러프-퍼지 추론 모델의 구성

The Structure of Rough-Fuzzy Inference Model

김두완, 정환묵
대구가톨릭대학교 컴퓨터 정보통신공학부
Kim Doo Ywan, Chung Hwan Mook
Faculty of Computer Information & Communication,
Catholic University of Taegu

ABSTRACT

대용량의 데이터베이스에서 효율적인 의사결정을 하기 위해서는 불필요한 지식을 제거한 지식베이스의 구축이 필요하다. 사용자의 언어적인 질의에 대해 대용량의 데이터베이스에서 불필요한 규칙을 제거한 최소지식베이스를 구축한다. 또한 불완전한 데이터베이스로부터 규칙들을 일반화한 근사함수에 기반하여 규칙 추출의 중요도를 나타낸다. 그리고 앞에서 생성된 최소지식베이스를 통해 언어적 변수에 대한 퍼지 연산을 수행하여 추론 값을 도출할 수 있는 모델을 제안한다.

1. 서론

데이터 마이닝은 방대한 데이터베이스로부터 지금까지 알려지지 않은 유용한 정보를 추출하기 위한 처리이다. 몇몇 모델들은 데이터베이스로부터 정보를 추출하고 표현하기 위해 사용되었다. 다른 표현들 역시 가능한 쉽게 식별하도록 만들었고 데이터에서 패턴들을 해석한다. 러프 집합 이론은 데이터베이스로부터 은닉 결정적 규칙들을 마이닝을 위해 최근에 고려할 수 있는 처리를 받아왔다. 연구자들은 또한 Pawlak에 의해 제안된 하나의 러프 집합 모델의 일반화를 제안했었다. 확률적 러프 집합 모델은 데이터베이스들로부터 확률적 규칙들을 추출하도록 만들었다[3].

본 논문에서는 불완전한 데이터를 고려하기 위해 믿음 함수를 이용해서 근사함수의 값을 구하여 임계치를 두어 최소 의사결정 지식베이스를 생성한다. 그리고 언어적 값을 소속도를 구하여 간단화 된 퍼지 논리 연산을 통해 변수의 값을 나타내는 모델을 제안한다.

먼저 2장에서는 일반적인 러프 이론과 러프-퍼지 집합을 소개하고 3장에서는 본 논문에서 제안한 러프 퍼지 규칙 모델을 소개하고 4장에서는 간단한 예를 통해 모델을 소개한다.

2. 러프-퍼지 집합

2.1 러프 이론

Pawlak에 의하여 제시된 러프 집합 이론은 데이터 분석과 데이터 마이닝에 대한 새로운 수학적 접근방법이며, 애매함과 불확실성을 다루기 위한 방법은 식별불능 관계와 근사공간에 근거한다. 러프 이론에서 제시된 부정확성 척도는 정보 부족에 따른 불완전 정도를 표현하며, 상한근사와 하한근사의 경계영역으로부터 발생되는 불확실성을 나타낸다[3].

러프 집합 이론을 이용하여 정보시스템을 분류하거나 추론규칙을 생성할 때 불완전 정보

는 러프 집합의 분류 능력을 감소시키고 정보 시스템의 불일치를 유발시키며 유도된 추론 규칙에 오류를 발생시킬 수가 있다.

퍼지 논리나 러프 논리 모두 데이터의 부분적 불확실성 관리에 효과적인 것으로 판단되나 현실 세계에서는 애매 모호성이 있으며 식별 불가능한 정보의 처리에 대한 러프-퍼지적 논리의 필요성이 대두되고 있다[4].

2.2 러프-퍼지 집합

X가 집합이고 R은 X상에 정의된 동치 관계이고 출력 클래스 $A \subseteq X$는 퍼지 집합이 된다고 하자. 러프-퍼지 집합은 쌍 $\langle \overline{R}(A), \underline{R}(A) \rangle$이고, 여기서 A의 하한근사 $\underline{R}(A)$와 상한근사 $\overline{R}(A)$는 정의된 소속 함수로 X/R의 퍼지 집합이다.

$$\mu_{\underline{R}(A)} = inf\{\mu_A(x) \mid x \in [x]_R\} \quad (2\text{-}1)$$
$$\mu_{\overline{R}(A)} = sup\{\mu_A(x) \mid x \in [x]_R\} \quad (2\text{-}2)$$

여기서, $\mu_{\underline{R}(A)}([x]_R)$과 $\mu_{\overline{R}(A)}([x]_R)$은 $\underline{R}(A)$와 $\overline{R}(A)$ 각각에서 $[x]_R$의 소속 값이다.

2.3 러프-퍼지 소속함수

퍼지 출력 클래스 $C_c = A \subseteq X$에 대해 패턴 $x \in X$의 러프-퍼지 소속 함수는 다음에 의해 정의된다.

$$l_{C_c}(x) = \frac{\|F \cap C_c\|}{\|F\|}$$

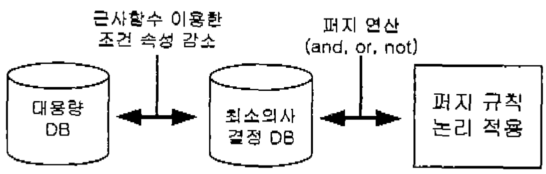
여기서 $F = [x]_R$과 $\|C_c\|$는 퍼지 집합 C_c의 카디널리티를 의미한다. 카디널리티를 결정하기 위한 한 가지 가능한 방법으로 다음과 같이 사용하는 것이다. : $\|C_c\| \overset{\text{def}}{=} \sum_{x \in X} \mu_{C_c}(x)$.

'∩'(교집합) 연산자에 대해,

$\mu_{A \cap B}(x) \overset{\text{def}}{=} \min\{\mu_A(x), \mu_B(x)\} \ \forall x \in X$를 사용한다. 그것은 러프-퍼지 소속 함수가 불확실한 데이터를 고려할 때 필요한 반면, 러프-퍼지 집합의 개념은 모호성 개념을 다룰 때 필수적이다.

3. 러프 퍼지 규칙 모델

대용량에 들어있는 속성들을 근사함수의 방법을 통해 불필요한 속성들을 감소해서 최소 지식 데이터베이스에 저장한다. 최소 지식 데이터베이스에 있는 속성들에 대해 일반적인 규칙을 퍼지 논리함수를 이용하여 규칙들을 생성한다. 또한 대용량 데이터베이스에 들어있는 불완전한 속성들을 일반화하기 위하여 Shafer에 의해 분류된 기본확률함수를 이용한 근사함수를 이용한다.



[그림 1] 러프 퍼지 규칙 모델

3.1 속성 감축을 위한 근사 함수

조건 속성들의 값에 대한 가능성 함수를 확장하여 각 값 $x_i \in D_{P_i}$에 대해, 그들이 클래스

$E_{|x_i|} \in U/R$일 때

$$E_{|x_i|} = \{e_m \mid x_i \in value_{P_i}(e_m)\} \quad (3\text{-}1)$$

하나의 원소가 아닌 값 $p_n \in 2^{D_{P_i}} - \{\emptyset\}$에 대해,

$$E_{p_n} = \bigcup_{x_i \in p_n} E_{|x_i|} \quad (3\text{-}2)$$

조건 속성의 규칙들을 E_P로 표현한다.

같은 방법으로, 각각의 값 $h_j \in D_Q$에 대해 집합 $H_{|h_j|}$는 다음처럼 정의된다.

$$H_{|h_j|} = \{e_m \mid h_j \in value_Q(e_m)\} \qquad (3\text{-}3)$$

하나의 원소가 아닌 값 $q \in 2^{D_Q} - \{\emptyset\}$에 대해,

$$H_q = \bigcup_{h_j \in q} H_{|h_j|} \qquad (3\text{-}4)$$

주어진 속성 값 $p \in 2^{P_i} - \{\emptyset\}$의 집합에 대해, 다음과 같이 근사 함수 $Pl: 2^{D_Q} \rightarrow [0,\ 1]$을 다음과 같이 표현한다. :

$$Pl(q) = \frac{|E_p \cap H_q|}{|E_p|} \qquad (3\text{-}5)$$

3.1.2 근사 결정 규칙 생성

많은 경우, 데이터 마이닝을 위해 사용된 데이터베이스들은 불완전하다. 데이터 마이닝에서 주요 이슈 중 하나가 속성들의 결손치(missing value)이다. 그것은 가능성 있는 값들의 집합에 의해 결손치들을 표현하는 것이 가능하다.

함수 $Pl: 2^{D_Q} \rightarrow [0, 1]$는 근사 함수로써 식 (3-5)를 얻는다. (3-5) 식이 의미하는 것은 근사 함수로 기본 확률 할당으로부터 유도될 수 있을 것이다. 결국, 근사적 규칙들은 다음과 같이 정의될 것이다. :

If $P_1 = x_1$ (and, or) ⋯ (and, or) $P_j = x_j$

then $Q = h_j$ ($Pl(q) = \frac{|E_p \cap H_q|}{|E_p|}$)

가능성 함수를 통해 나온 값을 임계값을 두어 최소의 규칙을 결정한다.

3.2 퍼지 의사결정 모델

최소 의사결정 DB에서의 언어적 변수에 대한 의사 결정의 퍼지 알고리즘은 다음과 같다.

① 입력 노드들에 대해 그것을 연결하는 가중치를 $W_1, W_2, \ldots, W_n$ 입력값을 $x_1, x_2, \cdots, x_i$이라면, 입력값과 가중치들의 쌍으로 표현한다.

$$[x_1,\ W_1],\ [x_2,\ W_2],\ \ldots,\ [x_i,\ W_i]$$

위의 쌍들의 첫 번째 요소의 값들에 의해 재배열한다.

$$[x_1',\ W_1'],\ [x_2',\ W_2'],\ \ldots,\ [x_i',\ W_i']$$

단 $x'_1(x) \le x_2(x) \le \ldots \le x'_i(x)$

② 서로 이웃하는 배열 요소 값들간의 차이 $d_1,\ d_2,\ \ldots,\ d_n$를 생성한다.

$$d_1 = x'_1(x)$$
$$d_j = x'_j(x) - x'_{j-1}(x) \qquad (3\text{-}6)$$
$$\text{for } j = 2,\ 3,\ \cdots,\ n$$

③ 가중치들의 합 $S_1,\ S_2,\ \cdots,\ S_n$를 생성한다.

$$S_i = \sum_{j=i}^{n} W_j' \qquad (3\text{-}7)$$

④ 임계값을 결정한다. $j = 1, 2, \cdots, n$에 대해

만일 $S_j < 1$ 이면, $S_j' = 0$

만일 $S_j \ge 1$ 이면, $S_j' = 1$

⑤ 소속함수 h_j를 계산한다.

$$h_j = \sum_{j=1}^{n} d_j S_j' \qquad (3\text{-}8)$$

4. 모의실험

다음과 같은 KR-시스템이 주어졌을 때

[표 1] 지식 베이스

U	P_1	P_2	P_3	Q
e_1	M	L	M	M
e_2	M	L	L	M
e_3	L	L	L	L
e_4	M, L	M	M	L
e_5	M	M	H	H
e_6	H	M, H	H	H
e_7	H	H	H	H

L : Low, M : medium, H : High

[표 1]의 지식베이스에서

$E_{P_{1(L)}} = \{e_3, e_4\}$, $E_{P_{1(M)}} = \{e_1, e_2, e_4, e_5\}$

$E_{P_{1(H)}} = \{e_6, e_7\}$

$E_{P_{2(L)}} = \{e_1, e_2, e_3\}$, $E_{P_{2(M)}} = \{e_4, e_5, e_6\}$

$E_{P_{2(H)}} = \{e_6, e_7\}$

$E_{P_{3(L)}} = \{e_2, e_3\}$, $E_{P_{3(M)}} = \{e_1, e_4\}$

$E_{P_{3(H)}} = \{e_5, e_6, e_7\}$

$H_{\{L\}} = \{e_3, e_4\}$, $H_{\{M\}} = \{e_1, e_2\}$

$H_{\{H\}} = \{e_5, e_6, e_7\}$

다음의 규칙들을 근사 규칙과 함께 나타낼 수 있다.

If $P_1 = M$ then $Q = M$ (1/2)

⋮

If $P_1 = L$ or $P_2 = M$ and $P_3 = M$ then $Q = L(1/1)$

⋮

If $P_1 = H$ and $P_2 = H$ and $P_3 = H$ then $Q = L(1/1)$

위 규칙들을 근사 값에 임계치를 두어 최소 의사결정 데이터베이스에 저장한다.

언어적 변수를 각각 소속도를 주어 퍼지 수로 확장하고 퍼지 연산을 한다.

최소 의사결정 데이터베이스의 규칙들 중 하나를 표현하면, $x_1 = 0.3$ and $x_2 = 0.5$ and $x_3 = 0.5$는 [0.3, 1], [0.5, 1], [0.5, 1]에서 [0.3, 1], [0.2, 1], [0.0, 1]로 재배열되어, 결국 $h_j = 0.5$가 된다.

이와 같은 방법으로 모든 규칙들을 나타낼 수 있을 것이다.

5. 결론

본 논문에서는 불완전한 데이터를 처리하기 위해, 근사 함수에 의해 구해진 근사치에 임계값을 두어 최소 의사 결정 DB를 구축한 후 퍼지 논리 연산을 수행하는 러프-퍼지 모델을 제안한다. 일반적으로 러프 집합은 데이터베이스에서 지식의 속성들을 감축하여 효과적인 의사결정 규칙을 생성하는데 유용하다. 여기서는 불완전한 데이터에서 규칙을 생성하고, 정확성의 척도를 나타내기 위해 근사 함수를 사용하였으며, 퍼지 논리 구현에서도 기존의 방법보다 좀 더 간단한 연산 방법을 사용하였다.

6. 참고문헌

[1] Glenn Shafer, "A Mathematical Theory of Evidence", Princeton University Press, 1975.

[2] 石塚 滿, "Dempster & Shafer の確率理論", 電子通信學會誌, Vol.66 , No.9

[3] P. J. Lingras 외 1, "Data Mining Using Extensions of the Rough Set Model", Journal of the American Society for Information Science, Vol.49 No.5, 1998

[4] 정구범, "러프집합을 이용한 불완전 정보처리 시스템에 관한 연구", 대구가톨릭대학교 대학원, 1999.

[5] 정환묵 외 1, "다치-뉴로 논리 모델의 구성", 한국퍼지 및 지능시스템학회 춘계학술대회, 제8권1호, 1998.