The Journal of The Institute of Internet, Broadcasting and Communication (IIBC)
Vol. 14, No. 2, pp.107-113, Apr. 30, 2014. pISSN 2289-0238, eISSN 2289-0246

http://dx.doi.org/10.7236/JIIBC.2014.14.2.107
JIIBC 2014-2-15

조건부 정보엔트로피에 의한 불완전 정보시스템의 불확실성 측정

Uncertainty Measurement of Incomplete Information System based on Conditional Information Entropy

박인규*

Inkyoo Park*

요 약 러프집합에서 식별불능의 관계와 근사공간의 개념을 이용해서 의사결정표로부터 최적화된 정보를 유도하게 된다. 그러나 일반적으로 결정표에서 데이터의 중복이나 비일관성은 피할 수 없기 때문에 속성의 중요성은 지식의 감축에서 매우 중요한 개념이다. 속성의 중요성에 대한 대수학적인 정의는 도메인중의 완전한 부분집합에 대한 해당 속성이 주는 영향을 고려하는 것이고, 정보이론적인 정의는 도메인 중의 불완전한 부분집합에 대한 해당 속성이 주는 영향을 고려하는 것이다. 따라서 속성 중요성은 정보이론적인 관점의 정의와 대수학인 관점의 정의가 분명하게 차이가 있다. 본 논문에서는 정보시스템의 조건속성과 결정속성에 포함될 수 있는 정보를 최적화하기 위한 정보이론적인 척도로써 러프집합을 이용한 조건부 정보엔트로피를 제안하고 그 효용성을 보인다.

Abstract The derivation of optimal information from decision table is based on the concept of indiscernibility relation and approximation space in rough set. Because decision table is more likely to be susceptible to the superposition or inconsistency in decision table, the reduction of attributes is a important concept in knowledge representation. While complete subsets of the attribute's domain is considered in algebraic definition, incomplete subsets of the attribute's domain is considered in information-theoretic definition. Therefore there is a marked difference between algebraic and information-theoretic definition. This paper proposes a conditional entropy using rough set as information theoretical measures in order to deduct the optimal information which may contain condition attributes and decision attribute of information system and shows its effectiveness.

Key Words : Rough Set, Indiscernibility Relation, Entropy, Uncertainty, Information System

Ⅰ. 서 론

러프집합은 부정확하고, 불확실하고 그리고 애매한 정보가 처리되는 정보시스템을 위한 효율적인 이론으로써 Pawlak교수가 1982년에 러프집합 이론을 창안한 이래 많은 연구자와 실무자가 관심을 가져왔음은 부인할 수 없다. 또한 러프집합은 식별불능(indiscernibility) 관계를 기초로 하고 근사화(approximation) 공간의 개념을 이용하여 대수학적으로 집합을 정의하고 있다.[1] 이러한 정의를 바탕으로 불완전 정보를 제거하여 원래의 데이터와 동일한 결과를 보장하여 감축된 데이터를 유도하는 토대를 제공한다. 이로 인해 데이터 마이닝, 기계학습, 지식

*정회원, 중부대학교 컴퓨터학과
접수일자 2014년 2월 26일, 수정완료 2014년 3월 30일
게재확정일자 2014년 4월 11일

Received: 26 February, 2014 / Revised: 30 March, 2014
Accepted: 11 April, 2014
*Corresponding Author: e-mail@fip2441g@gmail.com
Dept. of Computer Science, Joongbu University, Korea

발견, 데이터베이스 쿼리와 패턴인식과 같은 부정확한 지식에 대한 표현 및 추론 등의 연구에 사용되고 있다.[2]

임의의 대상에 대한 정보시스템을 구성하는 경우에 대상의 원소를 임의의 속성에 따라서 분류할 경우에는 데이터의 중복이나 기타 비 일관적인 데이터가 수반되기 마련이기 때문에 식별불능성에 의한 불확실성이 발생하게 된다. 따라서 각각의 동치류에 정확하게 하나의 원소가 분할된 경우에는 불확실성이 존재하지 않는다. 그러나 그렇지 않은 분할에서는 정보의 손실이 발생하게 된다. 따라서 이와 같은 불확실성은 하한근사와 상한근사를 통한 정확성의 척도나 부정확성의 척도에 의하여 대수학적으로 모델링될 수 있다. 그러나 분할된 동치류들이 동일한 부정확성의 척도를 가지고 있을지라도 동치류의 크기(cardinality)에 의해 불확실성이 다르게 나타난다. 결국 러프집합에서는 임의의 속성에 의해 발생된 동치류들에 의한 불확실성과 누락된 데이터의 처리와 같은 모든 불완전 정보를 대수학적으로 처리를 다루기에는 부족하다.

본 논문에서는 정보시스템에서 유용한 정보를 추출하기 위한 속성 값의 감축과정에서 발생되는 불완전성(incompleteness)을 해결하기 위한 정보 이론의 척도로써 러프 엔트로피(rough entropy)를 확장하여 조건부 정보엔트로피(conditional information entropy)를 제시한다.

Ⅱ. 러프집합과 정보 엔트로피

1. 러프집합

지식 표현 시스템은 쌍 $S=(U, A)$로 나타낸다. U는 공집합이 아닌 유한집합이며 전체집합이라 하고, A는 공집합이 아닌 원시속성(primitive attribute)들의 유한집합으로 조건속성 C와 결정속성 D로 구성된다. 속성 $B \subseteq A$의 모든 부분집합에 대하여 $V=\cup_{a\in A} V_B$이고 V_a는 원시속성 B의 영역이 된다. 식(1)과 같은 이진 관계 $IND(B)$를 정의 할 수 있으며 이를 식별 불가능관계(indiscernible relation)라 한다.[3]

$$ND\ B) = \{(x,y) \mid a(x) = a(y),\ \forall a \quad B\} \qquad (1)$$

식(2)와 같이 $IND(B)$는 동치 관계이다.

$$IND(B) = \qquad IND(a) \qquad (2)$$

$S=(U, A)$에서 $X \subseteq U$와 동치관계 $B \subseteq A$에 의해서 X의 하한근사(lower approximation)와 상한근사(upper approximation)는 각각 식(3)과 식(4)와 같이 $\overline{B}X$, $\underline{B}X$로 나타낸다.

$$B\,X = \{x_i \in U \mid [x]_B \subseteq X\} \qquad (3)$$

$$\overline{B}\,X = \{x_i \in U \mid [x]_B \cap X \neq \varnothing\} \qquad (4)$$

$BND(X)=\overline{B}X - \underline{B}X$를 X의 B-경계라고 하고, 임의의 집합이 부정확하다는 것은 경계영역이 존재하기 때문이고 경계영역이 커질수록 그 집합의 정확성은 떨어진다. 이 개념은 대수학적으로 식(5)와 같이 정확성 척도(accuracy measure)로 정의할 수 있다.

$$\alpha_B(X) = |\underline{B}\,X| / |\overline{B}\,X|,\ X \neq \varnothing \qquad (5)$$

여기서 $0 \leq \alpha_B(X) \leq 1$이 되고, 집합 X에 대한 지식의 불완전성의 정도는 식(6)과 같이 부정확성 척도(inaccuracy measure)로 정의 할 수 있다.

$$\rho_B(X) = 1 - \alpha_B(X) \qquad (6)$$

( )는 러프집합의 경계영역으로부터 발생되는 불완전성을 알기 위한 좋은 방법이지만, 부정확성 척도를 계산하기 위해서는 러프집합의 대수학적인 접근은 한계가 있다. 따라서 임의의 속성에 따른 불확실한 동치류에 대한 영향을 정보이론적인 관점에서 접근해야 할 필요성이 있다.

2. 정보 엔트로피

어떤 사건의 확률을 알고 있을 때 그에 대한 정보량을 측정할 수 있다. 어떤 통계량에서 각 미시적 상태 i의 확률을 p_i로 정의할 경우에 N개로 구성된 앙상블의 엔트로피 H는 다음과 같이 정의할 수 있다.[4]

$$H = -\sum_{i}^{n} p_i \ln p_i \qquad (7)$$

임의의 전체집합 U에 속하는 객체들은 임의의 집합 X와 값 V_X에 대하여 서로 다른 부분집합으로 분류 될 수 있다. $x \in V_x$ 에 의하여 정의되는 x의 부분집합을 정의 할 수 있다. 따라서 정보시스템은 하나의 통계적인 집단에 해당하고 X는 그에 따른 통계변수로 볼 수 있다. 따라서 X의 확률분포 $H(P(x))=E_{p(x)}[-ln(p(x))]$으로 $Ep(x)$는 X 확률에 대한 기대치이고 $p(x)=0$이면 $p(x)ln(p(x))=0$이다. $H(p(x))$를 $H(x)$로 나타내면 엔트로피는 $H(x) \geq 0$이고 속성집합 X에 대한 정보량 혹은 불확실성의 척도로 볼 수 있다. $p(x)$가 균일한 분포 즉, $p(x)=1/|V_x|$이고 $x \in V_x$에 대하여 엔트로피는 최대값 $ln|V_x|$를 가지며, P가 특정한 값 x_0 즉, $p(x_0)=1$이거나 $p(x_0)=0$는 엔트로피의 최소값으로 정보이론에서 엔트로피는 확률변수의 불확실성의 척도로 볼 수 있다.

따라서 지식에 존재하는 속성들은 식별불가능성에 의한 동치류의 서로 다른 크기로 인한 불확실성이 존재하기 마련이다. $P(X,Y)$는 X와 Y의 결합 확률분포(joint probability distribution)이고 조건부 확률분포로서 결국 $H(X|Y)$는 확률분포 $P(Y)$에 대하여 부집단 엔트로피 $H(X|Y)$의 기대치로 정의할 수 있다. 본 논문에서는 정보이론적인 접근 방법으로 조건속성 Y_j에 대한 결정속성 X_i의 조건부 정보엔트로피는 식(10)과 같이 정의할 수 있다.

$$\begin{aligned} X|Y_j) &= -\sum_{y} p(y) \sum_{x \in V_x} p(x|y) ln\,(p(x|y)) \quad (10) \\ &= -\sum_{y \in V_y} \sum_{x \in V_x} p(x,y) ln\,(p(x|y)) \\ &= -E_{P(X,Y)}\,[ln\,P(X,Y)] \\ &= -\sum_{j=1}^{n} P(Y_j) \sum_{i=1}^{m} P(X_i) ln P(X_i|Y_j) \end{aligned}$$

여기서 $=1,..,m$, $j=1,..,n$까지의 동치클래스의 개수이고, X_i와 Y_j는 조건속성과 결정속성에 대한 동치클래스이다. $|X_i \cap Y_j|/|Y_j|$는 Y_j에 대한 $(Y_j \cap X_i)$의 확률을 나타내며 러프집합 X_i와 공집합이 아닌 Y_j의 동치클래스와의 크기를 Y_j의 동치 클래스의 크기로 나눈 값에 해당한다. $(X_i \cap Y_j)/(Y_j)$는 Y_j는 전체집합에서 동치클래스 j에 있는 원소들의 수를 모든 동치 클래스들의 전체 원소들의 수이고, U는 전체집합의 수이다.

조건부 정보엔트로피에 의해서 동치류에 분포되어지는 granule의 특성을 정보이론적인 측면에서 러프집합 U={1,2,3,4,5,6,7}에서 임의의 동치류 $U/IND(E_1)$={{1,2,3,4},{5,6,7}}, $U/IND(E_2)$={{1,2},{3,4},{5,6,7}}, $U/IND(E_3)$={{1},{2},{3},{4},{5,6,7}}일 경우를 고려하자. 러프집합 X={1,4,5}에 대하여 식(10)에 의하여 동치관계 E_1, E_2, E_3의 조건부 정보엔트로피 $H(E|X)$는 다음과 같다.

$$\begin{aligned} H(E_1|X) &= H(\{1,2,3,4\}|X)+H(\{5,6,7\}|X) \\ &= -(4/7*3/7)ln(2/4))-(3/7*3/7)ln(1/3)) \\ &= 0.372 \\ H(E_2|X) &= H(\{1,2\}|X)+H(\{3,4\}|X)+H(\{5,6,7\}|X) \\ &= -(2/7*3/7)ln(1/2))-(2/7*3/7)ln(1/2)) \\ &\quad -(3/7*3/7)ln(1/3)) \\ &= 0.372 \\ H(E_3|X) &= H(\{1\}|X)+H(\{2\}|X)+\mathrm{H}(\{3\}|X) \\ &\quad +H(\{4\}|X)+H(\{5,6,7\}|X) \\ &= -(1/7*3/7)ln(1/1)-(1/7*3/7)ln(0/1) \\ &\quad -(1/7*3/7)ln(0/1)-(1/7*3/7)ln(1/1) \\ &\quad -(3/7*3/7ln(1/3)) \\ &= 0.324 \end{aligned}$$

E_1과 E_2는 X에 대한 각각의 동치류의 불확실성이 동일하게 나타났다. 이는 X={1,4,5}에 대하여 E_1과 E_2는 등가라는 것을 나타낸다. 다시 말해서 E_1={1,2,3,4}가 E_2={{1,2}, {3,4}}에 비해서 $P(E_1)$=0.5, $P(E_2)$=(0.5+0.5)/2 = 0.5로 확률이 같은 경우에 조건 엔트로피의 변화를 야기하지 않는다. $H(\{E_1\}/X)=H(\{E_2\}/X)$이면 합병된 등가의 동치류는 결정속성에 비해서 확률이 같다. 결국 E_3가 가장 안정적인 속성을 나타낸다고 할 수 있다. 계산된 결과를 통하여 조건부 정보엔트로피는 대수학적인 정의와 다르게 차이가 발생한다는 것을 알 수 있다. 따라서 조건부 정보엔트로피를 통하여 정보시스템의 동치류에 의한 불확실성을 측정할 수 있기 때문에 비일관적인 (inconsistent) 정보의 제거에 매우 유용하다고 할 수 있다.

Ⅲ. 불완전 정보시스템 모델링

1. 불완전정보시스템

정보 시스템 S={U, A}에서 속성들의 집합이 $A=C \cup \{d\}$이고, 결정속성(decision attribute)이 $d \notin C$일 경우에, $a \in C$에 대하여 $a : U \rightarrow V_a$를 만족하는 C를 조건속성(condition attribute)이고, V_a는 속성 값이다. $\forall a, d \in A$

일 때 객체 $(x, y) \in U \times U$에서 불완전 정보시스템(incomplete information system)의 조건은 $(a(x)=a(y)) \wedge (a(x) \neq a(y)), (a(x)=null) \vee (a(y)=null) \vee (d(x)=null) \vee (d(y)=null)$에 해당한다.[5]

Beaubouef와 Kryszkiewicz의 불완전 정보시스템의 제한은 $a(x) = a(y) \vee a(x) = null \vee a(y) = null$로써 속성의 동치관계에 대한 불완전성이 고려되고, 불완전 정보시스템의 원시정보를 변경하지 않으면서 지식을 감축을 이용하여 불완전한 정보를 제거하였다. 그러나 실제로는 조건속성에서 발생할 수 있는 불완전성은 결정속성에서도 동일하게 발생될 수 있으므로 결정속성의 불완전성에 대한 처리 방법도 고려되어야 한다.[6] 조건속성과 결정속성의 불완전성으로 인한 모든 경우를 고려하여 원시정보의 불완전성을 극복하는 방법을 조건부 정보엔트로피를 이용하여 제시한다.[7]

2. 결정속성의 조건부 엔트로피

조건부 정보 엔트로피 $H(X)$를 이용하여 *null*인 결정속성에 대하여 해당 객체의 조건속성의 조건부 엔트로피를 계산함으로써 결정속성의 값을 결정할 수 있고, 조건속성 값은 동일하나 결정속성 값이 불일치하는 결정부 속성 값도 결정할 수 있다. 따라서 러프집합의 객체 $x \in U$에 대한 결정속성의 조건부 정보엔트로피 $H_d(X|Y)$는 식(9)와 같이 정의할 수 있다.

$$X_i|Y_j) = -\sum_{j=1}^{n} \frac{|Y_j|}{|U|} \sum_{i=1}^{m} \frac{|X_i|}{|U|} ln \frac{|X_i \cap Y_j|}{|Y_j|} \quad (9)$$

여기서, $i=1,..,n$, $j=1,..,m$ 이다.

(x_t, d)의 결정속성 값 $a_d(x_t)$가 *null*이거나 또는 조건속성의 값이 동일하나 결정속성의 값이 다른 불일치 값일 때 대체되는 결정 속성 값은 결정부 속성의 조건부 엔트로피의 정의에 의해서 다음과 같이 결정된다.

(1) 가능한 결정속성 값$(a_d(x_1),..,a_d(x_{t-1})$에서 *null* 및 불일치하는 속성 값을 가지는 $a_d(x_t)$에 순차적으로 적용하여 $U/D=\{d_1,d_2,...,d_n\}$을 구한다.

(2) 조건부 속성의 개수 $j=1,...,m$과 동치류의 개수 $k=1,...,l$ 대하여 결정부 속성의 조건부 엔트로피 $\sum_j\sum_k(H_d(C_{jk}|d_1)),\ \sum_j\sum_k(H_d(C_{jk}|d_2)),..,\ \sum_j\sum_k(H_d(C_{jk}|d_n))$을 구한다.

(2) $\sum_j\sum_k(H_d(C_{jk}|d_1)),\ \sum_j\sum_k(H_d(C_{jk}|d_2)),..,\ \sum_j\sum_k(H_d(C_{jk}|d_n))$에서 j와 k에 대한 평균치가 *null*값을 대체할 수 있는 결정속성 값인 결정속성의 조건부 엔트로피가 된다.

(3) $a_d(x_t) = min\{mean(\sum_j\sum_k(H_d(C_{jk} \mid d_1))), mean(\sum_j\sum_k(H_d(C_{jk}|d_2))),...,mean(\sum_j\sum_k(H_d(C_{jk}|d_n)))\}$에 의하여 결정된 값으로 결정속성 $a_d(x_t)$의 *null*값 또는 불일치 값을 대체한다.

3. 조건속성의 조건부 엔트로피

조건속성의 조건부 엔트로피는 결정속성의 조건부 엔트로피 $H_d(X|Y)$에 기반을 두고 있으며 조건 속성 값이 *null*일 경우에 해당 객체의 *null*조건 속성 값만 계산함으로써 효과적으로 *null*값을 대체할 수 있다. 러프집합의 객체 $x \in U$에 대한 조건속성의 조건부 엔트로피 $H_c(X|Y)$는 식(10)과 같이 정의 할 수 있다.

$$H_c(X_i|Y) = -\frac{|Y|}{|U|} \sum_{j=1}^{n} \frac{|X_i|}{|U|} ln \frac{|X_i \cap Y|}{|Y|} \quad (10)$$

(x_t, a_c)의 결정속성 값 $a_c(x_t)$가 *null* 값일 때 대체되는 조건속성 값은 속성 관계 엔트로피를 이용하여 다음과 같이 결정한다.

(1) *null*값을 가지는 객체의 결정부 속성의 동치류 d와 *null*값을 가지는 조건속성 C_j의 동치류를 구한다.

(2) 결정부 속성의 동치류 d_j에 대하여 조건속성 C_i의 k개의 동치류의 조건부 정보엔트로피 $H_c(C_{i1}|d_j)$, $H_c(C_{i2}|d_j),...,H_c(C_{ik}|d_j)$를 구한다.

(3) $a_c(x_t) = min\{ H_c(C_{i1} \mid d),\ H_c(C_{i2}|d),\ ...,\ H_c(C_{ik}|d)\}$에 의하여 결정된 값으로 조건속성 $a_c(x_t)$의 *null*값 또는 불일치 값을 대체한다.

정보엔트로피 알고리즘은 객체관계 엔트로피와 속성관계 엔트로피의 처리절차를 결합한 것으로 정보시스템을 분석하여 조건속성과 결정속성의 불완전성에 의해서 객체관계 엔트로피와 속성관계 엔트로피를 구한다. 즉, 조건속성의 *null*속성 값에 대한 엔트로피를 계산하거나 또는 결정속성의 *null* 및 불일치 속성 값에 대한 엔트로피를 계산한다. 그리고 엔트로피결과를 비교하여 해당 불완전 속성값을 최적의 속성 값으로 대체시키는 속성값을 결정하는 기능을 수행한다.[8,9,10]

Ⅳ. 적용사례

어떤 게임에 등장하는 몬스터의 상태 천이규칙에 관한 정보들을 판단하는 정보시스템을 표 1과 같이 구성하였다. 조건속성에 해당하는 입력조건은 current, input과 weapon의 3개로 구성하였다. 입력조건에 의해서 몬스터의 다음상태인 결정속성이 결정되도록 하였다. 표 1의 몬스터의 상태 천이규칙의 의사결정표를 불완전 정보시스템으로 구성하기 위하여 표 2와 같이 표의 데이터를 범주형 데이터로 코드화하고 불완전 정보를 포함시켰다. 표 2에서 조건속성은 $\{c_1,c_2,c_3\}$항목이고 결정속성은 $\{d\}$항목이며, 객체는 10개의 항목으로 구성되어 있다. 그리고 결정속성 $\{x_{i10}, d\}$의 속성 값에 *null*이 포함되어 있는 불완전 정보시스템으로 가정한다면, $\{x_{i10}, d\}$의 *null*값을 대체할 수 있는 결정속성 값을 조건부엔트로피에 의하여 구할 수 있다.

1. 몬스터의 상태천이표
Table 1. State transition table for monster

index	current(c1)	input(c2)	weapon(c3)	output(d)
x_1	uncomfort	monster hurt	small	anger
x_2	uncomfort	player attack	medium	uncomfort
x_3	anger	monster remedy	small	anger
x_4	normal	player attack	large	uncomfort
x_5	uncomfort	player attack	large	uncomfort
x_6	anger	monster hurt	medium	anger
x_7	normal	player attack	large	uncomfort
x_8	normal	monster remedy	medium	uncomfort
x_9	anger	monster hurt	large	anger
x_{i10}	anger	monster remedy	large	anger

1. 결정속성의 조건부엔트로피

결정속성 $\{d\}$에서 결정속성 값의 유형은 {'2', '3'}이므로 $H_d(A$/output='2')과 $H_d(A$/output='3')의 조건부엔트로피를 계산한 다음, 그 결과를 비교하여 $\{d_{i10}\}$의 값 $H(x_{i10})$을 결정한다. 단, 상한 및 하한 근사를 구하는 것은 집합 X와 식별불능 관계 식별불가능 관계에 의하여 쉽게 산출될 수 있다. 또한 c_1과 d의 범주는 normal(1), uncomfort(2), anger(3)이고, c_2의 범주는 player attack(1), monster remedy(2), monster hurt(3)이고 c_3의 범주는 large(1), medium(2), small(3)이다.

표 2. NULL 결정 속성을 가지는 의사결정 표
Table 2. Decision table with NULL decision attribute

index	current(c1)	input(c2)	weapon(c3)	output(d)
x_1	2	3	3	3
x_2	2	1	2	2
x_3	3	2	3	3
x_4	1	1	1	2
x_5	2	1	1	2
x_6	3	3	2	3
x_7	1	1	1	2
x_8	1	2	2	2
x_9	3	3	3	3
x_{i10}	3	2	3	*null*

(1) $X_2=\{x_2, x_4, x_5, x_7, x_8, x_{i10}\}$에 대하여
H_d(current/output='2'):
$c_1(1)->d(1)=-3/10*6/10In(3/3)=0$
$c_1(2)->d(1)=-3/10*6/10In(2/3)=0.073$
$c_1(3)->d(1)=-4/10*6/10In(1/4)=0.333$
H_d(input/output='2'):
$c_2(1)->d(1)=-4/10*6/10In(4/4)=0$
$c_2(2)->d(1)=-3/10*6/10In(2/3)=0.073$
$c_2(3)->d(1)=-3/10*6/10In(0/3)=0.180$
H_d(weapon/output='2'):
$c_3(1)->d(1)=-3/10*6/10In(3/3)=0$
$c_3(2)->d(1)=-3/10*6/10In(2/3)=0.073$
$c_3(3)->d(1)=-4/10*6/10In(1/4)=0.333$
H(A/Output='2')=0.135+0.084+0.135=0.188

(2)$X_3=\{x_1, x_3, x_6, x_9, x_{i10}\}$에 대하여
H_d(current/output='3'):
$c_1(1)->d(1)=-3/10*5/10In(0/3)=0.150$
$c_1(2)->d(1)=-3/10*5/10In(1/3)=0.165$
$c_1(3)->d(1)=-4/10*5/10In(4/4)=0$
H_d((input/Output='3'):
$c_2(1)->d(1)=-4/10*5/10In(0/4)=0.2$
$c_2(2)->d(1)=-3/10*5/10In(2/3)=0.061$
$c_2(3)->d(1)=-3/10*5/10In(3/3)=0$
H_d(weapon/output='3'):
$c_3(1)->d(1)=-3/10*5/10In(0/3)=0.150$
$c_3(2)->d(1)=-3/10*5/10In(1/3)=0.165$
$c_3(3)->d(1)=-4/10*5/10In(4/4)=0$
$H(A$/Output='3')=0.105+0.087+0.105=0.099

결정부 속성의 조건부 정보엔트로피의 계산결과에 의하여 작은 값이 더 안정적이므로 {*d*}에 대한 x_{10}의 속성 값은 속성 값 '3'로 결정된다. {*d*}에 대한 x_{10}의 대체 속성 값의 정확성 여부는 표 1과 비교해 보면 알 수 있다. 즉, 두 표의 속성 값이 모두 '3'이므로, 조건부엔트로피의 수식과 계산결과는 정확하다고 볼 수 있다.

2. 조건속성의 조건부엔트로피

표 3은 표 2에서 사용된 표와 같으나, 조건 속성{x_6, *b*}의 속성 값에 *null*이 포함되어 있다고 가정하면, {x_6, *b*}의 가능한 조건속성 값을 조건부엔트로피에 의하여 구하게 된다.

3. NULL 조건 속성을 가지는 의사결정 표
Table 3. Decision table with NULL condition attribute

index	current(c1)	Input(c2)	Weapon(c3)	Output(d)
x_1	2	3	3	3
x_2	2	1	2	2
x_3	3	2	3	3
x_4	1	1	1	2
x_5	2	1	1	2
x_6	3	3	2	3
x_7	1	1	1	2
x_8	1	2	2	2
x_9	3	*null*	3	3
x_{10}	3	2	3	3

조건속성 {c2}에서 가능한 조건속성 값의 유형은 '1','2','3'이고, x9의 결정속성 값의 유형이 '3'이므로 Hc(c2(1)/'3'), Hc(c2(2)/'3')및 Hc(c2(3)/'3')의 조건부엔트로피를 계산한 다음, 그 결과를 비교하여 c2에 대한 x9의 속성 값을 결정한다.

Hc(input/output='3'):

(1) Hc(c2(1)/'3')=-5/10*5/10ln(1/5)=0.402

(2) Hc(c2(2)/'3')=-4/10*5/10ln(3/4)=0.058

(3) Hc(c2(3)/'3')=-3/10*5/10ln(3/3)=0

조건부 속성의 정보엔트로피의 계산결과를 토대로 작은 값이 영향력이 많은 속성이므로, c2에 대한 x9의 속성 값은 '3'으로 결정된다. c2에 대한 x9의 대체속성 값의 정확성 여부는 표 1과 비교해 보면 알 수 있다. 즉, 두 표의 속성 값이 모두 '3'이므로, 조건부엔트로피의 수식과 계산결과는 정확하다고 할 수 있다.

Ⅴ. 결 론

러프집합의 식별불능성에 대한 불확실성에 대하여 부정확성이라는 대수학적인 척도와 정보이론적인 엔트로피 척도를 비교분석하였다. 속성의 정보가 불완전하거나 일부의 데이터가 상충되어 비 일관적인 경우와 같은 불완전 정보시스템을 완전 정보시스템으로 전환시키기 위하여 러프집합 기반으로 정보이론의 척도인 엔트로피를 변형하여 조건부엔트로피를 제안하였다. 그리고 조건부엔트로피는 불완전 정보시스템의 조건속성과 결정속성에 포함될 수 있는 속성 값을 대체 가능한 값으로 결정될 수 있으며, 속성간의 가장 영향력이 있는 분할 속성을 추출하여 효율적인 분할을 할 수 있다. 또한, 조건부엔트로피는 지식베이스를 구성하기 위한 추론규칙의 정확성을 향상시키기 위한 전처리과정으로 수행될 수 있다.

References

[1] Kryszkieqicz, M., "Rules in incomplete information systems". Information Science, Vol. 113, No. 3-4, pp. 271-292, 1998.

[2] Lin, T. Y. ,and Cercone, N.(eds), "Rough sets and data mining-analysis of imperfect data", Boston:Klumer Academic Publishers, 1997

[3] Slowinski, R. and Stefanoqski, J., "Rough classification in incomplete information systems", Mathematical and Compute, Modeling, Vol. 12,, No. 10-11, pp.1347-1357, 1989

[4] Shannon, C., L., "The mathematical theory of communication", Bell System Technical Journal, Vol. 27, 1948

[5] Beaubouef, T., Petry, F. E. and Arora, G.,, "Information-theoretic measures of uncertainty for rough sets and rough relational databases", Information Science, Vol. 109, No. 1-4, pp. 185-195, 1998.

[6] Grzymala-Busse, J., "Knowledge Acquisition under Uncertainty-a Rough Set Approach. Journal of Intelligent and Robotic Systems, Vol. 1, pp.3-16, 1988

[7] KukBoh Kim, GuBeom Jeong and KyungOk Park, "The Study on Information-Theoretic Measures of Incomplete based on Rough Sets", Institute of Korean Multimedia Society, Vol. 3 No. 5, pp. 550-556, 2000

[8] Lin Sun, "Decision Table Reduction Method Based on New Conditional Entropy for Rough Set Theory", International Workshop on Intelligent Systems and Applications, pp. 23-24, May 2009

[9] J. Y. Kim, S S. Jo, K.K. Kim, S. H. Choi, Development of Localization and Three-dimensional hull map creation S/W for Underwater robot, Journal of Korean Institute of Information Technology, Vol.8 No.6 ,35-40, June 2010

[10] J. E. Chung, J. K. Ahn, A Study of Robust Design of FCM Gasket Using Taguchi Method, Journal of the Korea Academia-Industrial cooperation Society, v.14, no.7, 3177-3183, July 2013

소개

인 규(정회원)

- 1987년 : 연세대학교 공학석사
- 1997년 : 원광대학교 공학박사
- 1997년 ~ 현재 : 중부대학교 컴퓨터학과 교수
- 관심분야: 데이터마이닝, 소프트 컴퓨팅, 러프집합, 퍼지집합, 신경회로망