基于秩相关的属性约简

白 江,魏立力
BAI Jiang,WEI Li-li

宁夏大学 数学计算机学院,银川 750021
School of Mathematics & Computer Science,Ningxia University,Yinchuan 750021,China
E-mail:baijiang.com@163.com

BAI Jiang,WEI Li-li.Attribute reduction based on rank correlation.Computer Engineering and Applications,2009,45(6):66-68.

Abstract: In this paper,with respect to the ordinal variables,the definitions of attribute reduction of dominance-based rough set approach in ordered information systems and ordered decision tables are given.Then the approach of sorting objects comprehensively using rough set theory is brought in to obtain a series of ranks.Based on these ranks,nonparametric methods to analyze correlation between two attribute subsets are introduced,and Spearman rank correlation coefficient is used as a measure of attribute correlation.Based on this measure,a new method of attribute reduction of ordered information systems and ordered decision tables is presented without changing the ordinal information of the universe.Finally,the experiments show that the approach proposed is feasible and it provides a statistical evidence for rough set approach.

Key words: rough set;dominance relation;rank correlation;attribute reduction

摘 要:针对有序尺度变量,给出了有序信息系统与有序决策表在优势关系下的粗糙集约简定义;利用粗糙集方法将对象综合排序,进而得到一组秩;根据这些秩,运用非参数统计的思想研究了两个属性子集之间的相关性,并将 Spearman 秩相关系数作为属性相关性度量;在不改变总体序信息情况下,给出了基于此度量对有序信息系统与有序决策表进行约简的新方法。最后通过数值例子说明该方法是可行的,且为粗糙集方法提供了统计依据。

关键词:粗糙集;优势关系;秩相关;属性约简
DOI:10.3778/j.issn.1002-8331.2009.06.020 **文章编号:**1002-8331(2009)06-0066-03 **文献标识码:**A **中图分类号:**TP18

1 引言

粗糙集理论是波兰数学家 Z.Pawlak 于 1982 年提出的一种处理不精确、不确定与不完整数据的新的数学理论[1]。它以等价分类为基础定义了不确定性，在数据挖掘中有着广泛的应用。属性约简是粗糙集理论的核心问题之一,也是一个 NP 难问题。目前基于粗糙集方法的属性约简文献颇多,但使用粗糙集理论对信息系统进行分析的一个明显缺陷就是缺乏统计依据[2],因此在粗糙集中引入统计学方法已势在必行。魏玲等在文献[3]中指出,决策表中的对象是作为研究总体的样本来进行分析的,其分析结果如果没有统计检验作保证的话,就不能认为该分析反映了总体的信息。但该文单纯分析了单个条件属性与决策属性的关系其局限性太大。因此,发展适当的方法分析属性的综合作用对决策属性的影响，将是非常有意义的工作,这方面的工作已有文献[4-5]等。

经典粗糙集理论是以等价分类为基础,可用以处理名义尺度变量。然而现实数据中经常会出现有序情况,例如学生成绩有优、良、中、差之分,岩石的基性程度可分为超基性、基性、中性和酸性,人们的年龄可分为少年、青年、中年和老年等。对于这种有序尺度变量，如果不充分利用数据本身的序关系的话,就会损失有用信息。针对有序尺度变量,在不改变对象间的序信息情况下,给出了有序信息系统与有序决策表在优势关系下的粗糙集约简定义;利用粗糙集理论中的综合优势度,将对象综合排序,进而得到一组秩;根据所得的秩,从非参数统计思想出发，采用假设检验方法研究了两个属性子集间的相关性,并将 Spearman 秩相关系数作为属性相关性的度量。基于此度量提出了一种新的属性约简方法，且该方法不改变总体的序信息,使得粗糙集分析方法有了统计依据。

2 粗糙集理论与约简方法

定义 1[6] 称(U,A,F)为一个信息系统,其中有限集合 U 为对象集,即 $U=\{x_1,x_2,\cdots,x_n\}$。U 中的每个 $x_i(i\leqslant n)$,称为一个对象。而 A 为属性集,即 $A=\{a_1,a_2,\cdots,a_m\}$。A 中的每个 $a_j(j\leqslant m)$,称为一个属性。F 为 U 和 A 的关系集,即 $F=\{f_j:j\leqslant m\}$。其中 $f_j:U\rightarrow V_j(j\leqslant m)$,$V_j$ 为属性 a_j 的值域。

基金项目:国家自然科学基金(the National Natural Science Foundation of China under Grant No.60663003);宁夏自然科学基金(the Natural Science Foundation of Ningxia of China under Grant No.NZ0725)。

作者简介:白江(1983-),男,硕士研究生,主研方向为应用概率统计;魏立力(1965-),男,通讯作者,教授,主研方向为统计学、人工智能的数学基础。

收稿日期:2008-01-29 **修回日期:**2008-03-31

定义 2[6] 称(U,A,F,D,G)为目标信息系统或决策表,其中(U,A,F)是信息系统,A 称为条件属性集,D 称为目标属性集或决策属性集,即 $D=\{d_1,d_2,\cdots,d_p\}$。G 为 U 和 D 的关系集,即 $G=\{g_j;j\leqslant p\}$。其中 $g_j:U\to V_j'(j\leqslant p)$,$V_j'$为目标属性 d_j 的值域。

经典粗糙集模型面对的对象属性取值是无序的,然而在实际生活中存在大量数据是有序的情况。为此,给出如下有序信息系统的定义:

如果信息系统的条件属性值是有序的,则此时的信息系统(U,A,F)被称为有序信息系统。同理,若决策表的条件属性值和决策属性值都是有序的话,则决策表(U,A,F,D,G)称为有序决策表。

为处理这类型有序数据,Greco,Matarazzo 及 Slowinski 提出了基于优势关系下的粗糙集模型 DRSA[7-8](Dominance-based Rough Set Approach),它用优势关系来代替原来的不可分辨关系,即用优势类代替等价类。

定义 3 设(U,A,F)为有序信息系统,对于 $B\subseteq A$,令 $R_B^{\leqslant}=\{(x_i,x_j)\in U\times U:f_l(x_i)\leqslant f_l(x_j)\ (\forall a_l\in B)\}$,$B\subseteq A$ 称为有序信息系统上的优势关系。记

$$[x_i]_B^{\leqslant}=\{x_j\in U:(x_i,x_j)\in R_B^{\leqslant}\}=\{x_j\in U:f_l(x_i)\leqslant f_l(x_j)\ (\forall a_l\in B)\}$$

则$[x_i]_B^{\leqslant}$表示在属性集 B 条件下,优于对象 x_i 的所有对象集合,称为 x_i 的优势类。同理,可定义有序决策表上的优势关系,记

$$R_d^{\leqslant}=\{(x_i,x_j):g_d(x_i)\leqslant g_d(x_j)\}$$

称 $R_d^{\leqslant}$为决策属性 d 所确定的优势关系。若 $R_A^{\leqslant}\subseteq R_d^{\leqslant}$,则称有序决策表在优势关系下是协调的,否则不协调。

有序信息系统上的多个属性并不是同等重要的,如果去掉某个属性并不影响对象之间的序关系,那么这个属性是冗余的。在保证对象间的序信息不变的情况下,利用优势关系下的粗糙集方法对有序信息进行约简,删除冗余属性,从而简化了对象之间的比较。

定义 4 设(U,A,F)为一个有序信息系统,若 $B\subseteq A$,且满足:

(1)$R_B^{\leqslant}=R_A^{\leqslant}$

(2)$\forall b\in B,R_{B-\{b\}}^{\leqslant}\neq R_A^{\leqslant}$

则称 B 是有序信息系统在优势关系下的一个约简。

对于有序决策表,同样存在条件属性冗余的问题,条件属性的重要性程度与决策属性集 D 密切相关。总可以找到 A 的一个最小子集 A_0,使得用 A_0 来代替 A 不会丢失有序决策表的任何信息,从而得到一个简化的有序决策表。为简单起见,只讨论单目标情形,对于多目标情形可类似推广得到。

类似于有序信息系统的约简定义,同样可得到有序决策表$(U,A,F,\{d\},\{g_d\})$约简定义,即设 $B\subseteq A$,若优势关系 $R_B^{\leqslant}\subseteq R_d^{\leqslant}$成立,且对于任意 $b\in B$,优势关系 $R_{B-\{b\}}^{\leqslant}\subseteq R_d^{\leqslant}$不成立,则称 B 是有序决策表在优势关系下的一个约简。

对于优势类$[x_i]_B^{\leqslant}$和$[x_j]_B^{\leqslant}$,$(x_i,x_j\in U)$利用粗糙集方法可计算在属性集 B 条件下,对象 x_i 优于 x_j 的程度为

$$R_B(x_i,x_j)=\frac{\left|(\sim[x_i]_B^{\leqslant})\cup[x_j]_B^{\leqslant}\right|}{|n|}\tag{1}$$

其中$|\cdot|$表示集合元素的个数,n 表示对象的个数。

利用算术平均得到对象 x_i 在属性集 $B(B\subseteq A)$下的综合优势度为

$$R_B(x_i)=\frac{1}{n-1}\sum_{j\neq i}R_B(x_i,x_j)\tag{2}$$

$R_B(x_i)$的值越大,则对象 x_i 越具有优势,可以按照 $R_B(x_i)$由大到小的顺序将对象由优到劣排序。

引入综合优势度的目的在于对对象进行综合排序,这与加权排序法不同点在于无需提供数据集合之外的任何先验信息,直接从信息表出发,所以排序可以说是比较客观的。

3　基于统计证据的约简方法

本章将根据 Spearman 秩相关系数,给出属性集与属性子集相关性的非参数检验方法,以及基于此度量的属性约简方法。首先给出如下假设检验问题:

H_0:属性集 A 与属性子集 B 不相关↔H_1:属性集 A 与属性子集 B 正相关。

定义 5 称 $r_s(A,B)$为属性集 A 与 B 的 Spearman 秩相关系数,若

$$r_s(A,B)=1-\frac{6S}{n(n^2-1)}\tag{3}$$

其中 $S=\sum_{i=1}^{n}(r_i^A-r_i^B)^2$,$(B\subseteq A)$,$r_i^A$,$r_i^B$分别为在属性集 A 和 B 条件下用粗糙集方法排序后得到的 x_i 的秩。记 $r^A=(r_1^A,r_2^A,\cdots,r_n^A)$为在属性集 A 条件下排序得到的一组秩。

可以看出,此时将属性集 A、B 看作是两个变量 X、Y 来处理的。原来的非参数方法中考虑的两个变量的相关性,现在被替换为两个属性集的相关性。

如果排序结果中出现结(或称为同分)情况,即 $R_B(x_i)\approx R_B(x_j)$时,就让这些对象的秩都等于没有同分出现时它们应有秩的平均值,重新记 x_i 的秩为 $r_i^{A^*}$,$r_i^{B^*}$。如果同分的比例不大时,它们对 $r_s(A,B)$的影响可以忽略,仍可用原式计算。然而,当同分的比例较大时,$r_s(A,B)$的计算中就必须加入一个修正因子。此时,用下式来计算 $r_s(A,B)$:

$$r_s(A,B)=\frac{\dfrac{n(n^2-1)}{6}-\dfrac{1}{12}\left[\sum_i(\tau_{A,i}^3-\tau_{A,i})+\sum_j(\tau_{B,j}^3-\tau_{B,j})\right]-S^*}{2\sqrt{\left[\dfrac{n(n^2-1)}{12}-\dfrac{1}{12}\sum_i(\tau_{A,i}^3-\tau_{A,i})\right]\left[\dfrac{n(n^2-1)}{12}-\dfrac{1}{12}\sum_j(\tau_{B,j}^3-\tau_{B,j})\right]}}\tag{4}$$

其中,$S^*=\sum_{i=1}^{n}(r_i^{A^*}-r_i^{B^*})^2$,$\tau_{A,i}$,$\tau_{B,j}$分别表示 $r_i^{A^*}$,$r_i^{B^*}$ $(i,j\leqslant n)$的结统计量,即等于在一给定秩处同分的观察数。

对于 $r_s(A,B)$的显著性检验,如果是小样本数据,可直接查看 Spearman 秩相关系数检验临界值表,得到临界值 c_α,若 $r_s(A,B)\geqslant c_\alpha$,则拒绝原假设,称属性集 A 与属性子集 B 在 α 显著水平上是相关的。

如果是大样本数据,$r_s(A,B)$的显著性可以用统计量 t 来检验:

$$t=r_s(A,B)\sqrt{\frac{n-2}{1-r_s^2(A,B)}}\tag{5}$$

也就是说,对于大 n 值,由式(5)定义的值遵从 $df=n-2$ 的 t 分布。

定理 1 设(U,A,F)为一个有序信息系统,对于任意 $B\subseteq$

A,下列性质成立:

(1)$0\leqslant r_s(A,B)\leqslant 1$;

(2)若 B 是 A 的一个属性约简,则 $r_s(A,B)=1$。

证明 性质(1)显然成立,性质(2)的证明如下:

因为 B 是 A 的一个约简,所以 $R_B^{<}=R_A^{<}$,$R_{B-\{b\}}^{<}\neq R_A^{<}(\forall b\in B)$则

$$[x_i]_B^{<}=[x_i]_A^{<},[x_j]_B^{<}=[x_j]_A^{<}\quad(i,j\leqslant n)$$

由公式(1)知 $R_B(x_i,x_j)=R_A(x_i,x_j)$,再由公式(2)便得 $R_B(x_i)=R_A(x_i)$,所以利用综合优势度得到的排序结果是相同的,且 $r_i^A=r_i^B(i\leqslant n)$,进而由公式(3)得 $r_s(A,B)=1$。性质(2)得证。

根据以上属性相关性的定义,可将 Spearman 秩相关系数作为属性集相关性的度量,基于此度量给出相关性意义下的属性约简新方法,约简定义如下:

定义 6 设(U,A,F)为一个有序信息系统,如果存在 $B\subseteq A$,满足:

(1)$r_s(A,B)\geqslant c_\alpha$

(2)$\forall b\in B,r_s(A,B-\{b\})<c_\alpha$

则称 B 是 A 在 α 显著性水平上的一个约简。

定理 2 若 B 是 A 的一个属性约简,则 B 也是 A 在 α 显著性水平上的一个约简。

证明 若 B 是 A 的一个属性约简,由定理 1 的性质(2)知,$r_s(A,B)=1$,则 $r_s(A,B)\geqslant c_\alpha$,满足约简条件(1)。又因为 $R_{B-\{b\}}^{<}\neq R_A^{<}(\forall b\in B)$,所以

$$[x_i]_{B-\{b\}}^{<}\neq[x_i]_A^{<},[x_j]_{B-\{b\}}^{<}\neq[x_j]_A^{<}\quad(i,j\leqslant n)$$

再由公式(1)、(2)知,$R_{B-\{b\}}(x_i)\neq R_A(x_i)$,所以 $r_i^A\neq r_i^{B-\{b\}}$。进而由公式(3)知,必存在一个 α_0,使得 $r_s(A,B-\{b\})<c_{\alpha_0}\leqslant 1$ 成立,满足约简条件(2),则 B 也是 A 在 $\alpha=\alpha_0$ 显著性水平上的一个约简,定理得证。

反之未必成立,反例见例 1。

对于有序决策表中有些条件属性与制定决策是无关的,可以删除。为了使知识表达得到简化,需要保证有序决策表的决策能力不变的前提下,使它的属性集尽可能小。

定义 7 设$(U,A,F,\{d\},\{g_d\})$为一个有序决策表,如果存在 $B\subseteq A$,满足:

(1)$r_s(B,d)\geqslant r_s(A,d)\geqslant c_\alpha$

(2)$\forall b\in B,r_s(B-\{b\},d)<c_\alpha$

则称 B 是有序决策表在 α 显著性水平上的一个约简。

4 例子

通过下面的两个例子来说明如何利用 Spearman 秩相关系数来进行约简的。

例 1 设 a_1,a_2,a_3 分别表示学生课程:语文、数学、外语三个方面的评价,评价分为三个等级 1-差,2-良,3-优。可见该信息系统为有序信息系统,见表 1。

对于属性集 $A=\{a_1,a_2,a_3\}$,$B=\{a_2,a_3\}$,利用公式(1)、(2)计算对象 x_i 的综合优势度 $R_A(x_i)$ 与 $R_B(x_i)$,并分别按照 $R_A(x_i)$,$R_B(x_i)$ 的大小将对象排序为:

(1)$x_6\succ x_5\succ x_2\succ x_4\succ x_1\succ x_3$

(2)$x_6\succ x_5\succ x_4\succ x_2\succ x_1\succ x_3$

进而得到两组秩分别记为:

(1)$r^A=(5,3,6,4,2,1)$

(2)$r^B=(5,4,6,3,2,1)$

由公式(3)计算得 Spearman 秩相关系数 $r_s(A,B)=0.95$,查表得 $r_s(A,B)>c_{0.01}=0.943$,拒绝原假设,属性集 A 与 B 在 0.01 水平上显著相关。然后对 $B_1=\{a_2\}$条件下,同样用粗糙集方法排序后得到结果为:$x_5\succ x_1\approx x_2\approx x_6\succ x_3\approx x_4$,秩为 $r^{B_1}=(3,3,5.5,5.5,1,3)$,利用公式(4)计算得 $r_s(A,B_1)\approx 0.65<c_{0.01}=0.943$,接受原假设,属性集 A 与属性 a_2 不相关。同样计算得,$r_s(A,\{a_3\})\approx 0.49<0.943$,属性集 A 与属性 a_3 不相关。所以,$B=\{a_2,a_3\}$为有序信息系统在 0.01 显著性水平上的一个约简。依此方法同样验证了 $\{a_1,a_3\},\{a_1,a_2\},\{a_1\},\{a_3\}$ 均不为 A 在 0.01 显著性水平上的约简,所以该有序信息系统只有一个约简。

通过此例发现,在属性集 A 与 B 条件下的排序结果区别在于 x_2 与 x_4 上,这是因为 x_2 与 x_4 是不可比较的对象,因此这两种排序都是合理的。所以,在 α 显著性水平上的约简方法对这种情况有一定的容错功能。通过粗糙集方法发现 $R_A^{<}\neq R_B^{<}$,因此在粗糙集方法下是不可以再进行约简的,但本文的方法却可以在某种显著性水平上进行约简,放宽了约简的条件,且粗糙集约简是 α 显著性水平上约简的一种特殊情况。凡是粗糙集方法能约简的,本文的方法均能约简,且结果一致。反之,本文的方法能约简的,粗糙集方法未必能约简。

例 2 给出如下有序决策表,属性 d 为决策属性,表示学生总体评价水平,评价依然分为三个等级 1-差,2-良,3-优,见表 2。

表 1 学生评价信息系统

U	a_1	a_2	a_3
x_1	1	2	1
x_2	3	2	2
x_3	1	1	2
x_4	2	1	3
x_5	2	3	2
x_6	3	2	3

注:$f_1(x_5)=2$

表 2 有序决策表

U	a_1	a_2	a_3	d
x_1	1	2	1	1
x_2	3	2	2	3
x_3	1	1	2	1
x_4	2	1	3	2
x_5	3	3	2	3
x_6	3	2	3	3

注:$f_1(x_5)=3$

对于此例同样先采用粗糙集方法排序,然后根据 Spearman 秩相关系数进行约简,不同的是此时考虑的是属性子集与决策属性之间的相关程度。根据决策属性 d 将对象排序后得 $x_2\approx x_5\approx x_6\succ x_4\succ x_1\approx x_3$,进而得到一组秩为 $r^d=(5.5,2,5.5,4,2,2)$。根据例 1 已知排序结果,利用公式(4)计算条件属性集 $B=\{a_2,a_3\}$ 与决策属性 d 的 Spearman 秩相关系数,得 $r_s(B,d)=r_s(A,d)\approx 0.802>c_{0.10}=0.657$。所以,拒绝原假设,$B$ 与 d 在 0.10 水平上是显著相关的。当 $B_1=\{a_2\}$时,计算得 $r_s(a_2,d)\approx 0.657<c_{0.10}=0.657$,接受原假设,$a_2$ 与 d 不相关。同理可得,a_3 与 d 不相关。所以,$\{a_2,a_3\}$是有序决策表在 0.10 显著性水平上的一个约简。

若取 $B_2=\{a_1\}$时,利用粗糙集方法得到排序结果为:$x_2\approx x_5\approx x_6\succ x_4\succ x_1\approx x_3$,进而得到一组秩为 $r^{a_1}=(5.5,2,5.5,4,2,2)$,再根据公式(4)得条件属性 a_1 与决策属性 d 的 Spearman 秩相关系数 $r_s(a_1,d)=1>c_{0.10}$。所以 $B_2=\{a_1\}$也是有序决策表在 0.10

(下转 72 页)