The Journal of The Institute of Internet, Broadcasting and Communication (IIBC)
Vol. 14, No. 6, pp.47-54, Dec. 31, 2014. pISSN 2289-0238, eISSN 2289-0246

http://dx.doi.org/10.7236/JIIBC.2014.14.6.47
JIIBC 2014-6-8

베이지언 정보엔트로피에 의한 불완전 의사결정 시스템의 불확실성 향상

Uncertainty Improvement of Incomplete Decision System using Bayesian Conditional Information Entropy

최규석*, 박인규**

Gyoo-Seok Choi*, In-Kyu Park**

요 약 러프집합을 구성하는 식별불가능 관계를 표현하는 정보시스템에서 데이터의 중복이나 비일관성은 피할 수 없기 때문에 속성의 감축은 매우 중요하다. 러프집합이론에 있어서 일관적인 정보시스템과 비일관적인 정보시스템의 속성감축의 차이를 극복하고 자, 본 연구에서는 조건 및 결정속성에 대한 상관분석에 베이지언 사후확률을 적용한 새로운 불확실성 척도와 속성감축 알고리즘을 제안한다. 정보시스템의 불확실성에 대하여 제안된 척도와 기존의 조건부 정보엔트로피 척도를 비교해 본 결과, 정보시스템의 조건속성과 결정속성의 상호정보를 이용하여 속성간의 불확실성을 측정하는데 있어 제안된 방법이 조건부 정보엔트로피에 의한 방법보다 정확성이 있음을 보여준다.

Abstract Based on the indiscernible relation of rough set, the inevitability of superposition and inconsistency of data makes the reduction of attributes very important in information system. Rough set has difficulty in the difference of attribute reduction between consistent and inconsistent information system. In this paper, we propose the new uncertainty measure and attribute reduction algorithm by Bayesian posterior probability for correlation analysis between condition and decision attributes. We compare the proposed method and the conditional information entropy to address the uncertainty of inconsistent information system. As the result, our method has more accuracy than conditional information entropy in dealing with uncertainty via mutual information of condition and decision attributes of information system.

Keywords : Rough Set, Indiscernibility Relation, Conditional Information Entropy, Uncertainty, Bayesian Theory

Ⅰ. 서 론

데이터마이닝(data mining)은 데이터베이스나 정보저장소에 있는 대량의 데이터에서 유용한 지식을 찾고자 하는 요구에 따라 많은 관심을 받고 있으며, 최근 빅 데이터에서 지식을 발견하기 위한 연구가 활발하게 이루어지고 있다. 지식발견에 있어서 지식감축(knowlwdge reduction), 개념계층(concept hierarchy), 결정트리(decision tree)에 의한 규칙 귀납과 러프집합(rough set)에 대한 연구가 진전되어 왔다[1].

그 중에서도 러프집합은 부정확하고, 불확실하고 그리고 애매한 정보가 처리되는 정보시스템을 위한 효율적인

*종신회원, 청운대학교 컴퓨터학과
**정회원, 중부대학교 컴퓨터학과 (교신저자)
접수일자: 2014년 11월 11일, 수정일자: 2014년 12월 9일
게재확정일자: 2014년 12월 12일

Received: 11 November, 2014 / Revised: 9 December, 2014
Accepted: 12 December, 2014
**Corresponding Author: ikpark@joongbu.ac.kr
Dept. of Media & Software Engineering, Joongbu University, Korea

이론으로써 많은 유용한 결과를 가져 왔다.

또한 러프집합은 식별불능(indiscernibility) 관계를 기초로 하고 근사화(approximation) 공간의 개념을 이용하여 대수학적으로 집합을 정의하고 있다. 이러한 정의를 바탕으로 불완전 정보를 제거하여 원래의 데이터와 동일한 결과를 보장하여 감축된 데이터를 유도하는 토대를 제공한다. 이로 인해 데이터 마이닝, 기계학습, 지식 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(inconsistent) 데이터가 수반되기 때문에 식별불능성에 의한 불확실성(uncertainty)이 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 하한근사(lower approximation)와 상한근사(upper approximation)와)를 통한 정확성의 척도나 부정확성의 척도에 의하여 대수학적으로 모델링될 수 있다. 그러나 분할된 동치류(equivalence class)들이 동일한 부정확성의 척도를 가지고 있을지라도 동치류의 크기에 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 정보 이론의 척도로써 러프 엔트로피(rough entropy)를 확장하여 베이지언(Bayesian) 정보엔트로피(information entropy)를 적용함으로써 정보시스템에서 유용한 정보를 추출하기 위한 속성 값의 감축과정에서 발생되는 불확실성을 해결하여 시스템의 불완전성(incompleteness)을 향상시키는 방안을 제안한다.

II. 관련연구

1. 러프집합

임의의 대상의 유한집합인 전체집합에서 임의의 부분집합을 개념 또는 범주라고 하고 이들의 집합을 지식이라고 가정할 경우에, 객체들로 이루어진 어떤 범주(category)는 사용가능한 집합으로 정확하게 나타낼 수 없다. 이는 결국 집합을 통한 근사화를 통해서 가능하다. 임의의 저장소에 대한 지식을 표현하는 시스템은 U가 유한집합인 전체집합이고, 부분집합 X$\subseteq U$와 U의 식별 불가능한(indiscernible) 객체로 구성된 동치관계 $R\in IND(K)$을 때 $K=(U, \mathbf{R})$로 볼 수 있다. $\mathbf{P} \subseteq \mathbf{R}$ 이고 $\mathbf{P} \neq \varnothing$이면 $\cap \mathbf{P}$ 도 역시 동치관계가 되며, 이를 $IND(\mathbf{P})$라고 하고 $[x]_R$이 원소 $x\in U$를 포함하는 범주를 나타낼 경우 다음과 같이 정의한다.

$$[x]_{IND(B)} = \bigcap_{a\in P}[x]_R \tag{1}$$

$K=(U, \mathbf{R})$의 근사화를 위하여 R-하한근사와 R-상한근사를 다음과 같이 정의할 수 있다.

$$R\_X = \{x_i \in U \mid [x]_R \subseteq X\} \tag{2}$$

$$R^{-}X = \{x_i \in U \mid [x]_R \cap X \neq \varnothing\} \tag{3}$$

하한근사는 지식 R내에서 X의 원소로 확실하게 분류되는 U의 모든 원소들의 집합이고, 상한근사는 지식 R내에서 X의 원소로 분류될 가능성이 있는 U의 원소들의 집합이다. 그리고 R_X를 R에 대한 X의 긍정영역(positive region), $U-R^{-}X$를 부정영역(negation region) $NEG_R(X)$, $BN_R(X)=R^{-}X - R_X$를 경계영역(boundary region)라고 한다. 임의의 집합이 부정확하다는 것은 경계영역이 존재하기 때문이고 경계영역이 커질수록 그 집합의 정확성은 떨어진다. 이 개념은 대수학적으로 정확성의 척도(accuracy measure)를 다음과 같이 정의할 수 있다.

$$\alpha_R(X) = |R\_X|/|R^{-}X|, X \neq \varnothing \tag{4}$$

여기서 $0\leq a_R(X)\leq 1$이 되고, 집합 지식 X의 불확실성을 나타내는 부정확성의 척도(inaccuracy measure)를 다음과 같이 정의 할 수 있다.

$$\rho_R(X) = 1 - \alpha_R(X) \tag{5}$$

$\rho_R(X)$는 러프집합의 경계영역으로부터 발생되는 불확실성을 알기 위한 좋은 방법이지만, 식(5)의 대수학적인 부정확성 척도는 동치류의 크기에 따른 변별력을

완전하게 수용하지 못한다. 따라서 동치류간의 불확실성의 정확성을 향상시키기 위한 방안으로 정보이론적인 관점과 베이지언 확률의 관점에서 접근해야 할 필요성이 있다.

2. 정보 엔트로피

어떤 결과 값의 발생 가능도가 작아질수록 그 정보량은 커지고, 더 자주 발생할수록 그 정보량은 작아진다. 즉, 어떤 사건의 확률을 알고 있을 때 그에 대한 정보의 양을 측정할 수 있다. 어떤 통계량에서 각 상태 i의 확률을 p_i로 정의할 경우에 N개로 구성된 앙상블의 엔트로피 H는 다음과 같이 정의할 수 있다[4].

$$H = -\sum_{i}^{n} p_i \ln p_i \tag{6}$$

이와 같은 정보 엔트로피는 데이터베이스, 패턴인식, 의사결정 시스템과 같은 분야에서 불확실한 정보의 양을 측정하기 위하여 사용되고 있으며, 정보 엔트로피를 불확실성의 개념으로 보면 불확실성이 높아질수록 정보의 양은 더 많아지고 엔트로피는 더 커진다. 따라서 조건부 확률분포는 임의의 확률분포에 대한 엔트로피의 기대치로 정의할 수 있다. 결국, 정보 이론적인 접근 방법으로 조건속성 Y_j에 대한 결정속성 X_i의 조건부 확률 정보 엔트로피는 다음과 같이 정의할 수 있다[5].

$$H(X_i|Y_j) = -\sum_{j=1}^{n} P(Y_j) \sum_{i=1}^{m} P(X_i) \ln P(X_i|Y_j) \tag{7}$$

여기서 i=1,..,m, j=1,..,n까지의 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. 다음으로 동치류의 분할크기의 특성을 조건부 확률 정보 엔트로피를 이용하여 정보이론적인 측면에서 고려하여 보자.

U = {1,2,3,4,5,6,7}, $U/IND(E_1)$ = {{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}}. 임의의 러프집합 X={1,4,5}에 대하여 식(7)에 의하여 동치관계 E_1, E_2, E_3의 조건부 확률 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. 즉, 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 엔트로피의 변화를 야기하지 않는다[6]. 따라서 E_1과 E_2는 동치류의 분할 크기가 다를 경우에 불확실성이 동일하다는 것은 모순이 된다고 할 수 있다. 결국 E_3가 가장 안정적인 속성을 나타낸다.

3. 베이지언 정보엔트로피

정보시스템에서 처리되는 데이터에 대한 의사결정을 수행하기 위하여 신뢰할 수 있는 데이터의 저장은 필수적이다. 이와 같이 속성간의 상호의존성을 기반으로 한 정형화된 지식베이스를 구축하기 위하여 유용한 속성을 추출하는 방법이 중요한 문제가 된다. 문제는 조건부 속성의 분류에 있는 객체들 중에는 서로 다른 결론부에 속하는 객체가 존재한다. 이러한 일관성이 없는 객체들을 일관성이 있는 객체로 근사화를 통하여 분류하는 알고리즘이 필요하다. 이와 같이 비일관적인 객체를 근사화시키는 방법으로 많은 방법이 있지만 본 논문에서는 조건속성과 결정속성의 상호정보를 기반으로 상호정보의 변화량을 속성이 결정속성의 연관성의 정도를 인식하기 위하여 베이지안 정리를 적용하여 연관성의 정도에 대한

사후확률을 계산한다. 따라서 속성이 가지는 동치류들의 상호의존성을 바탕으로 영향력은 베이지안 사후확률 $P(x|E)$을 적용하여 얻을 수 있다. 결국, 조건속성과 결정속성간의 교집합 x가 주어진 상황에서 하나의 표본이 동치류 E에 속할 사후확률 $P(E|x)$을 구하는 문제로 볼 수 있다. 따라서 동치류 E에 속하면서 하나의 교집합 x를 가지는 경우의 확률은 $P(E \text{ and } x) = P(E)P(x|E) = P(x)P(E|x)$이 된다. 여기서 사후확률은 다음과 같이 정의 할 수 있다.

$$P(E_i|x) = \frac{P(E_i)P(x|E_i)}{\sum_{i=1}^{2} P(E_i)P(x|E_i)} \tag{8}$$

여기서, x는 두 개의 동치류간의 교집합, E는 동치류, 전체 집합에서 x에 대한 확률분포는 $P(x)$, 임의의 표본이 동치류에 속할 사전확률 $P(E)$이다. 전술한 동치류의 분할크기의 특성을 베이지안 정보엔트로피에 의한 정보이론적인 측면을 고려하기 위하여 러프집합 X={1,4,5}에 대하여 동치관계 E_1, E_2, E_3의 베이지안 정보엔트로피 $G(E|X)$는 식(8)에 의하여 다음과 같다.

$G(E_1|X)=G(\{1,2,3,4\}|X)+G(\{5,6,7\}|X)$
$= (-(4/7)ln(2/4)/(-(4/7)ln(2/4)-(3/7)ln(1/3))+$
$-(3/7)ln(1/3)/(-(3/7)ln(1/3)-(3/7)ln(1/3)))/2$
$= 0.478$
$G(E_2|X) = G(\{1,2\}|X)+G(\{3,4\}|X)+G(\{5,6,7\}|X)$
$= (-(2/7)ln(1/2)/(-(2/7)ln(1/2)-(3/7)ln(1/3))+$
$-(2/7)ln(1/2)/(-(2/7)ln(1/2)-(3/7)ln(1/3))+$
$-(3/7)ln(1/3)/(-(3/7)ln(1/3)-(3/7)ln(1/3)))/3$
$= 0.364$
$G(E_3|X) = G(\{1\}|X)+G(\{2\}|X)+G(\{3\}|X)$
$+G(\{4\}|X)+G(\{5,6,7\}|X)$
$= (-(1/7)ln(1/1)/(-(1/7)ln(1/1)-(3/7)ln(1/3))+$
$-(1/7)ln(1/1)/(-(1/7)ln(1/1)-(3/7)ln(1/3))+$
$-(3/7)ln(1/3)/(-(3/7)ln(1/3)-(3/7)ln(1/3)))/3$
$= 0.167$

X에 대해서 E_1, E_2와 E_3가 가지는 불확실성이 모두 다르게 계산되어 조건부 확률 정보 엔트로피에 의한 방법의 모순이 해결되었다. X={1,4,5}에 대하여 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 확률적으로 등가인 관계를 가지고 있지만, granule관점에서 보면 세 가지의 동치류들은 모두 다른 관계를 가지고 있다. 결국 granule가 작을수록 안정성을 가지고 있다는 것을 알 수 있고 E_3가 가장 안정적인 속성을 나타낸다고 할 수 있다. 결론적으로 일관적인 정보 시스템의 경우는 대수학 관점하의 속성감축은 감축후의 엔트로피가 변하지 않는 것을 보장할 수 없다. 반면에 일관적인 정보시스템의 경우는 정보론 관점하의 속성감축은 대수학 관점과 동일한 결과를 보여주었다.

따라서 베이지언 정보엔트로피를 통하여 정보시스템의 동치류에 의한 불확실성의 정확성을 보장할 수 있기 때문에 불완전한(incomplete) 정보 시스템에서 유용하다고 할 수 있다.

Ⅲ. 불완전 정보시스템 모델링

1. 결정속성의 베이지언 정보엔트로피

베이지언 정보엔트로피를 이용하여 *null*인 결정속성에 대하여 해당 객체의 조건속성의 베이지언 정보엔트로피를 계산함으로써 결정속성의 값을 결정할 수 있고, 조건속성 값은 동일하나 결정속성 값이 불일치하는 결정부속성 값도 결정할 수 있다[7,8,9,10,11]. 따라서 러프집합의 객체 $x \in U$에 대한 결정속성의 베이지언 정보엔트로피 H_o는 다음과 같이 정의할 수 있다.

$$H_o(X_i|X_i \cap d_j) = \sum_{i=1}^{n}\sum_{j=1}^{m} \frac{-P(X_i)ln|X_i \cap d_j|/|X_i|}{-\sum_{k=1}^{2} P(A_k)ln|X_i \cap d_j|/|A_k|} \tag{9}$$

여기서, i=1,..,n, j=1,..,m, t=1,..,l 이다. (X, d)의 결정속성 값 $a_d(x_t)$가 *null*이거나 또는 조건 속성의 값이 동일하나 결정속성의 값이 다른 불일치 값일 때 대체되는 결정 속성 값은 결정속성의 베이지언 정보엔트로피의 정의에 의해서 다음과 같이 결정된다.

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

단계 2. 조건속성의 개수(m)와 동치류의 개수(k)에 대하여 결정부 속성의 베이지언 정보엔트로피 $H_o(X_{jk}|X_{jk}\cap d_1), H_o(X_{jk}|X_{jk}\cap d_2),...,H_o(X_{jk}|X_{jk}\cap d_n)$을 구한다.

단계 3. $H_o(X_{jk}|X_{jk}\cap d_1)$, $H_o(X_{jk}|X_{jk}\cap d_2),...,H_o(X_{jk}|X_{jk}\cap d_n)$에서 j와 k에 대한 평균치가 *null*값을 대체할 수 있는결정속성 값인 결정속성의 베이지언 정보엔트로피 가 된다.

단계 4. $a_d(x_t)=min(mean(H_o(X_{jk}|X_{jk}\cap d_1), mean(H_o(X_{jk}|X_{jk}\cap d_2),...,mean(H_o(X_{jk}|X_{jk}\cap d_n)\}$에 의하여 결정된 값으로 결정속성 $a_o(x_t)$의 *null*값 또는 불일치 값을 대체다.

2. 조건속성의 베이지언 정보엔트로피

조건속성의 베이지언 정보엔트로피는 결정속성의 베이지언 정보엔트로피에 기반을 두고 있으며 조건속성 값이 *null*일 경우에 해당 객체의 *null*조건 속성 값만 계산함으로써 효과적으로 *null*값을 대체할 수 있다. $x\in U$에 대한 조건속성의 베이지언 정보엔트로피 H_c는 다음과 같이 정의 할 수 있다.

$$H_c(X_i|X_i\cap Y)=\sum_{i=1}^{n}\frac{-P(X_i)ln|X_i\cap Y|/|X_i|}{-\sum_{k=1}^{2}P(A_k)ln|X_i\cap Y|/|A_k|} \quad (8)$$

(x_t, a_c)의 결정속성 값 $a_c(x_t)$가 *null* 값일 때 대체되는 조건속성의 베이지언 정보엔트로피를 이용하여 다음과 같이 결정된다.

단계 1 *null*값을 가지는 객체의 결정속성의 동치류 d와 *null*값을 가지는 조건속성 X_j의 동치류를 구한다.

단계 2 결정속성의 동치류 d_j에 대하여 조건속성 X_i의 k개의 동치류의 베이지언 정보엔트로피 $H_c(X_{i1}|d_j)$, $H_c(X_{i2}|d_j),...,H_c(X_{ik}|d_j)$를 구한다.

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

Ⅳ. 적용사례

어떤 게임에 등장하는 몬스터의 상태 천이규칙에 관한 정보들을 판단하는 정보시스템을 표 1과 같이 구성하였다. 조건속성에 해당하는 입력조건은 current($c1$), input($c2$)과 weapon($c3$)의 세 개로 구성되고, 입력조건에 따른 몬스터의 다음상태가 결정속성이다. 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 $\{x_{10}, d\}$의 속성 값에 *null*이 포함되어 있는 불완전 정보시스템으로 가정한다면, $\{x_{10}, d\}$의 *null*값을 대체할 수 있는 결정속성 값을 최소 베이지언 정보엔트로피에 의하여 구할 수 있다.

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

index	c1	c2	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_{10}	anger	monster remedy	large	anger

1. 결정속성의 베이지언 정보엔트로피

결정속성 $\{d\}$에서 결정속성의 범주는 {'2', '3'}이므로 $H_d(A/\text{output='2'})$과 $H_d(A/\text{output='3'})$의 베이지언 정보엔트로피를 계산한 다음, 그 결과를 비교하여 $\{d_{10}\}$의 값 $H(x_{10})$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 결정속성을 가지는 의사결정 표
Table 2. Decision Table with NULL Decision Attribute

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

(1) X_2={x_2, x_4, x_5, x_7, x_8, x_{10}}에 대하여
H_d(current/output='1'):
$c_1(1)->d(1)$=-3/10*ln(3/3)/(-3/10*ln(3/3)-6/10*ln(3/6))=0
$c_1(2)->d(1)$=-3/10*ln(2/3)/(-3/10*ln(2/3)-6/10*ln(2/6))=0.1558
$c_1(3)->d(1)$=-4/10*ln(1/4)/(-4/10*ln(1/4)-6/10*ln(1/6))=0.3403
H_d(input/output='1'):
$c_2(1)->d(1)$=-4/10*ln(4/4)/(-4/10*ln(4/4)-6/10*ln(4/6))=0
$c_2(2)->d(1)$=-3/10*ln(2/3)/(-3/10*ln(2/3)-6/10*ln(2/6))=0.1558
$c_2(3)->d(1)$=-3/10*ln(0/3)/(-3/10*ln(0/3)-6/10*ln(0/6))=NaN
H_d(weapon/output='1'):
$c_3(1)->d(1)$=-3/10*ln(3/3)/(-3/10*ln(3/3)-6/10*ln(3/6))=0
$c_3(2)->d(1)$=-3/10*ln(2/3)/(-3/10*ln(2/3)-6/10*ln(2/6))=0.1558
$c_3(3)->d(1)$=-4/10*ln(1/4)/(-4/10*ln(1/4)-6/10*ln(1/6))=0.3403
H(A/Output='2')=(0.165+0.078+0.165)/3=0.136

(2)X_3={x_1, x_3, x_6, x_9, x_{10}}에 대하여
H_d(current/output='2'):
$c_1(1)->d(1)$=-3/10*ln(0/3)/(-3/10*ln(0/3)-5/10*ln(0/5))=NaN
$c_1(2)->d(1)$=-3/10*ln(1/3)/(-3/10*ln(1/3)-5/10*ln(1/5))=0.2906
$c_1(3)->d(1)$=-4/10*ln(4/4)/(-4/10*ln(4/4)-5/10*ln(4/5))=0
H_d((input/Output='2'):
$c_2(1)->d(1)$=-4/10*ln(0/4)/(-4/10*ln(0/4)-5/10*ln(0/5))=NaN
$c_2(2)->d(1)$=-3/10*ln(2/3)/(-3/10*ln(2/3)-5/10*ln(2/5))=0.2098
$c_2(3)->d(1)$=-3/10*ln(3/3)/(-3/10*ln(3/3)-5/10*ln(3/5))=0
H_d(weapon/output='2'):
$c_3(1)->d(1)$=-3/10*ln(0/3)/(-3/10*ln(0/3)-5/10*ln(0/5))=NaN
$c_3(2)->d(1)$=-3/10*ln(1/3)/(-3/10*ln(1/3)-5/10*ln(1/5))=0.2906
$c_3(3)->d(1)$=-4/10*ln(4/4)/(-4/10*ln(4/4)-5/10*ln(4/5))=0
H(A/Output='3')=(0.146+0.105+0.146)/3=0.132

결정속성의 최소 베이지언 정보엔트로피의 결과 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, 베이지언 정보엔트로피의 수식과 계산결과는 정확하다고 볼 수 있다.

2. 조건속성의 베이지언 정보엔트로피

표 3은 표 2에서 사용된 표와 같으나, 조건 속성{x_9, b}의 속성 값에 *null*이 포함되어 있다고 가정하면, {x_9, b}의 가능한 조건속성 값을 베이지언 정보엔트로피에 의하여 구하게 된다. 조건속성 {c_2}에서 가능한 조건속성 값의 유형은 '1','2','3'이고, x_9의 결정속성 값의 유형이 '3'이므로 $H_c(c_2(1)/'3')$, $H_c(c_2(2)/'3')$및 $H_c(c_2(3)/'3')$의 조건부엔트로피를 계산한 다음, 그 결과를 비교하여 c_2에 대한 x_9의 속성 값을 결정한다.

표 3. NULL 조건속성의 의사결정 표
Table 3. Decision Table with NULL Condition Attribute

index	c1	c2	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

H_c(input/output='3'):
(1) $H_c(c_2(1)/'3')$=-5/10*ln(1/5)/(-5/10*ln(1/5)-5/10*ln(1/5))=0.5
(2) $H_c(c_2(2)/'3')$=-4/10*ln(3/4)/(-4/10*ln(3/4)-5/10*ln(3/5))=0.3106
(3)$H_c(c_2(3)/'3')$=-3/10*ln(3/3)/(-3/10*ln(3/3)-5/10*ln(3/5))=0

조건속성의 베이지언 정보엔트로피의 계산결과를 토대로 작은 값이 영향력이 많은 속성이므로, c_2에 대한 x_9의 속성 값은 '3'으로 결정된다. c_2에 대한 x_9의 대체속성 값의 정확성 여부는 표 1과 비교해 보면 알 수 있다. 즉, 두 표의 속성 값이 모두 '3'이므로, 베이지언 정보엔트로피의 수식과 계산결과는 정확하다고 할 수 있다. 결국 비일관적인 정보시스템이 대한 베이지언 정보엔트로피에 의한 속성의 감축은 대수학적인 감축의 한계성을 극복할 수 있음을 보여주었다.

V. 결 론

본 논문에서는 불완전한 정보시스템에서 러프집합의 식별불능성에 대한 불확실성을 처리하기 위하여 대수적인 척도의 단점을 지양한 정보이론적인 베이지언 정보엔트로피 척도를 제안하였다. 제안된 척도는 지식베이스를 구성하기 위한 추론규칙의 정확성을 향상시키기 위한 전처리과정으로 수행될 수 있다. 결론적으로 불완전한 정보시스템의 대수학적인 속성의 감축은 감축후의 베이지언 정보엔트로피가 변하지 않는 것을 보장하지 않았기 때문에서 지식 감축의 정보이론적인 관점은 대수학적인 관점보다 포괄적이었다.

향후 제안된 척도를 이용한 알고리즘은 전자상거래와 같은 웹응용 분야에서 필요한 정보를 추출하고 분석하는 과정에서 속성간의 변별력을 정확하게 측정하여 상품의 선택에서 중요한 속성을 찾아냄으로써 사용자들이 빠르게 대처해야하는 경우에 효과적으로 이용될 수 있다고 사료된다.

References

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

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

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

[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] Baoxiang Liu, Ying Li, Lihong Li, Yaping Yu, “An Approximate Reduction Algorithm Based on Conditional Entropy”, Information Computing and Applications, Vol. 106, pp. 319-32, 2010

[10] Zhangyan Xu, Jianhua Zhou, Chenguang Zhang, “A Quick Attribute Reduction Algorithm Based on Incomplete Decision Table”, Information Computing and Applications, Vol. 391, pp. 499-508, 2013

[11] InKyoo Park, “Uncertainty Measurement of Incomplete Information System based on Conditional Information Entropy”, The Journal of The Institute of Internet, Broadcasting and Communication, Vol. 14, No. 2, pp. 107-113, 2014

※ 본 논문은 2013학년도 청운대학교 교내학술연구비에 의하여 지원되었음.

저자 소개

최 규 석(종신회원)

- 제9권 6호 참조
- 1991 ~ 1996 : (주)SK텔레콤 중앙연구원 책임연구원
- 현 : 청운대학교 컴퓨터학과 교수

<주관심분야 : 인공지능, ITS, 이동컴퓨팅>

박 인 규(정회원)

- 제10권 5호 참조
- 현 : 중부대학교 컴퓨터학과 교수

<주관심분야 : 데이터마이닝, 퍼지집합, 러프집합>