文章编号:1001-9081(2005)11-2645-02

基于 Rough set的有序信息表的排序问题研究

李 明,张保威
(兰州理工大学 计算机与通信学院,甘肃 兰州 730050)
(lim3076@163.com)

摘 要:针对有序信息表的排序问题,提出了总体排序的过程框架。将有序信息表转化为二进制信息表,运用粗糙集理论对信息表进行简化,在对属性值标准化的基础上构造有序信息表中实体的排序度量函数,根据度量函数值的大小进行排序。实例表明该过程框架在误差允许范围内是有效可行的。

关键词:粗糙集;有序信息表;属性约简;排序
中图分类号:TP311.13 **文献标识码**:A

Ordering objects in ordered information table based on rough set theory

LI Ming ZHANG Bao-wei
(School of Computer and Communication, Lanzhou University of Technology, Lanzhou Gansu 730050, China)

Abstract: Aiming at ordering objects in ordered information table, a framework based on rough set theory was proposed. An information system was firstly transformed into a binary one which would be simplified by rough set theory. According to the values of the weight function, the objects in the original information tables could be sorted. Experiments illustrate that the framework is feasible and effective.

Key words: rough set; ordered information table; attribute reduction; sorting

0 引言

在真实世界里,我们经常碰到对象的排序问题。在这类问题中,通常用一组属性来描述特定问题中的实体,不同的实体在某个特定的属性上会有一个排序,而所有实体也会有一个总体排序。例如,一个产品可以用一系列属性来描述,如价格、信誉度、市场占有率等,不同的产品在特定的属性上例如价格上根据价格的高低会有一个排序,而且所有的产品综合起来又会有一个排名。

在给定总体序的情况下,各个条件属性上的序与总体序的关系及各个属性之间的关系等问题,Y. Y. Yao[1], Greco, Matarazzo和 Slowinski[2,3]等人已经作了比较深入的研究。而在总体序未知的情况下,怎样由各个属性上的序关系得到总体的排序,传统的方法都要借助于领域专家的参与,这给实际应用增加了难度。针对这个问题,Toshihiro Kamishima, Shotaro Akaho[4]提出了一种贪心学习算法,给出了一个偏好函数,计算每一个实体与其他实体的偏好函数值,提取与其他实体偏好函数最大的实体,认为它优于其他实体,将其排在其他实体的前面,然后在剩余的实体中重复计算并提取,这样就得到了总体的排序,可是它有较高的时间复杂度。我们提出了一种总体排序的过程框架,首先将有序信息表转化成二进制信息表,利用粗糙集理论对信息表进行简化,计算简化后所有属性的重要性,在对属性值标准化的基础上构造有序信息表中实体的排序度量函数,根据排序度量函数值的大小进行排序。实例分析表明,提出的过程框架在误差允许范围内是有效可行的,而且与领域知识无关,并有效降低了时间复杂度。

1 相关概念及定义

定义 1 标准信息表是一个四元组:

$IT = (U, A_t, \{V_a \mid a \in A_t\}, \{I_a \mid a \in A_t\})$

其中 U是非空论域,A_t是属性集,V_a为值域,I_a为由论域到值域之间的映射(函数)。

为简便起见在这里我们仅讨论论域和属性有限的情况。

定义 2 U为非空论域,$\succ$为 U上的二元关系,$\succ$称为 U上的弱序关系如果满足以下两个条件:

1) 反对称性

$x \succ y \Rightarrow \neg(y \succ x)$

2) 负传递性

$\neg(x \succ y) \wedge \neg(y \succ z) \Rightarrow \neg(x \succ z)$

弱序关系 $\succ$ 所导出的等价关系定义如下:

$x \sim y \Leftrightarrow \neg(x \succ y) \wedge \neg(y \succ x)$

即 x等价于 y当且仅当 x与 y在 $\succ$ 上根据现有信息无法确定两者之间的序关系。

定义 3 有序信息表 OIT是一个二元组:

$OIT = (IT, \{\succ_a \mid a \in A_t\})$

其中 IT是一个标准信息表,$\succ_a$是 V_a上的序关系。对任意的 $x, y \in U$, $x \succ_a y \Leftrightarrow I_a(x) \succ I_a(y)$

在有序信息表中我们考虑的是实体之间的序关系,所以只需关心实体对 (x, y)。对任意的 $x, y \in U$,我们不需考虑实体本身之间的关系,即 $x \neq y$,故论域为:

$U \times U^+ = U \times U - \{(x, x) \mid x \in U\} = \{(x, y) \mid x, y \in U;\ x \neq y\}$

所以我们可以将有序信息表 OIT转化为二元信息表 DOIT:

$DOIT = (U \times U^+, A_t, \{V_a \mid a \in A_t\}, \{I_a \mid a \in A_t\})$

收稿日期:2005-05-13;**修订日期**:2005-07-14 **基金项目**:甘肃省教育厅科技基金项目(0416B_04)

作者简介:李明(1959-),男,河北束鹿人,副教授,主要研究方向:数据库、数据挖掘; 张保威(1980-),男,河南南阳人,硕士研究生,主要研究方向:数据挖掘、粗糙集.

$$I_a=\begin{cases}0, & x\succ_a y\\ 1, & \text{otherwise}\end{cases} \quad (1)$$

可以证明二进制信息表 DOIT保持了有序信息表 OIT上的序关系，而对 DOIT所做的任何不改变信息量的处理不会破坏原始信息表中实体之间的序关系。

Rough Set是由 Pawlak[5]提出的一种处理不精确、不确定数据的数学方法，下面对本文所用到的属性重要性进行定义。

定义 4 对任意的属性 $a\in A_t$，a的重要性 S(a)如下式：

$$S(a)=\mu_P(A_t)-\mu_{P-a}(A_t) \quad (2)$$

其中 $P,\ Q\subseteq A_t$，POS_PQ表示 Q的 P正域，$\mu_P(Q)=\frac{card(POS_PQ)}{card(U)}$。

定义 5 实体 O_x的排序度量函数定义如下：

$$W(O_x)=\sum_{i=1}^{n}S(a_i)\times v_{a_i}(O_x) \quad (3)$$

其中 $S(a_i)$表示属性 a_i的重要性，$v_{a_i}(O_x)$表示实体 O_x在属性 a_i上的标准化后的值，n为属性的个数。

定义 6 为了评价我们得到的排序与实际排序的误差，在本文中我们引用 Spearman[6]提出的评价系数 "ρ"，序 R_1和 R_2之间的系数 ρ为：

$$\rho=1-\frac{6\times\sum_{O_x\in U}(r(R_1,\ x)-r(R_2,\ x))^2}{(card\ U)^3-card\ U} \quad (4)$$

其中 r(R; x)表示实体 O_x在序 R中的位置，例如序 $R=O_3\succ O_1\succ O_2$中，r(R; 3)=1，r(R; 2)=3。R_1与 R_2的排序相同当且仅当 ρ=1，R_1与 R_2的排序相反当且仅当 ρ=-1。当 R_1，R_2分别为求出的序和实际排序时，ρ就变成了排序的准确系数，ρ与排序的准确率成正比。

2 总体排序的过程框架

设有序信息表 OIT有 m个实体元素，n个属性，则对 OIT总体排序的过程框架如下：

输入 有序信息表 OIT

输出 总体排序后的实体元素

step1 首先利用式(1)将有序信息表 OIT转化为二进制信息表 DOIT。

step 2 利用粗糙集理论对 DOIT进行简化：

首先求出 DOIT的所有约简，任取一个约简 R(设 R有 k个属性)，对任意的属性 $a_i\in R$求出 a_i的重要性 $S(a)=1-\mu_{P-a}(A_t)$，求出 R中所有属性的重要性。

对 OIT，保留 R中所有属性所对应的列，删除其他的属性得到有序信息表 OIT′。

step 3 对有序信息表 OIT′进行处理：

对任意的 $a_i\in R$，如果 V_{a_i}是数值且其上的序关系是普通意义上的数值大小关系，则转 step 4，否则转 step 5。

step 4 对每一个实体 O_j的属性值 $I_{a_i}(O_j)$，标准化：

$V_{a_i}(O_j)=\frac{I_{a_i}(O_j)-\min\{V_{a_i}\}}{\max\{V_{a_i}\}-\min\{V_{a_i}\}}$，转 step 6。

step 5 对所有的 $a_i\in R$，将其所对应的属性值排序：

$$I_{a_i}(O_{b1})\succ_{a_i}I_{a_i}(O_{b2})\succ_{a_i}\cdots\succ_{a_i}I_{a_i}(O_{bn}) \quad (5)$$

根据(5)式中的排位给每个实体所对应的属性由大到小均匀赋上 0到 1之间的数值。所赋值 $V_{a_i}(O_j)$满足：

$$V_{a_i}(O_j)>V_{a_i}(O_k)\Leftrightarrow I_{a_i}(O_j)\succ_{a_i}I_{a_i}(O_k)$$

step 6 对每一个实体 $O_j\in U$，求出序函数值：

$$w_j=w(O_j)=\sum_{1}^{k}S(a_k)\times V_{a_k}(O_j)$$

step 7 根据 w_j值由大到小的顺序，将所对应的实体元素排序，输出。如果存在实体 O_i，O_j，且 $w(O_i)=w(O_j)$，考察不属于约简 R的属性，取其中重要性最大的属性不妨设为 a_k，如果 $I_{a_k}(O_i)\succ I_{a_k}(O_j)$，则将 O_i排在 O_j之前，否则，将 O_i排在 O_j之后。

3 实例分析

表 1 有序信息表

产品	抽样率(a)	保质期(b)	价格(c)	重量(d)
P1	0.015	3 years	200	heavy
P2	0.010	3 years	300	very heavy
P3	0.025	3 years	300	light
P4	0.025	3 years	250	very light
P5	0.025	2 years	200	very light

假设 P1，P2，P3，P4，P5分别为 5个不同厂家生产的产品，a，b，c，d分别表示产品抽样率、保质期、价格和重量。

在各个属性上的序关系如下：

$0.025\succ_a 0.015\succ_a 0.010$

$3\ years\succ_b 2\ years$

$200\succ_c 250\succ_c 300$

$very\ light\succ_d light\succ_d heavy\succ_d very\ heavy$

step 1 利用(1)将有序信息表(表 1)转化为二进制信息表(表 2)。

表 2 二进制信息表

Object	a	b	c	d	Object	a	b	c	d
(1, 2)	1	0	1	1	(3, 4)	0	0	0	0
(1, 3)	0	0	1	0	(3, 5)	0	1	0	0
(1, 4)	0	0	1	0	(4, 1)	1	0	0	1
(1, 5)	0	1	0	0	(4, 2)	1	0	1	1
(2, 1)	0	0	0	0	(4, 3)	0	0	1	1
(2, 3)	0	0	0	0	(4, 5)	0	1	0	0
(2, 4)	0	0	0	0	(5, 1)	1	0	0	1
(2, 5)	0	1	0	0	(5, 2)	1	0	1	1
(3, 1)	1	0	0	1	(5, 3)	0	0	1	1
(3, 2)	1	0	0	1	(5, 4)	0	0	1	0

step 2 利用 Rough Set理论对表 2进行处理：求得表 2的其中一个约简 $R=\{b\ c\}$，属性 b，c的重要性分别为 S(b)=0.8，S(c)=0.2。对表 1进行处理，保留 R出现的属性，删除冗余的属性，得到表 3。

表 3 简化后的有序信息表

产品	b	c
P1	3years	200
P2	3years	300
P3	3years	300
P4	3years	250
P5	2years	200

表 4 标准化后的有序信息表

产品	b	c
P1	0.66	0.25
P2	0.66	0.75
P3	0.66	0.75
P4	0.66	0.50
P5	0.33	0.25

step 3，step 4，step 5 对表 3中的属性值进行标准化得到表 4。

(下转第 2664页)

推理。推理是从报警节点开始向上游进行的，LI的上游有 2个节点 FI_{in}和 FI_{out}，先检测这 2个节点的值有怎样的变化，发现 FI_{in}没有变化，只有 FI_{out}变小了，于是排除掉 FI_{in}方向存在相容通路的可能。从 LI向 FI_{out}反向推理，计算过程如下：

$$\Delta FI_{out}=EFI_{out}\circ\Delta LI=\begin{bmatrix}0&0&0.2&1\\0&0&0.8&0.2\\0.2&0.8&0&0\\1&0.2&0&0\end{bmatrix}\circ\begin{bmatrix}1\\0.1\\0\\0\end{bmatrix}=\begin{bmatrix}0\\0\\0.2\\1\end{bmatrix}$$

因此，FI_{out}的理论变化量减小很多，这与 FI_{out}实际的变化量相一致，因此得知支路 4是属于相容通路的支路，但由于节点 FI_{out}并非通路的首节点，所以还要以 FI_{out}为下游节点，向上游进行反向推理。推理过程同上。仅列出推理结果：

$\Delta V2=[0\quad 0\quad 0.3\quad 0.8]$，$\Delta V3=[0\quad 0\quad 0.2\quad 0.7]$

用加权法将其转化为实数：$\Delta V2=2\times0+1\times0+(-1)\times0.3+(-2)\times0.8=-1.9$，$\Delta V3=2\times0+1\times0+(-1)\times0.2+(-2)\times0.7=-1.6$。可以看出 $|\Delta V2|>|\Delta V3|$，因此，虽然推理出 2个潜在故障源，但可以认为 V2是实际故障源的概率大于 V3。因此，V2的优先级大于 V3，在核实故障源时要先核实 V2。

4 结语

本文提出了一种基于模糊矩阵运算的模糊 SDG故障诊断新方法，并利用该方法成功地对离心泵液位系统做了故障诊断，并对多个潜在故障源的优先级作了排序。这种方法的一个主要特点是模糊推理不是建立在隶属函数的基础上的，整个推理过程用的都是模糊矩阵的形式，这个特点对于计算机编程比较好实现，因此可以比较方便地将此方法移植到计算机自动故障诊断系统中。

此外，虽然支路模糊影响矩阵比支路影响度隶属函数更容易获得，但影响规则表参数 State的确定存在一定的难度，它直接关系到推理的准确性。本文案例中的影响规则表参数是通过在仿真器上做节点拉偏试验得出的，对于不具备仿真器的研究条件，可以考虑根据工程技术人员的经验确定影响矩阵。

参考文献：

[1] 张贝克. SDG实时推理机制、图形化软件平台及安全工程应用研究[D]. 北京化工大学，2004.

[2] 吴重光，夏 涛，张贝克. 基于符号定向图（SDG）深层知识模型的定性仿真[J]. 系统仿真学报，2003，15(10)：16－18.

[3] SHIOZAKI J，SHIBATA B，MATSUYAMA H，et al. Fault Diagnosis of Chemical Processes Utilizing Signed Directed Graphs－Improvement by Using Temporal Information[J]. IEEE Transactions on Industrial Electronics，1989，36(4)：469－474.

[4] WANG XZ，YANG SA，VELOSO E，et al. Qualitative Process Modelling－A Fuzzy Signed Directed Graph Method[J]. Computers & Chemical Engineering，1995，19(Supplement)：S735－S740.

[5] YU CC，LEE C. Fault Diagnosis Based on Qualitative/Quantitative Process Knowledge[J]. AIChE Journal，1991，37(4)：617－628.

[6] TARIFA EE，SCENNA NJ. Fault Diagnosis Direct Graphs and Fuzzy Logic[J]. Computers & Chemical Engineering，1997，21：649－654.

[7] WANG XZ，YANG SA，YANG SH，et al. The Application of Fuzzy Qualitative Simulation in Safety and Operability Assessment of Process Plants[J]. Computers & Chemical Engineering，1996，20：671－676.

[8] 青义学. 模糊数学入门[M]. 上海：知识出版社，1987.

[9] 刘叙华. 模糊逻辑与模糊推理[M]. 长春：吉林大学出版社，1989.

[10] 何平，王鸿绪. 模糊控制器的设计及应用[M]. 北京：科学出版社，1997.

[11] 孙庚山，兰西柱. 工程模糊控制[M]. 北京：机械工业出版社，1995.

[12] 吴重光. 化工仿真实习指南[M]. 北京：化学工业出版社，1999.

(上接第 2646页)

step 6 计算每一个实体元素的排序度量函数值：$w_1=0.578$，$w_2=0.678$，$w_3=0.678$，$w_4=0.628$，$w_5=0.314$。

step 7 根据度量函数值 w_j的大小，得到实体的排序为 P3≥P2≥P4≥P1≥P5。

在本例中，描述实体的属性共有 4个，利用粗糙集理论方法进行简化后只剩余属性 b、c这大大降低了计算的复杂度。考虑表中信息的实际意义亦可知，产品的抽样率 a及产品的重量 d对决定产品的名次影响不大。根据 Spearman[6]所提出的度量系数 ρ计算得到其值为 1，即几乎不存在误差。我们又分析了 Maclean's杂志 2000年加拿大大学排名的数据，共有 22个指标对医学/博士类大学进行排名[5]，这 22个指标产生了 22个单独的排名，Maclean's又给出了一个整体的排名，。用本文的方法进行简化后只剩余 7个属性，将这 7个属性所对应的值标准化，再由本身的信息计算其重要性，通过度量函数值的大小就可以得到总体的排名。

参考文献：

[1] SAI Y，YAO YY，ZHONG N. Data analysis and mining in ordered information tables[A]. Proceedings of IEEE International Conference On Data Mining[C]，2001. 497－504.

[2] GRECO S，MATARAZZO B，SLOWINSKI R. The use of rough sets and fuzzy sets in MCDM[A]. GAL T，HANNE T，STEWART T，ed. Advances in Multiple Criteria Decision Making[C]. Boston：Kluwer Academic Publishers，1999.

[3] GRECO S，MATARAZZO B，SLOWINSKI R. Rough approximation of a preference relation by dominance relations[J]. European Journal of Operational Research，1999，1(117)：63－83.

[4] KAMISHIMA T，AKAHO S. Learning from Ordered Examples[A]. Proceedings of the IEEE International Conference on Data Mining[C]，2002. 645－648.

[5] Maclean's University Guide[J]. Maclean，2000，(11).

[6] KENDALL M，GIBBONS JD. Rank Correlation Methods[M]. Oxford University Press，1990.

[7] PAWLAK Z. Rough Sets：Theoretical Aspects of Reasoning about Data[M]. Kluwer Academic Publishers，1991.

[8] GRECO S，MATARAZZO B，SLOWINSKI R. Rough sets theory for multicriteria decision analysis[J]. European Journal of Operational Research，2001，1(129)：1－47.

[9] 王国胤. Rough集理论与知识获取[M]. 西安：西安交通大学出版社，2001.