2002년도 대한전자공학회/대한전기학회 시스템 및 제어분야 합동 추계학술대회 논문집(2002.11.30)

러프집합과 계층적 구조를 이용한 규칙생성

김주영, 이철희
강원대학교 전기공학과

Rule Generation using Rough set and Hierarchical Structure

Ju-Young Kim, Chul-Heui Lee
Dept. of Electrical Engineering, Kangwon National University
E-mail : kjy0330@hanmir.com

Abstract

This paper deals with the rule generation from data for control system and data mining using rough set. If the cores and reducts are searched for without consideration of the frequency of data belonging to the same equivalent class, the unnecessary attributes may not be discarded, and the resultant rules don't represent well the characteristics of the data. To improve this, we handle the inconsistent data with a probability measure defined by support, As a result, the effect of uncertainty in knowledge reduction can be reduced to some extent. Also we construct the rule base in a hierarchical structure by applying core as the classification criteria at each level. If more than one core exist, the coverage degree is used to select an appropriate one among then to increase the classification rate. The proposed method gives more proper and effective rule base in compatibility and size. For some data mining example the simulations are performed to show the effectiveness of the proposed method.

1. 서론

현실세계에서 발생하는 많은 데이터들로부터 우리가 행하고자 하는 일에 도움이 되는 정보(information) 또는 지식(knowledge)을 찾아내는 것은 매우 중요하다[1]. 시스템 모델링 및 제어, 시계열 분석 및 예측, 의사결정 그리고 데이터 마이닝 등 폭넓고 다양한 분야에서 이러한 작업이 요구되고 있으며, 대부분의 경우 우리가 필요로 하는 정보 또는 지식은 규칙의 형태로 표현된다. 그런데 우리에게 제공되는 데이터에는 데이터의 중복, 오손, 비일관적인(inconsistent) 데이터의 존재, 불필요한 데이터의 포함 등으로 불확실성(uncertainty)이 존재하며, 적확한 지식의 도출이 쉽지 않다. 이처럼 불확실성을 갖는 데이터로부터 유용한 지식, 즉 규칙을 찾아내기 위해서는 데이터의 일반화(generalization) 능력을 갖춘 체계적이고 효율적인 규칙생성 방법의 적용이 필요한데, 최근 러프집합이론(rough set theory)[2][3][4]]을 이용한 연구들이 활발히 이루어지고 있다.

러프집합에 의한 규칙생성은 생성되는 규칙의 수가 많거나 모형을 구축하는데 사용되는 표본의 크기에 지나치게 민감한 의사결정나무기법이나, 결과만 제공할 뿐 어떻게 그러한 결과가 나왔는가에 대한 이유를 설명하지 못하는 신경망 기법과는 달리 명쾌한 규칙 생성과정과 적절한 양의 규칙으로 대상의 특성을 잘 묘사할 수 있게 해준다.

생성된 지식 또는 규칙이 대상을 보다 잘 묘사할 수 있으려면, 데이터에 포함된 불확실성의 영향을 최소한으로 줄이고 규칙간의 연계성이 잘 드러난 구조를 확보해야 한다. 따라서 본 논문에서는 러프집합이론의 비일관적 데이터에 대한 처리에 있어 확률적 개념[5]을 도입하여 데이터에서의 오차를 줄였으며, 규칙간의 연계성이 잘 드러나게 하기 위하여 규칙생성에 있어 계층적(hierarchical) 구조[6]를 사용하였다. 특히, 계층적 구조를 생성함에 있어 코어가 계층 분류의 기준이 되게 하였고, 그러한 코어가 여러개 존재할 경우 보다 효율적인 처리를 위하여 적용도를 기준으로 분류 기준이 되는 코어를 선택하는 방법을 사용하였다.

데이터로부터 규칙을 생성하는 절차는 크게 지식의 감축과 확률적 처리, 그리고 계층적 구조의 규칙생성으로 구성되어 있으며, 지식의 감축은 러프집합이론에서 사용하는 방법인 의사결정표(decision table)나 식별행렬을 이용하여 코어를 찾아내어 간략화 하는 것이다. 감축과정에서 일정 확률이상의 빈도를 나타내는 비일관적인 데이터를 일관적인 데이터로 취급하여 불확실성의 영향을 줄였다. 다음으로 계층적 구조를 갖는 규칙을 생성함에 있어 각 계층에 대한 분류 기준으로 코어만을 적용하여 계층적 구조를 생성하며 코어의 선택은 적용도가 높은 것을 최상위 단계에 적용함으로써 초기 분류 정확도를 향상시켜서 계산의 양을 줄여 명확하고 효율적인 구조를 갖도록 하였다.

2. 러프집합이론 [2]

Z. Pawlak에 의해 소개된 러프집합이론은 어떤 개념에 대해서 확실하게 그 개념에 속하는 하한 근사 공간과 속할 가능성을 가지는 상한 근사 공간을 집합을 통해서 나타낸다. 그 객체(object)들에 대한 식별불가능 관계가 러프집합이론의 수학적 기초이다. 러프집합에 대한 자세한 내용은 참고문헌을 참조하고 여기에서는 본 논문의 논리전개에 필요한 개념들을 간략히 설명한다[2][3][4].

2.1 러프집합의 주요 개념

2.1.1 식별불가능성(Indiscernibility)

모든 $a \in A$에 대하여,

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

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.

2.1.2 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.

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

집합 $BN_B(X) = B^*(X) - B_*(X)$은 X의 B-경계영역(boundary region)이라 부른다.

2.1.3 리덕트(reduct)와 코어(core)

$Q \subseteq P$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.

$$CORE(P) = \bigcap RED(P) \tag{3}$$

2.1.4 지식의 종속성

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.

$$k = \gamma(C, D) = \frac{card(POS_C(D))}{card\ U} \tag{4}$$

여기서, $POS_C(D) = \bigcup_{X \in U/D} C_*(X)$을 C에 대한 U/D의 긍정영역(positive region)이라 한다.

2.2 러프집합에서의 척도(measure) [2],[7]

러프집합이론에서는 지식의 불확실성을 경계영역으로 나타내고 있다. 집합의 경계영역이 커질수록 그 지식의 정확성(exactness)이 떨어진다.

정확성 척도(accuracy measure)는 다음과 같이 정의될 수 있다.

$$\alpha_B(X) = \frac{card\, B_*(X)}{card\, B^*(X)} \tag{5}$$

정확성 척도 $\alpha_B(X)$는 집합 X에 대한 데이터의 완전한 정도를 측정하도록 정의되었다.

집합의 근사화의 정확성을 나타내는 정의의 확장으로서 B에 의한 F의 근사화의 정확도(accuracy of approximation of F by B)는 다음과 같이 정의된다.

$$\alpha_B(F) = \frac{\sum card\, B_* X_i}{\sum card\, B^* X_i} \tag{6}$$

B에 의한 F의 근사화의 질 (quality of approximation of F by B)은 다음과 같이 정의된다.

$$\gamma_B(F) = \frac{\sum card\, B_* X_i}{card\, U} \tag{7}$$

분류의 근사화의 정확도는 지식 B을 이용하여 객체들의 분류를 할 때 가능한 옳은 판단의 비율을 나타낸다. 분류의 근사화 질은 지식 B을 이용하여 F의 동치류로 옳게 분류되는 개체들의 비율을 나타낸다.

러프집합에서 데이터 해석에 필요한 척도로 정확성과 함께 적용도라는 척도를 들을 수 있다. 정확성이라는 척도 $\alpha_R(D)$는 $R \rightarrow D$의 분류의 충분함 정도를 나타내는 척도라 할 수 있는 반면 적용도 $k_R(D)$은 $R \rightarrow D$의 필요성 정도(옳은 긍정 비율)를 나타내는 척도이다. 정의는 다음과 같다[8].

$$\alpha_R(D) = \frac{card\ (R_A \cap D)}{card\ (R_A)} \tag{8}$$

$$k_R(D) = \frac{card\ (R_A \cap D)}{card\ (D)} \tag{9}$$

3. 코어속성을 이용한 계층적 구조 규칙 생성

러프집합이론에서는 방대한 양의 데이터로부터 간결하고 유용한 지식을 얻기 위하여 의사결정표나 식별행렬을 이용하여서 코어와 리덕트를 찾아내어 지식을 감축한다. 본 논문에서는 지식 감축 과정에서 구하여진 코어를 계층적 지식 분류의 기준으로 적용하여 계층적 구조의 규칙을 생성하는 과정을 수행할 것이다.

3.1 지식 감축에서의 비일관적 데이터의 확률적 처리

의사결정표는 어떤 조건들이 만족되었을 때 어떤 의사결정을 취해야 하는가를 알려주는 일종의 규칙(prescription)이다. 코어와 리덕트 개념을 이용하여 의사결정표로부터 불필요한 속성과 속성값을 제거함으로써 간략화된 의사결정표를 얻을 수 있다.

의사결정표의 간략화 기법은 다음 3단계로 설명될 수 있다.

(1) 불필요한 속성들을 의사결정표에서 제거하여 차원을 줄인다.

(2) 같은 속성을 가지는 객체들을 통합한다.

(3) 불필요한 속성값을 제거한다.

의사결정표의 간략화 과정에서 비일관적인 데이터에 대한 일반적인 처리 방법은 이들을 모두 제거한 뒤 코어와 리덕트를 구하는 것이다. 그러나 이 경우 비일관적 데이터에 존재하는 불확실성의 영향을 피하기 위해 이들을 모두 배제함으로써 데이터에 포함된 유용한 정보의 유실도 가져오게 된다. 이를 보완하기 위하여 확률적으로 일정한 빈도이상 나타나는 비일관적인 데이터에 대해 일관적인 데이터로 간주하여 데이터에 대한 오차와 융통성을 부여함으로서 데이터에 포함된 불확실성의 영향을 최대한 줄이면서 효과적으로 규칙을 도출할 수 있다.

비일관적인 데이터의 빈도에 의한 확률은 아래와 같이 정의 할 수 있으며, 이 값이 $\beta(0.5 \le \beta \le 1)$보다 크면 일관적인 데이터로 간주한다.

$$P(r_i)=\frac{supp(r_i)}{\sum supp(r_j)} \tag{10}$$

여기서 $supp$는 규칙을 지지하는 객체의 수이고, r_j는 r_i와 조건부 속성이 같은 규칙들이다.

다음의 의사결정표의 예를 생각해 보자.

[illegible]	a_1	a_2	a_3	a_4	a_5	a_6	D
1	1	1	4	3	2	3	1
2	1	1	4	2	2	2	1
3	2	1	4	3	2	3	1
4	2	②	2	2	2	2	1
5	3	3	2	③	④	3	1
6	2	2	3	3	3	3	2
7	2	④	2	2	2	2	2
8, 9	3	3	2	②	4	2	2
10	4	4	1	4	3	4	2
11	4	4	1	④	4	4	2
12	3	3	2	3	③	3	3
13	3	3	2	2	4	2	3
14	4	4	1	③	4	4	3

표 5. 비일관적 데이터에 대한 처리의 예

표 1은 객체 8, 9, 13이 비일관적인 데이터이다. 비일관적 데이터에 대한 β를 0.6으로 정하고 확률적으로 처리하면 객체 8, 9는 2/(2+1)=2/3>0.6이므로 일관적인 규칙으로 처리하고 객체 13은 1/(1+2)=1/3<0.6이므로 비일관적인 데이터로 제거한다. 그리고 이에 대해 코어를 구하여 표 1에 표시하였다.

기존의 비일관적 데이터 처리후의 코어는 객체 4의 a_2=②, 객체 5의 a_5=④, 객체 7의 a_2=④, 객체 11의 a_4=④, 객체 12의 a_5=③ 객체 14의 a_4=③이고, 비일관적 데이터에 빈도의 의한 확률적 처리후의 코어는 기존의 비일관적 데이터 처리후 구해진 코어에 객체 5의 a_4=③과 객체 8의 a_4=②가 추가된다.

위에서 보여진 결과는 기존의 비일관적 데이터의 처리는 확률적인 데이터 처리에서 코어로 선택한 속성(객체 8의 코어)을 제거하는 결과를 초래한다. 그러므로 확률적 처리를 하면 데이터에 포함된 불확실성을 최소화함을 볼 수 있다.

3.3 계층적 구조의 규칙 생성

3.3.1 러프집합의 코어를 이용한 계층적 구조 규칙 생성

계층적 구조에서 전체집합에 대한 계층은 하나의 노드(node)가 하나의 클러스터(cluster)를 나타내는 나무 구조로써 설명될 수 있다. 계층적 구조에서 하위 레벨 클러스터의 모든 원소들은 상위 클러스터와 뿌리사이에 있는 모든 노드들에 포함된다. 최하위 레벨에서의 규칙 수는 객체 수만큼 생겨날 것이며, 규칙마다 속성의 개수만큼의 조건이 포함되어야 할 것이다. 따라서 전체 집합을 의사결정표로 나타낸 후 속성에서 코어를 찾아내어 그 코어를 가지고 전체집합을 계층적으로 분류한다. 많은 레벨로 계층을 나눌 경우 규칙의 수가 커지기 때문에 코어를 이용해 계층적 구조의 규칙을 찾아내게 된다. 코어를 이용하는 이유는 코어가 객체들을 나타내는 가장 핵심적인 속성이기 때문이다.

다음의 의사결정표의 예를 생각해 보자. [2]

U	a	b	c	d	e	f	D
1	3	3	2	②	2	4	I
2	③	2	2	②	2	4	
3	③	2	2	①	2	4	
4	②	2	2	1	1	4	II
5	②	2	2	2	1	4	
6	3	2	2	③	2	3	III
7	3	3	2	③	2	3	
8	4	3	②	3	2	3	
9	4	3	③	3	2	2	IV
10	4	4	3	3	2	2	
11	4	4	3	2	2	2	
12	4	3	3	2	2	2	
13	4	2	3	2	2	2	

표 6. rotary clinker kiln의 의사결정표

표 2는 [2]에 있는 rotary clinker kiln의 예이다. 러프집합으로 지식을 감축하는 과정은 [2]에 있는 방법에서 불필요한 리덕트 속성 b를 제거하지 않고 코어를 구한다. (표 2)

그림 1의 계층적 구조는 최상위 단계에 코어 속성 c를 적용하고 그 다음은 분류된 데이터의 코어와 리덕트를 적용하는 순으로 계층적 구조를 형성한다. 표 2는 계층적 구조의 의사 결정부 속성 D= I 의 정확도를 나타내고 있는 것이다. 정확도를 보면 알 수 있듯이 러프집합의 계층적 구조를 이용하여 규칙을 추출할 수 있음을 알 수 있다.

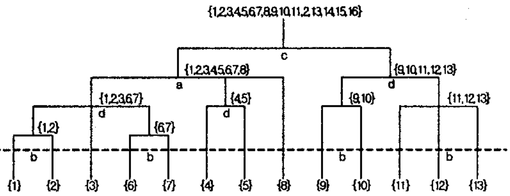

그림 1. 계층적 구조

D= I	하한근사	상한근사	정확도
1	Φ	U	0
2	Φ	{1,2,3,4,5,6,7,8}	0
3	Φ	{1,2,3,6,7}	0
4	{1,2,3}	{1,2,3}	1
5	{1,2,3}	{1,2,3}	1

표 7. 러프집합에 의한 계층적 근사화

그림 1을 보면 점선을 기준으로 위 부분은 코어 속성에 의해, 점선 밑 부분은 리덕트를 사용에 의해 규칙을 생성하는 것이고 표 3를 보면 코어만 이용한 경우의 정확도가 1이고 리덕트까지 이용한 경우도 같은 결과이다. 이것은 코어 속성만을 가지고도 계층적 분류가 가능하다는 것이다. 그러나 모든 상황에서 코어만을 가지고 분류가 가능한 것은 아니다. 코어 속성의 실제 존재하는 레벨의 곱이 의사결정부 레벨보다 작은 경우에는 불가능하다. 이것이 코어속성만을 가지고 의사결정을 하는 최소조건이고 이 조건을 만족할 경우 코어속성만을 가지고 규칙을 생성할

수 있다고 볼 수 있다.

3.3.2 코어 선택의 문제

코어 속성이 여러 개 존재할 경우 어느 속성을 선택하면 연계성이 높고 효율적인 구조를 갖는가에 대한 선택의 문제가 존재하게 된다.

식 (9)에서 나타낸 적용도는 속성이 의사결정부에 얼마만큼 필요한가를 나타내는 척도이므로 의사결정표에 간략화에서 구해진 각 코어에 대해 의사결정부 동치류에 대한 적용도를 계산하여 일정값 (δ)이상 되는 것을 선택한다.

$$k_R'(D) = k_R(D) \mid_{k_R(D_i) > \delta} \tag{11}$$

그런 다음 각 의사결정부에 대한 적용도를 합하여 전체 집합에 대한 적용도로 나타낸다. 이 때 전체 집합에 대한 의사결정부의 가중치를 곱한다.

$$k_R''(D_i) = \sum_{i=1}^{n} \frac{|D_i|}{|U|} k_R' \tag{12}$$

즉, 코어속성의 동치류들이 의사 결정부에 큰 영향을 주는 것의 합으로 표현한 것이다.

식 (12)에서 구한 전체에 대한 적용도가 높은 코어를 최상위에 적용하여 초기 분류 정확도를 향상시켜서 더 효율적이고 연계성이 높은 구조를 갖게 한다.

표 2에 대하여 적용도를 적용해 보면 a 속성의 적용도는 12/13이고, c 속성의 적용도는 13/13=1이고, d속성의 적용도는 8/13이다. 그림 2, 3은 최상위에 적용되는 코어 속성 c와 d를 적용하였을 때의 계층적인 구조이다. c 속성을 최상위로 분류한 경우의 D=Ⅳ의 규칙은 "If c=3 Then D=Ⅳ"이고 d 속성을 최상위로 분류한 경우의 D=Ⅳ의 규칙은 "If (d=3 & c=3) or (d=2 & c=3) Then D=Ⅳ"이 된다. c 속성을 최상위로 한 의사결정부 Ⅳ에서 보면 최소 알고리즘이 쉽게 얻어지는 반면 d 속성을 최상위로 한 의사결정부 Ⅳ를 보면 한번의 계산이 더 필요함을 알 수 있다. 그러므로 적용도가 높은 코어를 최상위에 적용함으로써 초기 분류율을 높일 수 있으며 효율적인 규칙을 생성할 수 있다.

이상의 규칙 생성과정을 정리하여 흐름도로 나타내면 그림 4와 같다.

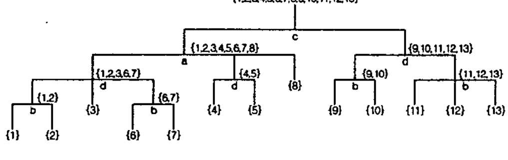

그림 2. 코어속성 c를 최상위로 했을 때 계층적 구조

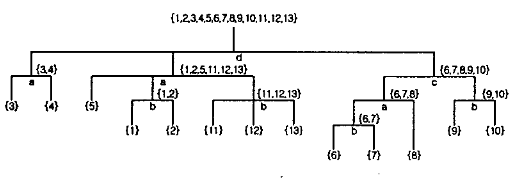

그림 3. 코어속성 d를 최상위로 했을 때 계층적 구조

4. 결론

본 논문에서는 러프집합을 이용하여 계층적인 구조를 갖는 규칙의 생성방법을 제안하였다. 제안된 방법은 비일관적인 데이터 처리에 확률적 개념을 도입하여 데이터의 포함된 불확실성의 영향을 최소한으로 줄이고 계층구조의 각 레벨에 적용도가 높은 코어를 분류 기준으로 적용함으로써 규칙간의 연계성이 높고 효율적인 규칙을 생성하도록 하였다.

따라서 제안된 규칙 생성 방법은 다른 방법에 비하여 규칙의 수가 적을 뿐만 아니라 조건부가 짧고 규칙의 형태가 간단하여 이해가 쉬우면서도 분류율이 좋다. 이러한 러프집합과 계층적 구조를 이용한 규칙 생성 방법은 시스템 모델링 및 제어, 시계열 분석 및 예측, 의사결정 그리고 데이터 마이닝 등 폭넓고 다양한 분야에서 적용이 가능할 것이다.

향후 연구과제로는 비일관적 데이터를 일관적 데이터로 취급하는 확률적 처리에 대하여 그에 합당한 패널티를 부여한 규칙을 생성하는 연구와 반복횟수, 중요도를 고려한 가중치를 주어 규칙을 생성하는 연구가 필요하다.

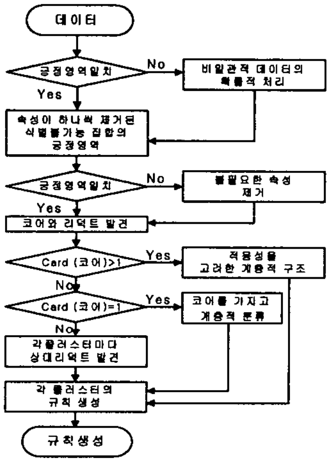

그림 4. 흐름도

[참 고 문 헌]

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

[2] 변증남, 방원철, 러프집합의 이론과 응용, 청문각, 1999

[3] Zdzislaw Pawlak, "Rough set approach to knowledge-based decision support", European Journal of Operational Research, Vol. 99, pp.48-57, 1997.

[4] Y.Y. Yao, "Rough Sets, Neighborhood Systems and Granular Computing", Proc. of the IEEE Canadian Conf. on ELetrical and Computer Engineering, pp.1553-1558, 1999.

[5] 권은아, 김홍기, "확률적 러프 집합에 기반한 근사 규칙의 간결화", 정보처리학회논문지(D) 제 8-D권 제 3호 pp.203-210, 2001

[6] 서선학, 이철희, "러프집합과 계층적 분류구조를 이용한 데이터마이닝에서의 분류지식 발견" 한국 퍼지 및 지능시스템학회 2002 vol.12, No.3, pp.202-209, 2002

[7] Z. B. Xu, J. Y. Liang, C. Y. Dang and K. S. Chin, "Inclusion degree: a perspective on measures for rough set data analysis", Information Sciences, Vol. 141, pp.193-314, 2002.

[8] Shusaku Tsumoto, "Extraction of Experts' Decision Rules from Clinical Databases Using Rough Set Model", Intelligent Data Analysis, Vol. 2, pp.215-227, 1998