基于二进制信息粒的数据挖掘算法研究

徐健锋[1] 刘 斓[2] 邱桃荣[2] 胡 然[1]
(南昌大学软件学院 南昌 330029)[1] (南昌大学信工学院 南昌 330029)[2]

摘 要 本文利用二进制数来表示基于粗糙集思想的信息粒的概念,定义了基于上述信息粒的一些粒计算的基本定义。并且提出了两个典型的基于粒计算的数据挖掘算法,即基于二进制信息粒的属性重要度算法和决策树算法。最后实验和分析验证了研究的有效性。

关键词 数据挖掘,信息粒,粗糙集,二进制

On Data Ming Algorithms Based on Binary Numeral Granular Computing

XU Jian-Feng[1] LIU Lan[2] QIU Tao-Rong[2] HU Ran[1]
(College of Software, Nanchang University, Nanchang 330029)[1] (College of Information Engineering, Nanchang University, Nanchang 330029)[2]

Abstract Firstly, the binary numeral is used to present each information granule base on rough set. Secondly, definition of granular computing and two algorithms base on granular to obtain the importance of attributes and construct decision tree. Finally an instance is proposed to prove effective of the algorithms.

Keywords Data ming, Information granule, Rough set, Binary numeral

人类在认知、推理和做决策都是在拥有了大量信息背景下进行的。由于人类的能力有限,常常把大量复杂信息按其各自的特征和性能将其划分成若干较简单的块,如此划分出来的每个块被看成一个信息粒。这种处理信息的过程被称为信息粒化。粒化涉及到整体到部分的划分,一个粒是通过不分明性、相似性或功能性被链接在一起而形成的对象块。本文中我们依据对象的不分明性即属性值相等进行粒化,把每个等价类看作一个粒。

表 1 S=(U,AT)

U	a	b	U	a	b
1	1	2	7	2	2
2	2	3	8	1	3
3	1	3	9	1	3
4	3	2	10	2	2
5	4	1	11	5	2
6	3	3	12	1	3

如表 1,给定一个信息系统 $S=(U,AT)$,其中 $U=\{1,2,\cdots,12\}$ 是对象的集合,$AT=\{a,b\}$ 是属性集。属性 a 或 b 均可以将 U 划分成互不相交的等价类,即信息粒。商集

$U/IND\{a\}=\{a_1,a_2,a_3,a_4,a_5\}$,

$U/IND\{b\}=\{b_1,b_2,b_3\}$。

其中:$a_1=\{1,3,8,9,12\}$,$a_2=\{2,7,10\}$,$a_3=\{4,6\}$,$a_4=\{5\}$,$a_5=\{11\}$;$b_1=\{5\}$,$b_2=\{1,4,7,10,11\}$,$b_3=\{2,3,6,8,9,12\}$。

商集中每个元素就是一个信息粒,a_1,a_2,b_3 等为信息粒的名称,等价类被定义为粒,显然是 Rough 集理论的扩展,因为实施粒计算比施行等价类计算要快得多。

粒计算是一种看待客观世界的世界观和方法论,它可把原来粗信息粒的大对象分割成若干细信息粒的小对象进行研究,也可把原来细信息粒的若干小对象组合成若干粗信息粒的大对象进行研究。即论域分为粒度粗细不同的空间,把信息粒作为问题求解的最基本单元。有关粒的计算规则可参考文[1,3]。

每个粒中元素对于它在全域 U 上的位置,即该元素的下标,本文用一个二进制数串来表示一个粒。如果一个个体对象 x_i 属于粒 a_j,那么表示该粒二进制数串中的第 i 位上置 1。显然,二进制数长度恰好等于 U 的基数,$U=111111111111$。上例的信息粒的二进制表示分别为:$a_1=101000011001$,$a_2=010000100100$,$b_2=100100100110$。

1 二进制信息粒的基本定义[1~3]

定义 1(粒的分解) 信息系统 $S=(U,AT=C\cup D)$,$B\in C$,若 B 的值域 $VB=\{b_1,b_2,b_3,\cdots,b_K\}$,值域的个数 $|VB|=k$,且 $|U/IND\{B\}|=k$,则 B 可按商集 $U/IND\{B\}$ 分解为 k 个二进制信息子粒。关于属性 D 则可按商集 $U/IND\{D\}$ 分解为 $|U/IND\{D\}|$ 个二进制信息子粒。

例如:按照定义 1,表 1 中属性 a 在 U 中各个对象相应的取值分别为{1,2,1,3,4,3,2,1,1,2,5,1}可分解为 5 个二进制信息粒如下:

$a_1=\{101000011001\}$,$a_2=\{010000100100\}$,$a_3=\{000101000000\}$,$a_4=\{000010000000\}$,$a_5=\{000000000010\}$。

属性 b 在 U 中各个对象相应的取值分别为 $b=\{2,3,3,2,1,3,2,3,3,2,2,3\}$可分解为 3 个二进制信息粒如下:

$b_1=\{000010000000\}$,$b_2=\{100100100110\}$,$b_3=\{011001011001\}$。

定义 2(二进制信息粒的粒度) 定义二进制信息粒 P 中 1 的个数为此信息粒的粒度记为 $|P|$。

例如:二进制信息 $a_1=\{101000011001\}$,$b_2=\{100100100110\}$的粒度分别为 5 和 6。

*)本课题获江西省科技厅项目[2006],江西省教育厅科技资助项目(赣教技[2006]31 号)资助。**徐健锋** 讲师,硕士,研究方向为粗糙集与数据挖掘。

定义 3(二进制信息粒的关联运算与二进制信息粒之间的关联度) 信息系统 $S=(U,AT=C\cup D)$,$\{C_1,C_2,\cdots,C_k\}\in C$,$c_i$ 为 C_i 的某个二进制信息子粒,则 $c_i\wedge c_j\wedge\cdots\wedge c_k$(其中$\wedge$代表二进制数的布尔"与"运算)称为粒的关联运算。其运算结果生成更小的信息粒。其粒度$|c_i\wedge c_j\wedge\cdots\wedge c_k|/|U|$称为 c_i AND c_j AND…AND c_k 同时发生的关联度 γ。其中$|U|$代表对象集合的个数。

例如:三个二进制信息粒:$a_1=\{101000011001\}$,$b_2=\{100101001101\}$,$c_3=\{110101000101\}$。进行关联运算如下所示:

$10100011001\wedge100101001101\wedge110101000101=100000000001$

$a_1\ b_2\ c_3$ 三个二进制信息粒之间的关联度 γ 为$|100000000000 1|/|U|=2/12=1/6$。

定义 4(基于二进制信息粒的决策规则一致性判断) 设 P,D 为 2 个二进制信息粒,设 P 为条件信息粒,D 为决策信息粒,条件粒 P 与决策粒 D 关联运算后得到的更小的信息粒的粒度 $P\wedge D$,决策规则 $P\rightarrow D$ 的确信度为 $\beta=|P\wedge D|/|P|$,则 if :$P\rightarrow D$ 的确信度 $\beta=1$,then $P\rightarrow D$ 是一致的或描述为真,else $P\rightarrow D$ 为不一致或描述为假。

例如:设二进制条件信息 $P=100000001001$ 与二进制决策信息粒 $D=1001001001101$ 进行关联运算 $100000001001\wedge 1001001001101=1000000001001$,可计算决策规则 $P\rightarrow D$ 的确信度为 $\beta=|P\wedge D|/|P|=1$,则 $P\rightarrow D$ 为一致的或描述为真。

设条件信息粒 $P=101000011001$ 与决策信息粒 $D=1001001001101$,进行关联运算如下:$101000011001\wedge 1001001001101=1000000001001$,可计算决策规则 $P\rightarrow D$ 确信度为 $\beta=|P\wedge D|/|P|=3/5$,则 $P\rightarrow D$ 为不一致或称为假。

2 基于二进制信息粒的数据挖掘决策规则算法

2.1 基于二进制信息粒的属性重要度算法

在数据挖掘算法中的一个重要的基础应用就是属性的重要度计算,它在许多经典算法中都占有非常重要的地位,例如:属性约简,决策树算法等应用。本文采用二进制信息粒的粒计算方法进行的属性的重要度计算,与传统的粗糙集属性重要度计算方法或信息熵算法相比具有相对简洁快速和便于程序实现的特点。

定义 5 $POS(C,D)$表示决策属性集 D 划分的等价集关于条件属性集 C 的下近似集的并集。

定义 6 $\gamma(C,D)$代表 2 个属性的依赖程度。

$\gamma(C,D)=|POS(C,D)|/|U|$

其中$|POS(C,D)|$表示该并集的元素个数。

定义 7 C 为条件属性集,D 为决策属性集,$a\in C$,属性 a 关于属性 D 的重要度定义为:$SGF(a,C,D)=\gamma(C,D)-\gamma(C-\{a\},D)$。$SGF(a,C,D)$表示 C 中缺少属性 a 后,导致不能被准确分类的对象在系统中占的比例。

2.1.1 算法描述

Step1 对信息系统 $S=(U,C\cup D)$的属性集 C 中各属性依照定义 1 进行二进制信息粒分解。属性集 D 根据属性集 D 的属性值组合也依照定义 1 进行二进制信息粒分解。

例如:条件属性 $C=\{C_1,C_2,C_3\}$,决策属性集合 D 并且 C 中各属性都只有 2 个属性值,D 属性集也只有 2 个属性值组合。

则 C 分解为:$c_{11},c_{12},c_{21},c_{22},c_{31},c_{32}$ 6 个二进制信息粒。D 分解为:d_1,d_2, 2 个二进制信息粒。

Step2 在各条件属性生成的各二进制信息粒集合中各提取一个二进制信息粒进行关联计算,并生成更小的粒。按照定义 4 用它与决策属性得到的二进制信息粒进行一致性判断:如有一致的,则计算其条件子粒的粒度。

Step3 在各条件属性生成的各二进制信息粒集合中的各提取一个二进制信息粒组成的所有组合重复 Step2 的步骤。

Step4 计算$|POS(C,D)|$。$|POS(C,D)|$等于 Step2 和 Step3 得到所有粒度的和。

Step5 计算 C 和 D 的依赖程度。

$\gamma(C,D)=|POS(C,D)|/|U|$

Step6 计算 C 分别除去某一个属性后的集合与和 D 的依赖程度:

设属性 $a\in C$ 重复 Step1－Step5 计算 $C-\{a\}$和 D 的依赖程度。

Step6 根据定义 7 计算条件属性 $a\in C$ 关于属性 D 的重要度 $SGF(a,C,D)=\gamma(C,D)-\gamma(C-\{a\},D)$,以及同样的方法计算其他各条件属性关于属性 D 的重要度。

2.1.2 算法分析

本算法的时间复杂度最差情况为 K^n 与分明矩阵方法等经典方法相同。但是当某些条件属性只考虑单一属性的情况,则时间复杂可减少为 K^{n-m},而同情况使用分明矩阵方法仍为 K^n。可见本算法在某些情况下优于分明矩阵方法。同时由于二进制计算的潜在优势,在实际应用中当 U 中对象数量较大而 C 的个数不太大的实际情况,本算法的也具有一定的优势。

2.1.3 实例 1

计算某信息系统各条件属性的重要度。

设某关于气候的信息系统:

A 代表天气:取值为晴(1),多云(2),雨(3)

B 代表气温:取值为冷(1),适中(2),热(3)

C 代表湿度:取值为高(1),正常(2)

D 代表风:取值为有(1),无(2)

E 代表气候类别 取值有 N,P。

表 2 气候的信息系统

A	B	C	D	E	A	B	C	D	E
1	1	1	2	N	1	2	1	2	N
1	1	1	1	N	1	3	2	2	P
2	1	1	2	P	3	2	2	2	P
3	2	1	2	P	1	2	2	1	P
3	3	2	2	P	2	2	1	1	P
3	3	2	1	N	2	1	2	2	P
2	3	2	1	P	3	2	1	1	N

Step1 对信息系统 $S=(U,\underline{C}\cup\underline{D})$的属性集 $\underline{C}\cup\underline{D}$ 中各属性依照定义 1 进行二进制信息粒分解(注:$\underline{C}=\{A,B,C,D\}$,$\underline{D}=\{E\}$)。

Step2 在各条件属性生成的各二进制信息粒集合中各提取一个二进制信息粒进行关联计算,并生成更小的粒。

Eg: $a_1\wedge b_1\wedge c_1\wedge d_1=a_1b_1c_1d_1=\{01000000000000\}$。

根据定义 4 :$a_1b_1c_1d_1\rightarrow e_1$ 为真,同时可计算出 $a_1b_1c_1d_1$ 的粒度$=|a_1b_1c_1d_1|=1$。

Step3 重复 Step2,解出所有二进制信息粒进行关联计算的组合为真的有 $a_2b_1c_1d_2\rightarrow e_1$;$a_1b_1c_1d_1\rightarrow e_1$;$a_2b_1c_1d_2\rightarrow e_2$;$a_3b_2c_1d_2\rightarrow e_2$;$a_3b_3c_2d_2\rightarrow e_2$;$a_3b_3c_2d_1\rightarrow e_1$;$a_2b_3c_2d_1\rightarrow e_2$;$a_1b_2c_1d_2\rightarrow e_1$;$a_1b_3c_2d_2\rightarrow e_2$;$a_3b_2c_2d_2\rightarrow e_2$;$a_1b_2c_2d_1\rightarrow e_2$;

$a_2b_2c_1d_1\rightarrow e_2$；$a_2b_1c_2d_2\rightarrow e_2$；$a_3b_2c_1d_1\rightarrow e_1$；

A			B			C		D		E	
a_1	a_2	a_3	b_1	b_2	b_3	c_1	c_2	d_1	d_2	e_1	e_2
1	0	0	1	0	0	1	0	0	1	1	0
1	0	0	1	0	0	1	0	1	0	1	0
0	1	0	1	0	0	1	0	0	1	0	1
0	0	1	0	1	0	1	0	0	1	0	1
0	0	1	0	0	1	0	1	0	1	0	1
0	0	1	0	0	1	0	1	1	0	1	0
0	1	0	0	0	1	0	1	1	0	0	1
1	0	0	0	1	0	1	0	0	1	1	0
1	0	0	0	0	1	0	1	0	1	0	1
0	0	1	0	1	0	0	1	0	1	0	1
1	0	0	0	1	0	0	1	1	0	0	1
0	1	0	0	1	0	1	0	1	0	0	1
0	1	0	1	0	0	0	1	0	1	0	1
0	0	1	0	1	0	1	0	1	0	1	0

Step4 计算条件属性集$\underline{C}$和决策属性集$\underline{D}$的依赖程度。

$|POS(\underline{C},\underline{D})|=|a_2b_1c_1d_2|+|a_1b_1c_1d_1|+|a_2b_1c_1d_2|+|a_3b_2c_1d_2|+|a_3b_3c_2d_2|+|a_3b_3c_2d_1|+|a_2b_3c_2d_1|+|a_1b_2c_1d_2|+|a_1b_3c_2d_2|+|a_3b_2c_2d_2|+|a_1b_2c_2d_1|+|a_2b_2c_1d_1|+|a_2b_1c_2d_2|+|a_3b_2c_1d_1|=14$

$\gamma(\underline{C},\underline{D})=|POS(C,D)|/|U|=1$

Step5 计算$\underline{C}-\{A\}$和$\underline{D}$的依赖程度。

$\gamma(\underline{C}-\{A\},\underline{D})=|POS(\underline{C}-\{A\},\underline{D})|/|U|=\{|b_1c_1d_1|+|b_3c_2d_2|+|b_2c_2d_2|+|b_2c_2d_1|+|b_1c_2d_2|\}/|U|=3/7$

Step6 条件属性A关于决策属性E的重要度。

$SGF(A,\underline{C},\underline{D})=\gamma(\underline{C},\underline{D})-\gamma(\underline{C}-\{A\},\underline{D})=4/7$

Step7 方法同上可分别计算出条件属性B,C,D关于决策属性集$\underline{D}$的重要度。

$SGF(B,\underline{C},\underline{D})=\gamma(\underline{C},\underline{D})-\gamma(\underline{C}-\{B\},\underline{D})=0$

$SGF(C,\underline{C},\underline{D})=\gamma(\underline{C},\underline{D})-\gamma(\underline{C}-\{C\},\underline{D})=0$

$SGF(D,\underline{C},\underline{D})=\gamma(\underline{C},\underline{D})-\gamma(\underline{C}-\{D\},\underline{D})=1-5/7=2/7$

2.2 二进制信息粒决策树构造算法

2.2.1 算法描述

Step1 信息系统$S=(U,AT=C\cup D)$。首先计算各条件属性重要度，可先采用二进制粒的属性重要度算法计算各条件属性重要度，如相同则利用信息论原理计算各条件属性的互信息[5]。

Step2 根据定义1将D按商集$U/IND\{D\}$分解为$J=|U/IND\{D\}|$个二进制决策信息子粒。$d_1,d_2,\cdots d_k$。

Step3 同样将互信息最大的条件属性C_i也分解为$K=|U/IND\{C_i\}|$个二进制条件信息子粒$a_1,a_2,\cdots a_k$。其余的条件属性留下待用。

Step4 根据定义4分别进行各条件信息子粒与各决策信息子粒关联运算，并判断$a_i\rightarrow d_j$是否为真。($a_i\in\{a_1,a_2,\cdots,a_k\}$，$d_j\in\{d_1,d_2,\cdots,d_k\}$)，为真的决策规则提出来，为获得的一条决策规则。为假的条件信息子粒留下待用，如全为真则可结束本算法。

Step5 根据余下的条件属性粒决定的各个等价类，分别计算余下的条件属性的重要度(不考虑本条件属性粒在前面Step4用过的条件属性)，每个类中重要度最大的C_j按商集$U/IND\{C_j\}$分解为$L=|U/IND\{C_j\}|$个二进制条件信息子粒$b_1,b_2,\cdots b_L$。

Step6 将Step4不能提取为决策规则余下的各二进制条件信息子粒与Step5对应得到的对于二进制条件信息子粒分别作关联运算$a_i\wedge b_j$，即获得一组新的条件关系子粒。并计算各新的条件关系子粒的关联度γ，关联度$\gamma<\&$删除，如全被删除则算法结束，否则goto Step4(&值可根据情况确定)。

2.2.2 实例2

采用二进制信息粒决策树构造算法从实例1气候的信息系统中提取决策规则。

Step1 对表2气候的信息系统计算方法可得各条件属性的重要度为：I(天气)=4/7、I(气温)=0、I(湿度)=0、I(风)=2/7。

Step2 决策属性E可分解为2个决策属性子粒：N：11000101000001，P：00111010111110。

Step3 将重要度最大的条件属性天气也分解为3个条件属性子：

A_1：110000011010 00，A_2：00100010000110，A_3：00011100010001。

Step4 根据定义4分别计算$A_1\rightarrow N$，$A_1\rightarrow P$，$A_2\rightarrow N$，$A_2\rightarrow P$，$A_3\rightarrow N$，$A_3\rightarrow P$的确信度只有$A_2\rightarrow P$的确信度为1：由此可判断$A_2\rightarrow P$为提取的第一条决策规则。A_2可除去。

Step5 根据A_1表达的等价类，首先计算重要度，结果相同。然后计算余下的互信息，得最大的条件属性是C。

根据A_3表达的等价类，首先计算重要度，结果相同。然后计算余下的互信息，得最大的条件属性是D。

Step6 将5求到的互信息最大的条件属性D也分解为3个条件属性子粒：

D_1：01000110001101，

D_2：10111001110010，并分别与A_3作关联运算，

$A_3\wedge D_1\rightarrow N$的确信度为1，(00011100010001∧01000110001101=00000100000001→11000101000001)。

$A_3\wedge D_2\rightarrow P$的确信度为1，由此可判断$A_3\wedge D_1\rightarrow N$，$A_3\wedge D_2\rightarrow P$为提取的决策规则。

Step7 将5求到的互信息最大的条件属性C也分解为3个条件属性子粒，C_1：11110001000101，C_2：00001110111010。并分别与A_1作关联运算$A_1\wedge C_1\rightarrow N$的确信度为1，$A_1\wedge C_2\rightarrow P$的确信度为1，由此可判断$A_1\wedge C_1\rightarrow N$，$A_1\wedge C_2\rightarrow P$为提取的决策规则。

Step8 最后可获得5条决策规则：

1：$A_2\rightarrow P$，2：$A_3\wedge D_1\rightarrow N$，3：$A_3\wedge D_2\rightarrow P$，4：$A_1\wedge C_1\rightarrow N$，5：$A_1\wedge C_2\rightarrow P$。

结束语 近年来二进制信息粒的研究已经有了很大的发展。本文给出二进制信息粒的一些典型的定义，提出了新型的基于二进制信息粒的属性重要度算法和二进制信息粒决策树构造算法，显示了其在数据挖掘领域的较广阔应用前景。上述二进制信息粒决策树构造算法较传统的ID3算法相比有效地提高了算法执行效率，克服了ID3算法的某些固有缺点，也给出了一种新的剪枝方法，在大数据量计算时将显示其优势。但由于篇幅的限制笔者将在后续的文章中发表这方面的进一步研究。可以相信这种二进制信息粒应用于快速挖掘算法的研究与应用仍将是一个重要的研究方向。

参考文献

1 刘清. Rough集及Rough集推理[M]. 北京：科学出版社，2001. 100～116

2 刘斓，刘清. 基于粒的二进制运算的关联规则提取方法[J]. 南昌大学学报(理科版)，2003，27(1)：98～101

3 Liu Qing，Jiang S L. Reasoning about Information Granules Based on Rough Logic. In：RSCTC 2002，LNAI 2475，2002. 139～143

4 刘宇熹. Rough集决策理论及其在KDD中的应用[D]：[南昌大学硕士学位论文]. 2003，6：34～36

5 陈文伟. 数据仓库与数据挖掘 [M]. 人民邮电出版社，2004