

二进制粒属性相关判定与应用

王红霞^{1,2}, 陈俊杰¹, 王志伟³

(1 太原理工大学 计算机科学与技术学院, 山西 太原 030002;

2 山西水利职业技术学院, 山西 太原 030027;

3 山西省气候中心, 山西 太原 030002)

摘 要: 采用二进制粒表示信息系统中的离散化属性值, 即对属性进行二进制粒化, 定义了基于二进制信息粒的匹配、合取及析取算子, 描述了二进制粒属性相关的定义, 提出了基于二进制粒的属性相关性判定矩阵的算法, 算法通过判定信息系统中属性的相关性, 为属性约简及关联规则的求解提供一种新方法. 如果相关属性同为条件属性, 则可进行约简, 选择其一; 如果有决策属性, 则可生成强关联规则; 该算法在气象数据的仿真实验中验证了其适用性.

关键词: 二进制; 粒度; 算子; 匹配矩阵; 属性相关; 气象数据

中图分类号: TP391

文献标识码: A

文章编号: 1000-7180(2012)02-0065-05

Application and Determination Attribute Correlation Trough Binary Granular

WANG Hong-xia^{1,2}, CHEN Jun-jie¹, WANG Zhi-wei³

(1 College of Computer Science and Technology, Taiyuan University of Technology, Taiyuan 030002, China;

2 Shanxi Conservancy Technical College, Taiyuan 030027, China;

3 Shanxi Climate Centre, Taiyuan 030006, China)

Abstract: By using the binary granular to show discretization attributes in the information system, this paper makes definitions for the binary granular matching, conjunction and disjunction operators, and the binary granular attribute correlation. This paper also discusses how to determine the matrix based on the binary granular attribute correlation, which is the new way to solve attribute reduction and association rules. If the correlation attribute is also the condition attribute, then reduction can be carried on. If there is the decision attribute, strong association rules are formed. This kind of algorithm is tested in the simulation experiment of meteorological data.

Key words: binary; granularity; algorithm operator; matching matrix; attribute correlation; meteorological data

1 引言

知识约简在 Rough 集理论中是非常重要的概念, 包括属性约简和属性值约简^[1-2]; 粒计算是一种看待客观世界的世界观和方法论, 它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, 即将原来的空间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^[3]; 二进制数存储效率高运算速度快.

文献[4-6]分别将粒计算与二进制运算相结合, 应用于不同算法中. 本文在粒与二进制数的对应关系上, 定义了二进制粒集的相关和属性相关的判定(匹配)矩阵, 并给出了相应的属性值相关判定算法. 该方法通过二进制数的异或、与、或等运算及匹配矩

收稿日期: 2011-05-11; **修回日期:** 2011-07-23

基金项目: 中国气象局技术推广项目(CMATG2006M26); 山西省水利厅项目(2130317-19); 山西省科技攻关项目(2007031106-1)

阵来判断属性之间的相关度,完全相关或匹配的属性可被约简或是强关联规则,不仅运算速度快,而且存储效率高,易于编程实现。

2 二进制粒相关定义

设 $S=(U,A)$ 是一信息系统,其中 $U=\{u_1, u_2, u_3, \dots, u_n\}$ 是个体域, $A=\{a_1, a_2, \dots, a_m\}$ 是属性集. 任意子集 $B \subseteq A$ 均可将 U 划分成互不相交等价类. 等价类被定义为粒. 对于论域中的每个信息粒用 n 位二进制串表示,其中 n 为论域的基数. 若个体属于该信息粒,则在二进制串中相应第 i 位置 1, 否则将相应位置 0, 即该个体不属于该信息粒,二进制中 1 的个数定义为该信息粒的大小^[4-6].

表 1 可数字化表示为表 2.

表 1 商务信息系统的知识表示(玩具)

U	颜色	大小	感觉	接触面	材料	价格
1	蓝	大	硬	不平滑	塑料	便宜
2	红	小	适度	平滑	长毛绒	贵
3	黄	小	软	粗糙	长毛绒	贵
4	蓝	中	软	粗糙	塑料	便宜
5	黄	小	硬	不平滑	塑料	便宜
6	绿	大	适度	平滑	木材	贵
7	黄	小	硬	不平滑	金属	贵
8	绿	小	适度	平滑	木材	中等

表 2 数字化表示的表 1

U	a	b	c	d	e	f
1	1	3	1	2	3	1
2	2	1	2	1	1	3
3	3	1	3	3	1	3
4	1	2	3	3	3	1
5	3	1	1	2	3	1
6	4	3	2	1	2	3
7	3	1	1	2	4	3
8	4	1	2	1	2	2

表 2 中:颜色— a ,大小— b ,感觉— c ,接触面— d ,材料— e ,价格— f ;数字信息:颜色 a : 蓝=1, 红=2, 黄=3, 绿=4; 大小 b : 大=3, 中=2, 小=1; 感觉 c : 硬=1, 适度=2, 软=3; 接触面 d : 平滑=1, 不平滑=2, 粗糙=3; 材料 e : 塑料=3, 木材=2, 长毛绒

=1, 金属=4; 价格 f : 便宜=1, 中等=2, 贵=3.

表中对应的信息系统 $S=(U, AT)$, $U=\{1, 2, 3, \dots, 8\}$ 为对象集合, $AT=\{a, b, c, d, e, f\}$ 为属性集. 任一属性均可以将 U 划分成互不相交的等价类, 即信息粒. 商集 $U/IND\{a\}$, 其中, $c_1=\{1, 5, 7\}$, $c_2=\{2, 6, 8\}$, $c_3=\{3, 4\}$, 每个元素就是一个信息粒, a_1, a_2 等为信息粒的名称. 每个信息粒二进制化后得相应的二进制粒.

基于信息系统 $S=(U, AT|AT=C \cup D)$, 引入概念:

定义 1 (论域的势) 论域 U 的个体数 l , 记 $|U|=l$.

表 2 中论域势为 8.

定义 2 (二进制粒) 若第 i 个体属于某信息粒, 则在二进制串中相应第 i 位置 1, 否则将相应位置 0, 即该个体不属于该信息粒, 生成二进制粒.

表 2 中, $a_1=10010000$, $c_1=10001010$, $c_2=01010101$, $c_3=00100000$, $d_1=01000101$, $e_4=00000010$.

定义 3 (二进制信息粒的粒度) 二进制信息粒 P 中二进制值 1 的个数定义为信息粒的粒度, 记 $|P|$.

表 2 中, $|a_3|=|00101010|=3$

定义 4 (二进制信息粒的匹配算子) $\{C_1, C_2, \dots, C_k\} \in AT$, c_i 为 C_i 的某个二进制信息粒, 则 $c_i \text{ xor } c_j \text{ xor } \dots \text{ xor } c_k$ 定义为信息粒的匹配运算. 其中: xor 为二进制粒按位“异或”运算, 也可表示为“ $\oplus$ ”.

c_2, d_1 匹配运算: $c_2 \oplus d_1 = 00000000$.

定义 5 (二进制信息粒绝对相关度) 信息粒绝对相关度 d_a 定义为 U 的势与信息粒匹配运算后的粒度差: $|U| - |c_i \text{ xor } c_j \text{ xor } \dots \text{ xor } c_k| = l - |c_i \text{ xor } c_j \text{ xor } \dots \text{ xor } c_k|$.

表 2 中 $d_a(c_2 \oplus d_1) = |U| - |c_2 \oplus d_1| = 8 - 0 = 8$.

定义 6 (二进制信息粒相对相关度) 信息粒相对相关度 d_r 定义为 $d_a/|U|$.

$d_r(c_2 \oplus d_1) = d_a(c_2 \oplus d_1)/|U| = 8/8 = 1$

定义 7 (二进制信息粒的析取运算) $\{C_1, C_2, \dots, C_k\} \in AT$, c_i 为 C_i 的某个二进制信息粒, 则 $c_i \wedge c_j \wedge \dots \wedge c_k$ 定义为信息粒的析取运算. 其中: $\wedge$ 为二进制粒按位“与”运算.

定义 8 (二进制信息粒的合取运算) $\{C_1, C_2, \dots, C_k\} \in AT$, $c_i, c_j, \dots, c_k$ 为 C_i 的二进制信息粒, 则 $c_i \vee c_j \vee \dots \vee c_k$ 定义为信息粒的合取运算. 其中: $\vee$ 为

二进制粒按位“或”运算。

表2中, $a_1 \wedge c_1 = 10000000$, $a_1 \vee a_2 = 11010000$ 。

3 基于二进制信息粒的属性相关性判定

3.1 属性相关性定义

信息系统 $S = (U, AT | AT = C \cup D)$, $\{C_1, C_2, \dots, C_k\} \in C$, 如果:

$$\forall c_{im} \in c_i, \exists c_{jm} \in c_j, \dots c_{im} \dots c_{jm} \dots = u_k \in U,$$

即属性列之间存在对应关系, 称存在这种对应关系的属性为相关属性。

表2中属性 c 的值和属性 d 的值存在对应关系: 1—2, 2—1, 3—3, 凡是属性 c 值为1所在的行, 属性 d 的值与之对应的值为2, 2对应1, 3对应3, 所以属性 c 和属性 d 为相关(匹配)属性。

对应二进制粒的匹配运算: $c_1 \oplus d_2 = c_2 \oplus d_1 = c_3 \oplus d_3 = 00000000$, 即3对信息粒之间的绝对匹配度为论域的势8, 相对匹配度为1。

3.2 二进制信息粒属性相关判定矩阵

需要判定相关性的属性, 离散化后离散量个数可能相同, 也可能不同, 且可能是多个属性之间判定其相关性, 以下分三种情况, 分别进行讨论。

3.2.1 属性值离散量个数相等的相关判定矩阵

判定矩阵也可称匹配矩阵。

$$|A_{m \times l} \oplus B'_{m \times l}| = \begin{vmatrix} a_1 \\ \vdots \\ a_m \end{vmatrix} (b_1 \quad \dots \quad b_m) \\ = \begin{vmatrix} |a_1 \oplus b_1| & |a_1 \oplus b_2| & \dots & |a_1 \oplus b_m| \\ \vdots & \vdots & \ddots & \vdots \\ |a_m \oplus b_1| & |a_m \oplus b_2| & \dots & |a_m \oplus b_m| \end{vmatrix}$$

表2: c 和 d 属性生成的判定矩阵结果为

$$\begin{vmatrix} 6 & 0 & 5 \\ 0 & 6 & 5 \\ 5 & 5 & 0 \end{vmatrix} \text{ 行列变换后为 } \begin{vmatrix} 0 & 6 & 5 \\ 6 & 0 & 5 \\ 5 & 5 & 0 \end{vmatrix} \text{ 为一对称矩阵,}$$

对角线元素为0。

定理1 判定矩阵中对角线元素为零的属性, 必是相关(匹配)属性。

推论: 属性判定矩阵经过行列变换后, 对角线元素为0, 则属性也必是相关(匹配)属性。

证明: 矩阵经过行列变换后性质不变。

3.2.2 属性值离散量个数不相等的相关判定

表3中属性 a 的离散量个数为4, 属性 b 的离散量个数为2, 判定该类属性值离散量个数不等的相关性判定算法步骤如下:

表3 $S = (U, AT)$

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

步骤:

step1 合取属性值多的二进制粒

对同一属性的二进制信息粒进行合取运算, 按位“或”, 信息粒的个数与少的一致;

Step2 计算相关判定矩阵;

Step3 如果不是相关属性, 即变化后对角线元素不为0, 则转 step1, 选择下一组合, 直至判定完所有组合或得解。

判定表3中属性 a 和 b 的相关性:

属性 a 有四个信息粒, 属性 b 有两个信息粒;

$$a_1 = 101000011000, a_2 = 010001100001,$$

$$a_3 = 000100000010, a_4 = 000010000100;$$

$$b_1 = 000110000110, b_2 = 1110011111001.$$

将 a 的四个粒进行合取, 生成两个粒。

显然: $a_1 \vee a_2 = 101000011000 \vee 010001100001 = 1110011111001$, $a_3 \vee a_4 = 000100000010 \vee 000010000100 = 000110000110$; 与 b 属性进行相关判定矩阵的计算, 得 a 与 b 相关。

属性 a 合取为两个信息粒的组合有: $(a_1 \vee a_3, a_2 \vee a_4)$, $(a_1 \vee a_2 \vee a_3, a_4)$, $(a_1 \vee a_2 \vee a_4, a_3)$, $(a_1 \vee a_3 \vee a_4, a_2)$, $(a_2 \vee a_3 \vee a_4, a_1)$

3.2.3 多属性的相关判定

表4中如需判定属性 a 、 b 和属性 c 的相关性, 即判定多个属性值的相关性。

表4 $S = (U, AT)$

U	a	b	c	U	a	b	c
1	1	2	0	7	2	2	0
2	2	1	1	8	2	1	0
3	1	1	1	9	1	2	0
4	2	2	0	10	1	1	1
5	1	2	0	11	2	1	1
6	1	2	0	12	2	2	0

其判定算法步骤如下:

step1 析取二进制粒,生成更小粒度的信息粒,即进行属性二进制信息粒的析取运算,按位“与”,生成新信息粒的粒度小于原来的信息粒度;

Step2 如果析取后属性粒之间的关系满足:

3.2.1 节中的属性值的离散量个数相等,则计算相关判定矩阵,否则,依据 3.2.2 节中的属性值的离散量个数不相等进行求解.

表 3 中属性 a 、 b 与属性 c 的相关性判定:

属性信息粒分别为: $a_1 = 101011001100$, $a_2 = 010100110011$; $b_1 = 011000010110$, $b_2 = 100111101001$; $c_1 = 100111111001$, $c_2 = 011000000110$.

属性间通过析取运算,生成粒度更小的二进制粒.

$a_1 \wedge b_1 = 001000000100$, $a_1 \wedge b_2 = 100011001000$, $a_2 \wedge b_1 = 010000010010$, $a_2 \wedge b_2 = 000100100101$.

组合属性有四个粒, c 属性 2 个粒,所以组合属性再进行合成运算后进行相关判定矩阵计算.

3.2.4 算法描述

基于二进制粒的属性相关性判定算法描述如下:

输入:相关分析的属性
输出:属性相关判定矩阵的结果
算法步骤:

Step1 属性二进制粒化;

Step2 如果多属性匹配,则相应属性的二进制粒合取,生成新二进制粒集;

Step3 判断二进制粒集中属性离散化量个数是否相同;

Step4 如果粒数相同,求解匹配矩阵;
否则,析取运算后生成相同的粒数求解;

Step5 对匹配矩阵进行行列变换,如果对角线元素为 0,则满足匹配条件,输入的待判定属性相关.

算法求解应用:

表 2 中输入属性 c 和 d ,结果匹配即两属性相关.
表 3 中属性 a 和 b 相关.
表 4 中属性 a 、 b 合取后和 c 属性相关.

4 仿真实验

本实验数据集来源于山西省气候中心,提供的山西省 109 观测站的气象资料主要要素值,其中包括:日最高气温、日最低气温、日平均气温、日照时数、日降水等,

实例 1 不同时域粒度的属性构造及相关性判定.

通过对气象站点安泽的要素值最高气温、最低气温的聚合运算,分别求得不同时间粒度:年、冬季(1,2,12 月)、夏季(6,7,8 月)四个聚合值,即求解 $\max(\text{最高气温,年})$, $\min(\text{最低气温,年})$, $\max(\text{最高气温,夏季})$, $\min(\text{最低气温,冬季})$

采用模糊 C—均值离散化后,分别对应属性: b , c , d , e , 其中 a 属性对应为年份. 构成的属性集见表 5.

表 5 不同时域粒度属性集构成 $S=(U,AT)$

a	b	c	d	e	a	b	c	d	e
1990	3	1	3	1	2000	3	2	3	2
1991	3	3	3	3	2001	3	4	3	4
1992	3	4	3	4	2002	5	2	5	2
1993	2	2	2	2	2003	1	2	1	2
1994	2	4	2	4	2004	2	4	2	4
1995	3	4	3	4	2005	5	3	5	3
1996	2	4	2	4	2006	3	4	3	4
1997	4	4	4	4	2007	3	5	3	5
1998	3	1	3	1	2008	2	2	2	2
1999	4	3	4	3	2009	4	3	4	3

二进制粒化后,分别以属性对 (b,c) , (b,d) , (b,e) , (c,d) , (c,e) 求解判定矩阵后得 (b,d) , (c,e) 为匹配矩阵,所以, b 和 d 属性, c 和 e 属性为相关属性.

结果解释与分析:

实际物理意义: b 属性对应年最高气温, d 属性对应夏季最高气温; c 属性对应年最低气温, e 属性对应冬季最低气温;两组属性相关性验证在四季气候特征明显的北方山西省,夏季最高气温即等同于年度最高气温,冬季最低气温等同于年度最低气温. 该研究方法对于空域气候相似,局地小气候相似提供新的方法,为作物引种和现代农业规划提供科学依据,具有一定的实用价值. 该结果还可应用于当地防御农业气象灾害,如高温热害、低温冷害等也有一定的借鉴作用.

实例 2 不同空域粒度的属性构造及相关性判定.

通过对山西省全省、运城地区、大同地区等不同空间区域,不同时间粒度年度、夏季、冬季等气象要素值最高气温、最低气温的聚合运算,求得不同时空粒度的属性值:年、冬季(1,2,12 月)、夏季(6,7,8

月)、山西省、运城地区、大同地区等共12个聚合值,即求解 $\max(\text{最高气温,年,空域}),\min(\text{最低气温,年,空域}),\max(\text{最高气温,夏季,空域}),\min(\text{最低气温,冬季,空域})$.其中参数空域值为:山西、运城、大同.

采用模糊C—均值离散化后,分别对应属性: b,c,d,e (表6),构成的属性集见表7.

表6 属性名称对照表

a	b	c	d	e
年份	全省年最高	全省年最低	运城年最高	大同年最低

表7 不同空域粒度属性集构成 $S=(U,AT)$

U	a	b	c	d	e	U	a	b	c	d	e
1	1972	5	2	5	2	20	1991	4	4	4	4
2	1973	4	4	4	4	21	1992	3	4	3	4
3	1974	3	3	3	3	22	1993	3	3	3	3
4	1975	3	4	3	4	23	1994	4	4	4	4
5	1976	3	4	3	4	24	1995	4	4	4	4
6	1977	4	4	4	4	25	1996	3	4	3	4
7	1978	3	3	3	3	26	1997	5	3	5	3
8	1979	4	5	4	5	27	1998	5	2	5	2
9	1980	3	4	3	4	28	1999	3	5	3	5
10	1981	4	3	4	3	29	2000	4	2	4	2
11	1982	2	4	2	4	30	2001	3	4	3	4
12	1983	1	3	1	3	31	2002	3	2	3	2
13	1984	1	3	1	3	32	2003	2	3	2	3
14	1985	4	3	4	3	33	2004	3	3	3	3
15	1986	3	3	3	3	34	2005	5	3	5	3
16	1987	2	5	2	5	35	2006	5	4	5	4
17	1988	3	3	3	3	36	2007	3	5	3	5
18	1989	3	4	3	4	37	2008	3	3	3	3
19	1990	3	5	3	5	38	2009	4	4	4	4

二进制粒化后,分别以属性对 $(b,c),(b,d),(b,e),(c,d),(c,e)$ 求解判定矩阵后得 $(b,d),(c,e)$ 为匹配矩阵,所以 b 和 d 属性, c 和 e 属性为相关属性.

结果解释与分析:
实际物理意义:受地理位置海拔高度等影响,山西省南暖北寒,全省的最高气温出现在南部运城地区,最低气温出现在北部朔州、大同,属性相关性判定结果与实际经验及实际数据,相互得到验证.
属性相关性判定可以为属性约简、离群数据挖

掘等提供一种新方法,同时也为气象数据中缺失数据的补空提供了一种新技术.

5 结束语

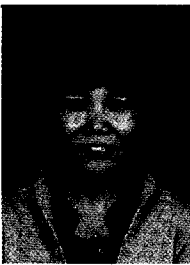
本文讨论了二进制粒在属性值相关上的应用,求解得属性相关性后,如果两者都为条件属性,则可两者择其一,进行相关数据挖掘算法的计算;如果其中一属性为决策属性,则得到强关联规则.通过二进制的异或、与和或运算及匹配矩阵来进行属性相关的判断,方法简单易行,快速可靠,在需要进行属性构造的信息系统中,运用该方法可快速进行原属性与构造后属性的相关判定,然后进行冗余属性的约简或在相关属性间进行预测.

文中算法在可应用于如气象等具有不同粒度层或时空特性明显的知识领域.在气象领域中,数据要素通过不同粒度的构造,生成属性可能与原属性有很强的相关性,文中的属性匹配方法,通过气象领域的应用,验证了算法的可行性及实用性.

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] Liu Qing, Jiang S L. Reasoning about information granules based on rough logic[C] // RSCTC 2002, LNAI 2475. USA: Malvern, 2002:139-143.
[3] Lin T Y. Granular computing, announcement of the BISC Special Interest Group on Granular Computing. 1997.
[4] 刘炯,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.
[5] 霍岩,张树雄,马垣. 基于粒的二进制运算的属性值约简[J]. 计算机与信息技术, 2006(1):29-30.
[6] 谢珺. 二进制粒神经网络研究及其在故障诊断中的应用[D]. 山西:太原理工大学,2009.

作者简介:



王红霞 女,(1966—),博士研究生,副教授. 研究方向为智能信息处理.

陈俊杰 男,(1956—),博士生导师,教授. 研究方向为数据库、智能信息处理.