

<http://dx.doi.org/10.7236/IIBC.2015.15.5.69>

IIBC 2015-5-8

사후확률에 기반한 근사 규칙의 생성

Creation of Approximate Rules based on Posterior Probability

박인규*, 최규석**

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

요약 본 논문에서는 데이터베이스의 정보시스템을 구성하는 속성을 감축하여 빠른 검색을 보장하는 제어규칙의 생성에 관한 연구이다. 일반적으로 정보시스템에는 불필요한 많은 속성들이 존재하고 있다. 이때 정보시스템의 객체들이 비일관적일 경우에는 응답의 정확성을 기대하기 어렵게 된다. 그러므로 본 논문에서는 러프엔트로피의 개념과 베이저언 사후확률을 적용하여 불필요한 속성을 제거하여 정보시스템을 간결화 하는데 주안점을 두었다. 제안된 알고리즘에서는 러프이론에 기반한 최적의 리덕트를 생성하는 과정에서 사후확률을 적용하여 결정속성에 대한 조건속성의 함의를 러프 엔트로피의 척도로 비교하여 영향력이 약한 속성을 제거하여 제어규칙을 간결하게 표현할 수 있다. 제안된 알고리즘을 신입사원의 채용에 적용하여 지식감축의 효용성을 보인다.

Abstract In this paper the patterns of information system is reduced so that control rules can guarantee fast response of queries in database. Generally an information system includes many kinds of necessary and unnecessary attribute. In particular, inconsistent information system is less likely to acquire the accuracy of response. Hence we are interested in the simple and understandable rules that can represent useful patterns by means of rough entropy and Bayesian posterior probability. We propose an algorithm which can reduce control rules to a minimum without inadequate patterns such that the implication between condition attributes and decision attributes is measured through the framework of rough entropy. Subsequently the validation of the proposed algorithm is showed through test information system of new employees appointment.

Key Words : Data Mining, Cluster Analysis, Uncertainty, Entropy, Rough Set

1. 서론

정보의 수요가 날로 커져감에 따라서 정보량을 줄이고 관련이 있는 정보만을 채집하는 일이 매우 복잡하게 되었다^[1,2,3,4]. 이러한 문제에 대한 해결책으로 데이터베이스에 존재하는 많은 정보로부터 관련된 정보를 채집할 수 있는 유용한 도구로 러프집합이론이 좋은 해결책이 되어 왔다. 러프집합 이론의 특징으로 첫째, 데이터에 숨

겨져 있는 사실(facts)을 해석할 수 있다. 둘째, 데이터에 대한 추가적인 정보가 필요 없다. 셋째, 최소한의 지식으로 전체 데이터를 나타낼 수 있다^[6,7].

또한, 사냥에 의하여 개발된 정보이론은 다양한 분야에서 정보의 량을 측정하는 도구로 실제로 널리 활용되고 있다. 엔트로피^[5]를 이용하여 어떤 정보를 표현하는 방식은 규칙베이스 시스템에서 데이터베이스, 퍼지 데이터베이스 쿼리(query), 데이터 할당과 분류와 같은 분야

*정회원, 중부대학교 컴퓨터학과

**중신회원, 청운대학교 컴퓨터학과(교신저자)

접수일자 : 2015년 7월 10일, 수정완료 2015년 8월 7일

게재확정일자 : 2015년 10월 9일

Received: 10 July, 2015 / Revised: 7 August, 2015 /

Accepted: 9 October, 2015

**Corresponding Author: lionel@chungwoon.ac.kr

Dept. of Computer Science, Chungwoon University, Korea

에서 활용되고 있다. 또한 이를 토대로 어떠한 의사결정에 필요한 정보를 정형화함으로써 유용한 정보를 도출할 수 있고, 어떤 객체나 상태, 의견 과정들 사이의 차이점을 추적할 수 있는 유용한 도구이다^[8].

러프집합을 활용하여 데이터베이스에서 지식을 추출하는데 활용되고 있는 분야에는 특징점 추출, 데이터의 감축, 데이터의 이산화(discretization) 등과 같이 많은 응용범위를 가지고 있다. 러프집합을 이용하여 임의의 데이터집합에서 불필요한 특징을 제거함으로써 전체적으로 집합의 차원을 줄여줌으로써 문제를 보다 단순화하여 필요한 특징만 추출할 수 있다. 데이터의 양이 날로 증가함으로 인하여 각각의 제어대상에 대하여 어떤 속성이 필요하고 그렇지 않은지를 정확하게 측정하는 것은 매우 어려운 일이다. 이러한 이유로 데이터의 양을 먼저 감소시키는 것이 중요하며 이를 위해서는 데이터에 대하여 중요한 특징을 추출하기 위한 동정(identification)이 선행되어야 한다. 이로 인하여 중요한 속성들을 추출함으로써 잡음과 같은 불필요한 데이터를 제거하여 적합한 데이터에 대한 제어규칙을 설정할 수 있다^[9]. 따라서 러프집합 이론의 리덕트와 같은 동정을 통하여 제어 대상에 관계하는 데이터에서 추출한 특징은 전체 데이터에 대한 특징과 동일하게 된다. 결국 제어에 대한 규칙의 수와 제어시간을 최적화 할 수 있다.

그러나 정보시스템은 객체의 속성 값이 크리스프(crisp)하거나 실수 값을 가지는 경우와 속성 값의 알갱이성(granularity)으로 인한 부정확함 또는 속성 값의 중복과 같이 식별 불가능성과 애매성으로 인한 정보의 불확실성(uncertainty)을 내포하기 마련이다. 일반적으로 러프집합에서는 근사영역을 통한 정확도(accuracy)와 거침(roughness)을 이용하여 불확실성을 해결한다. 따라서 동치류의 알갱이성이 거칠어지면 특정한 객체에 대한 정보의 양이 줄어들게 된다. 그러나 하한근사 영역의 객체들에 대한 식별불가능성 관계에 있는 알갱이 성으로 인한 불확실성에 대한 문제해결이 필요하다. 러프집합에서는 근사공간을 이용한 정확도와 거침을 이용하여 데이터의 불완전성을 처리하였으나, 중복되거나 누락된 데이터의 처리등 모든 불완전 정보를 다루기에는 부족하다.

본 논문에서는 기존의 정보이론에서 이용되는 러프 엔트로피개념과 사후확률을 병행하여 식별불가능성으로 인한 불완전성을 해결하기 위해 러프집합의 불확실성에 대한 새로운 척도를 제안한다.

II. 러프집합 이론

1. 러프집합

임의의 정보 시스템을 $I=(U, A)$ 을 고려하자. 여기서 U 는 유한한 공백 없는 집합, A 는 $a \in A$ 에 대해 $a:U \rightarrow V_a$ 를 만족하는 공백없는 유한한 집합이고 V_a 는 속성 a 가 가질 수 있는 값들의 집합이다. 모든 $P \subseteq A$ 에 대해 동치관계(equivalence relation) $IND(P)$ 가 할당되는데 $IND(P) = \{(x,y) \in U \times U \mid \forall a \in P, a(x) = a(y)\}$ 로 표현 가능하다. 관계 $IND(P)$ 는 P -식별불능관계라고 한다. U 의 일부는 모든 $IND(P)$ 의 동등 클래스의 패밀리어며 $U / IND(P)$ 혹은 U / P 로 나타낼 수 있다. 결국, 만약 $(x,y) \in IND(P)$ 이면, x 와 y 는 P 의 속성들로 식별 불가능한(indiscernible) 것이다.

$IND(P)$ 는 P 의 모든 원소관계를 포함하고 동치관계끼리의 교집합에 대하여 닫혀있는 동치관계들의 최소의 집합이다. $P \subseteq A$ 이고 $P \neq \emptyset$ 이면 $\cap P$ 즉, P 에 해당되는 모든 동치관계들의 교집합도 역시 동치관계가 되며, 이를 $IND(P)$ 로 나타내고 x 의 동치 클래스는 다음과 같다.

$$[x]_{IND(P)} = \bigcap_{R \in P} [x]_R \quad (1)$$

러프집합의 이론에 의하여 $\underline{P}X$ 와 $\overline{P}X$ 를 각각 X 의 P -하한근사(lower approximation)와 P -상한근사(upper approximation)라고 하며 다음과 같이 정의할 수 있다.

$$\begin{aligned} \underline{P}X &= \bigcup \{x \in UP(X) : P(X) \subseteq X\} \\ \overline{P}X &= \bigcup \{x \in UP(X) : P(X) \cap X \neq \emptyset\} \end{aligned} \quad (2)$$

하한근사는 집합 X 에 확실히 속하는 모든 원소들로 구성되고 상한근사는 집합 X 와 교집합이 공집합이 아닌 원소들로서 X 에 속할 가능성이 있는 원소들로 구성된다. 그리고 하한근사와 상한근사와의 차는 경계영역으로서 어느 쪽에도 속하지 않는 애매한 영역으로 다음과 같이 정의할 수 있다.

$$BND(X) = \overline{P}X - \underline{P}X \quad (3)$$

집합 X 에 대한 지식의 불완전성의 정도를 나타내는 불완전성 척도는 다음과 같다.

$$\alpha_R(X) = \frac{\text{card } \underline{R}X}{\text{card } \overline{R}X}, X \neq \emptyset \quad (4)$$

이는 경계영역에서 발생하는 불완전성을 나타내는 방법이지만 식별불가능 관계에 의한 불확실성을 완전히 처리하지 못한다.

2. 리덕트와 코어

지식의 감축에 있어 가장 기본적인 역할을 하는 것은 리덕트(reduct)와 코어(core)의 핵심적인 두 개념이다. 리덕트의 속성 집합은 스스로 데이터베이스의 지식을 특성화 할 수 있는 속성들의 부분집합이다. 리덕트는 속성들의 부분집합 $RED \subseteq P$ 인데, 이것은 $[x]_{RED} = [x]_P$ 이다. 즉, 줄어든 속성 집합 RED 에 의해 유도된 동등 클래스들은 전체 속성 집합 P 에 의해 유도된 동등 클래스 구조와 같다. 속성 집합 RED 는 최소형태이다. 모든 속성 $a \in RED$ 에 대해 $[x]_{(RED)-(a)} \neq [x]_P$ 라는 의미에서, 즉, RED 집합에서 동등 클래스 $[x]_P$ 를 바꾸지 않고서는 그 어떤 속성도 없앨 수 없다. 결국, 리덕트는 부분집합들의 충분 집합(sufficient set)이라고 생각될 수 있다. 즉, 카테고리 구조를 대표하기에 충분하다.

코어(core)는 모든 리덕트들에 공통적으로 있는 속성들의 집합이다. 또한 모든 적합한 리덕트들이 가지고 있는 속성들의 집합이다. 그러므로 동등 클래스 구조를 바꾸지 않고는 인포 시스템에서 제거될 수 없는 속성들로 이루어져 있고 필수 속성들이다. 필수적이라는 것은 범주구조를 대표할 수 있다.

3. 러프 엔트로피와 사후확률

정보시스템의 구성에서 일관적이지 않은 객체들이 존재할 경우에는 러프집합에서 정의되어 있는 하한근사, 상한근사와 긍정영역과 같은 대수학적인 러프집합의 척도를 이용한 객체의 분류는 정확성이 떨어진다. 이를 극복하기 위한 일환으로 다른 척도를 이용하여 우수한 결과를 보장하기 위한 휴리스틱한 방법을 이용하고 있다. 결과적으로 임의의 러프집합에서 특징을 추출하기 위한 대수학적인 방법과 정보 이론적인 방법은 서로 다른 결과로 등가관계(equivalent relation)를 가지고 있지 않다. 따라서 이러한 단점을 보완해주는 러프 엔트로피(rough entropy)는 다음과 같이 정의된다.

[정의 1] $I=(U,C,D,f)$, $B \subseteq C$. $U/B=\{B_1, B_2, \dots, B_n\}$ 와 $U/D=\{D_1, D_2, \dots, D_m\}$ 라고 가정하자. 여기서 $D_j \neq \emptyset$ 이다. D 에 대한 B 의 러프 엔트로피는 다음과 같다^[10].

$$H_B(D) = - \sum_{i=1}^n \frac{|B_i|}{|U|} \sum_{j=1}^k \ln \frac{|B_i \cap B_j|}{|B_i|} \quad (5)$$

여기서 $|B_i|/|U|$ 는 U 에서 B_i 의 확률이고 $B_{i1}, B_{i2}, \dots, B_{ik}$ 는 B_i 와 D_1, D_2, \dots, D_m 간의 교집합이다.

만약 U 는 한 도메인이며, P 는 U 에 있는 조건 속성 집합이고 D 는 결정 속성이고 도메인 U 는 P 에 $\{D\}$ 에 대해서 일치하면 P 중의 한 속성 r 은 P 가 결정속성 D 에 대해서는 불필요한 것이다. 이의 필요충분조건은 $H(\{D\}/P) = H(\{D\}/P - \{r\})$ 이다. 이러한 러프 엔트로피에 의한 방법은 조건속성에 대한 동치류의 사전확률만을 이용하고 있다.

본 연구에서는 조건속성과 결정속성의 정보를 모두 고려하기 위하여 베이지언 사후확률을 적용하여 러프 엔트로피를 구한다. 따라서 결정 모순에 있는 경계영역 $BND_C(D)$ 의 객체에 대해서 조건속성과 결정속성의 연관관계의 여부를 결정하는 사후확률은 다음과 같이 계산한다.

$$H([x]_B | [x]_{B \cap D_j}) = \frac{-P([x]_B) * \ln P([x]_{D_j} | [x]_B)}{\sum_{k \in \{B, D_j\}} -P([x]_k) * \ln P([x]_{D_j} | [x]_k)} \quad (6)$$

여기서 $[x]_{B \cap D_j}$ 는 조건속성과 결론속성의 동치류의 범주값을 만족하는 객체가 조건속성과 결정속성의 동치류에서 발견될 확률이고, $[x]_B$ 는 조건속성의 동치류의 전체 객체에 대한 확률이다. 또한, $P([x]_{D_j} | [x]_B)$ 는 조건속성이거나 결정속성의 동치류가 주어졌을 때, 조건속성과 결정속성의 범주값을 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 연관관계를 고려하여 이러한 불확실성을 모델링을 할 수 있다.

IV. 적용사례

1. 문제의 설정

어떤 기업체의 신입사원 채용 여부를 판단하는 정보

시스템을 표 1과 같이 구성하였다. 여기서 구분은 신입사원 지원 대상자이며, 조건 속성에 해당하는 평가 조건은 대학성적(x), 창의성(y), 사회성(z) 및 인성(w)의 3개 영역으로 구분하였다. 그리고 결정속성에 해당하는 판단결과(D)는 평가 조건에 따라 결정되었다. 표 1에서 객체 9와 10은 동일한 조건 속성에 대하여 서로 다른 결론 속성을 가지므로 표 1에 존재하는 데이터는 일관성을 가지고 있지 않다고 할 수 있다^[11].

표 1. (x, t, z, w, D)의 의사결정표
Table 1. Decision Table of (x, t, z, w, D)

객체	조건속성				결정속성
	x	y	z	w	D
1	A	P	3	A	1
2	A	P	1	S	1
3	P	P	1	A	1
4	P	R	3	A	2
5	A	R	2	A	2
6	P	R	3	P	3
7	S	R	3	P	3
8	S	N	3	P	3
9	S	N	2	S	2
10	S	N	2	S	1

표 2. 표 1의 감축 의사결정표
Table 2. Reduced Decision Table of Table 1

객체	조건속성				결정속성
	x	y	z	w	D
1	A	P	3	A	1
2	A	P	1	S	1
3	P	P	1	A	1
4	P	R	3	A	2
5	A	R	2	A	2
6	P	R	3	P	3
7	S	R	3	P	3
8	S	N	3	P	3

이 경우에 결론부의 속성에 부합하는 하한근사와 상한근사를 통하여 이러한 비일관성의 데이터를 처리할 수 있다. 표 1에서 주어진 데이터의 기본 범주는 {1},{2},{3},{4},{5},{6},{7},{8},{9},{10}로 구성된다. 표 1에 존재하는 세 가지의 개념 즉, {1,2,3,10}, {4,5,9}, {6,7,8}에 대하여 10개의 객체에 대한 하한근사는 {1,2,3,4,5,6,7,8}이고 상한근사는 {1,2,3,4,5,6,7,8,9,10}이다. 따라서 경계지역은 {9,10}이 된다. 따라서 객체 9와 10을 제거하여 표 2와 같이 구성된다.

2. 제어규칙의 발생

지식 시스템은 객체에 대한 성질들의 집합으로 볼 수 있기 때문에 속성-값(attribute-value)표로 구성되는 지식 시스템을 구축할 수 있다. 또한 이러한 집합을 이용하여 집합을 구성하는 데이터와 데이터 감축에 대한 종속성을 진술한 식을 이용하여 구할 수 있다. 결국 집합을 이용하여 어떤 식의 정규형(canonical form) 표현을 나타낼 수 있고, 식의 진위(true/false)를 가리기 위하여 식별 불가능성을 도입할 수 있다. 이러한 식별 불가능성을 이용한 방법은 데이터의 감축과 분석을 위한 알고리즘으로 사용되고 있다. 이러한 데이터 표는 의사결정 논리라는 하나의 모델로 볼 수 있고, 이를 이용하여 지식 시스템에서 이용 가능한 데이터로부터 결론을 도출하여 우리가 원하는 추론을 수행하기 위하여 리덕트와 코어를 이용하여 제어규칙을 발생할 수 있다.

표 3. 속성 x 의 사후확률 러프 엔트로피
Table 3. Rough Entropy of Attribute x

규칙(x, D)	Posterior Rough Entropy
$(x=A) \Rightarrow (D=1)$	$-3/8 \cdot \ln(2/3) / (-3/8 \cdot \ln(2/3) - 3/8 \cdot \ln(2/3)) = 0.5$
$(x=P) \Rightarrow (D=1)$	$-3/8 \cdot \ln(1/3) / (-3/8 \cdot \ln(1/3) - 3/8 \cdot \ln(1/3)) = 0.5$
$(x=P) \Rightarrow (D=2)$	$-3/8 \cdot \ln(1/3) / (-3/8 \cdot \ln(1/3) - 2/8 \cdot \ln(1/2)) = 0.704$
$(x=A) \Rightarrow (D=2)$	$-3/8 \cdot \ln(1/3) / (-3/8 \cdot \ln(1/3) - 2/8 \cdot \ln(1/2)) = 0.704$
$(x=P) \Rightarrow (D=3)$	$-3/8 \cdot \ln(1/3) / (-3/8 \cdot \ln(1/3) - 3/8 \cdot \ln(1/3)) = 0.5$
$(x=S) \Rightarrow (D=3)$	$-2/8 \cdot \ln(2/2) / (-2/8 \cdot \ln(2/2) - 3/8 \cdot \ln(2/3)) = 0$
평균계	0.485

표 4. 속성 y 의 사후확률 러프 엔트로피
Table 4. Rough Entropy of Attribute y

규칙(y, D)	Posterior Rough Entropy
$(y=P) \Rightarrow (D=1)$	$-3/8 \cdot \ln(3/3) / (-3/8 \cdot \ln(3/3) - 3/8 \cdot \ln(2/3)) = 0$
$(y=R) \Rightarrow (D=2)$	$-4/8 \cdot \ln(2/4) / (-4/8 \cdot \ln(2/4) - 2/8 \cdot \ln(2/2)) = 1.0$
$(y=R) \Rightarrow (D=3)$	$-4/8 \cdot \ln(2/4) / (-4/8 \cdot \ln(2/4) - 3/8 \cdot \ln(2/3)) = 0.665$
$(y=N) \Rightarrow (D=3)$	$-1/8 \cdot \ln(1/1) / (-1/8 \cdot \ln(1/1) - 3/8 \cdot \ln(2/3)) = 0$
평균계	0.424

표 5. 속성 z 의 사후확률 러프 엔트로피
Table 5. Rough Entropy of Attribute z

규칙(z, D)	Posterior Rough Entropy
$(z=3) \Rightarrow (D=1)$	$-5/8 \cdot \ln(1/5) / (-5/8 \cdot \ln(1/5) - 3/8 \cdot \ln(1/3)) = 0.709$
$(z=1) \Rightarrow (D=1)$	$-2/8 \cdot \ln(2/2) / (-2/8 \cdot \ln(2/2) - 3/8 \cdot \ln(2/3)) = 0$
$(z=3) \Rightarrow (D=2)$	$-5/8 \cdot \ln(1/5) / (-5/8 \cdot \ln(1/5) - 2/8 \cdot \ln(1/2)) = 0.853$
$(z=2) \Rightarrow (D=2)$	$-1/8 \cdot \ln(1/1) / (-1/8 \cdot \ln(1/1) - 3/8 \cdot \ln(2/3)) = 0$
$(z=3) \Rightarrow (D=3)$	$-5/8 \cdot \ln(3/5) / (-5/8 \cdot \ln(3/5) - 3/8 \cdot \ln(3/3)) = 1.0$
평균계	0.512

근사화를 통하여 확보한 일관성은 데이터에 대하여 제어규칙을 발생한다. 먼저 표 2의 데이터에 대한 리덕트는 $\{x,z,w\}$, $\{x,y,w\}$, $\{y,z,w\}$ 이고, 코어는 w 가 되어 가장 중요한 속성이 된다. 따라서 w 를 제외한 다른 속성에 대하여 속성의 중요도(confidence factor)를 계산하여 데이터를 보다 감축할 수 있다. 조건속성과 결정속성간의 의사결정규칙 $x \Rightarrow D$ 의 중요도는 x 를 포함하는 객체의 수에 대하여 xUD 를 포함하는 객체의 수에 대한 사후확률 러프엔트로피로 정의된다. 따라서 표 2에서 w 를 제외한 x , y 와 z 속성에 대한 사후확률 러프엔트로피는 식(6)에 의해 각각 표 3, 표 4 그리고 표 5와 같이 계산할 수 있다.

표 6. 표 3의 감축된 의사결정표
 Table 6. Reduced Decision Table of Table 3

객체	조건속성		결정속성
	y	w	D
1	P	A	1
2	P	S	1
3	P	A	1
4	R	A	2
5	R	A	2
6	R	P	3
7	R	P	3
8	N	P	3

표 7. 표 3의 감축된 의사결정표
 Table 7. Reduced Decision Table of Table 3

객체	조건속성		결정속성
	y	w	D
1	P	A	1
2	P	S	1
4	R	A	2
6	R	P	3
8	N	P	3

결국 x , y , z 속성별 사후확률 러프엔트로피는 각각 0.485, 0.424와 0.512인 곳을 알 수 있다. 따라서 y 속성이 가장 적은 값을 가짐으로써 가장 중요한 속성이라는 것을 알 수 있다.

표 8. 표 3의 코어
 Table 8. Core of Table 3

객체	조건속성		결정속성
	y	w	D
1	P	-	1
2	P	-	1
4	R	A	2
6	-	P	3
8	-	P	3

결국 원래의 속성집합 $\{x, y, z, w\}$ 에서 $\{y, w\}$ 속성으로 감축이 이루어졌다. 이에 대한 데이터는 표 6과 같이 구축되어진다. 표 6에서 속성 값의 중복을 제거하면 표 7을 얻을 수 있다. 표 7에서 코어를 이용하여 규칙의 일관성을 유지할 수 있다. w 속성의 A 를 제거하면 데이터의 일관성을 확보할 수 있다. 또한 y 속성의 R 을 제거하면 역시 일관성을 유지할 수 있다. 결국 표 8과 같이 객체에 대하여 코어를 확보하여 최적의 감축을 수행할 수 있다. 결국 표 9와 같이 사후확률 러프엔트로피를 이용하여 데이터의 감축을 수행하여 제어규칙은 $IF y \Rightarrow P THEN D \Rightarrow 1$, $IF y \Rightarrow R$ 와 $w \Rightarrow A THEN D \Rightarrow 2$ 와 $IF w \Rightarrow P THEN D \Rightarrow 1$ 이 된다.

표 9. 표 3의 최종 근사규칙
 Table 9. Final Approximate Rules of Table 3

객체	조건속성		결정속성
	y	w	D
1	P	-	1
4	R	A	2
6	-	P	3

V. 결론

본 논문에서는 조건속성과 결론속성간의 연관관계에서 연관정도를 결정하는 사후확률 러프엔트로피를 정의하였다. 제안된 척도는 조건부 확률 엔트로피에 의한 척도와의 비교 분석에서 속성간의 영향력에 의한 변별력에서 향상된 결과를 보였다. 속성값의 중복성을 제안된 척도를 이용하여 데이터베이스의 속성 데이터에서 검색의 시간을 줄여서 사용자에게 보다 빠른 응답시간을 보장할 수 있도록 불필요한 속성을 제거하여 지식을 효율적으로 감축할 수 있었다. 제안된 방법은 지식의 감축, 데이터 마이닝과 다른 음성인식과 같은 속성감축과 같은 속성 감축과 특징점 추출과 같은 분야에 적용될 수 있다. 제안된 방법을 보다 방대한 양의 빅데이터에 적용하여 기존의 방법들과 비교우위를 논할 필요가 있을 것으로 사료되어진다.

References

[1] Williams, Graham J. and Simoff, Simeon J. "Data

Mining Theory, Methodology, Techniques and Applications(Lecture Notes in Computer Science/Lecture Notes in Artificial Intelligence)", Springer, 2007

- [2] Ramakrishnan, Naren and Grama, Ananth Y., "Data Mining: From Serendipity to Science", IEEE Computer August Vol. 34-37, 1999
- [3] Han, Jiawei, Kamber, Micheline, "Data Mining: Concepts and Techniques", San Francisco CA, USA, Morgan, Kaufmann, Publishers, 2001.
- [4] Hand, D.J., Mannila, H., & Smyth, P. "Principles of Data Mining", Cambridge, MA:MIT Press, 2001
- [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] Pawlak, Z., "Rough sets", International Journal of Information Sciences, 11, pp. 341-356, 1982
- [7] Pawlak, Z., "Using Variable Precision Rough Set for Selection and Classification of Biological Knowledge Integrated in DNA Gene Expression", Journal of Integrative Bioinformatics, Vol. 9, No. 3, pp.1-17, 2012
- [8] Pal S.K., Skowron, "Rough Fuzzy Hybridization: A new trend in decision making", Springer Verlag, Berlin, 1999
- [9] 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
- [10] Lin S., Jiucheng X., Zhan'ao X. and Lingjun Z., "Rough Entropy-based Feature Selection and Its Application", Journal of International Computational Science, Vol. 8, No. 9, pp. 1525-1532, 2011
- [11] Inkyoo P., "The generation of Control Rules for Data Mining", The Journal of Digital Policy and Management, Vol. 11, No. 11, pp. 343-349, 2013

저자 소개

박 인 규(정회원)



- 제10권 5호 참조
- 현 중부대학교 컴퓨터학과 교수
<주관심분야 : 연상처리, 인공지능>

최 규 석(중신회원)



- 제9권 6호 참조
- 1991년 ~ 1996년 : (주)SK텔레콤 중앙연구원 책임연구원
- 현 청운대학교 컴퓨터학과 교수
<주관심분야 : 인공지능, ITS, 이동컴퓨팅>

※ 본 논문은 2015학년도 청운대학교 교내학술연구비에 의하여 지원되었음.