

러프집합과 Granular Computing을 이용한 분류지식 발견

최상철, 이철희
강원대학교 전기공학과

Discovering classification knowledge using Rough Set and Granular Computing

Sang-Chul Choi, Chul-Heui Lee
Department of Electrical Engineering, Kangwon National University

Abstract - There are various ways in classification methodologies of data mining such as neural networks, but the result should be explicit and understandable and the classification rules be short and clear. Rough set theory is an effective technique in extracting knowledge from incomplete and inconsistent information and makes an offer classification and approximation by various attributes with effect. This paper discusses granularity of knowledge for reasoning of uncertain concepts by using generalized rough set approximations based on hierarchical granulation structure and uses hierarchical classification methodology that is more effective technique for classification by applying core to upper level. The consistency rules with minimal attributes is discovered and applied to classifying real data.

1. 서 론

오늘날 발생하는 데이터의 양은 대단히 커지고 있다. 이런 방대한 양의 데이터로부터 유용한 정보를 추출하는 data mining은 현재 중요한 문제중의 하나가 되었다. data mining의 분류기법에는 신경망 등 여러 가지 방법들이 있으나, 그 결과가 명쾌하고 쉽게 이해할 수 있어야 하며 분류규칙이 간단해야 한다. 이러한 점에서 러프집합이론은 불충분하거나 일관성 없는 정보로부터 지식을 추출하는데 있어 효과적인 기법중의 하나이며, 여러 속성을 이용한 분류화 및 근사화를 효과적으로 제공한다. 러프집합이론은 불확실성, 부정확성 그리고 애매모호함을 다루기 위해 1980년대 이래로 Z. Pawlak에 의해 발전된 새로운 수학적 기법이다. 고전집합이론은 crisp sets를 다룬다면, 러프집합이론은 고전집합이론의 확장으로 고려될 수 있다. 러프집합이론에서 객체들은 몇몇 특징들에 대해서 식별할 수 없는 객체들의 클러스터들인 유사 클래스로 분류된다. 식별할 수 없는 분류들은 기본적인 지식의 빌딩블록(building blocks)이며, 데이터에서 숨겨진 패턴들을 찾는데 사용된다.^[1]

러프집합이론은 많은 이점을 가지고 있는데, 가장 중요한 이점은 퍼지집합이론에서의 소속정도나 value of possibility와 같은 데이터에 대한 사전 혹은 추가적인 정보를 필요로 하지 않는다는 것이다. 일어진 결과의 직접적인 해석을 제공한다는 것인데 이러한 것은 지식획득에 있어 매우 중요한 것이다.

지식발견과정에서 중요한 단계는 데이터 크기의 축소이다. 실제 데이터베이스 시스템에서 상황에 따라 많은 속성들과 기록들이 있지만, 실제로는 단지 몇 개의 속성들만이 필요하다. 만일 필요 없는 속성들이 제거될 수 있다면, 데이터를 분석하는데 있어 복잡도는 크게 감소될 수 있을 것이다.^[2]

본 연구는 일관성 없는 데이터에 러프집합이론을 적용

한 속성의 감축과 분류규칙의 발견과정으로 구성되어 있다. 속성감축의 과정은 판단결정과정에 있어 그다지 중요하지 않은 조건속성들을 제거하고 식별행렬을 이용하여 코어를 찾아내 계층적 구조의 상위단계에 적용하고 속성들에 대한 리덕트(reduct)를 구한다. 이러한 방법은 데이터의 크기를 줄일 수 있으므로 규칙의 수를 적게 할 수 있으며 규칙을 쉽게 찾아낼 수 있게 한다. 그리고 규칙발견 과정은 각각의 객체들의 속성값들간의 관계를 분석하고 불필요한 속성값들을 제거하여 최소의 규칙들을 유도한다.

2. 러프집합이론

같은 정보로써 특정화된 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 concepts)이라 하며, 전체집합에 대한 지식의 기본적인 알갱이(granule, atom)를 형성한다.^[3]

2.1 Indiscernibility and Granularity

유한집합이며 공집합이 아닌 전체집합 U 와 속성집합 A 가 주어졌다고 가정하자. 모든 속성 $a \in A$ 에 대해, 집합 V_a 는 속성 a 의 정의역이다. 정보시스템(information system) S 는 $S = (U, A)$ 로 정의된다. A 의 임의의 부분집합 B 는 U 상에서 하나의 이진관계(binary relation) I_B 를 결정하는데, 이를 식별 불가능관계(indiscernibility relation)라 하며, 다음과 같이 정의된다. 모든 $a \in A$ 에 대하여,

$$x I_B y \text{ if and only if } a(x) = a(y) \quad (1)$$

여기서 $a(x)$ 는 x 에 대한 속성값 a 를 나타낸다. I_B 는 동치관계(equivalence relation)이며, I_B 의 모든 동치류들의 집합은 U/I_B 혹은 간단히 U/B 로 나타내고, x 를 포함하는 I_B 의 동치류는 $B(x)$ 로 표현한다.

만일 (x, y) 가 I_B 에 속하면 x 와 y 는 B -식별불가능(B-indiscernible)이다. 관계 I_B 의 동치류(혹은 분할 U/B 의 블록)들은 B -기본개념(elementary concepts) 혹은 B -알갱이(granules)라고 부른다.

2.2 Approximations and Granularity

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.

$$B_*(X) = \bigcup_{x \in X} \{B(x) : B(x) \subseteq X\} \quad (2)$$

$$B^*(X) = \bigcup_{x \in X} \{B(x) : B(x) \cap X \neq \emptyset\}$$

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

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

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

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

러프소속함수는 다음과 같이 근사화와 경계영역을 정의하도록 사용될 수 있다.

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

러프소속함수는 다음의 식처럼 일반화될 수 있다.

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

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

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

2.3 Core and reduct

$Q \sqsubseteq P$ 가 독립이고 $U/I_Q = U/I_P$ 이면 Q 는 P 의 리덕트(reduct)라 하고 P 는 여러 개의 리덕트를 가질 수 있다. P 내의 모든 필요 불가결한 관계들의 집합을 P 의 코어(core)라 하고 다음과 같이 표현할 수 있다.

$$CORE(P) = \cap RED(P) \quad (6)$$

2.4 지식의 종속도(Dependency of Knowledge)

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

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

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

3. Hierarchical Granulation and Rule Extraction

3.1 규칙의 축소

객체들의 예는 판단 테이블의 형식으로 나타내어진다. 테이블의 행은 객체, 열은 조건속성과 판단속성을 나타낸다. 다음의 판단 테이블을 예로 들어본다.

표 1. 판단 테이블

U	a	b	c	d	e
1	0	1	0	0	0
2	1	1	0	1	0
3	1	0	0	0	1
4	1	0	0	1	1
5	1	1	0	2	2
6	2	2	1	2	2
7	2	3	2	2	2
8	2	0	1	3	1
9	2	0	1	3	2

표 1은 9개의 규칙, 4개의 조건속성 $\{a, b, c, d\}$ 그리고 1개의 판단속성 $\{e\}$ 를 가지고 있다. 코어와 리덕트를 이용하여 판단 테이블로부터 불필요한 속성과 속성값을 제거함으로써 간략화 된 규칙을 얻을 수 있다. 먼저 판단 테이블에서 긍정영역을 찾아보면 $POS_C(D) = \{1, 2, 3, 4, 5, 6, 7\}$ 으로 규칙 8, 9번은 비일관적인 데이터이므로 제거하여 일관성 있는 판단 테이블을 구성한다.

다음으로 불필요한 조건부 속성을 제거한다. 각각의 조건부 속성을 하나씩 제거해 나가면서 계산해보면 속성 c 만이 불필요한 속성임으로 c 열을 제거하여 표 2를 얻을 수 있다.

표 2. 불필요한 속성을 제거한 판단 테이블

U	a	b	d	e
1	0	1	0	0
2	1	1	1	0
3	1	0	0	1
4	1	0	1	1
5	1	1	2	2
6	2	2	2	2
7	2	3	2	2

3.2 Hierarchical Granulation

multi-layered granulation의 예가 판단 테이블에서 속성값을 기준으로 한 전체집합의 분할이다. 동치관계 E_1 을 속성집합으로 정의한다. 다음 동치관계들은 남아있는 속성들의 집합으로부터 속성들을 연속적으로 제거함으로써 정의될 수 있다. 반대로, E_m 은 속성들의 집합에 의한 동치관계이며, 다음 관계들은 연속적으로 속성들을 추가함으로써 얻을 수 있다.

즉, $E_1 \sqsubseteq E_2 \sqsubseteq \dots \sqsubseteq E_m$ 으로 표현될 수 있으며, equivalence granules의 일치하는 열은 다음 조건을 만족한다. $[x]_{E_1} \sqsubseteq [x]_{E_2} \sqsubseteq \dots \sqsubseteq [x]_{E_m}$ [4]

표 2에 대하여 $\{a, b, d\}$, $\{a, b\}$, $\{a\}$, \emptyset 의 속성집합들을 적용해보면, $I = E_{a, b, c} \subseteq E_{a, b} \subseteq E_a \subseteq E_\emptyset = U \times U$ 의 관계가 성립하고, 다음의 계층화된 granulation structure를 얻을 수 있다.

$$\begin{aligned} 4 &: \{1, 2, 3, 4, 5, 6, 7\}, \\ 3 &: \{(1), \{2, 3, 4, 5\}, \{6, 7\}\} \\ 2 &: \{(1), \{2, 5\}, \{3, 4\}, \{6, \{7\}\}\} \\ 1 &: \{(1), \{2\}, \{3\}, \{4\}, \{5\}, \{6\}, \{7\}\} \end{aligned}$$

최상위 분할은 동치관계 E_\emptyset 과 일치한다. 다른 레벨의 granulation에서 $2 = \{5, 6, 7\}$ 의 러프집합 근사화는 다음과 같다.

level	하한근사	상한근사	정확도
4	\emptyset	U	0
3	$\{6, 7\}$	$\{2, 3, 4, 5, 6, 7\}$	1/3
2	$\{6, 7\}$	$\{2, 5, 6, 7\}$	1/2
1	$\{5, 6, 7\}$	$\{5, 6, 7\}$	1

더 거친 알갱이(granulation)를 가지는 상위레벨일수록 더 정확한 러프집합 근사화를 가진다. 계층화된 알갱이를 사용하여 근사화를 위한 적당한 알갱이를 찾을 수 있다. 본 연구에서는 상위레벨에 코어(core)가 되는 속성을 적용하여 비슷한 분류의 알갱이를 찾은 후, 각각의 객체들에 대한 상대 리덕트(reduct)를 찾아 분류규칙을 찾아낸다.

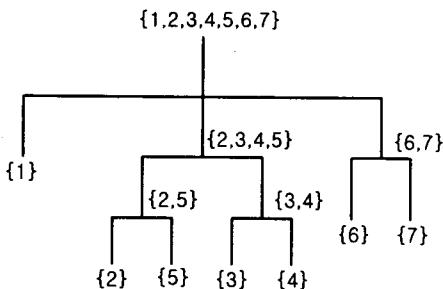


그림 1. Hierarchical Granulation

다음 단계에서 각 의사결정규칙에서 조건부 속성의 불필요한 속성을 줄여나가야 한다. 이를 위해 {1}, {2,3,4,5}, {6,7}에 대한 조건부 속성의 코어와 리덕트를 계산해야 한다.

{1}의 리덕트는 a 이다. {2,3,4,5}에 대한 리덕트를 구하면 다음과 같다.

$$F_{(2)} = \{[1]_a, [1]_b, [1]_d\}, [0]_e = \{2\}$$

불필요한 범주를 찾기 위해 한번에 한 개씩의 범주를 삭제하고 나머지 범주들의 교집합이 의사결정범주(decision category) $[0]_e = \{2\}$ 내에 포함되는지를 확인해야 한다.

$$[1]_a \cap [1]_b = \{2, 3, 4, 5\} \cap \{2, 5\} = \{2, 5\}$$

$$[1]_a \cap [1]_d = \{2, 3, 4, 5\} \cap \{2, 4\} = \{2, 4\}$$

$$[1]_b \cap [1]_d = \{2, 4\} \cap \{2, 5\} = \{2\}$$

따라서 리덕트(reduct)는 b, d 가 된다. 이런 식으로 나머지 객체들에 대한 리덕트를 구하면 다음과 같다.

표 3. 리덕트로 표현된 판단 테이블

U	a	b	d	e
1	0	-	-	0
2	-	1	1	0
3	1	0	-	1
4	1	0	-	1
5	-	-	2	2
6	2	-	2	2
7	2	-	2	2

위에서 구한 리덕트를 가지고 분류규칙을 발견하면 다음과 같다.

$$\begin{aligned} a_0 &\rightarrow e_0 \\ b_1 d_1 &\rightarrow e_0 \\ a_1 b_0 &\rightarrow e_1 \\ d_2 &\rightarrow e_2 \\ a_2 d_2 &\rightarrow e_2 \end{aligned}$$

규칙이 최소의 수가 되도록 공통이 되는 규칙들을 하나로 묶는다. 따라서 최소의 규칙은 다음과 같이 얻을 수 있다.

$$\begin{aligned} \text{if } a=0, & \quad \text{then } e=0 \\ \text{if } b=1 \text{ and } d=1, & \quad \text{then } e=0 \\ \text{if } a=1 \text{ and } b=0, & \quad \text{then } e=1 \\ \text{if } d=2, & \quad \text{then } e=2 \end{aligned}$$

3.3 WBC 데이터에 대한 응용

Wisconsin Breast Cancer 데이터는 위스콘신 대학 병원에서 1989년부터 1991년까지 수집한 데이터로 9개의 조건 속성(c_1, \dots, c_9)과 2개의 클래스(d_1, d_2)로 이루어진 1개의 판단속성으로 이루어져 있으며, 총 699개의 데이터로 이루어져 있다. (7) 학습 데이터로 369개의 데이터 중 불완전한 속성값을 갖는 14개의 데이터를

제외한 355개의 데이터가 사용되었으며, d_1 과 d_2 가 각각 189, 166개로 이루어져 있다. 시험 데이터로는 330개의 데이터가 사용되었다.

제안된 방법에 의해서 생성된 분류규칙의 수는 4개이며, 학습 데이터에 대한 분류율은 93.5%이고, 시험 데이터에 대한 분류율은 95.5%이다. 분류정도는 unit cost $C(d \rightarrow d')$ 을 사용하여 확인하였다.

$$C(d \rightarrow d') = \begin{cases} 1 & \text{if } d \neq d' \\ 0 & \text{otherwise,} \end{cases} \quad (8)$$

표 4. WBC 데이터에 대한 실험 결과

Rule	$n_1(A)$	$n_2(A)$	$C(A \rightarrow 1)$	$C(A \rightarrow 2)$
learnin.	1	175	9	0.4930
	2	14	157	0.0394
test	1	242	0	0.7333
	2	15	73	0.0455
				0.2212

4. 결 론

본 논문에서는 데이터 마이닝을 위한 분류기법으로 러프집합과 계층적인 분류방법을 기반으로 한 방법을 제안하였다. 제안된 방법은 계층구조의 상위레벨에 코어를 적용함으로써 상대 코어와 리덕트를 찾아내는데 있어 데이터의 크기를 줄여들게 하여 좀 더 효율적으로 분류규칙을 찾아내도록 하였기 때문에 부피가 큰 데이터일 경우 더욱 효과적이다.

WBC 데이터의 실험을 통한 결과에서 볼 수 있듯이 러프집합이론을 이용한 분류규칙 생성은 다른 방법에 비하여 규칙의 수가 적을 뿐만 아니라 조건부가 짧고 규칙의 형태가 간단하여 이해가 쉬우면서도 분류율이 좋다. 이러한 러프집합이론을 이용하여 분류규칙에 직접적으로 영향을 미치는 리덕트를 찾아내는 방법은 데이터 마이닝뿐만 아니라 제어기구를 구성하기 위한 규칙 생성에도 응용될 수 있다.

계층적인 분류방법의 경우, 이번 연구에서는 상위 한 단계만이 적용되었으나, 여러 단계를 사용할 경우 좀 더 정확한 결과를 얻을 수 있을 것이다. 하지만 분류규칙이 많아지고 복잡해질 수 있다. 따라서 계층적 구조의 단계와 규칙의 수를 적절히 조절하여 정확한 규칙을 찾아낼 수 있는 개선된 분류기법 연구가 필요하다.

(참 고 문 헌)

- Pawlak, Z., "Why Rough Sets?", Fuzzy Systems, Proc. of the 5th IEEE International Conf., Vol.2, pp.738-743, 1996
- Yanyi Yang, T.C.Chiam, "Rule Discovery Based On Rough Set Theory", Information Fusion, Proc. of the 3rd International Conf., Vol.1, pp. TuC4-11-16, 2000
- Pawlak, Z., "Granularity of Knowledge, Indiscernibility and Rough Sets", Fuzzy Systems Proc. IEEE World Congress on Computational Intelligence, Vol.1, pp.106-110, 1998
- Y.Y.Yao, "Stratified Rough Sets and Granular Computing", 18th International Conf. of the NAFIPS, pp. 800-804, 1999
- Y.Y.Yao, "Rough Sets, Neighborhood Systems, and Granular Computing", Proc. of the IEEE Canadian Conf. on Electrical and Computer Eng., pp.1553-1558, 1999
- M.B. Gorzalczany, Z.Piasta, "Neuro-fuzzy approach versus rough-set inspired methodology for intelligent decision support", Information Science, pp.45-68, 1999
- O.L.Mangasarian, W.H.Wolberg, "Cancer diagnosis via linear programming", SIAM News, Vol. 23, Number 5, pp.1&18, 1990