

基于广义概念格的广义粗近似空间中规则的发现与提取^{*}

张倩生¹ 周作领² 许绍元¹ 贾保果¹ 罗 俊¹

(中山大学数学与计算科学学院¹, 中山大学岭南学院² 广州510275)

Extract Rule from Generalized Rough Approximate Space Based on Generalized Concept Lattice

ZHANG Qian-Sheng¹ ZHOU Zou-Ling² XU Shao-Yuan¹ JIA Bao-Guo¹ LUO Jun¹

(Mathematics and Scientific Computer College¹, Linnan College², Zhongshan University, Guangzhou 510275)

Abstract This paper proposes a new method of constructing generalized concept lattice and producing rules from it in the generalized rough approximate space based on generalized similar relation which is more extensive than equivalent relation. Finally, a simple algorithm is presented to extract rules based on interesting measure.

Keywords Generalized rough approximation space, Concept lattice, Generalized similarity relation, Support (confidence) degree, Interesting measure.

1. 引言

在粗集理论中, 近似空间的概念给出了一种理解与处理知识的模型, 它将知识定义为不可区分关系族集, 但是, 近似空间中对象的分类只局限于等价关系, 过分苛刻要求对象间差异, 忽略了对对象间相似性, 不利于大规模数据集上的知识发现。为此, Slowinski R^[1]提出了利用对象间的一种更普遍存在的广义相似关系来进行近似分类的广义粗近似空间的概念。知识的另一种表示是概念格, Wille R^[2]等提出了根据一个二元关系建造相应概念格的基本思想, 根据二元关系建造相应格的 Hasse 图的基本方法见文[4], 但上述研究中, 或者只考虑了概念内含(属性子集), 或者只适应于近似空间的等价划分, 以等价类作为概念外延。而本文所提出的广义粗近似空间是基于广义相似关系的, 不存在等价类, 只存在广义相似类, 因此有必要重新定义其相应格和研究这种新格上的规则提取方法。

2. 广义粗近似空间和广义概念格

定义1 对象集 U 上的一个二元关系, 如果满足自反性, 则称之为一个广义相似关系; 具有自反性和对称性的二元关系是容差关系, 显然容差关系与等价关系都是特殊的广义相似关系。

定义2 设 R 为 U 上的一个广义相似关系, 则对 $\forall x \in U$, 其广义相似类可定义为:

$R^{-1}(x) = \{y \in U / x R y\}$; 其参考体类为 $R(x) = \{y \in U / y R x\}$ 。

定义3 已知 U 上的广义相似关系 R , 则 $\forall X \subseteq U$, 它的下, 上近似可分别定义为

$$R_-(X) = \{x \in U / R^{-1}(x) \subseteq X\}$$

$$R^+(X) = \{x \in U / R^{-1}(x) \cap X \neq \emptyset\} = \bigcup_{x \in X} R(x)$$

显然 $R_-(X) \subseteq X \subseteq R^+(X)$; 因此空间 U 可划分为正客体集 $R_-(X)$, 负客体集 $U/R^+(X)$ 和模棱两可集 $X/R_-(X)$, $R^+(X)/X$ 。

定义4 一个广义粗近似空间可看成是一个三元组 $K = (U, R, h)$, 其中 U 是一个非空对象有限全域, $R = \{R_1, R_2, \dots, R_n\}$, R_i 是 U 上的一个广义相似关系标识符, 映射 $h: P(R) \rightarrow TR$, P 为幂集符号, TR 是 U 上的所有广义相似关系的集, 即 U 上的多个广义相似关系可通过映射 h 产生 U 上一个新的广义相似关系; h 的具体定义方式不同, 相应的广义粗近似空间意义也不同。假设 $A \in P(R)$, 若 $(x, y) \in h(A) = S_A \in TR$, 则称 (x, y) 满足 A 所确定的广义相似关系 $h(A)$, 简记为 $x A y$, 否则写为 $\sim x A y$ 。

定义5 若映射 h 满足: 对 $\forall A, B \in P(R)$, $A \subseteq B$, $h(B) \subseteq h(A)$, 则称 h 是单调的; 若 h 满足: $\forall A, B \in P(R)$, $h(A \cup B) = h(A) \cap h(B)$, 则 h 是同构的, 显然 h 同构必定单调。

定义6 在 R 为广义相似关系标识符的形式背景中, 广义粗近似空间所对应的广义概念格的元素可定义为一个二元组 (A, Y) 。

其中①内涵 $A = f(Y)$ 满足 $A \in P(R) \wedge ((\forall (x, y) \in Y) x A y)$

②外延 $Y = g(A) = \{(x, y) \in U^2 : x A y\} = h(A)$

3. 广义概念格的构造

广义概念格的构造步骤大致分以下三步:

(1) 根据定义, 构造广义粗近似空间 $K = (U, R, h)$ 所有的广义概念格二元组 (A, Y) 。

(2) 对任两个二元组 $(A, Y), (B, Y')$, 如果 $Y = Y'$,

1. 若 h 是单调的, 且 $A \subseteq B$, 则这两个二元组可合并为 (A, Y) ;

2. 若 $A \not\subseteq B$ 且 $B \not\subseteq A$; 或者 $A \subseteq B$ 但 h 非单调, 则这两个二元组可合并为 $(A/B, Y)$ 每个合并二元组在广义概念格中用一个结点表示。

(3) 对任两个二元组 (A, Y) 和 (B, Y') ; 如果 $Y \subset Y'$ 且不存在其它二元组 (C, Y'') 使得 $Y \subset Y'' \subset Y'$; 则 (A, Y) 和 (B, Y') 间有直接先驱和后继关系, 在概念格中, 它们用箭头连接。

下面我们用一个实例来说明广义概念格的构造过程:

^{*} 本文得到国家自然科学基金(10171116), 教育部博士点基金(1999055810), 广东省自然科学基金(011221)资助。张倩生 博士生, 主要研究数据挖掘与知识工程。周作领 教授, 博导。

例 下面是一个描述一些鸟类某些特征的信息表(U, R);其中 $U = \{1, 2, 3, 4, 5\}$, 1表示“鸽子”, 2表示“企鹅”, 3表示“鹰”, 4表示“麻雀”, 5表示“雕”; 广义相似关系标识符集

$R = \{a, b, c, d, e\}$; a, b, c, d 可看成条件属性, e 可看成决策属性。

表1 信息表(U, R)

U \ R	a	b	c	d	e(%)
1	0.25	0.56	0.65	0.34	48
2	0.2	0.58	0.67	0.33	35
3	0.23	0.55	0.64	0.39	45
4	0.22	0.6	0.68	0.31	40
5	0.24	0.64	0.72	0.36	44

a 表示“眼”, 取值为两眼间距与头长比值。

b 表示“脚”, 取值为脚长与身长比值

c 表示“翅”, 取值为翅长与身长比值

d 表示“嘴”, 取值为嘴长与头长比值

e 表示“DNA”, 取值为 DNA 在其染色体中含量(%)

假设 h 的定义方式如下:

(1) $h(\phi) = U^2$

(2) $h(e) = \{(x, y) \in U^2 / 0 \leq |e(x) - e(y)| \leq 5\}$ 表示亲缘关系, 是一个公差关系。

$\forall r \in R / \{e\}, h(r) = \{(x, y) \in U^2 / 0 \leq r(x) - r(y) \leq \epsilon_r\}$;

其中 $\epsilon_a = 0.02, \epsilon_b = 0.05, \epsilon_c = 0.04, \epsilon_d = 0.1$;

$h(a)$ 表示眼相像这种广义相似关系, $h(b)$ 表示脚相像这一广义相似关系, $h(c)$ 表示翅相像的广义相似关系, $h(d)$ 表示嘴相像这个广义相似关系。

(3) $\forall B \in P(R), h(B) = \{(x, y) \in U^2 : (\forall b \in B) xby\}$ 是一广义相似关系。

h 显然是同构的, 由上述广义概念格构建方法, 即可得到其概念格 Hasse 图如下。

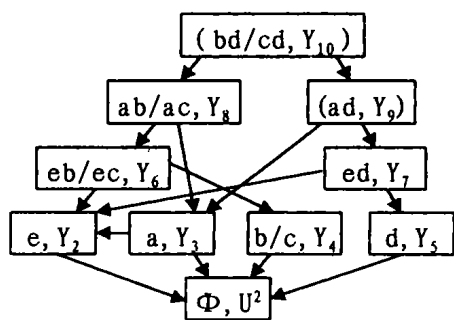


图1

其中: $Y_2 = \{(1, 1), (1, 3), (1, 5), (2, 2), (2, 4), (3, 1), (3, 3), (3, 4), (3, 5), (4, 2), (4, 3), (4, 4), (4, 5), (5, 1), (5, 3), (5, 4), (5, 5)\}$

$Y_3 = \{(1, 1), (1, 3), (1, 5), (2, 2), (3, 3), (3, 4), (4, 2), (4, 4), (5, 3), (5, 4), (5, 5)\}$

$Y_4 = \{(1, 1), (1, 3), (2, 1), (2, 2), (2, 3), (3, 3), (4, 1), (4, 2), (4, 3), (4, 4), (5, 4), (5, 5)\}$

$Y_5 = \{(1, 1), (1, 2), (1, 4), (2, 2), (2, 4), (3, 1), (3, 2), (3, 3), (3, 4), (3, 5), (4, 4), (5, 1), (5, 2), (5, 4), (5, 5)\}$

$Y_6 = \{(1, 1), (1, 3), (2, 2), (3, 3), (4, 2), (4, 3), (4, 4),$

$(5, 4), (5, 5)\} = Y_2 \cap Y_4$

$Y_7 = \{(1, 1), (2, 2), (2, 4), (3, 1), (3, 3), (3, 4), (3, 5),$

$(4, 4), (5, 1), (5, 4), (5, 5)\} = Y_2 \cap Y_5$

$Y_8 = \{(1, 1), (1, 3), (2, 2), (3, 3), (4, 2), (4, 4), (5, 4),$

$(5, 5)\} = Y_3 \cap Y_6$

$Y_9 = \{(1, 1), (2, 2), (3, 3), (3, 4), (4, 4), (5, 4), (5, 5)\}$

$= Y_3 \cap Y_7$

$Y_{10} = \{(1, 1), (2, 2), (3, 3), (4, 4), (5, 4), (5, 5)\} = Y_8$

$\cap Y_9$

4. 由广义概念格产生规则的原理

设规则中 conf 表示可信度(规则强度), $|Y|$ 是 Y 中元素个数, $N = |U^2|$, 那么由基于概率论中产生不精确规则的原则, 我们即可发现广义概念格产生规则的原理如下:

(1) 如果广义概念格某一结点是 $(A/B, Y)$ 形式, 则有规则: $\forall x, y \in U, \{xAy \Rightarrow xBy\}$

(2) 若广义概念格结点 (A, Y) 的直接后继结点中有 (B, Y') , 则有规则:

$\forall x, y \in U, \{xAy \Rightarrow xBy\}$ 和 $\{xBY \Rightarrow xAY, \text{conf} = |Y|/|Y'| \}$

(3) 若广义概念格结点 (A, Y) 的子结点有 $(A_1, Y_1), (A_2, Y_2)$ (可能还有其它子结点), 则有规则:

$\forall x, y \in U, \{xA_1y \Rightarrow xA_2y, \text{conf} = |Y_1 \cap Y_2|/|Y_1| \}$

注: 若 (A, Y) 无其它子结点, 则 $|Y_1 Y_2| = |Y|$

对图1运用上述原理即可得到它的全部规则, 表2仅列出它的部分规则及其含义。

表2

序号	规则	解释
1	$b \Rightarrow c$	脚相像的鸟类, 翅也相像; 反之亦然
2	$de \Rightarrow d$	嘴相像且有亲缘关系的鸟类, 必定嘴相像
3	$a \Rightarrow e$	眼相像的鸟类, 必有亲缘关系
4	$d \Rightarrow e \text{ conf} = 0.733$	嘴相像的鸟类, 有亲缘关系的可信度为0.733
5	$e \Rightarrow b/c \text{ conf} = 0.53$	有亲缘关系的鸟类, 脚或翅相像的可信度为0.53
6	$a \Rightarrow eb/ec \text{ conf} = 0.727$	眼相像的鸟类, 有亲缘关系且脚或翅相像的可信度为0.727
7	$a \Rightarrow ed \text{ conf} = 0.63$	眼相像的鸟类, 有亲缘关系且嘴相像的可信度为0.636
8	$d \Rightarrow ed \text{ conf} = 0.733$	嘴相像的鸟类, 则有亲缘关系且嘴相像的可信度为0.733
9	$ad \Rightarrow ed$	眼和嘴都相像的鸟类, 则必有亲缘关系且嘴相像
10	$a \Rightarrow ad \text{ conf} = 0.636$	眼相像的鸟类, 可推出眼与嘴都相像的可信度为0.636

我们很容易发现由该原理产生的规则数量过多, 显然有些规则是冗余的, 或实际利用价值不大, 为提高挖掘效率, 发掘出更简便易理解的规则, 仅有可信度是不够的, 还有必要引进规则支持度与兴趣度的概念^[6,7,10]。

假设两结点间有规则: $(B, Y) \Rightarrow (B', Y')$, 则定义其支持度 $\text{supp} = |Y \cap Y'|/|U^2| = |Y \cap Y'|/N$, 可信度 $\text{Conf} = |Y \cap Y'|/|Y|$, 兴趣度 $\text{Int} = N/|Y \cap Y'|/|Y||Y'|$; 易发现有些规则虽然可信度较高, 但实际利用价值不大, 例如: $(b/c, Y_4) \Rightarrow (d,$

Y_s), $\text{conf}=0.5 < \text{supp}(d) = |Y_s|/N$ (此时 $\text{Int} < 1$), 也即是说已知脚或翅相像的鸟类, 嘴相像的可能性为 0.5, 要比事先我们不知任何信息, 嘴相像的可能性还要小; 因此这条规则实际使用价值不大; 我们不妨考虑其使用价值更大的反面示例: $(b/c, Y_s) \Rightarrow (\bar{d}, Y_s)$, 其兴趣度 $\text{int} = N|Y_s \cap (U^2/Y_s)|/|Y_s| \cdot |U^2/Y_s| = 1.25 > 1$, 因此我们更偏向于使用这条反面规则, 或者说我们对这条反面规则更感兴趣。

另外为减少噪音对规则的影响, 有必要引进支持度门限 θ , 若结点 (B, Y) 满足 $|Y| \geq \theta|U^2|$, 则称之为频繁结点, 否则为非频繁结点; 在频繁结点间提取规则, 能有效地避免噪音。下面我们就给出一个基于兴趣度的在已构建好的广义概念格中提取无冗余规则的算法描述。

输入: 构建好的广义概念格 L , 支持度门限 minsup , 可信度门限 minconf , 最小兴趣度 minInt

输出: 包含负面属性的无冗余规则集 R

```

Begin
Step1.  $R = \emptyset, N = |U^2|$ ;
Step2. 找到格  $L$  中的所有频繁结点  $(A, Y)$ , 即所有外延项  $|Y| \geq N\text{minsup}$  的结点;
Step3. For all frequentnode  $(A, Y) \in L$  do
{if node  $(A, Y)$  is the form  $(B/C, Y)$  then  $R = R \cup \{B \Rightarrow C, \text{with } \text{supp} = |Y|/N, \text{conf} = 1, \text{int} = N/|Y|\}$ ;
else if  $(A, Y)$  has directsuccessornode  $(B, Y')$  then
{sup  $p = |Y|/N; \text{conf} = 1; \text{int} = N/|Y'|$ ;
if  $\text{Int} \geq \text{minInt}$  then
{if  $|Y|/|Y'| \geq \text{minconf}$  Then
 $R = R \cup \{A \Rightarrow B, \text{with } \text{supp}, 1, \text{int}\} \cup \{B \Rightarrow A, \text{with } \text{supp}, \text{conf} = |Y|/|Y'|, \text{int}\}$ ;
else
 $R = R \cup \{A \Rightarrow B, \text{with } \text{supp}, 1, \text{int}\}$ ;
}}
else  $(A, Y)$  have childnode  $(B_1, Y_1)$  and  $(B_2, Y_2)$  then
{sup  $p = |Y_1 \cap Y_2|/N; \text{conf} = |Y_1 \cap Y_2|/|Y_1|$ ;  $\text{Int} = N|Y_1 \cap Y_2|/|Y_1| \cdot |Y_2|$ ;
if  $\text{int} > \text{minInt}$  then
{if  $(\text{conf} \geq \text{minconf})$  then  $R = R \cup \{B_1 \Rightarrow B_2, \text{supp}, \text{conf}, \text{int}\}$ ;
else if  $\text{int} \leq 1$  then
 $\{S = |Y_1 \cap (U^2/Y_2)|/N; C = |Y_1 \cap (U^2/Y_2)|/|Y_1|; I = NC/|U^2/Y_2|$ ;
if  $s \geq \text{minsup}$  and  $c \geq \text{minconf}$  and  $I \geq \text{minInt}$  Then
 $R = R \cup \{B_1 \Rightarrow \bar{B}_2, \text{with } s, c, i\}$ ;
}}
}}
endfor
Step4. return( $R$ );
Step5. End.

```

若取 $\text{minsup} = 0.38$, $\text{minconf} = 0.72$, $\text{minInt} = 1.05$ 时, 通过该算法即可得如下:

确定性规则 $b \Rightarrow c$: with 0.48, 1, 2.08; $de \Rightarrow e$: with 0.44, 1, 1.47;

$de \Rightarrow d$: 0.44, 1, 1.66; $a \Rightarrow e$: 0.44, 1, 1.47;

(上接第112页)

实验。另外, 式中权重系数的确定问题仍未解决, 这也是我们下一步的研究方向之一。

参考文献

- 1 Sebastiani F. Machine learning in automated text categorization. ACM Computing Surveys, 2002, 34(1): 1~47
- 2 Vapnik V N. Statistical Learning Theory. New York: Wiley, 1998
- 3 Lewis D, Gale W. A sequential algorithm for training text classifiers. In: Proc. of the Seventeenth Annual Intl. ACM-SIGIR Conf. Research and Development in Information Retrieval, Springer-Verlag, 1994. 3~12

缺省(可能)性规则 $d \Rightarrow e$: 0.44, 0.73, 1.08; $d \Rightarrow de$: 0.44, 0.73, 1.67

讨论 本文提出的广义概念格, 优于文[9]所论述的一般概念格, 因为它适应于更普遍存在的广义粗近似空间, 而一般概念格只适应于基于等价类的近似空间。当然, 对基于等价划分的近似空间也可构造广义概念格, 但通过两者所提取规则的描述方式不同, 前者可描述到属性值, 而后者只能描述到属性广义相似与否。从本文可看出用广义概念格产生规则较为直观有效, 且一般只需 Y 的规模信息, 无需 Y 全信息, 这样最终可在结点外延项只记录 $|Y|$ 值, 从而大大减小存储空间, 另外映射 h 的定义方式关系到所对应广义概念格的建立及规则提取, 因此根据具体 h 的定义方式类型, 研究其所对应的广义概念格模型及规则提取是很有意义的。

参考文献

- 1 Slowinski R, et al. A generalized definition of rough approximations based on similarity. IEE Transaction on knowledge and data engineering, 2000, 12: 331~336
- 2 Mcdin D L, Gentner, et al. respect for similarity. Psychology Review, 1993, 100: 254~278
- 3 Wille R. Restructuring lattice theory; an approach based on hierarchies of concept. In: Rival I ed. dordrecht; Reidel. 1982. 445~470
- 4 Godin R, et al. Incremental concept formation algorithms based on galois lattices. Computation intelligence, 1995, 11(2): 246~267
- 5 Mollestad T. A rough set approach to data mining. [PhD Thesis]. the Norwegian univ of science and technology, norway, 1997
- 6 Klemettinen M, et al. finding interesting rules from large sets of discovered association rules. In: proc of the 3rd int'l conf. on information and knowledge management, USA. Acm press, 1994. 401~407
- 7 Savasere A, et al. mining for strong negative associations in a large database of customer transactions. In: proc of the 14th int'l conf. on data engineering. USA; IEEE Computer society press, 1998. 494~502
- 8 Oosthuizen G D. Rough sets and concept lattices, in: ziarko wped. Rough sets, and fuzzy sets and knowledge discovery. London: Springer-verlag, 1994. 24~31
- 9 Lwinski T B. Algebrac approach to rough sets. Bulletin of the polish academy of sciences; Mathematics, 1987, 35(9-10): 673~683

- 4 Tong S, Koller D. Support Vector Machine Active Learning with Applications to Text Classification. Journal of Machine Learning Research, 2001, 2: 45~66
- 5 Schohn G, Cohn D. Less is more: Active learning with support vector machines. In: Proc. of the Seventeenth Intl. Conf. on Machine Learning. 2000
- 6 Roy N, McCallum A. Toward Optimal Active Learning through Sampling Estimation of Error Reduction, ICML-2001. 441~448
- 7 Cauwenberghs G, Poggio T. Incremental and decremental support vector machine learning. Advances in Neural Information Processing Systems. 2001
- 8 程云鹏, 等. 矩阵论(第二版). 西北工业大学出版社, 1999
- 9 <http://kernel-machines.org/>