Rough 집합을 이용한 근사 패턴 분류

○
*최 성 해, **정 환 묵
*경주전문대학 전자계산과
**대구효성가톨릭대학교 전자정보공학부

Approximate Pattern Classification with Rough set

*Choi Soung Hea, **Chung Hwan Mook
*Kyongju Junior College
**Department of Computer Engineering, Catholic University of Taegu Hyosung

Abstract

In this paper, We propose the concept of approximate Classification in the field of two group discriminant analysis. In our approach, an attribute space is divided into three subspaces. Two subspaces are for given two group and one subspace is for a boundary area between the two groups. We propose Approximate Pattern Classification with Rough set. We also propose learning procedures of neural networks for approximate classification. We propose two weighting methods which lead to possibility analysis and necessity analysis. We illustrate the proposed methods by numerical examples.

1. 서 론

패턴분류는 선형식별 함수에의해 분류하며 또한, 계층형 신경망을 이용하며 패턴을 분류하는 근사식별을 분류가 가능하다. 한편, 비선형함수도 BP 알고리즘 학습에의해 구할 수 있다. 실제 Rumnelhart는 신경망에 의해 배타적 OR 문제나 우수성 식별 문제등의 비선형으로 두 그룹으로 식별하는 예가 발표하고 있다. 그러나 실제의 학습에서 과학습의 문제, 주어진 데이터를 완전하게 식별하는 신경망이 미학습의 데이터에 대해서는 유효하다고 할 수 없다. 이러한 신경망 학습의 효과는 학습시간의 관점에서 현실성이 없다. 러프 집합은 1982년 Z. Pawlak에 의해 제안 되었으며, 패턴분류를 데이터의 동치류에의해 평가 한다. 본 논문에서는 두 그룹에 속할 가능성이 있는 경계 영역에 있는 그룹으로 근사 식별을 하고, 패턴공간의 식별불가능한(indiscernibility relation) 경계영역(borderline)을 러프 집합을 이용하여 분류하고, 처리하는 방법을 제안한다. 기존의 패턴분류는 cut-off의 영역으로 표현되므로 결정 영역들간의 충첩된 영역을 분류할 수 없었다. 패턴 공간을 상한과 하한, 식별불가능한 경계영역으로 정의하여 신경망으로 학습하는 근사 패턴 분류 방법을 제안한다. 경계 영역의 존재를 반증한 근사식별을 하기 위해 신경망의 학습 알고리즘을 이용하여 수치 계산 예를 적용한다.

2. 러프 집합과 러프 멤버쉽 함수

2.1 러프 집합

러프 집합은 U가 전체집합일 때 R은 $U \times U$ 상의 동치관계로써 $A=(U,R)$를 근사공간(Approximation Space)라 한다. R은 식별불능관계(indiscernibility relation)가 된다. 이 때 근사공간 A를 동치관계 R로써 분할한 상공간이라 한다. X는 U의 부분집합으로써 X를 포함하는 A의 최소 정의 가능한 집합은 A에 대한 X의 상한근사(Upper Approximation)이며, 그 집한의 필연성(necessity)인 개체들이다. 이것을 $A^*(X)$로 나타낸다. 이와 같이 X에 포함되는 A의 최대 정의 가능한 집합을 A에 대한 X의 하한근사(Lower Approximation)하며, 그 집합에 속할 가능성(possibility)이 있는 개체들이다. 이것을 $A_*(X)$로 나타낸다.

$$A^*(X)=\{x\in U \mid A(x) \cap X\neq\emptyset\} \qquad (2.1)$$

$$A_*(X)=\{x\in U \mid A(x)\subset X\} \quad (2.2)$$

$B(X)$는 가능성의 그룹과 필연성의 그룹으로도 분류할 수 없는 그룹을 A에 대한 X의 경계집합이라 한다.

$$B(X)=A^*(X)-A_*(X) \quad (2.3)$$

$B(X)=\emptyset$이면, 집합 X는 속성 A에 대해 크리스프하고, $B(X)\neq\emptyset$이면, 집합 X는 속성 A에 대해 러프(rough)하다.

러프 집합을 특성지우는데는 두 가지 방법이 있다. 첫째는 accuracy measure이고, 둘째는 rough set의 분류이다. 계수(coefficient)는 집합의 경계영역이 얼머나 큰지를 표현하지만 경계 구조에 대해서는 업급하지 않는다. 반면에 러프 집합들의 분류는 경계영역의 크기에 관한 정보은 제공하지 않지만 경계영역이 어떻게 구성되었는가를 알 수 있다. 그러므로 러프집합 X는 근사 정확도(accuracy of approximation)라는 $\alpha_A(X)$의 특성을 나타낸다고 할 수 있다. 이 때 근사공간 $A=(U,R)$에 대한 집합 X의 근사 정도를 나타내는 $\alpha_A(X)$는 다음과 같다.

$$\alpha_A(X)=Card\,(A_*(X))\ /\ Card\,(A^*(X)) \quad (2.4)$$

단, $Card\,(X)$는 집합 X의 원소 수이다. 식 (2.4)에서 밝혀지듯이 $0\le\alpha_A(X)\le 1$이고, 만약 X가 A에 정의 가능한 집합이면 $\alpha_A(X)=1$이 되고, X는 근사공간 A에서 완전히 속하며, $\alpha_A(X)$의 값이 큰 만큼 근사에 가깝다.

또한, X_i는 U의 부분집합이고 $F=\{X_1,\cdots,X_n\}$는 U의 부분 집합의 군일 때 $F\subseteq R$, $F\neq 0$이면 $\bigcap F$도 동치관계가 되고 $IND(F)$는

$$[X]IND(F)=\bigcap_{R\in F}[X]_R \quad (2.5)$$

이때, F는 U의 분류가 된다. X_i는 F의 클래스가 된다. A에 대한 F의 상한에서의 근사와 하한에서의 근사는 다음과 같다.

$$A^*(F)=\{A^*(X_1),\ \cdots\ ,A^*(A_n)\} \quad (2.6)$$

$$A_*(F)=\{A_*(X_1),\ \cdots\ ,A_*(X_n)\} \quad (2.7)$$

A에대한 분류 $F=\{X_1,\ \cdots\ ,X_n\}$의 근사 정도는

$$\beta_A(F)=Card\,(\bigcup_{i=1}^{n}A_*(X_i))\ /\ Card(U) \quad (2.8)$$

모든 X_i가 A의 정의 가능한 집합이면 $\beta_A(F)=1$이 되고, 분류 F는 근사 공간 A에서 완전히 포함된다.

2.2 러프 멤버쉽 함수

러프집합의 근사 정도로써 신경망으로 학습시키기 위한 러프 멤버쉽함수 $\mu_X^A(x)$는 다음과 같다.

$$\mu_X^A(x)=|\ X\cap A(x)\ |\ /\ |\ A(x)\ | \quad (2.9)$$

이 때 $\mu_X^A(x)\in[0,1]$이다. 멤버쉽 함수 $\mu_X^A(x)$의 값은 조건적인 가능성의 종류이다. $x\in X$는 확실하게(certainty) 디그리에 포함된다고 할 수 있고, $1-\mu_X(x)$는 불확실하게(uncertainty) 디그리에에 포함된다. 러프 멤버쉽 함수는 근사와 경계 영역을 다음과 같다.

$$A_*(X)=\left\{x\in U\ :\ \mu_X^A(x)=1\right\} \quad (2.10)$$

$$A^*(X)=\left\{x\in U\ :\ \mu_X^A(x)>0\right\} \quad (2.11)$$

$$B(X)=\left\{x\in U\ :\ 0<\mu_X^A(x)<1\right\} \quad (2.12)$$

$\delta(x)$는 객체 x에 포함된 결정 규칙이다. 이 규칙의 신뢰성 벡터는 다음과 같다.

$$C(\delta(x))=\begin{cases}1, & \text{if}\ \ \mu_X^A(x)=0\ or\ 1\\ \mu_X^A(x), & \text{if}\ \ 0<\mu_X^A(x)<1\end{cases} \quad (2.13)$$

신뢰성 벡터와 같이 하나의 일관성 규칙과 비일관성 규칙을 유도할 수 있다. 또한, $0\le\beta<0.5$ 이며, β는 실수일 때 다음과 같이 근사한다.

$$A_{*\beta}(X)=\left\{x\in U\ :\ \mu_X^A(X)\ge 1-\beta\right\} \quad (2.14)$$

$$A_\beta^*(X)=\left\{x\in U\ :\ \mu_X^A(x)>\beta\right\} \quad (2.15)$$

만약 $\beta=0$이면 경계영역은 다음과 같다.

$$B^\beta(X)=A_\beta^*(X)-A_{*\beta}(X)=\left\{x\in U:\beta<\mu_X^A(x)<1-\beta\right\} \quad (2.16)$$

3. 근사 패턴 분류

패턴 분류는 미지의 데이터를 식별하기 위해 이미 알고 있는 데이터의 속성값에서 식별 규칙을 구하는 것이다. m개의 샘플 속성값이 전체집합 Ω(단 $\Omega\subset R^n$)상의 X_p ($p=1,2,\cdots,m$)로써 주어져 있다고 한다. 단, $X_p=(X_{p1},X_{p2},\cdots,X_{pn})$이고, 벡터 X_p는 n차원 속성 공간내에서의 위치를 나타낸다. 식별 분석은 주어진 데이터를 이용해서 전체집합 Ω를 다음과 같이 분할한다.

$$\Omega=\Omega_1\bigcup\Omega_2 \quad (3.1)$$

$$\Omega_1\bigcap\Omega_2=\emptyset \quad (3.2)$$

즉, 전체집합 Ω가 1그룹을 나타내는 부분집합 Ω_1과 2그룹을 나타내는 부분집합 Ω_2로 분할한다. 이때 다음 관계가 성립되면 Ω_1,Ω_2에 의한 전체집합 Ω의 분할은 주어진 데이터가 식별된다.

$$X_p\in\Omega_1 \quad \forall p \quad p\subset P_{G1} \quad (3.3)$$

$$X_p \in \Omega_2 \quad \forall p \quad p \subset P_{G2} \tag{3.4}$$

그러나 경계영역이 존재하는 식(2.3)에 의해 전체 집합이 분할하는 방법은 현실적으로는 어렵고 직관적으로 반영한다. 그러므로 본 논문에서는 두 영역 Ω_1과 Ω_2와의 사이에 경계영역 Ω_B가 존재하는 것을 가정한 근사 식별의 방법을 제안한다.

$$\Omega = \Omega_1 \cup \Omega_2 \cup \Omega_B \tag{3.5}$$

$$\Omega_1 \cap \Omega_2 = \Omega_1 \cap \Omega_B = \Omega_2 \cap \Omega_B = \emptyset \tag{3.6}$$

위의 영역을 이용하면 미지의 데이터 X_p의 식별은 다음의 식별 규칙이 성립한다.

R1 : If $X_p \in \Omega_1$ then p is G_1

R2 : If $X_p \in \Omega_2$ then p is G_2

R3 : If $X_p \in \Omega_B$ then p is G_1 or G_2

영역 Ω_1에 속하는 데이터 X_p는 1그룹(G1)으로 판단되고, 영역 Ω_2에 속하는 데이터은 2그룹(G2)으로 판단된다. 또, 영역 Ω_B에 속하는 데이터는 어느 쪽도 속하지 않는다고 판단된다.

본 논문에서는 신경망을 이용하여 다음의 조건을 만족한 영역 Ω_1, Ω_2, Ω_B를 구할 수 있다.

$$X_p \in (\Omega_1 \cup \Omega_B) \quad \forall p \quad p \subset P_{G1} \tag{3.7}$$

$$X_p \in (\Omega_2 \cup \Omega_B) \quad \forall p \quad p \subset P_{G2} \tag{3.8}$$

3개의 영역이 위의 조건을 만족하면 주어진 데이터는 식별규칙 R1, R2, R3에 모순이 일어나지 않는다.

4. 신경망의 학습 알고리즘

4.1 신경망에의한 식별 분석

신경망은 출력층에 1개의 유니트, 입력층에 n개의 유니트를 갖는 계층형 신경망으로써, 중간층 및 출력층 유니트의 입출력 함수는 시그모이드 함수를 이용한다. 또, n차원 벡터 X_p, ($p=1,2,\cdots,m$)가 입력될 때의 신경망에서 출력값을 O_p이다. O_p에 대한 교사신호 t_p를 다음과 같이 정할 수 있다.

$$t_p = \begin{cases} 1, & \forall p \quad p \subset P_{G1} \\ 0, & \forall p \quad p \subset P_{G2} \end{cases} \tag{4.1}$$

그러므로 신경망의 학습은 다음의 평가 함수의 최소화를 목적으로 이루어지고 있다.

$$E = \sum_{p=1}^{m} (t_p - O_p)^2 / 2 \tag{4.2}$$

학습 후 신경망에 대해 입력 벡터 X에 대응한 출력값을 $\mu_X^A(x)$로 나타낼 수 있으면 다음과 같이 전체집합 Ω의 2개로 분할을 할 수 있다.

$$\Omega_1 = \left\{ X \mid \mu_X^A(X) \geq 0.5, \quad X \in \Omega \right\} \tag{4.3}$$

$$\Omega_2 = \left\{ X \mid \mu_X^A(X) < 0.5, \quad X \in \Omega \right\} \tag{4.4}$$

신경망에서의 출력값 $\mu_X^A(x)$가 다음의 관계를 만족하면 신경망에 의해 주어진 모든 데이터는 다음과 같이 분류된다

$$\mu_X^A(X_p) \geq 0.5 \quad \forall p \quad p \subset P_{G1} \tag{4.5}$$

$$\mu_X^A(X_p) < 0.5 \quad \forall p \quad p \subset P_{G2} \tag{4.6}$$

4.2 근사 식별을 위한 신경망 학습

근사 식별을 하기위해 신경망에의한 식(3.5)-(3.8)의 조건을 바탕으로 3개의 영역 $\Omega_1, \Omega_2, \Omega_B$을 구하는 방법을 제안한다. 먼저, 신경망의 학습을 위한 영역 $\Omega_1, \Omega_2, \Omega_B$를 구하기 위해 주어진 m개의 데이터 사이에 다음의 관계를 만족해야 한다.

$$\mu^*{}_X^A(X_p) > 0.5 \quad \forall p \quad p \subset P_{G1} \tag{4.7}$$

$$\mu^*{}_X^A(X_p) \in [0,1] \quad \forall p \quad p \subset P_{G2} \tag{4.8}$$

$$\mu_*{}_X^A(X_p) \in [0,1] \quad \forall p \quad p \subset P_{G1} \tag{4.9}$$

$$\mu_*{}_X^A(X_p) < 0.5 \quad \forall p \quad p \subset P_{G2} \tag{4.10}$$

여기서 $\mu^*{}_X^A(X_p)$, $\mu_*{}_X^A(X_p)$는 신경망의 출력값이다. 식(4.7)-(4.10)에 의해 신경망의 데이터를 식별한다. 또한, Ω_B의 영역은 다음과 같다.

$$\bar{\mu}_X^A(X_p) = \left\{ \mu^*{}_X^A(X_p) + \mu_*{}_X^A(X_p) \right\} / 2 \tag{4.11}$$

따라서 식(3.7)-(3.8)을 만족하면 다음과 같다.

$$\Omega_1 \cup \Omega_B = \left\{ X \mid \bar{\mu}_X^A(X) > 0.25, \ X \in \Omega \right\} \tag{4.12}$$

$$\Omega_2 \cup \Omega_B = \left\{ X \mid \bar{\mu}_X^A(X) < 0.75, \ X \in \Omega \right\} \tag{4.13}$$

식(3.5)의 관계를 이용하면 영역 $\Omega_1, \Omega_2, \Omega_B$이 다음과 같이 얻어진다.

$$\Omega_1 = \Omega - (\Omega_2 \cup \Omega_B) = \left\{ X \mid \bar{\mu}_X^A(X) \geq 0.75, \ X \in \Omega \right\} \tag{4.14}$$

$$\Omega_2 = \Omega - (\Omega_1 \cup \Omega_B) = \left\{ X \mid \bar{\mu}_X^A(X) \leq 0.25, \ X \in \Omega \right\} \tag{4.15}$$

$$\Omega_B = \Omega - (\Omega_1 \cup \Omega_B) \cap (\Omega_1 \cup \Omega_B) = \left\{ X \mid 0.25 < \mu_X^A(X) < 0.75, \ X \in \Omega \right\} \tag{4.16}$$

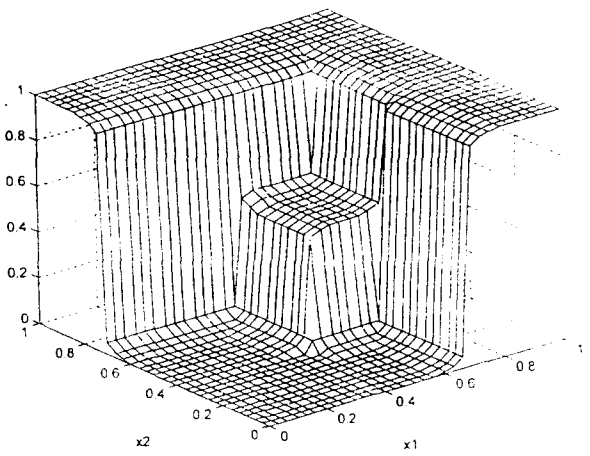
이와같이 영역 $\Omega_1, \Omega_2, \Omega_B$을 설정할 때 식(3.5)-(3.8)이 성립한다. 이상과 같이 식(4.7)-(4.10)을 만족하는 신경망이 만족하면 식(4.14)-(4.16)에 의해 얻어진 영역 $\Omega_1, \Omega_2, \Omega_B$를 이용해서 근사식별을 할 수 있다.

4.3 수치 적용 예

신경망의 학습은 각가의 데이터를 학습하는 순차 수정법을 이용하고, 학습의 속도를 빠르게 하기 위해 관성항을 도입했다. 또,학습 계수는 $\eta=0.25$로 써, 관성항에 관한 계수는 $\alpha=0.9$로 설정 했다. 이와 같은 데이터에 대해 중간층의 5개의 유니트를 갖는 신경망을 이용해서 각각의 데이터를 BP알고리즘으로 30000회 학습을 했다. 학습 후의 신경망에서의 출력값과 제안된 방법에 의한 $\Omega_1, \Omega_2, \Omega_B$의 영역으로 분류되는 출력값을 제시한다. 수치 예에서 전체집합 Ω를 [0,1]×[0,1]로써 60개의 데이터를 <표 1>과 같다. 본 논문에서 제안한 러프 집합을 적용하여 신경망으로 학습하여 분류된 결과는 [그림 2]과 같다. [그림 2]에서는 Ω_1은 1 그룹에 속하는 영역을, Ω_2는 2 그룹에 속하는 영역을, Ω_B는 두 그룹 사이의 경계 영역을 나타내고 있다. 제안된 방법의 결과에서 3개의 영역은 직관적인 인식과 일치하고 있음을 알 수 있다.

<표 1> 실험 데이터

	Class A		Class B	
NO	x_1	x_2	x_1	x_2
1	5.30	5.20	2.00	1.95
2	5.70	6.80	1.50	3.00
3	7.90	9.10	2.00	4.90
4	5.80	4.30	3.05	6.00
5	5.90	4.70	3.80	1.75
6	5.80	6.70	5.10	2.20
7	7.00	4.10	6.65	6.70
8	6.95	5.50	2.70	3.30
9	6.75	6.00	3.30	3.00
10	8.15	4.95	3.50	3.40
11	8.00	5.50	4.95	2.85
12	7.70	5.50	4.25	3.30
13	7.90	6.20	3.00	5.10
14	7.45	7.45	4.00	3.80
15	7.60	6.80	4.45	3.90
16	6.90	7.00	6.25	3.50
17	6.80	6.90	6.30	3.80
18	6.60	7.30	3.40	4.40
19	7.45	7.50	3.70	4.20
20	7.75	8.35	4.00	4.30
21	8.10	7.50	4.30	4.75
22	8.20	7.40	3.80	5.35
23	8.50	7.20	4.10	5.10
24	9.00	6.80	[illegible]	[illegible]
25	8.95	6.50	4.75	5.00
26	8.70	6.90	[illegible]	[illegible]
27	9.20	6.90	[illegible]	[illegible]
28	9.20	7.30	[illegible]	[illegible]
29	8.60	8.00	6.40	6.00
30	8.85	8.30	6.80	6.20



[그림 2] 제안된 방법에 의한 학습 영역

5. 결 론

본 논문에서는 패턴인식에서 분류하기 모호한 정보를 신경망을 적용하여 근사식별하는 방법을 제안했다. 신경망에 의한 종래의 학습은 많이 연구되고 있지만, 될 수 있는대로 짧은 계산 시간에서 주어진 데이터를 바르게 식별하는 것을 목적으로하고 있다. 제안된 방법에서는 다차원 속성 공간 내에서 어느 쪽의 그룹로 속할 가능성이 있는 경계 영역의 식별불가능한 관계를 러프 집합을 이용하여 분류하였으며, 패턴 공간을 상한과 하한, 식별불가능한 경계영역으로 정의하여 신경망으로 학습하는 근사 패턴 분류 방법을 제안했다. 마지막으로 제안된 방법을 이용하여 수치 적용의 예로써 근사 분류된 결과가 직관적인 판단과 일치함을 확인한다.

[참고 문헌]

1) S.M.Weiss and C.A.Kulikowski : Computer Systems That Learn,Morgan Kaufman,San Mateo,California, 1991.

2) N.P.Archer and S.Wang : "Fuzzy Set Representation of Neural Naetwork Classification Boundary",IEEE Trans,Systems, Man and Cybernetice,Vol.SMC-21,No.4,pp.735-742, 1991.

3) D.E.Rumelhart,J.L.McClelland and the PDP Research Group : Parallel Distributed Processing, Vol 1, 1986

4) A Mrozek, Rough sets and Dependency Analysis among Attributes in Computer Implementations of Expert's Inference Models, Int. J. of Man Machine Studies, Vol 30,