高技术通讯 2010 年 第 20 卷 第 4 期:366 ~ 370

doi:10.3772/j.issn.1002-0470.2010.04.007

基于广义粗糙集的知识约简方法研究①

张大勇② 张兆心 李 乔

(哈尔滨工业大学计算机科学与技术学院 哈尔滨 150001)

摘 要 针对粗糙集理论在知识约简中的实际需要,提出了建立在一般二元关系基础上的广义粗糙集知识约简方法。首先证明了广义粗糙集是经典粗糙集的一般性推广,而经典粗糙集是广义粗糙集的特例;然后以一般二元关系为分类基础,给出一般关系决策系统中的知识约简判定定理和辨识矩阵;最后根据实例提取最小的属性集,验证了该方法的实用性。该方法摆脱了二元等价关系对经典粗糙集的困扰,既保证了粗糙集理论在知识发现研究中的理论优势,又拓展了粗糙集理论在实际应用中的适用范围,具有较强的实用性。

关键词 知识约简,广义粗糙集,二元关系,判定定理

0 引 言

知识发现(knowledge discovery in databases, KDD)是一门新兴的交叉性学科,其主要目的就是从客观复杂、多样性的信息中抽取和精化能够被理解和使用的知识[1]。而知识约简(knowledge reduction)是知识发现的核心,是指在不影响知识表达能力的条件下,通过消除冗余知识获得知识库的简洁表达的方法。知识约简可分为属性约简和属性值约简,由于属性值约简相对较为简单,属性约简则困难得多,甚至到现在也还没有完全解决,仍然是研究重点,因此多数情况下将知识约简理解为属性约简[2]。经过 20 多年的发展,知识发现的研究范围已经涉及经济、工业、农业、军事、社会、商业等领域,知识发现的方法综合了数据库、人工智能、统计学、模式识别、机器学习、数据分析等领域的研究成果,发展成包括决策树法(decision tree)[3,4]、规则归纳方法(rules induction)[5]、神经网络方法(neural networks)[6,7]、统计方法(statistical approaches)[8]、K-最近邻技术(K-nearest neighbor, K-NN)[9]、模糊集理论(fuzzy sets theory)[10]等在内的许多方法。然而,上述方法需要预先给出某些特征和相关知识,对于处理不确定、不精确、不完全信息和知识,往往显得力不从心。

近些年来发展起来的粗糙集理论(rough set)通过属性约简算法,为拓展知识发现中对不确定性或不精确性信息的处理,提供了一条有效的解决途径,并在知识规则归纳、分类和聚类方面研究中取得了比较突出的成果[11-14]。但是随着粗糙集理论的发展,一些基础性的问题始终困扰着粗糙集理论在实际应用中的适用范围[15],如经典粗糙集是建立在等价关系基础之上的,只适合处理完备的信息系统,而对于现实中广泛存在的不完备信息系统是无法直接处理的。为此,本文将广义粗糙集理论应用于知识发现研究中,研究一般关系决策系统中的知识约简方法,理论上是经典粗糙集理论的推广,实际应用中是对经典粗糙集做更为一般性的转化,既保证了原有粗糙集理论在知识发现研究中的理论优势,又拓展了其在知识发现实际应用中的范围。

1 经典粗糙集理论基础

粗糙集理论是波兰数学家 Pawlak 于 1982 年提出的一种数据分析理论。粗糙集理论作为处理复杂系统的一个有效方法,其主要思想就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。应用该理论处理不确定性问题的最大优势是不需提供问题所需处理的数据集合之外的任何先验信息[16,17]。

① 全国高校博士点基金(No. 20070213008),教育部社科基金(No.07JC630027)和中国博士后基金(No. 20080440856)资助项目。
② 男,1975 年生,博士后,讲师;研究方向:复杂系统,人工智能;联系人,E-mial: yongzhhit@163.com
(收稿日期:2009-02-24)

经典粗糙集——Pawlak 集是建立在分类机制的基础上的,其将分类解释为特定空间上的等价关系,而且这种等价关系构成了对整个空间的划分,将知识理解为对数据的划分,并将每一个被划分的集合称为概念。对 Pawlak 集的定义如下:

设 $U=\{x_1,x_2,x_3,\cdots,x_n\}$ 是非空有限论域,$R\subseteq U\times U$ 是 U 上的二元等价关系,也称不可分辨关系,则序对 (U,R) 为 Pawlak 近似空间。若 $(x,y)\subseteq R$,$\forall(x,y)\in U\times U$,则称 x 和 y 关于 R 是不可区分的,等价关系 R 可以生成一个划分 U/R,并且 U 上的划分与 U 上的 R 间可以建立一一对应。若将 U 中的集合表示知识,则 (U,R) 成为知识库,U/R 表示基本概念或知识模块。

对于任意的 $X\subseteq U$ 不一定能用知识库中的知识来精确地描述,即 X 可能为不可定义集,则可选用 X 关于近似空间 (U,R) 的一对下近似 $\underline{R'}(X)$ 和上近似 $\overline{R'}(X)$ 表述:

$$\begin{aligned}\underline{R'}(X)&=\{x:[x]_R\subseteq X\}\\&=\bigcup\{[x]_R:[x]_R\subseteq X\}\\\overline{R'}(X)&=\{x:[x]_R\cap X\neq\varnothing\}\\&=\bigcup\{[x]_R:[x]_R\cap X\neq\varnothing\}\end{aligned}\qquad(1)$$

其中 $[x]_R$ 是 x 的 R 等价类,可以记为 $[x]_R=\{y\in U:(x,y)\in\mathbf{R}\}$;$\underline{R'}$ 和 $\overline{R'}$ 分别为 Pawlak 下近似算子与上近似算子;系统 $(2^U,\cap,\cup,\sim,\underline{R'},\overline{R'})$ 是一个 Pawlak 粗糙集代数系统,如图 1 所示。

从定义中可知,X 的下近似是包含在 X 中的最大可定义集合,而 X 的上近似是包含 X 的最小可定义集合。如果 X 的下近似与上近似相等,则 X 是可定义的;否则 X 是粗糙的。

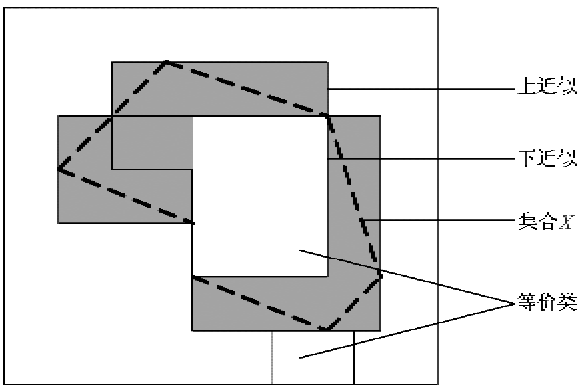

图 1　粗糙近似图

2　广义粗糙集的提出与定义

经典的粗糙集是定义在等价关系基础上的,只适用于处理离散型变量,对于连续型数据需要做离散化处理,把数值型属性转化为符号型属性。然而转换过程中不可避免地会带来信息损失,因此计算处理的结果很大程度上取决于离散化效果。为了解决这一问题,需要对 Pawlak 粗糙集做更为一般性的转化,而广义粗糙集理论的提出能够满足 Pawlak 粗糙集的推广性研究[18,19]。

定义 1　设 U 是有限非空的论域,$R\subseteq U\times U$ 为 U 上一个任意的二元关系,称 (U,R) 为广义近似空间,对于任意 $X\subseteq U$,X 关于近似空间 (U,R) 的下近似 $\underline{R}(X)$ 和上近似 $\overline{R}(X)$ 分别定义为

$$\begin{aligned}\underline{R}(X)&=\{x\in U:R_S(x)\subseteq X\}\\\overline{R}(X)&=\{x\in U:R_S(x)\cap X\neq\varnothing\}\end{aligned}\qquad(2)$$

X 关于近似空间 (U,R) 的正域 $POS_R(X)$,负域 $NEG_R(X)$ 和边界 $BN_R(X)$ 分别定义为

$$POS_R(X)=\overline{R}(X)$$

$$NEG_R(X)=\sim\overline{R}(X)=\{x\in U:R_S(x)\cap X=\varnothing\}$$

$$BN_R(X)=\overline{R}(X)-\underline{R}(X)$$

当 $\underline{R}(X)=\overline{R}(X)$ 时,称 X 关于近似空间 (U,R) 是可定义的,否则称 X 关于近似空间 (U,R) 是粗糙的,称系统 $(2^U,\cap,\cup,\sim,\underline{R},\overline{R})$ 为广义粗糙集代数。

可以证明,当二元关系 R 为等价关系时,广义粗糙集就变成为 Pawlak 粗糙集。按公式(1)、(2),当且仅当二元关系 R 为等价关系时,可得 $\underline{R}(X)=\underline{R'}(X)$ 且 $\overline{R}(X)=\overline{R'}(X)$。因此,Pawlak 粗糙集是广义粗糙集的特例,而广义粗糙集是 Pawlak 粗糙集的一般性推广。

3　关系决策系统的知识约简

本文以一般二元关系为分类基础,给出一般关系决策系统中的知识约简判定方法,即应用广义粗糙集理论在关系决策系统属性集中寻找一个最小的属性集,并确保最小属性集所确定的分类知识与用全体属性集所确定的分类知识相同。

3.1　关系决策系统的定义

定义 2　设 U 是有限非空的论域,$R=\{R_1,R_2,R_3,\cdots,R_n\}$ 是 U 上一簇二元关系,R、D 分

别代表条件属性和决策属性,则(U,R,D)是关系决策系统,$(IntR)_S(x)$为领域簇,且$(IntR)_S(x)=\bigcap_{i=1}^{n}(R_i)_S(x)$。

根据定义1和定义2,对于关系决策系统(U,R,D)可以作如下表述。

(1)任意子集$X\subseteq U$,X的下近似可以表示为:

$$\underline{R}(x)=\bigcup\{(IntR)_S(x):(IntR)_S(x)\neq\varnothing,(IntR)_S(x)\subseteq X\}$$

或

$$\underline{R}'(x)=\{x\in U:(IntR)_S(x)\neq\varnothing,(IntR)_S(x)\subseteq X\}$$

(2) D相对于R的正域为

$$POS_R(D)=\bigcup_{X\in U/D}\underline{R}(X)\text{ 或 }POS'_R(D)=\bigcup_{X\in U/D}\underline{R}'(X)$$

(3) D相对于R的空域和负域分别为

$$Nul_R(D)=\{x\in U:(IntR)_S(x)=\varnothing\}$$

$$Neg_R(D)=U-POS_R(D)-Nul_R(D)$$

(4)设$P\subseteq R$,则X相对于P的下近似为

$$\underline{P}(X)=\bigcup\{(IntP)_s(x):(IntP)_s(x)\neq\varnothing,(IntP)_s(x)\subseteq X\}$$

或

$$\underline{P}'(X)=\bigcup\{x\in\underline{R}'(X):(IntP)_s(x)\neq\varnothing,(IntP)_s(x)\subseteq X\}$$

(5) D相对于P的正域为

$$POS_P(D)=\bigcup_{X\in U/D}\underline{P}(\underline{R}(X))$$

或

$$POS'_P(D)=\bigcup_{X\in U/D}\underline{P}'(\underline{R}(X))$$

上述结论中$\underline{R}(X)$、$\underline{P}(X)$和$POS_P(D)$分别为$\underline{R}'(x)$、$\underline{P}'(X)$和$POS'_P(D)$的支撑域。而决策属性D对条件R的依赖程度为

$$k=\gamma_R(D)=\frac{|POS'_R(D)|}{|U|}$$

根据$POS'_R(D)$、$POS'_P(D)$、$\underline{R}(X)$和$\underline{P}'(X)$的定义可得:定理1对于关系决策系统(U,R,D),如$P\subseteq R$,则

$$POS'_R(D)=POS'_P(D)\Leftrightarrow\underline{R}'(\underline{R}(X))=\underline{P}'(\underline{R}(X)),\ \forall X\in U/D$$

3.2 知识约简的判定方法

定义3 设(U,R,D)是关系决策系统,$R_i\in\mathbf{R}$,如果满足条件:

(1) $POS'_{\{R\}}(D)\neq POS'_{(R-\{R_i\})}(D)$

或

(2) $Nul_R(D)\neq Nul_{(R-\{R_i\})}(D)$

则称R_i相对于D是R中必要的。

对于任意一个子集$P\subseteq R$,如果P中每一个元素相对于D都是P中必要的,并且P满足下列条件:

(1) $POS'_R(D)=POS'_P(D)$

(2) $Nul_R(D)=Nul_{(P)}(D)$

则称P是R的相对于D的一个等价子集,即P是R的相对于D的一个约简。R中所有相对于D的必要元素组成的集合称为R相对于D的核,记作$Core_D(R)$。

因此根据定义3和定理1可得知识约简的判定定理。

定理1 设(U,R,D)是关系决策系统,$P\subseteq R$,则P是R的相对于D的一个等价子集的充分必要条件是:对于任意$x,y\in U$,有

(1)如果$x\in Nul_R(D)$,则$y\notin(IntR)_S(x)\Rightarrow y\notin(IntP)_S(x)$;

(2)如果$x\in POS'_R(D)$并且$d((IntR)_S(x))=d(y)$,则$y\notin POS_R(D)\Rightarrow y\notin(IntP)_S(x)$;

(3)如果$x\in POS'_R(D)$并且$d((IntR)_S(x))\neq d(y)$,则$y\notin(IntR)_S(x)\Rightarrow y\notin(IntP)_S(x)$。

根据知识约简的判定定理可知,P是R的相对于D的一个等价子集的充分且必要条件是论域U关于P和R的空域、负域和正域对应不变。由此可得关系决策系统的辨识矩阵如下:设(U,R,D)是关系决策系统,$U=\{x_1,x_2,x_3,\cdots,x_n\}$,用$M(U,R,D)$表示一个$n\times n$的矩阵$(c_{ij})$,则$(c_{ij})$为$(U,R,D)$的辨识矩阵。对于任意$x_i,x_j\in U$,有

(1)如果$x_i\in Nul_R(D)$,则$c_{ij}=\{R\in\mathbf{R}:x_j\notin R_S(x_i)\}$;

(2)如果$x_i\in POS'_R(D)$且$d((IntR)_S(x_i))=d(x_j)$,则$c_{ij}=$

$$\begin{cases}R & x_j\in POS_R(D)\\ \{R\in\mathbf{R}:x_j\notin R_S(x_i)\} & x_j\notin POS_R(D)\end{cases}$$

(3) 如果$x_i\in POS'_R(D)$且$d((IntR)_S(x_i))\neq d(x_j)$,则$c_{ij}=\{R\in\mathbf{R}:x_j\notin R_S(x_i)\}$;

(4)如果$x_i\in Neg_R(D)$,则$c_{ij}=R$。

4 实例分析

评估证券市场交易者投资决策水平高低与信念、行为能力、规划和意图的关系,这4种要素用

(R_1, R_2, R_3, R_4) 表示,选择 10 位交易者作为试验样本。设 $U = \{x_1, x_2, x_3, \cdots, x_{10}\}$ 为试验主体集合。条件属性分别用{高,低}加以划分,根据实际收益情况确定 10 个行为主体的决策水平,可以描述如下:

$U/D = \{\{x_1, x_5, x_6\}$(很高),
$\{x_2, x_4, x_7\}$(高),$\{x_3, x_8\}$(一般),$\{x_9, x_{10}\}$(差)$\}$

根据实际记录得

$$\begin{bmatrix} R & R_1R_2R_3 & R_2R_4 & R & R & R & R & R & R & R \\ R & R_1R_2 & R & R & R & R & R_2R_4 & R_1R_2R_3 & R & R \\ R_1R_2R_3 & R & R & R & R & R & R & R_1R_3R_4 & R & R \\ R & R & R & R_2R_4 & R & R & R_1R_3 & R & R & R \\ R & R & R & R & R_1R_3R_4 & R & R & R & R & R \\ R & R & R & R & R & R_2 & R & R & R & R \\ R & R & R & R & R & R & R & R & R & R \\ R & R_1R_2R_4 & R & R & R & R & R & R & R & R \\ R & R & R & R & R & R & R & R & R & R \\ R & R & R & R & R & R_1R_3R_4 & R & R & R & R_2R_3R_4 \end{bmatrix}$$

$$f(U,R,D)(\overline{R_1}, \overline{R_2}, \overline{R_3}, \overline{R_4})$$
$$= \wedge \{\vee (c_{ij}) : i, j \leqslant 10, c_{ij} \neq \varnothing\}$$
$$= (R_1 \vee R_2 \vee R_3) \wedge (R_2 \vee R_4) \wedge (R_1 \vee R_2) \wedge (R_1 \vee R_3) \wedge (R_1 \vee R_3 \vee R_4) \wedge R_2 \wedge (R_1 \vee R_2 \vee R_4) \wedge (R_2 \vee R_3 \vee R_4)$$
$$= (R_1 \vee R_3) \wedge R_2$$
$$= (R_1 \wedge R_2) \vee (R_2 \wedge R_3)$$

于是得系统知识约简为 $\{R_1, R_2\}$ 与 $\{R_2, R_3\}$,从而可得 $Core(R) = \{R_2\}$。由此可知,虽然交易者的投资决策水平只是部分依赖于以上 4 种要素,但是在 4 个要素中,交易者行为能力对投资决策的提升有着重要的影响。

5 结 论

知识约简是知识发现中的基本问题也是研究的重点问题,虽然粗糙集理论因其不需要任何先前或额外的信息,对于处理模糊和不完全知识具有明显的优势,而被广泛地应用到知识约简中,但是由于经典粗糙集是建立在二元等价关系基础之上的,只适用于处理离散型变量,对于连续型数据需要做离散化处理,把数值型属性转化为符号型属性,然而转换过程中不可避免地会带来信息损失,造成经典粗糙集在解决实际问题中具有很大的局限性。为了解决这一问题,本文在证明广义粗糙集是经典粗糙集的一般性推广的基础上,提出了基于广义粗糙集的知识约简方法。该方法以一般二元关系为分类基础,根据一般关系决策系统中的知识约简判定定理和辨识矩阵,提取关系决策系统属性集中寻找一个最小的属性集,并结合实例验证了该方法能够确保最小属性集所确定的分类知识与用全体属性集所确定的分类知识相同。

基于广义粗糙集的知识约简方法不仅突破了经典粗糙集理论受到等价二元关系分类的限制,保证了原有粗糙集理论在知识发现研究中的理论优势,又拓展了其在知识发现实际应用中的适用范围。然而知识发现是一项复杂的系统工程,应用广义粗糙集对现实系统的知识获取和约简还有待于进一步完善,例如,如何处理系统内数据本身具有模糊性的问题,以及如何处理数据不完备和数据噪声对知识获取影响的问题,将成为今后的研究重点。

$Nul_R(D) = \{x_1, x_3, x_{10}\}$, $Neg_R(D) = \{x_7, x_9\}$

$POS_R(D) = \{x_1, x_2, x_3, x_4, x_5, x_6, x_8, x_{10}\}$

$POS'_R(D) = \{x_2, x_4, x_5, x_6, x_8\}$

而 $\gamma_R(D) = \dfrac{|POS'_R(D)|}{|U|} = 0.5$,则$(U, R, D)$的辨识矩阵为

参考文献

[1] Piatetsky-Shapiro G, Fayyad U, Smith P. From data mining to knowledge discovery: an overview. Advances in Knowledge Discovery and Data Mining, Menlo Park, California: AAAI/MIT Press, 1996. 1-35

[2] 黄兵. 基于粗糙集的不完备信息系统知识获取理论与方法:[博士学位论文]. 南京:南京理工大学自动化学

院,2004.7-8
[3] Quinlan J R. Induction of decision trees. *Machine Learning*, 1986,(1):81-106
[4] Brodley C E, Utgoff P E. Multivariate decision trees. *Machine Learning*,1995, 19 (1):45-77
[5] Agrawal R, Shafer J C. Parallel mining of association rules. *IEEE Transaction on Knowledge and Data Engineering*,1996, 8(6):962-969
[6] Lu H J, Setiono R, Liu H. Effective data mining using neural networks. *IEEE Transaction on Knowledge and Data Engineering*,1996,8(6):957-961
[7] Fu L M. A neural-network model for learning domain rules based on its activation function characteristics. *IEEE Transaction on Neural Networks*,1998,9(5):787-795
[8] D Heckerman. Bayesian networks for data mining. *Data Mining and Knowledge Discovery*, 1997, 1(1): 79-119
[9] Fisher D. Optimization and simplication of hierarchical clustering. In: Proceedings of the 1st International Conference on Knowledge Discovering and Data Mining (KDD'95), Montreal, Canada: AAAI Press,1995.118-123
[10] Ying-hua L, George A C Ⅲ, Stephen V C, et al. Nonlinear system input structure identification: two stage fuzzy curves and surfaces. *IEEE Transactions on Systems, Man, and Cybernetics*,1998,28(5):678-684
[11] Lingras P J, Yao Y Y. Data mining using extensions of the rough set model. *Journal of the American Society for Information Sciences*,1998,49:415-422
[12] 管涛,冯博琴.模糊目标信息系统上的知识约简方法.软件学报,2004,15(10):1470-1478
[13] 王艳丽,郭景杰,傅恒志.钛合金冷坩埚定向凝固过程温度场数值模拟.哈尔滨工业大学学报,2008,40(11):1808-1810
[14] 张雪峰, 田晓东, 张庆灵.基于粗糙集理论和层次分析的数据约简. 东北大学学报:自然科学版,2008,29(1):21-32,42
[15] Beynon M J, Peel M J. Variable precision rough set theory and data discretization: an application to corporate failure prediction. *Omega*, 2001, 29(6): 561-576
[16] Pawlak Z, Skowron A. Rudiments of rough sets. *Information Sciences*, 2007,177(1):3-27
[17] Pawlak Z, Skowron A. Rough sets: some extensions. *Information Sciences*,2007,177 (1):28-40
[18] Zhu W. Generalized rough sets based on relations. *Information Sciences*, 2007,177(22):4997-5011
[19] Yao Y Y. Relational interpretations of neighborhood operators and rough set approximation operators. *Information Sciences*, 1998,(1-4):239-259

A method of knowledge reduction based on generalized rough sets

Zhang Dayong, Zhang Zhaoxin, Li Qiao
(School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001)

Abstract

To meet the practical demand of the rough set theory in knowledge reduction, the paper establishes a method of knowledge reduction based on generalized rough sets. Firstly, the paper proves that an important value of generalized rough sets is based on arbitrary binary relations on a universal set, which may extend applications of the classical rough set theory, and then presents the decision theorem of knowledge reduction and discernible matrix based on some general binary relations. Finally, the validity of the method is verified by the application of a practical knowledge system, which can accurately abstract a minimal attribute set. The major contributions of this paper are the method of knowledge reduction based on generalized rough sets may overcome the shortage of the classical rough set theory, and extend many practical applications in various areas.

Key words: knowledge reduction, generalized rough sets, binary relations, decision theorem