文章编号:1001 - 9081(2009)07 - 1758 - 02

基于 spearman 秩相关的序值决策系统约简

祁云嵩[1],谢 军[2]
(1. 江苏科技大学 计算机科学与工程学院,江苏 镇江 212003; 2. 南京理工大学 计算机科学与技术学院,南京 210094)
(qys@ujs.edu.cn)

摘 要:约简是知识发现的重要过程。经典的基于等价关系的粗糙集理论,没有考虑系统取值的序值性,并且对数据噪声较为敏感。提出了一个基于 spearman 秩相关分析的序值决策系统约简方法,该方法通过各属性对被决策个体的 spearman 秩次的影响来确定约简结果。实验结果表明,该方法不但考虑了系统属性值的序值关系,并且对数据噪声不敏感,因而更符合实际应用的要求。

关键词:粗糙集;知识约简;序值决策系统;spearman 秩相关
中图分类号: TP18 **文献标志码**:A

Reduction in ordered decision system based on spearman rank correlations

QI Yun-song[1], XIE Jun[2]
(1. *School of Computer Science and Engineering, Jiangsu University of Science and Technology, Zhenjiang Jiangsu* 212003, *China*;
2. *School of Computer Science and Technology, Nanjing University of Science and Technology, Nanjing Jiangsu* 210094, *China*)

Abstract: Reduction is an important knowledge discovery process. The original rough set theory based on equivalence relations does not consider attributes with preference-ordered domains and is sensitive to data noise. Based on spearman rank correlation analysis, a method was introduced to reduction ordered decision systems. This method computes the reduction according to the impact of certain attributes on the spearman rank correlations between objects. Experimental results show that this method is effective as it not only takes into account the preference relationship, but also improves noise immunity, and thus is more suitable for practical application.

Key words: rough set; knowledge reduction; ordered decision system; spearman rank correlation

0 引言

数据约简[1-4]是知识发现的重要过程,是寻求最简知识表达的重要手段。在粗糙集理论研究中求解数据约简已有众多的研究成果[5-7],但这些研究大多是基于等价关系的约简方法。基于等价关系的粗糙集约简方法在实际应用中受到了较大的限制。首先,基于等价关系的粗糙集约简对噪声数据非常敏感。在标准粗集约简的定义下,信息表中即使仅仅一个对象被污染,整个系统的不可分辨关系都会改变,从而可能得到完全不同的约简结果。其次,基于等价关系的粗糙集约简方法没有考虑信息系统各属性值的序值关系(可比较大小的关系),然而,在实际应用中很多系统的属性取值具有序值性。如:学生综合成绩评定、项目投资效益评估、产品市场份额分析等。文献[8, 9]提出的基于优势关系的粗糙集方法(Dominance-based Rough Set Approach,DRSA)很好地解决了该问题,弥补了经典粗糙集在解决此类问题时的缺陷。DRSA 用优势关系取代了经典粗糙集中的等价关系来考虑多标准下的对象之间的优劣信息。这种基于优势关系的粗糙集约简方法虽然在序值决策系统约简上取得了一定的进展[10, 11],但它仍未从根本上突破粗糙集理论的等价关系的局限性——对系统噪声数据较为敏感。这种局限性的根源来自于该理论缺乏统计学的理论基础[12]。众所周知,知识发现的基础是数据样本,如果没有统计学上的证据,那么从数据样本中得出的结论其泛化能力有限。本文依据无参数统计理论,提出了一个基于 spearman 秩相关分析的序值决策系统约简方法,该方法不但考虑了系统属性值的序值关系,其结果在数据噪声干扰下也较为稳定,因而具有较好的实际应用前景。

1 Spearman 秩相关序值决策系统分析

一个决策系统是一个四元组 $S = \langle U, AT \cup D, V, f \rangle$,其中:$U$ 为论域,AT 为条件属性集,D 为决策属性,$V = \bigcup_{a \in AT \cup D} V_a$,$f$ 为 $U \times (AT \cup D) \rightarrow V$ 的信息函数,$\forall a \in AT \cup D, x \in U$, $f(x,a) \in V_a$(V_a 是属性 a 的值域)。如果一个决策系统在各个属性上的取值均是可比较大小的,则该决策系统称为序值决策系统。基于 spearman 秩相关的序值决策系统分析是根据系统中各个个体与某一标准个体的关联度大小来决策的。

在单个属性(设为第 j 个属性)意义下,序值决策系统中个体的关联系数定义为论域中两个个体在该属性指标下接近程度的一种度量,记为 $E(j)$。显然,指标关联系数 $E(j)$ 越大,两个个体在相应属性指标下越接近。因此,可构造一标准个体(例如,该个体在各属性指标下取最大值),通过测量论域中个体与标准个体的接近程度对被决策个体作出决策。这里,记 $E_i(j)$ 为论域中第 i 个个体在第 j 个属性指标下与标准个体的接近程度,即关联系数。

仅凭单个指标的关联系数 $E_i(j)$ 还难以对个体作出合理的决策,为计算论域中个体在所有属性指标下的综合评价值,这里引入相关度的概念,记为 λ_i。相关度 λ_i 定义为论域中第 i 个个体综合所有属性指标的关联系数 $E_i(j)$ 与各属性指标的

收稿日期:2009 - 01 - 05;**修回日期**:2009 - 03 - 02。 **基金项目**:国家自然科学基金资助项目(60773172)。
作者简介:祁云嵩(1967 -),男,江苏如皋人,副教授,博士研究生,主要研究方向:模式识别、生物信息学; 谢军(1973 -),男,湖南衡阳人,讲师,博士研究生,主要研究方向:人工智能、粗糙集。

重要程度 w_j 秩次一致性的一种度量。若某一个体与标准个体的相关度 λ_i 越大,则说明该个体关联系数 $E_i(j)$ 的结构分布与指标权重 w_j 结构分布吻合越好,该个体与标准个体的关联程度也越高。据此,序值决策系统的决策问题便简化为关联度数值的大小比较。

1.1　单个指标关联系数的计算

取序值决策系统中各属性值域的最大值构造标准个体 x_0(考虑到不同属性取值的数量级不同,在数据处理前常对序值决策系统各属性的取值作标准化处理,这时,标准个体 x_0 各属性的取值为1),以 x_0 为参照序列,以论域中第 i 个个体 x_i 为比较序列,则 x_0 与 x_i 在第 j 个属性上的关联系数 $E_i(j)$ 定义为[13]:

$$E_i(j) = \frac{\delta_{\min} + \rho\delta_{\max}}{\delta_{i0}(j) + \rho\delta_{\max}} \tag{1}$$

式中 $\delta_{\min} = \min_i \min_j |x_0(j) - x_i(j)|$、$\delta_{\max} = \max_i \max_j |x_0(j) - x_i(j)|$、$\delta_{i0}(j) = |x_0(j) - x_i(j)|$、$\rho \in [0,1]$ 为分辨系数,其作用是削弱最大绝对差数值太大而产生的影响,用以提高关联系数之间的差异显著性,本研究中取值为 0.3。

关联系数 $E_i(j)$ 描述了标准个体 x_0 与被决策个体 x_i 间第 j 个属性标准下的取值的接近程度,$E_i(j)$ 越大,说明两个体在该属性标准下越相似。

1.2　相关度的计算

设序值决策系统中各条件属性的重要性指标(权重系数)为 w_j($j = 1, 2, \cdots, n$,n 为条件属性的个数),根据非参数统计理论,各属性的关联系数 $E_i(j)$ 与属性权重序列的 spearman 秩相关系数定义为标准个体 x_0 与被决策个体 x_i 的相关度 λ_i:

$$\lambda_i = 1 - \frac{6\sum_{j=1}^{n}(a_j - b_j)^2}{n(n^2 - 1)} \tag{2}$$

式中,a_j 表示第 j 个属性权重在权重向量中由大到小的排序数,b_j 表示关联系数 $E_i(j)$ 在其相应的序列中的排序数。若 $\lambda_i = 1$ 表示属性权重 w_j 与关联系数 $E_i(j)$ 之间的秩次完全相同,呈正相关;$\lambda_i = -1$ 表示两者之间的秩次完全相反,呈负相关;$\lambda_i = 0$ 则表示两者之间的秩次完全无关。显然,λ_i 值越大,属性权重 w_j 与关联系数 $E_i(j)$ 之间的相关性越好,被决策个体 x_i 与标准个体 x_0 越相似。如果是分类决策,则相关度越接近的两个个体越有可能属于同一类;如果是优劣判断,则相关度越大的样本越优异(假设系统中各属性指标取值越大越优异)。

2　基于 spearman 秩相关的序值系统约简

根据前述讨论,系统中个体与标准个体的相关度可以作为个体特征的一种度量,而这种相关度的计算取决于个体在各属性标准下的取值。如果系统中的某个条件属性不影响个体间这种相关度的秩次,则可认为该属性是冗余的。据此,可以给出基于这种相关度的序值决策系统约简算法:

1) 构造序值决策系统的标准个体 x_0;

2) 根据式(1)计算属性集 AT 的单个指标关联系数;

3) 根据式(2)计算各个个体 x_i 与 x_0 的 spearman 秩相关系数并构成向量 λ;

4) 从属性集 AT 减去某个属性 a;重复步骤1) ~ 3),得到相应的相关系数向量 γ;

5) 计算 λ、γ 的 spearman 秩相关系数 k,如果 k 大于给定的阈值,则属性 a 是可约简的。

3　算例

本文选用如表 1 所示的决策系统对算法进行测试。对照前述系统定义,其中论域 $U = \{x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8\}$,$AT = \{a_1, a_2, a_3, a_4, a_5, a_6\}$,$d$ 为决策属性。显然,该系统为序值决策系统。设属性的重要性顺序为 $a_1 > a_2 > a_3 > a_4 > a_5 > a_6$,在对系统条件属性取值标准化后构造标准个体 $x_0 = \{1, 1, 1, 1, 1, 1\}$。我们分别计算了全体属性集以及在全体属性集中分别减去某个属性的情况下各个体与标准个体的 spearman 秩相关度。计算结果如表 2 所示,其中 λ_0 为全体属性集下的各个体与标准个体的 spearman 秩相关度,λ_i 为全体属性集中减去属性 a_i 时各个体与标准个体的 spearman 秩相关度。

表 1　序值决策系统 S

U	a_1	a_2	a_3	a_4	a_5	a_6	d
x_1	0	0.1	3.0	4.2	1.5	1.4	2
x_2	0.1	0.2	1.5	4.0	1.6	1.5	1
x_3	3.0	3.0	0.1	3.2	0.5	0.6	1
x_4	1.5	3.2	0	3.3	0.2	0.1	3
x_5	2.0	0.3	3.2	3.5	1.6	1.6	2
x_6	4.0	0.5	3.4	4.6	1.7	1.8	3
x_7	3.2	1.7	4.2	1.5	1.9	2.0	0
x_8	3.5	4.2	4.4	3.5	2.0	2.1	0

表 2　系统在特定属性集下各个体与标准个体相关度

U	λ_0	λ_1	λ_2	λ_3	λ_4	λ_5	λ_6
x_1	0.0286	0.1000	0.1000	0	-0.7000	0.3000	0.3000
x_2	-0.0143	0.2250	0.2250	0	-0.7750	0.0750	0.1000
x_3	-0.1000	0	0	0.0750	-0.9250	0.0750	0.0750
x_4	0.4857	0.1000	0.7000	0.7000	0.1000	0.6000	0.6000
x_5	0.0714	0.1750	0.1750	0.0750	-0.6250	0.3000	0.3000
x_6	0.2571	0.7000	0.1000	0.4000	-0.3000	0.3000	0.3000
x_7	-0.4857	-0.3000	-0.8000	-0.6000	-0.1000	-0.5000	-0.5000
x_8	-0.3857	-0.5000	-0.2750	-0.2750	-0.5000	-0.2750	-0.2750

表 3 列出了 λ_i 与各个体的决策属性取值向量与表 2 计算的相关度向量间的 spearman 秩相关度 k。对照上述算法,如果取相关度 k 的阈值为 0.95,则 $\{a_1, a_2, a_3, a_4, a_5\}$,$\{a_1, a_2, a_3, a_4, a_6\}$ 为该序值决策系统的两个约简;如果取相关度 k 的阈值为 0.89,则系统约简为 $\{a_1, a_2, a_4, a_5\}$,$\{a_1, a_2, a_4, a_6\}$ 或 $\{a_1, a_2, a_3, a_4\}$。同时实验还发现,数据的轻微变化(< 0.1)并不影响约简结果,这说明这种基于 spearman 秩相关分析的约简算法具备一定的抗干扰能力。

表 3　λ_i 与决策值之间的 Spearman 秩相关度 k

λ_0	λ_1	λ_2	λ_3	λ_4	λ_5	λ_6
0.9762	0.7202	0.7202	0.8929	0.2619	0.9583	0.9524

4　结语

本文从统计学的角度出发,提出了一个基于 spearman 秩相关性分析的数据约简方法。与传统的粗糙集中基于等价关

(下转第 1763 页)

JBPM 引擎管理。系统改造者只需修改过程定义文件(JPDL),即可以很小的代价定制 JSPMP 平台管理的业务流程(任务报告过程)。下面给出对应于图 7 的过程定义文件(JPDL)片段:

```
<start-state name="开始状态">
  <transition to="个人报告"/>
</start-state>
<task-node name="个人报告" create-tasks="false"
  signal="last-wait">
  <task name="PM_TASK_MEMBER_REPORT"/>
  <event type="node-enter">
    <action class="DispatchTaskAction"> <taskName>PM_
      TASKMEMBER_REPORT</taskName>
      <swimlaneName>taskMember</swimlaneName>
    </action>
  </event>
  <transition name="任务报告" to="任务报告"/>
</task-node>
```

可见,JSPMP 平台可以充分发挥工作流的理论与技术优势,为类似的 J2EE 遗留系统定制灵活的业务过程。

5　讨论

JSPMP 平台既有优势也存在潜在问题。优势包括:1)坚持非侵入原则,设计出 JSPMP 平台与遗留系统间的接口,避免直接修改遗留系统。2)日志功能,JSPMP 平台维护系统状态,集中记录每一次状态变化,小到一次表单提交,大到一次业务过程的结束。分析这些日志,既有助于发现系统瓶颈,也可以分析系统业务特点。3)业务和状态数据分离,遗留系统记录系统的业务数据,JSPMP 平台记录系统的状态信息,避免侵入遗留系统的同时解决了一些工作流应用的难题,例如回退就可以使用分支实现而不会丢失业务数据。

潜在问题:1)不完全支持移除活动节点的过程定制,JPDL 虽可以定义缺失业务活动的过程定义,例如没有“审批报告”的任务报告过程,但只有在系统中不依赖此活动数据的前提下才可行,局限性很明显。2)新增活动定义,JPDL 只能使用遗留系统已有的业务活动定制已有的业务过程。如果用户需要增加业务单元,就不能直接使用 JSPMP 平台。但是,JSPMP 平台的可扩展性保证,它能立即利用系统新增的业务活动。

6　结语

本文介绍了利用工作流引擎定制遗留系统业务过程定义的平台 JSPMP。通过实际应用案例,证明 JSPMP 可为以过程为中心的 J2EE 遗留系统增加工作流定制支持。然而,JSPMP 平台的适用范围被系统的技术架构与实现所限制,有必要进一步扩展 JSPMP 平台,使其能够应用于更多类型的遗留系统。为了做到这一点,可以扩展 JSPMP 平台自身,增加一些组件,不完全依赖于遗留系统已有的逻辑单元,例如独立的数据共享组件、角色控制组件等。

参考文献:

[1] 袁峰,李明树. 基于从 SPEM 到 XPDL 的转换支持软件过程的执行[J]. 软件学报,2007,18(9):2141-2152.

[2] 罗海滨,范玉顺,吴澄. 工作流技术综述[J]. 软件学报,2000,11(7):899-907.

[3] 方进,王铁成,石志宽. 基于 UML 的工作流建模[J]. 计算机工程与设计,2004,25(9): 1572-1575.

[4] BECKER J, MUEHLEN M Z. Workflow application architectures: classification and characteristics of workflow-based information systems[EB/OL]. [2008-11-20]. http://www.workflow-research.de/Publications/PDF/JOBE. MIZU. MAGI-WFHandbook (2002).pdf

[5] BRAHE S, SCHMIDT K. The story of a working workflow management system[C]// Proceedings of the 2007 international ACM Conference on Supporting Group Work. New York: ACM, 2007:249-258.

[6] 徐斌. 支持异地协同遗留系统再工程的软件工程方法研究[D]. 杭州:浙江大学,2004.

[7] 詹剑锋,程虎. 基于 Mobile Agent 技术的遗留系统再工程方法[J]. 软件学报,2002, 13(12):2343-2348.

(上接第 1759 页)

系的数据约简方法相比,该方法不但考虑了决策系统中属性取值的序值性,还避免了基于等价关系的属性约简方法对数据噪声过于敏感的不足。此外,我们还可以通过改变相关度阈值来调节算法的约简精度,从而使算法能在数据约简及保持数据完备性之间求得平衡,这一特性更符合实际应用的要求。

参考文献:

[1] 唐彬, 李龙澍, 李伟, 等. 一类对 Jenolek 属性约简算法的新的改进方法[J]. 系统仿真学报,2005,17(05):1087-1091.

[2] 邓大勇, 黄厚宽, 李向军. 不一致决策系统中约简之间的比较[J]. 电子学报,2007,35(02): 252-255.

[3] 支天云, 苗夺谦. 二进制可辨矩阵的变换及高效属性约简算法的构造[J]. 计算机科学,2002,29(2):140-143.

[4] 蒙祖强,史忠植. 一种新的基于简化二进制可辨矩阵的相对约简算法[J]. 控制与决策,2008, 23(9) : 976-980.

[5] 董超俊, 刘智勇, 刘贤坤. 基于粗糙集的区域交通控制交通量属性约简[J]. 系统仿真学报,2006, 18(6): 1524-1528.

[6] TSUMOTO SHUSAKU. Automated extraction of medical expert system rules from clinical databases based on rough set theory [J]. Information Sciences, 1998, 112(1): 67-84.

[7] 黄金杰, 李士勇, 左兴权. 一种 T-S 型粗糙模糊控制器的设计与仿真[J]. 系统仿真学报,2004, 16(3): 480-484.

[8] GRECO S, MATARAZZO B, SLOWINSKI R. Rough approximation by dominance relations[J]. International Journal of Intelligent Systems, 2002, 17(2): 153 - 171.

[9] SHAO MINGWEN, ZHANG WENXIU. Dominance relation and rules in an incomplete ordered information system[J]. International Journal of Intelligent Systems, 2005,20(1): 13 - 27.

[10] YANG XIBEI, YANG JINGYU, WU CHENG, *et al*. Dominance-based rough set approach and knowledge reductions in incomplete ordered information system[J]. Information Sciences, 2008,178(4): 1219-1234.

[11] QI YUNSONG, SUN HUAIJIANG, YANG XIBEI, *et al*. Approach to approximate distribution reduct in iincomplete ordered decision system[J]. Journal of Information and Computing Science, 2008, 3(3): 189-198.

[12] TSUMOTO S. Statistical evidence for rough set analysis[C]// Proceedings of the 2002 IEEE International Conference on Fuzzy Systems. New York: IEEE, 2002: 757-762.

[13] 邓聚龙. 灰色系统理论教程[M]. 武汉:华中理工大学出版社,1990.