文章编号：1000－5471(2008)06－0009－04

基于粗糙集理论的 Bayes 粗糙模型研究①

马 旭， 陈志恩， 田彦山

宁夏师范学院 数学与计算机科学系，宁夏 固原 756000

摘要：从经典的概率粗糙集推广模型出发，参照 Bayes 因子，研究了一种 Bayes 粗糙集模型，该模型不受先验概率的影响，从而减小了决策失误的风险.

关 键 词：粗糙集；Bayes 因子；粗糙集模型

中图分类号：O212.8 **文献标识码**：A

自从 Pawlak1982 年提出粗糙集理论[1]以来，粗糙集的理论和应用都取到了迅速的进展. 近年来，粗糙集理论已成为人工智能领域中的一个新学术热点，在机器学习、知识获取、知识发现和决策分析等领域得到了广泛的研究与应用. 目前，对粗糙集的研究主要集中在粗糙集的模型推广[2]，问题的不确定性研究，粗糙集的算法研究以及与人工智能其它方向的关系的研究等方面.

在经典的概率粗糙集推广模型中，往往通过将后验概率与 2 个确定参数或与先验概率比较来定义粗糙集，或直接通过比较目标概念的后验概率来定义粗糙集. 然而当 2 个目标概念的先验概率显著不同时，2 个后验概率的比较也失去了意义. 本文参照 Bayes 因子，研究了一种 Bayes 粗糙集模型. 该模型不受先验概率的影响，从而减小了决策失误的风险.

1 粗糙集理论的基本概念

相关的概念可以参见文献[2－5].

定义 1 设 U 是一个非空的有限集合，R 是定义在 U 上的一簇等价关系，则 $K=(U, R)$ 称为一个近似空间，而 U 则称为一个论域. 近似空间也称为知识库. 一种知识就是对于论域的一种划分，从而对应一种等价关系.

定义 2 设 U 是一个论域，$C=\{X_1, X_2, \cdots, X_n\}$ 称为 U 的一个划分，若：

(1) $X_1 \cup X_2 \cup \cdots \cup X_n = U$

(2) $X_i \cap X_j = \phi$，$i \neq j$，$i, j = 1, 2, \cdots n$

则，x_i，$i = 1, 2, \cdots n$. 称为元知识或基本知识.

定义 3 设 U 是一个论域，R 是定义在 U 上的一个等价关系，则将 R 产生的 U 上的划分 $\{[x]_R \mid x \in U\}$ 记作 U/R，其中 $[x]_R$ 为包含 x 的一个等价类，称为 U 的一个基本集. 特别地，ϕ 也称为 U 的一个基本集.

① 收稿日期：2008－04－16

基金项目：宁夏自然科学科研基金资助项目(NZ08113)；宁夏高校科研基金资助项目(20060237)；宁夏师范学院科研基金资助项目(Z070011).

作者简介：马 旭(1971－)，男，宁夏固原人，副教授，主要从事软件理论与计算数学的教学与研究.

定义 4　设 $K=(U, R)$ 是一个近似空间，$X\subseteq U$，R 是 U 的任意一个等价关系，则集合

$$\underline{R}(X)=\{x\in U\mid [x]_R\cap X=\phi\}$$

$$\overline{R}(X)=\{x\in U\mid [x]_R\subseteq X\}$$

分别称为 X 的 R 上近似集和 X 的 R 下近似集. 若 $\underline{R}(X)=\overline{R}(X)$，称 X 为可定义集，否则称 X 为粗糙集，$POS_R(X)=\underline{R}X$ 称为 X 的 R 正域；$NEG_R(X)=U-\overline{R}X$ 称为 X 的 R 负域.

定义 5　称 $S=(U, A, V, f)$ 为一知识表达系统，其中：U 为研究对象的非空有限集合，称为论域；A 为属性的非空有限集合，$A=C\cup D$，$C\cap D=\phi$，C 为条件属性集，D 为决策属性集；$V=\bigcup\limits_{a\in A}V_a$，$V_a$ 是属性 a 的值域；$f: U\times A\to V$ 是一个信息函数，它为每个对象的每个属性赋予一个信息值，即 $\forall a\in A$，$x\in U$，$f(x, a)\in V_a$.

知识表达系统也称为信息系统. $S=(U, A, V, f)$ 也可表示为 S=(U, A).

定义 6　设 S=(U, A)是一个信息系统，其中，U 是论域；A 是有限个属性的非空集合，可进一步划分为条件属性集 C(用于描述对象)和决策属性集 D(用于描述分类)，且满足 $A=C\cup D$，$C\cap D=\phi$. 称信息系统 $I=(U, C\cup D)$ 为决策表.

2　粗糙贝叶斯模型

定义 7　令 P 为定义在论域 U 上可测子集所构成的 σ-代数 $M(U)$ 上的概率测度. 用 U 的子集 X 表示一个随机事件，称为概念，$0\leqslant P(X)\leqslant 1$. 则称 $P(X)$，$P(X\mid E)$，$P(E\mid X)$ 分别为概念 X 基于知识 E 的先验概率、后验概率和逆概率：

$$P(X)=\frac{\mathrm{Card}(X\cap U)}{\mathrm{Card}(U)},\ P(X\mid E)=\frac{\mathrm{Card}(X\cap E)}{\mathrm{Card}(E)},\ P(E\mid X)=\frac{\mathrm{Card}(X\cap E)}{\mathrm{Card}(X)}$$

其中 Card(*) 表示一个集合的基数. 对任意 X_l，$X_k\subseteq U$，$l\neq k$，记

$$B_l^k=\frac{P(E\mid X_k)}{P(E\mid X_l)}\left(\frac{P(E\mid X_k)}{P(E\mid X_l)}=\frac{P(X_k\mid E)/P(X_k)}{P(X_l\mid E)/P(X_l)}=\frac{P(X_k\mid E)/P(X_l\mid E)}{P(X_k)/P(X_l)}\right)$$

B_l^k 为基于数据集的后验概率与先验概率的比率，称 B_l^k 为贝叶斯因子.

贝叶斯因子用于刻画已知知识 E 对概念 X_k 的相对支持度. 相对后验概率定义的粗糙集，它的优点在于可靠性高且有很强的数据敏感性.

在只有 2 个决策类的信息系统中，用 B_0^1 表示贝叶斯因子，即 $B_0^1=\dfrac{P(E\mid X_1)}{P(E\mid X_0)}$，$B_0^1\geqslant 0$. 若 $B_0^1>1$，表明在已知知识 E 下，概念 X_1 相对于 X_0 的确定性程度大；若 $0<B_0^1<1$，结果相反；若 $B_0^1=1$，不能判断 X_1 与 X_0 的确定性程度. 若 $P(E\mid X_0)=0$ 则表达式无意义.

因此为了便于描述概念 X_1 相对于 X_0 的完全性程度，可采用下述一般准则知识 E 对概念 X_k 的相对支持度：

$$P(E\mid X_0)\leqslant \varepsilon_1^0 P(E\mid X_1)\qquad \varepsilon_0^1\in[0, 1) \tag{1}$$

式(1)中当 $0<\varepsilon_0^1<1$ 时，表示 X_1 的确定性程度大于 X_0；当 $\varepsilon_1^0=0$ 时，表示 X_1 为可定义集；当 $\varepsilon_0^1\neq 0$ 时，$B_0^1=\dfrac{1}{\varepsilon_0^1}$.

按照式(1)，在只有 2 个决策类的信息系统里，可根据显著性阈值 $\varepsilon\in[0, 1)$ 来定义目标概念的正域、负域及边界区域，因为这种模型与贝叶斯因子有关，所以称这种模型为贝叶斯粗糙集模型.

定义 8　设决策表 $I=(U, A\cup D)$，$U/D=\{X_0, X_1\}$，$B\subseteq A$，对任意参数 $\varepsilon=(\varepsilon_0^1, \varepsilon_1^0)$，$\varepsilon_0^1$，$\varepsilon_1^0\in[0, 1)$，定义 X_1 的粗糙贝叶斯正域、负域和边界域(简称 RB − 区域)分别为：

(1) $BAYPOS_B^{\varepsilon}(X_1)=\bigcup\{E\in U/B: P(E\mid X_0)\leqslant \varepsilon_1^0 P(E\mid X_1)\}$

(2) $BAYNEG_B^{\beta}(X_1)=\bigcup\{E\in U/B: P(E\mid X_1)\leqslant \varepsilon_0^1 P(E\mid X_0)\}$

(3) $BAYBND_B^{\varepsilon}(X_1) = \bigcup \{E \in U/B: P(E|X_0) > \varepsilon_1^0 P(E|X_1) \wedge P(E|X_1) > \varepsilon_0^1 P(E|X_0)\}$ (2)

类似可写出 X_0 的 RB − 区域.

说明：

1) 显著性阈值 $\varepsilon = (\varepsilon_0^1, \varepsilon_1^0)$ 中允许 $\varepsilon_0^1 = \varepsilon_1^0$，此时可用一个符号 ε 来代替 2 个系数.

2) 若 $\varepsilon_0^1 \neq 0$ 且 $\varepsilon_1^0 \neq 0$，则 RB − 正域、负域及边界域也可分别用：$B_0^1 \geqslant 1/\varepsilon_1^0$，$B_1^0 \geqslant 1/\varepsilon_0^1$ 及 $\max\{B_0^1\varepsilon_1^0, B_1^0\varepsilon_0^1\} < 1$ 来表示.

3) 若 $\varepsilon = (0, 0)$，则 RB − 区域为经典粗糙集区域.

4) 粗糙贝叶斯模型通过直接比较 2 个目标概念来评价概念的确定性程度，当 $\varepsilon_0^1 \to 1$ 时，该模型就变为粗糙集模型.

定理 1　已知 $I = (U, A \cup D)$，$U/D = \{X_0, X_1\}$，令 $\varepsilon = (\varepsilon_0^1, \varepsilon_1^0)$，对任意 $\varepsilon_0^1, \varepsilon_1^0 \in [0, 1)$，$RB$ − 区域对应于相应的 $VPRS$ − 区域，并且相应参数为：

$$\alpha_0^{\varepsilon} = \varepsilon_1^0 P(X_0)/(\varepsilon_1^0 P(X_0) + P(X_1)),\ \beta_0^{\varepsilon} = P(X_0)/(P(X_0) + \varepsilon_0^1 P(X_1))$$

$$\alpha_1^{\varepsilon} = \varepsilon_0^1 P(X_1)/(\varepsilon_0^1 P(X_1) + P(X_0)),\ \beta_1^{\varepsilon} = P(X_1)/(P(X_1) + \varepsilon_1^0 P(X_0))$$

证　由贝叶斯公式：

$$P(X_1|E) = \frac{P(E|X_1)P(X_1)}{P(E|X_1)P(X_1) + P(E|X_0)P(X_0)}$$

若 $P(E|X_0) \leqslant \varepsilon_1^0 P(E|X_1)$，则

$$\frac{P(E|X_1)P(X_1)}{P(E|X_1)P(X_1) + P(E|X_0)P(X_0)} \geqslant \frac{P(E|X_1)P(X_1)}{P(E|X_1)P(X_1) + \varepsilon_1^0 P(E|X_1)P(X_0)}$$

比较可得

$$P(X_1|E) \geqslant \frac{P(X_1)}{P(X_1) + \varepsilon_1^0 P(X_0)} = \beta_1^{\varepsilon}$$

其它几个参数可类似证明.

说明：定理 1 中所得 α_0^{ε}、β_1^{ε}、α_1^{ε}、β_0^{ε} 满足：

1) $0 \leqslant \alpha_0^{\varepsilon} < P(X_0) < \beta_0^{\varepsilon} \leqslant 1$，$0 \leqslant \alpha_1^{\varepsilon} < P(X_1) < \beta_1^{\varepsilon} \leqslant 1$.

2) $\alpha_0^{\varepsilon} + \beta_1^{\varepsilon} = 1$，$\alpha_1^{\varepsilon} + \beta_0^{\varepsilon} = 1$.

3　应用举例

例　如表 1 所示，已知信息系统 $I = (U, A \cup D)$，$U = \{x_1, x_2, \cdots x_{20}\}$，$A = \{a_1, a_2, \cdots, a_5\}$，$U/D = \{X_0, X_1\}$，$B \subseteq A$，$B = \{a_1, a_3\}$，则：

$$X_0 = \{x_1, x_2, x_5, x_7, x_{10}, x_{11}, x_{14}, x_{15}, x_{17}, x_{18}\}$$

$$X_1 = \{x_3, x_4, x_6, x_8, x_9, x_{12}, x_{13}, x_{16}, x_{19}, x_{20}\}$$

$$U/B = \{E_1, E_2, E_3, E_4, E_5\}$$

其中：$E_1 = \{x_1, x_{11}, x_{12}, x_{16}, x_{17}\}$，$E_2 = \{x_2, x_{10}\}$，$E_3 = \{x_3, x_6, x_8, x_9, x_{15}, x_{19}, x_{20}\}$，$E_4 = \{x_4, x_7, x_{13}\}$，$E_5 = \{x_5, x_{14}, x_{18}\}$.

若取系数 $\varepsilon_0^1 = \varepsilon_1^0 = 1/5$，则 RB − 区域如下：

$$BAYPOS_B{}^{1/5}(X_1) = \{x_3, x_6, x_8, x_9, x_{15}, x_{19}, x_{20}\} = \{E_3\}$$

$$BAYNEG_B{}^{1/5}(X_1) = \{x_2, x_5, x_{10}, x_{14}, x_{18}\} = \{E_2, E_5\}$$

$$BAYBND_B{}^{1/5}(X_1) = \{x_1, x_4, x_7, x_{11}, x_{12}, x_{13}, x_{16}, x_{17}\}$$

根据数据表 $P(X_1) = 1/2$，则进而可求得相应的 VPRS− 区域的上下阈值 $\alpha_1^{1/5}$，$\beta_1^{1/5}$ 如下：

$$\alpha_1^{1/5} = \frac{1/5 \cdot 1/2}{1/5 \cdot 1/2 + (1 - 1/2)} = \frac{1}{6} \qquad \beta_1^{1/5} = \frac{1/2}{1/2 + 1/5(1 - 1/2)} = \frac{5}{6}$$

表1 决策系统

U	a_1	a_2	a_3	a_4	a_5	d	U	a_1	a_2	a_3	a_4	a_5	d
x_1	1	1	0	1	2	0	x_{11}	1	2	0	0	2	0
x_2	0	0	0	2	2	0	x_{12}	1	1	0	1	2	1
x_3	2	2	2	1	1	1	x_{13}	0	1	2	2	2	1
x_4	0	1	2	2	2	1	x_{14}	2	1	1	0	2	0
x_5	2	1	1	0	2	0	x_{15}	2	2	2	1	1	0
x_6	2	2	2	1	1	1	x_{16}	1	1	0	1	2	1
x_7	0	1	2	2	2	0	x_{17}	1	1	0	1	2	1
x_8	2	2	2	1	1	1	x_{18}	2	1	1	0	2	0
x_9	2	2	2	1	1	1	x_{19}	2	2	2	1	1	1
x_{10}	0	0	0	2	2	0	x_{20}	2	2	2	1	1	1

4 小 结

本文从经典的概率粗糙集推广模型出发，参照贝叶斯因子，研究了一种贝叶斯粗糙集模型，该模型不受先验概率的影响，从而减小了决策失误的风险.

参考文献：

[1] Pawlak. Z. Rough sets [J]. International Journal of Computer and Information Sciences, 1982, 11 : 314－356.
[2] 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.
[3] 张文修，吴伟志，梁吉业，等. 粗糙集理论与方法 [M]. 北京：科学出版社，2001.
[4] 张文修，梁 怡，吴伟志，等. 信息系统与知识发现 [M]. 北京：科学出版社，2003.
[5] 徐宗本，张讲社，郑亚林. 计算智能中的仿生学 [M]. 北京：科学出版社，2003.

A Rough Model of Bayes Based on Rough Sets Theory

MA Xu, CHEN Zhi-en, TIAN Yan-shan

Department of Mathematics and Computer Science, Ningxia Teacher's College, Guyuan Ningxia 756000, China

Abstract: Starting from a classical probability rough set extending model, consulting the Bayes factor, a rough model of Bayes has been studied. This model is not effected by prior probability, therefore the risk of the fault-making policy is diminished.

Key words: rough set; Bayes factor; rough set model

责任编辑 张 枸