粗糙集理论中概念与运算的信息表示*

苗夺谦[1,2] 王 珏[2]

[1](山西大学数学系 太原 030006)

[2](中国科学院自动化研究所 北京 100080)

E-mail: dqmiao@deer.sxu.edu.cn

摘要 粗糙集理论对知识进行了形式化定义,为知识处理提供了一套严密的分析工具,但在代数表示下,粗糙集理论的本质不易被理解,并且,尚无高效的知识约简算法.该文首先建立了知识与信息之间的关系;然后,在此基础上给出了粗糙集理论中概念与运算的信息表示;最后,证明了知识约简在信息和代数两种不同表示下是等价的.这些结论有助于人们深刻理解粗糙集理论的本质,同时,为寻找高效的知识约简算法奠定了基础.

关键词 粗糙集,知识表示,知识约简,信息熵,等价性.

中图法分类号 TP18

粗糙集(Rough Set)理论是由Z. Pawlak于1982年提出的[1].这一理论为处理具有模糊、不精确或不完全信息的分类问题提供了一种新的工具.其主要思想是,在保持信息系统分类能力不变的前提下,通过知识约简,导出问题的决策或分类规则.目前,粗糙集理论已被应用于机器学习、故障诊断、控制算法获取、过程控制以及关系数据库中的知识获取等各种应用领域,并取得了很大成功[2~4].

粗糙集理论中所有的概念和运算都是通过代数学的等价关系和集合运算来定义的,我们称之为粗糙集理论的代数表示.在代数表示下,粗糙集理论的很多概念与运算的直观性较差,人们不容易理解其本质.另外,在此表示下,目前还没有关于知识约简的高效算法.同一问题在不同知识表示下的算法难度是不同的[5].

1 粗糙集理论中主要概念的代数表示

为了在第3节中证明本文给出的信息表示与粗糙集理论的代数表示之间的等价性,本节有必要先介绍一下该理论的一些主要概念与运算,现将这些概念的代数表示罗列如下[1]:

设U是一个论域,R是U上的一个等价关系.U/R表示R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,$x\in U$.一个知识库就是一个关系系统$K=(U,P)$,其中U为论域,P是U上的一个等价关系族.如果$Q\subseteq P$,且$Q\neq\varnothing$,则$\cap Q$(Q的所有等价关系的交)也是一个等价关系,记作$IND(Q)$.

定义1.1. 设$K=(U,P)$和$K_1=(U,Q)$是两个知识库.如果$IND(P)=IND(Q)$,则称K和K_1(或P和Q)是等价的,记作$K\cong K_1$(或$P\cong Q$).

知识库K和K_1等价,意味着K和K_1具有相同的基础类,因而它们具有相同的表达能力.

定义1.2. 设U为一个论域.P为定义在U上的一个等价关系族,$R\in P$.如果$IND(P-\{R\})=IND(P)$,则称关系R在P中是不必要的(多余的);否则,称R在P中是必要的.

不必要的关系在知识库中是多余的.如果将它从知识库中去掉,不会改变该知识库的分类能力.相反,若从

* 本文研究得到国家863高科技项目基金、国家青年基金和山西省青年基金资助.作者**苗夺谦**,1964年生,博士,副教授,主要研究领域为粗糙集理论,人工智能与模式识别.**王珏**,1948年生,研究员,博士生导师,主要研究领域为人工智能与模式识别.

本文通讯联系人:苗夺谦,太原 030006,山西大学数学系

本文1996-10-09收到原稿,1998-03-03收到修改稿

知识库中去掉一个必要的关系，则一定改变该知识库的分类能力.

定义 1.3. 设 U 为一个论域，P 为定义在 U 上的一个等价关系族，$R\in P$. 如果每个关系 $R\in P$ 在 P 中都是必要的，则称关系族 P 是独立的；否则，称 P 是相依的.

对于相依的关系族来说，其中包含有多余关系，可以对其约简；而对于独立的关系族，去掉其中任何一个关系都将破坏知识库的分类能力.

定义 1.4. 设 U 为一个论域，P 为定义在 U 上的一个等价关系族. P 中所有必要关系组成的集合，称为关系族 P 的核，记作 $CORE(P)$.

定义 1.5. 设 U 为一个论域，P,Q 为 U 上的两个等价关系族，且 $Q\subseteq P$.

如果 (1) $IND(Q)=IND(P)$；

(2) Q 是独立的，

则称 Q 是 P 的一个约简.

如果知识 Q 是知识 P 的约简，那么，U 中通过知识 P 可区分的对象，同样可以用知识 Q 来区分. 知识约简是粗糙集理论中最重要的概念.

2 知识与信息熵的关系

粗糙集理论从新的视角对知识进行了定义，把知识看作是关于论域的划分，从而使得对知识能够进行严密的分析与处理. 本节我们将对粗糙集理论中的知识作新的理解，建立知识与信息熵的关系. 这是本文所讨论的信息表示的基础.

设 U 为一个论域，P,Q 为 U 上的两个等价关系(即知识). 我们认为 U 上任一等价关系都可以看作是定义在 U 的子集组成的 σ-代数上的一个随机变量. 其概率分布可通过如下方法来确定.

设 P,Q 在 U 上导出的划分分别为 X,Y：

$$X=\{X_1,X_2,\ldots,X_n\},\qquad Y=\{Y_1,Y_2,\ldots,Y_m\},$$

则 P,Q 在 U 的子集组成的 σ-代数上定义的概率分布为

$$[X;p]=\begin{bmatrix}X_1 & X_2 & \ldots & X_n\\ p(X_1) & p(X_2) & \ldots & p(X_n)\end{bmatrix},\qquad [Y;p]=\begin{bmatrix}Y_1 & Y_2 & \ldots & Y_m\\ p(Y_1) & p(Y_2) & \ldots & p(Y_m)\end{bmatrix},$$

其中 $p(X_i)=\dfrac{|X_i|}{|U|},i=1,2,\ldots,n;p(Y_j)=\dfrac{|Y_j|}{|U|},j=1,2,\ldots,m$；符号 $|E|$ 表示集合 E 的基数.

有了知识概率分布的定义之后，根据信息论可以定义知识的熵与条件熵的概念. 知识 P 的熵 $H(P)$ 定义为

$$H(P)=-\sum_{i=1}^{n}p(X_i)\log p(X_i).$$

知识 Q 相对于知识 P 的条件熵 $H(Q|P)$ 定义为

$$H(Q|P)=-\sum_{i=1}^{n}p(X_i)\sum_{j=1}^{m}p(Y_j|X_i)\log p(Y_j|X_i).$$

3 主要概念与运算的信息表示

在上节中，我们建立了粗糙集理论中的知识与信息熵的关系，从而使我们能够从信息的角度对粗糙集理论的主要概念与运算进行表达. 本文称此表达为粗糙集理论的信息表示. 本节给出了该理论的主要概念与运算的信息表示，并证明了知识约简在信息与代数两种不同表示下是等价的[6].

定理 3.1. 设 U 是一个论域，P,Q 是 U 上的两个等价关系族，若 $IND(P)=IND(Q)$，则 $H(P)=H(Q)$.

证明：因为 $IND(P)=IND(Q)$，所以，P,Q 在 U 的子集组成的 σ-代数上确定的概率分布相同. 故有 $H(P)=H(Q)$. □

注意：定理 3.1 的逆未必成立.

本定理说明，两个代数表示下等价的知识库具有相同的信息量.

定理 3.2. 设 U 是一个论域，P,Q 是 U 上的两个等价关系族，且 $P\subseteq Q$. 若 $H(P)=H(Q)$，则 $IND(P)=$

$IND(Q)$.

证明：因为 $P\subset Q$，所以 $IND(P)\supset IND(Q)$. 下面证明

$$IND(P)\subseteq IND(Q). \tag{1}$$

令　　$U/IND(P)=\{A_1,A_2,\ldots,A_n\}$，　　$U/IND(Q)=\{B_1,B_2,\ldots,B_m\}$.

用反证法，假设式(1)不成立，则至少存在一个 $A_{i_0}\in U/IND(P)$，对任何 $B_j\in U/IND(Q)$，都有 $A_{i_0}\not\subset B_j$，$j=1,2,\ldots,m$. 从而存在正整数 $K(2\leq K\leq m)$，使得 $A_{i_0}\cap B_j\neq\varnothing$，且

$$0<p(B_j|A_{i_0})=\frac{|A_{i_0}\cap B_j|}{|A_{i_0}|}<1,\qquad j=1,2,\ldots,K,$$

$$H(Q)=H[P+(Q-P)]=H(P)+H[(Q-P)|P].$$

因为 $H(P)=H(Q)$，所以，$H[(Q-P)|P]=0$. 即

$$H(Q|P)=0, \tag{2}$$

又因为　$H(Q|P)=-\sum_{i=1}^{n}p(A_i)\sum_{j=1}^{m}p(B_j|A_i)\log p(B_j|A_i)\geq -p(A_{i_0})\sum_{j=1}^{K}p(B_j|A_{i_0})\log p(B_j|A_{i_0})>0$，

这与式(2)矛盾！故假设不成立，结论得证. □

本定理说明，当两个知识库存在包含关系时，由知识信息量的相等可得出它们在代数表示下是等价的.

定理 3.3. 设 U 是一个论域，P 是 U 上的一个等价关系族. 一个关系 $R\in P$ 在 P 中是不必要的(多余的)，其充分必要条件为 $H(R|P-\{R\})=0$.

证明：(必要性) 设 $R\in P$ 在 P 中是不必要的，由定义知下式成立

$$IND(P-\{R\})=IND(P).$$

由定理 3.1 可知，$H(P-\{R\})=H(P)$.

因为　　$H(P)=H[(P-\{R\})+\{R\}]=H(P-\{R\})+H(R|P-\{R\})$，

所以，　　$H(R|P-\{R\})=0$.

(充分性) 设 $H(R|P-\{R\})=0$.

因为　　$H(P)=H[(P-\{R\})+\{R\}]=H(P-\{R\})+H(R|P-\{R\})$，

所以，　　$H(P-\{R\})=H(P)$.

又显然有　　$P-\{R\}\subset P$.

由定理 3.2 知，$IND(P-\{R\})=IND(P)$. 故 $R\in P$ 在 P 中是不必要的. □

本定理说明，在代数表示下，不必要的知识在知识库中没有提供新的信息；反之亦然.

推论. $R\in P$ 在 P 中是必要的充分必要条件为 $H(R|P-\{R\})>0$.

定理 3.4. 设 U 是一个论域，P 是 U 上的一个等价关系族. P 是独立的充分必要条件为：对任意 $R\in P$，都有 $H(R|P-\{R\})>0$.

证明：由独立性的定义及定理 3.3 的推论可证.

定理 3.5. 设 U 是一个论域，P 是 U 上的一个等价关系族. $Q\subseteq P$ 是 P 的一个约简的充分必要条件为下列两个条件成立：

(1) $H(Q)=H(P)$；

(2) 对任意的 $q\in Q$，有 $H(q|Q-\{q\})>0$.

证明：$Q\subseteq P$ 是 P 的一个约简的充分必要条件为

$$IND(Q)=IND(P), \tag{3}$$

且

$$Q\text{ 是独立的.} \tag{4}$$

由定理 3.2 知，式(3)成立的充分必要条件是 $H(Q)=H(P)$(因为 $Q\subset P$).

由定理 3.4 知，式(4)成立的充分必要条件是，对任意的 $q\in Q$，有 $H(q|Q-\{q\})>0$.

故结论成立. □

本定理说明,对于知识约简而言,本文所给出的信息表示与原来的代数表示完全等价.但信息表示比代数表示更加直观;而且在前者表示的基础上,能够导出高效的知识约简算法.

4 结 论

目前,关于粗糙集理论的研究与应用受到了国际人工智能界越来越多的关注.本文把定义在论域上的等价关系看作是其上的一个随机变量,从而建立了知识与信息熵之间的关系.我们从信息的角度给出了粗糙集理论中主要概念与运算的信息表达.证明了知识约简在信息与代数两种不同表示下是等价的.这些结论有助于人们深刻理解粗糙集理论的本质,而且为寻找高效的知识约简算法奠定了基础.

本文只对不考虑决策的情况进行了讨论.关于有决策的情况以及由此得到的有关知识约简的一种高效算法将另文介绍.

致谢 衷心感谢审稿者提出的宝贵建议!

参考文献

1 Pawlak Z. Rough Sets: Theoretical Aspects of Reasoning about Data. Amsterdam: Kluwer Academic Publishers, 1991. 6~42

2 Ziarko W. Introduction to the special issue on rough sets and knowledge discovery. International Journal of Computational Intelligence, 1995,11(2):223~226

3 苗夺谦,王珏.基于粗糙集的多变量决策树构造方法.软件学报,1997,8(6):425~431
(Miao Duo-qian, Wang Jue. Rough sets based approach for multivariate decision tree construction. Journal of Software, 1997,8(6):425~431)

4 王珏,苗夺谦,周育健.关于 Rough Set 理论与应用的综述.模式识别与人工智能,1996,9(4):337~344
(Wang Jue, Miao Duo-qian, Zhou Yu-jian. Rough set theory and its application: a survey. Chinese Journal of Pattern Recognition and Artificial Intelligence, 1996,9(4):337~344)

5 王珏,袁小红,石纯一等.关于知识表示的讨论.计算机学报,1995,18(3):212~224
(Wang Jue, Yuan Xiao-hong, Shi Chun-yi *et al*. A discussion on knowledge representation. Chinese Journal of Computers, 1995,18(3):212~224)

6 苗夺谦.Rough Set 理论及其在机器学习中的应用研究[博士学位论文].中国科学院自动化研究所,1997
(Miao Duo-qian. Rough sets and its application in machine learning[Ph. D. Thesis]. Institute of Automation, The Chinese Academy of Sciences, 1997)

An Information Representation of the Concepts and Operations in Rough Set Theory

MIAO Duo-qian[1,2] WANG Jue[2]

[1](*Department of Mathematics Shanxi University Taiyuan* 030006)

[2](*Institute of Automation The Chinese Academy of Sciences Beijing* 100080)

Abstract Rough set theory proposes a formal definition of knowledge and provides a series of tools to deal with knowledge. However, in the algebraic representation of this theory, it is difficult to understand the essence of rough set theory, and efficient algorithm of knowledge reduction has not been found. In this paper, a relationship between knowledge and information is set up, and then based on the relationship an information representation of the concepts and operations about rough set theory is given. Finally, the equivalence properties between information representation and algebraic representation of knowledge reduction are proved. These conclusions are helpful for people to understand the essence of rough set theory and essential to seek new efficient algorithm of knowledge reduction.

Key words Rough set, knowledge representation, knowledge reduction, information entropy, equivalence property.