小样本条件下基于属性权重 Shapley 值分配的粗糙集决策模型

李志远, 刘思峰†, 杜俊良, 方志耕, 陶秋澄

(南京航空航天大学 经济与管理学院, 南京 211106)

摘　要: 小样本条件下根据粗糙集理论构建的决策规则, 受数据来源偶然性误差影响较大, 个别数据样本难以反映真实知识关系. 为解决小样本条件下粗糙集决策规则可信度未知的问题, 提出信息区分量、属性影响方向等概念, 运用 Shapley 值法进行进行属性权重分配, 求取每个属性对决策结果的影响方向, 进而得出决策规则的参考信度, 以寻求真实可信且适合工程实际的决策规则. 最后经实例论证了新方法的可行性以及对数据来源误差的分辨能力.

关键词: Shapley 值; 粗糙集; 决策规则; 信度

中图分类号: TP182　　　**文献标志码:** A

DOI: 10.13195/j.kzyjc.2020.1709

Rough Set Decision-making Model based on Shapley Value Assignment of Attribute Weight under the Condition of Small Sample

LI Zhiyuan, LIU Sifeng†, DU Junliang, FANG Zhigeng, TAO Qiucheng

(College of Economics and Management, Nanjing University of Aeronautics and Astronautics, Nanjing 211106, China)

Abstract: The decision rules based on rough set theory under the condition of small sample are greatly affected by the chance error of the initial data, and individual data samples are difficult to reflect the true knowledge relationship. To solve the problem of unknown reliability of rough set decision rules under the condition of small sample, concepts such as the amount of information distinction and the influence direction of attributes are proposed, and attribute weights are assigned using the Shapley value method. The influence direction of each attribute on the decision result is obtained, and the reference reliability of the decision rule is obtained to seek credible and suitable decision-making rules for engineering. Finally, the feasibility of the new method and the ability of discriminating the error of the data source are demonstrated through cases.

Keywords: Shapley value; rough set; decision rule; reliability

0　引　言

波兰学者 Pawlak 提出的粗糙集理论[1], 提供了一种处理不确定性问题的途径. 粗糙集理论中, 通过对论域数据集合按不同属性进行分类, 在经计算分类质量求约简等步骤后, 可得到相应的决策规则. 在经典粗糙集基础上, 许多学者也提出各类拓展的粗糙集模型, 如概率粗糙集[2-3] 、多粒度粗糙集[4-5] 等.

传统的粗糙集模型是在理想的数据环境中建立的[6], 在现实数据为小样本情形的应用中受限. 由于论域中数据集合的数量局限性, 小样本条件下极个别数据样本的存在会使得最终的决策规则受较大数据偶然性误差, 尤其是对确定性规则影响较大. 从统计学角度来看, 小样本数据源的偶然因素造成了粗糙集论域集合的代表性误差[7], 即部分现象不能代表整体规律, 最终会影响决策规则的准确性. 从工程实际的角度来看, 实际的输出往往达不到决策规则中的理想置信度. 因此, 建立适用于实际的粗糙集决策规则十分重要. 在粗糙集决策规则上, 现有的研究多偏向于多属性决策. Greco 等[8] 在经典粗糙集模型上进行改进, 提出了基于属性优势区分的优势粗糙集模型, 实质上是一种多属性决策方法. 而后大量学者对优势粗糙集进行改进, Wojciech Kotłowski 等[9] 针对噪声情况下粗糙近似的缺点提出具有单调性约束的概率优势粗糙集模型. 何亚群等[10] 提出一种拓

收稿日期: 2020-12-08；**修回日期:** 2021-07-12.

基金项目: 国家自然科学基金项目 (面上项目)(72071111)；国家科技部科技创新引智基地项目 (G20190010178)；中央高校基本科研业务费 (NC2019003)；南京航空航天大学研究生创新基地 (实验室) 开放基金项目 (kfjj20200908).

†通讯作者. E-mail: sfliu@nuaa.edu.cn.

展优势关系,解决了不完备偏好信息下的多属性决策问题. 近年来,计算机及人工智能技术的进步使得粗糙集决策出现了许多新的研究方向,诸如: “三支决策”[11-13]、“粒计算”[14-17]等. Xu[18]提出一种基于属性粒度和属性值粒度的多粒度粗糙集模型(MRAAV),为在多粒度粗糙集模型中构建多粒度空间提供了方法. Luo 等[19]为解决与不完整信息有关的语义问题,提出了 Pawlak 粗糙集分析的四步模型. Khan 等[20]定义了概率犹豫模糊粗糙集 (PHFR) 的运算法则,并提出基于 PHFR 对应的决策方法. Xue[21]等将直觉模糊集和多粒度粗糙集结合,给出多粒度直觉模糊粗糙集模型并基于三支决策模型设计了决策规则提取算法. 陈泽华[22]基于极概念和极概念格,提出无冗余属性的决策规则获取算法. 骆公志[23]提出基于覆盖的加权多重代价决策粗糙集模型以弥补传统决策粗糙集模型忽略了多重代价矩阵重要性的不足. 上述研究均未考虑到现实中由于各种原因导致的数据集不完备的因素而致使的决策误差. 唐建国[24]针对信息不完备情形给出决策规则可信度的概念,认为合理的可信度定义应与信息的完备程度有关. Ramasubramanian[25]综合考虑信息不完备和信息不相容两类可信度影响因素,重新设计了可信度计算方法. 朱劼[26]基于这两类可信度设计了用于评价临床检验专家系统质量的可信度计算模块. 上述研究的可信度概念虽然能为决策规则的准确性提供参考,但可信度本身是基于每个样本完全准确的角度出发,仍不能解决小样本中可能存在错误样本、偶然样本等问题. 从决策规则的提取上作出改进是该问题的另一种解决办法. 骆公志等[27]提出了限制扩展优势关系以降低分类决策误差率,但该方法对于单一属性极端情况个体并不适用. 陶志[28]由此提出一种优势关系比限制扩展优势关系宽松的先验概率优势关系决策模型. 李旭等[29]首次提出了带权的决策表,为粗糙集属性约简提供了新的思路. 虽然这些研究考虑到了小样本条件下的数据误差问题,但仍未考虑到属性间的相互影响,因此无法充分挖掘出给定的数据集合中的关系. 属性带权决策表给解决小样本情形问题提供了有效思路.

为解决上述问题将 Shapley 值法引入粗糙集属性权重分配中. 1953 年 Shapley[30]提出的 Shapley 值法,其思想是从个体对集体贡献量的角度进行“利益”分配,在处理群体分配决策问题中具有严格、公正的优点. Shapley 值法被证明是一种解决集体中个体“利益”分配的行之有效的方法,在多个领域得到了应用[31-34]. 本文将利用 Shapley 值法考虑属性间的联合作用,有望得出合理、客观的属性权重分配方法,进而生成适合实际应用的决策规则.

基于上述分析,本文从提出适用于实际的决策规则、减小小样本条件下数据来源偶然性误差的目的出发,重点考虑小样本数据集的输出信度,利用 Shapley 值法建立适用于小样本条件背景的带权决策规则模型,并通过实验说明,验证了该模型具有的优势,为粗糙集的应用提供了参考价值.

1 条件属性集的信息区分量及区分力

针对原始信息系统中数据集合数量有限而导致蕴含信息量匮乏的问题,本文提出一种有利于对数据深层关系进行挖掘的模型. 对于一个粗糙集信息系统而言,不同对象因具备不同的条件属性组合而致使决策属性不同,进而被归为不同类. 换个角度而言,可以视为不同条件属性相组合而产生具备不同决策属性的对象. 基于上述思想,本文从条件属性集对决策属性造成影响的角度出发,先后探究条件属性的影响权重和条件属性对决策属性的影响方向,进而推导出决策规则的参考信度计算方法,为小样本数据集识别不可信数据提供方法.

条件属性集的影响权重应根据条件属性集对决策属性的影响程度来分配,而为描述每个条件属性对决策属性的影响程度,本节采用条件属性集的信息区分量和区分力概念来刻画这种影响程度.

定义 1 对于信息系统 $S=(U, C\cup D, V, f)$,设条件属性集 Q 和 R, Q 和 R 均为条件属性集 C 的子集,且两者互补. 若存在对象 u_i 和 u_j,且 u_i 与 u_j 属性 R 完全一致,而属性 Q 不一致,使得 u_i 与 u_j 的决策属性 D 不同,则称 u_i 和 u_j 是由属性集 Q 可区分的. 亦即: 对给定的 $Q\subseteq C, R\subseteq C$, $R=C-Q$, 以 u_i/Q 表示对象 u_i 的属性集 Q 取值,若 $\exists\, u_i$、$u_j\subseteq U$,有 $u_i/Q\neq u_j/Q, u_i/R=u_j/R$,并且 $u_i/D\neq u_j/D$,则称对象 u_i 和 u_j 是由属性集 Q 可区分的. u_i 和 u_j 被 Q 所区分,记作 $(u_i\mid u_j)_Q$.

定义 2 若对于给定的条件属性集 Q,共有 k 组 u_i 和 u_j 能被 Q 区分 $(i,j=1,2,...,n)$,则称条件属性集 Q 的信息区分量为 k,记作 $D_f(Q)=k$.

定理 1 若信息系统中共有 n 组对象,则该系统中任意一个条件属性集 H 的信息区分量 $D_f(Q)$ 均满足 $0\leqslant D_f(Q)\leqslant \dfrac{n(n-1)}{2}$.

证明 从 n 组对象中选取一对,至多存在 $C_n^2=\dfrac{n(n-1)}{2}$ 种组合. 当 $D_f(Q)=\dfrac{n(n-1)}{2}$ 时,说明属性集 H 是唯一影响决策属性的属性集,其余属性集对决策属性集无影响. 证明成立.

定义 3 称 $C_d=\dfrac{2k}{n(n-1)}$ 为对应条件属性集的信息区分力.

C_d 反映了一个条件属性集在整体中的个体影响程度.C_d 越大, 代表该条件属性集在对决策属性的影响上越占主导地位.

条件属性集的信息区分量体现了该条件属性集在整个信息系统中对决策属性的影响程度, 其本质是条件属性集对决策属性所造成的影响量. 在同一个信息系统中, 条件属性集的信息区分量越大, 其对决策属性的影响程度就越高.

2 基于 Shapley 值的属性权重分配

每个条件属性对决策属性的影响程度是有限的, 当所有条件属性集组成为整体, 即全集时, 其信息区分力最强,影响程度最高. 但以全集划分论域时,实际上是每个条件属性均参与了对决策属性的影响, 即可以视为每个条件属性均对决策属性有一定影响程度, 看成作出了“影响贡献”. 为厘清各个条件属性对决策属性作出的“影响贡献”, 进而按“影响贡献”求取条件属性的权重, 本文提出使用 Shapley 值分配条件属性权重.

定义 4 [32] 对于一个合作总体集合 I, 设 I 的任意合作组合子集 s 均对应一个合作收益函数 $v(s)$, 且满足 $v(s_1\cup s_2)\geqslant v(s_1)+v(s_2), s_1\cap s_2=\varnothing$, 则在合作总体中每个成员 i 从合作中应得的收益, 即 Shapley 值为:

$$\phi_i=\sum_{i\in s_i}\frac{(n-|s_i|)!(|s_i|-1)!}{n!}[v(s_i)-v(s_i\backslash i)] \quad (1)$$

式 (1) 中, s_i 为总体合作集合 I 中所包含成员 i 的所有子集, $|s_i|$ 为合作组合 s_i 中的成员数目, $s_i\backslash i$ 为 s_i 合作组合中除去 i 成员的组合, 后续将式 $\dfrac{(n-|s_i|)!(|s_i|-1)!}{n!}$ 表示为 $w(|s_i|)$.

基于 Shapley 值的属性权重分配方法为：将所有条件属性组成的集合看作合作总体, 条件属性的信息区分量作为收益函数. 若干个条件属性合并一起将论域划分, 计算得到的信息区分量称为这些条件属性的合作收益. 运用 Shapley 值法的前提是其参与对象满足相应条件, 即任意参与对象的“合作”收益不小于各自“单干”收益之和.

命题 1 将条件属性视为参与对象, 信息区分量视为收益时, 任意条件属性的“合作”信息区分量不小于各自“单干”信息区分量之和.

证明 设有总体条件属性集合 U, 条件属性集合 P、Q、R, 决策属性集合 D. 且 $P, Q, R\subseteq U$, $P\cap Q=\varnothing, P\cap R=\varnothing, Q\cap R=\varnothing, P\cup Q\cup R=U$. 设存在对象 u_i 与 u_j, w_i 与 $w_j, u_i, u_j, w_i, w_j\in U$, 且对象 u_i 与 u_j 被 P 区分, 对象 w_i 与 w_j 被 Q 区分, 即 $(u_i\mid u_j)_P, (w_i\mid w_j)_Q$. 若对象 u_i 和 u_j 的属性集 P 取值相同则记为 $u_i\,/\,P=u_j\,/\,P$. 首先证明 $P\cup Q$ 的信息区分量至少等于 P 和 Q 各自信息区分量之和: 依 $(u_i\mid u_j)_P$ 可知对象 u_i 和 u_j 满足 $u_i\,/\,P\neq u_j\,/\,P$, $u_i\,/\,(Q\cup R)=u_j\,/\,(Q\cup R), u_i\,/\,D\neq u_j\,/\,D$. 中间式亦等价于 $u_i\,/\,Q=u_j\,/\,Q$、$u_i\,/\,R=u_j\,/\,R$. 所以由属性相同与否的判定方法有 $u_i\,/\,(P\cup Q)\neq u_j\,/\,(P\cup Q)$. 于是对象 u_i 和 u_j 满足 $u_i\,/\,(P\cup Q)\neq u_j\,/\,(P\cup Q), u_i\,/\,R=u_j\,/\,R$, $u_i\,/\,D\neq u_j\,/\,D$, 由定义 1 有 $(u_i\mid u_j)_{P\cup Q}$. 同理由 $(w_i\mid w_j)_Q$ 有 $(w_i\mid w_j)_{P\cup Q}$. 故得被 P 或 Q 区分的对象组合也必能被 $P\cup Q$ 区分, 所以由定义 2 知 $P\cup Q$ 的信息区分量至少等于 P 和 Q 各自信息区分量之和. 其次再证明 $P\cup Q$ 的信息区分量可能大于 P 和 Q 各自信息区分量之和: 若对于上述四个对象中, 恰巧有对象 u_i 与 w_i, 满足 $u_i\,/\,R=w_i\,/\,R$, $u_i\,/\,(P\cup Q)\neq w_i\,/\,(P\cup Q), u_i\,/\,D\neq w_i\,/\,D$, 则有 $(u_i\mid w_i)_{P\cup Q}$, u_i 和 w_i 对 $P\cup Q$ 贡献了 1 个信息区分量. 下面分为三种情形探讨 u_i 和 w_i 对 P 以及 Q 贡献的信息区分量: (1) u_i 与 w_i 能被 P 或 Q 之一区分, u_i 和 w_i 对 P 和 Q 共贡献 1 个信息区分量; (2) u_i 与 w_i 不能被 P 或 Q 任何一个区分, 无信息区分量贡献; (3) u_i 与 w_i 能既能被 P 区分又能被 Q 区分, 共贡献 2 个信息区分量. 对于 (1) 情形, 条件属性集 $P\cup Q$ 的信息区分量仍等于 P 和 Q 的信息区分量之和; 对于 (2) 情形, 条件属性集 $P\cup Q$ 的信息区分量则超越了 P 和 Q 的信息区分量之和. (3) 情形矛盾, 不可能存在, 因为给定的条件属性集 P、Q、R 是互补且互不相交的, 故式 $u_i\,/\,(Q\cup R)=w_i\,/\,(Q\cup R)$ 与式 $u_i\,/\,(P\cup R)=w_i\,/\,(P\cup R)$ 不可能同时成立, 若同时成立则可推得 $u_i\,/\,(P\cup Q\cup R)=w_i\,/\,(P\cup Q\cup R)$ 成立, 即 $u_i\,/\,U=w_i\,/\,U$ 成立, 意味着对象 u_i 与 w_i 的所有条件属性值完全一致, 这与 u_i, w_i 能被 P 和 Q 区分矛盾. 因此 $P\cup Q$ 的信息区分量可能大于 P 和 Q 各自信息区分量之和. 综上, 任意条件属性的“合作”信息区分量不小于各自“单干”信息区分量之和, 得证.

定义 5 设有条件属性 $C_j, j=1,2,...,n$, 则称

$$\eta_j=\frac{\phi_{C_j}}{\sum\limits_j\phi_{C_j}} \quad (2)$$

为条件属性 C_j 对决策属性的影响权重. 式 (2) 中, ϕ_{C_j} 为条件属性 C_j 的 Shapley 值.

3 小样本条件下基于参考信度的粗糙集决策模型

本文模型主要针对小样本条件, 因为小样本表现出大概率的偶然性和极小概率的必然性, 由少量样本组成的研究对象相较大样本而言, 其不确定性更高, 出现错误数据的影响相对更大. 对于粗糙集来说, 小样本条件下导出的决策规则出错的影响也更严重. 而当拥有大量样本时, 一条决策规则将有多个支持样本, 错误样本带来的影响被正常样本所稀释. 基于此, 为充分利用小样本信息, 寻求小样本总体的内在规律, 甄别样本中的偏差或错误数据, 本文提出还原未知信息的条件属性影响方向概念, 并以此提出参考信度及决策规则可信临界判定条件.

3.1 考虑还原未知信息的属性影响方向

在求得各条件属性的权重后, 仍需考虑各单个属性对决策属性的影响方向或程度, 才能生成决策规则对信息系统总体进行评估. 考虑单个条件属性, 对于拥有同样此属性的不同对象集而言, 其对应决策属性的影响结果一般是多样化的, 产生这一点主要是此属性与其余属性共同作用的结果. 因此, 在考虑单个条件属性对决策属性的影响时, 需考虑该属性与其他各类属性所有搭配情况下的决策属性作用指向方向.

基于上述考虑, 设对于条件属性集 P, Q 和 R, 三者互补且无相交, 若属性集 Q 和 R 分别有 m 和 n 种不同取值情形, 则显然考虑属性集 P 对决策属性的影响方向时应分 $m{\cdot}n$ 种情况. 而一般而言, 给定的信息系统往往不具备齐全的信息, 部分属性组合的决策属性是未知的. 因此, 需要对未知信息进行还原. 下面提出如下方法计算条件属性对决策属性的影响方向.

定义 6 对于给定的信息系统 $S=(U,C\cup D,V,f)$, C 为条件属性集, D 为决策属性集. 设有条件属性 $P, P\subseteq C$, 则称

$$\vec{d}_{u_i,P}=\frac{Num_{u_i,P}}{\tau_P}\cdot\vec{D}_{u_i} \tag{3}$$

为对象 u_i 中条件属性 P 对决策属性 D 的影响方向. 其中 $Num_{u_i,P}$ 为对象集中已经出现包含对象 u_i 在内的, 满足条件属性 P 取值与对象 u_i 相同且其余条件属性组合互不重复、决策属性取值与对象 u_i 相同的对象个数; τ_P 为除 P 外其余条件属性之间取值的应有组合个数; $\vec{D}_{u_i}$ 为对象 u_i 的决策属性取值.

若信息系统中存在理论上应有而实际未出现的属性取值组合, 则需对式 (3) 进行修正. 文献 [24] 指出这是信息不完备的情况, 应对缺损的信息进行补全处理. 在小样本情形的粗糙集里, 样本 (对象) 数量一般只有十几个甚至几个, 所以缺损信息的情况较为常见. 提出适用于小样本情形的属性影响方向修正算式为

$$\vec{d}_{u_i,P}=\frac{Num_{u_i,P}+\dfrac{1}{m}\delta_P}{\tau_P}\cdot\vec{D}_{u_i} \tag{4}$$

式 (4) 中, δ_P 为除条件属性 P 外其余条件属性之间取值组合应有而未出现的个数; m 为决策属性取值类数.

3.2 参考信度及决策规则可信临界判定条件

受数据偶然性影响, 小样本数据可能混有偏差数据甚至错误数据, 在此条件下粗糙集得出的决策规则需要一个能够衡量数据和决策规则真实性的参考标准.

定义 7 在粗糙集导出的决策规则中, 设有所有条件属性 $C_j, j=1,2,...,n$, 称

$$p_i=\sum_{j=1}^{n}\frac{\vec{d}_{u_i,C_j}\cdot\eta_j^{\mathrm{T}}}{\vec{D}_{u_i}} \tag{5}$$

为决策规则 i 的参考信度 p_i. 其中 u_i 为决策规则 i 所取自的对象.

定理 2 设信息系统 $S=(U,C\cup D,V,f)$, 决策属性 D 共有 m 种取值, 则在由 S 推导出的决策规则表中, 当

(1) 决策规则 i 的参考信度 $p_i\leqslant\dfrac{1}{m}$ 时, 该决策规则不可信;

(2) 决策规则 i 的参考信度 $p_i>\dfrac{1}{m}$ 时, 该决策规则可信;

证明 决策规则是否可信是看该规则是否符合样本总体的内在规律, 其临界情况是决策属性取值是随机取值, 与条件属性无关. 设想将信息不完备系统里未出现的样本视为假想样本, 该样本所有属性的取值均未在样本总体中出现. 依 3.1 节缺损信息补全的思想, 其决策属性应按随机取值补全. 于是可将该假想样本视为决策规则是否可信的临界样本. 依式 (4) 和式 (5) 知该临界样本 u_0 作为决策规则的参考信度为

$$\begin{aligned}p_0&=\sum_{j=1}^{n}\frac{\vec{d}_{u_0,C_j}\cdot\eta_j^{\mathrm{T}}}{\vec{D}_{u_0}}\\&=\sum_{j=1}^{n}\frac{Num_{u_0,C_j}+\dfrac{1}{m}\delta_{C_j}}{\tau_{C_j}}\cdot\eta_j^{\mathrm{T}}\end{aligned}$$

上式中 $Num_{u_0,C_j}=0$, $\delta_{C_j}=\tau_{C_j}$, 因为该样本所有属性的取值均未在样本总体中出现. 于是临界样本的参考信度为 $p_0=\dfrac{1}{m}$. 当决策规则 i 的参考信度满

足 $p_i \leqslant p_0 = \dfrac{1}{m}$ 时, 由于样本不变, 所以 $\dfrac{1}{m}\delta_{C_j}$、τ_{C_j} 不变, 只能是 Num_{u_i,C_j} 偏小, 也即满足决策规则 i 这类条件属性组合所得出的决策属性是 $\vec{D}_{u_i}$ 的支持数偏少, 甚至不如假想随机样本支持数多, 意味着该决策规则偏离了样本总体水平, 违背了样本总体的内在联系和规律. 同理可得 $p_i > \dfrac{1}{m}$ 的含义是该样本符合样本总体的内在规律, 是可信的. 所以 (1)、(2) 两条判断准则成立, 得证.

基于定义 7 和定理 2, 可以依据决策规则的参考信度是否达标来判断出决策规则是否可信, 识别出其取自样本是否为误差样本甚至错误样本.

3.3 时间复杂度分析

参考信度涉及信息区分量、Shapely 值和属性影响方向的计算. 时间复杂度主要产生在信息区分量及属性影响方向的计算上, 计算这两者本文使用下列算法.

算法 1 基于属性权重 Shapley 值分配的粗糙集决策模型信息区分量算法.

输入: 信息系统 $S = (U, C \cup D, V, f)$, 对象集 $U = \{u_i\}$, 条件属性集 $C = \{C_j\}$, $j \in \{1, 2, ..., p\}$, 决策属性 D.

输出: 条件属性集子集的信息区分量集 $D_f = \{D_f(C'_1), D_f(C'_2), ..., D_f(C'_{2^p-1})\}$. C'_k 为条件属性集的非空子集.

step 1: 对条件属性集 C 的 $2^p - 1$ 个非空子集 C'_k, $k \in \{1, 2, ..., 2^p - 1\}$ 分别构造 p 维计数向量 $Count_k = (Count_k(j))$, $Count_k(j) = \begin{cases} 0, C_j \notin C'_k \\ -1, C_j \in C'_k \end{cases}$. 并合并得 $(2^p - 1) \times p$ 维计数矩阵 $Count = [Count_1; Count_2; ...; Count_k]$. 建立 p 维零向量 $X = (X(j))$, $X(j) = 0$. 建立 C'_k 的信息区分量向量, 为初始零量 $D_f = (D_f(C'_k))$, $D_f(C'_k) = 0$.

step 2: 比较对象 u_1 与 u_2, if $u_1 / D = u_2 / D$, then 跳过 step 2; else, for $j \in \{1, 2, ..., p\}$, 计算向量 X, 其中 $X(j) = \begin{cases} 1, u_1/C_j \neq u_2/C_j \\ 0, u_1/C_j = u_2/C_j \end{cases}$.

step 3: for $k \in \{1, 2, ..., 2^p - 1\}$, 令 $Count'_k = Count_k + X$. If $Count'_k$ 中存在 1 元素, then $D_f(C'_k) + 1$.

step 4: 重置 $X = (X(j)), X(j) = 0$, 遍历其余对象: 比较 u_1 与 $u_3, ..., u_{n-1}$ 与 u_n, 重复 step 2 和 step 3. 返回 $D_f = (D_f(C'_k))$.

算法 2 基于属性权重 Shapley 值分配的粗糙集决策模型属性影响方向算法.

输入: 信息系统 $S = (U, C \cup D, V, f)$, 对象集 $U = \{u_i\}$, 条件属性集 $C = \{C_j\}$, $j \in \{1, 2, ..., p\}$, 决策属性 D.

输出: 每个对象的属性影响方向向量 $\vec{d} = (\vec{d}_{u_i,C_j})$

step 1: 选定对象 u_1, 令 $U' = U - \{u_1\}$, for each $u_m \in U'$, $m \in \{2, 3, ..., n\}$, 计算同时满足 $u_1 / C_1 = u_m / C_1$, $u_1 / (C - C_1) \neq u_m / (C - C_1)$, $u_1 / D = u_m / D$ 且不重复的对象集个数 Num_{u_1,C_1}.

step 2: 计算其余条件属性应有的取值组合总数 τ_P, 以及未出现的取值组合个数 δ_P, 运用式 (4) 进行属性影响方向修正.

step 3: 将对象 u_i 的其余条件属性 C_j 依次替换 C_1, 重复 step 1 和 step 2.

step 4: 选定其余对象, 重复 step 1～step 3.

对于算法 1, step 2 需依次比较两组对象的所有属性值, 共 p 个, step 3 需依次遍历计数矩阵的每一行, 共 $2^p - 1$ 行, step 4 重复迭代了 $\dfrac{n(n-1)}{2}$ 次, 所以求得算法 1 的时间复杂度近似为 $O(n^2 2^p)$.

对于算法 2, step 1 遍历其余对象集执行 $n - 1$ 次, Step3 替换条件属性迭代了共 p 次, Step4 重复迭代 n 次, 所以求得算法 2 的时间复杂度近似为 $O(pn^2)$.

根据时间复杂度分析结果, 算法 1 的时间复杂度对条件属性个数 p 而言为指数级, 原因是当条件属性个数很大时, 计算 Shapley 值所需的条件属性组合存在组合爆炸的问题. 但实际上, 粗糙集决策规则是建立在经属性约简后的数据集基础之上, 其条件属性个数一般较少, 不会存在组合爆炸问题, 因此上述算法仍然适用.

3.4 建模步骤

基于上述分析, 提出含参考信度作为参考的决策规则搭建步骤如下:

步骤 1 收集待导出决策规则的数据集, 求出各条件属性及其所有属性组合的信息区分量;

步骤 2 计算各条件属性的 Shapley 值, 将其进行归一化处理得出各属性的权重 η_j.

步骤 3 求各条件属性在所有对象集中对决策属性的影响方向 $\vec{d}_{u_i,P}$. 即单独考察某个条件属性, 分析该条件属性下决策属性的取值分布情况.

步骤 4 计算粗糙集输出决策规则中各对象的决策参考信度 p_i, 依定理 2 判断决策规则是否可信.

4 实例分析

本节通过实例验证所提小样本条件下基于属性权重 Shapley 值分配的粗糙集决策模型的有效性, 选取经典粗糙集、变精度粗糙集和文献 [25] 所提可信度模型作对比.

例 1 实验使用 kaggle 数据集, 数据集名为

Gender Classification, 对象数共 66 个, 条件属性集包括"Favorite Color"、"Favorite Music Genre"、"Favorite Beverage"、"Favorite Soft Drink", 决策属性为"Gender", 取值为二值型. 该数据集取自 21 个国家不同专业的大学生, 旨在探寻个人兴趣偏好与性别的关系. 实验硬件平台为 Intel(R) Core(TM) i5-5200U, 2.2Hz 和 4G 内存, 所有求解均使用 matlab 编程运行.

首先对该数据集进行约简, 得知四个条件属性均为核属性. 导出的决策规则共 62 条, 除第 5、18、39、44 条的支持数为 2 以外, 其余决策规则的支持数均为 1. 依式 (1) 和式 (2) 求得条件属性集的 Shapley 值向量为 [198,321.5,328.33,241.17], 条件属性权重向量为 [0.1818,0.2952,0.3015,0.2215]. 由于决策属性为二值型, 故依据定理 2, 知样本可信的临界参考信度为 50%. 本文的模型与经典粗糙集、变精度粗糙集 ($\beta=70\%$) 和文献 [25] 所提可信度模型对比结果如图 1 所示.

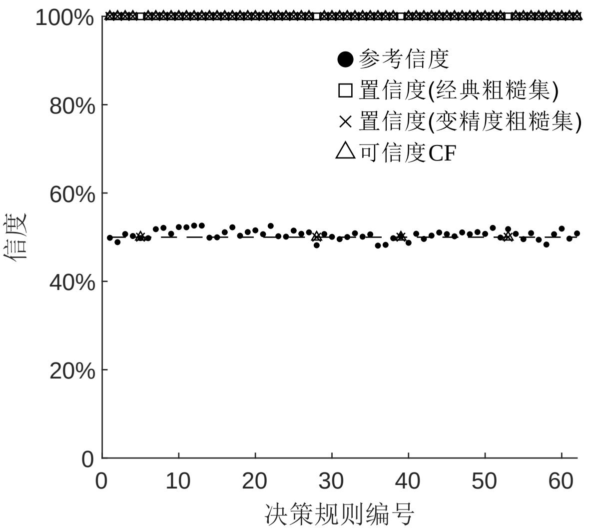

图 1 经典粗糙集、变精度粗糙集、可信度模型和本文模型在决策信度上对比一

由图 1 所示的实验结果可知, 经典粗糙集得出的 62 条决策规则置信度均为 100%, 变精度粗糙集和文献 [25] 所提模型给出的决策规则置信度/可信度除部分为 50% 外, 其余均为 100%. 出现 50% 的情况是因为决策属性存在明显矛盾. 本文模型计算得到的 62 条决策规则参考信度均在 50% 水平波动, 识别出参考信度低于 50% 的不可信样本决策规则数有 18 条. 由此可见, 该模型具备一定的误差样本识别能力.

上述实验数据集由于样本数较多, 不属于小样本范畴, 且四个条件属性的取值组合数较多, 为 $3\times7\times6\times4$ 种, 总体并没有显著的内在规律, 因此运用本文所提模型求得的参考信度都较低, 样本间参考信度差异不大.

例 2 从前例中提取部分样本, 组成具备显著内在规律的新小样本集, 并加入一条测试样本 (表 1 中标 * 样本) 以测试模型的错误样本识别能力, 该测试样本的条件属性与对象 1、2 等类似, 而决策属性则与具备这类条件属性的对象完全相反, 是不符合样本总体规律的一条. 该小样本集如表 1 所示.

表 1 具备显著内在规律的小样本集

对象	Favorite Color	Favorite Music Genre	Favorite Beverage	Favorite Soft Drink	Gender
1	Neutral	Pop	Vodka	Coca_Cola / Pepsi	F
2	Neutral	Rock	Wine	Coca_Cola / Pepsi	F
3	Cool	Rock	Vodka	Coca_Cola / Pepsi	F
4	Cool	Rock	Wine	Coca_Cola / Pepsi	F
5	Cool	Rock	Wine	Fanta	F
6	Cool	Pop	Wine	Wine	F
7	Warm	Pop	Wine	Fanta	M
8	Warm	Pop	Wine	Fanta	M
9	Warm	Pop	Vodka	Fanta	M
10*	Neutral	Rock	Vodka	Coca_Cola / Pepsi	M

对表 1 样本求约简, 得条件属性"Favorite Color"、"Favorite Music Genre"、"Favorite Beverage"是一个约简. 导出决策规则共 8 条, 第 8 条为测试样本所导出. 求得三个条件属性的 Shapley 值向量为 [12, 6.5, 5.5], 属性权重向量为 [0.5, 0.2708, 0.2292]. 本例中本文模型与经典粗糙集、变精度粗糙集 ($\beta=70\%$) 和文献 [25] 所提可信度模型对比结果如图 2 所示.

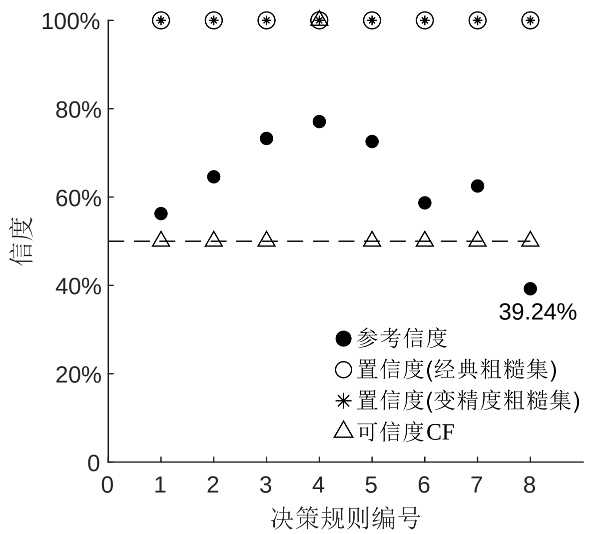

图 2 经典粗糙集、变精度粗糙集、可信度模型和本文模型在决策信度上对比二

图 2 显示经典粗糙集和变精度粗糙集的 8 条决策规则置信度均为 100%, 不能识别出测试样本的错误; 文献 [25] 所提可信度模型求得的可信度虽然能识别出第 8 条测试样本是不完全可信 (可信度 =50%), 但它对其他正常样本的可信度也多为

50%, 以致正常样本与错误样本间不能很好地横向比较以筛选错误样本; 本文所提模型求得的正常样本参考信度均显著大于 50%, 测试样本的参考信度为 39.24% 远小于 50% 的可信临界线.

例 2 结果表明, 本文所提模型可以有效地识别出偏离样本总体规律的错误样本, 能对决策规则的可信程度判断起到一定作用. 同时, 例 1 和例 2 的对比从侧面说明本文模型在样本总体具有较强内在规律时应用效果更为显著.

5 结论

小样本数据因数据偶然性等原因容易出现错误样本, 运用粗糙集导出的决策规则真实可信程度未知. 本文提出信息区分量、属性影响方向等概念, 并通过 Shapley 值计算属性影响权重, 最后推导参考信度的计算方法和决策规则的可信临界判定条件, 基于此建立了适用于小样本条件的基于属性权重 Shapley 值分配的粗糙集决策模型. 结果表明, 该模型可以对决策规则的可信程度给出参考, 且能有效辨别出样本中偏离样本总体规律的偏差数据或错误数据. 另外, 本文模型在样本总体具有较强内在规律时应用效果更为显著. 然而本文需要求取所有的条件属性组合, 其组合个数对模型的时间复杂度影响较大, 接下来可以寻求条件属性组合爆炸造成时间复杂度较高的解决办法.

参考文献 (References)

[1] Pawlak Z. Rough sets: Theoretical aspect of reasoning about data[M]. London: Kluwer Academic Publishers, 1991: 9-42.

[2] Wong S K M, Ziarko W. Comparison of the probabilistic approximate classification and the fuzzy set model[J]. Fuzzy Sets and Systems, 1987, 21(3): 357-362.

[3] Wei L L, Zhang W X. Probabilistic rough sets characterized by fuzzy sets[J]. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 2004, 12(1): 47-60.

[4] Yang X B, Xu S, Dou H L, et al. Multigranulation rough set: A multiset based strategy[J]. International Journal of Computational Intelligence Systems, 2017, 10(1): 277 – 292.

[5] Yu P Q, Li J J, Lin G P. Vector-based approaches for computing approximations in multigranulation rough set[J]. The Journal of Engineering, 2018, 2018(16): 1538-1543.

[6] 杨臻, 邱保志. 混合信息系统的动态变精度粗糙集模型 [J]. 控制与决策, 2020, 35(02): 297-308.
(Yang Z, Qiu B Z. Dynamic variable precision rough set model of mixed information system[J]. Control and Decision, 2020, 35(02): 297-308.)

[7] 王云峰. 统计学原理: 理论与方法 [M]. 复旦大学出版社, 2010.
(Wang Y F. Principles of statistics: theory and methods[M]. Fudan University Press, 2010.)

[8] Greco S, Matarazzo B, Slowinski R. Rough approximation by dominance relation[J]. International Journal of Intelligent Systems, 2012, 17(2): 153-171.

[9] Kotłowski W, Dembczyński K, Greco S, Słowiński R. Stochastic dominance-based rough set model for ordinal classification[J]. Information Sciences, 2008, 178(21): 4019-4037.

[10] 何亚群, 胡寿松. 不完全信息的多属性粗糙决策分析方法 [J]. 系统工程学报, 2004, 19(2): 117-120.
(He Y Q, Hu S S. Rough analysis method of multi_attribute decision making with incomplete information[J]. Journal of Systems Engineering, 2004, 19(2): 117-120.)

[11] 胡峰, 张苗, 于洪. 基于三支决策的主动学习方法 [J]. 控制与决策, 2019, 34(04): 718-726.
(Hu F, Zhang M, Yu H. An active learning method based on three-way decision model[J]. Control and Decision, 2019, 34(04): 718-726.)

[12] Liang D C, Pedrycz W, Liu D, et al. Three-way decisions based on decision-theoretic rough sets under linguistic assessment with the aid of group decision making[J]. Applied Soft Computing Journal, 2015, 29: 256-269.

[13] Xu J F, Miao D Q, Zhang Y J, et al. A three-way decisions model with probabilistic rough sets for stream computing[J]. International Journal of Approximate Reasoning, 2017, 88: 1-22.

[14] 庞阔, 陈思琪, 宋笑迎, 等. 基于粒计算的语言概念决策形式背景分析 [J]. 山东大学学报 (工学版), 2018, 48(06): 74-81.
(Pang K, Chen S Q, Song X Y, et al. Linguistic concept formal decision context analysis based on granular computing[J], Journal of Shandong University(Engineering Science), 2018, 48(06): 74-81.)

[15] Kok V J, Chan C S. GrCS: Granular Computing-Based Crowd Segmentation[J]. IEEE Trans Cybern, 2017, 47(5): 1157-1168.

[16] 徐怡, 王泉，霍思林. 粒计算中基于属性分类的形式概念属性约简 [J]. 控制与决策, 2018, 33(12): 2203-2207.
(Xu Y, Wang Q, Huo S L. Formal concept attribute reduction model based on attribute classification relation[J]. Control and Decision, 2018, 33(12): 2203-2207.)

[17] Niu J J, Huang C C, Li J H, et al. Parallel computing techniques for concept-cognitive learning based on granular computing[J]. International Journal of Machine Learning and Cybernetics, 2018, 9(11): 1785-1805.

[18] Yi X. Multigranulation rough set model based on granulation of attributes and granulation of attribute values[J]. Information Sciences, 2019, 484: 1-13.

[19] Luo J F, Fujita H, Yao Y Y, et al. On modeling similarity and three-way decision under incomplete information in rough set theory[J]. Knowledge-Based Systems, 2019,

191. 105251.

[20] Khan M A, Ashraf S, Abdullah S. et al. Applications of probabilistic hesitant fuzzy rough set in decision support system. Soft Comput, 2020, 24: 16759-16774.

[21] Xue Z A, Zhao L P, Sun L, et al. Three-way decision models based on multigranulation support intuitionistic fuzzy rough sets[J]. International Journal of Approximate Reasoning, 2020, 124, 147-172.

[22] 陈泽华, 宋波, 闫继雄, 柴晶. 基于概念格的不完备信息系统最简规则提取算法 [J]. 控制与决策, 2019, 34(05): 1011-1017.
(Chen Z H, Song B, Yan J X, et al. Concise rule extraction algorithm of incomplete information system based on concept lattice[J], Control and Decision, 2019, 34(05): 1011-1017.)

[23] 骆公志, 许鑫鑫. 基于覆盖的多重代价粗糙决策分析方法 [J]. 计算机科学, 2019, 46(05): 209-213.
(Luo G Z, Xu X X. Multi-cost decision- theoretic rough set based on covering approximate space[J]. Computer Science, 2019, 46(05): 209-213.)

[24] 唐建国. 粗糙集理论处理不完备信息的可信度分析 [J]. 控制与决策, 2002(02): 255-256.
(Tang J G. Reliability analysis to deal with imperfect information by rough set theory[J]. Control and Decision, 2002(02):255-256.)

[25] Ramasubramanian P, Sureshkumar V, Nachiyappan S, et al. Computing Rule Confidence using Rough set and Data Mining[J]. Research Journal of Information Technology, 2010, 2(2): 39.

[26] 朱劼. 医院检验信息系统中的专家系统及其推理机的设计和实现 [D]. 哈尔滨工业大学, 2012.
(Zhu J. Design and implementation of inference engine component and expert system of LIS[D]. Harbin Institute of Technology, 2012.)

[27] 骆公志, 杨晓江, 周德群. 基于限制扩展优势关系的粗糙决策分析模型 [J]. 系统管理学报, 2009, 8(4): 391-396.
(Luo G Z, Yang X J, Zhou D Q. Rough analysis model of multi-attribute decision making Based on limited extended dominance relation[J]. Journal of Systems & Management, 2009, 8(4): 391-396.)

[28] 陶志, 卞文静. 基于先验概率优势关系的粗糙决策分析模型 [J]. 中国民航大学学报, 2013, 31(04): 60-64.
(Tao Z, Bian W J. Rough analysis model based on prior probability dominance relation[J]. Journal of Civil Aviation University of China, 2013, 31(04): 60-64.)

[29] 李旭, 荣梓景, 任艳. 带权决策表的属性约简 [J]. 计算机工程与应用, 2020, 56(12): 54-59. (Li X, Rong Z J, Ren Y. Attribute reduction on weighted decision table. Computer Engineering and Applications, 2020, 56(12): 54-59.)

[30] Shapley L S. Stochastic Games[J]. Proceedings of the National Academy of Sciences of the United States of America, 1953, 39(10): 1095-1100.

[31] 高太光, 王庆, 黄敏, 等. 基于团队合作博弈的自动协商模型 [J]. 控制与决策, 2020, 35(02): 285-296.
(Gao T G, Wang Q, Huang M, et al. Automated negotiation model with collaborative offering of team[J]. Control and Decision, 2020, 35(02): 285-296.)

[32] 曹颖赛, 刘思峰, 方志耕, 等. 系统可靠性退化的单元责任 Shapley 值分配模型 [J]. 机械工程学报, 2017, 53(20): 202-208.
(Cao Y S, Liu S F, Fang Z G, et al. Shapley value allocation model of responsibility during the degradation of system reliability[J]. Journal of Mechanical Engineering, 2017, 53(20): 202-208.)

[33] 陈星兴, 陆朝阳, 刘方毅. 中心性分析与 Shapley 值法在确定恐怖分子网络节点重要性中的应用 [J]. 安全与环境学报, 2019, 19(05): 1676-1684.
(Chen X X, Lu Z Y, Liu F Y. Application of centrality analysis and Shapley value method in determining the importance of the terrorist net-work nodes[J]. Journal of Safety and Environment, 2019, 19(05): 1676-1684.)

[34] Rezaei M, Derhami V. Improving LNMF Performance of Facial Expression Recognition via Significant Parts Extraction using Shapley Value[J]. Journal of Artificial Intelligence and Data Mining, 2019, 7(1): 17-25.

作者简介

李志远 (1997−), 男, 硕士生, 从事不确定性系统理论、可靠性分析等研究, E-mail: zyli2625129@163.com;

刘思峰 (1955−), 男, 教授, 博士生导师, 从事灰色系统理论、复杂装备研制管理等研究, E-mail: sfliu@nuaa.edu.cn;

杜俊良 (1994−), 男, 博士生, 从事粗糙集、冲突分析等研究, E-mail: dujunliang1994@163.com;

方志耕 (1962−), 男, 教授, 博士生导师, 从事可靠性、复杂装备研制管理等研究, E-mail: zhigengfang@163.com.

陶秋澄 (1999−), 男, 本科生, 从事工业工程、复杂装备研制管理等研究, E-mail: 3143389511@qq.com.