计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2018,27(4):117−123 [doi: 10.15888/j.cnki.csa.006277]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于概率覆盖决策粗糙集模型的中医菜谱分析①

雷雪梅, 谢依彤

(北京科技大学 计算机与通信工程学院, 北京 100083)
通讯作者: 谢依彤, E-mail: xin9910@163.com

摘　要: 针对营养决策表规则提取中规则矛盾多、覆盖样例冗余多, 导致有效规则遗漏的问题, 提出概率覆盖决策粗糙集模型. 首先, 对决策粗糙集相关理论进行简要介绍, 给出对应的属性约简和值约简理论和算法. 然后, 在决策粗糙集基础上, 提出概率覆盖模型, 根据值约简需求提出一、二、三度覆盖矩阵, 以解决规则矛盾和冗余问题. 最后, 通过中医菜谱数据提取营养学规则实验, 证明所提模型可有效解决规则矛盾问题, 相比其他常用规则提取模型, 概率覆盖模型所得规则约简力度较高, 矛盾个数较少.

关键词: 中医营养学; 决策粗糙集; 覆盖关系; 值约简; 规则提取

引用格式: 雷雪梅,谢依彤.基于概率覆盖决策粗糙集模型的中医菜谱分析.计算机系统应用,2018,27(4):117–123. http://www.c-s-a.org.cn/1003-3254/6277.html

Analysis of Chinese Medicine Recipe Based on Probabilistic Coverage Decision Rough Set Model

LEI Xue-Mei, XIE Yi-Tong

(School of Computer and Communication Engineering, University of Science and Technology Beijing, Beijing 100083, China)

Abstract: A probabilistic coverage decision-theoretic rough set (PCDTRS) model is proposed in this study to deal with the two main issues in rule acquisition from decision table, *i.e.*, contradiction of extracted rules and redundancy of override sample. Firstly, the basic theories of the decision-theoretic rough set (DTRS) model including the attribute and value reduction algorithms are presented. Subsequently, the probabilistic coverage model is raised based on the DTRS model, and three levels covered matrixes meeting the needs of value reduction are proposed to resolve the aforementioned problems. Finally, the results of a series of experiments on Chinese cookbook nutrition illustrate the feasibility and effectiveness of the PCDTRS model. Compared with other models, the reduction strength and the number of conflicting rules using the PCDTRS model are higher and fewer respectively.

Key words: Chinese medicine nutrition; decision-theoretic rough set (DTRS); probability of covered relation; value reduction; extracting rules

近年来快节奏的生活方式引发的亚健康人群数激增现象, 引起人们对养生学的重视, 饮食营养成为突破口, 现存的西医营养学过分注重化学成分 (如卡路里、碳水化合物等) 对人体的影响, 割裂了人体统一和谐的身体系统[1], 而中医营养学从人体整体把握饮食对人的影响, 已成为主流. 中医营养学起源于中医药学“药食同源”, 人工流传记载为主, 其中大量营养属性值存在元素多元化、数据缺失、数据值模糊的问题, 规则提取后的知识冗余度高且存在冲突项.

应用粗糙集理论进行规则提取可以从已知的各种不完备的数据信息中快速获取知识和规律, 包括属性约简和值约简两个步骤. 对这个问题的探索已经存在

① 收稿时间: 2017-07-03; 修改时间: 2017-07-17; 采用时间: 2017-07-28; csa 在线出版时间: 2018-03-31

一些方法, 文献[2]提出了基于加权变精度容差粗糙集模型; 文献[3]提出了一种基于粗糙集理论的值约简算法; 文献[4]提出了改进不可区分关系的值约简算法; 文献[5]提出一种基于差异关系的变精度粗糙集知识约简算法; 文献[6]等提出了基于广义优势决策函数的决策规则获取方法; 文献[7]改进了邻域粗糙集模型, 并将其应用到故障诊断规则提取中. 然而在分析现有规则提取的算法及数据特点时发现, 算法注重对决策属性之间的关系提取, 对得到矛盾规则不具备良好的处理力能.

因此, 引入具有噪声容忍机制的决策粗糙集模型, 使用 α-正域约简启发式算法约简属性, 克服无关属性对决策值的干扰. 在决策表值约简过程中, 引入可覆盖关系和活跃值等概念, 构建待定矩阵和覆盖矩阵, 提出概率覆盖模型, 处理矛盾的知识规则, 得到普适性较高的规则.

1 决策粗糙集模型

决策粗糙集[8]引入了 Bayes 风险决策理论, 使其在不确定性知识获取和数据处理中具有更加可靠的理M 论依据和语义解释[9], 是对经典 Pawlak 粗糙集理论模型缺乏容错能力的概率拓展.

定义 1. 给定菜谱信息系统$S=(U,C\cup D,f)$, U 为菜谱标号, C 和 D 分别为中医营养学属性集与是否养肠胃的决策属性集, f 是信息函数, $X\subseteq U$, 设行为集合$\{a_P,a_B,a_N\}$表示接受菜谱 x 为养肠胃菜谱的决策行为、待定决策行为、拒绝决策行为, $\lambda_{PP},\lambda_{BP},\lambda_{NP}$表示$x\in X$ 时采取行为a_P,a_B,a_N 所产生的风险损失, $\lambda_{PN},\lambda_{BN},\lambda_{NN}$表示$x\in\neg X$时采取行为$a_P,a_B,a_N$所产生的风险损失. 根据贝叶斯最小风险决策原则, 将其划分为三个部分, 正域$\mathrm{POS}(X)$、边界域$\mathrm{BND}(X)$、负域$\mathrm{NEG}(X)$, 决策粗糙集有如下决策规则:

1) if $P(X|[x]_C)\geq\alpha, x\in\mathrm{POS}(X)$

2) if $\beta<P(X|[x]_C)<\alpha, x\in\mathrm{BND}(X)$

3) if $P(X|[x]_C)\leq\beta, x\in\mathrm{NEG}(X)$

其中, $\begin{cases}\alpha=\dfrac{\lambda_{PN}-\lambda_{BN}}{\lambda_{BP}-\lambda_{PP}+\lambda_{PN}-\lambda_{BN}}\\ \beta=\dfrac{\lambda_{BN}-\lambda_{NN}}{\lambda_{NP}-\lambda_{BP}+\lambda_{BN}-\lambda_{NN}}\end{cases}$, 且$\alpha,\beta\in[0,1]$.

由此, 定义决策粗糙集的下、上近似分别为:

$$\underline{apr}_C^{(\alpha,\beta)}(\mathrm{X})=\{x\in U|P(X|[x]_C)\geq\alpha\} \tag{1}$$

$$\overline{apr}_C^{(\alpha,\beta)}(X)=\{x\in U|P(X|[x]_C)>\beta\} \tag{2}$$

其中$P(X|[x]_C)=\frac{|X\cap[x]_C|}{|[x]_C|}$, $|\cdot|$表示集合中元素的基数, $[x]_C$ 为 x 在属性集 C 下的等价类.

定义 2. $S=(U,C\cup D,f)$为一个菜谱决策表, $\alpha\in[0,1]$, 若属性子集$B\subseteq C$满足下面 2 个条件:

(1) 正域非减性, $|\mathrm{POS}_B^\alpha(D)|\geq|\mathrm{POS}_C^\alpha(D)|$;

(2) 属性独立性, 对任意$\alpha\in B$,

$$|\mathrm{POS}_{B-\{a\}}^\alpha(D)|<|\mathrm{POS}_B^\alpha(D)|$$

称属性子集 B 为属性全集 C 的一个决策粗糙集 α-正域约简[10].

定义 3. 设$S=(U,C\cup D,f)$为一个菜谱决策表, $\alpha\in[0,1]$为条件概率阈值, $a\in C$为单个属性, 则属性 a 的 α-正域全局重要度[11]定义为:

$$\gamma_a^\alpha=\frac{\left|\mathrm{POS}_{\{a\}}^\alpha(D)\right|}{|U|} \tag{3}$$

2 基于决策粗糙集的概率覆盖模型

本节将建立基于决策粗糙集的概率覆盖模型, 解决中医学数据噪声产生的规则矛盾现象.

2.1 属性约简

原决策表中过多的不重要属性会影响规则提取的有效性, 因此, 本文采用决策粗糙集 α 正域约简算法[11]对属性进行约简. 基于决策粗糙集的属性约简引入了决策损失函数确定阈值参数, 改进粗糙集零错误容忍率的局限性. 调整 α 的取值对属性重要度排序, 由专家确定最优属性组合模型, 进而确定 α 取值.

2.2 值约简

从决策表中得到的规则不一定是完全正确的, 原因有二: 一是上百年的中医传承记载, 样本数据本身可能含有矛盾信息, 即条件相同但决策不同; 二是营养数据离散化过程中, 由于菜谱中医属性值是计算求和所得, 忽略各原料之间相互作用, 因此属性程度概念模糊, 分类划分可能导致误差. 从而归纳得到的知识矛盾性大, 准确率低, 鲁棒性弱, 因此, 在决策表值约简过程中, 引入可覆盖关系和活跃值概念, 构建覆盖矩阵.

定义 4. $S=(U,C\cup D,f)$为一个菜谱决策表, $x_i,x_j\in U\times(C\cup D)$, a_i^S、$a_j^S\in C(i\neq j)$分别表示在决策表中任意两个不相等的样本对应的属性值. 缺省集合$T_{a_i}^S=\left\{x_i-a_i^S\right\}$,$T_{a_j}^S=\left\{x_j-a_j^S\right\}$表示在决策表 S 中 x_i 和 x_j 去掉a_i^S,a_j^S后剩余的数据集合:

$$T_{a_i}^S \underline{\cup} a_i^S = x_i$$
$$T_{a_j}^S \underline{\cup} a_j^S = x_j$$

$\underline{\cup}$ 表示 $T_{a_i}^S$ 与 a_i^S 有序合并. 例如: $x_i=\{1,0,2,0\}$, $a_i^S=\{2\}, T_{a_i}^S=\{1,0,0\}$, 有 $T_{a_i}^S \underline{\cup} a_i^S=\{1,0,0\}\underline{\cup}\{2\}=\{1,0,2,0\}=x_i$.

定义 5. $S=(U,C\cup D,f)$ 为一个菜谱决策表, d_i, $d_j\in \mathrm{D}$, 对于菜谱 U 中的任意两条记录 x_i 和 x_j 有:

$$\begin{cases} Tol\left(x_i,x_j\right): T_{a_i}^S \underline{\cup} d_i = T_{a_j}^S \underline{\cup} d_j \\ \overline{Tol}\left(x_i,x_j\right): T_{a_i}^S = T_{a_j}^S \ \&\ d_i \neq d_j \end{cases} \tag{4}$$

对表中的一条菜谱记录 x_i, 去掉其中的一个属性值 a_i^S, 缺省集合 $T_{a_i}^S$ 可以决定决策属性, 说明 a_i^S 不重要, 两条菜谱有相似性, 可以融合, 称 x_i、x_j 为可覆盖关系 $Tol\left(x_i,x_j\right)$; 如果缺省集合不能决定决策属性, 称 x_i 和 x_j 为不可覆盖关系 $\overline{Tol}\left(x_i,x_j\right)$.

定义 6. 设 $S=(U,C\cup D,f)$ 为一个菜谱决策表, 论域 $U=\{x_1,x_2,\cdots,x_n\}, C=\{c_1,c_2,\cdots,c_m\}, D=\{d\}, x_{ij}\in U\times C$.

$$\mathrm{M}(PT)=\left(c_{ij}\right)_{m\times n}\ i=1,2,3,\cdots,m;\ j=1,2,3,\cdots,n \tag{5}$$

定义 $\mathrm{M}(PT)$ 为待定矩阵, c_{ij} 为矩阵元素:

$$c_{ij}=\begin{cases} ?, & T_{a_i}^S \neq T_{a_j}^S \\ *, & Tol\left(x_i,x_j\right) \\ x_{ij}, & \overline{Tol}\left(x_i,x_j\right) \end{cases} \tag{6}$$

其中, "*"为 x_i, x_j 为覆盖关系时的值, "?"为待定矩阵中待评价的值, 定义矩阵元素标 "?"为活跃值 b_{ij}, x_{ij} 是原决策表 S 的值.

定义 7. 一度覆盖矩阵 $\mathrm{M}_{S1}(PT)=(U,C\cup D,f)$, $\mathrm{M}_{S1}(PT)\supseteq\left\{*,x_{ij}\right\}$, 对待定矩阵 $\mathrm{M}(PT)$ 的活跃值 b_{ij} 再判定:

$$b_{ij}=\begin{cases} x_{ij}^S, & T_{b_i}^{\mathrm{M}(PT)}=\emptyset \quad \overline{Tol}\left(x_i,x_j\right) \\ *, & T_{b_i}^{\mathrm{M}(PT)} \neq T_{b_j}^{\mathrm{M}(PT)} \quad Tol\left(x_i,x_j\right) \end{cases} \tag{7}$$

其中 x_{ij}^S 代表在决策表 S 中的取值.

定义 8. 二度覆盖矩阵 $\mathrm{M}_{S2}(PT)=(U,C\cup D,f,\omega)$, $\mathrm{M}_{S2}(PT)\supseteq\left\{*,\omega_i,x_{ij}\right\}$, ω 为菜谱库 U 中规则 R 所覆盖的样例数目: $\omega=[x_i]_R$.

定义 9. 三度覆盖矩阵 $\mathrm{M}_{S3}(PT)=(U,C\cup D,f,\omega,Support,Confidence)$, $\mathrm{M}_{S3}(PT)\supseteq\left\{*,\omega_i,x_{ij},\psi_S,\psi_C\right\}$; $\forall x_i,x_j\in C\cup D$ 且满足 $\overline{Tol}\left(x_i,x_j\right)$, $R|\mathrm{M}_{S2}(PT)|$ 表示矩阵 $\mathrm{M}_{S2}(PT)$ 的规则数目.

错误覆盖率:

$$\theta\left(x_i,x_j\right)=\frac{\left|[x_i]_R-[x_j]_R\right|}{\max\left([x_i]_R,[x_j]_R\right)} \tag{8}$$

约简力度:

$$\mu=\frac{R|\mathrm{M}_{Sk}(PT)|-R|\mathrm{M}_{S(k+1)}(PT)|}{R|\mathrm{M}_{Sk}(PT)|} \tag{9}$$

其中, $k=1,2$.

$[x]_C\in U/C$, $[x]_D\in U/D$, 得到的一个规则 $f([x]_C,C)\to f([x]_D,D)$, $[x]_C\cap[x]_D\neq\emptyset$.

支持度:

$$\psi_S=|[x]_C\cap[x]_D|/|U| \tag{10}$$

置信度:

$$\psi_C=|[x]_C\cap[x]_D|/|[x]_D| \tag{11}$$

为了获得有用的规则, 需要保留重要的属性值, 删减非重要的属性值, 直至形成规则, 值约简流程图如图 1 所示.

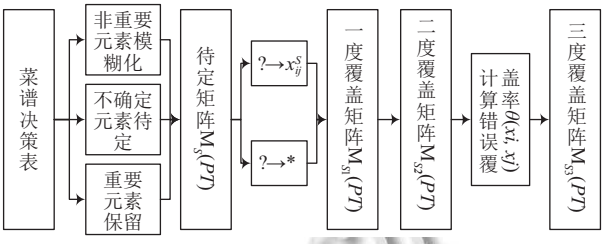

图 1 值约简流程图

首先初步对菜谱决策表进行重定义, 对表中的一条菜谱记录 x_i, 去掉其中的一个属性值 a_i^S, 校验剩下的属性集合 $T_{a_i}^S$ 能否决定决策属性: 如果 $T_{a_i}^S$ 可以独自推出决策属性, 则 x_i 与决策表中某条记录 x_j 构成可覆盖关系 $Tol\left(x_i,x_j\right)$, 即去掉 a_i^S 不影响分类, 那么 a_i^S 没有价值, 置为符号"*"; 如果 $T_{a_i}^S \neq T_{a_j}^S$, 单条记录不能说明 a_i^S 的重要性, a_i^S 是待评价的值, 设置为活跃值, 用符号"?"代替; 如果 $T_{a_i}^S = T_{a_j}^S$, 但决策属性值不同, 即产生冲突, 说明去掉 a_i^S 就不能得到正确的规则, 那么需要保留重要属性值 a_i^S. 此过程后可以得到待定矩阵 $\mathrm{M}(PT)$.

对待定矩阵 $\mathrm{M}(PT)$ 中待定的活跃值 b_{ij} 再进行判定, 即可得到一度覆盖矩阵 $\mathrm{M}_{S1}(PT)$: 如果一条记录中, 除了活跃值, 其他属性值皆被标记为"*", 说明在本条记录中活跃值重要, 则将其还原为原决策表的属性值; 如果记录中除去活跃值和"*"值属性, 剩下的属性组合

可以推出决策条件, 则判定此活跃值 b_{ij} 不重要, 标记为“*”, 如果产生矛盾, 则说明属性值重要, 还原决策表属性值.

对一度覆盖矩阵去重处理, 并记录每条模糊规则的覆盖范围 ω, 构成二度覆盖矩阵.

二度覆盖矩阵中存在矛盾项 (即条件属性相同但决策属性不同), 定义错误覆盖率描述矛盾的两条规则的覆盖范围的差异程度, 认为低于所设 θ 的两条矛盾规则的较小项为误差, 进行舍弃. 错误覆盖率的设置: 选取在决策规则的覆盖范围较高且约简力度较大时的取值, 即可以得到覆盖样本最多且矛盾项最少的规则, 实现提取规则质量最优, 得到三度覆盖矩阵.

得到三度覆盖矩阵后, 仍存在错误覆盖率区分不开的少数冲突规则, 使用支持度、置信度进行规则约简, 置信度表示此条件属性的样例中满足此决策条件的比率, 支持度表示满足此规则的样例数占全部样本数的百分比, 比较冲突规则的支持度、置信度, 删减可信度较低的规则, 消除噪声样例造成的影响.

2.3 概率覆盖模型的算法描述

概率覆盖模型融合了决策粗糙集属性约简和值约简算法, 概率是指 α 正域约简中的 α, 覆盖是指控制值约简中误差范围的错误覆盖率 θ, 对原始矩阵不断进行精简覆盖.

2.3.1 属性约简

输入: 菜谱决策表 S, 给定的概率阈值 α
输出: 营养属性约简集 R

(1) 令初始约简属性集$R=\emptyset$;
(2) 根据式子 (3) 计算每个属性的 α-正域重要度γ_a^α;
(3) 将属性按照重要度由大到小排列, 令其为 P;
(4) 在$|\mathrm{POS}_R^\alpha(D)|<|\mathrm{POS}_C^\alpha(D)|$时循环:
(5) 令 α 为 P 中重要度最高的属性, 将 α 放入约简属性集 R 中, 即 $R=R\cup\{a\}$;
(6) 当 R 不满足定义 2 时, 对$\forall a\in R$, $B=R$, 若$|\mathrm{POS}_{B-\{a\}}^\alpha(D)|\geq|\mathrm{POS}_B^\alpha(D)|$时, $R=R-\{a\}$;
(7) 输出约简集 R.

2.3.2 规则提取

输入: 属性约简后的菜谱决策表$S=(U,C\cup D,f)$, θ
输出: 三度覆盖矩阵$\mathrm{M}_{S3}(PT)$

(1) 对$\forall x_i,x_j\in S(i\neq j)$;
(2) 选定 x_i 中一条属性$c_k\in C_k(k=1,2,3,\cdots,m)$, 计算缺省值$T_{a_i}^S$, 重复
① 判定$T_{a_i}^S$和其他记录 x_i 的缺省值$T_{a_j}^S$关系;
② 根据式 (5)、(6) 对决策表当前条件属性值重定义, 得待定矩阵 $\mathrm{M}(PT)$;
(3) 对$\forall x_i\in\mathrm{M}(PT)$
根据式 (7), 对$\mathrm{M}(PT)$的活跃值 b_{ij} 重定义;
(4) 得到一度覆盖矩阵$\mathrm{M}_{S1}(PT)$;
(5) 对$\forall x_i,x_j\in\mathrm{M}_{S1}(PT)$
if x_i==*
删掉 x_i 所在记录
else if $x_i==x_j$
记录$\omega_{x_i}=\omega_{x_i}+\omega_{x_j}$, 并删掉 x_j;
(6) 得到二度覆盖矩阵$\mathrm{M}_{S2}(PT)$;
(7) 对$\forall x_i,x_j\in\mathrm{M}_{S2}(PT)$
if $c_i=c_j\&\&d_i\neq d_j\&\&\theta(x_i,x_j)>\theta$
删减覆盖范围较小的菜谱项;
(8) 剩下的菜谱规则构成三度覆盖矩阵$\mathrm{M}_{S3}(PT)$;
(9) 输出三度覆盖矩阵$\mathrm{M}_{S3}(PT)$.

3 实验应用

中医营养学目前还没有专门的算法分析, 而中医医药学有很多研究方法[12]: 专家系统、关联规则、神经网络、遗传算法、粗糙集、决策树, 如文献[13]等将关联规则应用于中医肝病处方用药分析, 文献[2]提出基于加权变精度容差粗糙集模型对中医处方的研究. 西医营养学现存一些探究方法, 文献[11,14,15]提出了 Apriori 算法、BP 神经网络对营养元素的数据挖掘.

将本文的算法和模型应用到中医营养学菜谱分析, 以食物“四气五味”为切入口, 应用概率覆盖模型, 挖掘出对养肠胃好的决策规则, 从中医营养学的角度对养肠胃的人群提供指导性建议. 具体应用流程设计如图 2 所示.

3.1 条件属性和决策属性的选取

食物性能与药物性能一致, 包括性味归经、升浮沉降、补泻等[16], 而中医营养学的研究侧重食物“四气”对人体的调和作用, 所谓“四气”, 即寒、热、温、凉四种性质, 另有不寒不热、不温不凉的饮食, 属于平性. 食物的温热寒凉平属性记载来源于《本草图经》、《植物名实图考》、《中国药植志》、《本草纲目》、《古今医鉴》等经典古书籍对食物类别的区分记载, 整理如表 1 所示, 另收集《钱家鸣教你养肠胃就该这样吃》[17]中菜谱食材及其重量, 对一条菜谱有原料$y_i(i=1,2,3,\cdots)$, 其原料对应的重量 m_i, 所选每一种主要材料的含量均大于 10 g, 即以 10 g 为最小计量单位, 根据公式: $\sum_1^i \frac{m_j}{10}$计算菜谱的温热寒凉平五个属性值, 如表 2 部分所示. 对于预防一种类型的疾病来说, 这几种属性具有某种程度的紧密联系的特性, 在数据

分析上具有可挖掘的意义.

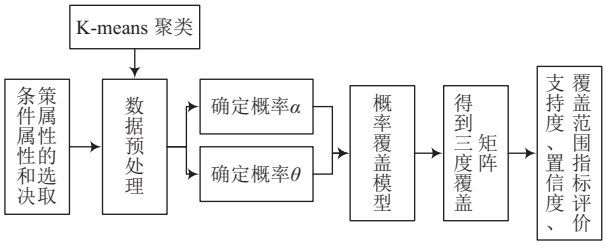

图 2 养肠胃菜谱中医营养分析流程图

表 1 部分原料属性表

原料编号	原料名称	温性	热性	寒性	凉性	平性
1	南瓜	1	0	0	0	0
2	小米	0	0	0	1	0
4	大米	0	0	0	0	1
5	红枣	1	0	0	0	0
6	鲜花生	0	0	0	0	1
…	…	…	…	…	…	…

表 2 部分原始菜谱信息表

菜谱编号	菜谱名称	温性	热性	寒性	凉性	平性	决策
1	南瓜小米粥	30	0	0	0	10	1
2	红枣花生大米粥	60	0	0	0	12	1
3	凉拌芹菜	0	0	0	20	0	1
4	红烧肉	0	0	30	0	83	0
5	焦香辣椒	0	32.5	0	0	0	0
6	扣肉	0	0	25	10	51.1	0
…	…	…	…	…	…	…	…

3.2 数据预处理

本文研究目的是识别四气五味的菜谱对肠胃疾病的功能作用, 把属性值极其相似的归并成一类, 在一定程度上保留离散类的分布特征, 探究其属性等级对肠胃疾病的影响力, 因此使用 K-means 聚类算法[18]对决策信息表进行数据离散化, K-means 聚类含义表如表 3.

表 3 聚类含义表

符号: 含义	温性	热性	寒性	凉性	平性
1: 低	0–15	0–10	0–10	0–10	0–10
2: 中	15–40	10–30	10–25	10–40	10–30
3: 高	40–100	30–100	25–100	40–100	30–100

经过 K-means 离散化处理后的菜谱决策表如表 4 所示, $x_1, x_2, \cdots, x_n$表示每条菜谱标号, 本文选取菜谱原料的寒、热、温、凉、平五种中医属性为研究对象, 称为条件属性集合 $C=\{a, b, c, d, e\}$, 菜谱是否具有养肠胃功能作为决策属性 $D=\{1, 0\}$.

表 4 菜谱决策表

U	a	b	c	d	e	DD
x_1	2	1	1	1	2	1
x_2	3	1	2	1	2	1
x_3	1	1	1	2	1	1
x_4	1	1	3	1	2	0.
x_5	1	3	1	1	1	0
x_6	1	1	3	2	3	0
…	…	…	…	…	…	…

3.3 参数选取

3.3.1 确定概率取值

α=1 根据定义 2, 其中正域为以一定的概率 (大于阈值) 正确分类的属性对象集, 本节探求在何值时, 属性对决策的分类正确率最好. α 取值[0, 1](取值间隔为 0.1), 时, 为经典粗糙集模型, 以公式 (1) 计算每个属性的 α-正域集合, 并由公式 (3) 计算当前阈值下 5 个属性的重要度排序趋势如图 3 所示.

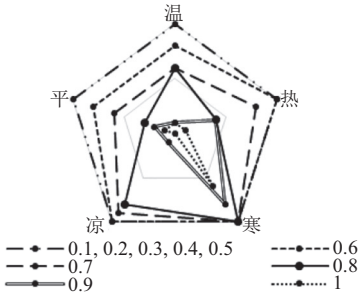

图 3 α 阈值分布图

随着 α 的减小, 各属性重要度增加直至顶峰, 虽然分类精度允许一定程度的误差, 可以提高含误差的属性对决策的正确率, 但不能一味降低 α 值, 否则, 会使得各个属性都重要度全部提高, 失去属性约简的意义, 因此 α 为 0.1~0.6 时失去分类意义. 当设置 α 取值接近 1 时, 分类精度几乎不允许误差, 导致经典粗糙集出现的缺乏容错能力缺陷, 从而舍弃 α 为 0.9、1 的取值.

经专家经验: e 属性代表菜谱的平性特征, 不论何种菜谱都具备广泛的平值属性, 因此平性不具备评价菜谱特性的标准, 选用平性重要度最小时的 α 作为概率覆盖模型的概率取值, 因此最优的温热寒凉趋势为 α=0.8, 得重要度排序如表 5 所示.

表 5 属性重要度

属性	a	b	c	d	e
重要度	0.1812	0.0437	0.5687	0.2437	0.0313
重要度排序	3	4	1	2	5

3.3.2 确定错误覆盖率取值

分析属性值在决策表中出现的规律, 寻找对决策属性影响力最大的属性值, 约简冗余属性值, 舍弃冲突规则. 由公式 (8) 可知错误覆盖率描述一对冲突矛盾规则覆盖范围的差距性, 如果一条规则的覆盖范围达不到另一条矛盾规则的半数以上, 差距过小, 不能判定其中的矛盾为误差所致, 因此设置错误覆盖率的取值范围为[0.5, 1], 取值间隔为 0.1, 控制错误覆盖率 θ, 描绘约简力度与错误覆盖率的关系, 如图 4 所示.

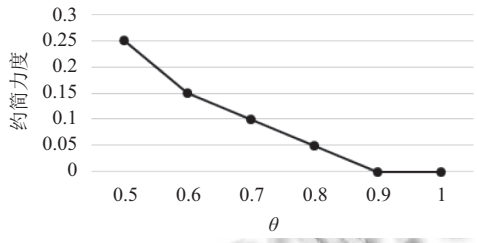

图 4 约简力度与错误覆盖率的关系

为了提取高质量的知识规则, 由图 4 所示设置概率覆盖模型的错误覆盖率为 0.5, 此时所得模型具备最优性能.

3.3.2 建立概率覆盖模型进行规则提取

由 3.3.1 节可确定 0.8-正域约简集合 R={a, b, c, d}, 精简后的菜谱决策表如表 6 所示.

表 6 精简后的菜谱决策表

U	a	b	c	d	D
x_1	2	1	1	1	1
x_2	3	1	2	1	1
x_3	1	1	1	2	1
x_4	1	1	3	1	0
x_5	1	3	1	1	0
x_6	1	1	3	2	0
…	…	…	…	…	…

使用属性约简后的决策表进行值约简, 代入概率覆盖模型得二度覆盖矩阵有 20 条规则, 如表 7 所示, 其中含有 7 对矛盾规则: 3 和 11、4 和 13、5 和 12、8 和 16、10 和 19、14 和 17、15 和 18, 矛盾规则占比 35%, 矛盾规则覆盖样例 33%, 矛盾规则现象较为突出, 如果完全舍弃矛盾项, 则所得规则少, 甚至有可能造成知识断层, 因此在分析菜谱营养成分数据时, 需要对矛盾规则进行评估和精简.

设置膳食属性错误覆盖率阈值为 0.5, 按照决策表值约简的算法步骤, 低于错误覆盖率阈值的矛盾规则中的较小项判定为误差, 经公式 (8) 计算, 删减其中 5 对错误覆盖率率低于 0.5 的较小矛盾规则: 10、11、12、15、16, 约简后得 15 条规则的三度覆盖矩阵, 如表 8 所示.

表 7 二度覆盖矩阵决策表

R	温	热	寒	凉	D	ω
1	*	1	1	*	1	70
2	1	1	2	1	1	8
3	2	2	1	1	1	7
4	1	1	3	2	0	2
5	1	1	2	2	1	7
6	*	*	2	3	1	1
7	3	2	*	*	1	3
8	1	1	3	1	0	14
9	*	3	*	*	0	28
10	2	1	2	1	0	1
11	2	2	1	1	0	3
12	1	1	2	2	0	1
13	1	1	3	2	1	1
14	1	1	3	3	0	2
15	3	1	1	1	0	1
16	1	1	3	1	1	6
17	1	1	3	3	1	1
18	3	1	1	1	1	4
19	2	1	2	1	1	3
20	3	*	2	*	1	1

表 8 三度覆盖矩阵决策表

R	温	热	寒	凉	D	ω	Ψ_s	Ψ_c
1	*	1	1	*	1	70	0.4375	0.6422
2	1	1	2	1	1	8	0.0500	0.0734
3	2	2	1	1	1	7	0.0438	0.0642
4	1	1	3	2	0	2	0.0125	0.0183
5	1	1	2	2	1	7	0.0438	0.0642
6	*	*	2	3	1	1	0.0063	0.0092
7	3	2	*	*	1	3	0.0188	0.0275
8	1	1	3	1	0	14	0.0875	0.2745
9	*	3	*	*	0	28	0.1750	0.5490
13	1	1	3	2	1	1	0.0063	0.0196
14	1	1	3	3	0	2	0.0125	0.0392
17	1	1	3	3	1	1	0.0063	0.0092
18	3	1	1	1	1	4	0.0250	0.0367
19	2	1	2	1	1	3	0.0188	0.0275
20	3	*	2	*	1	1	0.0063	0.0092

三度覆盖矩阵决策表是相对比较精准的决策表, 由表 8 中分析可知, 仍存在两对错误覆盖率难以取舍的矛盾规则 4 和 13、14 和 17, 比较规则 4 和 13 支持度和置信度, 0.0125>0.0063, 0.0183<0.0196, 规则 4 的支持度远高于规则 13, 二者的置信度相差不大, 因此舍弃规则 13. 规则 14 的置信度和支持度均大于规则 17, 删减规则 17.

经过概率覆盖模型及评价指标分析后, 终得到

13 条决策规则. 规则 1、2、3、8、9 相对其他规则而言具有较高的支持度, 在样本数据中也具有较高的覆盖范围, 具有较高的可信性. 规则 1、2、3, 7、18、19、20 显示寒凉性低、温性高且微热的菜谱有利于养肠胃, 可以暖胃健脾, 促进血液循环, 益气补血, 安神抗寒, 如羊肉萝卜粥、生姜羊肉粥、土豆炖牛肉等, 都属于温补的膳食[19]. 由规则 4、8、14 所得, 寒性高的食物会刺激肠胃, 胃肠膜黏硬, 造成肠胃负担, 不利于肠胃吸收. 规则 9 显示过热的食物损害肠胃, 回归菜谱样例分析, 豆沙炸糕、炸大扁丸子等菜谱不利于肠胃消化吸收. 所得规则揭示了中医寒热调和的原理, 所以, 养肠胃人群要远离大寒过热的食物, 食用菜谱时也要选择少寒多温稍热的菜谱进行调节[20].

3.4 实验总结

对得到规则质量的评价指标有: 矛盾规则对数, 提取规则条目, 非矛盾规则的覆盖率. 对比本文提出的概率覆盖模型、传统的基于粗糙集的值约简算法和文献[2]针对中医方剂提出的基于加权变精度容差粗糙集模型. 本文构建的概率覆盖模型提取得到较为精简的规则数, 规则中具备较少的矛盾冲突, 非矛盾规则的覆盖率较高, 如图 5 所示.

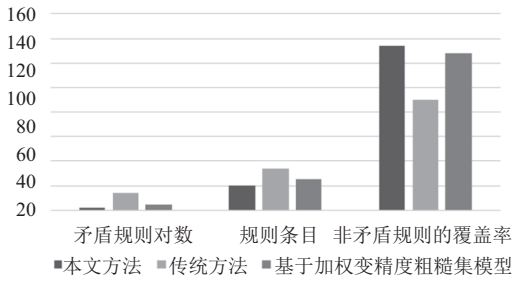

图 5 实验对比图

另抽取 50 条除标签菜谱样例为测试数据, 以文中相同的处理方法进行数据处理, 应用本文所得概率覆盖模型, 给菜谱样例标记决策属性, 对比原标签, 有 41 条标记正确, 正确率达 82%, 可以给养肠胃人群饮食给予指导.

4 结束语

本文基于决策粗糙集 α-正域约简算法, 利用中医经验确定属性重要度的 α 概率, 构建待定矩阵, 一度、二度、三度覆盖矩阵提取决策规则, 引入错误覆盖率对值约简规则提取进行了改进, 所得规则与中医医学知识相吻合. 实验结果表明, 该模型提取的矛盾规则数量相对较少, 得出的中医营养规则具有可信度, 对菜谱是否养肠胃具有辨识度, 可用于中医营养学饮食指导. 本文模型还可以应用到政府信访系统、师生教学评价、植物生长监测等领域, 根据不同应用背景控制模型参数, 研究矛盾规则的取舍问题.

参考文献

1 申杰, 杨联河, 唐华伟. 对中西医结合营养学的思忖. 第七届全国中西医结合营养学术会议论文资料汇编. 北京, 中国. 2016. 24–26.
2 佘侃侃, 胡孔法, 王珍. 基于加权变精度容差粗糙集模型的属性约简及应用研究. 计算机科学, 2014, 41(S2): 351–353.
3 常犁云, 王国胤, 吴渝. 一种基于 Rough Set 理论的属性约简及规则提取方法. 软件学报, 1999, 10(11): 1206–1211.
4 杨振峰, 郭景峰, 常峰. 一种基于粗集的值约简方法. 计算机工程, 2003, 29(9): 96–97.
5 焦娜. 基于差异关系的变精度粗糙集知识约简算法研究. 计算机科学, 2015, 42(5): 265–269. [doi: 10.11896/j.issn.1002-137X.2015.05.053]
6 韦碧鹏, 吕跃进, 李金海, 等. 不完备不协调序决策系统的属性约简与规则提取. 计算机科学, 2013, 40(S2): 160–164.
7 索明亮. 基于粗糙集的故障预测及诊断技术在卫星中的应用[硕士学位论文]. 哈尔滨: 哈尔滨工业大学, 2013.
8 Yao YY, Wong SKM, Lingras P. A decision-theoretic rough set model. Proceedings of the 5th International Symposium on Methodologies for Intelligent Systems. North-Holland, NY, USA. 1990. 17–24.
9 李华雄, 周献中, 李天瑞, 等. 决策粗糙集理论及其研究进展. 北京: 科学出版社, 2011: 1–91.
10 Yao YY. Decision-theoretic rough set models. International Conference on Rough Sets and Knowledge Technology. Toronto, Canada. 2007. 1–12.
11 刘盾, 姚一豫, 李天瑞. 三枝决策粗糙集. 计算机科学, 2011, 38(1): 246–250.
12 张璐, 雷雪梅. 基于粒子群优化 BP 神经网络的养肠胃菜谱判定. 计算机科学, 2016, 43(S2): 63–66, 72.
13 罗悦, 温川飙, 严小英. 基于专家系统的中医辨证论治信息表示方法研究. 中国数字医学, 2016, 11(7): 37–40.
14 寇文心. 智能营养配餐系统及其核心算法的研究[硕士学位论文]. 北京: 北京工业大学, 2015.
15 张云渡. 数据挖掘技术在营养配餐系统中的应用研究[硕士学位论文]. 北京: 北京工业大学, 2014.
16 宋京美, 吴嘉瑞, 姜迪, 等. 基于数据挖掘的中医治疗肝病处方用药规律分析. 中国实验方剂学杂志, 2015, 21(22): 218–221.
17 钱家鸣. 钱家鸣教你养肠胃就该这样吃. 北京: 中国轻工业出版社, 2014.
18 张友海, 李锋刚. Kmeans 算法的 Spark 实现及优化. 西安文理学院学报 (自然科学版), 2017, 20(3): 18–20, 32.
19 翁维健. 中医饮食营养学. 上海: 上海人民出版社, 2008. 1–207.
20 Chapmannovakofski K. Summer is different: What that means for nutrition educators. Journal of Nutrition Education and Behavior, 2016, 48(7): 436. [doi: 10.1016/j.jneb.2016.05.010]