한국 퍼지 및 지능시스템 학회 2002, Vol. 12, No. 3, pp. 202-209

러프집합과 계층적 분류구조를 이용한 데이터마이닝에서 분류지식발견

Discovering Classification Knowledge for Data Mining using Rough Sets and Hierarchical Classification Structure

이철희 · 서선학*
Chul-Heui Lee, Seon-Hak Seo*

강원대학교 전기전자정보통신 공학부
강원대학교 대학원 전기공학과*

요 약

본 논문은 제어 시스템에서 규칙기반과 데이터 마이닝에서의 분류규칙의 명료함에 대해 다룬다. 대용량의 데이터로부터 유용한 정보를 얻어내는 데이터 마이닝은 중요한 이슈가 되고 있다. 인공지능에 기반을 둔 데이터 마이닝 분류기법에는 신경망, 의사결정나무 등 여러 가지가 있지만 그 결과는 명확하고 이해하기 쉽고 분류규칙이 간단명료해야 한다. 러프집합이론은 불충분하고 비일관적인 데이터로부터 의미있는 지식을 추출하는데 효과적인 기법이고, 다양한 속성들을 효과적으로 사용함으로써 분류와 근사화에 대한 좋은 해법을 제시한다. 본 논문에서는 러프집합이론의 근사화를 이용하여 알갱이 속에 숨겨져 있는 지식들을 찾아내는데 있어 효과적인 접근을 하였으며, 최상위 레벨에 코어를 적용하여 계층적 분류를 함으로써 대량의 데이터를 효율적으로 처리할 수 있도록 하였다. 제안된 분류방법은 정보시스템의 해석을 용이하게 하고 최소의 분류규칙을 만든다.

Abstract

This paper deals with simplification of classification rules for data mining and rule bases for control systems. Data mining that extracts useful information from such a large amount of data is one of important issues. There are various ways in classification methodologies for data mining such as the decision trees and neural networks, but the result should be explicit and understandable and the classification rules be short and clear. The rough sets theory is an effective technique in extracting knowledge from incomplete and inconsistent data and provides a good solution for classification and approximation by using various attributes effectively. This paper investigates granularity of knowledge for reasoning of uncertain concepts by using rough set approximations and uses a hierarchical classification structure that is more effective technique for classification by applying core to upper level. The proposed classification methodology makes analysis of an information system easy and generates minimal classification rules.

Key words : Rough Sets theory, Hierarchical Classification Structure, Data Mining

1. 서 론

산업화 · 정보화와 더불어 여러 가지 형태로 저장되는 데이터의 양은 기하급수적으로 증가되어 왔다. 그러나 이러한 데이터의 무제한적인 증가는 우리가 원하는 정보를 찾아내는 일을 보다 어렵게 만들고 있는 것이 현실이다. 왜냐하면 우리는 데이터로부터 의미 있는 지식(knowledge)을 찾아내고자 하는 것이 목적인데 반하여 실제적으로는 오히려 데이터만 계속 쌓이고 있는 상황이기 때문이다. 따라서 대용량의 데이터로부터 의미 있는 지식을 찾아내는 데이터마이닝(data mining)은 현재 중요한 문제중의 하나가 되고 있다.[1]

접수일자 : 2001년 12월 18일
완료일자 : 2002년 6월 3일

인공지능에 기반을 둔 데이터마이닝의 분류기법에는 신경망, 의사결정나무 등 여러 가지 방법들이 있다.[1][2] 분류기법을 통해서 구하는 분류규칙의 수는 데이터마이닝의 실행 목표에 따라 다르지만 많은 규칙을 포함하는 모형보다 간단한 규칙으로 구성된 모형을 더 선호한다. 왜냐하면 비록 많은 규칙을 포함하는 모형이 더 정확하더라도, 기술 작업을 이해하고 설명하고 그리고 지식발견을 위해서는 간단한 규칙이 용이하기 때문이다.[3] 이러한 점에서 의사결정나무기법은 생성되는 분류규칙의 수가 많거나 모형을 구축하는데 사용되는 표본의 크기에 지나치게 민감하다는 단점이 있다. 그리고 신경망 기법의 경우 분류나 예측 결과만을 제공할 뿐 어떻게 그러한 결과가 나왔는가에 대한 이유를 설명하지 못하며, 입력변수의 수가 너무 많으면 망을 형성하는데 더욱 오랜시간이 걸리며 예측력도 감소한다.

본 논문에서는 기존 기법들이 안고 있는 문제점들을

개선하고자 방대하고 불분명한 자료 및 정보를 해석하는 데 있어서 여러 속성을 이용한 분류화 및 근사화를 효과적으로 제공하는 러프집합이론(Rough Sets theory)[4][5]을 사용하여 최소의 분류규칙을 발견함으로써 그 결과가 명쾌하고 쉽게 이해될 수 있도록 하였다. 사실 지식발견 과정에서 데이터 크기의 축소는 매우 중요하다. 데이터베이스에서 상황에 따라 많은 속성과 기록들이 있지만 지식발견에는 단지 몇 개의 속성들만이 실제적인 역할을 담당한다. 만일 필요없는 속성들이 제거될 수 있다면 데이터를 분석하는 데 있어 복잡도는 크게 감소될 수 있다는[6] 러프집합에 의한 접근은 이외에도 얻어진 결과에 대한 직접적인 해석을 제공하는 이점을 지니고 있다. 본 논문에서는 러프집합을 이용하여 효율적으로 분류규칙을 찾아내기 위해 계층적 분류구조를 사용하였다. 계층적 분류구조를 사용함으로써 객체(object)들에 대한 리덕트(reduct)를 구하는 계산의 양을 줄여 명확하고 효율적인 구조를 갖도록 하였다.

데이터로부터 분류규칙을 찾아내는 절차는 크게 속성의 감축과 분류규칙의 발견과정으로 구성되어있다. 전자의 속성감축의 과정은 조건속성(condition attribute)에 대한 판단속성(decision attribute)의 긍정영역(positive region)을 구해 판단결정과정에 있어 그다지 중요하지 않는 조건속성들을 발견하여 제거하고 리덕트를 구하거나 식별행렬(discernibility matrix)을 이용하여 코어를 찾아낸다. 후자의 분류규칙 발견과정은 코어를 계층적 분류구조의 최상위 단계에 적용함으로써 초기 분류 정확도를 향상시켜서 상대 리덕트를 좀 더 효율적으로 구할 수 있도록 하였으며 각각의 객체들의 속성값들간의 관계를 분석하고 불필요한 속성값들을 제거하여 최소의 분류규칙들을 유도하였다.

2. 러프집합이론[5],[7],[8]

같은 정보로써 특징화된 객체들은 그 객체들에 대한 유용한 정보들의 관점에서 볼 때 식별 불가능하다. 이렇게 생겨난 식별불가능 관계가 러프집합이론의 수학적 기초이다. 모든 식별 불가능한 객체들의 집합을 기본개념(elementary concept)이라 하며, 전체집합에 대한 지식의 기본적인 알갱이(granule, atom)를 형성한다.[9] 러프집합에 대한 자세한 내용은 참고문헌을 참조하도록 하고, 여기에서는 분류지식 발견에 이용되는 핵심 개념만 요약하였다.

2.1 식별불가능성(Indiscernibility)

유한집합이며 공집합이 아닌 전체집합 U와 속성집합 A가 주어졌다고 가정하자. 모든 속성 $a \in A$에 대해, 집합 V_a는 속성 a의 정의역이다. 정보시스템(information system) S는 $S=(U, A)$로 정의된다. A의 임의의 부분집합 B는 U상에서 하나의 이진관계(binary relation) I_B를 결정하는데, 이를 식별 불가능관계(indiscernibility relation)라 하며, 다음과 같이 정의된다. 모든 $a \in A$에 대하여,

$$xI_By \text{ if and only if } a(x) = a(y) \qquad (1)$$

여기서 $a(x)$는 x에 대한 속성값 a를 나타낸다. I_B는 동치관계(equivalence relation)이며, I_B의 모든 동치류들의 집합은 U/I_B 혹은 간단히 U/B로 나타내고, x를 포함하는 I_B의 동치류는 $B(x)$로 표현한다.

만일 (x, y)가 I_B에 속하면 x와 y는 B-식별불가능(indiscernible)이다. 관계 I_B의 동치류(혹은 분할 U/B의 블록)들은 B-기본개념(elementary concepts) 혹은 B-알갱이(granules)라고 부른다. 기본개념은 본질(reality)에 대한 지식의 기본 빌딩블럭(building blocks, concepts)이다.

2.2 근사화(Approximation)

부분집합 $X \subseteq U$와 동치관계 $B \in U/I_S$을 써서 두 집합 B-하한근사(lower approximation)와 B-상한근사(upper approximation)를 각각 다음과 같이 정의한다.

$$\begin{aligned} B_*(X) &= \bigcup_{x \in U}\{B(x) : B(x) \subseteq X\} \\ B^*(X) &= \bigcup_{x \in U}\{B(x) : B(x) \cap X \neq \emptyset\} \end{aligned} \qquad (2)$$

집합 $BN_B(X) = B^*(X) - B_*(X)$은 X의 B-경계영역(boundary region)이라 부른다. X의 경계영역이 공집합, $BN_B(X) = \emptyset$,이면 X는 B에 대해서 정확하게 분류되며, $BN_B(X) \neq \emptyset$이면 B에 대해서 러프(부정확)하다고 한다.

러프집합은 또한 러프 소속함수(rough membership function)을 사용하여 정의될 수 있다.

$$\mu_X^B = \frac{card(B(x) \cap X)}{card(B(x))} \qquad (3)$$

여기서, $\mu_X^B(x) \in [0, 1]$이다. 소속함수 $\mu_X^B(x)$의 값은 일종의 조건부 확률이며, x가 X에 속할 확실성의 정도(degree of certainty)로서 해석될 수 있다.

러프소속함수는 다음과 같이 근사화와 경계영역을 정의하도록 사용될 수 있다.

$$\begin{aligned} B_*(X) &= \{x \in U : \mu_X^B(x) = 1\}, \\ B^*(X) &= \{x \in U : \mu_X^B(x) > 0\}, \\ BN_B(X) &= \{x \in U : 0 < \mu_X^B(x) < 1\} \end{aligned} \qquad (4)$$

또한, 러프소속함수는 다음의 식처럼 일반화될 수 있다.

$$\mu(X, Y) = \frac{card(X \cap Y)}{card\ X} \qquad (5)$$

여기서, $X, Y \subseteq U, X \neq \emptyset$이고 $\mu(\emptyset, Y) = 1$이다.

함수 $\mu(X, Y)$는 러프 포함의 예이며, X가 Y에 포함되는 정도를 나타낸다. 즉 $\mu(X, Y) = 1$이면 $X \subseteq Y$이다.

2.3 리덕트(reduct)와 코어(core)

지식의 리덕트는 현재의 지식내에서 나타나는 모든 기본적인 범주들을 정의하기에 충분한 지식의 필수적인 부분이고, 코어는 어떤 의미에서 지식의 가장 중요한 부분이라 할 수 있다. $Q \subseteq P$가 독립이고 $U/I_Q = U/I_P$이면 Q는 P의 리덕트(reduct)라 하고 P는 여러 개의 리덕트를 가질 수 있다. P내의 모든 필요 불가결한 관계들의 집합을 P의 코어(core)라 하고 다음과 같이 표현할 수 있다.

$$CORE(P) = \cap RED(P) \tag{6}$$

코어의 개념은 두가지로 쓰일 수 있는데, 첫째 코어는 모든 러덕트에 포함되고 계산이 자명하므로 모든 리덕트의 계산을 위한 기초가 될 수 있다. 둘째 코어는 지식의 가장 특징적인 부분의 집합으로 해석될 수 있어서 지식을 감축할 때 빠트릴 수 없다.

2.4 지식의 종속도(Dependency of Knowledge)

속성집합 D의 모든 속성값이 속성집합 C의 속성값들을 유일하게 결정하면 D는 C에 완전 종속되며, $C \Rightarrow D$로 나타낸다. $C, D \subset A$에 대해서 C에 대한 D의 속성 의존도(dependency degree of attributes) k $(0 \le k \le 1)$는 다음과 같이 정의한다.

$$k = \gamma(C, D) = \frac{card(POS_C(D))}{card\ U} \tag{7}$$

여기서, $POS_C(D) = \bigcup_{X \in U/D} C_*(X)$을 C에 대한 U/D의 긍정영역(positive region)이라 한다. D의 C-긍정영역은 분류 U/C에 의하여 나타내어지는 지식을 이용하여 U/D의 동치류에 잘 분류될 수 있는 전체집합 U의 모든 객체들의 집합이다. k=1이면 D는 C에 완전 종속된다라고 하고, $0 < k < 1$이면 D는 C에 러프 종속된다 라고 한다. 또한 k=0이면 D는 C에 완전 독립이라고 한다.

3. 계층적 분류구조를 이용한 분류지식 발견

3.1 지식의 감축(Knowledge reduction)

정보시스템은 의사결정표(decision table)로 표현될 수 있다. 의사결정표는 어떤 조건들이 만족되었을 때 어떤 의사결정(행동)을 취해야 하는가를 알려주는 일종의 규칙이다. 대부분의 의사결정문제는 의사결정표로 정형화할 수 있으므로 이는 의사결정에 있어 매우 유용하다.

의사결정표의 간략화는 많은 응용에 있어 기본적으로 중요하다. 의사결정표의 조건속성의 감축은 지식의 감축으로 연결된다. 간략화된 의사결정표는 적은 수의 조건들로 똑같은 의사결정을 내릴 수 있다. 이러한 간략화를 통해 불필요한 조건을 확인하지 않아도 되고, 어떤 응용분야에서는 결국 나중에 간단한 방법으로 얻게 될 결론을 주는 고비용의 검사를 수행하지 않아도 된다.

의사결정표의 간략화 기법은 다음과 같은 3단계로 설명될 수 있다.

첫째, 불필요한 속성들을 의사결정표에서 제거하여 차원을 줄인다.

둘째, 같은 속성을 가지는 객체들을 통합한다.

셋째, 불필요한 속성값을 제거한다.

3.2 Granular Computing

전체집합의 과립화(granulation)는 비슷한 원소들을 알갱이로 그룹 짓는 것과 연관되어있다. 알갱이 관점에서 보면, 우리는 전체집합의 부분집합에 의해 나타내어지는 개념의 근사화를 다루고 있는 것이다. 러프집합에서 사용되는 알갱이 구조는 전형적으로 전체집합에 대한 분할이다.

U에 대한 이진관계는 Cartesian product $U \times U$의 부분집합으로서 해석될 수 있다. 집합의 포함관계를 사용하여 U에 대한 동치관계들의 순서를 정의할 수 있다. 만일 $E_1 \subset E_2$이면, 동치관계 E_1은 다른 동치관계 E_2보다 더 세밀(fine)하다고 할 수 있으며, 혹은 E_2가 E_1보다 거칠(coarse)다고 말할 수 있다. 동치 알갱이의 관점에서 보면, 더 세밀한 관계는 거친 관계보다 더 작은 알갱이들을 만든다.

모든 $x \in U$에 대하여, $[x]_{E_1} \subseteq [x]_{E_2}$ (8)

E_2의 동치 알갱이는 사실상 E_1의 몇몇 동치 알갱이들의 합집합이다. 동치관계 $E_1 \subset E_2$에 의해 유도된 러프집합 근사화들간의 관계는 다음과 같다.

$$\begin{aligned} &E_2(X)_* \subseteq E_1(X)_* \subseteq X \\ &X \subseteq E_1(X)^* \subseteq E_2(X)^* \end{aligned} \tag{9}$$

즉, 더 세밀한 동치관계는 더욱 단단한(tighter) 근사화를 유도한다. 정확성 척도 측면에서 보면, 이것은 다음을 나타낸다.

$$\alpha E_2(X) \le \alpha E_1(X) \tag{10}$$

여기서, $\alpha = \frac{card\ B_*(X)}{card\ B^*(X)}$이다. 하지만 이 관계의 역이 항상 성립하지는 않는다.

위의 내용을 두 개 이상의 동치 관계들로 확장시켜 m개의 동치관계들을 고려해보자.

$$E_1 \subseteq E_2 \subseteq \cdots \subseteq E_m \tag{11}$$

동치 알갱이들은 다음 조건을 만족한다.

$$[x]_{E_1} \subseteq [x]_{E_2} \subseteq \cdots \subseteq [x]_{E_m} \tag{12}$$

이러한 조건들은 전체집합에 대한 다층 알갱이구조(multi-layered granulation structure)를 유도하게 된다. 다른 계층의 알갱이들이 같을 수도 있지만, 하나의 분할은 다른 분할보다 더 세밀하거나 거칠게 된다. 일련의 러프집합 근사화는 다음을 만족한다.

$$\begin{aligned} &E_m(X)_* \subseteq \cdots E_2(X)_* \subseteq E_1(X)_* \subseteq X \\ &X \subseteq E_1(X)^* \subseteq E_2(X)^* \subseteq \cdots \subseteq E_m(X)^* \\ &\alpha E_m(X) \le \cdots \le \alpha E_2(X) \le \alpha E_1(X) \end{aligned} \tag{13}$$

3.3 계층적 분류구조와 과립화

본 논문에서 제안한 계층적 분류구조에서 전체집합에 대한 계층은 하나의 노드(node)가 하나의 클러스터(cluster)를 나타내는 나무 구조로써 설명될 수 있다. 개념적으로 계층은 전체집합 U의 연속적인 top-down 분해라고 볼 수 있으며, 반대로 계층적 구조는 작은 클러스터들로부터 큰 클러스터들로의 연속적인 bottom-up 결합으로 볼 수 있다. 계층적 구조에서 하위레벨 클러스터의 모든 원소들은 상위 클러스터와 뿌리 사이에 있는 모든 노드들에 포함된다. 최하위 레벨까지 계층화시킬 경우 모든 객체들이 정확하게 속성들을 가지고 분류규칙을 찾아낼 수 있으나, 이러한 것은 아무런 의미가 없는 일이다. 즉, 최하위 레벨에서의 분류규칙 수는 객체 수만큼

생겨날 것이며, 또한 하나의 분류규칙마다 속성의 개수만큼의 조건이 포함되어야 할 것이다. 따라서 본 논문에서는 전체집합을 의사결정표로 나타낸 후 속성에서 코어를 찾아내어 그 코어를 가지고 전체집합을 계층적으로 분류한다. 많은 레벨로 계층을 나눌 경우 분류규칙의 수가 커지기 때문에 뿌리와 코어를 이용해 나누어진 클러스터들만을 가지고 분류규칙을 찾아내게 된다. 분류를 코어를 이용해 나누는 이유는 코어가 객체들을 나타내는 가장 핵심적인 속성이기 때문이다.

다음의 의사결정표의 예를 생각해보자.

표 1. 정보시스템의 의사결정표
Table 1. Decision table

Object	a_1	a_2	a_3	a_4	a_5	a_6	a_7	D
1	2	2	3	1	1	4	1	1
2	1	1	4	2	2	2	2	1
3	2	4	2	2	2	4	2	2
4	2	1	4	3	2	2	3	1
5	2	2	3	3	3	3	3	2
6	4	4	1	3	2	4	3	3
7	1	1	4	3	2	1	3	1
8	3	3	2	3	3	4	3	3
9	4	4	1	4	3	4	4	2
10	2	2	2	2	2	4	2	1
11	2	2	3	2	2	3	2	1
12	4	4	1	4	4	4	4	2
13	3	3	2	4	4	3	4	1
14	3	3	2	2	2	4	2	2
15	1	2	1	3	4	2	1	1
16	1	2	1	3	4	2	1	3

표 1의 정보시스템은 16개의 규칙, 7개의 조건부속성 C= $\{a_1, a_2, a_3, a_4, a_5, a_6, a_7\}$ 그리고 3개의 클래스를 가지는 1개의 의사결정부속성 D= $\{D\}$을 가지고 있다. 이 의사결정표에서 동치류를 갖는 동치 관계들의 집합을 구하면 다음과 같다.

$U/I_{a_1}=\{\{2,7,15,16\},\{1,3,4,5,10,11\},\{8,13,14\},\{6,9,12\}\}$

$U/I_{a_2}=\{\{2,4,7\},\{1,5,10,11,15,16\},\{8,13,14\},\{3,6,9,12\}\}$

$U/I_{a_3}=\{\{6,9,12,15,16\},\{3,8,10,13,14\},\{1,5,11\},\{2,4,7\}\}$

$U/I_{a_4}=\{\{1\},\{2,3,10,11,14\},\{4,5,6,7,8,15,16\},\{9,12,13\}\}$

$U/I_{a_5}=\{\{1\},\{2,3,4,6,7,10,11,14\},\{5,8,9\},\{12,13,15,16\}\}$

$U/I_{a_6}=\{\{7\},\{2,4,15,16\},\{5,11,13\},\{1,3,6,8,9,10,12,14\}\}$

$U/I_{a_7}=\{\{1,15,16\},\{2,3,10,11,14\},\{4,5,6,7,8\},\{9,12,13\}\}$

$U/I_C=\{\{1\},\{2\},\{3\},\{4\},\{5\},\{6\},\{7\},\{8\},\{9\},\{10\},\{11\},\{12\},\{13\},\{14\},\{15,16\}\}$

$U/I_D=\{\{1,2,4,7,10,11,13,15\},\{3,5,9,12,14\},\{6,8,16\}\}$

조건부속성에 대한 의사결정부속성의 긍정영역을 찾아보면

$$POS_C(D)=\{1,2,3,4,5,6,7,8,9,10,11,12,13,14\}$$

임을 알 수 있다. 따라서 비일관적인 데이터인 규칙 15, 16을 제거하여 일관성이 있는 의사결정표를 구성한다. 다음으로 조건속성에 대해서 속성을 하나씩 제거해 나가면서 불필요한 속성인지 필요한 속성인지를 확인해 리덕트(reduct)와 코어(core)를 구하게 된다. 다른 방법으로는 식별행렬을 이용하여 구할 수 있다. 결과적으로 4개의 리덕트 $\{a_2, a_4\}, \{a_2, a_7\}, \{a_1, a_2, a_5\}, \{a_2, a_5, a_6\}$를 얻을 수 있으며, 모든 리덕트의 교집합이 코어이므로 속성 $\{a_2\}$가 코어가 된다. 이러한 코어는 분류를 위한 가장 중요한 속성이라는 것을 의미한다. 따라서 코어는 분류의 근사화 특성을 감소시키지 않기 위해서 속성 집합으로부터 제거될 수 없다. 그림 1은 계층적 분류구조를 가지고 속성에 따라 정보시스템을 분해한 것이다. 첫 단계에서는 코어의 속성을 가지고 나누었고, 차례대로 나머지 리덕트의 속성을 가지고 분해한 것이다.

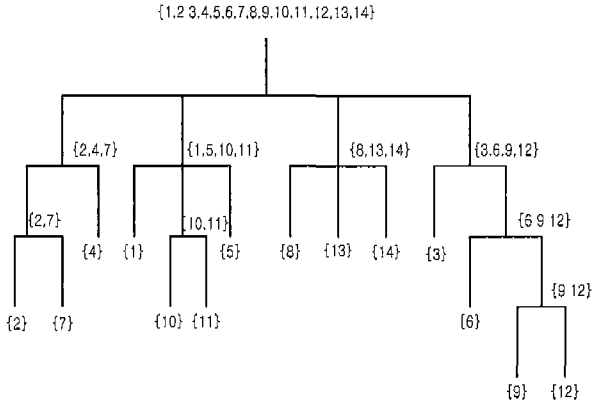

그림 1. 계층적 알갱이화
Fig. 1. Hierarchical Granulation

그림 1에서 다음과 같이 계층화된 알갱이 구조를 얻을 수 있다.

6 : {{1,2,3,4,5,6,7,8,9,10,11,12,13,14}}
5 : {{1,5,10,11},{2,4,7},{8,13,14},{3,6,9,12}}
4 : {{1,5,10,11},{2,7},{3},{4},{8,13,14},{6,9,12}}
3 : {{1},{2},{3},{4},{5},{6},{7},{8},{9,12},{10,11},{13},{14}}
2 : {{1},{2},{3},{4},{5},{6},{7},{8},{9},{10,11},{12},{13},{14}}
1 : {{1},{2},{3},{4},{5},{6},{7},{8},{9},{10,11},{12},{13},{14}}

최상위 분할은 동치관계 $E_{\varnothing}$와 일치한다. 다른 레벨의 알갱이화에서 판단속성 1={1,2,4,7,10,11,13}의 러프집합 근사화는 다음과 같다.

더 거친 알갱이를 가지는 상위레벨일수록 더 정확한 러프집합 근사화를 가진다. 계층화된 알갱이를 사용하여 근사화를 위한 적당한 알갱이를 찾을 수 있다. 본 연구에서는 상위레벨에 코어가 되는 속성을 적용하여 비슷한 분류의 알갱이를 찾은 후, 각각의 클러스터에 대한 상대리덕트를 찾아 분류규칙을 찾아낸다. 상대리덕트를 찾아보면,

속성 a_2=1의 클러스터 : $\{a_2\}$
속성 a_2=2의 클러스터 : $\{a_2, a_4\}, \{a_2, a_5\}$
속성 a_2=3의 클러스터 : $\{a_2, a_4\}, \{a_2, a_5\}$
속성 a_2=4의 클러스터 : $\{a_2, a_4\}$ 이다.

표 2. 러프집합에 의한 계층적근사화
Table 2. Hierarchical approximation with rough set

level	하한근사	상한근사	정확도
6	∅	U	0
5	{2,4,7}	{1,2,4,5,7,8,10,11,13,14}	3/10
4	{2,4,7}	{1,2,4,5,7,8,10,11,13,14}	3/10
3	{1,2,4,7,10,11,13}	{1,2,4,7,10,11,13}	1
2	{1,2,4,7,10,11,13}	{1,2,4,7,10,11,13}	1
1	{1,2,4,7,10,11,13}	{1,2,4,7,10,11,13}	1

표 3. 정보시스템의 분류규칙
Table 3. Decision table of information system

Rule	a_2	a_4	D
1	1	-	1
2	-	1	1
3	2	2	1
4	3	4	1
5	2	3	2
6	3	2	2
7	4	2	2
8	4	4	2
9	3	3	3
10	4	3	3

구한 상대 리덕트 중에서 최소의 분류규칙을 생성하기 위해서 리덕트 $\{a_2, a_4\}$의 속성만으로 구성된 의사결정표를 구한다. 이러한 의사결정표는 동일한 속성값을 가지는 객체들이 존재하고, 이러한 객체들은 서로 합쳐서 최소의 객체들을 가지도록 한다.

위의 분류규칙에서 규칙이 최소의 수가 되도록 공통이 되는 규칙들을 하나로 묶으면 다음의 규칙을 얻을 수 있다.

If a_2=1 or a_4=1 then D=1
If a_2=2 and a_4=2 then D=1
If a_2=3 and a_4=4 then D=1
If a_2=2 and a_4=3 then D=2
If a_2=3 and a_4=2 then D=2
If a_2=4 and a_4≠3 then D=2
If a_2≠2 and a_4=3 then D=3

위의 결과에서 볼 수 있듯이 고려된 정보시스템은 10개의 If-then 규칙으로 구성된 분류지식을 갖는다. 제안된 분류기법으로 러프집합의 근사화 성질을 이용하여 불필요한 속성을 제거하고 가장 핵심적인 정보를 지니고 있는 코어를 찾을 수 있었으며, 이를 계층적 분류구조에 적용함으로써 알갱이 성질을 갖는 데이터에서 지식을 추출할 수 있었다. 이러한 분류기법으로 지식의 감축을 이루어 최소의 분류규칙을 찾아낼 수 있었으며, 계층적 분류구조가 지식분류에 있어 효과적인 접근방식이라 할 수 있다.

3.4 분류규칙 발견 알고리즘

이상의 분류규칙 발견과정을 정리하여 흐름도를 나타내면 그림 2와 같다.

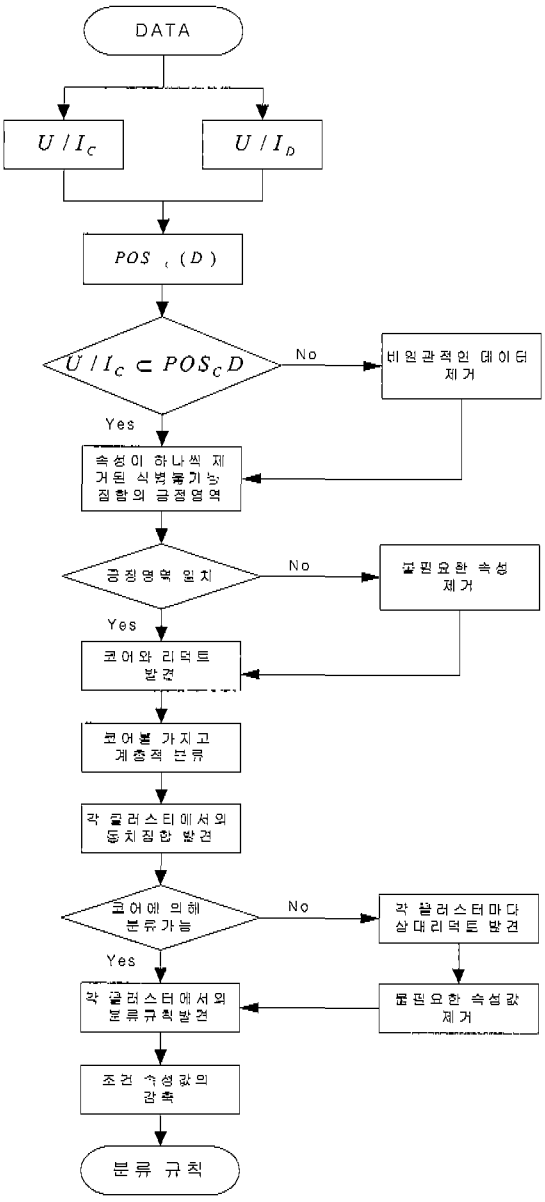

그림 2. 분류규칙 발견 알고리즘
Fig. 2. Flowchart of classification rule discovery

4. Simulation

시뮬레이션을 위한 데이터로 Wisconsin Breast Cancer Database[10]를 사용하였다. 이 데이터는 위스콘신 대학병원의 William H. Wolberg 박사가 1989년부터 1991년까지 수집한 데이터로 9개의 조건속성과 2개의 클래스(양성, 악성)로 이루어진 1개의 판단속성으로 이루어진 총 699개의 데이터로 구성되어있다. 속성분류규칙을 찾아내기 위한 학습데이터로 369개의 데이터를 선택했으며, 이중 불완전한 속성값을 갖는 14개의 데이터를 제외한 355개의 데이터가 학습데이터로 사용되었다. 그리고 시험데이터로 330개의 데이터가 사용되었다.

제안된 방법에 의해서 생성된 분류규칙의 수는 4개이며, 다음과 같다.

1. if $(a_2 \leq 2)$ *then* $d = class\ 1$
2. if $(a_2 \geq 3$ and $a_3 = 1)$ *then* $d = class\ 1$
3. if $(a_2 \geq 3$ and $a_3 \geq 2)$ *then* $d = class\ 4$
4. if $(a_1 \geq 9)$ *then* $d = class\ 4$

구해진 분류규칙에 대해서, 학습 데이터의 분류율은 93.5%이고, 시험 데이터에 대한 분류율은 95.45% 이다.

표 4. WBC 데이터에 대한 학습 데이터 결과
Table 4. Result of learning data for WBC data

class	원래 데이터	분류된 데이터	정확한 분류	잘못된 분류	정확도
class 1	189	184	175	9	93.5%
class 2	166	171	157	14	

표 5. WBC 데이터에 대한 시험 데이터 결과
Table 5. Result of test data for WBC data

class	원래 데이터	분류된 데이터	정확한 분류	잘못 된분류	정확도
class 1	257	242	242	0	95.5%
class 2	73	88	73	15	

결과에서 볼 수 있듯이 WBC 데이터에 대한 분류규칙은 4개로 이루어져 있으며 시험 데이터의 분류율이 95.5%로 양호한 결과를 보여주었다. 이러한 결과는 본 논문에서 최소의 분류규칙을 가지고 지식을 추출하려는 목적과 부합되는 것이다. 기존의 분류기법을 사용한 것과 비교하면 표 6과 같다.

표 6. WBC 데이터에 대한 분류기법들의 정확도
Table 6. Accuracy of classification methodologies for WBC data

분류기법	데이터	정확도 및 분류규칙수
뉴로-퍼지분류기법	학습데이터 시험데이터	은닉층의 수 = 10 100.0% 95.0%
C4.5	학습데이터 시험데이터	규칙수 = 16 96.8% 95.6%
CN2	학습데이터 시험데이터	규칙수 = 30 100.0% 94.4%
계층적 러프 분류기법	학습데이터 시험데이터	규칙수 = 4 93.5% 95.45%

계층적 러프분류기법을 뉴로-퍼지분류기법, C4.5 그리고 CN2와 비교해 보았다.[11] 뉴로-퍼지분류기법의 경우, 학습시는 100%의 정확도를 보여주었으나 시험시는 95%의 정확도를 가졌다. 양호한 결과를 나타내는 분류기법이지만 은닉층의 수를 결정하는 것이 문제이다. C4.5와 CN2 분류기법은 표 6에서 볼 수 있듯이 계층적 러프분류기법과 정확도는 비슷하지만 규칙의 수가 너무 많다는 단점이 있다.

위의 기법들에 반해 계층적 러프분류기법은 다른 기법들과 비슷한 정확도를 가지면서 분류규칙의 수가 매우 적다는 것을 알 수 있다. 이는 분류기법에서 나온 결과를 이해하기 매우 쉽다는 것을 의미하며, 분류의 속도가 더욱 빠르게 된다. 또한 하나의 규칙에 들어가는 조건절의 수 또한 적다는 장점이 있다. 그리고 계층적 분류구조를 사용함으로써 신경망 기법에 비해 설계가 쉬워 모델을 구축하는데 걸리는 시간이 빠르다는 이점을 가지게 된다.

다음으로 Lymphography domain data를 가지고 시뮬레이션을 해 보았다. Lymphography domain data[12]는 Institute of Oncology의 대학의료센터로부터 Zwitter와 Soklic에 의해서 제공된 자료이다. 객체의 수는 148개이며, 18개의 조건속성 $\{a_1, \cdots, a_{18}\}$과 4개의 클래스(normal find, metastases, malign lymph, fibrosis)로 이루어진 한 개의 판단속성 $\{d\}$으로 이루어져 있으며, 속성들에 대한 자세한 내용은 부록에 수록하였다. 분류규칙을 생성시키기 위한 학습 데이터로 100개의 데이터를, 시험 데이터로 48개의 데이터를 사용하였다. 제안된 방법에 의해 분류된 규칙의 수는 28개이며, 분류결과는 다음과 같다.

표 7. Lymphography domain 데이터에 대한 학습 데이터 결과
Table 7. Result of learning data for lymphography domain data

class	원래 데이터	분류된 데이터	정확한 분류	잘못된 분류	정확도
class 1	1	1	1	0	81%
class 2	53	51	43	8	
class 3	43	46	35	11	
class 4	3	2	2	0	

표 8. Lymphography domain 데이터에 대한 시험 데이터 결과
Table 8. Result of test data for lymphography domain data

class	원래 데이터	분류된 데이터	정확한 분류	잘못된 분류	정확도
class 1	1	1	1	0	81.25%
class 2	28	23	21	2	
class 3	18	23	16	7	
class 4	1	1	1	0	

이 데이터에 대한 기존의 기법을 사용한 결과를 보면,

Assistant의 경우 76%의 정확도를 보였으며, Simple Bayes 기법은 83%, CN2는 78-82%, Experts는 85% 그리고 AQ15의 경우는 80-82%의 정확도를 나타내었다.[13]

Lymphography domain 데이터의 경우, 분류규칙을 찾아내기 까다로운 데이터이다. 결과에서 볼 수 있듯이 분류규칙의 수가 28개이다. 이는 데이터가 어느 특정 속성에 크게 영향을 받지 않고 여러 속성들에 걸쳐서 그 특성이 정해질 경우, 러프집합을 이용하여 불필요한 속성을 찾아내어 그 속성을 제거함으로써 지식의 감축을 이루어내게 하는 것을 어렵게 만든다. 또한 이 데이터는 특정 패턴을 가지고 분류가 이루어지지 않기 때문에 찾아낸 속성들을 가지고 최소의 분류규칙을 만드는데 어려움이 있다. 이러한 점은 속성들이 갖는 속성값의 분류가 크지 않기 때문이다. 즉, 속성의 수는 많았지만 각각의 속성들이 갖는 속성값의 범위가 크지 않았다는 것이다. 그리고 속성 선정의 문제도 있을 수 있다. 결과를 예측하기 위해 선택된 속성들이지만 이러한 속성들이 제시된 문제에 대한 정확한 지식을 갖고 있지 않다는 얘기이다. 정확도를 높이기 위해 계층적 구조에서 한 단계 하위구조를 사용할 수 있지만 이 경우 분류규칙의 수가 늘어난다는 단점이 있다.

따라서 이에 대한 대안으로 반복적인 분류방식이 있을 수 있다. 즉, 본 논문에서 제안한 분류기법으로 분류규칙을 찾은 후 그 정확도가 낮을 경우, 잘못 분류된 데이터들을 다시 계층적 분류기법으로 분류규칙을 찾는 것이다. 이 경우 역시 분류규칙의 수가 늘어나고 규칙의 조건절도 많아질 것이지만 정확도는 향상될 것이다. 이 경우 주의할 것은 처음 분류 때 사용되었던 속성들은 반복 분류 때 제외되어야 한다는 것이다. 즉, 처음에 데이터를 분류할 때 사용되었던 속성들을 다시 사용하게 될 경우, 첫 번째 반복에서 나온 규칙과 두 번째 반복에서 나온 규칙들이 서로 모순을 일으킬 수 있기 때문이다.

표 9. Lymphography Domain 데이터에 대한 분류기법들의 정확도

Table 9. Accuracy of classification methodologies for Lymphography Domain data

분류기법	데이터	정확도 및 분류규칙수
AQR	학습데이터	규칙수 = 102 100.0%
	시험데이터	76%
CN2 (90% threshold)	학습데이터	규칙수 = 97 100.0%
	시험데이터	78%
(95% threshold)	학습데이터	규칙수 = 86 99%
	시험데이터	81%
(99% threshold)	학습데이터	규칙수 = 73 91%
	시험데이터	82%
계층적 러프 분류기법	학습데이터	규칙수 = 28 81%
	시험데이터	81.25%

5. 결 론

본 논문에서는 컴퓨터 사용의 보편화에 따라 데이터의 수집과 저장이 용이해지면서 데이터로부터 유용하고 이해 가능한 정보를 추출하기 위해 러프집합이론과 계층적 분류구조를 데이터마이닝에서의 분류기법으로 사용한 분류지식 추출 방법을 제안했다. 비일관적이고 불충분한 데이터를 일관적인 지식으로 바꾸기 위해서 러프집합이론의 근사화를 이용하여 비일관적인 데이터를 제거하였으며, 의미 있는 지식으로 바꾸기 위해 리덕트와 코어를 이용하여 불필요한 속성을 제거하였다. 이 과정에서 계층적 granular computing 기법을 사용하여 알갱이 속에 숨겨져 있는 지식들을 찾아내는데 있어 효과적인 접근을 하였으며, 최상위레벨에 코어를 적용하여 계층적 분류를 함으로써 대량의 데이터를 효율적으로 처리할 수 있도록 하였다.

따라서 제안된 계층적 러프분류기법은 기존의 인공지능 기법인 신경망이나 의사결정나무에 비해 간단한 분류규칙을 유도해 낼 수 있으며, 분류과정을 쉽게 이해할 수 있으므로 모델에 대한 해석이 용이하다. 이러한 기법은 데이터마이닝에서 뿐만 아니라 제어를 위한 규칙 테이블의 규칙 간소화 등에 적용될 수 있다.

향후 연구과제는 계층적 분류구조에서 다른 레벨간의 분류율 연구와 반복적 분류구조에 대한 연구이다. 하위레벨로 갈수록 분류가 더 세밀하게 이루어지기 때문에 최소의 규칙을 얻는 것보다 정확한 분류를 원할 경우 이러한 성질을 이용하여 최적의 분류규칙을 찾아낼 수 있을 것이다. 또한 반복적 분류구조의 경우, 분류가 잘못된 데이터들에 대해 기존의 속성과는 다른 공통된 속성을 찾아내어 분류규칙을 찾아낼 수 있는 연구가 필요하다.

참 고 문 헌

[1] I. Witten, E.Frank, Data Mining, Morgan Kaufmann Publisher, 2000.

[2] A. Berson, S. Smith, K. Thearling, Building Data Mining Applications for CRM, McGraw-Hill, 1999.

[3] C. Olaru, L. Wehenkel, "Data Mining", *IEEE Computer Application in Power*, Vol. 12, No. 3, pp. 19-25, 1999.

[4] Zdzislaw Pawlak, "Why Rough Sets?", Proc. of the 5th IEEE International Conf. on Fuzzy Systems, Vol. 2, pp. 738-743, 1996.

[5] B. Walczak, D.L. Massart, "Rough sets theory", Chmometrics and Intelligent laboratory Systems, Vol. 47, No. 1, pp. 1-16, 1999.

[6] Y. Yang, T.C. Chiam, "Rule Discovery On Rough Set Theory", Proc. of the 3rd International Conference on Information Fusion, Vol.1, TuC4-11-16, 2000.

[7] Zdzislaw Pawlak, "Granularity of Knowledge, Indiscernibility and Rough Sets", Proc. of the *IEEE International Conf. on Fuzzy Systems : FUZZ-IEEE'98*, Vol. 1, pp. 106-110, 1998.

[8] 변증남, 방원철, 러프집합의 이론과 응용, 정문각, 1999.
[9] Y.Y. Yao, "Rough Sets, Neighborhood Systems, and Granular Computing", Proc. of the IEEE Canadian Conf. on Electrical and Computer Engineering, pp. 1553-1558, 1999.
[10] W.H. Wolberg, O.L. Mangasarian, "Multisurface method of pattern separation for medical diagnosis applied to breast cytology", Proc. of the National Academy of Sciences, USA., Vol. 87, pp. 9193-9196, 1990.
[11] M. Gorzalczany, Z. Piasta, "Neuro-fuzzy approach versus rough-set inspired methodology for intelligent decision support", Information Sciences, Vol. 120, No. 1, pp. 45-68, 1999.
[12] M. Zwitter, M. Slklic, "Lymphography Domain", the University Medical Centre, Institute of Oncology, 1988.
[13] P. Clark, T. Niblett, "The CN2 Induction Algorithm", Machine Learning Journal, Vol. 3. pp.261-283, 1989.

저 자 소 개

이철희(Chul-Heui Lee)
1983년 : 서울대학교 전기공학과(공학사)
1985년 : 서울대학교 전기공학과(공학석사)
1989년 : 서울대학교 전기공학과(공학박사)
1994년~1995년 : 미국 IONA대학 방문교수
1990년~현재 : 강원대학교 전기전자정보통신공학부 교수

관심분야 : Soft computing & Computational intelligence (신경망, 퍼지 시스템, 유전자 알고리즘), 지능제어 및 신호처리

서선학(Seon-Hak Seo)
1995년 : 강원대학교 전기공학과(공학사)
1997년 : 강원대학교 전기공학과(공학석사)
1997년~현재 : 강원대학교 전기공학과 박사과정

관심분야 : Soft computing(퍼지제어, 신경망, 유전자 알고리즘)