对 Bayesian 粗糙集模型的讨论*)

闫德勤
(辽宁师范大学计算机系　大连 116029)

摘　要　变精度粗糙集模型是对传统的(Pawlak)粗糙集模型的一个重要拓展,但变精度模型中需要设定人为参数不利于信息的客观体现。Bayesian 粗糙集模型是基于变精度和概率论的思想最新提出的无参数模型。对 Bayesian 粗糙集模型进行了分析,指出了其中的不足,提出了一种改进形式。

关键词　粗糙集,变精度模型,Bayesian 粗糙集模型

A Discussion to the Bayesian Rough Set Model

YAN De-Qin
(Department of Computer Science, Liaoning Normal University, Dalian 116029)

Abstract　In this paper, the newly proposed Bayesian rough set model is analyzed, and the defect of the model is discussed, further more, a modified form of the model is proposed.

Keywords　Rough sets, Variable precision rough set model, Bayesian rough set model

1　引言

粗糙集理论[1,5]在计算机应用的很多领域如数据分析、机器学习、知识发现等有着重要的应用,对其理论方法及其应用的研究不断产生有价值的成果而成为国内外当前的一个研究热点。为使粗糙集对信息处理和利用更有效率,Ziarko 改进了传统的粗糙集模型,提出了著名的变精度模型[2]。变精度粗糙集模型增强了粗糙集的适应能力,受到广泛的重视。然而,变精度模型中上下近似的界定依赖人为设置的精度参数。这是一个不足之处。虽然 Katzberg 和 Ziarko 在文[3]中对此不足做了改进,但仍然没有解决依赖人为参数决定上下近似的问题。为此,Slezk 和 Ziarko 最近提出了一种 Bayesian 粗糙集模型[4]。该模型以概率表示为基础,给出了正域、负域和边界的定义方法。

本文在对变精度粗糙集进行一定程度的讨论基础上,对 Bayesian 粗糙集模型进行了分析,指出了其中存在的问题。最后,给出了改进的方法。

2　关于 Pawlak 模型和变精度模型

设 $U=\{x_1,x_2,\cdots,x_n\}$ 是一有限集,称为论域,R 是 U 上的一个等价关系,U/R 表示在 U 上导出的所有等价类;$[x]_R$ 表示包含元素 x 的 R 的等价类,$x\in U$。

Pawlak 粗糙集模型基于传统的粗糙集定义方法[1,5]:

对任集合 $X\subseteq U$

$R_-(X)=\{x\in U|[x]_R\subseteq X\}$

$R^-(X)=\{x\in U|[x]_R\cap X\neq\varnothing\}$($\varnothing$为空集)

分别称 $R_-(X)$ 与 $R^-(X)$ 为 X 的 R 下近似 X 和 R 的上近似。上下近似不相等时,称对于 R,X 为粗集。在省略 X 时由 (R_-,R^-) 表示粗糙集。

$Bn_R(X)=R^-(X)-R_-(X)$ 称为 X 的 R 边界,也称为粗糙集的边界。$Pos_R(X)=R_-(X)$ 称为 X 的 R 正域,也称为粗糙集的正域。$neg_R(X)=U-R_-(X)$ 为粗糙集的负域。

基于这样定义的粗糙集理论由 Pawlak 教授于 1982 年首次提出[1],其形式称为基本的粗糙集模型(或 Pawlak 模型),现又称为传统的粗糙集模型。

Pawlak 模型由等价类严格限制了边界的范围,为信息更多地得到利用,Ziarko 提出了变精度粗糙集模型[2]。

变精度粗糙集模型基于如下的定义:

设 X 和 Y 为非空有限论域 U 上的子集,把 X 错误分类到 Y 的度量 $C(X,Y)$ 定义为

$C(X,Y)=1-card(X\cap Y)/card(X)$

(若,$card(X)$ 定义 $C(X,Y)=0$)

其中,$card(\cdot)$ 表示基数。

设等价关系 R 在论域 U 上的等价类为 $\tilde{R}=\{E_1,E_2,\cdots,E_k\}$,变精度粗糙集 β-下近似(β-正域)和 β-上近似定义为

$\underline{R}_\beta(X=\bigcup\{E\in\tilde{R}:C(E,X)\leqslant\beta\}$,

$\overline{R}(X)=\bigcup\{E\in\tilde{R}:C(E,X)<1-\beta\}$。

相应地变精度粗糙集 β-边界区域和 β-负域为

$BNR_\beta(X)=\bigcup\{E\in\tilde{R}:\beta<C(E,X)<1-\beta\}$,

$NEGR_\beta(X)=\bigcup\{E\in\tilde{R}:C\{E,X\}\geqslant 1-\beta\}$。

由于这样定义的变精度模型依赖参数 β,一般简写为 $VPRS_\beta$(依赖 β 的变精度粗糙集)。

可见变精度粗糙集相对于传统粗糙集扩大了正域的范围,减小了边界域的范围。应用变精度粗糙集可以最大限度地获取边界信息,同时应用 β-上近似的过滤也可减少一定的噪声干扰。

然而,参数 β 使得变精度模型比传统的模型更灵活的同时,也产生了一定的限制:一旦参数 β 确定,上下近似同时受到一个参数的制约。为此,Katzberg 和 Ziarko 在文[3]中提

*)国家自然科学基金(60372071)资助;辽宁省教育厅高等学校科学研究基金(2004C031)资助;辽宁师范大学校基金资助。闫德勤　博士,教授,主要研究领域为模式识别、数据挖掘等。

出了一种改进的变精度粗糙集模型(简称 $VPRS_{l,u}$ 模型)。

$VPRS_{l,u}$ 模型中提供了两个人为的可变参数:l 和 u,满足条件 $0\leqslant l<P(X)<u\leqslant 1$。$P(X)$ 是以论域 U 为样本空间,集合 X 为样本点的概率表示。相关定义如下:

设等价关系 R 在论域 U 上的等价类为 $\tilde{R}=\{E_1,E_2,\cdots,E_k\}$。对于 $X\subset U$:

u-正域为

$$POS_u(X)=\bigcup\{E\in\tilde{R}:P(X|E)\geqslant u\}$$

l-负域为

$$NEG_l(X)=\bigcup\{E\in\tilde{R}:P(X|E)\leqslant l\}$$

(l,u)-边界为

$$BND_{l,u}(X)=\bigcup\{E\in\tilde{R}:l<P(X|E)<u\}$$

其中,$P(X|E)$ 为条件概率。

下面我们把 $VPRS_\beta$ 和 $VPRS_{l,u}$ 两个模型做一下比较。

由于 $C(X,Y)=1-card(X\cap Y)/card(X)$ 可以得到

$C(E,X)=1-card(E\cap X)/card(E)=1-P(X|E)$,

因此,在 $VPRS_\beta$ 模型下,β-正域、负域和边界可写为:

$$\underline{R}_\beta(X)=\bigcup\{E\in\tilde{R}:P(X|E)\geqslant 1-\beta\}$$

$$NEGR_\beta(X)=\bigcup\{E\in\tilde{R}:P(X|E)\leqslant\beta\}$$

$$BNR_\beta(X)=\bigcup\{E\in\tilde{R}:\beta<P(X|E)<1-\beta\}$$

β-上近似定义可写为:

$$\overline{R}_\beta(X)=\bigcup\{E\in\tilde{R}:P(X|E)>\beta\}$$

在 $VPRS_{l,u}$ 模型下,可以得到对于 $X\subset U$,上近似的表示:

$$POS_l(X)=\bigcup\{E\in\tilde{R}:P(X|E)>l\}$$

由此可见 $VPRS_{l,u}$ 模型是把 $VPRS_\beta$ 模型中的参数 β 和 $1-\beta$ 分别用两个参数 l 和 u 代替的一种改进。改进后的模型使得粗糙集边界更有弹性。

3 关于 Bayesian 粗糙集模型

虽然 $VPRS_{l,u}$ 模型具有很大程度的灵活性,但其中的参数是人为的,为了克服这一缺陷 Slezk 和 Ziarko 最近提出了一种新的模型:Bayesian 粗糙集模型[4]。

Bayesian 粗糙集模型基于贝叶斯(Bayesian)推理的思想定义出关于集合 X 的正域、负域和边界:

$$POS^*(X)=\bigcup\{E\in\tilde{R}:P(X|E)>P(X)\}$$

$$NEG^*(X)=\bigcup\{E\in\tilde{R}:P(X|E)<P(X)\}$$

$$BND^*(X)=\bigcup\{E\in\tilde{R}:P(X|E)=P(X)\}$$

但是在 Bayesian 粗糙集模型的定义中我们看到一些不足:

(1)当 $P(X)$ 很小时,满足 $P(X|E)>P(X)$ 的所有等价类 E 都成为了正域的集合,这就意味着正域中有一些等价类含有很少的 X 中的元素。这不符合定义正域的思想。

(2)当 $P(X)$ 很大时,使用负域的定义会损失大量的信息。这也不符合改进传统粗糙集模型以获得更多信息的思想。

(3)对于边界的定义,Slezk 和 Ziarko[4] 认为边界区域完全与 X 无关,并且从概率事件的角度认为 X 与其它等价类 (E) 事件是相互独立的。事实上这样的认为是不合理的。这是因为边界区域中存在着与 X 有关的信息,而信息量的比例与等价类有关。在该模型中,边界的定义存在的另一个问题是满足 $P(X|E)=P(X)$ 条件的等价类在实际计算中很少,因此在很多情况下,边界区域是空集。这不符合改进传统粗糙集模型更柔性化的思想。

针对以上 3 点不足我们做如下改进尝试:

定义 1 设等价关系 R 在论域 U 上的等价类为 $\tilde{R}=\{E_1,E_2,\cdots,E_k\}$。$X$ 为论域 U 上的集合,令 $K(X)=\max\{P(X),1-P(X)\}$,关于集合 X 的正域、负域和边界定义为:

$$POS(X)=\bigcup\{E\in\tilde{R}:P(X|E)>K(X)\}$$

$$NEG(X)=\bigcup\{E\in\tilde{R}:P(X|E)<1-K(X)\}$$

$$BND(X)=\bigcup\{E\in\tilde{R}:1-K(X)<P(X|E)<K(X)\}$$

结论 本文讨论了 Bayesian 粗糙集模型以及相关变精度模型。指出了 Bayesian 粗糙集模型存在的不足,提出了改进方法。对 Bayesian 粗糙集模型的进一步研究和应用具有重要性。

参 考 文 献

1 Pawlak Z. Rough set. International Journal of Computer and Information Science, 1982, 11(5): 341~356
2 Ziarko W. Variable precision rough set model. Journal of Computer and System Science, 1993, 46(1): 39~59
3 Katzberg JD, Ziarko W. Variable precision extension of rough sets. Fundamenta Informaticae, 1996, 27: 155~168
4 Slezk D, Ziarko W. The investigation of the Bayesian rough set model. International Journal of Approximate Reasoning, 2005, 40: 81~89
5 曾黄麟. 粗集理论及其应用(修订版). 重庆:重庆大学出版社, 1998

(上接第 161 页)

当条件属性中包含的标准多于一个的时候,多个标准会使条件属性部分的知识粒子更加复杂,如何更有效地推得决策规则;

根据条件属性和决策属性部分的知识粒子推导规则的算法并不容易给出,本文根据提出的原则形成规则,是否存在能得到完备规则集且时间效率较高的算法;

含序粗集方法中,如何有效地获得决策表的简式以及核;

当存在基于支配原则不一致时,如何消除之及进行化简;

如何从决策表中推导得出标准的有序信息,而不是由决策者作为先验知识给出等。

参 考 文 献

1 Pawlak Z. Rough sets: theoretical aspects of reasoning about data. In: System Theory, Knowledge Engineering and Problem Solving, Kluwer Academic Publishers, Dordrecht, 1991, 9
2 Greco S, Matarazzo B, Slowinski R. Rough sets methodology for sorting problems in presence of multiple attributes and criteria. European Journal of Operational Research, 2002, 138: 247~259
3 Greco S, Matarazzo B, Slowinski R. Rough approximation of a preference relation by dominance relations. European Journal of Operational Research, 1999, 117: 63~83
4 Greco S, Matarazzo B, Slowinski R. Rough sets theory for multicriteria decision analysis. European Journal of Operational Research, 2001, 129: 1~47
5 Slowinski R, Greco S, Matarazzo B. Rough Set Analysis of Preference-Ordered Data. Rough Sets and Current Trends in Computing, 2002. 44~59
6 Gediga G, Dumtsch I. Approximation quality for sorting rules. Computational Statistics & Data Analysis, 2002, 40
7 Sun Chengmin, Liu Dayou, Sun Shuyang. Containing Order Rough Set Methodology. In: Proceedings of 2005 International Conference on Machine Leaning and Cybernetics, vol 3 of 9, Guangzhou, China, 2005. 1722~1727