无限论域上粗糙集的拓扑结构

乔全喜[1,2] 秦克云[1]

(西南交通大学数学系 成都 610031)[1] (河南理工大学数学与信息科学学院 焦作 454000)[2]

摘 要 讨论了当论域不限制是有限集时满足自反、传递关系的广义近似空间中的近似算子的拓扑结构;证明了论域上满足自反、传递关系的集合与其上所有的拓扑的集合是一一对应的;指出了该拓扑空间的拓扑基。

关键词 粗糙集,拓扑,近似算子,自反关系,传递关系

Topological Structure of Rough Sets in Infinite Universes

QIAO Quan-xi[1,2] QIN Ke-yun[1]

(Department of Mathematics, Southwest Jiaotong University, Chengdu 610031, China)[1]

(School of Mathematics & Information Science, Henan Polytechnic University, Jiaozuo 454000, China)[2]

Abstract We investigated the topological structure of approximation operators satisfied reflexive and transitive relations on the universe which is not restricted to be finit. It was proved that there is a one-to-one correspondence between the set of all reflexive and transitive relations and the set of all topologies. It gave the base of topological space.

Keywords Rough set, Topology, Approximation operator, Reflexive relation, Transitive relation

波兰数学家 Pawlak 于 1982 年[1]提出的粗糙集理论,是继 Zadeh 的模糊集理论之后处理不确定性概念的又一新的数学工具[1,2]。粗糙集概念刚提出来就得到了包括 Zadeh 在内的许多学者的肯定和高度重视。Zadeh 还把它列入他新提倡的软计算的基础理论之中。经过二十多年的发展,粗糙集理论在数据分析和知识发现等领域已经有许多的应用。在理论研究方面,人们研究了它的代数结构、格序结构、拓扑结构。我们认为 Pawlak 正是受了拓扑理论的启发才引入粗糙集这一概念的,因此研究不同粗糙集模型的拓扑结构有利于我们利用比较成熟的拓扑理论来解决粗糙集中的问题。我们知道,确定性概念和不确定性概念的本质区别是二者有无边界。论域中的元素对于确定性的概念是黑白分明、非此即彼的,不存在边界;而不确定性概念有一个灰色地带,我们称之为边界,该区域内的元素对于所讨论的概念来说既属于又不属于。拓扑理论告诉我们,一个拓扑空间(U,T)中的任何一个集合X都对应着它的内部$i(X)$和闭包$c(X)$,我们称内部与闭包之间的区域为X的边界。据此,Pawlak 由论域U上的等价关系定义了一个近似空间$A=(U,R)$,限定集合的下近似集与上近似集。陈德刚等[3]讨论了 Pawlak 粗糙集模型的拓扑性质;Yao[4]研究了当论域为有限集时满足自反、传递关系的粗糙集模型的拓扑结构;对于模糊粗糙集以及模糊粗糙集构成的拓扑空间也有大量文献论及[5-10]。本文讨论当论域不限制是有限集时满足自反、传递关系的粗糙集模型的拓扑结构,并给出它的拓扑基。

1 Pawlak 粗糙集模型的拓扑结构

定义 1[11]设U是一个有限集,称为论域。R是U上的一个等价关系,称$A=(U,R)$为 Pawlak 近似空间,记为"PAS"。$[x]_R$表示$x\in U$所在的等价类。对于任意$X\subseteq U$,X关于近似空间(U,R)的下近似$\underline{R}(X)$与上近似$\overline{R}(X)$分别定义为

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

$$\overline{R}(X)=\{x\in U|[x]_R\cap X\neq\varnothing\}$$

显然

$$\underline{R}(X)=\bigcup\{Y\in U/R|Y\subseteq X\}$$

$$\overline{R}(X)=\bigcup\{Y\in U/R|Y\cap X\neq\varnothing\}$$

如果$\underline{R}(X)\neq\overline{R}(X)$,那么称$X$为一个 Pawlak 粗糙集,记作$(\underline{R}(X),\overline{R}(X))$。

在 PAS 中,将U/R中元素的有限并称作A的合成集。所有合成集所组成的集族记为$Com(A)$,即

$$Com(A)=\{X\subseteq U|\forall x\in X,[x]_R\subseteq X\}$$

定义 2 设$A=(U,R)$是一个 PAS 空间,对于任意$X\subseteq U$,X关于近似空间(U,R)的下近似$\underline{R}(X)$与上近似$\overline{R}(X)$分别定义为

$$\underline{R}(X)=\bigcup\{G\subseteq X;G\in Com(A)\}$$

$$\overline{R}(X)=\bigcap\{G\supseteq X;G\in Com(A)\}$$

在 PAS 中,将$[x]_R$称作"砖块"。$Com(A)$中的元素是由砖块粘成的,将粘成后的砖块称为"方块"。那么,从定义 2 可以看出,下近似$\underline{R}(X)$是含于X的"方块"之并或者说含于X的最大"方块";上近似$\overline{R}(X)$是包含X的"方块"之交或者说是包含X的最小"方块"。

定理 1[1] 在 PAS 中,定义 1 与定义 2 中的上、下近似是等价的。

定理 2[1] 设$A=(U,R)$为 Pawlak 近似空间,则$(U,Com(A))$是一个拓扑空间,$B=\{[X]_R|x\in U\}$是$Com(A)$的一个

到稿日期:2010-11-18 返修日期:2011-02-26 本文受国家自然科学基金项目(60875034)资助。

乔全喜 博士生,主要研究领域为智能控制及应用,E-mail:qiaoqx@hpu.edu.cn;秦克云 博士,教授,主要研究领域为代数逻辑与智能信息处理。

基，并且拓扑 $Com(A)$ 中的每个集合既是开集又是闭集。

由定理 2 可知，拓扑空间 $(U,Com(A))$ 中的任意集合 X 的内部 $i(X)$ 就是 X 的下近似 $\underline{R}(X)$；其闭包 $C(X)$ 就是 X 的上近似 $\overline{R}(X)$；X 的边界就是粗糙集 $(\underline{R}(X),\overline{R}(X))$ 的边界。

定理 3[3] 设 (U,T) 是一个有开划分的拓扑空间，且该划分是它的一个基，则存在一个等价关系 R，使 $(U,T)=(U,Com(A))$。

定理 4 设 Γ 是 U 上所有等价关系的集合，Σ 是 U 上所有具有开划分的拓扑空间的集合，则 Γ 与 Σ 是同势的。

证明：由定理 2 和定理 3 可得。

2 广义粗糙集模型和拓扑空间

本小节试图将上一小节的结论推广至广义粗糙集模型。

定义 3[11] 设 U 是一个集合，称为论域。R 是 U 上的一个二元关系，称 $A=(U,R)$ 为广义近似空间，记为“GAS”。对于 $\forall x\in U$ 而言，$R_s(x)=\{y\in U|(x,y)\in R\}$ 称为 x 关于 R 的右邻域；$R_p(x)=\{y\in U|(y,x)\in R\}$ 称为 x 关于 R 的左邻域。对于任意 $X\subseteq U$，X 关于 A 的上、下近似可定义为

$$\underline{R}(X)=\{x\in U|R_s(x)\subseteq X\}$$

$$\overline{R}(X)=\{x\in U|R_s(x)\cap X\neq\varnothing\}$$

显然，若 R 为等价关系，则 $R_s(x)=[x]_R$，我们将 $R_s(x)$ 称为 GAS 的“砖块”。故 GAS 中的上、下近似算子是 PAS 中上、下近似算子概念的推广。

定义 4 设 $A=(U,R)$ 是一个广义近似空间，$X\subseteq U$，如果 X 满足

$$\forall x\in X,R_s(x)\subseteq X$$

则称 X 是右合成集。论域 U 上所有右合成集所组成的集族记为 $\mathcal{T}_R$，即

$$\mathcal{T}_R=\{X\subseteq U|\ \forall x\in X,R_s(x)\subseteq X\}$$

同理可定义左合成集族 ${}_R\mathcal{T}$，即

$${}_R\mathcal{T}=\{X\subseteq U|\ \forall x\in X,R_p(x)\subseteq X\}$$

$\mathcal{T}_R$、${}_R\mathcal{T}$ 是由“砖块”粘成的“大块”的集合。

注意：与 PAS 中不同的是，在 GAS 中不是所有的“砖块”都可以粘成“大块”的。

定理 5 设 $A=(U,R)$ 是一个广义近似空间，则 $\mathcal{T}_R$ 是 U 上的一个拓扑。

证明：(1)显然 $\varnothing,U\in\mathcal{T}_R$。

(2)设 $A,B\in\mathcal{T}_R$，若 $x\in(A\cap B)$，则 $x\in A$ 且 $x\in B$；从而 $R_s(x)\subseteq A,R_s(x)\subseteq B$，因此 $R_s(x)\subseteq(A\cap B)$，故 $(A\cap B)\in\mathcal{T}_R$。

(3)设 $\{A_i;i\in I\}\in\mathcal{T}_R$，由 $x\in\bigcup_{i\in I}A_i$ 可得 $\exists i_0\in I$，使得 $x\in A_{i_0}\subseteq\bigcup_{i\in I}A_i$，因此 $R_s(x)\subseteq A_{i_0}\subseteq\bigcup_{i\in I}A_i$，所以 $\bigcup_{i\in I}A_i\in\mathcal{T}_R$。

由(1)－(3)可知 $\mathcal{T}_R$ 是 U 上的一个拓扑。同理可证，${}_R\mathcal{T}$ 也是 U 上的一个拓扑。

定理 6 设 $A=(U,R)$ 是一个广义近似空间，$\mathcal{T}_R$ 与 ${}_R\mathcal{T}$ 是一对互余的拓扑。

证明：设 $X\in\mathcal{T}_R$，下面证明 $X^c\in{}_R\mathcal{T}$。$\forall z\in X^c$，假设 $R_p(z)\not\subseteq X^c$，那么有 $R_p(z)\cap X\neq\varnothing$，因此 $\exists b\in X$ 且 $b\in R_p(z)$，即 $\exists b\in X$ 且 bRz。即 $\exists b\in X,z\in R_s(b)\subseteq X$。这与 $z\in X^c$ 矛盾。所以，$X^c\in{}_R\mathcal{T}$。同理可证，若 $X\in{}_R\mathcal{T}$，则 $X^c\in\mathcal{T}_R$。因此，$\mathcal{T}_R$ 与 ${}_R\mathcal{T}$ 是一对互余的拓扑。

由定理 6 可知，当 $A=(U,R)$ 为 PAS 时，$\mathcal{T}_R={}_R\mathcal{T}=\{X\subseteq U|\ \forall x\in X,[x]_R\subseteq X\}$，即其上的每一个闭集同时也是开集。

定义 5 设 $A=(U,R)$ 是一个广义近似空间，$X\subseteq U$，X 关于 A 的上、下近似分别定义为

$$\underline{apr_R}(X)=\bigcup\{G\subseteq X;G\in\mathcal{T}_R\}$$

$$\overline{apr_R}(X)=\bigcap\{H\supseteq X;H\in{}_R\mathcal{T}\}$$

显然，$\underline{apr_R}(X)$ 是含于 X 的 $\mathcal{T}_R$“大块”之并，即含于 X 的最大开集，对拓扑空间 $\mathcal{T}_R$ 来说即是 X 的内部，即 $\underline{apr_R}(X)=i(X)$；$\overline{apr_R}(X)$ 是包含 X 的 ${}_R\mathcal{T}$“大块”之交，即包含 X 的最小闭集，对拓扑空间 $\mathcal{T}_R$ 来说即是 X 的闭包，即 $\overline{apr_R}(X)=C(X)$。显然，X 是开集当且仅当 $\underline{apr_R}(X)=X$；X 是闭集当且仅当 $\overline{apr_R}(X)=X$。

不难证明，当 $A=(U,R)$ 为 PAS 时，定义 3 与定义 5 是等价的。而当 $A=(U,R)$ 是 GAS 时，定义 3 与定义 5 是不等价的。

例 1 设 $U=\{a,b,c,d,e\}$，$R=\{(a,a),(a,e),(b,c),(b,d),(c,e),(d,a),(d,e),(e,e)\}$，则 $R_s(a)=R_s(d)=\{a,e\}$，$R_s(b)=\{c,d\}$，$R_s(c)=R_s(e)=\{e\}$，因此

$$\mathcal{T}_R=\{U,\varnothing,\{e\},\{a,e\},\{a,d,e\},\{a,c,d,e\}\}$$

$${}_R\mathcal{T}=\{U,\varnothing,\{b\},\{b,c\},\{b,c,d\},\{a,b,c,d\}\}$$

设 $X=\{b,c\}$，则 $\underline{apr_R}(X)=\varnothing,\overline{apr_R}(X)=\{b,c\}$，而 $\underline{R}(X)=\varnothing,\overline{R}(X)=\{b\}$。

引理 1 设 $A=(U,R)$ 是一个广义近似空间，如果二元关系 R 满足传递性，那么对 $\forall x\in U$，有 $R_s(x)\in\mathcal{T}_R$。

证明：$\forall y\in R_s(x)$，有 $(x,y)\in R$；若 $z\in R_s(y)$，则 $(y,z)\in R$。由 R 的传递性知，$(x,z)\in R$，因此 $z\in R_s(x)$。所以，$R_s(y)\subseteq R_s(x)$，这就证明了 $R_s(x)\in\mathcal{T}_R$。

同理可证，当二元关系 R 满足传递性时，$\forall x\in U$，有 $R_p(x)\subseteq{}_R\mathcal{T}$。

推论 1 当二元关系 R 满足传递性时，$\{R_s(x);x\in U\}$ 是拓扑 $\mathcal{T}_R$ 的一个基。

定理 7 当 R 满足自反性与传递性时，定义 3 与定义 5 是等价的。

证明：首先，由 R 的自反性知 $\underline{R}(X)\subseteq X$。其次，$\underline{R}(X)$ 是含于 X 的开集。这是因为 $\forall x\in U$，由自反性有 $x\in R_s(x)$，所以 $\{x\in U;R_s(x)\subseteq X\}\subseteq\bigcup\{R_s(x);R_s(x)\subseteq X\}$。另一方面，$\forall y\in\bigcup\{R_s(x);R_s(x)\subseteq X\}$，$\exists z,R_s(z)\subseteq X$，使得 $y\in R_s(x)$。又由 R 的传递性有 $R_s(y)\subseteq X$；所以，$y\in\{x;R_s(x)\subseteq X\}$；因此

$$\bigcup\{R_s(x);R_s(x)\subseteq X\}\subseteq\{x\in U;R_s(x)\subseteq X\}$$

所以，$\underline{R}(x)=\{x\in U;R_s(x)\subseteq X\}=\bigcup\{R_s(x);R_s(x)\subseteq X\}$；$\underline{R}(x)$ 是含于 X 的开集。

再次，$\forall G\subseteq X,G\in\mathcal{T}_R$ 有 $\forall x\in G,R_s(x)\subseteq X$，因此 $x\in\underline{R}(X)$，所以 $G\subseteq\underline{R}(X)$。综上所述，$\underline{R}(X)$ 是含于 X 的最大开集。即

$$\underline{R}(X)=\underline{apr_R}(X)=\bigcup\{G\subseteq X;G\in\mathcal{T}_R\}$$

同理可证 $\overline{R}(X)=\overline{apr_R}(X)=\bigcap\{H\supseteq X;H\in{}_R\mathcal{T}\}$。因此，定义 3 与定义 5 是等价的。

由定理 7 知道，对于论域 U 上任意一个满足自反、传递的二元关系 R，一定存在一个拓扑空间 $(U,\mathcal{T}_R)$，且该拓扑空间中的内部算子正好就是下近似算子，其闭包算子正好就是上近似算子。

定理 8 设 (U,T) 是任意一个拓扑空间，则存在一个 U 上的自反和传递的二元关系 R，满足 $(U,T)=(U,\mathcal{T}_R)$。

证明：在 U 上定义二元关系 R：$xRy\Leftrightarrow x\in C(\{y\})$，显然 R 是自反的。下面证明其传递性。

如果 xRy, yRz，那么 $x\in C(\{y\}), y\in C(\{z\})$，所以 $\{y\}\subseteq C(\{z\})$。由于 $C(\{y\})$ 是包含集合 $\{y\}$ 的最小闭集，因此 $C(\{y\})\subseteq C(\{z\})$，故 $x\in C(\{z\})$，即 xRz。这就证明了 R 是传递关系。

综上所述，当 U 不限制是有限集时，U 上的所有拓扑的集合与 U 上所有满足自反、传递关系的集合之间是一一对应的。

结束语 本文对加拿大学者 Y. Y. Yao 所做的工作进行了改进与扩展。Yao 研究了当论域为有限集时满足自反、传递关系的粗糙集模型的拓扑结构。本文则讨论了当论域不限制是有限集时满足自反、传递关系的广义粗糙集空间中的上、下近似算子的拓扑结构；证明了 U 上满足自反、传递关系的集合与 U 上所有的拓扑的集合是同势的，并指出了该拓扑空间的基为 $\{R_s(x); x\in U\}$。对于满足自反、对称关系的粗糙集模型以及模糊粗糙集模型的拓扑结构，我们将另文讨论。

参 考 文 献

[1] Pawlak Z. Rough Sets[J]. International Joural of Computer and Information Science, 1982, 11: 341-356

[2] Pawlak Z. Rough sets: Theoretical Aspects of Reasoning About Data[M]. Boston: Kluwer Academic Publishers, 1991

[3] 陈德刚，张文修. 粗糙集和拓扑空间[J]. 西安交通大学学报，2001(12)

[4] Yao Y Y. Two views of the theory of rough sets in finite universes[J]. International Journal of Approximate Reasoning, 1996, 15(4): 291-317

[5] Qin K, Pei Z. On the topological properties of fuzzy rough sets [J]. Fuzzy Sets and Systems, 2005, 151(3): 601-613

[6] Morsi N N, Yakout M M. Axiomatics for fuzzy rough sets[J]. Fuzzy Sets and Systems, 1998, 100(1-3): 327-342

[7] Lashin E F, Kozae A M, Abo K A, et al. Rough set theory for topological spaces [J]. International Journal of Approximate Reasoning, 2005, 40(1/2): 35-43

[8] Kuncheva L I. Fuzzy rough sets: Application to feature selection [J]. Fuzzy Sets and Systems, 1992, 51(2): 147-153

[9] 秦克云，乔全喜. 粗糙集的拓扑结构[J]. 计算机科学，2007(12): 161-162

[10] 秦克云，裴峥，杜卫锋. 粗糙近似算子的拓扑性质[J]. 系统工程学报，2006(1): 81-85

[11] 张文修，等. 粗糙集理论与方法[M]. 北京：科学出版社，2001

(上接第 201 页)

从图 2 可以看出，大多数表单都达到了较为理想的拟合效果（$IS\_DWI<0.2$ 或 $IS\_DWI>0.8$），少部分表单以相对粗糙的概率对表单做出了正确识别（$0.2<IS\_DWI<0.5$ 或 $0.5<IS\_DWI<0.8$）。这说明：(1)本文选取的区分指标是合理的；(2)本文提出的机器学习方法可以根据这些特征指标对表单进行准确分类，特别是在简单深网入口与搜索引擎的区分上，达到了令人满意的效果。

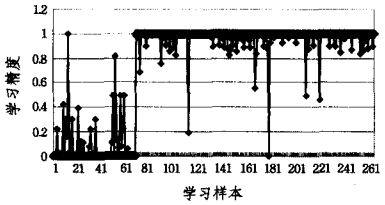

图 2 学习结果图

但从图中也可看出，有极个别表单被错误分类，通过对这些表单进行分析，发现它们都采用了较为特殊的设计，如在 amazon 网站中，允许用户在一个 textarea 中输入关于图书的搜索信息，而这与论坛表单的特征非常类似，致使系统做出了错误判断。如何对这部分表单进行识别，进一步完善该算法的动态适应性，还需要做进一步研究。

结束语 不同于以往的基于规则的判定方法，本文从抽取深网入口与非深网入口的可区分特征入手，采用机器学习的思路，提出了一种利用神经网络进行深网入口自动识别的方法，实验证明了该方法的有效性。在下一步研究中，我们拟结合考虑表单及所在页面的语义信息，以进一步提高该识别方法的准确率。

参 考 文 献

[1] Ghanem T M, Aref W G. Databases Deepen the Web [J]. IEEE Computer, 2004, 73(1): 116-117

[2] Bergman MK. The deep Web: Surfacing hidden value. Technical Report, BrightPlanet[EB/OL]. 2001. http:// www. brightplanet. com/pdf/deepwebwhitepaper. pdf

[3] 刘伟，孟小峰，孟卫一. Deep Web 数据集成研究综述[J]. 计算机学报，2007, 30(9): 1475-1489

[4] He B, Tao T, Chang K C C. Organizing structured Web sources by query schemas: A clustering approach [C]// Proc. of the 13th Conf. on Information and Knowlege Management. Washington: ACM Press, 2004: 22-31

[5] Wu P, Wen J R, Liu H, et al. Query selection techniques for efficient crawling of structured Web sources [C] // Proc. of the 22nd Int'l Conf. on Data Engineering. Atlanta: IEEE Computer Society, 2006: 47-56

[6] Raghavans, Garcia-Molina H. Crawling the hidden Web [C]// Proc. of the 27th Int'l Conf. on VLDB. Italy: Rome, 2001: 129-138

[7] Bergholz A, Chidlovskii B. Crawling for domain-specific hidden Web resources[C]// Proc. of the Int'l Conf. on Web Information Systems Engineering. Roma: IEEE Computer Society, 2003: 125-133

[8] Lage J P, da Silva A S, Golgher P B, et al. Automatic generation of agents for collecting hidden Web pages for data extraction [J]. Data & Knowledge Engineering, 2004, 49(2): 177-196

[9] Cope J, Craswell N, Hawking D. Automated discovery of search interfaces on the Web [C]// Proc. of 14th Conf. on Database technologies. Australian Computer Society, 2003: 181-189

[10] Barbosa L, Freire J. Combining classifiers to identify online databases[C]// Proc. of the 16th Int'l Conf. on WWW. New York: ACM Press, 2007: 431-440

[11] Wang Hui, Liu Xian-wei, Zuo Wan-li. Using classifiers to find domain specific online databases automatically[J]. Journal of Software, 2008, 19(2): 246-256

[12] Anthony M. Probabilistic Analysis of Learning in Artificial Neural Networks: The PAC model and its variants[J]. Neural Computing Surveys, 1997, 1: 1-47

[13] UIUC Web integration repository[EB/OL]. 2010. http://metaquerier. cs. uiuc. edu/ repository/

[14] 黄勤，龚海清，刘金亨，等. 基于改进的遗传神经网络入侵检测系统[J]. 重庆理工大学学报：自然 科学版，2010, 24(2): 83-86