

# 基于粗糙集的间接学习

Indirect Learning Based on Rough Set Theory

米据生<sup>1,2</sup> 张文修<sup>1</sup>

(西安交通大学理学院信息与系统科学研究所 西安710049)<sup>1</sup>

(河北师范大学数学与信息科学学院 石家庄050016)<sup>2</sup>

**Abstract** In this paper, the concept of composition of approximation spaces is given, and the relationships between the approximation operators in the composition space and in the two approximation spaces are discussed. Then we discuss the relationship between the qualities of indirect learning and direct learning. Also we study the structures of the approximation operators under the relation closures. Finally, we generalize the concept of composition to the case of general fuzzy rough set model.

**Keywords** Rough set, Composition of approximation spaces, Indirect learning, Approximation operator, Fuzzy set

粗糙集理论为处理不精确、不确定数据提供了一个新的数学方法,已被广泛应用于人工智能、模式识别与智能信息处理等计算机领域。

在 Pawlak 粗糙集模型<sup>[1,2]</sup>中,等价关系起着至关重要的作用。任意给定一个概念(论域的子集),人们不一定能用知识库中的知识(等价类)来精确地描述,这时就用关于这个概念构成的集合的一对上、下近似来描述。但是,对等价关系的严厉要求限制了粗糙集理论的应用,因此许多学者从各方面对 Pawlak 粗糙集模型作了推广<sup>[3-5]</sup>。

在处理多个信息系统的复杂问题时,经常会遇到有多个近似空间的情形。如果一个近似空间的属性集正好就是另一个近似空间的对象集,那么就需要考察这两个近似空间的复合问题了。通过复合,我们可以用间接知识刻画不确定性概念。在通常情况下,间接学习的学习质量低于直接学习的学习质量。但是,当中间知识能精确刻画导师的知识,或学习者的能力区包含于中间知识的能力区时,间接学习不会影响学习质量。如何定义两个近似空间的复合?复合空间的近似算子与原来空间的近似算子之间的关系如何?本文从一个方面回答了这些问题,然后讨论一般二元关系的关系闭包所生成的近似算子的结构。最后把近似空间复合问题推广到模糊近似空间的情形。

## 1. 广义粗糙集模型

设  $U$  与  $W$  是两个非空有限集合,称为论域。 $U$  的所有经典子集全体组成的集合记为  $P(U)$ 。任何子集  $R \subseteq U \times W$  都称为从  $U$  到  $W$  的二元关系。如果  $U=W$ ,则称  $R$  为  $U$  上的二元关系。

称  $R$  是自反的,如果  $\forall x \in U \rightarrow (x, x) \in R$ ;称  $R$  是对称的,如果  $\forall x, y \in U, (x, y) \in R \rightarrow (y, x) \in R$ ;称  $R$  是传递的,如果  $\forall x, y, z \in U, (x, y) \in R, (y, z) \in R \rightarrow (x, z) \in R$ ;称  $R$  是等价的,如果  $R$  是  $U$  上的自反、对称、传递的二元关系。

对于从  $U$  到  $W$  的二元关系  $R$ ,定义邻域算子  $f_R: U \rightarrow P(W)$

$$f_R(x) = \{y \in W : (x, y) \in R\}$$

利用邻域算子  $f_R$ ,我们可以定义一对由  $P(W)$  到  $P(U)$  的集合算子  $L_R$  与  $H_R$ :

$$L_R(X) = \{x \in U : f_R(x) \subseteq X\} \quad \forall X \in P(W)$$

$$H_R(X) = \{x \in U : f_R(x) \cap X \neq \emptyset\} \quad \forall X \in P(W)$$

我们称  $(L_R(X), H_R(X))$  为由  $R$  产生的  $X$  的粗糙集。 $L_R$  与  $H_R$  分别称为广义下、上近似算子,  $(U, W, R)$  称为广义近似空间<sup>[2]</sup>。

当  $R$  是  $U$  上的等价关系时,容易知道:  $f_R(x)$  就是  $x$  所在的等价类,此时所得到的粗糙集模型称为 Pawlak 粗糙集模型<sup>[1]</sup>。

设  $L, H$  为由  $P(W)$  到  $P(U)$  的两个算子,如果对  $\forall A \in P(W)$ ,有  $L(-A) = -H(A)$ ,则称  $L$  与  $H$  是对偶算子。容易验证,上面所定义的广义下、上近似算子是一对对偶算子。更进一步地,我们有下面的结果:

**引理1**<sup>[6]</sup> 设  $L, H$  是由  $P(W)$  到  $P(U)$  的一对对偶算子,则存在二元关系  $R \subseteq U \times W$ ,使得  $L=L_R, H=H_R$  的充分必要条件是: (1)  $LU=U, H\emptyset=\emptyset$ ; (2)  $L(A \cap B) = LA \cap LB, H(A \cup B) = HA \cup HB$ 。

为了方便,我们把满足引理1的算子称为一对对偶的近似算子。下面考虑近似算子与二元关系的对应性。为此引入如下概念<sup>[7]</sup>:

设  $\Phi$  是  $P(U)$  到  $P(U)$  的算子。称  $\Phi$  是单调的,如果  $\forall X, Y \in P(U), X \subseteq Y \rightarrow \Phi(X) \subseteq \Phi(Y)$ ;称  $\Phi$  是嵌入的,如果  $\forall X \in P(U), X \subseteq \Phi(X)$ ;称  $\Phi$  是闭的,如果  $\forall X \in P(U), \Phi(\Phi(X)) \subseteq \Phi(X)$ ;称  $\Phi$  是对称的,如果  $\forall x, y \in U, x \in \Phi(\{y\}) \rightarrow y \in \Phi(\{x\})$ ;称  $\Phi$  是闭包算子,如果  $\Phi$  是单调、嵌入且是闭的;称  $\Phi$  是对称闭包算子,如果  $\Phi$  是闭包算子且是对称的。

**引理2**<sup>[2,7]</sup> 设  $L, H$  是  $P(U)$  到  $P(U)$  的一对对偶的近似算子,则存在  $U$  上的二元关系  $R$ ,使得  $L=L_R, H=H_R$  且  $R$  是自反的  $\Leftrightarrow H$  是嵌入的;  $R$  是对称的  $\Leftrightarrow H$  是对称的;  $R$  是传递的  $\Leftrightarrow H$  是闭的;  $R$  是等价的  $\Leftrightarrow H$  是对称闭包算子。

由于二元关系与集合算子之间有如上关系,我们很容易想到,能够利用算子的复合或二元关系的复合来方便地定义近似空间的复合。

## 2. 近似空间的复合

**定义1** 设  $A=(U, V, R), B=(V, W, S)$  是两个广义近似空间,  $A$  与  $B$  的复合定义为广义近似空间  $A \otimes B=(U, W,$

$T$ ). 其中  $T$  是二元关系  $R$  与  $S$  的复合, 即  $T=R \circ S$ .

下面的定理说明: 复合空间的近似算子等于这两个近似空间的近似算子的复合.

**定理1** 设  $A=(U, V, R), B=(V, W, S)$  是两个广义近似空间,  $A \otimes B=(U, W, T)$  是  $A$  与  $B$  的复合, 则: (1)  $L_T=L_R \circ L_S$ ; (2)  $H_T=H_R \circ H_S$ . 其中运算  $\circ$  为算子的复合.

证明: (1)  $\forall X \subseteq W$ , 需证  $L_T X=L_R(L_S X)$ .  $\forall x \in L_T X$ , 则  $f_T(x) \subseteq X$ , 下证  $f_R(x) \subseteq L_S X$ .

事实上: 若  $y \in f_R(x)$ , 则  $(x, y) \in R$ . 又  $\forall z \in f_i(y)$ , 有  $(y, z) \in S$ , 从而  $(x, y) \in R \circ S$ , 即  $z \in f_T(x) \subseteq X$ . 由  $z \in f_i(y)$  的任意性知  $f_i(y) \subseteq X$ , 故  $y \in L_S X$ . 再由  $y \in f_R(x)$  的任意性知:  $f_R(x) \subseteq L_S X$ . 因此  $x \in L_R(L_S X)$ , 从而  $L_T X \subseteq L_R(L_S X)$ .

反之,  $\forall x \in L_R(L_S X)$ , 则  $f_R(x) \subseteq L_S X$ . 于是,  $\forall y \in f_R(x)$ , 有  $f_i(y) \subseteq X$ . (\*)

若  $z \in f_T(x)$ , 则  $(x, z) \in T$ , 从而存在  $y \in V$ , 使得  $(x, y) \in R$  且  $(y, z) \in S$ , 即  $y \in f_R(x)$ ,  $z \in f_S(y)$ . 由 (\*) 式即得  $f_S(y) \subseteq X$ . 因此  $z \in X$ . 由  $z \in f_T(x)$  的任意性知  $f_T(x) \subseteq X$ , 即  $x \in L_T X$ . 因此  $L_R(L_S X) \subseteq L_T X$ .

这样便证明了  $L_T X=L_R(L_S X), \forall X \in P(W)$ .

(2) 由上、下近似算子的对偶性及(1)的结果立即可得.

**定理2** 设  $R$  与  $S$  是  $U$  上的自反的二元关系,  $A=(U, R), B=(U, S), A \otimes B=(U, T)$ , 则  $T$  是自反的, 且对  $\forall X \subseteq U$ , 有: (1)  $L_T X \subseteq L_R X \subseteq X \subseteq H_R X \subseteq H_T X$ ; (2)  $L_T X \subseteq L_S X \subseteq X \subseteq H_S X \subseteq H_T X$ .

证明: 易知  $T$  是自反的. (1) 由引理2及上、下近似算子的对偶性得:  $L_R X \subseteq X \subseteq H_R X$ . 又由定理1及引理2得:  $L_T X=L_R(L_S X) \subseteq L_R X; H_T X=H_R(H_S X) \subseteq H_R X$ , 因此(1)成立. (2) 同理可证.

在知识学习中, 间接学习是学习者通过中间知识学习导师的知识. 设  $A, B$  与  $D$  分别是其属性集, 则直接学习的学习质量<sup>[3]</sup>为:  $k^* = |Pos_A D|/|U|$ , 间接学习的学习质量为  $k = |Pos_{A \otimes B} D|/|U|$ . 其中,  $Pos_{A \otimes B} D = \bigcup_{E \in U/D} A \otimes B(E) = \bigcup_{E \in U/D} \underline{A}(\underline{B}E)$ . 由下近似的性质易知: 间接学习的学习质量低于直接学习的学习质量. 但是, 当导师的知识能由中间知识精确表达时,  $k^* = k$ . 另外, 若学习者的能力区包含于中间知识的能力区, 即  $Pos_A D \subseteq Pos_B D$ , 则对  $\forall E \in U/D$ , 有  $\underline{A}E \subseteq \underline{B}E$ , 从而  $\underline{A}\underline{A}E \subseteq \underline{A}\underline{B}E$ . 再由引理3即可推出  $\underline{A}E = \underline{A}\underline{B}E$ . 这样,  $Pos_{A \otimes B} D = Pos_A D$ . 这时间接学习的学习质量就等于直接学习的学习质量.

### 3. 关系闭包下的近似空间

设  $R$  是  $U$  上的二元关系,  $U$  上包含  $R$  的最小的自反、对称、传递关系分别称为  $R$  的自反闭包、对称闭包和传递闭包, 并分别记作  $r(R), s(R)$  和  $t(R)$ . 我们熟知:

$$r(B) = R \cup I, s(R) = R \cup R^{-1}, t(R) = \bigcup_{i=1}^n R^i$$

其中  $n$  为  $U$  中的元素个数,  $I = \{(x, x) : x \in U\}, R^{-1} = \{(x, y) : (y, x) \in R\}$ .

下面我们来考察近似空间  $(U, r(R)), (U, s(R))$  和  $(U, t(R))$  中的近似算子与  $(U, R)$  的近似算子之间的关系. 先给出一个定理:

**定理3** 设  $R, S$  是  $U$  上的二元关系, 则  $f_{R \cup S}(x) = f_R(x)$

$\cup f_S(x)$ .

$$\begin{aligned} \text{证明: } f_{R \cup S}(x) &= \{y \in U : (x, y) \in R \cup S\} \\ &= \{y \in U : (x, y) \in R \text{ 或 } (x, y) \in S\} \\ &= \{y \in U : (x, y) \in R\} \cup \{y \in U : (x, y) \in S\} \\ &= f_R(x) \cup f_S(x) \end{aligned}$$

**定理4** 设  $R$  是  $U$  上的二元关系, 则对  $\forall X \subseteq U$ , 有:

$$\begin{aligned} (1) L_{r(R)} X &= L_R X \cap X; H_{r(R)} X = H_R X \cup X; \\ (2) L_{s(R)} X &= L_R X \cap L_R^{-1} X; H_{s(R)} X = H_R X \cup H_R^{-1} X; \\ (3) L_{t(R)} X &= \bigcap_{i=1}^{\infty} L_R^i X; H_{t(R)} X = \bigcup_{i=1}^{\infty} H_R^i X. \end{aligned}$$

其中  $L_R^k$  是算子  $L_R$  的  $k$  次复合. 特别地, 当  $R$  是自反关系时, 有

$$L_{r(R)} X = L_R^* X; H_{r(R)} X = H_R^* X$$

证明: (1) 由定理3得:

$$\begin{aligned} L_{r(R)} X &= \{x \in U : f_{r(R)}(x) \subseteq X\} = \{x \in U : f_R \cup I(x) \subseteq X\} \\ &= \{x \in U : f_R(x) \cup f_I(x) \subseteq X\} \\ &= \{x \in U : f_R(x) \subseteq X\} \cap \{x \in U : f_I(x) \subseteq X\} = L_R X \cap X \\ H_{r(R)} X &= \{x \in U : f_{r(R)}(x) \cap X \neq \emptyset\} = \{x \in U : [f_R(x) \cup f_I(x)] \\ &\quad \cap X \neq \emptyset\} \\ &= \{x \in U : f_R(x) \cap X \neq \emptyset\} \cup \{x \in U : f_I(x) \cap X \neq \emptyset\} \\ &= H_R X \cup X \end{aligned}$$

(2) 与(3)类似于(1)可证.

当  $R$  是自反关系时, 由引理2知,  $L_R^k$  关于  $k$  是递减的, 而  $H_R^k$  关于  $k$  是递增的. 因此利用(3)的结果可得:  $L_{t(R)} X = L_R^* X; H_{t(R)} X = H_R^* X$ .

### 4. 模糊近似空间的复合

设  $U, W$  是非空有限论域,  $U$  的模糊子集全体记作  $F(U)$ . 任何映射  $R: U \times W \rightarrow [0, 1]$  都称为由  $U$  到  $W$  的模糊关系.

对于  $\forall X \in F(W)$ , 定义  $X$  的下、上近似为  $U$  上的模糊子集  $L_R X$  与  $H_R X$ , 其隶属函数为:

$$\begin{aligned} L_R X(u) &= \min_{y \in W} \max\{1 - R(u, y), X(y)\}, u \in U \\ H_R X(u) &= \max_{y \in W} \min\{R(u, y), X(y)\}, u \in U \end{aligned}$$

称  $(L_R X, H_R X)$  为广义模糊粗糙集,  $L_R$  与  $H_R$  为广义模糊下、上近似算子,  $(U, W, R)$  为广义模糊近似空间. 这里广义模糊下、上近似算子的定义与文[8]中是不同的.

容易证明: 广义模糊下、上近似算子具有如下性质:

**定理5** (1)  $L_R X = -H_R(-X)$ ; (2)  $L_R W = U; H_R \emptyset = \emptyset$ ; (3)  $L_R(X \cap Y) = L_R X \cap L_R Y, H_R(X \cup Y) = H_R X \cup H_R Y$ ; (4) 若  $X \subseteq Y$ , 则  $L_R X \subseteq L_R Y, H_R X \subseteq H_R Y$ .

对于两个广义模糊近似空间  $A=(U, V, R), B=(V, W, S)$ ,  $A$  与  $B$  的复合定义为广义近似空间  $A \otimes B=(U, W, T)$ . 其中  $T$  为  $R$  与  $S$  的复合  $T=R \circ S$ , 即  $T$  是由  $U$  到  $W$  的模糊关系. 对  $\forall (u, w) \in U \times W$

$$T(u, w) = \max_{y \in V} \min\{R(u, y), S(y, w)\}$$

我们证明下面的结果:

**定理6** 设  $A=(U, V, R), B=(V, W, S)$  是两个广义模糊近似空间,  $A \otimes B=(U, W, T)$  是  $A$  与  $B$  的复合, 则: (1)  $L_T = L_R \circ L_S$ ; (2)  $H_T = H_R \circ H_S$ .

证明: 先证(2). 对  $\forall X \in F(W), u \in U$ , 有

$$(H_R \circ H_S) X(u) = H_R(H_S X)(u)$$

$$= \max_{y \in V} \min\{R(u, y), H_S X(y)\}$$

(下转第104页)

期的药品作为不同类别,可分为8类。有27个训练样本和23个测试样本,训练样本集和测试样本集为医学院先后提供的两批数据。

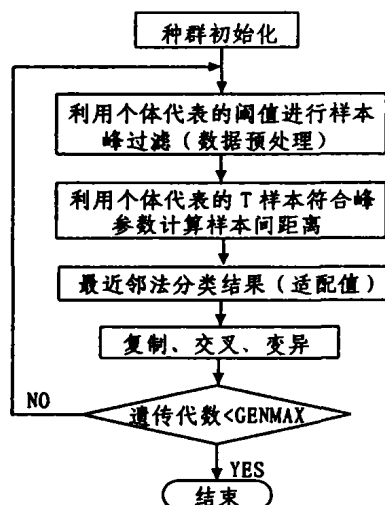


图6 遗传算法优化参数流程图

采用 Leave-One-Out 的验证方法<sup>[10]</sup>对训练分类器进行验证,即每次从训练样本集中取一个未验证过的样本做为测试样本,其他的样本做为训练样本。这样依次下去,就可以得到训练集中所有样本的最优的测试结果,从而可以得到一个训练样本分类正确率。

训练样本分类正确率 =  $\frac{\text{所有训练样本中被正确分类的数目}}{\text{所有训练样本的数目}}$

最后用测试集样本对分类器进行测试,可以得到测试样本分类正确率。

实验结果如表1所示。

表1 样本按不同产地不同日期的分类结果

| 类别     | 训练<br>样本数 | 训练样本<br>分类正确率 | 测试<br>样本数 | 测试样本<br>分类正确率 |
|--------|-----------|---------------|-----------|---------------|
| pa0828 | 3         | 100%          | 3         | 100%          |
| pa0911 | 3         | 100%          | 3         | 100%          |
| pa0925 | 3         | 100%          | 3         | 100%          |
| xj0820 | 6         | 100%          | 2         | 100%          |
| xj0905 | 3         | 100%          | 3         | 100%          |
| xj1005 | 3         | 100%          | 3         | 100%          |
| zjpa98 | 3         | 100%          | 3         | 66.7%         |
| zjxj98 | 3         | 100%          | 3         | 100%          |
| 总分类正确率 | 27        | 100%          | 23        | 95.65%        |

结果分析:

1)虽然样本数和一般的分类相比比较少,但是由于此领域研究取数据耗时间长,因此样本一般都不多<sup>[2~6]</sup>,而且由于本方法并未采用相关的统计信息,只是采用了最近邻法进行分类,而最近邻法的分类正确率通常随着样本数的增加而增加,因此可以预见当样本数增加时,分类正确率不会有很多降低。

2)关于分类错误样本的详细说明:zjpa98的一个测试样本被错误分类到zjxj98,从色谱图上看,这两类样本的确很相似,当样本数足够多时,可以采用多级分类器的方法解决此问题,即先进行粗分类(如选取较大峰面积过滤阈值T),这时可将不同类的相似样本分到一个粗类里,然后再进行细致的分类(如选取较小的峰面积过滤阈值T)。

3)用遗传算法收敛很快,本文采用了20代,每代30个体的规模很快得到结果,在第一代,就可以得到最优结果,在第三代就明显收敛。而且遗传算法的评价函数为训练样本集的分类正确率,可以保证得到较高的训练样本分类准确率,经过多次验证,效果都很好。

4)所见文献提到的中药识别,都未细致到对不同收获日期不同产地的同属样品进行鉴别。本文提出的方法可以进行此种鉴别,并能得到较好结果,那么可以推测这种方法对不同种属药品进行鉴别可以得到更好的识别结果。

## 参考文献

- 1 罗旭,毕开顺,等. 中药质量化学模式识别研究的进展. 药科学报, 1993, 28(12): 936~940
- 2 乔延江,王玺,等. 人工神经网络在中药蟾酥化学模式识别特征提取中的应用. 药科学报, 1995, 30(9): 698~701
- 3 吴玉柱,罗旭,等. 中药大黄质量的化学模式识别. 药科学报, 1991, 26(2): 132~138
- 4 王秀坤,李家实,等. 苦参质量的化学模式识别. 中国中药杂志, 1996, 21(4)
- 5 刘谦光,陈战国,等. 西洋参质量的化学模式识别. 中草药, 1999, 30(11)
- 6 杨维稼,杨健,等. 挥发油未知生药的气相色谱—计算机检索. 中日友好医院学报, 1996, 10(4)
- 7 边肇祺,张学工,等. 模式识别. 北京:清华大学出版社, 2000, 第二版
- 8 Kohavi R, John G H. The Wrapper Approach, In Feature Selection for Knowledge Discovery and Data Mining, H. Liu & H. Motoda(eds.), Kluwer Academic Publishers, pp 33~50
- 9 陈国良. 遗传算法及其应用. 北京:人民邮电出版社, 1996
- 10 Fukunaga K. Introduction to statistical pattern recognition. Boston: Academic Press, c1990, 2nd ed.

(上接第97页)

$$\begin{aligned}
 &= \max_{y \in V} \min \{R(u, y), \max_{z \in W} \min \{S(y, z), X(z)\}\} \\
 &= \max_{y \in V} \max_{z \in W} \min \{R(u, y), \min \{S(y, z), X(z)\}\} \\
 &= \max_{z \in W} \min \{\max_{y \in V} \min \{R(u, y), S(y, z)\}, X(z)\} \\
 &= \max_{z \in W} \min \{R \cdot S(u, z), X(z)\} = H_T X(u)
 \end{aligned}$$

故  $H_T = H_R \cdot H_S$ 。再证(1): 由定理5及(2)的结果可得:

$$\begin{aligned}
 \forall X \in F(W) (L_R \cdot L_S) X &= L_R(-H_S(-X)) \\
 &= -H_R(H_S(-X)) = -H_T(-X) = L_T X
 \end{aligned}$$

故  $L_T = L_R \cdot L_S$ 。

## 参考文献

- 1 Pawlak Z. Rough sets[J]. International Journal of Computer and Information Sciences, 1982, 11(5): 205~218
- 2 张文修,吴伟志,梁吉业,等. 粗糙集的理论与方法[M]. 北京: 科学出版社, 2001
- 3 曾黄麟. 粗糙集理论及其应用[M]. 重庆: 重庆大学出版社, 1998
- 4 王珏,苗夺谦,周育健. 关于 Rough set 理论与应用的综述[J]. 模式识别与人工智能, 1996, 9(2): 337~344
- 5 祝峰,何华灿. 粗糙集的公理化[J]. 计算机学报, 2000, 23(3): 330~333
- 6 Yao Y Y. Constructive and algebraic methods of the theory of rough sets[J]. Information Sciences, 1998, 109(1): 1~47
- 7 Thiele H. On axiomatic characterizations of crisp approximation operators[J]. Information Sciences, 2000, 129(1): 221~226
- 8 Morsi N N, Yakout M M. Axioms for fuzzy rough sets[J]. Fuzzy Sets and Systems, 1998, 100(2): 327~342