X_3:期中成绩等级	Mtermrate	离散型
X_4:课程学分	C-credits	离散型
X_5:课程类别	Course type	离散型
X_6:课程难易度	Coursedif	离散型
X_7:不及格记录	flunk ratio	数值型

表1中的特征属性 Depart-code 的取值域为{jsjkx、jsjtx、zdh、xxaq},分别代表计算机科学与技术专业、计算机通讯专业、自动化专业、信息安全专业;特征属性 Gender 的取值域为{male、female};特征属性 Mtermrate 的取值域为 {A、B、C、D};特征属性 Coursetype 的取值域为{required、optionnal};特征属性 coursedif 经过离散化处理后的取值域为{easy、midd、diff、vdiff };特征属性 flunkratio 经过离散化处理后的取值域为{no、low、high、vhigh};类别属性的取值范围为{vgood、good、pass、fail}。

3.2 分类规则挖掘结果

取群体初始规模为100,适应度函数:$F(r) = a * S(r) + b * A(r) + c * C(r)$,其中 $a = b = c = 1$。对1000个数据记录实施 AGA 算法,挖掘结果如F(每条规则后赋有相应的适应度值、支持度、置信度和覆盖度):

Mtermrate (B C) Coursetype(optional)=>pass(1.70,0.29,0.81,0.62);

Mtermrate (D) Coursedif(diff)=>fail(1.83,0.31,1.1,0.51);

Coursetype(optional)Coursedif(easy)=>pass(2.21,0.59,0.67,0.89);

Mtermrate (C D) Coursetype (required) Coursedif(vdiff)=>fail(1.28,0.31,0.49,0.31);

Mtermrate (A B) Coursedif(diff)flunkratio(no)=>good(1.76,0.44,0.72,0.73);

Mtermrate (A) Coursedif(easy midd)flunkratio(no)=>vgood(1.86,0.43,0.71,0.7);

Mtermrate (C) Coursedif(diff)flunkratio(hogh vhigh)=>fail(1.24,0.37,0.69,0.54)。

3.3 与简单遗传算法(SGA)的性能比较

习惯上将 J. H. Holland 提出的遗传算法称为简单遗传算法(SGA),SGA 中参数设置如表2所示。

表2 简单遗传算法(SGA)中参数设置

项目	值
种群大小	100
代数	50/100/200/300/400/500
交叉概率	0.5
变异概率	0.01
选择操作	赌轮方式

仿真平台 Matlab 的版本为 Matlab2006a。在遗传代数100/200/300/400/500的情况,对提出的 AGA 算法与 SGA 算法在平均出错率方面进行了对比,实验结果如图3所示。

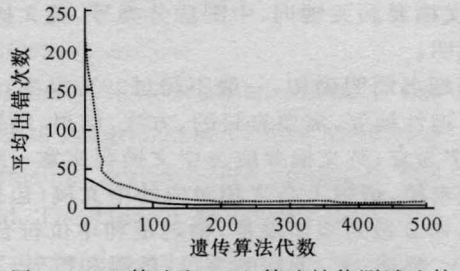

图3 AGA 算法和 SGA 算法性能测试比较

·····:SGA 算法;——:AGA 算法。

由图3可以看出,由于 AGA 算法是在 Apriori 算法的基础上进行的,在遗传代数较小的情况下,出错率要比简单遗传算法小得多,随着遗传代数的增大,平均出错次数也会线性减少,但是减少率明显下降。如图3中遗传代数为50,100时,执行 AGA 算法的出错率比执行一般遗传算法所产生的出错次数要小得多。当遗传代数较大时,基于 Apriori 算法的初始化结果对搜索的加速作用会变得越来越小。当遗传代数大到一定程度时,基于 Apriori 初始化结果相当于一组随机的规则,AGA 算法转变成简单遗传算法。由图3可以看出随着遗传代数的增加,执行 AGA 算法所产生的平均出错次数越来越接近执行简单遗传算法,这说明 AGA 算法适用于遗传代数较小的情况。但是由于一般情况下遗传代数比较少,因此该算法具有实用性。

参考文献:

[1] Holland J H. Adaptation in natural and artificial

system[M]. Ann Arbor: The University of Michigan Press, 1975.

[2] Quinlan J R. Induction of decision trees[J]. Machine Learning, 1986, 1(1): 81-106.

[3] Quinlan J R. C4. 5: Programs for machine learning [M]. San Francisco: Morgan Kaufmann Publishers Inc, 1993: 170-247.

[4] Breiman L, Friedman J, Olshen R, et al. Classification and regression trees [M]. San Francisco: Morgan Kaufmann Publishers Inc, 1993.

[5] Jiawei Han, Micheline Kamber. 数据挖掘概念与技术 [M]. 范明, 孟小峰, 译. 北京: 机械工业出版社, 2007: 189

[6] Agrawal R, Srikant R. Fast algorithms for mining association rules: proceedings of the 20th International Conference Very Large Data Bases (VLDB'94) [C]. Santiago, Chile, 1994: 487-499.

(责任编辑: 韦廷宗)

《广西科学院学报》投稿要求和注意事项

1. 文稿可以寄打印稿，也可以将电子文稿直接发送到本刊邮箱(gxkxyxb@gmail. com)。本刊接受方正小样文件，. TXT，. DOC，. WPS，. TEX 文件。文稿文责自负，附有不一稿多投的证明或说明函件。为了便于联系，文稿请注明联系电话、E-mail 地址和详细的通信地址。

2. 文稿务必论点明确，论据可靠，数据准确，文字精炼。每篇论文(含图、表、公式、参考文献等)一般不超过8000字(研究简报不超过2000字)。文稿必须包括题目(中英对照)、工作单位(中英对照)和电子信箱、邮政编码、中文摘要和关键词、中图法分类号、英文摘要和英文关键词，正文，致谢(必要时)，参考文献，表格和插图及其说明。

3. 文稿题名简明确切，一般不超过20个汉字；摘要要用第三人称书写，不使用“本文”、“作者”等做主语，尽量写成报道性摘要，需要有目的、方法、结果、结论的内容，不重复本学科领域已经成为常识的内容，一般以不超过400字为宜；英文摘要应与中文摘要文意一致，并符合英文语法规范，以不超过250个实词为宜。

4. 英文来稿，请附上与之相对应的中文稿(包括题名页，正文，致谢，参考文献，表格和插图及其说明)。

5. 文稿务必做到写作规范，物理量和单位符合国家标准和国际标准。稿件中的外文字母和符号必须分清大、小写，正、斜体；上、下标的字母、数码和符号，其位置高低区别应明显可辨；外文缩略词和容易混淆的外文字、符号请在第一次出现时注明中文名称。

6. 文稿中只需附必要的图、表、照片。图中文字、符号要注明清楚，并与正文一致。照片请用光面相纸印出，要求清晰、层次分明。图、表、照片应注明序号和插入文内的位置。图、照片大小一般以80mm×50mm 或160mm×100mm 为宜。

7. 参考文献只需择主要者列入，未公开发表的资料请勿引用。文献序号请按文中出现先后为序编排。书写格式，期刊：“序号 作者姓名(不超过3人者全部写出，超过者只写前3名，后加‘等’或‘et al.’。外文姓前名后，名缩写，不加缩写点，姓名用大写字母). 文章题目[J]. 期刊名(外文可缩写，不加缩写点)，出版年，卷(期)：起止页码.”；如果期刊无卷号，则为“年(期)：起止页码”。专著：“序号 作者(英文姓名用大写). 书名[M]. 版本(第一版不写). 出版地：出版单位(国外出版单位可用标准缩写，不加缩写点)，出版年：起止页码.”

8. 本刊编辑部可以对文稿进行规范性删改。如作者不允许，务请在来稿中注明。

9. 请作者自留底稿，投到本刊的文稿无论刊登与否不再退稿。本刊编辑部收到稿件，即寄发收稿回执。收到本刊收稿回执2个月内，本刊编辑部会告之文稿是否录用或修改，若超过期限请向本刊编辑部咨询。

10. 自治区、省(部)级以上重大科研项目及攻关项目，国家863计划项目，自然科学基金资助项目，开放实验室研究项目和拟到国际学术会议上宣读的论文优先发表，请作者投稿时注明，并写清项目编号。

11. 文稿不得侵犯他人版权，如有侵权，由投稿者负完全责任。

12. 文稿一经采用，酌收版面费；刊登后，付稿酬含(《中国学术期刊(光盘版)》、中国期刊网、万方数据网及台湾华艺 CEPS 中文电子期刊服务网等网络发行的的稿酬，并同时赠送每位作者1本样刊。

13. 本刊入编《中国学术期刊(光盘版)》、中国期刊网、万方数据网及台湾华艺 CEPS 中文电子期刊数据库。作者如果不同意将论文入编上述数据库，请在来稿时声明，本刊将作适当处理。