基于遗传编程的分等级规则挖掘模型

周　崎[1],王　斌[2]
(1.盐城工学院 广电处,江苏 盐城　224002; 2.盐城工学院 工程中心,江苏 盐城　224002)

摘要:根据分等级规则挖掘的具体要求,结合遗传编程的特点建立了一种新型的规则挖掘模型。对规则的编码方式和遗传算子进行了详细的设计,构造了满足遗传编程挖掘的树型染色体规则结构,并综合考虑支持度、可信度和相关度3个指标和等级系数等约束因素来设计适应度函数。最后通过实例对基于遗传编程的分等级规则挖掘模型的可靠性进行了验证。

关键词:遗传编程;分等级规则;数据挖掘
中图分类号:TP392　　**文献标识码**:A　　**文章编号**:1671-5322(2008)02-0039-05

规则挖掘的目的是发现大规模数据集中项集之间有趣的关联或相关关系[1]。目前国内外对基于人工智能技术的规则发现进行了积极深入的研究,提出了粗糙集[2]、决策树[3]、遗传算法[4]和人工神经网络[5]等规则发现方法。但在许多应用中,仅在低层的数据项之间,很难找出强关联规则,可能在不同等级的项之间找出的规则比仅在同等级原始数据之间的关联度更高一些,而上述方法在分等级的规则发现方面又存在着种种不足之处。因此本文提出一种基于遗传编程(genetic programming, GP)的分等级规则挖掘方法,它能有效解决挖掘数据量庞大及数据之间存在复杂的拓扑关系、位置关系和度量关系关系。

遗传编程是遗传算法(genetic algorithms, GAs)的扩展。GP利用自然进化的原理进行程序的进化。它与GAs的最大区别在于其个体是可以执行的程序而非字符串[6]。在标准GP技术中,程序表示为语法树的形式。GP可以用来生长"树"。这些树代表一些有组织的指令与函数,可以用来对给定的初始结构进行结构与参数操作,从而让简单的规则树成长为满足需要的规则树[7,8]。因为在GP中,个体的长度是开放式的变长表示,所以特别适合分等级规则的开放式的结构空间搜索。

1　基于GP的分等级规则挖掘模型

基于GP的分等级规则挖掘是运用GP的自适应寻优及智能搜索技术,获取与客观事实相容的问题的解。GP首先将问题的可能的解进行编码,形成染色体(GP树),构成初始群种。根据评价函数计算每个GP树适应值,复制适应值高的染色体(GP树),并通过遗传操作(选择、交叉、变异)来产生新的染色体种群(见图1)。

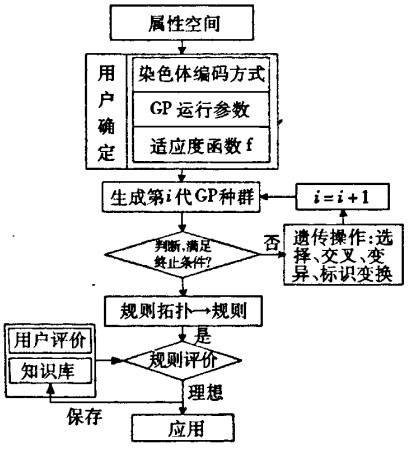

图1　分等级规则拓扑GP挖掘运算流程

Fig. 1　GP-Based Hierarchical Rules Mining Process

收稿日期:2008-02-15

作者简介:周崎(1981-),男,江苏盐城人,助教,南京农业大学硕士研究生,主要研究方向为数据库挖掘。

该模型主要具有以下特点:

(1)模型根据个体适应性分等级地组织成多样性群体,而个体的竞争是在同级之间进行的。

(2)模型中个体的交换是单方向的,在等级交换结构中允许个体从低适应性群体向高适应性群体移动。

(3)模型中不同适应性水平群体的进化平衡被维持。低适应性群体往往探索新的搜索领域并将它们的后代送往更高适应性的群体进行进一步的进化。

(4)该模型在不消除适应度低个体的前提下维持大量的适应度高的个体,从而有效的避免了早熟现象。

(5)当适应度低的个体能持续进行彻底地搜索时,一旦它们产生适应度高的后代,后代会立即进入适应度高的层次,来完成和其它的高适应性个体的重组,保证了同水平情况下的公平竞争。在低适应性群体中随机地插入新的个体,以帮助探索新的搜索领域,并减少搜索陷入局部最优解的可能。

(6)该模型具有自适应性,能根据其特征适当的分配搜索能力,并且该模型和当前的其它技术兼容,来改善进化算法的搜索能力。

2 基于GP的规则挖掘准备过程

2.1 数据准备过程

根据规则挖掘的要求需要对数据仓库中的数据进行抽取,本文采用SQL语句来完成数据的初步选取,如以下语句表示从用户数据库中抽取有关企业ID、企业类型、规模、应用服务类型方面的信息来进行分析。

Select Enterprise_ID, Enterprise. type, Scale, economic status, Service_type from Database_asp. user

group by Enterprise. ID

having (Entreprise. type = Metal products Manufacturing') and (Scale = Large)

and (economic status > = 3) and (service_No. > = 2) and (Service_type = ERP')

2.2 分等级规则的GP编码

将上一阶段清洗好的数据进行分类,并按数据的等级、类别来建立完整的条件树。条件树中以实线连接的部分称为前件集,其中每一个节点 $P_i(i=1,2,\cdots m)$ 均为前件,下标的位数 N 表明该前件所处的层数。树中虚线连接的部分称为后件集,其中每一个节点 $R_i(i=1,2,\cdots\cdots m)$ 均为后件(如图2所示)。$C_i(i=1,2,\cdots m)$ 是赋不同节点数值的系数。前件与前件之间的关系默认为逻辑与的关系。该条件树中每一节点的位置都是固定不变的,在规则挖掘过程中,不参与挖掘的条件均被赋予 *NULL* 值,但其在GP树中所处的位置保持不变。

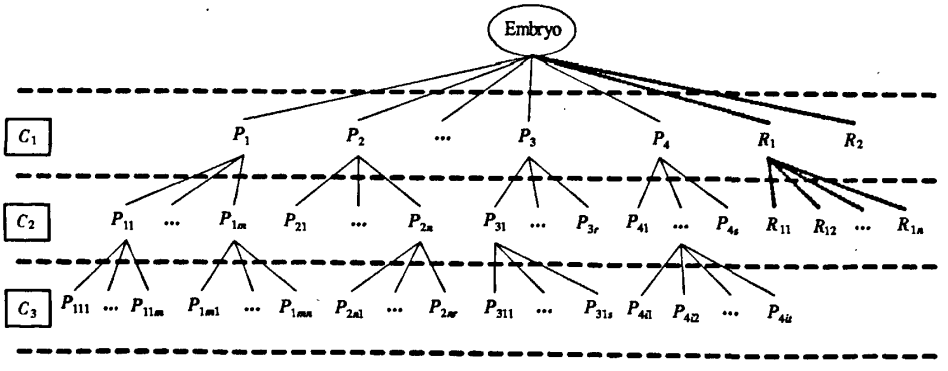

图2 分等级规则的GP编码模式

Fig. 2 GP-Based Hierarchical Rules Coding Model

2.3 产生初始种群

产生的初始种群是通过随机算法生成的树型程序,初始程序的数量预先设定。本文中初始种群是从完整的条件树拓扑中根据一定的约束条件抽取的拓扑,而建立的规则空间。规则包含两部分内容:前件(If部分,即条件属性)和后件(Than部分,即目标属性)。此外,染色体个体(chromosome)表示的预测规则都是模糊的。规则的每一

属性都具有两条对应的属性值(No_Attr/No_GoalAttr。其中:No_Attr是某个属性在整个待挖掘数据集中的出现的频率,No_GoalAttr是目标属性在数据集中出现的频率)。染色体GP树由众多条件属性组成,其中一个节点对应表示一个条件属性及其值域。建立初始GP树时要求:①完整条件树末层的每一分支只选取唯一节点作为该拓扑的终端;②初始解空间中所有GP树必须以完整条件树的末层终端为终端;③GP树中终端的父节点在挖掘过程中只进行遗传操作,在计算适应度时,该节点条件将不参与计算。

3　遗传操作

3.1　选择

GP使用选择算子对群体中的个体进行优胜劣汰操作,本文采用轮盘选择法,即假设每个个体的适应度为$f_i(i=1,2,3,\cdots,n)$,种群的总适应度为$\sum_{i=1}^{n} f_i$,则个体k的选择概率为$f_k/\sum_{i=1}^{n} f_i$。选择操作建立在对个体的适应值进行评价的基础之上。选择操作的主要目的是为了避免基因缺失、提高全局收敛性和计算效率。

3.2　交叉

交叉运算是指按某种方式相互交换两染色体GP树上对应的部分,从而形成两个新的个体(见图3)。由于本文中GP树上不同位置的节点与不同类型的条件相对应,因此交叉操作只能在两个体对应的节点和层上进行,即:GP树A上第i层的第j个节点只能与GP树B上第i层的第j个节点进行互换。

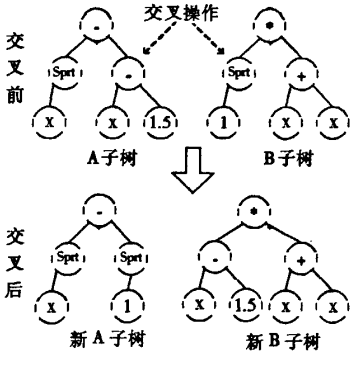

图3　交叉操作

Fig.3　Crossover Operation

3.3　变异

本文采用的是基于均匀变异的改进方法,即以某一变异概率在种群中随机选择变异个体,选中后,将这个个体GP树的每一都依次进行变异,并且子树的每一最高等级的子节点都被随机赋予同一父节点下相同等级的其他某个子节点的值,这样进行依次变异之后就可以保证每个属性值都将存在,如图4所示。

4　适应度函数设计

关联规则挖掘的任务就是要发现能够反映记录属性之间的关联或者是关系,这些规则应该具有一定的支持度、相关度。对挖掘出来的规则我们从支持度、可信度、相关度三个方面来综合评价。为让潜在的规则获得较高的适应度值,使其在选择竞争中获得更大的生存和交叉的机会。在这里定义函数:

$$n(k) = C_i(W_S S(k) + W_C C(k) + W_R R(k)) \tag{1}$$

其中:k为染色体代数;W_S,W_C,W_E分别为$S(k)$、$C(k)$、$R(k)$的权重,且$0 \leqslant W_S, W_C, W_R \leqslant 1$,$W_S$、$W_C$、$W_R$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;$S(k)$为规则支持度,$S(k)=P(CondAttr \cap GoalAttr)$,其中CondAttr为前件集,$CondAttr$为后件集;$C(k)$为规则的可信度,$c(k)=P(GoalAttr \mid CondlAttr)$,表示在前件集$CondlAttr$出现的前提下后件集$CondlAttr$出现的概率;$R(k)$为规则的相关度,相关度计算明确地表明前件后件之间的依赖关系是相互促进还是相互制约的。相关度计算公式如下:

$$R(k) = p(m)/(P(CondAttr) \cdot P(GoalAttr)) \tag{2}$$

其中:$P(m)=P(CondAttrUGoalAttr)$。如果$R(k)=1$,则说明前件后件相互独立;如果$R(k)>1$,则说明前件后件正相关;如果$R(k)<1$,则说明前件后件负相关。

从挖掘出的关联性属性集中提取出关联规则,需要建立以下提取标准:满足值$S(k) \geqslant \min S(k)$、$C(k) \geqslant \min C(k)$、$R(k) \geqslant \min R(k)$的规则才能继续进行适应度评价,否则舍去。其中$\min S(k)$、$\min C(k)$、$\min R(k)$是用户给定阈值。关联规则的评价是一个多目标决策问题,决策者通常难以准确地给出各个目标的具体值。不同等级规

则的 $n(k)$ 值差别较大,通常 GP 树中父节点对应条件产生的规则,其 $S(k)$、$C(k)$、$R(k)$ 等值均明显高于由子节点对应条件产生的规则。为避免或有意突出某个等级的规则,适应度函数可以表示为:$f(k)=n(k)\cdot\sum_{i=1}^{m}C_i/m$,不同等级的条件属性设定相应的参数 C_i,其中 i 为等级标识码,m 表示规则中不同等级的前件个数。

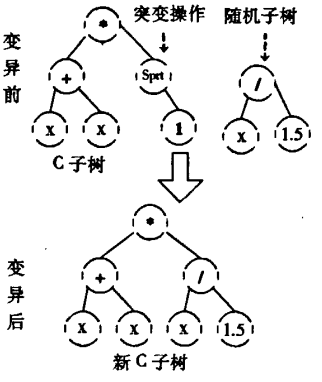

图4　变异操作

Fig.4　Mutation Operation

5　实例

例如从某银行数据库系统中挖掘顾客的性别、年龄、职业、收入等与贷款信用度之间的关系。实验运行参数设计如下:种群规模:200;GP 树最大深度:4;初始深度:4;最大节点数:100;交叉率:0.9;变异率:0.1;最大繁殖代数:100。

GP 挖掘结果如下:

图5所示的 GP 树形图表明了如下两条规则。经挖掘计算规则1、规则2的3个单项指标值均高于设定的阈值,且获得了最高的适应度函数值。

规则1:

Age(40,50]^Vocation(officer)^Income(3000,4000] = > Credit(high),

$S(k)=0.316$、$C(k)=0.967$、$R(k)=1.972$,$f(k)=3.416$

规则2:

Age(Old)^Vocation(retiree)^Income(1500,2500] = > Credit(high),

$S(k)=0.193$、$C(k)=0.992$、$E(k)=2.365$,$f(k)=3.392$

6　总结

本文首次将在拓扑结构搜索方面优势明显的遗传编程方法应用到分等级规则挖掘中,并根据规则挖掘的要求与特点,对传统遗传编程的选择、交叉、变异算子和适应度函数的设计进行了相应的改进,提出了一个基于遗传编程的分等级规则挖掘模型。与其他挖掘算法比较,本方法可以更加高效地挖掘和更加清楚地描述分等级规则。

参考文献:

[1] Tsai,Chen Chienming. Mining interesting association rules from customer databases and transaction databases[J]. Information Systems, 2004,29(8):685-696.

[2] BI Yaxin, Terry A, MCCLEAN Sally. A rough set model with ontologies for discovering maximal association rules in document collections[J]. Knowledge-Based Systems. 2003,16(5-6):243-251.

[3] Carvlho D, Freitas A. A hybrid decision tree/ genetic algorithm method for data mining[J]. Information Sciences, 2004, 163(1-3):13-35.

[4] Wang bin, Xie qingsheng. mining fuzzy association rules from the ASP database[C]. proceedings of E-ENGDET 2006, 5th international conference on e-Engineering & digital enterprise technology. Guiyang,2006.

[5] Carpenter G, Martens S, Ogas O. Self-organizing information fusion and hierarchical knowledge discovery: a new framework using ARTMAP neural networks[J]. Neural Networks. 2005,18(3):287-295.

[6] Koza J R. Genetic Programming: On the Programming of Computers by Means of Natural Selection[M]. The MIT Press, 1992.

[7] 李少波,陈茜,胡建军.基于分等级搜索的可持续进化算法研究[J].中国机械工程,2006,17(11):1162-1165.

[8] 李少波,胡建军,谢庆生,等.基于遗传编程(GP)与键合图的机电系统自动设计[J].系统仿真学报,2002,14(11):1513-1516.

Hierarchical rules mining model based on genetic programming

ZHOU Qi[1], WANG Bin[2]

(1. Division of Radio and Television Yancheng Institute of Technology, Jiangsu Yancheng 224002, China;
2. Engineering Research Center, Yancheng Institute of Technology, Jiangsu Yancheng 224002, China)

Abstract: According to the requirements of mining hierarchical rules and the characteristic of the genetic programming (GP), a hierarchical rule mining pattern based on GP has been proposed. To construct the chromosome tree of the hierarchical rules that is necessary for GP, the coding pattern and genetic operators of the hierarchical rule mining based on GP has been designed. To design the fitness function, three indicators – support, confidence, and correlativity, and some other constraints such as hierarchical coefficients were considered. Finally, a case has been employed to demonstrate the reliability of the GP – based hierarchical rules mining model.

Keywords: genetic programming; hierarchical rules; data mining

(上接第38页)

[9] Oja E. A simplified neuron model as a principal components analyzer[J]. Journal of Mathematical Biology, 1982,15: 267 – 273.

[10] Zhao L, Suzuki H, Nakagawa S. A Comparison Study of Probability Functions in HMMs through Spoken Digit Recognition, IEICE, TRANS. INF and SYST. 1995(6):669 – 675.

[11] 新美康永. 音声认识[M]. 日本共立出版社,1987.

Study on Speaker Recognition Under Noise Environment Based on PCANN/PDP

XIA Shu-lan

(Department of Experiment Teaching, Yancheng Institute of Technology, Jiangsu Yancheng 224051, China)

Abstract: This paper presents the method for speaker recognition based on PCANN/PDP mixed configuration. It adopts a sequence of frames as input of speaker recognition system to introduce the correlation between frames. In allusion to the noise characteristics of small relativity, small contribution in main component of voice signal, it adds PCANN for compressing voice main parameters in the preceding part of speaker recognition system. At the same time, it advances speaker recognition of probability DP. Through the experiments of speaker – independent in noise environment, the validity of the proposed approach could be verified.

Keywords: speaker recognition; noise; DP matching algorithm; principal components analysis neural network