러프집합에 의한 불완전 데이터의 처리에 관한 연구

A Study on the Processing of Imprecision Data by Rough Sets

정 구 범*, 김 두 완**, 정 환 묵**
*상주대학교 교양과정부
**대구효성가톨릭대학교 전자정보 공학부

Jeong Gu Beom*, Kim Doo Ywan** Chung Hwan Mook**
*Dept. of Liberal Arts, Sangju National University
**Faculty of Electronics & Information Engineering,
Catholic University of Taeguhyosung

요 약

일반적으로 러프집합은 지식베이스 시스템에서 근사공간을 이용한 불확실한 데이터의 분류, 추론 및 의사결정 등에 사용된다. 지식베이스 시스템의 데이터 중에서 연속적인 구간 특성을 갖는 정량적 속성값이 불연속적일 때 중복 또는 불일치 등의 불확실성이 발생된다.

본 논문은 러프집합의 정량적 속성값들을 정성적 속성으로 변환시킬 때 식별 불가능 영역에 있는 정량적 속성값들을 명확한 경계를 갖는 보조구간으로 분리하여 불확실성을 제거함으로써 러프집합의 분류능력을 향상시키는 방법을 제안한다.

Ⅰ. 서 론

러프집합(*rough set*) 이론은 Pawlak에 의하여 제안되었다. 지식베이스 시스템에서 상한 및 하한 근사(*upper and lower approximation*) 개념을 이용하여 데이터의 속성에 대한 관계를 정형화함으로써 불확실하거나 부정확한 정보를 취급하기 위한 접근 방법을 제공하며, 불확실한 정보의 관리, 기계학습, 지식 발견, 부정확한 지식에 대한 표현 및 추론 등의 연구에 사용되고 있다.

러프집합 이론의 장점은 (1)데이터의 분류를 효과적으로 수행하고 (2)불확실하거나 부정확한 데이터를 취급하기 위한 수학적 기법

을 제공하며, (3)수학적으로 정의된 지식을 분석하거나 추론할 수 있게 한다[1,6].

러프집합에서 취급되는 불확실성에 대한 근원은 비교적 명확하다. 즉, 속성값을 사용함으로써 제한적으로 식별 가능한 객체에 의하여 발생된 애매함을 다루고 있다. 동일한 조건속성의 값을 갖는 객체들이 다른 결정 클래스(*decision class*)에 속해 있는 경우 조건속성에 관한 결정 클래스들을 정확히 기술하려는 지식베이스 시스템에서 불일치가 발생되는데, 러프집합에서는 지식베이스 시스템의 각 [객체, 속성]의 쌍이 정확하고 유일한 값을 갖는다고 가정함으로서 이러한 불일치를 해결한다. 그러나 실세계에서 이러한 [객체, 속성]이 갖는 값은 여러 가지 이유로 인하여 불확실할 수도 있다. 특히 연속적인 구간 특성을 갖는 정량적 속성의 불확실성에 대해서는 러프집합의 근사공간(*approximation space*)으로는 처리하기가 어렵다.

본 논문에서는 러프집합의 지식베이스 시스템에서 발생되는 여러 가지 불확실성 중에서 정량적 속성의 불확실성을 해결하는 방법을 제안함으로써 러프집합의 분류능력을 향상시키고자 한다.

II. 러프집합의 기본개념과 불확실성

2.1 지식베이스 시스템

지식베이스 시스템 $S = \{U, A, V\}$라 하자. 객체들의 유한집합 $U = \{x_1, x_2, \cdots, x_n\}, U \neq 0$ 이며, A 는 기본속성들의 유한집합이다.
A 에 있는 속성들은 조건속성 C 와 결정속성 D 로 분류되며, $A = C \cup D$, $V = \cup_{p \in A} V_P$, 이고, V_P는 기본속성 P의 영역이 된다.

속성들의 모든 부분집합 $P(P \subseteq A)$와 임의의 원소 x_i, $x_j \in U$ 라 하면, 식별 불가능 관계(*indiscernibility relation*)인 이진관계 $IND(P)$ 를 정의하면 다음과 같다.

$$IND(P)=\{(x_i, x_j) \in U \times U : \forall p \in P,\ p(x_i)=p(x_j)\}$$

여기서 x_i, x_j 는 지식베이스 시스템 S 에서 속성 P의 집합에 의하여 식별 불가능하다고 말한다. 그리고 $P(x)$는 객체 x에 할당된 속성 P의 값으로서, $IND(P)$는 모든 $P \subseteq A$ 에 대하여 U 에서 식별 불가능한 동치관계(*equivalence relation*)가 되며, 다음과 같은 관계가 성립된다.

$$IND(P) = \bigcap_{p \in P} IND(p)$$

지식베이스 시스템 $S = \{U, A, V\}$이고, $R \subseteq A$ 가 동치관계라면, 순서쌍 $AS = (U, R)$을 근사공간이라 한다. U의 원소 x_i에 대하여 $IND(P)$에서 x_i의 동치 클래스는 다음과 같다.

$$[x_i]_{IND(P)} = \bigcap_{R \in P} [X]_R$$

$X \subseteq U$ 라 하면, AS에서 X의 하한근사 $\underline{R}X$ 와 상한근사 $\overline{R}X$ 는 다음과 같다.

$$\underline{R}X = \{ x_i \in U \mid [x]_R \subseteq X \}$$

$$\overline{R}X = \{ x_i \in U \mid [x]_R \cap X \neq 0 \}$$

집합 $BN_R(X) = \overline{R}X - \underline{R}X$ 를 X의 R-경계(R-*boundary*)라 한다.

객체들의 집합 X에 대한 정확성 척도는 다음과 같다.

$$\alpha_R(X) = \frac{card\ \underline{R}}{card\ \overline{R}}, \quad X \neq 0$$

여기서 $0 \leq \alpha_R(X) \leq 1$이 된다. 부정확성 척도 $\rho_R(X) = 1 - \alpha_R(X)$ 는 X의 "R-대략적 정의"가 된다. 이러한 부정확성에 대한 수치적 표현은 근사 개념에 근거하고 있으며, 지식베이스 시스템의 제한된 분류 능력을 명시적으로 나타낸다.

2.2 정량적 속성의 불확실성

지식베이스 시스템에서 발생할 수 있는 여러 가지 불확실성의 유형은 다음과 같다.

(1) [객체, 속성]의 값을 알 수 없는 경우
(2) [객체, 속성]의 값이 여러 개인 경우
(3) [객체, 속성]의 값이 부정확한 경우
(4) 정량적 속성의 값이 불연속적인 경우

전문가 시스템과 같은 실세계의 응용에서 불확실한 데이터의 처리가 명확하지 못할 경우 효율적인 시스템 구성이 곤란해지고 추론 결과에 대한 신뢰성 역시 떨어지게 된다. 이에 따라 러프집합에서 불확실성을 해결하기 위한 연구가 계속되어 왔으며, 접근방법은 주로 러프집합의 근사개념과 정확성 척도를 이용하여 불확실성 모델을 일반화시키거나 부분적으로 퍼지집합 이론을 적용하였다[4].

본 논문에서는 유형 (4)에 대한 불확실성을 처리하고자 한다. 그 이유는 연속적인 구간 특성을 갖는 정량적 속성의 불확실성이 러프집합 분석을 위한 데이터 분류과정에서 가장 기본적인 문제가 됨에도 불구하고 간과되어 왔기 때문이다.

일반적으로 지식베이스 시스템에서 사용되는 속성은 정량적인 것과 정성적인 것으로 구분할 수 있다. 수치(*numeric*)를 최초 값으로 갖는 정량적 속성은 영역이 "10에서 100 사이에 분포된 값 등"과 같이 주로 구간과 관련된 부분집합이 되는 반면, 정성적 속성의 영역은 "low, medium, high, very high"와 같이 퍼지집합의 언어변수와 같은 정성적 용어의 유한집합이 된다. 정량적 속성의 값은 러프집합 분석에서 대부분 직접 사용되므로 정량적 속성들은 러프집합 분석 이전에 정성적 용어로 변환되어야 한다.

최초의 정량적 값들은 정성적 용어에 대한 보조구간(*subinterval*)으로 분리된다. 이들 보조구간을 나타내는 정성적 용어는 관련된 응용분야의 표준, 약정 및 관례에 따라 전문가에 의하여 주로 만들어지게 된다. 그러나 정성적 용어가 나타내는 보조구간들은 임의적으로 분리된 경우가 많고 보조구간들 간의 경계에는 정량적 속성값들이 중복되거나 불일치되는 등 모호한 경우가 발생되기 때문에 러프집합의 분석 결과에 영향을 줄 수 있다[5].

따라서 데이터를 분류할 때 보조구간들의 중복 가능한 경계를 보다 명확히 구분하여 정량적 속성의 불확실성을 최대한 배제하기 위한 방안이 필요하다.

Ⅲ. 러프집합의 정량적 속성값 변환

3.1 정량적 속성의 불확실성에 대한 모델화

러프집합에서 불확실성을 취급하기 위한 가장 기본적인 방법은 불연속성의 정량적 속성을 취급하는 유형 (4)의 불확실성을 모델화하는 것이다.

정량적 속성값은 "low temperature, normal temperature, … high temperature 등"과 같이 본래의 값과는 다른 수준에 해당하는 정성적 용어로 변환된다. 이러한 정량적 속성은 "low, medium, high 등"과 같은 범위기호 (*range symbol*)로 변환시킬 수도 있다. 이러한 범위기호들은 정량적 속성값들이 분산되어 있는 실제 구간(속성의 최초 영역)을 보조구간들로 분리함으로써 모델화 한다. 이때 범위기호에 주어진 보조구간은 러프집합의 근사공간에 포함되는 정량적 속성값의 집합이 된다.

러프집합의 근사공간 내에 포함된 정량적 속성값의 불확실성은 범위기호로 나타나는 보조구간들 간에 서로 중복된 경계영역으로 존재한다. 이러한 경계영역은 러프집합의 식별불가능 영역으로 존재하며 그림.1과 같다.

정량적 속성값을 정성적 속성으로 변환하기 위해서는 그림 1에서와 같이 두 집합의 경계

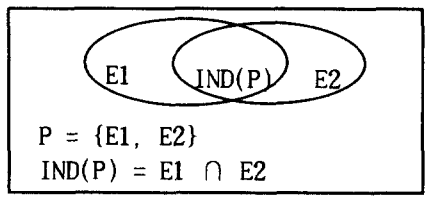

그림 1. 정량적 속성값을 갖는 두 집합간의 중복 영역

영역에서 발생된 불확실성의 데이터 즉, 식별 불가능 영역인 $IND(P)$에 속하는 교집합 영역의 데이터를 정성적 속성을 갖는 범위기호의 보조구간 영역으로 분리해야 한다. 이를 위하여 본 논문에서는 하이퍼박스(*hyperbox*) 개념을 이용하였다[2,3].

3.2 정량적 속성값 분류

본 논문에서 사용된 하이퍼박스는 러프집합에서 식별 불가능 영역의 정량적 속성값을 순환적으로 분류하는 규칙을 정의한다. 각 규칙은 활성(*active*) 하이퍼박스와 억제(*inhibit*) 하이퍼박스로 구성된다. 활성 하이퍼박스는 데이터의 존재 영역을 정의하며, 이 영역에 포함된 데이터는 정량적 속성값을 정성적 속성으로 변환시켰을 때 해당 범위기호의 보조구간에 속하는 데이터가 된다. 억제 하이퍼박스는 클래스간의 중복 데이터 영역이 되며, 활성 하이퍼박스에서 데이터의 존재를 제한하는 역할을 한다.

이러한 하이퍼박스는 순환적으로 정의된다. 우선적으로, 러프집합의 근사공간에 있는 데이터의 최소 및 최대 값을 계산함으로써 활성 하이퍼박스 영역을 결정한다. 만일 활성 하이퍼박스 영역에서 클래스 X_i와 클래스 X_j의 영역이 중복된다면, 중복 영역을 억제 하이퍼박스로 정의한다. 만일 억제 하이퍼박스에서 클래스 X_i와 X_j에 대한 데이터가 존재한다면, 이들 클래스에 대한 추가의 활성 하이퍼박스를 정의한다. 반복해서, 이들 추가 활성 하이퍼박스 사이에 중복 데이터가 다시 존재한다면, 억제 하이퍼박스의 중복 영역을 계속 정의한다. 이 방법을 이용하면 순환적으로 활성 하이퍼박스의 중복 데이터를 분할하기 때문에 다른 클래스간의 중복을 해결할 수 있다. 그 과정은 그림 2와 같다. 여기서 A_i, A_j 는 활성 하이퍼박스를, I_{ij}는 억제 하이퍼박스를 나타낸다.

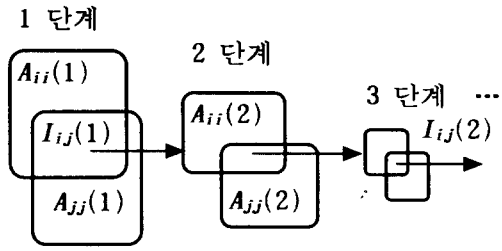

그림 2. 활성 및 억제 하이퍼박스의 순환적 정의

러프집합의 근사공간에 포함된 정량적 속성값의 집합인 X_i의 1단계 활성 하이퍼박스인 $A_{ii}(1)$을 정의하면 다음과 같다.

$$A_{ii(1)} = \{x \mid LV_{iik}(1) \le x_k \le HV_{iik}(1), \; k=1,\ldots,m\}$$

여기서, x_k 는 입력벡터 x 의 k 번째 원소를, $LV_{iik}(1)$은 $x \in X_i$ 에 대한 x_k 의 최소 값을, $HV_{iik}(1)$은 $x \in X_i$ 에 대한 x_k 의 최대 값을 나타낸다.

만일 활성 하이퍼박스 $A_{ii}(1)$과 $A_{jj}(1)$이 중복되었다면, 그림 2와 같이 $I_{ij}(1)$로 표시된 1단계의 억제 하이퍼박스를 정의한다.

$$I_{ij}(1) = \{x \mid LW_{ijk}(1) \le x_k \le HW_{ijk}(1), \; k=1,\ldots,m\}$$

$$LV_{iik}(1) \le LW_{ijk}(1) \le HW_{ijk}(1) \le HV_{iik}(1)$$

억제 하이퍼박스 $I_{ij}(1)$의 최소 및 최대 값은 다음과 같이 정의된다.

i) $LV_{jjk}(1) \le LV_{iik}(1) \le HV_{jjk}(1) < HV_{iik}(1)$
$LW_{ijk}(1)=LV_{iik}(1)$, $HW_{ijk}(1)=HV_{jjk}(1)$

ii) $LV_{iik}(1) < LV_{jjk}(1) \le HV_{iik}(1) \le HV_{jjk}(1)$
$LW_{ijk}(1)=LV_{jjk}(1)$, $HW_{ijk}(1)=HV_{iik}(1)$

iii) $LV_{jjk}(1) \le LV_{iik}(1) \le HV_{iik}(1) \le HV_{jjk}(1)$
$LW_{ijk}(1)=LV_{iik}(1)$, $HW_{ijk}(1)=HV_{iik}(1)$

iv) $LV_{iik}(1) < LV_{jjk}(1) \le HV_{jjk}(1) < HV_{iik}(1)$
$LW_{ijk}(1)=LV_{jjk}(1)$, $HW_{ijk}(1)=HV_{jjk}(1)$

X_i에 속한 어떤 데이터가 $I_{ij}(1)$에 존재한다면, $I_{ij}(1)$에 있는 데이터에 근거한 X_k 의

최소 및 최대 값을 계산함으로써 억제 하이퍼박스인 $I_{ij}(1)$ 안에 있는 $A_{ij}(2)$를 2단계의 활성 하이퍼박스로 나타낼 수 있다.

$$A_{ii}(2) = \{x \mid LV_{ijk}(2) \le x_k \le HV_{ijk}(2), \; k=1,\ldots,m\}$$

여기서, $x \in X_i$이며, x 는 $I_{ij}(1)$에서 존재한다. 그리고 $LV_{ijk}(2)$는 x_k 의 최소값, $HV_{ijk}(2)$는 x_k 의 최대 값이 된다.

$$LW_{ijk}(1) \le LV_{ijk}(2) \le x_k \le HV_{ijk}(2) \le HW_{ijk}(1)$$

활성 하이퍼박스 $A_{ij}(2)$와 $A_{ji}(2)$가 중복되었다면, $I_{ij}(2)$로 표시된 2단계의 억제 하이퍼박스로 중복된 영역을 정의한다.

$$Iij(2) = \{x \mid LW_{ijk}(2) \le x_k \le HW_{ijk}(2), \; k=1,\ldots,m\}$$

$$LV_{ijk}(2) \le LW_{ijk}(2) \le HW_{ijk}(2) \le HV_{iik}(2)$$

중복된 데이터가 여전히 남아 있을 경우 클래스간에 중복된 데이터가 존재하지 않을 때까지 위와 같은 방법을 순환적으로 계속한다. 이러한 절차에 따라 분류된 클래스에 대하여 적절한 범위기호를 부여하면 정량적 속성의 불확실한 데이터에 대한 정성적 속성 변환이 완료된다.

Ⅳ. 결 론

본 논문에서는 러프집합의 불연속적인 정량적 속성값을 정성적 속성으로 변환할 때 식별불가능 영역에서 발생되는 불확실성을 제거하는 방안을 제시하였다.

연속 구간의 특성을 갖는 정성적 속성의 범위기호에 대한 보조구간에서 인접된 보조구간과의 중복 영역을 제거하기 위하여 하이퍼박스 개념을 적용하였다.

본 논문에서의 제시된 방법은 러프집합에서 분류능력을 향상시키고, 광범위한 데이터의 분류 문제를 해결하는데 유용할 것으로 기대된다.

Ⅴ. 참고 문헌

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

[2] Shigeo Abe.,Ming-Shong Lan, "A Method for Fuzzy Rules Extraction Directly from Numerical Data and Its Application to Pattern Classification", IEEE Trans. on Fuzzy Systems, vol. 3, no. 1, pp. 18-28, Feb. 1995.

[3] Simpson, P. K., "Fuzzy min-max neural networks-Part 1: Classification," IEEE Trans. Neural Networksm, vol. 3, no. 5, pp. 776-786, Sept. 1992.

[4] Slowinski, R., and Stefanowski, J., "Handling Various Types of Uncertainty in the Rough Set Approach", Proc. Intl. Workshop on Rough Sets and Knowledge Discovery, Banff, 1993.

[5] Slowinski, R., and Stefanowski, J., "Rough classification in incomplete information systems". Mathematical and Compute. Modelling, vol. 12, no. 10/11, pp 1347-1357, 1989.

[6] Ziarko, W., "Rough Sets and Knowledge Discovery: An Overview", Proc. Intl. Workshop on Rough Sets and Knowledge Discovery, Banff, 1993.

[7] Xiaohua Hu., Nick Cercone and Jiawei Han, "An Attribute-Oriented Rough Set Approach for Knowledge Discovery in Databases", Proc. Intl. Workshop on Rough Sets and Knowledge Discovery, Banff, 1993.