Rough 집합을 이용한 사례베이스에 관한 연구

○
*최 성 혜, **정 환 묵
*경주전문대학 전자계산과
**대구효성가톨릭대학교 전자정보공학부

A Study on Reducsion of CBR Using Rough set

*Choi Soung Hea, **Chung Hwan Mook
*Kyongju Junior College
**Department of Computer Engineering, Catholic University of Taegu Hyosung

요 약

실세계에서 존재하는 대부분의 지식은 다양한 패턴들로 구성되어있다. 본 논문에서는 사례베이스 추론 (Case-Based Reasoning : CBR)에서 다중의 의미를 갖는 불확실한 지식을 쉽게 표현할 수 있는 러프 집합을 이용하여 지식의 함축의 의미를 갖는 지식을 간략화하는 방법을 제안한다. 전문가의 지식 구조를 명확화 하는 데는 많은 노력이 필요하고 지식획득의 병목현상이 일어난다. 이러한 문제점을 해결하기 위해 많은 사례의 수를 러프 집합의 성질을 이용하여 사례를 동치 클래스로 분류하여 사례의 수를 감소하므로써 CBR의 기능을 향상 시킨다.

1. 서 론

rough set은 1982년 Z. Pawlak[4]에 의해 제안 되었으며, 전문가에 의한 분류를 데이터에 동치류에 의해 평가 할 수 있다. 러프 집합의 응용으로는 Z. Pawlak[5]에 의한 심장병에 관한 의사의 판단과 환자의 검사 데이터와의 정합성의 해석을 하고있다. A. Morzek[1]은 러프 집합의 개념을 이용하여 If -Then 규칙을 구성하고, 감소된 규칙에 의해 제어에 적용하고 있다. 또한, 환자의 검사 데이터에 의한 지식을 정보시스템이라 하고, 데이터의 속성의 수를 감소화하는 연구가 Z. Pawlak에 의해 이루어지고 있다. 본 논문에서는 전문가 시스템에서 지식 표현의 다중의 의미를 갖는 불확실한 정보를 쉽게 표현 할 수 있는 러프 집합을 이용하여 사례베이스 시스템에 적용하는 방법을 제안한다. 많은 수의 사례들을 함축의 의미를 갖는 동치류의 사례로 간략화하는 방법을 제시한다..

2. 러프 집합과 정보 시스템

2.1 러프 집합

Z. Pawlak에 의한 러프 집합은 U가 전체집합일 때 R은 $U \times U$ 상의 동치관계로써 $A=(U,R)$를 근사공간(Approximation Space)라 한다. 또, R은 식별불능관계라 하며, 만약 $x, y \in U$ 혹은 $(x,y) \in R$이면 x와 y는 A에 대해 식별불가능하다. 이 때 R의 동치류를 A의 기본집합이라 한다. 단, 공집합도 기본집합이다. A의 기본 집합이 유한 개의 합집합일 때도 A로 정의할 수 있다. 근사공간 A를 동치관계 R로써 분할한 상공간은 [그림 1]과 같다.

X는 U의 부분집합으로써 X를 포함하는 A의 최소 정의 가능한 집합은 A에 대한 X상한에서 근사(Upper Approximation)하다. 이것을 $A^*(X)$로 표시한다. 이와 같이 X에 포함되는 A의 최대 정의 가능한 집합을 A에 대한 X의 하한에서 근사(Lower Approximation) 하다고 한다. 이것을 $A_*(X)$로 표시한다.

$$B(X) = A^*(X) - A_*(X) \tag{2.1}$$

A에 대한 X의 경계집합이라 한다. [그림 1]은 U가 R에 의해 16개의 기본 집합 $E_1, \cdots, E_{16}$의 동치류(equivalence class) 로 나누어진 것이다. 이 때

$$A^*(X) = E_1 \cup \cdots \cup E_{12} \tag{2.2}$$
$$A_*(X) = E_6 \cup E_7 \tag{2.3}$$
$$B(X) = E_1 \cup \cdots \cup E_5 \cup E_8 \cup \cdots \cup E_{12} \tag{2.4}$$

즉, (2.2),(2.3)식은 다음과 같이 정의할 수 있다.

$$A^*(X)=\{E_i \mid E_i \cap X \neq \emptyset\} \quad (2.5)$$

$$A_*(X)=\{E_i \mid E_i \subset X\} \quad (2.6)$$

근사공간 $A=(U,R)$에 대한 집합 X의 근사 정도는

$$\alpha_A(X)= Card\ (A_*(X))\ /\ Card\ (A^*(X)) \quad (2.7)$$

단, $Card\ (X)$는 집합 X의 원소 수이다. 식 (2.7)에서 밝혀지듯이 $0 \le \alpha_A(X) \le 1$이고, 만약 X가 A의 정의 가능한 집합이면 $\alpha_A(X)=1$이 되고, X는 근사공간 A에서 완전하게 표현할 수 있다. $\alpha_A(X)$의 값이 큰 만큼 근사에 가깝다. X_i는 U의 부분집합이고 $F=\{X_1,\cdots,X_n\}$는 U의 부분 집합의 군이다. $F\subseteq R$, $F\neq 0$이면 $\bigcap F$도 동치관계가 되고 $IND(F)$는

$$[X]IND(F)=\bigcap_{R\in F}[X]_R \quad (2.8)$$

이때 F와 U의 분류 X_i는 F의 클래스이다. A에 대한 F의 상한에서의 근사와 하한에서의 근사는 다음과 같다.

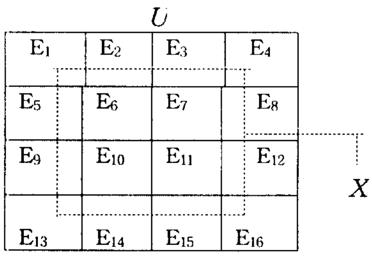
$$A^*(F)=\{A^*(X_1),\ \cdots\ ,A^*(A_n)\} \quad (2.9)$$

$$A_*(F)=\{A_*(X_1),\ \cdots\ ,A_*(X_n)\} \quad (2.10)$$

A에 대한 분류 $F=\{X_1,\ \cdots\ ,X_n\}$의 근사 정도는

$$\beta_A(F) = Card\ (\ \bigcup_{i=1}^{n} A_*(X_i)\ )\ /\ Card(U) \quad (2.11)$$

모든 X_i가 A의 정의 가능한 집합이면 $\beta_A(F)=1$, 분류 F는 근사 공간 A에서 완전히 표현할 수 있다.



[그림 1] 근사공간 $A=(U,R)$

2.2 정보 시스템의 간소화

정보 시스템을 $S=(X,Q,V,\Psi,F)$로 나타낸다. X는 대상의 집합, Q는 속성의 집합, V는 속성치의 집합, Ψ는 $\Psi: X\times Q \to V$의 기술 함수, F는 전문가에 의해 주어진 분류이다 Ψ는 <표 1>에서 나타내는 데이터베이스 이다. 정보 시스템 S의 예를 <표 1>과 같다.

<표 1> 정보 시스템의 예

F	U	q1	q2	q3	q4
X1	x1	0	0	0	0
	x2	0	1	0	2
	x3	1	1	0	1
X2	x4	1	1	0	1
	x5	0	1	1	2

$X=\{x_1,\cdots,x_5\}$, $Q=\{q_1,\cdots,q_4\}$

$V=\bigcup_{q\in Q} V_q$ 이며, 이 때

$V_{q1}=V_{q2}=V_{q3}=\{0,1\}$, $V_{q4}=\{0,1,2\}$가 된다. 또한, x_i와 x_j가 q에 대해 식별할 수 없는 것을 $\Psi(x_i,q)=\Psi(x_j,q)$라 정의하고, 이것을 동치관계 $ind(q)$를 이용해서 $(x_i,x_j)\in ind(q)$로 나타낸다. x_i와 x_j가 Q의 부분집합 P에 대해 식별할 수 없는 것을 $(x_i,x_j)\in ind(P)$ 로 나타낸다. 이것은

$$ind(P)=\bigcap_{q\ni p} ind(P) \quad (2.12)$$

특히 $P=Q$일 때 $(x_i,x_j)\in ind(Q)$이면 x_i,x_j는 정보 시스템 S에 대해 식별할 수 없게 된다. $ind(q)$는 동치관계에 있으므로 X의 분할을 상집합이라 한다.

$$X\ /\ ind(q)=\{[x_i] \mid x_i\in X\} \quad (2.13)$$

라 나타낸다. <표 1>에서는

$$X\ /\ ind(q_4)=\{[x_1],[x_2,x_5],[x_3,x_4]\} \quad (2.14)$$

가 된다. 단 $[x_2,x_5]$는 x_2와 x_5가 q_4에 관해 식별할 수 없음을 의미한다. 그러므로 속성의 종속성에 따라 정보 시스템을 다음과 같이 감소화를 정의 할 수 있다.

(1) 속성 p는 속성 q에 종속일 때 $p\to q$, $p\to q \Leftrightarrow ind(q)\subset ind(p)$로 정의한다.
(2) p와 q는 독립일 때 $\Leftrightarrow p\to q$, $q\to p$ 의 어느 쪽도 독립이 아니다.
(3) 부분집합 $P\subset Q$가 독립일 때 $\Leftrightarrow$ 모든 $P'\subset P$에 대해 $ind(P')\supset ind(P)$.
(4) 부분 집합 $P\subset Q$가 종속일 때 $\Leftrightarrow ind(P')=ind(P)$가 된다. $P'\subset P$가 존재
(5) 부분집합 $P\subset Q$가 P에 대해 여분 일 때 $\Leftrightarrow ind(P-P')=ind(P)$
(6) 부분집합 $P\subset Q$가 Q에 대해 여분일 때 $\Leftrightarrow Q-P$가 여분이고 P는 독립이다.

(6)에서 Q의 간략화를 P로써, P에 대응하는 정보 시스템을 $S'=(X,P,V,\Psi')$로 나타낸다. 이것을 간략화한 정보 시스템이라 한다. Q의 간략화는 다음의 성질을 이용할 수 있다.

(a) 부분 집합 P가 독립일 때 $\Leftrightarrow$ 모든 i에 대해 $ind(P_i)\supset P$, 단 $P_i=P-\{p_i\}$.
(b) 부분집합 P가 Q에 대해 여분일 때 $\{p_i\}$가 $Q-P$에서 여분이면 $P\cup\{p_i\}$는 Q에 대해 여분

성질 (a)에 의해 속성 집합의 P가 독립인지 아닌지를 비교할 수 있다. 성질 (b)에서 여분의 속성을 하나씩 감소할 수 있다. <표 1>의 Q의 간략화는 $\{q_1,q_2,q_3\}$와 $\{q_3,q_4\}$이다.

3. 간략화 방법

지식을 축소된 카테고리로 정형화함으로써 간략화 할 수 있다. 근사 공간 A에 대한 $F=\{X_1,\cdots,X_n\}$을

분류할 때 속성 부분집합 $P' \subset F$에 대해

$$\beta_A\, ind(P') \le \beta_A\, ind(P) \qquad (3.1)$$

$\beta_A\, ind(P)$는 속성의 부분 집합 F에 대한 식(2.11)과 같다. 또한 $P' \subset F$가 Q의 간략화를 할 수 있는 것은 $Q-F$가 여분이다. 즉, R이 독립이고, $P \subseteq R$이면 F도 독립이다. Q가 독립적이고, $IND(Q)=IND(P)$이면 $Q \subseteq F$는 F의 reducter가 된다. 이 때 F에서 필요한 모든 집합 F의 CORE라하며 $CORE(P)$로 나타낸다. $RED(P)$는 F의 축소된 모든 군이다

$$CORE(P) = \bigcap RED(P) \qquad (3.2)$$

Q의 간략화 알고리즘은 다음과 같다.

[step 1] 속성의 집합 Q에 대해 분류한 여분의 속성 q_i를 탐색. 여분의 속성이 없으면 [step 4].

[step 2] Q를 $Q-\{q_i\}$로 치환. 즉, $Q \triangleq Q-\{q_i\}$

[step 3] 모든 속성 $\{q_i\} \subset Q$가 Q에 대해 독립이면 [step 4]로,그렇지 않으면 [step 1].

[step 4] 얻어진 Q는 간략화된 것이다.

이상에서 Q의 간략화에서 정보 시스템을 감소화 할 수 있다. <표 1>에서 분류 F가 $F=\{X_1, X_2\}$일 때

$X_1=\{x_1, x_2, x_3\}$, $X_2=\{x_4, x_5\}$로써

$P=\{q_3, q_4\}$로 한다. 먼저 $Q=(q_1, \cdots, q_4)$이므로

$$A_*(X_1)=\{x_1, x_2\}, \quad A^*(X_1)=\{x_1, x_2, x_3, x_4\},$$
$$A_*(X_2)=\{x_5\}, \quad A^*(X_2)=\{x_3, x_4, x_5\} \qquad (3.8)$$

된다. $\beta_A\, ind(Q)=3/5=0.6$이다. 한편 $P=\{q_4, q_4\}$가 되면 X_1, X_2의 상한에서 근사하면 하한에서 근사하는 것은 $\beta_A\, ind(P)=3/5=0.6$가 된다.

$$A_*(X_1)=\varnothing, \quad A^*(X_1)=\{x_1, x_2, x_3, x_4\}$$
$$A_*(X_2)=\{x_5\}, \quad A^*(X_2)=\{x_1, x_2, x_3, x_4, x_5\}$$
$$A_*(X_1)=\{x_1\}, \quad A^*(X_2)=\{x_1, x_2, x_3, x_4, x_5\}$$
$$A_*(X_2)=\varnothing, \quad A^*(X_2)=\{x_2, x_3, x_4, x_5\}$$ 이므로

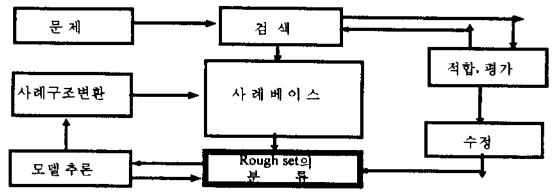
$$\beta_A\, ind(P') = 0.2 < \beta_A\, ind(P)$$
$$\beta_A\, ind(P'') = 0.2 < \beta_A\, ind(P)$$

가 된다. F는 분류에서 독립이므로 Q의 분류로 간략화로 기본집합의 수를 감소화함으로써 정보 시스템을 감소시킬 수 있다.즉, 동치 클래스 내에서는 기본집합을 결합할 수 있지만 다른 클래스간에는 기본 집합을 결합할 수 없다.

4. 사례베이스의 간소화

4.1 사례베이스

전문가의 경험적인 지식은 규칙으로 명확히 기술하기 어려운 경우가 있어 전문가 시스템을 구축하는데 어려움이 있다 유사한 과거의 사례를 검색하여 검색된 사례를 결과에 이용하여 현재의 문제를 해결한다.전문가의 지식 구조는를 명확화 하는데는 많은 노력이 필요하고 지식획득의 병목현상이 일어난다. 이러한 해결을 위해 지식에 비해 획득이 쉬운 과거의 사례를 직접 이용함으로써 해를 유도하는 사례를 기반으로하는 추론(Case-Based Reasoning : CBR)의 연구가 이루어지고 있다. 그러나 CBR 시스템을 구축하기 위해서는 사례를 특징지우는 색인(슬롯트) 그룹으로 구성되는 사례를 획득하는 것이 곤란하다. 즉, 사례 구조 획득의 병목현상이 지식 획득 의 병목현상이 되므로 시스템의 성능을 저하 시키는 요인(바이어스)이 된다 그러므로 본 논문에서는 사례의 수집 및 분류를 러프 집합의 이용하여 사례를 동치 클래스로 분류하여 사례의 수를 감소하므로써 CBR 의 기능을 향상 시킨다. 사례구조 변환을 가지는 CBR 시스템은 [그림 2]와 같다.



[그림 2] CBR의 구성도

4.2 사례베이스의 간소화

지식 표현은 데이터 테이블 (Knowledge Representation System : KRS)을 이용하여 동치관계를 표현한다. 간염과 감기의 진단 사례의 KRS는 <표 2>와 같다. KRS은 경험, 과거의 사례, 측정 등으로 관찰된 과거의 사례들로 구성된다. 정형화된 지식 표현의 KRS는 $S=(U, A)$로 나타내며 모든 조건속성 $a \in A$는 전체함수 $a: U \rightarrow V_a$ 이고, V_a는 a의 값에 대한 부분집합이다. 조건속성의 모든 부분집합 $B \subseteq A$를 속성이라하면 $IND(B)$는 동치관계로써 다음과 같이 정의된다

$$IND(B) = \bigcap_{a \in B} IND(a) \qquad (4.1)$$

KRS인 $K=(U, A)$에서 $C \subseteq A$는 조건속성을, $D \subseteq A$는 결정속성이며 $T=(U, A, C, D)$로 나타낸다. KRS는 지식의 간략화 방법으로는 조건속성의 감소하여, 중복행의 제거하고, 속성값의 간략화.

<표 2> 간염, 감기의 사례

증상 \ 사례	C1	C2	C3	C4	C5	C6	C7
sneeze	0	1	0	1	0	0	0
sore-throat	0	0	1	0	0	1	1
cough	0	0	1	0	0	0	1
headache	1	1	0	0	0	1	1
pneumonia suspicion	1	1	1	1	0	0	0

	C1	C2	C3	C4	C5	C6	C7	C8
sneeze	0	1	1	1	1	0	0	0
sore-throat	0	1	0	1	0	0	1	1
cough	0	1	0	1	1	0	0	1
headache	1	1	1	0	1	0	1	1
cold	1	1	1	1	1	0	0	0

조건속성 $\{a, b, c, d\}$와 결정속성 $\{f\}$는 <표 3>과 같다

<표 3>조건 속성과 결정 속성

	a	b	c	d	f
증상	sneeze	sore-th.	cough	headac.	pne-sus, cold
0	false	false	false	false	false
1	true	true	true	true	true

KRS 의 $S=(U, A)$의 의사결정은 <표 4>과 같다.

<표 4> KRS

X	a	b	c	d	Y
X_1	0	0	0	1	1 1
X_2	1	1	1	1	0 1
X_3	1	1	1	0	0 1
X_4	1	0	0	1	1 1
X_5	1	0	1	1	0 1
X_6	0	0	0	0	0 0
X_7	0	1	0	1	0 0
X_8	0	1	1	1	0 0
X_9	0	1	1	0	1 0
X_{10}	1	0	0	0	1 0

<표 4>의 의사결정의 속성 및 속성값의 논리값이 $f(00) \to y_1$, $f(01) \to y_2$, $f(10) \to y_3$, $f(11) \to y_4$ 일 때

$X_1 = a_0b_0c_0d_0 \to y_4 \quad \cdots \quad X_{10} = a_1b_0c_0d_0 \to y_3$

(1) 조건속성의 제거

$IND(R) = IND(R - \{R\})$의 조건속성의 독립성 판단

① $\{a\}$는 불제거(indispensavle) : 독립적, false (core)

$X_1 = b_0c_0d_0 \to y_1$, $\quad X_8 = b_0c_0d_0 \to y_3$

$X_3 = b_1c_1d_1 \to y_1$, $\quad X_4 = b_1c_1d_1 \to y_2$

$X_5 = b_1c_1d_0 \to y_2$, $\quad X_7 = b_1c_1d_0 \to y_3$

② $\{b\}$ 불제거 :독립적, false (core)

$X_2 = a_0c_0d_1 \to y_1$, $\quad X_9 = b_0c_0d_1 \to y_4$

③ $\{c\}$ 불제거 : 독립적이므로 false (core)

$X_6 = a_1b_0d_1 \to y_2$, $\quad X_{10} = a_1b_0d_1 \to y_4$

④ $\{d\}$ 불제거 : 독립적, false (core)

$X_1 = a_0b_0c_0 \to y_1$, $\quad X_9 = a_0b_0c_0 \to y_4$

$X_3 = a_0b_1c_1 \to y_1$, $\quad X_7 = a_0b_1c_1 \to y_3$

$X_8 = a_1b_0c_0 \to y_3$, $\quad X_{10} = a_1b_0c_0 \to y_4$

이상의 간략화된 조건속성의 값은 <표 5>와 같다.

<표 5> 간략화된 속성값

X	a	b	c	d	Y
X_1	0	×	×	0	y_1
X_2	0	1	×	×	y_1
X_3	0	×	×	1	y_1
X_4	1	×	×	×	y_2
X_5	1	×	×	×	y_2
X_6	×	×	1	×	y_2
X_7	0	×	×	0	y_3
X_8	1	×	×	0	y_3
X_9	×	0	×	1	y_4
X_{10}	×	×	0	1	y_4

(2) 속성값 제거

$\bigcap (F - \{X_i\}) = \bigcap F$조건에 따라 속성 값 제거

① $\{X_1\}$에서 $\{c\}$의 모든 속성 제거

$\{a_0, b_0, d_0\} \to y_1$

② $\{X_2, X_3\}$에서 $\{c\}$의 모든 속성 값 제거

$\{a_0, b_1, d_1\} \to y_1$

③ $\{X_4, X_5, X_6\}$에서 $\{b, d\}$의 모든 속성 값 제거

$\{a_1, c_1\} \to y_2$

④ $\{X_7\}$에서 $\{c\}$의 모든 속성 값 제거

$\{a_0, b_1, d_0\} \to y_3$

⑤ $\{X_8\}$에서 $\{c\}$의 모든 속성 값 제거

$\{a_1, b_0, d_0\} \to y_3$

⑥ $\{X_9, X_{10}\}$에서 $\{a, b\}$의 모든 속성 값 제거

$\{b_0, c_0, d_1\} \to y_4$

감소화된 KRS는 <표 6>과 같다.

<표 6> 최종적인 속성 값

X	a	b	c	d	Y
X_1	0	0	×	0	y_1
X_2, X_3	0	1	×	1	y_1
X_4, X_5, X_6	1	×	1	×	y_2
X_7	0	1	×	0	y_3
X_8	1	0	×	0	y_3
X_9, X_{10}	×	0	0	1	y_4

(3) 중복행 제거

: $CORE(P \to Q) = \bigcap RED(P \to Q)$에 따라 동일한 속성 값을 갖는 행을 결합한다. <표 6>에서 감소된 사례를 정규화시키고 결정 클래스로 결합하여 최종 사례가 되며. $a'b'd' \to y_1, ac \to y_2, b'c'd \to y_4$과 $a'bd' \vee ab'd' \to y_3$, 이 결정된다. <표 2>의 간소화된 사례는 <표 7>과 같다.

<표 7> 간략화된 간염, 감기의 사례

증상 사례	C1	C2	C3	C4	C5	증상 사례	C1	C2	C3	C4
sneeze	0	1	0	1	×	sneeze	0	0	1	×
sore-thr	0	0	1	0	0	sore-thr	0	1	×	0
cough	×	×	×	×	0	cough	×	×	1	0
headache	1	1	0	0	1	headache	0	0	×	0
pne.-suspi	0	0	1	1	1	cold	0	1	1	1

5. 결 론

전문가가 확립되어 있지 않은 과거의 사례를 이용하는 사례베이스 추론은 추론의 효율을 높일 수 있지만 많은 사례의 획득으로 지식의 병목현상을 일으켰다. 본 논문에서는 이러한 문제점을 해결하기 위하여 Rough set의 성질을 이용하여 사례를 함축하여 지식을 표현함으로써 지식의 축소 및 최소화된 사례를 구축할 수 있다. 적용의 예로써 의료진단의 간단한 사례로써 사례의 간략화를 적용하였다. Rough set은 지식 표현이 단순하며, 이해가 쉽고, 알고리즘이 용이하여 구현이 쉬우므로 많은 전문가 시스템에서 적용될것이다.

참고문헌

[1] A. Morzek : Rough sets and Dependency Analysis among Attributes in Computer Implementations of Expert's Inference Models, Int. J. of Man-Machine Studies, Vol. 30,p 457-473,1989.

[2] Bryniarski. E., "A Proposal of the Approximation theory of Sets", Norte Dame Journal of Formal Logic, 1991

[3] Gao,J. M. and Makamura, A. "A semantic Decision Method for the Logic of Indisceribility Relation", Fundermanta Informaticase, 1991

[4] Pawlak. Z., "Rough Sets", International Journnal of Computer annd Information Sciences, 11, pp.341-356, 1982

[5] Pawlak. Z., "Rough Sets - Theoretical Aspects of Reasoning about Data", Kluwer, 1991