http://dx.doi.org/10.14400/JDC.2015.13.6.157

범주형 데이터의 러프집합 분석을 통한 의사결정 향상기법

박인규
중부대학교 컴퓨터·게임공학과

An Improvement of the Decision-Making of Categorical Data in Rough Set Analysis

In-Kyu Park
Dept. of Computer-Game, Joongbu University

요 약 최적의 의사결정시스템에서 효율적인 정보검색을 위해서는 정보의 감축이 필수적이다. 다양한 종류의 데이터에 존재하는 유용한 정보를 찾는 데이터 마이닝 기법에 대한 많은 연구가 활발하게 진행되어 왔고 타 산업과의 융복합을 위한 빅데이터 활용이 높아져 가고 있다. 유용한 지식의 발견을 위한 여러 가지 기법들이 특징을 가지고 있지만 단점이 존재하기 마련이다. 따라서 그러한 특징을 수렴하는 하나의 새로운 기법이 필요하다. 본 논문에서는 베이지언 정리를 이용하여 정보의 대수학적인 확률을 측정하고 이 확률에 대하여 정보엔트로피를 계산함으로써 정보의 불확실성을 계산한다. 제안된 척도를 기반으로 불필요한 속성을 제거하여 최소의 리덕트를 생성하고 이를 기반으로 결정규칙을 유도하는 알고리즘을 제안한다. 제안된 방법의 효율성을 위하여 콘텍트 렌즈를 결정하는 실험을 통하여 기존방법과 비교 결과, 본 연구가 의사결정의 유용성면에 있어 일반성이 있음을 보인다.

주제어 : 융복합, 데이터 마이닝, 러프집합, 베이지언 정리, 정보 엔트로피

Abstract An efficient retrieval of useful information is a prerequisite of an optimal decision making system. Hence, A research of data mining techniques finding useful patterns from the various forms of data has been progressed with the increase of the application of Big Data for convergence and integration with other industries. Each technique is more likely to have its drawback so that the generalization of retrieving useful information is weak. Another integrated technique is essential for retrieving useful information. In this paper, a uncertainty measure of information is calculated such that algebraic probability is measured by Bayesian theory and then information entropy of the probability is measured. The proposed measure generates the effective reduct set (i.e., reduced set of necessary attributes) and formulating the core of the attribute set. Hence, the optimal decision rules are induced. Through simulation deciding contact lenses, the proposed approach is compared with the equivalence and value-reduct theories. As the result, the proposed is more general than the previous theories in useful decision-making.

Key Words : Convergence and Integration, Data Mining, Rough Set, Bayesian Theory, Information Entropy

Received 20 March 2015, Revised 28 April 2015
Accepted 20 June 2015
Corresponding Author: In-Kyu Park(Joongbu University)
Email: fip2441g@gmail.com

ISSN: 1738-1916

1. 서론

예측되는 정보나 데이터를 미리 수집, 가공, 처리하여 찾기 쉬운 형태로 축적해 놓은 데이터베이스로부터 요구에 적합한 정보를 신속하게 접근할 수 있는 정보시스템의 구축은 이미 일반화되어 있다. 따라서 지식의 대용량화에 따른 정보검색의 효율성 또한 부인할 수 없다. 이에 대한 일환으로 인공지능(artificial intelligence) 기법을 이용한 정보시스템에 관한 연구들이 상당히 유용한 결과를 보여 왔다[4,7].

실질적인 의사결정에 활용하기 위하여 데이터베이스에 존재하는 패턴의 관계를 찾아내어 유용한 정보를 추출해 내는 방법으로 여러 가지의 통계적 접근법, 신경망(neural network), 인공지능, 전문가시스템(expert system), 퍼지논리, 패턴인식(pattern recognition), 기계적 학습(machine learning)과 불확실성 추론(reasonging with uncertainty)기법들이 사용되어 왔다[외국논문]. 따라서 실질적인 의사결정 규칙들을 도출하기 위하여 불완전한 지식의 처리, 모순이 있는 정보의 취급과 임의의 지식을 다양한 수준으로 표현할 수 있어야 한다. 정보시스템의 구성요소인 데이터베이스에 존재하는 유용한 데이터의 정제와 추출을 위하여 부정확한 데이터로 부터의 지식추론에 핵심적인 데이터간의 불확실한 관계를 밝히는 것이 관건이다. 불확실성(uncertainty)을 취급하는 방법에는 확률이론이 전통적으로 사용되어 왔으며 가장 잘 알려져 있다. 그 중의 하나로 불확실한 정보의 처리 과정에서 설정된 가설에 주어진 관찰 증거로 신뢰도를 이끌어 내는 베이지언 정리(Bayesian theory)가 있다[2].

이러한 방향의 연구에 러프집합(rough set)이 적합하다는 충분한 근거가 제시되어 왔으며 데이터를 지식으로 변환해 주는 논리이다. 반면에, 러프집합은 지식을 명확하게 나타내기에는 전체집합이 너무 세밀하게 구별되어 있고, 여러 범주에 포함되어 있는 전체집합이 너무 거칠게 나뉘어져서 데이터 중의 어떤 객체는 식별 불가능한 상태에 있을 수 있다[11].

지식을 표현하는 기본적인 개념은 분류(classification)와 범주(category)에 기초하고 있다. 분류는 추론, 학습과 의사결정에 있어서 중요한 문제라고 할 수 있다. 임의의 지식기반(knowledge base)내에 존재하는 지식을 표현하는 기본 단위는 객체이고 이들의 부분적인 모임을 개념 또는 범주라고 한다. 일반적으로 객체들로 이루어진 집합인 어떤 범주를 사용가능한 지식으로 정확하게 나타낼 수 없다. 따라서 객체들을 분할하는데 범주의 애매성(vagueness)이 발생하게 되어 다른 집합들로 집합에 대한 근사화(approximation)가 수반되어 진다. 그러나 이러한 조건부 확률(conditional probability)에 의존한 베이지언 계산법은 확실성하에서 잘 정의된 조건(가설과 증거)들을 취급한 확률을 갱신해 가는 것이므로, 베이지언 스키마(schema)에는 확실성 관리 대상 중 가장 중요한 부정확성(inexactness), 부정합(inconsisency), 부적절(irrelavance), 불완전(incompleteness)을 수용에는 한계가 있다[3].

본 논문에서는 베이지언 정리에 정보의 특이성에 대한 불확실성을 측정하는 정보 엔트로피(information entropy)를 결합하여 베이지언 확률 정보엔트로피(Bayesian probability information entropy) 척도를 제안하고, 이를 기반으로 최적의 정보시스템을 구성하기 위한 속성 리덕트(reduct)생성 알고리즘과 제어규칙 생성 알고리즘을 제안한다.

본 논문의 2장에서는 러프집합의 근사공간과 리덕트 및 코어를 정리하였으며 3장에서는 엔트로피 기반 속성 리덕트 생성알고리즘과 제어규칙 생성 알고리즘을 제안한다. 또한 4장에서는 제안된 알고리즘을 이용하여 안경 결정사례에 적용하여 규칙을 추출한다. 마지막장에서는 결론 및 향후 연구 방향을 제시한다.

2. 러프집합이론

지식에 대한 개념은 기본적으로 분류와 범주를 기반으로 하고 있고, 공집합이 아닌 유한집합 U에 대한 임의의 부분집합 X를 U내의 범주라고 한다. 범주는 지식을 표현하는 객체들로 구성되지만 이들이 항상 정확하게 정의되지 않을 수도 있다. U내의 여러 개의 분류들의 집합을 U의 지식기반이라 하고 분류에 해당하는 동치관계를 기반으로 지식을 표현할 수 있다.

$P \subseteq R$이고 $p \neq \varnothing$이면 P에 해당하는 모든 동치관계들의 교집합은 동치관계가 되며 $IND(P)$는 표기 하고 $IND(P) = \{(x_1,x_2) \in U*U \mid \forall a \in P, f(x_1,a) = f(x_2,a)$이다. 즉, P의 식별불가능관계(indiscernibility relation)라고 한

다. 따라서 정의되지 않는 범주는 애매성을 가지므로 부분집합 $X \subseteq U$와 동치관계 $R \in IND(R)$을 이용하여 두 개의 집합 R-하한근사(R-lower approximation)와 R-상한근사(R-upper approximation)를 다음과 같이 정의한다[10].

$$\underline{R}X = \cup \{x \in U | R(X) : R(X) \subseteq X \quad (1)$$
$$\overline{R}X = \cup \{x \in U | RB(X) : R(X) \cap X \neq \varnothing$$

$X \subseteq U$이고 R이 동치관계(equivalence relation)일 경우에 X가 R-기본범주들의 합집합이면 X가 R-정의 가능하다(R-definable)라고 하고, 그렇지 않으면 X가 R-정의 불가능하다(R-undefinable)라고 하고 X를 R-러프집합(R-rough set)이라고 한다. X가 R-정의 가능집합이 되도록 하는 동치관계 $R \in IND(K)$가 존재하면 집합 $X \subseteq U$는 k내에서 정확하다고 하고, 어떤 $R \in IND(K)$에 대해서도 X가 R-러프하면 X는 k내에서 러프하다고 한다. 또한 하한근사와 상한근사와의 차는 경계영역으로서 어느 쪽에도 속하지 않는 애매한 영역은 다음과 같이 정의할 수 있다.

$$BND(X) = \overline{B}X - \underline{B}X \quad (2)$$

이 영역에 속하는 범주는 하한근사와 상한근사의 두 정확한 범주를 기반으로 러프하게 정의할 수 있다. 즉, 부정확한 범주에 해당되며 이에 대한 척도는 부정확한 정도로서 다음과 같이 정의할 수 있다.

$$\rho_A(X) = 1 - \frac{card\ \underline{A}X}{card\ \overline{A}X}, X \neq \varnothing \quad (3)$$

이는 경계영역에서 발생되는 부정확성을 나타내는 척도이지만 식별불가능 관계에 의한 부정확성을 완전히 처리하지 못한다.

따라서 정보시스템의 효율적인 검색을 위한 지식의 감축은 유용한 대안이고, 리덕트와 코어(core)를 이용하여 지식기반 내의 유용한 범주들의 집합은 보존되고 불필요한 동치관계를 제거하여 원래의 지식과 동일한 지식의 표현이 가능하다. 결국, 지식의 리덕트는 지식내에 존재하는 모든 지식의 범주들을 정의하기에 충분한 지식의 필수적인 부분이고, 코어는 어떤 의미에서 지식의 가장 중요한 부분이다. 리덕트에 존재하는 모든 필수 불가결한 속성들의 집합은 리덕트의 코어이고 다음과 같이 나타낸다.

$$CORE(B) = \cap_{R \in RED(B)} R \quad (4)$$

여기서 $RED(B)$는 B의 모든 리덕트의 집합으로서 속성들에 대한 최소한의 집합으로 객체들을 속성들에 의하여 분별할 수 있다.

3. 제어규칙의 모델링

3.1 다중 속성 리덕트

특정 분야의 정보를 나타낸 데이터베이스 시스템에서 유용한 정보를 검색하여 필요한 의사 결정을 내리기 위해서는 신뢰할 수 있는 데이터의 저장은 물론, 속성간의 상호 의존성을 분석한 최소의 사실과 규칙에 대한 정형화된 지식베이스의 구축이 필요하다. 이러한 지식은 분류 즉, 분할과 밀접한 관련이 있고 지식을 구성하는 범주들은 분류를 구성하는 동치류에 해당한다. 따라서 이러한 동치류들의 알갱이성(granularity)은 분류와 밀접한 관련이 있기 때문에 이러한 알갱이성에 의한 분류의 불확실성이 중요한 단서가 되어 왔다. 예를 들어, 임의의 확률을 알고 있을 때 그에 대한 정보량에 대한 불확실성을 정보 엔트로피를 이용하여 측정할 수 있다. 어떤 통계량에서 상태 i의 확률을 p_i로 정의할 경우에 n개의 정보에 대한 엔트로피 E는 다음과 같이 정의할 수 있다[8].

$$E = -\sum_{i}^{n} p_i \ln p_i \quad (5)$$

엔트로피와 물확실성은 비례하기 때문에 베이지언 확률에 변별력을 부여하기 위하여 엔트로피 개념을 결합하면 동치류들의 변별력이 향상되었다. 가령, U = {1,2,3,4,5,6,7}에서 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 경우에 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, 각 동치류들의 알갱이 관점에서 보면 다른 관계를 가지고 있다. 그러나 정보엔트로피를 계산하면 X에 대해서 E_1, E_2와 E_3가 가지는 불확실성이 모두 다르게 계산되어 각 부분집합이 서로 상이하다. 결국 동

치류들의 알갱이성이 작을수록 애매성이 적기 때문에 E_3가 가장 안정적이라고 할 수 있다.

본 논문에서는 결정속성의 결정규칙별로 분할하여 조건속성의 동치류를 구성하여 조건부와 결정부의 연관되어 있는 대수학적인 확률을 구하는데 베이지언 정리를 이용하였다. 또한 임의의 의사결정에 대하여 해당하는 조건부와 결론부의 속성간의 존재하는 결합확률에 대하여 정보 엔트로피를 적용하여 다음과 같은 베이지언 확률 정보엔트로피를 얻을 수 있다.

$$H(S|R) = \frac{-P(S)*lnP(T|S)}{\sum_{k\in\{B_i,D_j\}} -P(IND(k))*lnP(T)|IND(k))} \quad (6)$$

여기서 R=$IND(B_i \cap D_j)$는 조건속성과 결론속성의 동치류의 범주 값을 만족하는 객체가 조건속성과 결정속성의 동치류에서 발견될 확률이고, S=$IND(B_i)$는 조건속성의 동치류의 전체 객체에 대한 확률이고, T=$IND(D_j)$는 결정속성의 동치류의 전체 객체에 대한 확률이다. 또한, $P(T|S)$는 조건속성이나 결정속성의 동치류가 주어졌을 때, 조건속성과 결정속성의 범주 값을 만족하는 객체가 발생할 조건부 확률 즉, 우도(likelihood)이다. 따라서 러프집합에서 다루어지고 있는 불확실성은 식별 불가능한 관계에서 발생하는 불확실성으로 동치류들을 구성하는 객체들을 구별할 수 없기 때문에 조건속성과 결정속성의 동치류의 연관정도를 고려하여 이러한 불확실성을 모델링을 할 수 있다.

이러한 베이지언 확률 정보엔트로피에 의한 다중 속성리덕트는 데이터의 패턴이 같은 전체 조건속성에 대한 최소 부분집합으로써 최초의 정보시스템의 일반적인 특성을 유지한다. 따라서 조건부와 결정부간의 대수학적인 함의(implicantion)에 정보엔트로피를 적용하여 최적의 속성리덕트를 결정하기 위한 알고리즘을 [Fig. 1]에 나타내었다. 알고리즘의 전위선택(forward selection)은 우선 결정 및 조건속성 간의 엔트로피의 값이 낮은 조건속성부터 후보리덕트가 되어 이 후보리덕트 RED=$RED \cup a$로 추가하는 부분이다. 연관성이 높은 조건속성이 속성리덕트의 후보가 되어 후위제거(backward elimination)의 $r_{RED}(D)$=$r_C(D)$의 조건에 의해 최초의 지식 베이스의 일관성을 조사하게 된다. 이와 같은 절차에 의해 정보의 손실이 없는 속성요인을 계속적으로 리덕트에 부가하여, 여분의 조건속성이 제거된 최적의 다중 속성 리덕트를 추출할 수 있다.

Attribute reduction by Bayesian probability entropy

Input: inforation system $S=(U,C,D,V,f)$
$X=\{x_1, x_2, ...,x_n\}$, $x_j(1\le j\le n)$
$A=\{A_1,A_2,...,A_p\}$, $A_l(1\le l\le p)$
Output: optimal attribute reduct $RED_D(C)$

Determine C, D
Compute lower and upper approximations and core
Compute each H for $a\in C$ w.r.t. $d\in D$
for l = 1 to $p(1\le l\le p)$
  for k=1 to n_d
    for j=1 to n_l
    $h = -\frac{\{|a_i^{(j)}||a_i^{(j)}\in DOM(A_l)\}}{n} ln \frac{\{|a_i^{(j)}||(a_i^{(j)},a_d^{(k)})\}}{\{|a_i^{(j)}||a_i^{(j)}\in DOM(A_l)\}}$
    endfor
  endfor
  $H_l = H + h$
endfor
Select an attribute a the lowest H_l value in C
$RED_D(C)=CORE \cup a$

[Fig. 1] Optimal reduct extraction algorithm

3.2 러프 제어규칙의 발생

지식베이스에서 베이지언 확률 엔트로피에 의한 다중 속성 리덕트 알고리즘에 의해 간소화된 조건부 속성을 기반으로 결정시스템을 구성하고 불일치하는 결정 규칙들에 대하여 최소의 결정 일관성 규칙을 추출한다. 다중 속성 리덕트에 의해 생성된 결정규칙들에는 포함관계와 조건속성의 부분적 중복이 있을 수 있다[9].

즉, 두 개의 규칙(r_a, r_b)의 조건부를 $cond(r_a)$, $cond(r_b)$라 하고, 결론부를 $dec(r_a)$, $dec(r_b)$라 할 때 $cond(r_a) \supseteq cond(r_b)$이고 $dec(r_a)$= $dec(r_b)$이면 규칙 r_a는 규칙 r_b를 논리적으로 포함한다고 한다. 이때 논리적으로 포함되는 규칙은 제거될 수 있다. 또한 $cond(r_a) \supseteq cond(r_b)$이고 $dec(r_a) \neq dec(r_b)$이면 규칙 r_a와 규칙 r_b는 불일치 결정규칙이라고 정의한다[1,5].

논리적인 포함관계와 $cond(r_a) \supseteq cond(r_b)$이고 $dec(r_a) \neq dec(r_b)$일 경우에 $supp(r_a)/(supp(r_a)+supp(r_b)) > \beta$이면 규칙 r_a는 규칙 r_b를 확률적으로 포함한다고 가정하여 결정규칙을 간소화한다. 여기서 $supp$는 규칙을 지지하는 객체의 수이다. 따라서 결정속성에 대하여 최소의 결정규칙을 생성하는 결정규칙 알고리즘은 [Fig. 2]에 나타낸다.

Decision rule generation by Bayesian probability entropy

Input: inforation system $S=(U,C,D,V,f)$
$X=\{x_1, x_2, ...,x_n\}$, $x_j(1\le j\le n)$
$A=\{A_1,A_2,...,A_p\}$, $A_l(1\le l\le p)$

Output: optimal control rules
Row reduction according to the duplication
CORE extraction
for l=1 to p
  for j=1 to n_l-1
    if $IND(a_l(j))\ge 2$
      if $IND(a_l(j)) \neq IND(a_{l+1}(j))$
        remove $f(a_{l+1},IND(a_l(j)))$
        endif
    endif
  endfor
endfor
Merge category values of conditional part

[Fig. 2] Control rules generation algorithm

4. 실험 및 고찰

제안된 알고리즘을 이용하여 안경상이 환자가 콘텍트 렌즈를 사용하기가 적합한가 하는 문제를 고려해 보고자 한다. 이에 관한 모든 가능한 의사결정들이 <Table 1>에 나와 있다. 여기서 조건 속성은 나이(a), 시야(b), 난시(c) 및 눈물 분비율(d)이고 e는 의사결정을 나타내는 속성이다. 속성 e는 안경상의 의사결정을 나타내며 속성 값은 *1*(딱딱한 컨텍트 렌즈), *2*(부드러운 컨텍트 렌즈) 와 *3*(컨텍트 렌즈 불필요)으로 구성되어 있다. 나이속성의 범주 값은 *1*(젊음), *2*(노안 이전) 그리고 *3*(노안), 시야의 범주는 *1*(근시), *2*(원시), 난시의 경우는 *1*(정상), *2*(난시) 그리고 눈물 분비율의 범주는 *1*(감소), *2*(정상)이다. 먼저 <Table 1>의 데이터에 대한 코어는 {*a*, *b*, *c*, *d*}이다. 그러나 코어를 구성하는 각 속성이 가지는 속성간의 기여도에 차이가 있음을 알 수 있다. 따라서 최소의 의사결정 규칙을 추출하기 위하여 일환으로 베이지언 정리를 이용하여 결정속성의 범주 값에 대하여 조건속성의 범주 값의 함의(implication)에 대한 일종의 기여도에 해당하는 확률을 구하여 이 값이 임의의 임계값을 만족하지 못하는 확률을 가지는 속성을 코어에서 제거하였다.

<Table 1>에 대한 각각의 조건속성의 베이지언 러프 엔트로피는 <Table 2,3,4,5>를 통하여 0.848, 0.467, 0.437과 0.352임을 알 수 있다.

〈Table 1〉 A incomplete decision table

index	condition				decision
	a	b	c	d	e
1	1	1	2	2	1
2	1	2	2	2	1
3	2	1	2	2	1
4	3	1	2	2	1
5	1	1	1	2	2
6	1	2	1	2	2
7	2	1	1	2	2
8	2	2	1	2	2
9	3	2	1	2	2
10	1	1	1	1	3
11	1	1	2	1	3
12	1	2	1	1	3
13	1	2	2	1	3
14	2	1	1	1	3
15	2	1	2	1	3
16	2	2	1	1	3
17	2	2	2	1	3
18	2	2	2	2	3
19	3	1	1	1	3
20	3	1	1	2	3
21	3	1	2	1	3
22	3	2	1	1	3
23	3	2	2	1	3
24	3	2	2	2	3

〈Table 2〉 Rough entropy of attribute *a*

rule of a	bayesian rough entropy
(x=1)⇒(D=1)	-8/24*ln(2/8)/(-8/24*ln(2/8)-4/24*ln(2/4))=0.8
(x=2)⇒(D=1)	-8/24*ln(1/8)/(-8/24*ln(1/8)-4/24*ln(1/4))=0.75
(x=3)⇒(D=1)	-8/24*ln(1/8)/(-8/24*ln(1/8)-4/24*ln(1/4))=0.75
(x=1)⇒(D=2)	-8/24*ln(2/8)/(-8/24*ln(2/8)-2/24*ln(1/2))=0.888
(x=2)⇒(D=2)	-8/24*ln(2/8)/(-8/24*ln(2/8)-2/24*ln(1/2))=0.888
(x=3)⇒(D=2)	-8/24*ln(1/8)/(-8/24*ln(1/8)-2/24*ln(1/2))=0.923
(x=1)⇒(D=3)	-8/24*ln(1/8)/(-8/24*ln(1/8)-3/24*ln(1/3))=0.834
(x=2)⇒(D=3)	-8/24*ln(2/8)/(-8/24*ln(2/8)-3/24*ln(2/3))=0.901
(x=3)⇒(D=3)	-8/24*ln(2/8)/(-8/24*ln(2/8)-3/24*ln(2/3))=0.901
average	0.848

〈Table 3〉 Rough entropy of attribute *b*

rule of b	bayesian rough entropy
(y=1)⇒(D=1)	-12/24*ln(3/12)/(-12/24*ln(3/12)-4/24*ln(3/4))=0.935
(y=2)⇒(D=1)	-12/24*ln(1/12)/(-4/24*ln(1/12)-4/24*ln(1/4))=0.843
(y=1)⇒(D=2)	-12/24*ln(2/12)/(-2/24*ln(2/12)-5/24*ln(1/5))=0.728
(y=2)⇒(D=2)	-12/24*ln(3/12)/(-4/24*ln(3/12)-5/24*ln(2/5))=0.784
(y=1)⇒(D=3)	-12/24*ln(7/12)/(-2/24*ln(7/12)-15/24*ln(7/15))=0.361
(y=2)⇒(D=3)	-12/24*ln(8/12)/(-2/24*ln(8/12)-15/24*ln(8/15))=0.340
average	0.467

〈Table 4〉 Rough entropy of attribute *c*

rule of c	bayesian rough entropy
(z=2)⇒(D=1)	-12/24*ln(4/12)/(-12/24*ln(4/12)-12/24*ln(4/12))=0.5
(z=1)⇒(D=2)	-12/24*ln(5/12)/(-12/24*ln(5/12)-12/24*ln(5/12))=0.5
(z=1)⇒(D=3)	-12/24*ln(7/12)/(-12/24*ln(7/12)-15/24*ln(7/15))=0.407
(z=2)⇒(D=3)	-12/24*ln(8/12)/(-12/24*ln(8/12)-15/24*ln(8/15))=0.340
average	0.437

〈Table 5〉 Rough entropy of attribute *d*

rule of d	bayesian rough entropy
(z=2)⇒(D=1)	-12/24*ln(4/12)/(-12/24*ln(4/12)-12/24*ln(4/12))=0.5
(z=2)⇒(D=2)	-12/24*ln(5/12)/(-12/24*ln(5/12)-12/24*ln(5/12))=0.5
(z=1)⇒(D=3)	-12/24*ln(12/12)/(-12/24*ln(12/12)-15/24*ln(12/15))=0
(z=2)⇒(D=3)	-12/24*ln(3/12)/(-12/24*ln(3/12)-15/24*ln(3/15))=0.408
average	0.352

따라서 엔트로피 값이 적을수록 기여도가 크기 때문에 *a*속성이 임계값(0.5)을 만족하지 못하므로 코어는 {*b*, *c*, *d*}로 결정되어 <Table 1>은 <Table 6>과 같이 간략화 되고 <Table 6>의 의사결정규칙에 대한 코어 값들이 <Table 7>에 나타나 있다.

〈Table 6〉 Reduced decision table of 〈Table 1〉

index	condition			decision
	b	c	d	e
1	1	2	2	1
2	2	2	2	1
3	1	1	2	2
4	2	1	2	2
5	1	1	1	3
6	1	2	1	3
7	2	1	1	3
8	2	2	1	3
9	2	2	2	3
10	1	1	2	3

〈Table 7〉 Reduced decision table of 〈Table 1〉

index	condition			decision
	b	c	d	e
1	*	2	2	1
2	*	1	2	2
3	1	*	1	3
4	2	*	1	3
5	2	2	2	3
6	1	1	2	3

결과적으로 원래의 주어진 표와 등가인 <Table 7> 의 최소 의사결정규칙을 얻을 수 있다. *는 속성의 'don't care'값을 의미한다. 즉, 원래의 표에서 주어진 의사결정을 내리는데 필요한 조건에 대한 최소의 집합만이 <Table 7>에 포함되어 있다. 상기의 결과는 오로지 하나의 해에 해당할 뿐이고 일반적으로 다수의 해가 존재할 수 있다. 제안된 방법에 대한 비교우위를 위하여 두 개의 기존 방법의 결과를 <Table 8>과 <Table 9>에 나타내었다.

〈Table 8〉 Decision table of 〈Table 1〉 by equivalences

index	condition				decision
	a	b	c	d	e
1	1	*	2	2	1
2	*	1	2	2	1
3	1	*	1	2	2
4	2	*	1	2	2
5	*	2	1	2	2
6	*	*	*	1	3
7	2	2	2	*	3
8	3	1	1	*	3
9	3	2	2	*	3

<Table 8>은 <Table 1>의 모든 의사결정 규칙에 대하여 참과 식별불가능성을 모든 조건부 속성에 적용하여 얻은 최소의사결정 규칙이다. <Table 9>는 러프집합 이론의 기본적인 값리덕트(value reduct) 알고리즘을 이용하여 추출된 최소의사결정 규칙이다.

〈Table 9〉 Decision table of 〈Table 1〉 by Value reduct

index	condition			decision
	a	c	d	e
1	*	2	2	1
2	*	1	2	2
3	*	2	*	3
4	3	*	*	3

<Table 8>은 모든 속성을 가지고 결정규칙을 추출하기 때문에 최적화에는 거리가 있고 가능한 의사결정 규칙을 보여주고 있다고 할 수 있다. <Table 9>의 의사결정 *3*에서는 <Table 8>의 의사결정 *3*에 해당하는 결정의 조건속성에 'don't care'가 50%이상 존재하기 때문에 원래의 조건부에 대한 최적화라고 하기에는 거리가 있다. 결국 세 가지의 최소의 의사결정규칙에는 <Table 9> ⊂ <Table 7> ⊂ <Table 8>의 포함관계가 존재한다고 할 수 있다. 따라서 제안된 방법의 의사결정규칙이 최적화에 근접하다고 할 수 있다. 의사결정 *1*과 *2*에서는 조건속성 *c*와 *d*가 코어임을 알 수 있다. 또한 의사결정 *3*에서는 기존의 값리덕트 알고리즘의 경우를 제외한 두 개의 최소의사결정 규칙을 통하여 조건속성의 *b*와 *c*의 코어가 존재하고 다른 코어에는 *b*와 *d*가 존재하는 것을 알 수 있다.

5. 결론

본 논문에서는 조건부 속성과 결정부 속성과의 연관관계의 신뢰도를 기반으로 지식베이스를 정제하여 효율적인 정보 검색에 관해 기술하였다. 러프집합의 다중 리덕트의 생성과 경계영역의 비결정성 객체의 인식에 대한 문제점을 해결하기 위해서, 조건부와 결정부 속성간의 함의에 대하여 베이지언 확률 엔트로피를 이용하여 두 속성의 상관성이 높은 유일한 최적의 속성리덕트 생성 알고리즘과 최소의 의사결정규칙을 생성하는 알고리즘을 제안하였다. 결과적으로 기존의 방법보다 불필요한 속성을 제거하고 일반성을 확보할 수 있었다.

제안된 방법에서는 데이터베이스에 내재된 중요한 규칙을 발견하므로 여러 가지의 의사결정 업무에 적용될 수 있다. 또한, 의사결정규칙의 유도를 통하여 전문가 시스템 등의 지식 베이스의 구축에 활용될 수 있다. 향후, 제안된 결정규칙 추출 알고리즘을 빅데이터를 위한 의사결정 시스템으로 구축할 필요가 있다고 사료된다.

REFERENCES

[1] E.A. Kweon, H.G. Kim, "Simplication of control rules using probabilistic rough set", Journal of Information Processing, Vol. 8-D, No. 3, pp.203-210, 2001.

[2] E. Satake, A. V. Murray, "Teaching an Application of Bayes' Rule for Legal Decision-Making: Measuring the Strength of Evidence", Journal of Statistics Education, Vol. 22, No.1, 2014.

[3] H.S. Kim, H.G. Kim, S.B. Lee, "Retrieval of fuzzy information based on probabilistic rough sets", Journal of Information Science, Vol. 25, No. 9, pp. 1431-1441, 2005.

[4] G. J. Williams and S. J. Simoff, "Data mining theory, methodology, Techniques and Applications (Lecture Notes in Computer Science/Lecture Notes in Artificial Intelligence)", Springer, 2007.

[5] In-Kyoo Park. "The generation of control rules for data mining", The Journal of Digital Policy & Management, Vol. 11, No.1, pp.343-349, 2013.

[6] R. Vashist, M.L. Garg, "Rule generation based on reduct and core: a rough set approach", International Journal of Computer Applications, Vol. 29, No. 9, pp. 0975-8887, Sept. 2011.

[7] S.K. Pal and A. Skowron, "Rough Fuzzy Hybridization: A new trend in decision making", Springer Verlag, Berlin, 1999.

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

[9] Y. C. Tsai, C. H. Cheng and J. R. Chang, "Entropy-based fuzzy rough classification approach for extracting classification rules", Expert Systems with Applications Vol. 31, pp. 436-443, 2006.

[10] Z. Pawlak, "Rough sets", International Journal of Information Sciences, Vol.11, No. 5, pp. 341-356, 1982.

[11] Z. Pawlak, "Rough sets: Theoretical aspects of reasoning about data", Kluwer Academic Publishers, 1991.

[12] J. Liang, Z. Shi, D. Li, M.J. Wierman, "Information entropy, rough entropy and knowledge granulation in incomplete information systems", International Journal of Genreal Systems, Vol. 35, No. 6, pp. 641-654, December, 2006.

[13] Q. Shen, R. Jensen, "Rough sets, their Extensions and Applications", International Journal of Automation Computing, Vol. 4, No. 1, pp. 100-106, 2007.

[14] Chua Hong Siang, Sanghyuk Lee, "Information Management by Data Quantification with FuzzyEntropy and Similarity Measure", Journal of the Korea Convergence Society, Vol. 4, No. 2, pp. 35-41, 2013.

[15] Sang-Hyun Lee, "A Study on Determining Factors for Manufacturers to Distributors Warehouse in Supply Chain", Journal of the Korea Convergence Society, Vol. 4, No. 2, pp. 15-20, 2013.

박 인 규(Park, In Kyu)

- 1985년 2월 : 연세대학교 공학석사
- 1997년 2월 : 원광대학교 공학박사
- 1997년 3월 ~ 현재 : 중부대학교 컴퓨터·게임공학과 교수
- 관심분야 : 데이터마이닝, 러프집합, 퍼지집합, 지능형 시스템
- E-Mail : fip2441g@gmail.com