러프 소속 함수를 이용한 수치 속성의 이산화와 근사 추론

(Discretization of Numerical Attributes and Approximate Reasoning by using Rough Membership Function)

권 은 아 † 김 홍 기 ††
(Eun-Ah Kwon) (Hong-Gi Kim)

요 약 본 논문에서는 저장 데이타베이스의 정보 시스템을 정제하여 이해 가능한 정보로 전환하고 새로운 객체를 근사 추론할 수 있도록 하기 위해 러프 소속 함수 값의 개념을 도입한 계층적 근사 분류 알고리즘을 제안한다. 제안하는 알고리즘은 근사 추론의 한 방법인 퍼지 추론 방법의 언어적 불확실성을 속성의 퍼지 소속 함수 값으로 나타내고 조건 속성의 소속 함수 값의 합성에 의해 근사 추론하는 방법을 이용하였으며 퍼지 소속 함수 값 대신에 러프 소속 함수 값을 이용하도록 제안하였다. 이는 퍼지 소속 함수 값을 이용하여 퍼지 규칙을 생성하는 과정을 생략할 수 있는 장점이 있다. 또한 정보 시스템 내의 속성 중에서 수치 속성에 대한 이산화 방법을 연구하고 이것 또한 러프 소속 함수 값과 정보이론의 무질서도의 개념을 이용한 수치 속성의 이산화를 제안하였다. 제안된 알고리즘을 이용하여 패턴 분류 문제에 표준적으로 사용되는 IRIS 데이타에 대한 실험결과 96%-98%의 분류율을 나타냈으며 다른 실험 데이타에서도 기존 알고리즘과 비교하여 수치 이산화나 근사 추론 모두 우수함을 보였다.

Abstract In this paper, we propose a hierarchical classification algorithm based on rough membership function which can reason a new object approximately. We use the fuzzy reasoning method that substitutes fuzzy membership value for linguistic uncertainty and reason approximately based on the composition of membership values of conditional attributes. Here, we use the rough membership function instead of the fuzzy membership function. It can reduce the process that the fuzzy algorithm using fuzzy membership function produces fuzzy rules. In addition, we transform the information system to the understandable minimal decision information system. In order to do, we study the discretization of continuous valued attributes and propose the discretization algorithm based on the rough membership function and the entropy of the information theory. The test shows a good partition that produce the smaller decision system. We experimented the IRIS data etc. using our proposed algorithm. The experimental result with IRIS data shows 96%-98% rate of classification.

1. 서 론

보편화된 컴퓨터 사용으로 현실 세계에서 발생하는 많은 데이타들이 데이타베이스로 저장되고 있다. 이렇게 저장된 데이타베이스는 그 자체가 다음의 의사결정에 도움이 되는 정보가 담겨져 있는 하나의 정보시스템을 이룬다. 그러나 저장 데이타베이스에는 중복 데이타가 존재할 수 있고 결정 정보에 필요 없는 여러 요소가 있을 수 있으므로 데이타베이스를 정제하여 보다 직관적이고 이해 가능한 정보 시스템으로의 전환이 필요하다.

데이타 마이닝 (Data mining : DM)은 데이타 베이스로부터의 지식발견(Knowledge Discovery in Databases : KDD)이라고도 하는데 대규모의 데이타베이스 내에 숨겨져 있는 고급 정보를 추출해서 의사결정, 예측, 예보에 응용하고자 하는 기법으로 최근 2000년대의 데이타베이스 응용기술로 주목을 받고 있는 기술분야이다[1,2]. 이러한 데이타 마이닝 연구분야중 하나인 분류는 과거에 발생되어 저장되어 있는 데이타베이스 정보

† 정 회 원 : 주성대학 컴퓨터계열 교수
eunahk@ns.jsc.ac.kr
†† 종신회원 : 충북대학교 전기전자컴퓨터공학부 교수
hgkim@cbucc.chungbuk.ac.kr
논문접수 : 2000년 4월 17일
심사완료 : 2001년 7월 14일

로부터 새로운 정보 객체를 분류해낼 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[4,5,6]는 그룹의 값에 따라 그루핑하는 분류규칙을 생성해내는 것으로 일단 트리가 만들어진 후에는 일반적으로 새로운 데이타를 분류하는 데 쓰인다. 트리분류기는 루트 노드에서 결정에 가장 관계가 많은 속성을 택하여 그 속성으로 일단 분류하고 하위 노드에서 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.

Z. pawlak에 의해 소개된 러프 집합 이론[7,8,9]은 어떤 개념에 대해서 확실하게 그 개념에 속하는 하한 근사 공간과 속할 가능성을 가지는 상한 근사 공간을 집합을 통해서 나타낸다. 이 때 조건 속성에 대한 결정 속성 동치류의 하한 근사의 합집합의 카디날리티로 이들 정보 객체(object)들의 조건 속성과 결정 속성간의 의존도(dependency)를 결정하고 이 의존도에 기여하는 일련의 조건 속성의 집합을 구함으로써 꼭 필요한 속성(리덕트)을 도출한다. 이 속성만으로 표현된 최소 결정 시스템의 각각의 투플이 하나의 규칙이 된다.

정보 시스템을 구성하는 투플은 여러 속성으로 구성되어 있으며 그 속성의 도메인에 따라 스칼라 속성(예 : 색상-빨강,노랑,…)과 연속 수치 속성으로 나뉘어진다. 대부분의 분류 규칙 마이닝 알고리즘은 스칼라 속성에 대해 고안되어 있으나 실생활에서의 정보는 연속 수치 속성이 더 일반적이고 연속 수치 값은 값의 범위가 너무 넓으므로 이 속성에 대한 이산화가 필수적이다[10, 11,12,13]. 일련의 수치 속성 이산화는 결정 클래스에 무관하게 설정되는 단점이 있고 결정 클래스를 고려한 구간 설정 알고리즘도 구간 설정에 따른 결정 충돌이나 최소 결정 시스템에 나타나지 않는 공간에서의 추론시 이를 해결할 방법이 필요하다.

따라서 본 논문에서는 수치 속성을 포함하는 정보 시스템에서의 수치 속성의 적절한 이산화 뿐만 아니라 최소 결정 시스템 내에 포함되지 않는 공간에서의 결정의 근사 추론을 위한 알고리즘을 제안한다. 근사 추론의 한 방법인 퍼지 추론 방법은 속성의 언어적 불확실성을 속성의 퍼지 소속 함수 값으로 나타내고 조건 속성의 소속 함수 값의 합성에 의해 근사 추론을 한다[14,16]. 본 논문에서는 이러한 퍼지 추론 방법을 이용하나 속성의 퍼지 소속 함수를 이용하는 대신에 각 조건 속성이 결정 속성의 러프 집합에 속하는 정도의 확률적 러프 소속 함수 값으로 이용하여 조건 속성의 합성방법에 의한 추론에 활용하도록 하였다. 이는 한 객체의 속성에 의해 하나의 결정 클래스에 완전히 속하는 하나의 집합에 대한 근사의 정도를 의미하는 것으로 퍼지 규칙 생성과 같은 과정을 수행할 필요가 없다.

본 논문의 2장에서는 러프 집합과 정보 시스템 내에서의 러프 근사 공간에 대하여 정리하였으며 3장에서는 기존의 수치 속성의 이산화 알고리즘을 설명한다. 4장에서는 러프 소속 함수 값에 의한 수치 이산화와 기존 알고리즘과의 비교를 논하며 5장에서는 근사 추론을 위한 제안 알고리즘을 설명한다. 6장에서는 제안된 알고리즘을 이용하여 IRIS 데이타에 대한 실험결과를 기존 알고리즘과 비교하며 마지막장에서는 결론 및 향후 연구 방향을 제시한다.

2. 러프 집합

2.1 러프 집합의 정의

Z. pawlak에 의해 소개된 러프 집합 이론[7,8,9]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 R 를 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 결정(decision)에 대해 정보 객체를 분류하는 경우, 동일한 기본집합 내에 있으면서도 서로 다른 결정을 나타내는

경우가 발생할 수 있다. 이러한 결정상의 불일치(inconsistency)를 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.

하위 근사에 속하는 원소는 전체 집합 내에서 개념 X를 분명하게 나타내고, 상위 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 영역(boundary region)이 된다. 그림 1은 러프 집합의 근사 영역을 나타낸 것이다

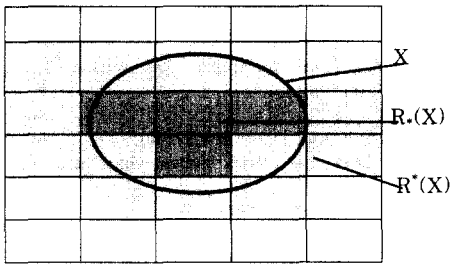

그림 1 러프집합의 근사영역

이 때 한 개념 X의 정확도는 상한 근사에 대한 하한 근사의 비로 표현될 수 있고 이를 개념 X의 근사 정확도라 하며 아래와 같이 정의한다.

$$\alpha_R(X)=\frac{|R_*(X)|}{|R^*(X)|}$$

여기서 $|X|$는 집합 X의 원소의 수를 나타내며 정확도는 $0\le\alpha_R(X)\le 1$의 범위를 가진다.

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

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

2.2 정보시스템

2.2.1 정보 시스템

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. 데이타베이스의 하나의 테이블은 이러한 정보 시스템의 한 형태로 다음 표 1은 mileage를 결정 클래스로 하는 자동차정보의 저장 테이블이다[9].

표 1 정보 시스템 예

make	fuel	disp	weight	cyl	power	turbo	comp	trans	mileage
usa	efi	medium	876	6	high	yes	high	auto	medium
usa	efi	medium	1100	6	high	no	medium	manu	medium
usa	efi	medium	1589	6	high	no	high	manu	medium
usa	efi	medium	987	6	high	no	medium	manu	medium
usa	efi	medium	1096	6	high	no	high	manu	medium
usa	efi	medium	867	6	high	no	medium	manu	medium
usa	efi	medium	1197	6	high	no	high	manu	medium
usa	efi	medium	798	6	high	yes	high	manu	high
usa	efi	medium	1056	4	medium	no	medium	manu	medium
usa	efi	medium	1557	6	high	no	medium	manu	low
japan	2-bbl	small	786	4	low	no	high	manu	high
usa	2-bbl	small	1098	4	low	no	high	manu	medium
usa	2-bbl	small	1187	4	medium	no	high	auto	medium
japan	efi	small	1023	4	low	no	high	manu	high
japan	efi	medium	698	4	medium	no	medium	manu	high
usa	efi	medium	1123	4	medium	no	medium	manu	medium
japan	efi	small	1094	4	high	yes	high	manu	high
japan	2-bbl	small	1023	4	low	no	medium	manu	high
usa	efi	medium	980	4	high	yes	medium	manu	medium
usa	efi	medium	1600	6	high	no	medium	auto	low
usa	efi	medium	1002	6	high	no	medium	auto	medium
usa	efi	medium	1098	4	high	no	medium	auto	medium
japan	efi	small	1039	4	medium	no	high	manu	high
usa	efi	small	980	4	medium	no	high	manu	high
usa	efi	small	1000	4	medium	no	high	manu	high

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 1] R(C)와 R(D)를 각각 조건 속성 C와 결정 속성 D의 동치관계를 만족하는 동치류의 집합이라고 할 때 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.

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

위에서 정의한 각 분할 R(di)의 하한 근사의 합집합 $U_{di \in R(D)} R_{*C}(d_i)$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 하여 그림으로 나타내면 다음 그림 2와 같다

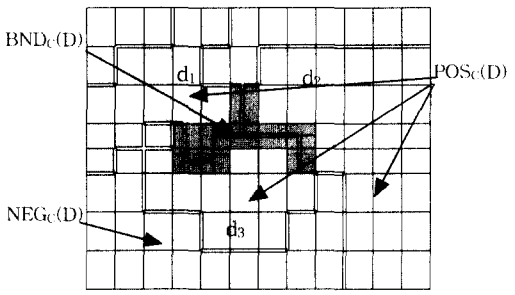

그림 2 정보 시스템의 근사 영역

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 종속(dependent)되어 있고 따라서 조건 속성에 의해 모든 객체가 결정지어지는 결정 정보 시스템이 만들어진다. 이것으로 조건 속성 C와 결정 속성 D간의 의존도를 나타내는 식은 다음과 같다[7,9].

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

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

위의 표 1에서의 조건 속성과 결정 속성간의 의존도는 1이므로 이 정보 시스템의 조건 속성으로 하나의 결정을 분류하는 데에는 아무 문제가 없음을 보인다.

그러나 표 1에서의 조건 속성 중에서 속성{fuel, disp, cyl}을 제외시킨 나머지 속성{make, weight, power, comp, trans}과 결정 속성간의 의존도도 역시 1임을 알 수 있다.

이것으로써 다음의 조건 속성 {fuel, disp, cyl}은 결정 속성을 나타내는 데에 꼭 필요한 속성이 아니라는 것이다.

이것으로 우리는 다음과 같은 정의를 생각할 수 있다[8,9].

[정의 2] 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 속성 리덕트가 될 필요충분 조건은 다음과 같다.

(1) K(B,D) = K(C,D)

(2) K(B,D) ≠ K(B-{a},D), a∈B □

속성 리덕트는 구분 메트릭(discernibility matrix)을 사용하여 구할 수 있다[7,9]. 모든 속성 리덕트를 발견하는 것은 NP-Complete 문제이며 하나의 리덕트를 생성하기 위한 시간 복잡도는 O(an + a log a)이다[7,9]. 다음 그림 3은 하나의 속성 리덕트를 구하는 알고리즘이다.

이 때 사용되는 significance value 로는 χ^2 값[1]을 이용하는데 significance value가 높은 속성이 결정 속성과의 연관관계가 높음을 의미한다

```
Algo. 속성 리덕트 생성
input  : R1, set of attributes C, K(C,D).
output : A reduct (SM)

compute the significance value for each a_i∈C
sort the value
SM ← 0
while K(SM, D) ≠ K(C,D) do
    select an a_i with highest value in C;
    SM = a_i ∪ SM;
    compute K(SM,D)
end while
N = |SM|
for I=0 to N-1 do
    remove a_i from SM
    compute K(SM,D)
    if K(SM,D) ≠ K(C,D) then
        SM = SM ∪ a_i
    end if
end for
```

그림 3 속성 리덕트 생성 알고리즘

2.3 근사 결정 규칙

정보 시스템에서 속성 리덕트 생성 알고리즘에 의해 최소 결정 시스템을 결정하고 그 최소 결정 시스템의

1) 표 4의 분포도표를 참조하여

$\chi^2 = \sum_i \sum_j \frac{(n_{ij} - E_{ij})^2}{E_{ij}}$, 여기서 $E_{ij} = \frac{n_{i.} \times n_{.j}}{N}$

하나의 투플이 결정 규칙으로 만들어진다.

최소 결정 시스템의 전체 공간은 그림 2와 같이 서로 소인 $POS_C(D)$, $NEG_C(D)$, $BND_C(D)$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 하나의 분류 결정 규칙으로 만들어진다. 그러나 결정 클래스가 하나 이상이면 $BND_C(D)$영역에 나타나고 이런 경우 다음 정의의 확률적 러프 집합에 기초하여 근사 결정 규칙을 생성한다[9,15].

[정의 3] Pawlak의 러프 집합$(U/R, \cup \cap \sim R_*(X), R^*(X))$의 확장인 확률적 러프 집합의 상하한 근사의 정의는 다음과 같다.

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

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

결과적으로 그림 2의 $BND_C(D)$의 영역이 이 확률적 러프 집합의 연산의 대상이 되며, α값의 설정에 따라 결정 규칙의 하한 근사 영역$POS_C(D)$이 되어 결정 규칙에 포함된다. 만약 정보 시스템내에 나타난 적이 없는 경우의 객체는 $NEG_C(D)$영역에 들게 되고 이러한 객체는 결정 불가능한 객체가 된다.

결정 규칙을 만들어 내는 것은 기존의 데이타에 의한 규칙 추출의 의미로도 중요하지만 새로운 객체에 대한 분류를 하는 것으로도 중요하다. 러프 집합에 근거한 분류는 실제 저장 데이타에서 나온 것으로만 상한 공간과 하한 공간으로 만들어 내는 것으로 저장 데이타에 포함되지 않은 영역은 모두 음역에 속하게 된다. 본 연구에서는 이러한 음역의 추론을 위한 방법을 연구하고자한다.

[9]는 규칙의 분류 정확도를 떨어뜨리지 않으면서 최대의 조건 속성을 제거하는 규칙의 일반화 방법을 제안하였으며 이것은 저장 데이타에 포함되지 않은 negative 영역을 포함한 최대의 영역에 대한 규칙으로 처리 될 수 있다. 따라서 규칙의 중복성과 불일치성을 다음과 같이 정의하고 그림 4와 같이 규칙의 일반화 방법을 제안하였다.

[정의 4] 두 개의 규칙(ri, rj)의 조건부를 cond(ri), cond(rj)라 하고 결론부를 dec(ri), dec(rj)라 할 때

cond(rj)⊇cond(rj)이고 dec(ri) = dec(rj)이면 규칙 ri는 규칙 rj를 논리적으로 포함한다고 정의하고 cond(rj)⊇cond(ri)이고 dec(ri)≠dec(rj)이면 규칙 ri와 규칙 rj는 결정 불일치 규칙이라고 정의한다. □

이 일반화에 드는 복잡도를 계산해보면 a의 조건속성과 n의 투플이 있을 때 한 규칙에 대해 SIG 계산에

Algo. 결정 규칙 생성을 위한 규칙의 일반화

```
input : 최소정부 시스템의 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 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 4 일반화 규칙 생성 알고리즘

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 $O((2an^2)\cdot n^2) = O((2a+1)n^2) = O(an^2)$이다.

이때 각 일반화된 규칙을 지지(support)하는 저장 데이타에 있는 투플의 수를 계산하여 이 지지값이 크면 규칙의 더 큰 확신(confirmation)을 나타낸다. 그러나 이것은 저장 데이타의 빈도수에 따른 분류이지 결정에 대한 속성의 관계를 표현하기에는 부족한 점이 있다.

이러한 관점에서 본 연구에서는 저장 데이타의 속성 합성을 통하여 그 저장 데이타를 가장 잘 표현할 수 있는 속성 합성을 찾아내어 근사 영역으로의 분류방법을 제안한다.

3. 수치 속성의 일반화

표 1의 정보 시스템에서 각 속성의 significant value를 구하고 그림 3의 속성 리덕트 추출 알고리즘에 의해 최소 결정 시스템을 구하면 다음과 같다.

그러나 이것은 수치 속성인 weight에 너무 의존적이므로 규칙의 가치가 없을 수 있다. 이렇듯 연속 속성은 값은 값의 범위가 너무 넓고 수치 값 하나 하나에 대한 규칙 생성은 값의 오차를 고려할 때 별 의미가 없을 수 있

표 2 정보 시스템에서 각 속성의 significant value(χ^2 값)

속성	χ^2 값
make	14.03509
fuel	0.67032
disp	10.80798
weight	47.71825
cyl	7.47783
power	7.46173
turbo	0.67032
comp	5.03154
trans	4.01785

표 3 표 1의 정보 시스템의 최소 결정 시스템

weight	disp	mileage
698	medium	high
786	small	high
798	medium	high
980	small	high
1000	small	high
1023	small	high
1039	small	high
1094	small	high
867	medium	medium
876	medium	medium
980	medium	medium
987	medium	medium
1002	medium	medium
1056	medium	medium
1096	medium	medium
1098	medium	medium
1098	small	medium
1100	medium	medium
1123	medium	medium
1187	small	medium
1197	medium	medium
1589	medium	medium
1557	medium	low
1600	medium	low

으므로 이 속성에 대한 일반화가 필수적이다. 즉 수치 범위를 이산화 하는 것으로 적절한 분할(good partition)과 분할의 수는 규칙의 수와 규칙에 포함되는 조건 속성의 수에 큰 영향을 미친다. 즉 구간의 수가 너무 많으면 규칙의 수가 너무 많아지고 구간의 수가 너무 적으면 규칙에 포함되는 조건 속성의 수가 너무 많아진다. 최소 결정 시스템의 구현에 대단히 중요한 역할을 한다.

본 장에서는 수치 속성에 대한 여러 가지 이산화 일고리즘을 살펴본다.

3.1 동등 간격 구간(equi-width-interval)과 동등 빈도 구간

이산화의 가장 보편화된 방법은 동등 간격 구간(equi-width-interval)과 동등 빈도 구간(equi-frequency-interval)이다. 동등 간격 구간은 [a,a+W], (a+W, a+2W], ... ,(a+(n-1)W,b], 여기서 a는최소값, b는 최대값이며 W=(b-a)/n 이다. 이때 n은 사용자에 의해 주어진 매개변수이다.

동등 빈도 구간은 근사적으로 같은 수의 값을 갖는 구간으로 나눈다.

이 두 방법은 대단히 간단한 반면에 조건 속성과 결정 속성간의 관계를 고려하지 않고 있다. 즉 다른 결정값을 가지고 있더라도 같이 그룹될 가능성이 많으며 그러한 점에서 어떤 대책을 고려치 않고 있다.

3.2 chiMerge 알고리즘

chiMerge 알고리즘[10]은 χ^2 통계치에 따른 연속 속성을 이산화하도록 고안하였다. 이것은 초기 단계와 상향(bottom-up) 머지 과정으로 구성되어 있으며 이 과정에서 사용자가 정의한 χ^2 임계치를 넘을 때까지 인접한 구간을 계속적으로 병합하는 과정으로 구간을 나눈다. 이 과정에서 최소 구간수와 최대 구간수, 두 개의 매개 변수를 사용한다. 이 때 χ^2 임계치가 너무 작으면 작은 구간들이 병합되지 아니하고 너무 크면 너무 많은 구간이 병합되므로 임계치 값의 영향이 크다.

3.3 D-2 algorithm

D-2 algorithm[11]은 결정 트리 알고리즘이 사용하는 [6]의 무질서도를 이용하여 무질서도를 최소화하는 구간으로 전체 구간을 이분화하고 그러한 이분화 방법을 계속 수행하여 각 분할에 의한 정보 이득(information gain)이 임계치를 넘으면서 최대 구간수를 초과하지 않는 동안 계속 분할하는 방법으로 이산화 한다. 그러나 이 방법은 하나의 속성에 대해 구간을 나누기 위해 매번 무질서도를 측정해야 한다.

3.4 interval classification

interval classification[13]은 우선 동등 빈도 구간으로 나눈 다음 각 구간에 대해 최다 결종 그룹(winner group)을 정하고 그것의 최다 결정 정도(strength)를 구하여 같은 최다 결종 그룹과 같은 최다 결정 정도를 가진 구간을 묶어서 처리하는 방법을 사용한다.

그러나 이것은 처음 구간을 나누는 동등 빈도 구간 설정에 의존적이며 최다 결정 정도가 약한(weak) 부분에서 너무 많은 구간으로 나뉠 가능성을 배제하지 못한다.

위의 여러 방법은 각기 장점을 가지고는 있으나 본 연구에서 의도하는 최소 결정 시스템에서 표현되지 않

는 공간에서의 추론을 위한 방안이 고려되지 않고 있다.

그러므로 본 연구에서는 수치 속성에 있어서 수치 속성의 효율적인 구간 결정뿐만 아니라 전체 공간을 표현할 수 있는 방안을 제안하고자 한다. 이를 위해서 속성의 언어적 불확실성을 퍼지 집합의 개념을 이용하여 정량적으로 표현하고 추론할 수 있는 수단으로 활용되는 퍼지 규칙에서의 추론 방안을 도입한다. 우선 4장에서는 러프 소속 함수 값에 의한 구간 이산화를 제안하고, 5장에서는 퍼지 규칙에서의 속성합성 방법을 4장에서의 러프 소속 함수 값으로 사용한 근사 추론을 제안한다.

4. 러프 소속 함수 값에 의한 구간 이산화

4.1 러프 소속 함수 구하기

러프 소속 함수를 구하기 위해 각 분할에 대해 수치 속성값이 가지는 러프 소속 값 $\{(x_i, \mu_{x_i})\}$을 구한다. 이때 연속 수치 값이나 수치 값의 범위가 너무 클 경우는 임의의 구간으로 버켓팅하여 그 구간에서의 값으로 결정한다. 보통 수치변량을 우선처리 하기 위한 구간 설정으로 동등 간격 또는 동등 빈도를 많이 사용한다. 러프 소속 값이 확률에 근거한 것이am로 일정빈도를 무시할 수 없으므로 본 연구에서는 우선 동등 빈도 처리를 우선한다. 이 때 그 구간의 대표값 $x_i = ave\frac{\sum(x_i * f_i)}{\sum(f_i))}$로 정하고 그 구간의 모든 x값은 xi로 대체되어 μ_{x_i}가 구해진다.

이렇게 구해진 $\{(x_i, \mu_{x_i})\}$를 두 변량사이의 상관관계로 하고 통계적 분석 방법으로 두 변량사이의 회귀 방정식을 구하여 이를 조건 속성이 결정 속성 값에 대한 러프 소속 함수 식으로 결정한다. 회귀 방정식은 최소자승 법을 사용하여 다음 식에서 보이는 것과 같은 최적(best fit) 함수 식을 구하였다.

일원 4차 회귀 방정식의 모델을 설정할 경우 이것은 $f(x) = a + bx + cx^2 + dx^3 + ex^4$의 형태가 되고 자료값 xi와 f(xi)에 대응하는 곡선 $f(x) = a + bx + cx^2 + dx^3 + ex^4$상의 $f(x_i)$의 값 $f(x_i) = a + bx_i + c\,x_i^2 + d\,x_i^3 + e\,x_i^4$과의 편차의 제곱합을 최소로 하는 계수(coefficient)를 구한다.

$\sum_{i=0}^{n} d_i^2 = \sum_{i=0}^{n}\{y_i - (a + bx_i + cx_i^2)\}^2$에서 다음 정규방정식을 만든다.

$$na + b\sum_{i=1}^{n} x_i + c\sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} y_i$$
$$a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} x_i^2 + c\sum_{i} x_i^3 = \sum_{i=1}^{n} x_i y_i$$
$$a\sum_{i=1}^{n} x_i^2 + b\sum_{i=1}^{n} x_i^3 + c\sum_{i} x_i^4 = \sum_{i=1}^{n} x_i^2 y_i$$

이 정규방정식을 연립으로 풀어서 우리가 구하고자하는 회귀 곡선을 구할 수 있다.

아래 그림 5는 표 1에서의 수치 속성 weight의 러프 소속 함수 그래프이다.

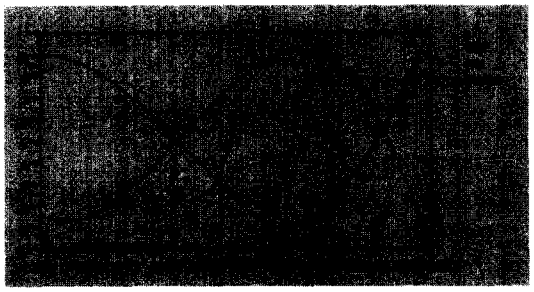

그림 5 표 1에서의 수치 속성 weight의 러프 소속 함수 그래프

4.2 정보이론에 근거한 효율적인 구간 결정

구간 결정에 있어서는 정보이론의 무질서도(entropy)에 근거한 정보 이득(information gain)과 이득 비율(gain ratio)를 이용한다.

무질서도는는 어떤 값에 따른 확률 분포가 나타내는 무질서의 정도를 보이는 것으로 결정 값에 따른 확률 분포 P가 $(P_1, P_2, ..., P_n)$일 때 이 분포가 보이는 info(P)를 P의 무질서도라 정의하였다. 이는 확률 분포 P가 (0.5,0.5)이라면 info(P)는 1이 되고 P가 (1.0)이라면 info(P)는 0이 되는 것으로 즉, 확률 분포가 균일하면 이 값은 큰 값을 가진다.

다음 표 4는 하나의 조건 속성에 대한 상이한 값 $(V_1, V_2, ...)$과 결정 속성의 동치류의 원소$(D_1, D_2, ...)$과의 관계 원소수를 표로 나타낸 것이다.

n_{ij}는 속성 V_j와 결정 D_i를 갖는 객체의 수이며, N은 전체 수를 의미한다. 이때 $n_{i.}$은 $\sum_j n_{ij}$ 이며 $n_{.j}$는 $\sum_i n_{ij}$ 이다.

표 4 객체의 분포도표

	V_1	V_2	...	V_j	...	
D_1						$n_{1.}$
D_2						$n_{2.}$
...						
D_i				n_{ij}		$n_{I.}$
...						
	$n_{.1}$	$n_{.2}$		$n_{.1}$		N

위의 표 4 에서 $P_{ij}=\frac{n_{ij}}{N}$, $P_{i.}=\frac{n_{i.}}{N}$, $P_{.j}=\frac{n_{.j}}{N}$, $P_{i|j}$를 $\frac{n_{ij}}{n_{.j}}$이라 하고 각 무질서도를 정의하면 다음과 같다.

1. $H_D=-\sum_i P_{i.}\log P_{i.}$는 결정 클래스의 무질서도이다.
2. $H_C=-\sum_j P_{.j}\log P_{.j}$는 하나의 주어진 조건 속성 값들의 무질서도이다.
3. $H_{DC}=-\sum_i\sum_j P_{ij}\log P_{ij}$ 는 결정 클래스와 조건 속성 값의 결합에 의한 무질서도이다.
4. $H_{D|C}=H_{DC}-H_C$ 는 주어진 조건 속성에서의 결정 클래스의 무질서도이다.

[5,6]은 객체의 클래스에 대한 주어진 속성이 나타내는 정보 값으로 아래와 같은 gain을 정의하였다.

[정의 5] 하나의 조건 속성에 의해 얻어지는 정보 이득은 위의 각 무질서도 정의 표현을 사용하여 다음과 같이 정의한다.

gain= H_D + H_C - H_{DC} = H_D - $H_{D|C}$ □

이는 조건 변수 x에 대해 정보 객체 t를 t1, t2,,,tn 으로 분류한다면 t의 한 요소의 클래스를 정의하기 위해 필요한 information(정보)은 다음 식과 같이 ti의 한 요소의 클래스를 정의하기 위해 필요한 information의 평균 가중치가 된다.

$$info(x,t)=\sum_{i=1}^{n}\frac{|t_i|}{|t|}*info(ti)$$

따라서 t의 한 요소의 클래스를 정의(identify)하기 위해 필요한 information과 결정 속성 x값이 주어진 후에 t의 한 요소의 클래스를 정의하기 위해 필요한 information의 차이는 결정 속성 d에 의한 정보 이득이 됨을 나타낸다.

또한 정보 이득이 많은 상이한 값을 갖는 속성에 잇어서 더 나은 값을 갖게되는 경향이 있음을 발견하고 다중치 속성에 대한 정보 이득의 치우침을 없애기 위해서 gain ratio를 다음과 같이 정의하였다.

[정의 6] 하나의 조건 속성에 의해 얻어지는 정보 이득 비율은 위의 각 무질서도 정의 표현을 사용하여 다음과 같이 정의한다.

gain ratio = gain/Hc □

이 논문에서는 수치 이산화에 있어서 최대의 정보 이득 또는 정보 이득 비율을 갖는 러프 소속 함수값(i)을 찾고 그 값을 기준으로 수치 속성을 이산화하는 방법을 택하였다. 이것은 각 이산화 방법들이 임계치를 외부에서 지정해야 하는 문제점을 해결하는 방법이 된다. 또한 수치값이 아닌 러프 소속 함수값을 기준으로 이산화하는 것으로 러프 소속 함수값의 범위가 [0,1]인 것을 감안하면 범위가 일정치 않은 수치값을 기준으로 이산화하는 것보다 훨씬 간단해 진다.

이산화 구간의 결정은 모든 결정에 대한 러프 소속 함수가 최대의 정보 이득 또는 정보 이득 비율을 갖는 러프 소속 함수값(i)과 교차하는 교차점(cp1,cp2,...)을 구하고 그 교차점으로 다음과 같이 이산화 한다. [cp0,cp1), [cp1,cp2), [cp2,cp3), [cp3,cp4), [cp4,cpn].

그림 6은 이 논문에서 제안하는 러프 소속 함수 값을 이용한 수치속성 이산화 알고리즘이다.

Algo. 러프 소속 함수 값을 이용한 수치속성 이산화.

input : 정부 시스템
output :러프 소속 함수, 이산화 구간.

for each 수치 속성에 대해
  러프 소속 값 $\{(x_i, \mu_{x_i})\}$을 계산한다.
  $x_j=ave\frac{\sum(x_j*f_j)}{\sum(f_j))}$ 계산한다.
  for each 결정 분할 D_i 에 대해
    러프 소속 함수의 회귀방정식
    $\mu^R_{D_i}(x)=a+bx+cx^2+dx^3+\ldots+ex^n$를 구한다.
  end for
  for i=0 to 1
    $\mu^R_{D_i}(x)=i$의 해에 의한
      객체의 분포도표를 구한다.
    computer significant value
  end for
  최대 gainratio를 갖는 i 선택
  $\mu^R_{D_i}(x)=i$의 해로 수치 구간 결정

그림 6 러프 소속 함수 값에 의한 구간 이산화 알고리즘

4.3 기존 연구와 비교

다음 표 5, 표 6은 [9]에서의 표 1의 정보 시스템 예에서 수치 속성 weight를 임의로 정한 수치 이산화 구간 설정과 그로 인한 최소 결정 시스템이다.

표 7과 표 8은 본 논문에서 제안한 러프 소속 함수 값에 의한 weight 수치 속성의 이산화 구간 설정과 그 구간에 의한 최소 결정 시스템이다.

표 5 전문가에 의한 weight 수치 값의 이산화

속성	이산화 값	수치 구간
weight	10001	0 … 800
	10002	801 … 1200
	10003	1201 … 1600

표 6 표 5의 이산화 구간 설정에 의한 최소 결정 시스템

make	weight	power	comp	trans	mileage
japan	10001	low	high	manu	high
japan	10001	medium	medium	manu	high
usa	10001	high	high	manu	high
japan	10002	high	high	manu	high
japan	10002	low	high	manu	high
japan	10002	low	medium	manu	high
japan	10002	medium	high	manu	high
usa	10002	high	high	auto	medium
usa	10002	high	high	manu	medium
usa	10002	high	medium	auto	medium
usa	10002	high	medium	manu	medium
usa	10002	low	high	manu	medium
usa	10002	medium	high	auto	medium
usa	10002	medium	high	manu	high
usa	10002	medium	medium	manu	medium
usa	10003	high	high	manu	medium
usa	10003	high	medium	auto	low
usa	10003	high	medium	manu	low

표 7 본 연구에서 제안한 러프 소속 함수값에 의한 weight 수치 값의 이산화

속성	이산화 값	수치 구간		
weight	10001	698	···	1058.8
	10002	1058.9	···	1509.8
	10003	1509.9	···	1600

표 8 표 7의 이산화 구간 설정에 의한 최소 결정 시스템

make	trans	weight	comp	mileage
japan	manu	10001	high	high
japan	manu	10001	medium	high
japan	manu	10002	high	high
usa	auto	10001	high	medium
usa	auto	10001	medium	medium
usa	auto	10002	high	medium
usa	auto	10002	medium	medium
usa	auto	10003	medium	low
usa	manu	10001	high	high
usa	manu	10001	medium	medium
usa	manu	10002	high	medium
usa	manu	10002	medium	medium
usa	manu	10003	high	medium
usa	manu	10003	medium	low

우리는 이것에서 리덕트의 수와 투플의 수가 훨씬 작아졌음을 발견할 수 있다.

같은 데이타로 D-2 방법과 IC(interval classifier) 방법에 의한 수치 이산화 구간 설정은 각각 표 9와 표10과 같으며 그로 인한 리덕트의 수와 투플의 수의 비교는 표 11과 같다.

표 9 D-2 알고리즘에 의한 weight 수치 값의 이산화

속성	이산화 값	수치 구간		
weight	10001	698	···	798
	10002	867	···	1096
	10003	1098	···	1600

표 10 초기 구간을 4±2, winer strength를0.6인 IC 알고리즘에 의한 weight 수치 값의 이산화

속성	이산화 값	수치 구간		
weight	10001	698	···	980
	10002	987	···	1039
	10003	1056	···	1600

표 11 기존 연구와의 이산화 결과 비교

이산화 방법	리덕트 수	투플 수
제안 알고리즘	4	14
D-2 알고리즘	6	18
IC 알고리즘	5	18

또한 IBM QUEST Project에서 무료로 배포하는 Synthetic data generator를 이용하여 생성된 data에 대한 실험의 결과 표 12와 같은 결과를 얻었다.

사용된 data의 수는 구현 시스템의 사정상 1000개의 data 생성에서 [13]에서 사용한 function 3을 적용하고 A그룹과 B그룹의 수를 동일하게 한 400개의 data로 실험하였다.

표 12 Synthetic data에 대한 이산화 결과 비교

이산화 방법	리덕트 수	투플 수
제안 알고리즘	3	27
D-2 알고리즘	3	31
IC 알고리즘	3	28

위의 결과로 제안 알고리즘이 다른 알고리즘에 비해 뒤떨어지지 않음을 알 수 있으며, 이 연구에서 제안한 알고리즘은 단순히 수치 이산화에만 사용되는 것이 아니라 러프 집합의 음역에 대한 객체 추론시 속성의 추

이 정보로 이용되어 추론율을 향상시키는데 도움이 되는 장점이 있다.

5. 계층적 데이타 근사 추론 방법

저장 데이타로부터 구해진 최소 결정 시스템에서 최적 속성 리덕트의 조건 속성 C의 동치류의 객체는 결정 분할 D의 동치류에 따라 그림 2의 세 가지 영역에 속한다. 결정 속성의 개수가 n개이고 각 속성이 갖는 상이한 값의 종류가 m개라면 전체 m^n개의 영역으로 나뉘어지게 되고 이 때 저장 데이타에 속하지 않는 영역은 전부 음역에 속하게 된다.

러프 집합에서는 하한 근사 공간을, 그리고 확률적 러프 집합 개념을 도입하여 상한 근사 공간에서의 근사 결정 규칙을 도출해 내지만 저장 데이타에 포함되지 않은 음역에 대한 근사 규칙을 추론해 내지는 못한다.

본 연구에서는 이러한 영역에서 퍼지 집합개념을 도입하여 근사 추론코자한다.

5.1 퍼지 추론

다음의 식은 클래스 d_i를 결론부로 하는 퍼지 규칙의 한 형태이다.

if c_{i1} is U_{i1} and c_{i2} is U_{i2} and ... and c_{in} is U_{in} then class d_i (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.

본 연구에서는 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 활용하도록 하였다.

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.

5.2 속성 합성 분류 함수 결정

본 논문에서는 속성의 합성연산 방법을 새로이 정의하지 않고 한 응용에 있어서 양역의 객체에 대하여 가장 영향력 있는 합성연산 방법을 택할 것을 제안한다. 합성 연산 방법으로 max-min, max-product, max-average방법 중에서 양역의 객체에 대한 결정에 대한 최고의 일치율을 보이는 연산식을 채택하기 위해 다음과 같은 과정을 수행한다.

Algo. 최적 속성 합성 연산식 결정 알고리즘

input : 정부 시스템
output : 최적 속성 합성 연산식

러프 집합에 의한 양역 객체에 대한 최소 정보 시스템을 구한다.
for each 합성 방법
최소 정보 시스템에 있는 객체에 대하여
각 속성의 러프 소속 함수 값의 합성 연산을 적용
count 일치율
end for
일치율이 가장 높은 합성 연산식을 채택

그림 7 최적 속성 합성 연산식 결정 알고리즘

5.3 계층적 데이타 근사 추론

러프 집합에서 하한 근사 공간에서의 분류와 확률적 러프 집합에서의 경계역에서의 객체 분류와 함께 이 연구에서의 제안하는 계층적 데이타 근사 추론의 알고리즘은 그림 8과 같다.

즉 정보 시스템으로부터 얻어지는 양역의 최소 정보

Algo. 계층적 데이타 근사 추론의 알고리즘.

input : 정부 시스템
output : 근사 결정

저장 정보 시스템으로부터 수치 속성에 대해 이 논문에서 제안한 러프 소속 함수 값에 의한 수치 이산화 알고리즘으로 러프 소속 함수와 이산화 구간을 설정한다.
러프 집합에 의한 양역의 최소 정보 시스템을 구한다.
if 새로운 객체가 최소 정보 시스템에 존재
then
그 객체의 결정을 채택한다.
else
수치 속성에 대해서는
러프 소속 함수의 회귀방정식
$\mu^{R}_{D_i}(x) = a + bx + cx^2 + dx^3 + \ldots + ex^n$에 의한 값계산
스칼라 속성에 대해서는
$\mu^{R(C)}_{R(D_i)}(x \in R(C_j)) = \frac{|R(D_i) \bigcap R(C_j)|}{|R(C_j)|}$ 계산
이 논문에서 제안한 최적 속성 합성 연산식 결정 알고리즘에 의해 선택된 합성 연산식으로 값 계산
가장 큰 값을 갖는 결정을 채택한다.
end if

그림 8 계층적 데이타 근사 추론의 알고리즘

시스템에서 새로운 객체에 대한 결정 정보를 얻어내고 만약 새로운 객체가 정보 시스템에서 가지고 있지 않은 음역의 객체이면 양역의 객체를 기준으로 구한 각 속성의 결정에 대한 추이 정보의 합성 연산식으로 그 결정을 추론케 하는 것이다.

6. 실험 결과

본 연구에서 제안하는 러프 소속 함수 값에 의한 계층적 근사 추론의 분류율이 기존의 알고리즘에 비해 높음을 검증하기 위해 패턴 분류 문제에 표준적으로 사용하는 데이타인 IRIS데이타[17]에 대하여 실험하였다. 실험에 사용된 IRIS데이타는 UCI 기계학습 데이타베이스로부터 얻을 수 있다.

iris 데이타는 setosa, versicolor, virginica의 3개의 클래스로 구성되어 있는 데이타로 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 데이타로 되어있다. 이들 각 수치 속성에 대해 본 연구에서 제안한 러프 소속 함수 값에 의한 수치 속성을 이산화하고, 이 이산화에 의한 러프 집합의 최소 결정 정보시스템을 구한다. 또한 150개의 데이타를 실험 데이타로 앞 절에서 제안한 계층적 근사 추론 알고리즘에 의한 분류율 성능을 평가한다. 저장 데이타에 따른 성능을 평가하기 위해 저장 데이타의 수를 21, 30, 60, 90개로 하여 실험하였다.

그림 9는 저장 데이타의 수가 90개 이때의 각 특성에 대한 러프 소속 함수 그래프이다.

그림 9의 러프 소속 함수로부터 본 연구에서 제안한 수치 이산화 알고리즘에 의한 각 수치 속성의 구간설정은 다음 표 13과 같다.

표 13 각 속성에 대한 수치 이산화 구간

속성	이산화 값	수치 구간		
sepal_length	10001	4.3	···	5.225
	10002	5.226	···	6.545
	10003	6.546	···	7.7
sepal_width	10001	2.0	···	2.84
	10002	2.85	···	3.44
	10003	3.45	···	4.4
petal_length	10001	1.0	···	2.14
	10002	2.15	···	4.99
	10003	5.0	···	6.7
petal_width	10001	0.1	···	0.46
	10002	0.47	···	1.66
	10003	1.67	···	1.78
	10004	1.79	···	2.5

위의 구간 분류에 의한 속성 리덕트를 구하고 그로 인한 최소 결정 시스템은 다음과 같다.

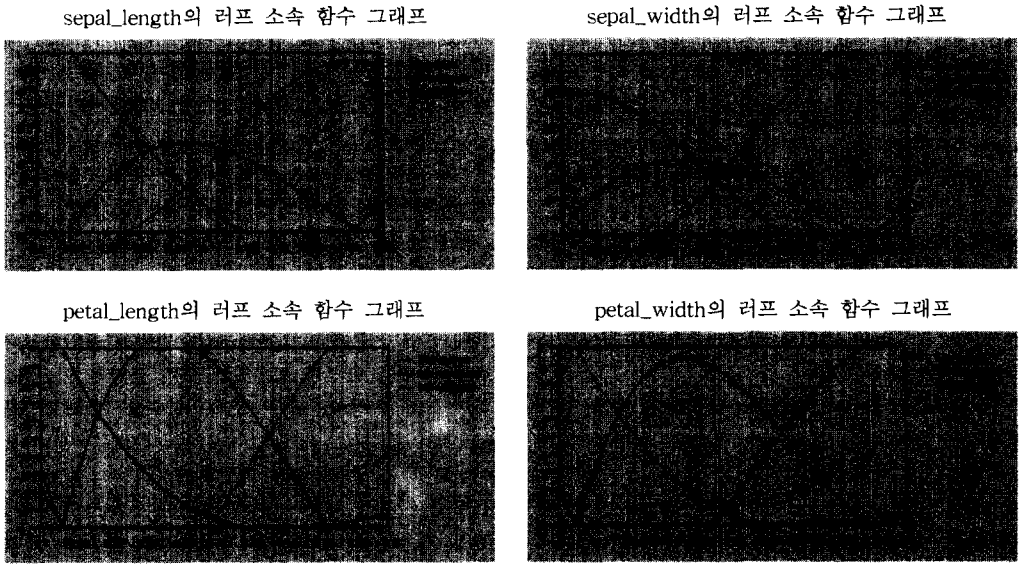

그림 9 IRIS 데이타 수치 속성의 러프 소속 함수 그래프

표 14 IRIS 데이타의 최소 결정 시스템

petal_length	petal_width	class
10001	10001	Iris-setosa
10002	10002	Iris-versicolor
10003	10002	Iris-virginica
10003	10003	Iris-versicolor
10003	10004	Iris-virginica

최소 결정 시스템으로 만들어낸 일반화 규칙은 다음과 같다.

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

다음 표 15는 본 논문에서 제안한 계층적 러프 소속 함수 값에 의한 객체 분류율의 결과와 퍼지규칙, 일반화 규칙에 의한 객체 분류율과 비교한 것이다. 앞의 두 알고리즘의 결과는 논문[16]에서 얻은 것이다.

표 15 IRIS 데이타에 대한 각 방법에 의해 생성된 규칙의 수와 분류율

저장 데이터의 수	Jang	간결한 퍼지규칙	일반화 규칙에 의한 분류	제안된 방법에 의한 분류
21	32 88.3%	3 91.3%	3 93.3%	3 96%
30	36 91.6%	3 95.3%	4 94.6%	6 96.6%
60	46 93.3%	3 96%	3 94.6%	5 97.3%
90	46 95.0%	3 96%	5 95.3%	5 98%

이것으로 본 연구에서 제안한 방법에 의한 분류의 분류율이 많이 향상된 것을 알 수 있다. 특히 저장 데이타 수에 따른 오차가 훨씬 적은 것으로 나타나 대용량의 정보 시스템으로부터의 마이닝 알고리즘으로나 근사 분류에 적합한 것으로 평가된다.

또한 Wisconsin Breast Cancer 데이타와 수치 구간 이산화의 실험에 사용된 synthetic data로 실험을 하였으며 그 결과는 표 16과 같다.

Wisconsin Breast Cancer 데이타는 양성과 음성 종양 두개의 클래스를 나타내는 699개의 데이타로 구성되어 있다. 각 데이타는 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 데이타의 실험에서와 같은 러프 소속 함수 값에 의한 수치 속성 이산화방법을 택하였다. 실험에 사용된 데이타는 missing value를 가진 16개의 데이타를 제외한 683개이다. 표 15의 C4.5와 CI의 결과는 [18]에서 얻은 것이며 ID3와 IC의 결과는 [13]에서 얻은 것이다.

표 16 실험 데이타에 대한 기존 방법과의 비교(규칙의 수와 분류율)

실험 데이타	제안된 방법에 의한 분류	C4.5	CI	ID3	IC
Wisconsin Breast Cancer	24 99.4%	94.8%	95.8%	-	-
synthetic data (function 3 적용)	27 98.0%	-	-	95.0%	94.5%

7. 결론 및 향후 연구 방향

본 논문은 보편화된 컴퓨터의 사용으로 현실세계에서 발생하는 많은 저장 데이타로부터 유용한 정보를 추출해 내는 데이타 마이닝에 대한 연구이며 새로운 객체에 대한 근사 추론에 관한 연구이다. 기존의 대표적인 데이타 마이닝 알고리즘의 경우 불필요한 속성이 포함될 가능성이 있으며 이를 러프 집합에 의한 속성 리덕트로 해결한다. 그러나 속성 리덕트로 해결한다고는 하나 저장 데이타에 포함되지 않은 음역의 객체 공간에서의 근사 분류는 불가능하다. 이의 해결로 규칙을 일반화 시켜서 해당 규칙의 지지율로 근사 분류를 추론 하지만 분류율이 높지 못하며 일반화 과정에 따른 시간이 소용된다.

따라서 본 논문에서는 양역에서의 분류 규칙을 찾아내고 이를 이용하여 저장 데이타에 나타나지 않은 음역에 대한 적절한 속성 합성 함수를 찾아냄으로써 계층적 객체의 근사 분류를 추론해 내는 방법을 제안하였다.

또한 수치 속성에 필수적인 수치 구간의 이산화를 위하여 각 속성의 러프 소속 함수를 구하고 최대의 gainratio를 갖는 러프 소속 함수 값으로 수치 구간을 이산화 하는 알고리즘을 제안하였다.

실험 결과 리덕트의 수와 투플의 수가 감소하여 최소 결정 시스템을 구하는데 적절하였으며 저장 데이타에 따른 분류 규칙의 도출과 새로운 객체에 대한 분류율에

서도 높은 결과를 가져옴을 알 수 있었다. 향후 연구 과제로는 양역의 최소 결정 시스템의 규칙의 수를 줄이며 규칙을 보다 간결하게 하기 위한 러프 규칙에 관한 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, pp. 866-883, 1996.

[3] 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, pp. 18-32, 1996.

[4] Breiman, L., Friedman, J. H., Olshen, R. A. and Stone, C. J., *Classification and Regression Tree*, Wadsworth, Belmont, 1984.

[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] Shan, N., Hamilton,H., Ziarko, W. and Cercone, N., "Discretization of Continuous Valued Attributes in Attrribute-Value Systems," Fifth Rough Sets, Fuzzy Sets, and Machine Discovery RFSD'96, pp. 74-81. 1996.

[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, pp. 560-573, 1992.

[14] Dlir, G. J. and Folger, T.I., *Fuzzy sets uncertainty and information*, Prentice Hall, pp. 188-192, 1992.

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

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

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

[18] Zijian Zheng, "Constructing Nominal X-of-N Attributes," Proceedings of the 14th IJCAI, Morgan Kaufmann, pp. 1064-1070, 1995.

권 은 아

1981년 이화여자대학교 수학과(학사). 1985년 이화여자대학교 대학원 수학과(이학석사). 2000년 충북대학교 대학원 전자계산학과(이학박사). 1994년 ~ 현재 주성대학 컴퓨터정보공학부 부교수. 관심분야는 데이타마이닝, 퍼지 및 러프 이론, 시스템 추론, 데이타모델링 등

김 홍 기

1961년 연세대학교 수학과 학사. 1975년 연세대학교 교육대학원 응용수학 교육학과(교육학 석사). 1985년 중앙대학교 대학원 응용수학(이학박사). 1980년 ~ 현재 충북대학교 전기전자컴퓨터공학부 교수. 관심분야는 퍼지 이론, 정보통신