확률적 러프 집합에 기반한 근사 규칙의 간결화

권 은 아[†] · 김 홍 기[††]

요 약

본 논문에서는 저장 데이터베이스의 정보 시스템을 정제하여 새로운 객체를 근사 추론하기 위한 규칙 생성에 관한 연구이다. 이 때 많은 수의 규칙 생성은 의사 결정자로 하여금 직관적인 판단을 어렵게 하며 의사 결정 시 부가되는 시간적인 단점도 있다. 그러므로 본 논문에서는 확률적 러프 이론에 기반하여 규칙을 최대한 간결화 하는데 주안점을 두었다. 제안하는 알고리즘은 러프 이론에 기반한 최적 리덕트를 생성하는 과정에 확률적 개념을 도입하여 리덕트 생성에서부터 어느 정도의 허용치를 부여함으로써 기존의 규칙 생성 알고리즘의 근사 결정 규칙을 보다 간결하게 표현할 수 있다. 이 과정에서 제안한 확률적 최소 리덕트 생성 알고리즘은 기존의 리덕트를 더욱 작게하여 추론에 필요한 조건 속성의 수를 최소화하였고 이는 확률적 근사 결정 규칙의 생성 과정에서 시간 복잡도에 따른 시간을 줄일 수 있다. 제안된 알고리즘을 이용하여 패턴 분류 문제에 표준적으로 사용되는 IRIS 데이터와 Wisconsin Breast Cancer 데이터에 대해 실험하였으며 허용된 분류율 하에서 규칙의 수와 간결함의 정도를 기존 알고리즘과 비교하였다.

Reduction of Approximate Rule based on Probabilistic Rough sets

Eun-Ah Kwon[†] · Hong-Gi Kim[††]

ABSTRACT

These days data is being collected and accumulated in a wide variety of fields. Stored data itself is to be an information system which helps us to make decisions. An information system includes many kinds of necessary and unnecessary attribute. So many algorithms have been developed for finding useful patterns from the data and reasoning approximately new objects. We are interested in the simple and understandable rules that can represent useful patterns. In this paper we propose an algorithm which can reduce the information in the system to a minimum, based on a probabilistic rough set theory. The proposed algorithm uses a value that tolerates accuracy of classification. The tolerant value helps minimizing the necessary attribute which is needed to reason a new object by reducing conditional attributes. It has the advantage that it reduces the time of generalizing rules. We experiment a proposed algorithm with the IRIS data and Wisconsin Breast Cancer data. The experiment results show that this algorithm retrieves a small reduct, and minimizes the size of the rule under the tolerant classification rate.

키워드 : 데이터마이닝(Data Mining), 근사추론(Approximate Reasoning), 러프 집합(Rough Set)

1. 서 론

보편화된 컴퓨터 사용으로 현실 세계에서 발생하는 많은 데이터들이 데이터베이스로 저장되고 있다. 이렇게 저장된 데이터베이스는 그 자체가 다음의 의사결정에 도움이 되는 정보가 담겨져 있는 하나의 정보시스템을 이룬다. 그러나 저장 데이터베이스에는 중복 데이터가 존재할 수 있고 결정 정보에 필요 없는 여러 요소가 있을 수 있으므로 데이터베이스를 정제하여 보다 직관적이고 이해 가능한 정보 시스템으로의 전환이 필요하다.

데이터 마이닝 (Data mining : DM)은 데이터 베이스로부터의 지식발견(Knowledge Discovery in Databases : KDD)이라고도 하는데 대규모의 데이터베이스 내에 숨겨져 있는 고급 정보를 추출해서 의사결정, 예측, 예보에 응용하고자 하는 기법으로 최근 2000년대의 데이터베이스 응용기술로 주목을 받고 있는 기술분야이다[1, 2]. 이러한 데이터 마이닝 연구분야중 하나인 분류는 과거에 발생되어 저장되어 있는 데이터베이스 정보로부터 새로운 정보 객체를 분류해 낼 수 있는 분류 규칙을 생성해 내는 기법이다[2, 3].

분류 규칙을 생성해내는 방법으로써는 통계(statistics), 신경망(neural network), 결정 트리(decision tree)등이 있으며 가장 많이 사용되는 것은 결정 트리에 기초한 트리분류기이다. 트리분류기[4-6]는 그룹의 값에 따라 그루핑하는 분류 규칙을 생성해내는 것으로 일단 트리가 만들어진 후에는 일반적으로 새로운 데이터를 분류하는데 쓰인다. 트리분류기는

† 정 회 원 : 주성대학 컴퓨터정보공학부 교수
†† 정 회 원 : 충북대학교 컴퓨터학과 교수
논문접수 : 2000년 8월 12일, 심사완료 : 2001년 4월 12일

루트 노드에서 결정에 가장 관계가 많은 속성을 택하여 그 속성으로 일단 분류하고 하위노드에서 그 분류를 결정할 수 없으면 그와 같은 과정을 되풀이 해 나감으로써 트리를 확장시켜 나간다. 이 때 속성 중에 수치 속성이 있다면 대부분 그 속성으로 하여금 얻어지는 정보 이득(information gain)이 크게 되므로 이 속성이 분류 규칙에 꼭 들어가게 된다. 그러나 이 속성이 다른 속성에 의해 꼭 필요하지 않을 수 있으므로 이러한 연구가 최근의 러프 집합으로 연구되고 있다[7-9].

Z. pawlak에 의해 소개된 러프 집합 이론[7-9]은 어떤 개념에 대해서 확실하게 그 개념에 속하는 하한 근사 공간과 속할 가능성을 가지는 상한 근사 공간을 집합을 통해서 나타낸다. 하한 근사 영역은 하나의 결정을 갖는 결정 데이터 영역이 되지만 상한 근사 공간에서 하한 근사 공간을 제외한 경계역이 비 결정성 데이터 영역이 된다. 이러한 비 결정성 데이터 영역에 있어서 대수적 러프 집합의 확장인 확률적 러프 집합의 정의를 도입하여 근사 결정 데이터 영역으로 확장한다.

정보 시스템의 기초가 되는 저장 데이터베이스에는 결정 정보에 필요 없는 여러 요소가 있을 수 있으므로 이를 정제할 필요가 있다. 이러한 정제 방법중의 하나로 리덕트 도출 방법[9]을 사용하고 있는데 이는 조건 속성에 대한 결정 속성 동치류의 하한 근사 카디날리티로 이들 정보 객체들의 조건 속성과 결정 속성간의 의존도를 결정하고 이 의존도에 기여하는 일련의 조건 속성의 집합을 구해 내는 것이다.

이 때 한 조건 속성 동치류가 정보 시스템에 나타나는 빈도를 표현하는 지지도(support)를 고려하지 않고 리덕트를 도출하는 방법은 측정 오차에 대한 고려를 하지 않는 것으로서 불필요한 속성을 배제하지 못하는 결과를 초래한다. 이는 정보시스템으로부터 감소 정보 시스템을 도출하는 데에 장애가 되며 많은 수의 규칙을 만드는 요인이 된다.

본 논문에서는 최대한 작은 감소 정보 시스템을 도출하기 위해 리덕트 도출방법에서부터 확률적 러프 집합의 이론을 도입한 확률적 속성 리덕트 생성 알고리즘을 제안한다.

제안된 알고리즘에 의한 확률적 속성 리덕트는 확률적으로 허용할 수 있는 최소 데이터 탐색 공간인 감소 정보 시스템을 얻을 수 있게 한다.

또한 이는 정보 시스템을 구성하는 여러 속성중에서 일반적인 연속 수치 값이 그 필요성[10-13] 에 의해 이산화 되었을 때 구간 설정에 따른 결정 충돌의 허용 범위를 확률적으로 허용함으로써 보다 효율적인 감소 정보 시스템을 얻을 수 있게 한다.

본 논문의 2장에서는 확률적 러프 집합과 정보 시스템 내에서의 확률적 러프 근사 공간에 대하여 정리하였으며 3장에서는 확률적 속성 리덕트 생성 알고리즘을 제안한다. 4장에서는 규칙생성에서의 확률적 러프 집합의 개념을 도입한 일반화 알고리즘을 제안하며 5장에서는 러프 집합에서의 음역에 대한 근사 추론 알고리즘을 소개하고 본 논문에서의 계층적 근사 추론 알고리즘을 설명한다. 6장에서는 제안된 알고리즘을 이용하여 IRIS데이터와 Wisconsin Breast Cancer 데이터에 대한 실험 결과를 비교하고 마지막장에서는 결론 및 향후 연구 방향을 제시한다.

2. 확률적 러프 집합

2.1 확률적 러프 집합의 정의

Z. pawlak에 의해 소개된 러프 집합 이론[7-9]은 어떤 개념에 대해서 확실하게 그 개념에 속하는 것과 속할 가능성을 가지는 것을 집합을 통해서 나타내고 있다. 이들 정보 객체들은 각 정보 객체를 나타내는 속성들에 의해 표현되며, 주어진 정보에 의해서 서로 구별할 수 없는 경우, 이들 정보 객체들이 구분 불가능한 동치(indiscernibility) 관계에 있다고 정의한다. 정보 객체 x, y, z가 동치 관계 R를 만족한다면, 이들은 다음 세 가지 성질을 만족한다.

1) 반사적(reflexive) : xRx
2) 대칭적(symmetric) : xRy -> yRx
3) 추이적(transitive) : xRy and yRz -> xRz

이러한 동치 관계에 의해 정보 객체 집단은 동치류(equivalence class)로 나뉘어 질 수 있으며, 이들 동치류내의 원소의 집합을 기본(elementary) 집합[U/R] 이라 하고, 이 기본집합에 의해 정의되는 집합 공간을 근사(approximation) 공간 Apr=(U, R)이라고 한다. 여기서 U는 정보 객체의 전체 집합이고, R 는 U상에 정의된 동치 관계를 나타낸다. 근사 공간상에 하나의 결정에 대해 정보 객체를 분류하는 경우, 동일한 기본집합 내에 있으면서도 서로 다른 결정을 나타내는 경우가 발생할 수 있다. 이러한 결정상의 불일치를 나타내기 위해서 러프 집합에서는 두 가지 근사를 정의한다. 하나는 결정에 의해 나타내어지는 개념 X에 항상 포함되는 기본집합으로 정의되는 하한 근사(lower approximation)이고 다른 하나는 개념 x와 일치하는 부분이 하나라도 존재하는 모든 기본집합으로 정의되는 상한 근사(upper approximation)인데, 이를 집합으로 나타내면 다음과 같다.

$$R_*(X) = \{x \in U \mid R(x) \subseteq X\}$$
$$R^*(X) = \{x \in U \mid R(x) \cap X \neq \emptyset\}$$

여기서 R(x)는 한 정보 객체x가 속한 기본집합, 즉 동치류를 나타낸다.

이와 같은 상. 하한 근사를 이용한 (U/R, ∪, ∩, ~ , $R_*(X)$, $R^*(X)$)를 Pawlak의 러프 집합이라 한다.

하한 근사에 속하는 원소는 전체 집합 내에서 개념 X를

분명하게 나타내고, 상한 근사에 속하는 원소는 개념 X를 표현할 수 있는 가능성이 존재하고 있음을 나타낸다. 따라서 이들 사이의 차집합 $R^*(X)$-$R_*(X)$는 개념 X를 애매하게 정의하는 원소들을 나타내는 경계 영역이 된다.

한 원소 x가 개념 X에 속하는 정도를 다음과 같은 러프 소속 함수 $\mu_X^R(x)=\frac{|X\bigcap R(x)|}{|R(x)|}$로 정의한다. 이렇게 정의되는 러프 소속 함수는 의 $0\le\mu_X^R(x)\le 1$ 범위를 갖는다. 앞에서 정의한 두 근사를 러프 소속 함수에 의해 정의하면 다음과 같다.

$$R_*(X)=\{x\in U \mid \mu_X^R(x)=1\}$$
$$R^*(X)=\{x\in U \mid \mu_X^R(x)>0\}$$

러프 집합에서 경계영역에 대한 모호성에 대해 다음 정의의 확률적 러프 집합에 기초하여 개념의 근사적 표현을 허용한다[9, 14].

[정의 1] Pawlak의 러프 집합(U/R, ∪, ∩, ~, $R_*(X)$, $R^*(X)$)의 확장인 확률적 러프 집합의 상하한 근사의 정의는 다음과 같다.

$$R_{*\beta}(X)=\{x\in U \mid \mu_X^R(x)\ge 1-\beta\}$$
$$R_{\beta}^*(X)=\{x\in U \mid \mu_X^R(x)>\beta\}$$

이때 인자 β는 근사 결정 규칙을 생성하기 위한 결정 임계치이다. □

이는 측정치의 오차나 소수의 측정치에 대한 하한 근사의 개념에 어느 정도 허용함을 나타낸다.

2.2 정보시스템

정보 시스템 S는 다음과 같은 구성요소로 구성되어 정보 시스템내의 지식을 표현하게 된다. S=(U, C, D, V, f)에서 U는 정보 객체 전체의 집합을 의미하고 C는 정보 객체를 나타내는 속성 중에서 조건 속성, D는 결정 속성을 나타낸다. 또 V는 각 속성의 도메인을 보이고 f는 정보 객체가 가지는 하나의 속성에 해당하는 값을 표현하는 정보 함수이다. 데이터베이스의 하나의 테이블은 이러한 정보 시스템의 한 형태이다.

정보 시스템에서의 결정 속성이 조건 속성에 어느 정도 종속(dependent)되어 있는가를 결정하기 위해 러프 집합 이론을 이용하여 다음과 같은 C에 대한 D의 하한 근사를 정의한다.

[정의 2] R(C)와 R(D)를 각각 조건 속성 C와 결정 속성 D의 동치관계를 만족하는 동치류의 집합이라고 할 때 러프 집합의 근사 공간 Apr = (U, R(C))내의 C에 대한 분할 $R(d_i)\in R(D)$의 하한 근사 $R_{*C}(d_i)$와 상한 근사 $R^*_C(d_i)$는 다음과 같이 정의된다.

$$R_{*C}(d_i)=\bigcup_j\{x\in R(c_j) \mid x\subseteq R(d_i)\}$$
$$R^*_C(d_i)=\bigcup_j\{x\in R(c_j) \mid x\bigcap R(d_i)\neq\emptyset\}$$ □

정보시스템에서의 근사 영역 역시 확률적 러프 집합에 기초하여 근사적 데이터 영역으로 표현하기 위해 C에 대한 분할 $R(d_i)\in R(D)$의 확률적 러프 집합의 하한 근사 $R_{*\beta C}(d_i)$와 확률적 러프 집합의 상한 근사 $R^*_{\beta C}(d_i)$는 다음과 같다.

$$R_{*\beta C}(d_i)=\bigcup_j\{x\in R(c_j) \mid \mu_{d_i}^R(x)\ge 1-\beta\}$$
$$R^*_{\beta C}(d_i)=\bigcup_j\{x\in R(c_j) \mid \mu_X^R(x)>\beta\}$$

위에서 정의한 각 분할 $R(d_i)$의 확률적 러프 집합의 하한 근사의 합집합 $U_{d_i\in R(D)}R_{*\beta C}(d_i)$을 C에 대한 D의 양역(POS : positive region)이라 하고 각 분할 R(di)의 상한 근사의 합집합에서 각 분할 R(di)의 하한 근사의 합집합을 제외한 영역을 경계역(BND : boundry region)이라 한다. 또한 전체 공간중에 양역이나 경계역에 속하지 않는 영역을 음역(NEG : Negative Region)이라 한다.

임의의 객체가 POS에 속하면 그 객체의 조건 속성에 의해 하나의 결정 클래스가 결정된다는 것을 의미하므로 POS의 객체 수가 전체 집합의 수와 같을 때 결정 속성은 조건 속성에 종속되어 있고 따라서 조건 속성에 의해 모든 객체가 결정지어지는 결정 정보 시스템이 만들어 진다. 본 논문에서는 조건 속성 C와 결정 속성 D간의 의존도를 나타내는 식을 [7, 9]에서 제안한 것으로 사용한다.

$$K(C,D)=\frac{card(POS_C(D))}{card(U)}$$

만약 K(C, D) = 1 이면 조건 속성 C와 결정 속성 D간의 의존도는 근사적 함수관계이고 K(C, D) = 0 이면 D에 있는 속성 값의 하나도 C의 속성 값으로부터 근사적으로나마 결정될 수 없음을 나타낸다.

3. 확률적 속성 리덕트

조건 속성 중에서 어떤 속성을 제외한 속성으로도 결정 클래스를 분류할 수 있을 때가 있다. 이렇게 제외할 수 있는 속성을 여분의 속성으로 취급한다. 여러 조건 속성을 가지고 있는 정보 시스템에서 여분의 속성을 제외한 필수 불가결한 속성을 찾아내는 것은 대단히 중요한 일이다.

[정의 3] K(C, D)를 조건 속성 c와 결정 속성 d 간의 의존도라할 때 속성B($\subset$ C)가 속성의존도 k(C, D)연산에 기초하여 D에 대한 C의 속성 리덕트가 될 필요 충분 조건은 다음과 같다.

(1) K(B, D) = K(C, D)
(2) K(B, D) ≠ K(B-{a}, D), a∈B □

대용량의 정보 시스템에서 결정 클래스간의 충돌을 해결하기 위해서 러프 집합을 확장한 확률적 러프 집합을 정의하였다. 허용치를 두고 있는 확률적 러프 집합에서의 리덕트의 결정으로 위의 정의는 만족할 만한 것이 아니다. [9]는 확률적 러프 집합에서의 β-속성리덕트를 다음과 같이 정의하였다.

[정의 4] 다음과 같이 정의된 러프집합의 하한 근사와 확률적 러프집합의 하한근사

$$R_{\bullet C}(d_i)=\bigcup_j\{x\in R(c_j)\mid x\subseteq R(d_i)\}$$

$$R_{\bullet \beta C}(d_i)=\bigcup_j\left\{ x\in R(c_j)\mid \frac{card(x\cap \overline{(d_i)})}{card(x)}\le \beta\right\}$$

를 각각 $\underline{C}D_i$와 $\underline{C}_\beta D_i$ 로 표현했을 때 속성B(⊂ C)가 결정 속성 D에 대한 C의 β-속성리덕트가 될 필요 충분 조건은 다음과 같다.

(1) $\forall D_i\in R(D),\ \underline{B}(\underline{C}_\beta D_i)=\underline{C}_\beta D_i$
(2) $\forall A\subset B, \exists D_i\in R(D),\ \underline{A}(\underline{C}_\beta D_i)\neq \underline{C}_\beta D_i$ □

이러한 속성 리덕트는 구분 메트릭(discernibility matrix)을 사용하여 구할 수 있다[7, 9].

예로써 <표 1>과 같은 정보시스템이 있다고 가정하자. a, b, s, d는 조건 속성이고 e는 결정 속성이다.

〈표 1〉 정보 시스템의 예

a	b	c	d	e
1	1	1	1	1
1	1	1	1	1
1	1	2	1	1
1	1	3	1	1
2	2	2	2	1
1	1	1	1	2
2	2	2	1	2
1	1	1	2	2
1	1	3	1	2

$\beta=\frac{1}{3}$으로 했을 때 [정의 4]에 의한 구분 메트릭으로 β 속성리덕트를 구하면 a, c, d 또는 b, c, d 이다.

〈표 2〉 속성리덕트 ($\beta = \frac{1}{3}$)

b	c	d	e
1	1	1	1
1	2	1	1
2	2	2	1
2	2	1	2
1	1	2	2

그러나 이것을 분류률을 기준으로 보았을 경우위의 예에서 속성 c는 여분의 속성이 된다. 즉 분류률을 고려한 속성 리덕트 a, d 또는 b, d가 된다.

〈표 3〉 분류률을 고려한 속성리덕트

b	d	e
1	1	1
2	2	1
2	1	2
1	2	2

그러므로 본 논문에서는 분류률을 고려한 속성리덕트를 구하기 위한 방안을 제안하고자 한다.

본 논문에서는 확률적 러프 집합에 근거하여 확률적 속성 리덕트에 대한 정의를 다음과 같이 정의한다.

[정의 5] K(C, D)를 조건 속성 c와 결정 속성 d 간의 의존도라할 때 속성B(⊂ C) 가 속성의존도 k(C, D)연산에 기초하여 D에 대한 C의 확률적 속성 리덕트가 될 필요 충분 조건은 다음과 같다.

(1) $K(B, D)\ge K(C, D)*\beta$
(2) $K(B-a, D)\ < K(C, D)*\beta,\ a\in B$ □

다음 (그림 1)은 위의 정의에 의해 본 논문에서 제안하는 하나의 확률적 속성 리덕트를 구하는 알고리즘이다.

이 때 사용되는 조건 속성에 대한 결정속성과의 관계중요도(significance value)로는 χ^2 값을 이용하는데 조건 속성에 대한 결정속성과의 관계중요도가 높은 속성이 결정속성과의 연관관계가 높음을 의미한다. 조건 속성에 대한 결정속성과의 관계중요도로 χ^2 값을 이용하는 대신에 정보이론에 기반한 정보 이득을 사용할 수 도 있다.

χ^2 값은 의 조건 속성에 대한 상이한 값($V_1, V_2, \cdots$)과 결정 속성의 동치류의 원소($D_1, D_2, \cdots$)와의 관계 원소수를 <표 4>로 나타냈을 때 $\chi^2=\sum_i\sum_j\frac{(n_{ij}-E_{ij})^2}{E_{ij}}$ 이다. 여기서 $E_{ij}=\frac{n_{i.}\times n_{.j}}{N}$ 이다.

〈표 4〉 객체의 분포도표

	V_1	V_2	⋯	V_j	⋯	
D_1	n_{11}					$n_{1.}$
D_2						$n_{2.}$
⋯						
D_i				n_{ij}		$n_{i.}$
⋯						
	$n_{.1}$	$n_{.2}$		$n_{.J}$		N

Algo. 확률적 속성리덕트 생성

input : Rl, set of attributes C, K(C, D).
output : A reduct (SM)
compute the significance value for each a∈C

```
sort the value
SM ← 0
while K(SM, D) < K(C, D) * β do
   select an a with highest value in C ;
   SM = a ∪ SM ;
compute K(SM, D)
end while
N = |SM|
for I = 0 to N-1 do
   remove ai from SM
   compute K(SM, D)
   if K(SM, D) < K(C, D) * β then
     SM = SM ∪ ai
   end if
end for
```

(그림 1) 확률적 속성리덕트 생성 알고리즘

4. 규칙의 일반화를 통한 확률적 근사 결정 규칙 생성

정보 시스템에서 확률적 속성 리덕트 생성 알고리즘에 의해 분류 정확도의 허용치를 둔 최소 결정 시스템을 결정하고 그 최소 결정 시스템의 하나의 투플이 결정 규칙으로 만들어진다.

최소 결정 시스템의 전체 공간은 서로소인 $POS_C(D)$, $NEG_C(D)$, $BND_C(D)$의 세 근사영역으로 구분된다. 임의의 투플이 어느 하나의 결정 클래스에 속하게 되면 그 투플은 $POS_C(D)$영역에 들게 되며 이 경우 그 투플이 그대로 다음과 같은 형태의 분류 결정 규칙으로 만들어진다.

$$\text{if } c_{i1} \text{ is } U_{i1} \text{ and } c_{i2} \text{ is } U_{i2} \text{ and } \cdots \text{ and } c_{in} \text{ is } U_{in} \text{ then } class_{di}$$

이렇게 만들어진 규칙 중에는 논리적으로 하나의 규칙이 다른 규칙을 포함하고 있을 수 있다. 즉 두 개의 규칙(r_i, r_j)의 조건부를 $cond(r_i)$, $cond(r_j)$라 하고 결론부를 $dec(r_i)$, $dec(r_j)$라 할 때 $cond(r_i) \supseteq cond(r_j)$이고 $dec(r_i) = dec(r_j)$이면 규칙 r_i는 규칙 r_j를 논리적으로 포함한다고 한다. 이때 논리적으로 포함되는 규칙은 제외될 수 있다.

또한 $cond(r_i) \supseteq cond(r_j)$이고 $dec(ri) \neq dec(r_j)$이면 규칙 r_i와 규칙 r_j는 결정 불일치 규칙이라고 정의한다.

본 논문에서는 러프 집합에서 결정 불일치 영역을 임계치를 두어 확률적 러프 집합으로 확장한 것과 같이 이 결정 불일치 규칙을 다음 [정의 6]과 같이 확률적 근사 결정 규칙으로 정의하고 확률적 포함 규칙으로 확장한다.

[정의 6] 두 개의 규칙(r_i, r_j의 조건부를 $cond(r_i)$, $cond(r_j)$라 하고 결론부를 $dec(r_i)$, $dec(r_j)$라 할 때 $cond(r_i) \supseteq cond(r_j)$이고 $dec(r_i) = dec(r_j)$이면 규칙 r_i는 규칙 r_j를 논리적으로 포함한다고 정의하고 $cond(r_i) \supseteq cond(r_j)$이고 $dec(r_i) \neq dec(r_j)$일 때 $supp(r_i)/(supp(r_i) + supp(r_j)) > \beta$ 이면 규칙 r_i는 규칙 r_j를 확률적으로 포함한다고 정의한다. 여기서 supp는 규칙을 지지하는 객체의 수이다. □

위에서 정의한 것과 같이 최대의 조건 속성을 제거하는 확률적 근사 결정 규칙 생성 알고리즘은 다음 (그림 2)와 같다.

```
Algo. 확률적 근사 결정 규칙 생성
input : 최소정보시스템의 규칙 Rule
output : 일반화된 규칙 MRule.
MRule = 0 ; N = |Rule|
for i = 0 to N - 1 do
   r = ri
   M = |r|
   규칙 r의 각 조건 속성에 대한 결정속성과의 관계중요도 SIG를
      계산. SIG(Ci) = p(Ci)(p(D|Ci) - p(D))
   SIG에 대해 조건 속성의 오름차순으로 정렬
   for j = 0 to M - 1 do
      규칙 r의 j번째 속성 aj 제거
      if r 이 다른 규칙 rn ∈ RULE 과 확률적 결정 불일치 then
         제거한 속성 aj를 다시 포함
      end if
   end for
   규칙 r에 논리적 포함인 규칙 r'∈ MRULE 제거
   if rule r 이 r'∈ MRULE에 논리적 포함이 아니면
      MRULE ← r∪MRULE
   end if
end for
```

(그림 2) 확률적 근사 결정 규칙 생성 알고리즘

이 알고리즘에 드는 복잡도를 계산해보면 a 의 조건속성과 n의 투플이 있을 때 한 규칙에 대해 SIG 계산에 O(an), 속성제거에 O(an)이므로 전체 투플에 대해 $O(2an^2)$이 된다.

또한 중복 규칙을 없애는데 $O(n^2)$이 되므로 위의 알고리즘에 대한 시간 복잡도는 $O((2an^2) + n^2) = O((2a + 1)n^2) = O(an^2)$이다.

이것은 속성 리덕트의 속성의 수 a과 최소 정보시스템의 투플의 수 n에 종속적이므로 앞장에서 제안한 확률적 리덕트에 의해 그 시간을 최소화 할 수 있다.

이로써 제안한 확률적 리덕트와 확률적 포함 관계로의 확률적 근사 결정 규칙은 대용량의 정보시스템에서의 데이터 마이닝에서 허용된 규칙의 임계치안에서 가장 간략화된 규칙으로 도출된다.

이 규칙으로 대용량의 정보 시스템에서 경계역의 결정을 허용된 오차 범위내에서 근사 추론된다.

본 논문에서는 분류율을 향상시키기 위해서 정보시스템에 나타나지 않은 음역에 대해 정보 시스템의 양역의 객체 추론에 기반하여 근사 추론 할 수 있도록 하였다.

[15]는 정보시스템의 음역 객체에 대해 퍼지 집합개념을 도입한 근사 추론 알고리즘을 제안하였다.

5. 퍼지 집합개념을 도입한 근사 추론 알고리즘

다음의 식은 클래스 di를 결론부로 하는 퍼지 규칙의 한 형태이다.

$$\text{if } c_{i1} \text{ is } U_{i1} \text{ and } c_{i2} \text{ is } U_{i2} \text{ and } \cdots \text{ and } c_{in} \text{ is } U_{in} \text{ then } class_{di} \quad (CF_i)$$

이러한 퍼지 규칙의 도출의 결정트리 생성 알고리즘에 퍼지 개념을 결합하여 퍼지 결정트리를 생성하고 퍼지 결정 트리로부터 퍼지 결정 규칙을 만든다[16].

한 객체의 입력 조건에 대한 각 규칙의 추론 값은 $\min_k(\mu_{U_{ik}}) * CF_j$으로 입력조건에 대한 퍼지 소속 값 중 가장 작은 값과 규칙의 CF를 곱한 값이된다. 그 중에 가장 큰 값을 갖는 결론 $concl_d = \max_j \min_k(\mu_{U_{ik}}) * CF_j$을 그 객체의 근사 결론으로 정한다.

이는 근사 결정의 max-min 방법으로 이 외에 max-product, max-average가 있다.

이 때 속성의 언어적 불확실성을 퍼지 소속 함수값으로 이용하는 대신 조건 속성이 결정 속성의 러프 집합에 속하는 정도의 확률적 러프 소속 함수값으로 이용하여 조건 속성의 합성방법에 의한 추론에 활용한다. 이는 한 객체의 임의의 속성 c_{ij}가 결정 분할 d_i에 완전히 속하는 어떤 집합 U_{ij}에 대한 근사의 정도를 의미하는 것으로 퍼지 규칙 생성과 같은 과정을 수행할 필요가 없다.

또한 속성의 합성연산 방법을 새로이 정의하지 않고 한 응용에 있어서 양역의 객체에 대하여 가장 영향력 있는 합성연산 방법을 택할 것을 제안하였다.

제안된 근사 추론 알고리즘은 정보 시스템내의 음역에 대한 객체에 대한 근사 추론으로 실제 도출된 규칙으로의 추론이 우선된 후에 이루어지는 계층적 데이터 근사 추론으로 사용된다.

본 연구에서의 계층적 데이터 근사 추론 알고리즘은 다음 (그림 3)과 같다.

```
Algo. 계층적 데이터 근사 추론의 알고리즘.
input : 정보시스템
output : 근사 결정
저장 정보 시스템으로부터 수치 속성에 대해 러프 소속 함수 값에
    의한 수치 이산화 알고리즘으로 러프 소속 함수와 이산화 구간
    을 설정한다.
확률적 러프 집합에 의한 양역의 최소 결정 시스템을 구한다.
if 새로운 객체가 최소 결정 시스템에 존재
  then
    그 객체의 결정을 채택한다.
  else
    최적 속성 합성 연산식 결정 알고리즘에 의해 선택된 합성 연
    산식에 의해 가장 큰 값을 갖는 결정을 채택한다.
end if
```

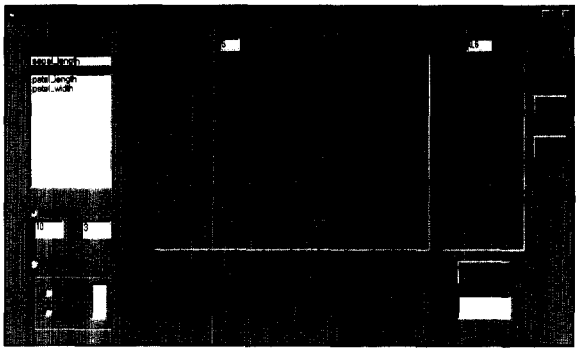
(그림 3) 계층적 데이터 근사 추론의 알고리즘

6. 실험 결과

본 연구에서 제안하는 확률적 리덕트 생성과 확률적 추론 규칙의 간략화 정도를 검증하기 위해 패턴 분류 문제에 표준적으로 사용하는 데이터인 IRIS 데이타와 Wisconsin Breast Cancer 데이터[17]에 대하여 실험하였다. 실험을 위한 구현 프로그램은 Windows 98환경에서 Visual Basic 6.0을 사용하였으며 정보 시스템을 위한 데이터베이스로는 Microsoft Access를 사용하였다. 실험 결과로는 확률적 속성 리덕트의 크기와 규칙의 수뿐만 아니라 앞장에서 제안한 계층적 근사 추론 알고리즘에 의한 분류율의 결과를 보인다.

실험에 사용된 IRIS 데이터와 Wisconsin Breast Cancer 데이터는 UCI 기계학습 데이터베이스로부터 얻을 수 있다.

IRIS 데이터는 setosa, versicolor, virginica의 3개의 클래스로 구성되어 있는 데이터로 꽃받침(sepal)의 길이(sepal_length)와 폭(sepal_width), 꽃잎(petal)의 길이(petal_length)와 폭(petal_width)의 수치적 특성으로 기술되어 있는 150개의 데이터로 되어있다. 이들 각 수치 속성에 대한 수치 속성 이산화 방법으로는 [15]에서 제안한 러프 소속 함수 값에 의한 수치 속성 이산화 방법을 택하였다. 아래 (그림 4)는 러프 소속 함수에 의한 속성 sepal_width의 이산화 알고리즘을 수행하는 과정의 그림이다.



(그림 4) 러프 소속 함수에 의한 이산화 알고리즘 수행 과정

다음 <표 5>는 본 논문에서 제안한 확률적 속성 리덕트의 크기와 규칙의 수 그리고 계층적 근사 추론 알고리즘에 의한 분류률의 결과를 러프 집합을 기반으로한 최소 결정 시스템에 의한 분류와 일반화 규칙에 의한 분류와 비교한 것이다.

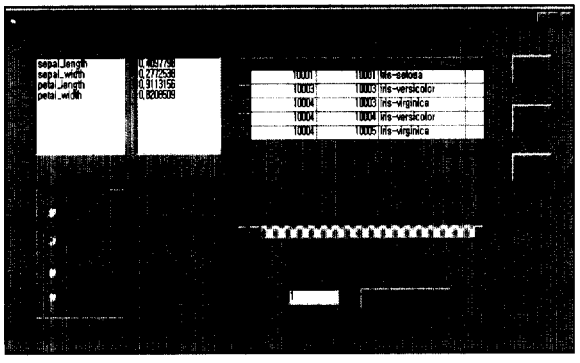
〈표 5〉 각 방법에 의해 생성된 규칙의 수와 분류율

저장 데이터수	항 목	최소결정 시스템에 의한 분류	일반화 규칙에 의한 분류	제안된 방법에 의한 분류 $\beta = 0.95$	제안된 방법에 의한 분류 $\beta = 0.93$
150개	리덕트수	3	3	2	2
	규 칙 수	14	6	4	4
	분 류 율	98.70%	98.0%	96.7%	96.7%
90개	리덕트수	2	2	1	1
	규 칙 수	5	5	3	3
	분 류 율	95.3%	98.0%	94.6%	94.6%

학습 데이터의 수로는 전체 데이터의 수인 150개와 90개를 가지고 실험하였다.

위 실험의 속성 리덕트 생성에 있어서 본 논문에서 제안한 확률적 러프 집합에 기반한 확률적 속성 리덕트를 도출할 수 있게 하기 위해 근사 정확도의 임계치를 부여할 수 있다. 임계치가 1인 경우 대수적 러프 집합을 기반한 속성 리덕트가 생성된다.

(그림 5)는 IRIS 데이터에서 각 조건 속성의 결정 속성에 대해 조건 속성에 대한 결정속성과의 관계중요도로 선택한 정보 이득 값을 보여주며 임계치를 1로 했을 때의 속성 리덕트와 그것으로 도출된 최소 결정 시스템을 보여준다.



(그림 5) 속성 리덕트 도출과 최소 결정 시스템

최소 결정 시스템으로 만들어낸 일반화 규칙은 다음과 같다. 여기서 10001, 10002,…은 수치값을 수치 구간화 알고리즘에 의해 이산화된 구간 값이다.

if petal_width is 10001 then Iris-setosa
if petal_length is 10002 then Iris-versicolor
if petal_length is 10003 and petal_width is 10002 then Iris-virginica
if petal_width is 10003 then Iris-versicolor
if petal_width is 10004 then Iris-virginica

실험 결과 리덕트의 수와 규칙의 수가 작아 진 것을 확인할 수 있다. 위의 IRIS 데이터는 그 데이터의 수가 작아서 우리는 다음의 Wisconsin Breast Cancer 데이터로 실험을 하였다.

Wisconsin Breast Cancer 데이터는 양성과 음성 종양 두개의 클래스를 나타내는 699개의 데이터로 구성되어 있다. 각 데이터는 조직의 두께(Clump Thickness), 셀 크기의 균일 정도(Uniformity of Cell Size), 셀 모양의 균일 정도(Uniformity of Cell Shape), Marginal Adhesion, Single Epithelial Cell Size, Bare Nuclei, Bland Chromatin, Normal Nucleoli, Mitoses의 9개의 수치적 특성으로 기술되어 있다. 이들 각 수치 속성에 대한 수치 속성 이산화 방법도 IRIS 데이터의 실험에서와 같은 러프 소속 함수 값에 의한 수치 속성 이산화방법을 택하였다. 실험에 사용된 데이터는 missing value를 가진 16개의 데이터를 제외한 683개로 실험하였다.

다음 <표 6>은 학습데이터의 수를 전체 데이터(683개)와 410개, 274개로 실험한 결과이다. 분류율은 전체 데이터(683개)를 학습 데이터로 도출된 규칙과 분류 알고리즘에 적용했을 때의 분류율 이다.

분류 결과 리덕트와 규칙의 감소를 가져옴으로써 대용량의 데이터에서 분류 정확도의 허용치 안에서 최소의 리덕트와 규칙을 도출함을 보인다. 확률적 러프 이론에 기반한 근사 규칙의 간결화 알고리즘과 음역에 대한 근사 추론 알고리즘을 사용 시 규칙의 수와 리덕트의 수가 훨씬 줄어든 반면에 분류율에서는 기존의 알고리즘에 비해 약간 낮게 나왔다. 그러나 이 분류율은 사용자가 지정한 분류오차의 허용치 내의 값으로 만약 분류율의 허용치를 높이면 분류율은 높일 수 있으나 규칙의 수와 리덕트의 속성 수는 많아질 것이다.

〈표 6〉 각 방법에 의해 생성된 규칙의 수와 분류율

저장데이터수	항목	최소결정 시스템에 의한 분류	일반화 규칙에 의한 분류	제안된 방법에 의한 분류 $\beta = 0.95$	제안된 방법에 의한 분류 $\beta = 0.93$
683개	리덕트수	7	7	5	2
	규칙수	168	24	14	5
	분류율	99.4%	99.4%	98.2%	96.0%
410개	리덕트수	6	6	4	3
	규칙수	93	23	12	7
	분류율	98.6%	97.2%	98.0%	97.2%
274개	리덕트수	5	5	2	1
	규칙수	54	13	5	3
	분류율	96.5%	96.2%	97.1%	96.6%

또한 객체와 속성의 수가 많은 대용량의 정보시스템에서의 리덕트의 수와 규칙의 수가 현저한 차이를 보임으로써 제안된 방법이 대용량의 정보시스템에서의 규칙 간결화에 많은 도움이 될 것으로 보인다.

7. 결론 및 향후 연구 방향

본 논문은 보편화된 컴퓨터의 사용으로 현실세계에서 발생하는 많은 저장 데이터로부터 유용한 정보를 추출해내는 데이터 마이닝에 대한 연구이며 새로운 객체에 대한 근사 추론에 관한 연구이다. 기존의 대표적인 데이터 마이닝 알고리즘의 경우 불필요한 속성이 포함될 가능성이 있으며 이를 러프 집합에 의한 속성 리덕트로 해결한다.러프 집합은 결정 불일치 부분에 대하여 어느 정도의 허용치를 두어 결정하는 확률적 러프 집합으로 확장 정의되었다. 본 논문에서는 이를 이용하여 속성 리덕트를 도출해 내는 과정에서 확률적 러프 집합에서의 허용치를 도입한 확률적 속성 리덕트를 추출하는 알고리즘을 제안하였다. 또한 이 과정에서 만들어지는 최소 결정 시스템으로부터의 규칙을 규칙의

일반화 과정을 통하여 가장 간단하면서도 이해할 수 있는 규칙으로의 생성 알고리즘을 제안하였다. 이러한 규칙은 러프 집합의 하한 근사와 경계역에 대한 것으로 학습 데이터에 포함되지 않은 음역의 객체 공간에서의 근사 분류는 고려되지 않았다. 따라서 본 논문에서는 양역에서의 분류 규칙을 이용하여 저장 데이터에 나타나지 않은 음역에 대한 적절한 속성 합성 함수를 찾아냄으로써 계층적 객체의 근사 분류를 추론해 내는 방법을 사용하였다.

실험 결과 리덕트의 수와 투플의 수가 감소하여 최소 결정 시스템을 구하는데 적절하였으며 기존 알고리즘과 비교하여 허용된 분류율하에서 규칙의 수와 간결함의 정도를 최소화하는 우수함을 보였다. 또한 객체와 속성의 수가 많은 대용량의 정보시스템에서의 리덕트의 수와 규칙의 수가 현저한 차이를 보임으로써 제안된 방법이 대용량의 정보시스템에서의 규칙 간결화에 많은 도움이 될 것으로 보인다. 향후 연구 과제로는 하나의 최소 리덕트를 추출해 내는 것만이 아니라 가능한 최소 리덕트 중에서 가장 효율적인 최소 리덕트를 도출해내기 위한 연구가 이루어져야 할 것이다.

참 고 문 헌

[1] Fayyad, U. M., Piatesky-Shapiro, G., Smyth, P., "From Data mining to Knowledge Discovery : An Overview," in Advances in Knowledge Discovery and Data Mining, Fayyac, U. M., Piatesky-shapiro, G., Smyth, P., pp.1-34, MIT Press, 1996.

[2] Chen. M. S., Han, J., and Yu, P. S., "Data Mining : An overview from Database Perspective," IEEE TKDE, Vol.8, No.6, 1996.

[3] Agrawal, R., et al., "An Internal Classifier for Database Mining Applications," Proceedings of the 18th VLDB Conference, 1992.

[4] Mehta, M., Agrawal, R. and Rissanen, J., "SLIQ : A Fast Scalable Classifier for Data Mining," Proc. of the Fifth Int'l Conference on Extending Database Technology, Avignon, France, March 1996.

[5] Quinlan, J. R., "Induction of Decision Trees," Machine Learning, 1, pp.81-106, 1986.

[6] Quinlan, J. R., *C4.5 : Programs for Machine Learning*, Morgan Kaufmann Publishers, 1993.

[7] Pawlak, Z., *Rough sets : Rough Sets : Theoretical Aspects of Reasoning About Data*, A Kluwer Academy Publisher, 1991.

[8] Pawlak, Z., "Rough Sets Present state and Further prospects," Intelligent Automation and Soft Computing, Vol.2, pp.96-102. 1996.

[9] Lin, T. Y. and Cercone, N., *Rough Sets and Data Mining : Analysis of imprecise data*, Kluwer Academic Publisher, 1997.

[10] Catlett, J., "On changing Continuous Attributes into Order Discrete Attributes," European Working Session on Learning. Springer-Verlag. pp.164-178, 1991.

[11] Kerber, R., "ChiMerge : Discretization of Numeric Attributes," Proceedings of AAAI-92, pp.123-128. 1992.

[12] Skowron, A. and Nguyen, H. S., "Quantization of Real Value Attributes," Proceedings of the Second Joint Annual Conference on Information Sciences, Wrightsville Beach, North Carolina, USA, 1995.

[13] Agrawal, R., Ghosh, S. ,Imielinski, T., Iyer, B. and Swami, A., "An Interval Classifier for Database Mining Applications," Proceedings of the 18th VLDB Conference Vancouver, British Columbia, Canada, 1992.

[14] Ziarko, W., "Variable Precision Rough Set Model," Journal of Computer and System Sciences, Vol.46, pp.39-59, 1993.

[15] 권은아, 김홍기, "Discretization of Continuous Valued Attributes and Approximate Reasoning based on Rough Membership Function," submitted.

[16] 민창우, 김명원, 김수광, "간결한 퍼지 규칙을 생성하는 데이터 마이닝 알고리즘", 정보과학회 논문지(B), 26권 11호, pp.1559-1565, 1999.

[17] http : //www.ics.uci.edu/~mlearn/MLRepository.html

권 은 아

e-mail : eunahk@jsc.ac.kr

1981년 이화여자대학교 수학과(학사)

1985년 이화여자대학교 대학원 수학과 (이학석사)

2000년 충북대학교 대학원 전자계산학과 (이학박사)

1994년~현재 주성대학 컴퓨터정보공학부 부교수

관심분야 : 데이터마이닝, 퍼지 및 러프 이론, 시스템 추론, 데이터모델링 등

김 홍 기

e-mail : hgkim@cbucc.chungbuk.ac.kr

1961년 연세대학교 수학과 졸업

1985년 중앙대학교 대학원 응용수학과 (이학박사)

1986년~1987년 캘리포니아 주립대학 교환교수

현재 충북대학교 컴퓨터과학과 교수

관심분야 : 퍼지 및 카오스이론, 계산이론, 신경회로망, 유전진화 알고리즘, 지능시스템, 자연어처리 등