三枝决策粗糙集

刘 盾[1] 姚一豫[2] 李天瑞[3]
(西南交通大学经济管理学院 成都 610031)[1]
(Department of Computer Science, University of Regina, Regina, Saskatchewan, S4S 0A2)[2]
(西南交通大学信息科学与技术学院 成都 610031)[3]

摘 要 从贝叶斯理论出发,介绍基于三枝决策粗集理论。首先讨论在期望风险最小决策的语义下决策粗集理论基本模型的构建过程。其次,分析决策粗集三枝决策方法在不同概率区间犯错的可能性,并通过其与二枝决策及Pawlak粗集三枝决策的差异,给出决策粗集三枝决策方法优于其他两种决策方法的成立条件。最后,提供一种利用决策粗集三枝决策解决实际问题的方法。

关键词 决策粗集理论,贝叶斯过程,三枝决策,二枝决策
中图法分类号 TP18 **文献标识码** A

Three-way Decision-theoretic Rough Sets

LIU Dun[1] YAO Yi-yu[2] LI Tian-rui[3]
(Department of Economics and Management, Southwest Jiaotong University, Chengdu 610031, China)[1]
(Department of Computer Science, University of Regina, Regina, Saskatchewan S4S 0A2, Canada)[2]
(Department of Information Science and Technology, Southwest Jiaotong University, Chengdu 610031, China)[3]

Abstract A model of three-way decision-theoretic rough sets (DTRS) was presented based on the Bayesian decision theory. Based on the minimum expected risk, a detailed formulation of DTRS was given. Different types of errors in several probability intervals were examined. The conditions under which DTRS three-way method is superior to the Pawlak three-way method and two-way method were identified. DTRS three-way model was discussed for solving the practical decision problems.

Keywords Decision-theoretic rough set theory, Bayesian decision procedure, Three-way decision making, Two-way decision making

1 引言

粗糙集理论是一种处理不确定性和不精确性问题的新型数学工具。它自1982年由Pawlak提出以来[1],无论是在理论还是在应用上都取得很多重要成果。经典的Pawlak粗集利用等价关系将论域划分为若干等价类,而这些等价类将整个论域分为3个部分:完全属于某个集合的所有等价类构成正域、可能但不完全属于某个集合的所有等价类构成边界域、完全不属于某个集合的所有等价类构成负域。基于3个区域,Yao等人在文献[2-4]中提出了三枝决策规则,探讨了粗糙集的一个新的语义。从正域里获取的正规则用来接受某事物(acceptance);从负域里获取的负规则用来表示拒绝某事物(rejection);落在边界域上的规则表示需要进一步观察,即延迟决策(deferment)。这种将论域分为3部分的决策方式,很好地描述了人类在解决实际决策问题时的思维模式[2-4]。

Pawlak粗集并没有考虑到决策规则的容错性,完全正确和确定的规则才能进入正域。基于此,作为Pawlak粗集的更一般形式,一系列概率粗集模型如0.5-概率粗集模型[5]、决策粗集模型[6,7]、变精度粗集模型[8]、贝叶斯粗集模型[9-10]等相继提出。在概率粗集模型中,具有较高正确可能性的等价类会进入正域,而不满足较低划分阈值的等价类将会进入负域,介于两者之间的等价类则在边界域中。这使得论域被相应地分为具有某种容错能力的3个区域,形成具有容错性的概率三枝决策过程。考虑到不同的概率阈值会导致不同的决策结果,因而如何选择合理的概率成为解决问题的关键。此外,Yao在文献[15]中,从微观和宏观两个层面探讨了三枝决策的优越性,通过比较决策粗集三枝决策与二枝决策及Pawlak粗集三枝决策的差异,给出了它们各自成立的条件,为人们研究三枝决策粗集模型提供了坚实的理论支撑。

基于上述结果,本文分别从3个方面来介绍三枝决策粗集模型。首先,简要介绍概率粗集模型和决策粗集模型;其次,详细阐述三枝决策思想在实际决策过程中的优越性;最后,给出一种在实际问题中利用决策粗集模型三枝决策解决实际问题的方法。本文的主要工作着重于对已有文献进行总

到稿日期:2010-03-09 返修日期:2010-05-07 本文受国家自然科学基金(60873108,70971062),西南交通大学博士创新基金(200907),西南交通大学优秀博士论文培育基金(2009LD)资助。

刘 盾(1983-),男,博士生,主要研究方向为粗集决策、数据挖掘等,E-mail:newton83@163.com;姚一豫(1964-),男,教授,主要研究方向为粗集理论、网络智能、粒计算等;李天瑞(1969-),男,教授,博士生导师,主要研究方向为智能信息处理、数据挖掘等。

结，并给出了笔者对决策粗集理论研究现状的理解，回答了“为什么用”和“怎样用”决策粗集三枝决策方法来解决现实问题。

2 粗集理论的基本概念

在本节中，主要介绍 Pawlak 粗集、相关概率粗集的基本概念和定义[1-7,11,12]。

定义 1 假设 U 是一个有限的非空集，R 是定义在 U 上的一个等价关系，记 $apr=(U,R)$ 为近似空间。U 在等价关系 R 下的划分记为 $U/R=\{[x]_R \mid x\in U\}$，$[x]$ 是包含 x 的等价类。对 $X\subseteq U$，其上下近似可定义为

$$\underline{apr}(X)=\{x\in U\mid [x]\subseteq X\}$$
$$\overline{apr}(X)=\{x\in U\mid [x]\cap X\neq\varnothing\}$$

上下近似将论域分为 3 个部分：正域 $POS(X)$、边界域 $BND(X)$ 和负域 $NEG(X)$，其定义分别为

$$\begin{aligned} POS(X)&=\underline{apr}(X)\\ BND(X)&=\overline{apr}(X)-\underline{apr}(X)\\ NEG(X)&=U-\overline{apr}(X)\end{aligned}\tag{1}$$

由正域中元素导出的规则表示确定属于 X 的规则，由负域中元素导出的规则表示确定不属于 X 的规则，而由边界域导出的规则表示可能属于 X 的规则，这体现了“三枝决策”的基本思想。但是 Pawlak 粗集并没有考虑到规则的容错性，这就需要引进条件概率、概率粗集等相关概念[11-13]。

定义 2 假设 $S=(U,A,V,f)$ 是一个信息表，$\forall x\subseteq U, X\subseteq U$，令

$$\Pr(X\mid [x])=|[x]\cap X|/|[x]|\tag{2}$$

式中，$|\cdot|$ 表示集合中元素的基数，$\Pr(X\mid [x])$ 表示分类的条件概率。

定义 3 假设 $S=(U,A,V,f)$ 是一个信息表，$X\subseteq U$，$0\leqslant\beta<\alpha\leqslant 1$，则 (α,β)-下近似，(α,β)-上近似可定义为

$$\begin{aligned}\underline{apr}_{(\alpha,\beta)}(X)&=\{x\in U\mid \Pr(X\mid [x])\geqslant\alpha\}\\ \overline{apr}_{(\alpha,\beta)}(X)&=\{x\in U\mid \Pr(X\mid [x])>\beta\}\end{aligned}\tag{3}$$

同样地，(α,β)-上下近似将论域分为 3 个部分：(α,β)-正域 $POS_{(\alpha,\beta)}(X)$、(α,β)-边界域 $BND_{(\alpha,\beta)}(X)$ 和 (α,β)-负域 $NEG_{(\alpha,\beta)}(X)$，其定义分别为

$$\begin{aligned} POS_{(\alpha,\beta)}(X)&=\{x\in U\mid \Pr(X\mid [x])\geqslant\alpha\}\\ BND_{(\alpha,\beta)}(X)&=\{x\in U\mid \beta<\Pr(X\mid [x])<\alpha\}\\ NEG_{(\alpha,\beta)}(X)&=\{x\in U\mid \Pr(X\mid [x])\leqslant\beta\}\end{aligned}\tag{4}$$

当 $\alpha=1,\beta=0$ 时，上述模型转化成 Pawlak 粗集模型；当 $\alpha=\beta=0.5$ 时，上述模型转化成 0.5-概率粗集模型。然而，Pawlak 粗集模型和 0.5-概率粗集模型只是两种极值条件下的三枝决策模型。如何选取合理的 α 和 β 值，对于实际决策问题意义重大。基于此，Yao 和 Wong 将贝叶斯决策过程引入概率粗集模型，提出了决策粗集理论[9]。在决策粗集理论中，α 和 β 值能在总体风险最小的条件下直接计算得出，这为概率粗集模型给出了一种语义上的解释。

3 决策粗集理论的基本模型

决策粗集理论的基本模型可参考文献[2-4,6,7,11,12,14]。其基本思想为：假设 $S=(U,A,V,f)$ 是一个信息表，$\Omega=\{w_1,w_2,\cdots,w_m\}$ 为 m 个有限的状态集，$A=\{a_1,a_2,\cdots,a_n\}$ 为 n 个有限的行动集。$\Pr(w_i\mid x)$ 表示 x 在状态 w_i 下的条件概率。令 $\lambda(a_j\mid w_i)$ 为在状态 w_i 下采取行动 a_j 的损失或代价。对于某元素 x，如果采取行动 a_j，其期望损失为

$$R(a_j\mid x)=\sum_{i=1}^{m}(a_j\mid w_i)\Pr(w_i\mid x)\tag{5}$$

一般地，对于给定的描述 x，令 $\tau(x)$ 为一个决策规则，它是描述空间到 A 的一个函数，$\tau(x)\in A$。令 $\mathcal{R}$ 是在给定一个决策规则 τ 下的总体期望风险，它可表示为

$$\mathcal{R}=\sum_{x\in U}R(\tau(x)\mid x)\Pr(x)\tag{6}$$

式中，$\Pr(x)$ 为 x 的先验概率，$R(\tau(x)\mid x)$ 则为采取不同行动 x 的条件风险。按照贝叶斯决策过程，需要选取使得总体期望风险 $\mathcal{R}$ 达到最小的决策行为作为最佳行动方案。如果有多个决策使得 $\mathcal{R}$ 达到最小，则根据实际情况选择其中之一。

决策粗集模型是基于贝叶斯决策过程的。基于三枝决策的思想，决策粗集模型利用 2 个状态集和 3 个行动集来描述决策过程。状态集 $\Omega=\{X,\neg X\}$ 分别表示某事件属于 X 和不属于 X，行动集 $A=\{a_P,a_B,a_N\}$ 分别表示接受某事件、延迟决策和拒绝某事件 3 种行动。考虑到采取不同行动会产生不同的损失，记 $\lambda_{PP},\lambda_{BP},\lambda_{NP}$ 分别表示当 x 属于 X 时采取行动 a_P,a_B 和 a_N 下的损失；同样地，记 $\lambda_{PN},\lambda_{BN},\lambda_{NN}$ 分别表示当 x 不属于 X 时采取行动 a_P,a_B 和 a_N 下的损失。因此，根据式(1)，采取 a_P,a_B 和 a_N 3 种行动下的期望损失可分别表示为

$$\begin{aligned} R(a_P\mid [x])&=\lambda_{PP}\Pr(X\mid [x])+\lambda_{PN}\Pr(\neg X\mid [x])\\ R(a_B\mid [x])&=\lambda_{BP}\Pr(X\mid [x])+\lambda_{BN}\Pr(\neg X\mid [x])\\ R(a_N\mid [x])&=\lambda_{PN}\Pr(X\mid [x])+\lambda_{NN}\Pr(\neg X\mid [x])\end{aligned}\tag{7}$$

根据贝叶斯决策准则，需要选择期望损失最小的行动集作为最佳行动方案，于是可得到如下 3 条决策规则：

(P)：若 $R(a_P\mid [x])\leqslant R(a_B\mid [x])$ 和 $R(a_P\mid [x])\leqslant R(a_N\mid [x])$ 同时成立，则 $x\in POS(X)$；

(B)：若 $R(a_B\mid [x])\leqslant R(a_P\mid [x])$ 和 $R(a_B\mid [x])\leqslant R(a_N\mid [x])$ 同时成立，则 $x\in BND(X)$；

(N)：若 $R(a_N\mid [x])\leqslant R(a_P\mid [x])$ 和 $R(a_N\mid [x])\leqslant R(a_B\mid [x])$ 同时成立，则 $x\in NEG(X)$。

由 $\Pr(X\mid [x])+\Pr(\neg X\mid [x])=1$，上述规则只与概率 $\Pr(X\mid [x])$ 和相关的损失函数 λ 有关。此外，考虑到接受正确事物的损失不大于延迟接受正确事物的损失，且这两者都小于拒绝正确事物的损失；同样地，拒绝错误事物的损失不大于延迟拒绝错误事物的损失，且这两者都小于接受错误事物的损失，则一个合理的假设为 $0\leqslant\lambda_{PP}\leqslant\lambda_{BP}\leqslant\lambda_{NP}$，$0\leqslant\lambda_{NN}\leqslant\lambda_{BP}\leqslant\lambda_{PN}$。据此，可计算 3 条决策规则$(P)$，$(B)$，$(N)$的条件为：

对于规则(P)而言，

$$\begin{aligned}&R(a_P\mid [x])\leqslant R(a_B\mid [x])\\ \Leftrightarrow&\Pr(X\mid [x])\geqslant\frac{(\lambda_{PN}-\lambda_{BN})}{(\lambda_{PN}-\lambda_{BN})+(\lambda_{BP}-\lambda_{PP})}\\ &R(a_P\mid [x])\leqslant R(a_N\mid [x])\\ \Leftrightarrow&\Pr(X\mid [x])\geqslant\frac{(\lambda_{PN}-\lambda_{NN})}{(\lambda_{PN}-\lambda_{NN})+(\lambda_{NP}-\lambda_{PP})}\end{aligned}$$

对于规则(B)而言，

$$\begin{aligned}&R(a_B\mid [x])\leqslant R(a_P\mid [x])\\ \Leftrightarrow&\Pr(X\mid [x])<\frac{(\lambda_{PN}-\lambda_{BN})}{(\lambda_{PN}-\lambda_{BN})+(\lambda_{BP}-\lambda_{PP})}\\ &R(a_B\mid [x])\leqslant R(a_N\mid [x])\\ \Leftrightarrow&\Pr(X\mid [x])\geqslant\frac{(\lambda_{BN}-\lambda_{NN})}{(\lambda_{BN}-\lambda_{NN})+(\lambda_{NP}-\lambda_{BP})}\end{aligned}$$

对于规则(N)而言，

$$R(a_N\mid [x])\leqslant R(a_P\mid [x])$$

$$\Leftrightarrow \Pr(X|[x])<\frac{(\lambda_{PN}-\lambda_{NN})}{(\lambda_{PN}-\lambda_{NN})+(\lambda_{NP}-\lambda_{PP})}$$

$$R(a_N|[x])\leqslant R(a_B|[x])$$

$$\Leftrightarrow \Pr(X|[x])<\frac{(\lambda_{BN}-\lambda_{NN})}{(\lambda_{BN}-\lambda_{NN})+(\lambda_{NP}-\lambda_{BP})}$$

令

$$\alpha=\frac{(\lambda_{PN}-\lambda_{BN})}{(\lambda_{PN}-\lambda_{BN})+(\lambda_{BP}-\lambda_{PP})}=\left(1+\frac{\lambda_{BP}-\lambda_{PP}}{\lambda_{PN}-\lambda_{BN}}\right)^{-1}$$

$$\beta=\frac{(\lambda_{BN}-\lambda_{NN})}{(\lambda_{BN}-\lambda_{NN})+(\lambda_{NP}-\lambda_{PP})}=\left(1+\frac{\lambda_{NP}-\lambda_{BP}}{\lambda_{BN}-\lambda_{NN}}\right)^{-1} \quad (8)$$

$$\gamma=\frac{(\lambda_{PN}-\lambda_{NN})}{(\lambda_{PN}-\lambda_{NN})+(\lambda_{NP}-\lambda_{PP})}=\left(1+\frac{\lambda_{NP}-\lambda_{PP}}{\lambda_{PN}-\lambda_{NN}}\right)^{-1}$$

由规则(B)可知，$\alpha>\beta$，则$\frac{\lambda_{BP}-\lambda_{PP}}{\lambda_{PN}-\lambda_{BN}}<\frac{\lambda_{NP}-\lambda_{BP}}{\lambda_{BN}-\lambda_{NN}}$。此外，由不等式$\frac{b}{a}>\frac{d}{c}\Rightarrow\frac{b}{a}>\frac{b+d}{a+c}>\frac{b}{c}(a,b,c,d>0)$，有$\frac{\lambda_{BP}-\lambda_{PP}}{\lambda_{PN}-\lambda_{BN}}<\frac{\lambda_{NP}-\lambda_{PP}}{\lambda_{PN}-\lambda_{NN}}<\frac{\lambda_{NP}-\lambda_{BP}}{\lambda_{BN}-\lambda_{NN}}$。因而，有$0\leqslant\beta<\gamma<\alpha\leqslant 1$。这样，规则$(P)$，$(B)$，$(N)$可重写为：

$(P1)$：如果，$\Pr(X|[x])\geqslant\alpha$，则$x\in POS(X)$；

$(B1)$：如果$\beta<\Pr(X|[x])<\alpha$，则$x\in BND(X)$；

$(N1)$：如果$\Pr(X|[x])<\beta$，则$x\in NEG(X)$。

由此可见，决策粗集模型不仅给予了概率粗集模型一种基于贝叶斯最小风险下的语义解释，而且两个参数α和β可以直接计算得出。在变精度粗集模型中[7]，两个参数完全由专家给定，没有语义解释。因此，很多研究工作着重于用实验方法来获取其参数值。从数学上讲，变精度粗集模型用到了决策粗集模型的一个结果（即$\alpha+\beta=1$，$\alpha>0.5$且$\alpha>\beta$），它可以看作是决策粗集模型的一种特殊情形[2-4]。由于决策粗集模型给予了概率粗集模型一种语义解释，因而它更能代表概率粗糙集的思想。在实际应用中，人们可以简单地使用决策粗糙集的结果，而不去考虑所依赖的损失函数，但应该认识到两个阈值与决策的风险密切相关。

4 决策粗集理论的三枝决策思想

本节的讨论主要是基于文献[15]展开的。考虑到决策粗集理论是基于三枝决策的，参数α和β将论域分为了3个区间。为了简便起见，记这种三枝决策模式为(α,β)-三枝决策。相对于Pawlak三枝决策（$\alpha=1,\beta=0$，$(1,0)$-三枝决策）和γ-二枝决策而言，(α,β)-三枝决策具有其独特的优势。下面，从统计学的角度来诠释以上观点。

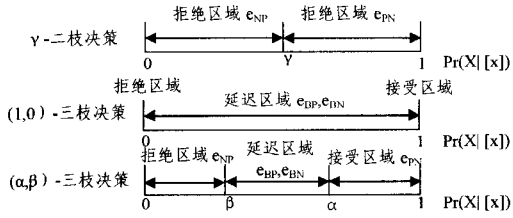

图1 3种决策方式对比图[15]

对于γ-二枝决策，其错误有两类：其一为拒真错误（e_{NP}），将正确的事物判为错误的；其二为采伪错误（e_{PN}），将错误的事物判为正确的。对于$(1,0)$-三枝决策，它不会产生拒真和采伪错误，但其延迟错误也有两类：延迟拒真错误（e_{BP}）和延迟采伪错误（e_{BN}）。对于(α,β)-三枝决策，以上4种情况都可能发生。在接受区域可能造成采伪错误，在拒绝区域可能造成拒真错误，在延迟区域可能造成延迟拒真错误和延迟采伪错误。以上3种类型各自的错误类型可由图1来说明。

通过图1发现，上述参数将整个条件概率分布区间分为了6个区域：0，$(0,\beta]$，(β,γ)，$[\gamma,\alpha)$，$[\alpha,1)$，1。表1形象地给出了上述3种决策方式在6个区域中各自的错误类型。

$\Pr(X\|[x])$	$(1,0)$-三枝决策	(α,β)-三枝决策	γ-二枝决策
0	无错误	无错误	无错误
$[0,\beta]$	延迟错误 e_{BP}，e_{BP}	拒真错误 e_{NP}	拒真错误 e_{NP}
(β,γ)	延迟错误 e_{BP}，e_{BP}	延迟错误 e_{BP}，e_{BP}	拒真错误 e_{NP}
$[\gamma,\alpha]$	延迟错误 e_{BP}，e_{BP}	延迟错误 e_{BP}，e_{BP}	采伪错误 e_{BN}
$(\alpha,1)$	延迟错误 e_{BP}，e_{BP}	采伪错误 e_{BN}	采伪错误 e_{BN}
1	无错误	无错误	无错误

在表1中，虚框部分为(α,β)-三枝决策分别与$(1,0)$-三枝决策、γ-二枝决策错误类型的不同之处。(α,β)-三枝决策分别与$(1,0)$-三枝决策在区间$(0,\beta]$和$[\alpha,1)$不同；(α,β)-三枝决策与二枝决策在区间(β,α)不同。$(1,0)$-三枝决策过于保守，它不允许任何拒真错误和采伪错误发生。条件概率不为1或0的事件一律延迟决策，是一种“厌恶风险”的决策方式；γ-二枝决策过于激进，它不考虑延迟决策，因而犯拒真错误和采伪错误的可能性大大增加，这属于一种“偏好风险”的决策方式。而(α,β)-三枝决策处于$(1,0)$-三枝决策和γ-二枝决策的中间过程，它可看为一种“风险中性”的决策方式，这符合人们的理性思维，体现了一种决策过程中的“中庸之道”。

再者，结合第2节的分析，表2给出了3种决策方式各自成立的条件。在表2中，将条件(C_1)代入式(6)－式(8)可得$\alpha=1,\beta=0,0<\gamma<1$，根据决策准则$(P)$，$(B)$，$(N)$，此时为$(1,0)$-三枝决策；同样地，将条件$(C_1')$和$(C_2')$代入式(6)－式(8)，易得$0\leqslant\beta<\gamma<\alpha\leqslant 1$。根据决策准则$(P)$，$(B)$，$(N)$，参数$\alpha$和$\beta$起作用，参数$\gamma$失效，此时为$(\alpha,\beta)$-三枝决策；将条件$(C_1'')$和$(C_2'')$代入式(6)－式(8)，易得$\gamma>\alpha,\gamma<\beta$，根据决策准则(P)，(B)，(N)，参数$\gamma$起作用，参数$\alpha$和$\beta$失效，此时为$\gamma$-二枝决策。表2给出了在考虑行动损失情况下3种决策方式的成立条件。下面给出(α,β)-三枝决策方式比其他两种决策方式更优的条件。相对于文献[15]中从微观和宏观层面得到的结果，本文从“错误分类”角度得到了同样的结论。

表2 3种决策方式的成立条件

条件	$(1,0)$-三枝决策	(α,β)-三枝决策	γ-二枝决策
	(C_1)	(C_1')	(C_1'')
条件1	$\lambda_{PP}=\lambda_{BP}<\lambda_{NP}$ $\lambda_{NN}=\lambda_{BN}<\lambda_{PN}$	$\lambda_{PP}<\lambda_{BP}<\lambda_{NP}$ $\lambda_{NN}<\lambda_{BN}<\lambda_{PN}$	$\lambda_{PP}\leqslant\lambda_{BP}<\lambda_{NP}$ $\lambda_{NN}\leqslant\lambda_{BN}<\lambda_{PN}$
		(C_2)	(C_2')
条件2	—	$\frac{\lambda_{BP}-\lambda_{PP}}{\lambda_{PN}-\lambda_{BN}}<\frac{\lambda_{NP}-\lambda_{BP}}{\lambda_{BN}-\lambda_{NN}}$	$\frac{\lambda_{BP}-\lambda_{PP}}{\lambda_{PN}-\lambda_{BN}}\geqslant\frac{\lambda_{NP}-\lambda_{BP}}{\lambda_{BN}-\lambda_{NN}}$

定理1 在条件(C_1')和(C_2)同时成立的情况下，(α,β)-三枝决策方式比$(1,0)$-三枝决策方式更优。

证明：假设$a_1=|\{x\in U|\Pr(X|[x])=1\}|$，$a_2=|\{x\in U|x\in X,\alpha\leqslant\Pr(X|[x])<1\}|$，$a_3=|\{x\in U|x\in X,\beta<\Pr(X|[x])<\alpha\}|$，$a_4=|\{x\in U|x\in X,0<\Pr(|[x])\leqslant\beta\}|$；$b_1=|\{x\in U|\Pr(X|[x])=0\}|$，$b_2=|\{x\in U|x\in\neg X,0<\Pr(X|[x])\leqslant\beta\}|$，$b_3=|\{x\in U|x\in\neg X,\beta<\Pr(X|[x])<\alpha\}|$，$b_4=|\{x\in U|x\in\neg X,\alpha\leqslant\Pr(|[x])\leqslant 1\}|$。其中，$|\cdot|$表示集合元素的基数。由上述定义，$a_2+b_4=a_4+b_2=|U|$，则$\frac{a_2}{a_2+b_4}\geqslant\alpha$，$\frac{a_4}{a_4+b_2}\geqslant\beta$。

下面从犯错的损失大小来比较方法的优劣。由于把正确的

事情放入正域，把错误的事情放入负域，都不产生犯错的损失，则 $\lambda_{PP}=\lambda_{NN}=0$。由式(6)和式(7)，$\frac{\alpha}{1-\alpha}=\frac{\lambda_{PN}-\lambda_{BN}}{\lambda_{BP}}$，$\frac{\beta}{1-\beta}=\frac{\lambda_{BN}-\lambda_{NP}}{\lambda_{BP}}$。代入上面两式，有 $\lambda_{BN}\cdot a_2\geqslant(\lambda_{PN}-\lambda_{BN})b_4$，$(\lambda_{NN}-\lambda_{BP})a_4\leqslant\lambda_{BN}\cdot b_2$。

由表1可知，(α,β)-三枝决策和$(1,0)$-三枝决策的不同之处在$(0,\beta]$和$[\alpha,1)$两个区间。对于前者，其在$(0,\beta]$区间可能犯拒真错误 e_{PN}，其损失为 $\lambda_{PN}\cdot a_4$；在$(\alpha,1]$可能犯采伪错误 e_{PN}，其损失为 $\lambda_{PN}\cdot b_4$；则其在两个区间的总损失记为 $Cost^{\alpha,\beta}_{(0,\beta]\cup[\alpha,1)}=\lambda_{NP}\cdot a_4+\lambda_{PN}\cdot b_4$。对于后者，其在$(0,\beta]$区间可能犯延迟错误 e_{BP}，e_{BN}，其损失为 $\lambda_{BP}\cdot a_4+\lambda_{BN}\cdot b_2$；其在$(0,\beta]$区间可能犯延迟错误 e_{BP}，e_{BN}，其损失为 $\lambda_{BP}\cdot a_2+\lambda_{BN}\cdot b_4$；则其在两个区间的总损失记为 $Cost^{(1,0)}_{(0,\beta]\cup[\alpha,1)}=\lambda_{BP}\cdot(a_2+a_4)+\lambda_{BN}\cdot(b_2+b_4)$。两者的总损失相减，得 $Cost^{(1,0)}_{(0,\beta]\cup[\alpha,1)}-Cost^{\alpha,\beta}_{(0,\beta]\cup[\alpha,1)}=[\lambda_{BP}\cdot a_2-(\lambda_{PN}-\lambda_{BN})\cdot b_4)]+[\lambda_{BN}\cdot b_2-(\lambda_{NP}-\lambda_{BP})\cdot a_4)]\geqslant 0$。故$(1,0)$-三枝决策方式总损失比$(\alpha,\beta)$-三枝决策方式更大，后者在$(C_1')$和$(C_2)$下更优。

定理2 在条件(C_1')和(C_2)同时成立的情况下，(α,β)-三枝决策方式比γ-二枝决策方式更优。

证明：设 $c_1=|\{x\in U|x\in X,\alpha\leqslant\Pr(X|[x])\leqslant 1\}|$，$c_2=|\{x\in U|x\in X,\gamma\leqslant\Pr(X|[x])<\alpha\}|$，$c_3=|\{x\in U|x\in X,\beta<\Pr(X|[x])<\gamma\}|$，$c_4=|\{x\in U|x\in X,0\leqslant\beta\}|$；$d_1=|\{x\in U|x\in\neg X,0\leqslant\Pr(X|[x])\leqslant\beta\}|$，$d_2=|\{x\in U|x\in\neg X,\beta<\Pr(X|[x])<\gamma\}|$，$d_3=|\{x\in U|x\in\neg X,\gamma\leqslant\Pr(X|[x])<\alpha\}|$，$d_4=|\{x\in U|x\in\neg X,\alpha\leqslant\Pr([x])\leqslant 1\}|$。由上述定义，$d_2+c_3=d_3+c_2=|U|$，则 $\gamma\leqslant\frac{c_2}{c_2+d_3}<\alpha$，$\beta<\frac{c_3}{d_2+c_3}<\gamma$。同理，从犯错损失的角度，把正确的事情放入正域，把错误的事情放入负域，都不产生犯错的损失，则 $\lambda_{PP}=\lambda_{NN}=0$。将式(6)、式(7)代入，有 $\lambda_{BP}\cdot c_2<(\lambda_{PN}-\lambda_{BN})\cdot d_3$，$(\lambda_{NP}-\lambda_{BP})\cdot c_3>\lambda_{BN}\cdot d_2$。

由表1可知，比较(α,β)-三枝决策和γ-二枝决策的不同之处在(β,γ)和$[\gamma,\alpha)$两个区间。对于前者，其在(α,β)可能犯延迟错误 e_{BP}，e_{BN}，其损失为 $\lambda_{BP}\cdot c_3+\lambda_{BN}\cdot d_2$；其在$[\gamma,\alpha)$可能犯延迟错误 e_{BP}，e_{BN}，其损失为 $\lambda_{BP}\cdot c_2+\lambda_{BN}\cdot d_3$；则其在两个区间的总损失记为 $Cost^{\alpha,\beta}_{(\beta,\gamma)\cup[\gamma,\alpha)}=\lambda_{BP}(c_2+c_3)+\lambda_{BN}\cdot(d_2+d_3)$。对于后者，其在$(\beta,\gamma)$区间可能犯拒真错误 e_{NP}，其损失为 $\lambda_{NP}\cdot c_3$；其在$[\gamma,\alpha)$区间可能犯采伪错误 e_{PN}，其损失为 $\lambda_{PN}\cdot d_3$；则其在两个区间的总损失记为 $Cost^{\gamma}_{(\beta,\gamma)\cup[\gamma,\alpha)}=\lambda_{NP}\cdot c_3+\lambda_{PN}\cdot d_3$。则两者的总损失相减，得 $Cost^{\gamma}_{(\beta,\gamma)\cup[\gamma,\alpha)}-Cost^{(\alpha,\beta)}_{(\beta,\gamma)\cup[\gamma,\alpha)}=[(\lambda_{NP}-\lambda_{BP})\cdot c_3-\lambda_{BN}\cdot d_2]+[(\lambda_{PN}-\lambda_{BN})d_3-\lambda_{BP}\cdot c_2]>0$。故$\gamma$-二枝决策方式比$(\alpha,\beta)$-三枝决策方式总损失更大，后者在$(C_1')$和$(C_2)$下更优。

5 决策粗集理论的应用过程

三枝决策思想已经在医学、工程、管理、石油勘探等领域得到了广泛的应用[3,4]。Pauker和Kassirer利用两个阈值“检验参数”和“检验-治疗参数”将医疗诊断问题分为3类：治疗、进一步检查、不需治疗[16]。Woodward和Naylor将三枝决策用到产品检验上：接受不需检测、直接拒绝和需要进一步检测[17]。Li等将文本分类过程分为3种情形：相关的文本、不相关的文本和可能相关的文本[18]。Zhao等将三枝决策思想用于电子邮件的过滤中，并提出一个通用信息过滤模型[19]。Zhou等将三枝决策思想用到邮件分类上，所有邮件被分为3类：正确邮件、垃圾邮件和可疑邮件[20]。Yusgiantoro指出所有的油田可能具有3种特征：富有高可能性矿床、有待被开采和开发的油田；具有低可能性碳氢化合物特征，没有石油储备的油田；可能有矿床存在，需要进一步考察的油田[21]。此外，Macmillan认为石油开采有3种情况：钻井、不钻井、需要更多资料再决定是否钻井[22]。由此可见，三枝决策将研究事物分为3个部分，这真实地逼近了实际决策过程。下面，结合决策粗集理论，给出一种利用三枝决策解决现实问题的方法。

由本文第2节决策粗集理论可知，(α,β)-三枝决策与两部分参数相关。正如规则$(P1)$-$(N1)$所示，右边部分的参数α,β可由采取不同行动的6个损失函数 λ_{PP}，λ_{BP}，λ_{NP}，λ_{PN}，λ_{BN}，λ_{NN}给出，而左边的条件概率 $\Pr(X|[x])$可通过信息表直接计算得出。下面给出 $\Pr(X|[x])$的具体计算过程。

根据贝叶斯理论，条件概率 $\Pr(X|[x])$可由先验概率 $\Pr(X)$和相似比$\frac{\Pr([x]|X)}{\Pr([x])}$计算得到：$\Pr(X|[x])=\Pr(X)\cdot\frac{\Pr([x]|X)}{\Pr([x])}$。其中，$\Pr(X)$又可从全概率公式得到，$\Pr([x])=\Pr(X)\cdot\Pr([x]|X)+\Pr(\neg X)\cdot\Pr([x]|\neg X)$，则

$$\Pr(X|[x])=\frac{\Pr(X)\cdot\Pr([x]|X)}{\Pr([x])}=\frac{\Pr(X)\cdot\Pr([x]|X)}{\Pr(X)\cdot\Pr([x]|X)+\Pr(\neg X)\cdot\Pr([x]|\neg X)}\tag{9}$$

基于信息表的角度，式(9)中的 $\Pr(x)$即是基于决策属性划分下状态集 X 的概率，$\Pr(x)$即是基于条件属性划分下等价类$[x]$的概率。由于等价类$[x]$可表示成具有相同属性值的元素的集合，为简便起见，记 $\Pr(x)=\Pr(v_1,v_2,\cdots,v_m)$，其中 m 为条件属性的个数，v_i 为$[x]$在第 i 个条件属性下的取值。同理，$\Pr([x]|X)=\Pr(v_1,v_2,\cdots,v_m|X)$。

考虑到 $\Pr([x])$和 $\Pr([x]|X)$的展开式，假设信息系统中的属性相互独立，则

$$\Pr([x])=\Pr(v_1,v_2,\cdots,v_m)=\prod_{i=1}^{m}\Pr(v_i)\tag{10}$$

$$\Pr([x]|X)=\Pr(v_1,v_2,\cdots,v_m|X)=\prod_{i=1}^{m}\Pr(v_i|X)\tag{11}$$

可以看到，计算条件概率 $\Pr([x]|X)$的相关参数 $\Pr(X)$，$\Pr(v_i)$，$\Pr(v_i|X)$都能够从信息系统中直接得出。

值得注意的是，为了计算方便，可对式(10)和式(11)作Logit变换，这样可将等式右端的相乘计算变为相加计算，从而大大提高运算的效率。至此，对于实际决策问题，可按上述分析分别计算相关参数，并利用规则$(P1)$，$(B1)$，$(N1)$对研究对象进行分类和决策。

总而言之，在决策粗集理论的(α,β)-三枝决策过程中，条件概率 $\Pr(X|[x])$可完全从信息系统计算得出，它是通过机器学习得到的，是客观的；阈值 α 和 β 是通过行动损失参数给出的，它是通过人类经验得到的，是主观的。利用 α 和 β 去验证条件概率 $\Pr(X|[x])$的正确性，利用条件概率 $\Pr(X|[x])$去指导 α 和 β 设置的合理性，两者相辅相成，互为补充。因而，决策粗集理论体现了一种主观和客观相结合、“人机合一”的思想。

结束语 本文从贝叶斯理论及其决策过程出发，概述了决策粗集理论提出的背景、模型、语义、优势和应用价值。首先，考虑到不同行动所产生的损失不同，决策粗集引入贝叶斯决策过程，给出了基于贝叶斯最小风险下的三枝决策模型及其规则判定准则。其次，基于统计学假设检验角度，分析$(\alpha,$

β)-三枝决策、(1,0)-三枝决策和γ-二枝决策各自犯错的详细情况，给出(α,β)-三枝决策优于其他两者的成立条件。最后，利用贝叶斯理论来诠释决策粗集三枝决策方法在实际问题中的应用过程。相对于其他概率粗集理论，决策粗集理论具有其深刻的研究背景和独有的语义解释。可以预见，决策粗集理论不仅能够给予决策问题以理论指导，而且能在实际应用和实践中发挥越来越重要的作用。

参 考 文 献

[1] Pawlak Z. Rough sets [J]. International Journal of Computer and Information Sciences, 1982, 11: 341-356

[2] Yao Y Y. Three-way decision: an interpretation of rules in rough set theory [J]. LNAI, 2009(5589): 642-649

[3] Yao Y Y. Three-way decisions with probabilistic rough sets [J]. Information Sciences, 2010, 180: 341-353

[4] Yao Y Y. Two semantic issues in a probabilistic rough set model [J]. Fundamenta Informaticae, Manuscript, 2009

[5] Pawlak Z, Wong S K M, Ziarko W. Rough sets: probabilistic versus deterministic approach [J]. Inter. Journal of Man-Machine Studies, 1988, 29: 81-95

[6] Yao Y Y, Wong S K M. A decision theoretic framework for approximating concepts [J]. Inter. Journal of Man-machine Studies, 1992, 37: 793-809

[7] Yao Y Y. Decision-theoretic rough set models [J]. Lecture Notes in Artificial Intelligence, 2007(4481): 1-12

[8] Ziarko W. Variable precision rough set model [J]. Journal of Computer and System Sciences, 1993, 46: 39-59

[9] Slezak D. Rough sets and Bayes factor [J]. LNCS Transactions on Rough Sets III, 2005: 202-229

[10] Slezak D, Ziarko W. The investigation of the Bayesian rough set model [J]. International Jounral of Approximate Reasoning, 2005, 40: 81-91

[11] Yao Y Y. Probabilistic approaches to rough sets [J]. Expert Systems, 2003, 20: 287-29

[12] Yao Y Y. Probabilistic rough set approximations [J]. International Jounral of Approximate Reasoning, 2008, 49: 255-271

[13] Ziarko W. Probabilistic approach to rough set [J]. International Jounral of Approximate Reasoning, 2008, 49: 272-284

[14] Duda R, Hart P. Pattern Classification and Scene Analysis [M]. New York: Wiley Press, 1973

[15] Yao Y Y. The superiority of three-way decision in probabilistic rough set models [J]. Information Sciences, Manuscript, 2010

[16] Pauker S, Kassirer J. The threshold approach to clinical decision making [J]. The New England Journal of Medicine, 1980, 302: 1109-1117

[17] Woodward P, Naylor J. An application of Bayesian methods in SPC [J]. The Statistician, 1993, 42: 461-469

[18] Li Y, Zhang C, Swan J. An information filtering model on the Web and its application in JobAgent [J]. Knowledge-Based Systems, 2000, 13: 285-296

[19] 赵文清，朱永利，高伟. 一个基于决策粗糙集理论的信息过滤模型 [J]. 计算机工程与应用，2007, 43(7): 185-187

[20] Zhou B, Yao Y Y, Luo J. A Three-way decision approach to email spam filtering [C]// The 23th Canadian AI. 2010: 28-39

[21] Macmillan F. Risk, Uncertainty and Investment Decision-making in the Upstream Oil and Gas Industry [D]. UK: University of Aberdeen, 2000

[22] Yusgiantoro P, Hsiao F. Production-sharing contracts and decision making in oil production [J]. Energy Economics, 1993, 10: 245-256

(上接第 216 页)

从表 4 可以看出，本系统的召回率较高，说明本系统对于应该识别的零指代项的识别率较高，且完备性较好，但准确率只有 57.3%，说明本系统对于不具有零指代关系的零形代名词的识别率也较高。从表 2 中可以看出，训练和测试的时候正负例的选取比较平衡。从实验数据可以看出，本文所研究的基于树核函数零指代项识别的正确率较基于规则系统的正确率要高，说明本文研究的结构化信息在零指代项识别这个阶段具有一定的研究价值。

结束语 目前，中文零指代消解是中文指代消解研究领域中一个热点课题，作为零指代消解第一阶段的工作——零指代项识别是必不可少的。本文参考 Zhao 和 Ng(2007)的基于机器学习的零指代消解方法，实现了一个基于树核函数的中文零指代项识别系统。为了能够获得包含零形代名词的结构化信息，进行了语料人工标注，构建了一个基准语料库，并通过各种裁剪策略得到包含零形代名词的句法分析树，将其作为结构化特征交由 SVM 进行训练。同时本文实现了一个基于规则的零形指代项识别的系统。通过实验结果比较，本文所研究的基于树核函数的零指代消解具有一定的价值。

参 考 文 献

[1] Zhao Shanheng, Ng H T. Identification and Resolution of Chinese Zero Pronouns: A Machine Learning Approach[C]// ACL. 2007: 541-550

[2] Ferr′andez A, Peral J. A computational approach to zero-pronouns in Spanish[C]// ACL. 2000: 166-172

[3] Yeh Ching-long, Chen Yi-chun. Zero anaphora resolution in Chinese with shallow parsing[J]. Journal of Chinese Language and Computing, 2004

[4] Converse S. Pronominal Anaphora Resolution in Chinese [D]. Department of Computer and Information Science, University of Pennsylvania, 2006

[5] Iida R, Inui K, Matsumoto Y. Exploiting syntactic patterns as clues in zero-anaphora resolution[C]// Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for ComputationalLinguistics COLING-ACL2006. 2006: 625-632

[6] Wu Dian-song, Liang T. Zero anaphora resolution by case-based reasoning and pattern conceptualization [J]. Expert Systems with Applications: Expert Syst, 2008, 36(4): 7544-7551

[7] Zhou G D, Kong F. Global Learning of Noun Phrase Anaphoricity in Coreference Resolution via Label Propagetion[C]// EMNLP. 2009: 978-986

[8] Yang X F, Su J, Tan C L. Kernel-based Pronoun Resolution with Structured Syntactic Knowledge[C]// ACL. 2006: 41-48

[9] Yeh Ching-long, Chen Yi-chun. Using Zero Anaphora Resolution to Improve Text Categorization[C]// Proceeding of PACLIC17. Sentosa, Singapore

[10] Kong F, Zhou G D, Zhu Q M. Employing the Centering Theory in Pronoun Resolution from the Semantic Perspective[C]// EMNLP. 2009: 987-996

[11] 张威，周昌乐. 汉语语篇理解中元指代消解初步[J]. 软件学报，2002, 1(13): 732-738

[12] 王厚峰. 指代消解的基本方法和实现技术[J]. 中文信息学报，2002(6): 9-17