Rough Set을 이용한 퍼지 규칙의 생성

Extraction of Fuzzy Rules from Data using Rough Set

조영완, 노흥식, 위성윤, 이희진, 박민용
Young Wan Cho, Heung Sik Noh, Sung Yoon Wi, Hee Jin Lee, Mignon Park

Dept. of Electronic Eng. Yonsei Univ.
134 Shinchon-dong, Seodaemun-ku, Seoul, Korea
Tel : 361-2868 Fax : 312-4584
E-mail : mignpark@bubble.yonsei.ac.kr

요 약

Rough Set theory suggested by Pawlak has a property that it can describe the degree of relation between condition and decision attributes of data which don't have linguistic information. In this paper, by using this ability of rough set theory, we define a occupancy degree which is a measure can represent a degree of relational quantity between condition and decision attributes of data table. We also propose a method that can find an optimal fuzzy rule table and membership functions of input and output variables from data without linguistic information and examine the validity of the method by modeling data generated by fuzzy rule.

1. 서론

퍼지 이론은 멤버쉽 함수의 정의와 퍼지 규칙의 생성에 있어서 인간 조작자의 경험과 지식에 의존함으로써 미세 조정에 많은 시행착오를 거쳐야 하는 어려움이 있었다. 또한, 조작자의 조작 행위, 지식 또는 경험을 언어 변수화하여 제어 규칙을 만들어내는 과정은 비교적 용이하였으나 경험과 지식이 가미되지 않은 순수한 입·출력 데이터쌍으로부터 입·출력 관계를 기술하는 퍼지 규칙을 생성하는데는 여러가지 어려움이 있었고 이에 대한 많은 연구가 진행중이다.

기존의 퍼지 모델링 방법은 패턴 인식 기법을 이용한 접근 방식[1], 입력 공간의 퍼지 분할을 기반으로 후건부를 선형식으로 표현하는 TSK 모델 접근 방식[3][4][5], NN(뉴럴 네트워크)의 학습 능력, GA (Genetic Algorithm)의 최적해 탐색 능력 등을 이용한 모델의 최적 파라메터 탐색법[6][7] 등이 있다. 패턴 인식을 이용한 접근 방식은 데이터의 클러스터링에 의존하므로 클러스터링 기법에 민감하고 입력 공간을 퍼지 분할하여 퍼지 모델링을 수행하는 TSK 모델은 선형성이 강한 모델의 묘사력은 뛰어나지만 비선형성에 대해서는 그 알고리즘이 상당히 복잡해지고 묘사력도 떨어지는 단점이 있다.

한편, Pawlak이 제안한 Rough Set 기법[8][9]은 조건부 속성과 결론부 속성간의 연관성을 수치화하여 표현할 수 있는 장점이 있으므로 언어 정보가 부족한 데이터쌍의 조건부와 결론부 사이의 규칙을 정량화하기에 적합하다. 본 논문에서는 입·출력 속성의 관련성을 나타내는 Rough Set 기법을 이용하여 퍼지 규칙표에서 조건부 속성과 결론부 속성의 연관성을 나타내는 척도를 제안하고 이로부터 언어 정보에 대한 사전 지식이 없는 입·출력 데이터쌍을 묘사할 수 있는 최적의 퍼지 규칙과 멤버쉽 함수를 찾아낼 수 있는 정사영 분할법을 제시하고 이와 같이 구성된 퍼지 규칙이 주어진 데이터를 얼마나 잘 묘사할 수 있는지를 확인한다.

2. Rough Set[8]

Pawlak에 의해 제안된 Rough Set[9]은 데이터의 불확실성을 다루기 위한 수학적 이론으로 집합의 상한

근사(upper approximation)과 하한 근사(lower approximation)에 대한 정의에서 출발하여 Rough Fuzzy 제어, 모델링과 시스템 동정(identification), 실험 데이터로부터의 규칙 발견 등에 이용되고 있다. 어떤 집합의 하한 근사는 그 집합과 "확실히" 관련있는 개체들로 구성되고 상한 근사는 그 집합과 관련될 "가능성"있는 원소들로 구성되며 상한 근사와 하한 근사의 차이는 경계 영역을 구성하게 된다.

2.1 여러가지 용어의 정의

Rough Set 이론은 논의 영계에 있는 모든 개체는 그것을 기술하기 위하여 정보와 관련지어 생각할 수 있다는 가정에 근거를 두고 있으며 이를 다루기 위해 다음과 같은 용어들을 정의하여 사용한다. 데이터가 표1과 같이 표현되어 있을 때 세로축의 각 데이터 즉, 개체(object)는 가로축의 속성(attribute)이라는 정보의 관점에서 표현된다.

개체를 포함하는 전체 집합 U와 속성들의 집합 A, 그리고 각 개체들의 각각의 속성 $a \in A$에 대한 속성값 V_a가 주어져 있을 때, 속성 A의 모든 부분 집합 B에 대하여 불구분 관계(indiscernibility relation)라는 2진 관계 $I(B)$를 다음과 같이 정의한다.

모든 $a \in B$에 대해 $a(x) = a(y)$ 이면 $x\ I(B)\ y$로 나타낸다. 이때, $a(x)$는 원소 x의 속성 a에 대한 속성값 V_a를 의미한다.

이와 같은 불구분 관계에 의해 전체 집합 U의 모든 원소는 구분 불가능한 원소들을 포함하는 클래스들로 분할되어지는데 속성 집합 B에 의한 분할을 $U/I(B)$ 또는 간단히 U/B로 나타내며 원소 x를 포함하고 있는 하나의 등가 클래스(equivalence class), 즉 U/B의 분할 영역을 $B(x)$라고 표현한다. x와 y가 $I(B)$의 등가 클래스에 속해 있다는 것은 속성 집합 B의 관점에서는 개체 x, y를 구별할 수 없음을 의미하는 것으로 x와 y는 B-구분불가능(B-indiscernible) 이라고 말한다. 임의의 집합 X를 속성으로 묘사하기 위하여 상한 근사와 하한 근사를 다음과 같이 정의한다.

전체 집합 U의 부분 집합 X에 대하여 다음의 연산으로 정의되는 두 집합 $B_*(X)$, $B^*(X)$를 각각 X의 B-하한, B-상한 근사라고 한다.

$$B_*(X) = \{ x \in U \mid B(x) \subseteq X \}$$

$$B^*(X) = \{ x \in U \mid B(x) \cap X \neq \phi \}$$

$BR_B(X) = B^*(X) - B_*(X)$를 X의 B-경계 영역이라고 하는데 이는 집합 X가 속성 B의 관점에서 어느 정도 정확하게 관련되는가의 정도를 나타내는 의미를 가지고 있으므로 $BR_B(X) = \phi$이면 집합 X는 속성 B에 대해 크리스프하다고 하고 $BR_B(X) \neq \phi$이면 집합 X는 속성 B에 대해 러

attribute / object	error	error derivative	output	attribute / object	error	error derivative	output
d1	NB	NB	PB	d11	NB	NB	PS
d2	NS	NB	PB	d12	NS	NB	PB
d3	NS	NS	PS	d13	NB	ZE	ZE
d4	NB	NS	PS	d14	ZE	ZE	ZE
d5	NB	NB	PB	d15	ZE	PS	NS
d6	NB	NS	PB	d16	PS	ZE	NS
d7	NS	ZE	PS	d17	ZE	PB	NS
d8	NB	ZE	PS	⋮	⋮	⋮	⋮
d9	NS	ZE	PS				
d10	NS	NS	PS	dN	PB	PB	NB

NB : Negative Big, NS : Negative Small, ZE : Zero, PS : Positive Small, PB : Positive Big

표1 Rough Set에서의 규칙표

프(Rough)하다고 한다. 다시 말해서 Rough Set X는 근사 정확도(accuracy of approximation)라고 불리는 다음의 $\alpha_B(X)$로 그 특성이 나타난다고 말할 수 있다.

$$\alpha_B(X) = \frac{|B_*(X)|}{|B^*(X)|}$$

이때, $|B^*(X)|$와 $|B_*(X)|$는 각각 $B^*(X)$, $B_*(X)$의 원소의 갯수를 나타낸다.

위의 정의에서 알 수 있듯이 Rough Set은 집합 X 자체만의 정의가 아니라 속성 집합과 관련된 정의라는 사실에 유념할 필요가 있다.

2.2 속성의 부분 의존도와 일관성 척도

퍼지 규칙을 발견하기 위하여 데이터의 전건부에 해당하는 속성과 후건부에 해당하는 속성의 관련성을 나타내야 하는 것과 같이 데이터 분석을 위해 속성들 사이의 관련성을 파악하는 것이 중요할 때가 있다. 이를 위하여 부분 의존도(partial dependency)라고 불리는 속성들간의 의존도를 다음과 같이 정의하여 사용한다.

D와 C가 A의 부분 집합일 때,

$$k = \frac{|POS_C(D)|}{|U|},$$

$$POS_C(D) = \bigcup_{X \in U/I(D)} C_*(X)$$ *이면 D는 C에 k정도로 의존한다고 한다.*

이때, $POS_C(D)$는 C에 대한 분할 $U/I(D)$의 양의 영역(positive region)이라고 하는데 이는 속성 C의 관점에서 나누어진 어떤 클래스에 속한 모든 원소가 분할 $U/I(D)$로 만들어지는 어떤 영역에 속해 있을 때 이들 원소들로 구성됨을 의미하므로 "C의 어떤 클래스이면 D의 어떤 클래스가 된다라고 말할 수 있는 정도"를 나타낸다. 다시 말해서 C에 대한 D의 부분 의존도 k는 전체 집합의 모든 원소중 C라는 속성들로 고려해도 $U/I(D)$로 분할된 영역에 정당하게 분류되어지는 원소의 비율을 의미한다. 개념적으로 전체 집합 U의 원소가 C이면 D라고 말할 수 있는 정도를 나타내고 있으므로 퍼지 규칙에 있어서 전건부 속성과 후건부 속성의 관련성을 묘사하는데 적합하다고 할 수 있다. 표1에서 살펴보면 속성 error와 error derivative는 조건부 속성 C, 속성 output은 결론부 속성 D에 해당되고 각 행은 전건부 속성에 대한 후건부 속성의 결정 규칙을 나타내고 있다. 표1에서 데이터가 d1부터 d17까지 17개의 데이터만 존재한다고 할 경우 조건부가 (NS, NB)이면 결론부는 항상 PB이 되는데(d2, d12) 비해 d1, d5, d11은 조건부 속성이 (NB, NB)로 같은데도 불구하고 다른 결론부 속성값을 나타내고 있다. 전자의 규칙을 일관적(consistent)이라고 하고 후자의 규칙을 비일관적(inconsistent)이라고 하는데 일관성 척도(consistent measure)를 사용해서 규칙의 일관성의 정도를 나타낸다.

C와 D가 각각 조건부 속성, 결론부 속성일 때

$$\gamma(C, D) = \frac{|POS_C(D)|}{|U|}$$ *를 일관성 척도라고 한다.*

정의에서 알 수 있듯이 일관성 척도는 데이터 표에서 모든 데이터에 대한 일관적 규칙으로 표현된 데이터의 수의 비로 나타낸다.

3. 퍼지 규칙과 멤버쉽 함수의 생성

주어진 데이터쌍에 대하여 이를 묘사할 수 있는 퍼지 규칙을 발견해 낸다는 것은 데이터에 대한 언어 변수를 정의하고 정의된 언어 변수로 입·출력 데이터의 관계를 표현하는 규칙 표를 작성하는 것으로 작성된 표는 주어진 데이터쌍들을 묘사함에 있어서 일관성이 있어야 한다. 표1에서 살펴보았듯이 d1, d5, d11의 경우, 조건부 속성(NB, NB)에 대하여 다른 결론부 속성값(PB, PS)을 나타낸다는 것은 퍼지 규칙의 일관성이 결여되는 것으로 일관성을 높이기 위해서는 데이터에 해당되는 언어 변수를 변화시키거나 퍼지 규칙을 변화시키는 방법을 모색해야 할 것이다.

본 논문에서는 개량된 일관성 척도를 정의하여 퍼지 규칙의 전건부와 후건부 사이의 일관성을 높일 수 있도록 멤버쉽 함수를 정의함과 동시에 퍼지 규칙을 생성할 수 있는 방법을 제안한다. 입력 공간(조건부 속성) $X \times Y$와 출력 공간(결론부 속성) Z가 다음 식(1)과 같은 방식으로 크리스프 분할되었다고 하자.

$$X_i \cap X_j = \phi, \quad i \neq j$$

$$X_1 \cup X_2 \cup \cdots \cup X_l = X$$

$$x_i \le X_i < x_{i+1} \qquad \text{식(1)}$$

크리스프 분할된 집합 X_i, Y_i, Z_i에 대한 퍼지 집합 FX_i, FY_i, FZ_i에 대한 멤버쉽 함수는 각각 다음과 같이 정의된 삼각형 모양의 구조를 갖는다.

$\mu_{FX_i}(x)$의 중심은 $\dfrac{x_i + x_{i+1}}{2}$,

$$\mu_{FX_i}(x_i) = \mu_{FX_i}(x_i) = 0.5 \qquad \text{식(2)}$$

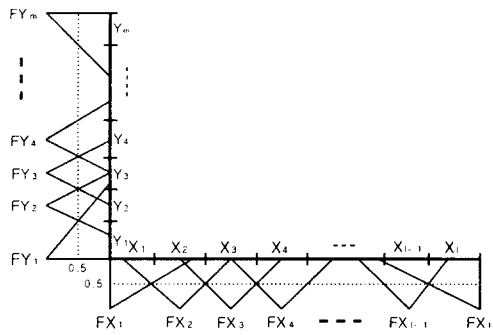

그림1 전건부 변수 공간의 퍼지 분할

입·출력 공간이 분할되었을 때 데이터를 묘사하는 퍼지 규칙을 생성한다는 것은 다음과 같은 함수 f를 결정하는 것과 같다.

$$f : FX \times FY \rightarrow FZ$$

이 때,

$$FX = \{ FX_1, FX_2, \cdots, FX_l \}$$

$$FY = \{ FY_1, FY_2, \cdots, FY_m \}$$

$$FZ = \{ FZ_1, FZ_2, \cdots, FZ_n \}$$

함수 f를 결정함에 있어서 전건부 변수와 후건부 변수의 대응에서 일관성이 있어야 하므로 이를 위하여 본 논문에서는 전건부 속성의 후건부 집합 Z_i에 대한 점유도(occupancy degree)를 정의하여 사용한다.

$$occup_{X_i \times Y_j}(Z_k) = \frac{| C(d) \cap X_i \times Y_j |}{| X_i \times Y_j |} \qquad \text{식(3)}$$

이때, Z_k는 결론부 속성 Z에 의해 분할된 클래스 즉, $Z_k \in U/I(Z)$이고, $| X_i \times Y_j |$는 $X_i \times Y_j$ 공간에 존재하는 데이터 수, $C(d)$는 데이터 $d \in Z_k$를 포함하는 전건부 속성 집합 C에 의한 등가 클래스이다.

최적화된 룰 테이블을 작성하는 것은 분할된 입·출력 공간에 대한 최적 사상 f를 결정하는 것으로 다음과 같이 정해진다.

$$f(X_i, Y_j) = Z_k \qquad \text{식(4)}$$

이때, $\max_{Z_d \in U/I(D)} occup_{X_i \times Y_j}(Z_d) = Z_k$

즉, 후건부 변수의 각 분할 영역에 대해 점유도가 최대가 되게 하는 전건부 분할 공간을 대응시킨다. 이와 같은 사상 f는 입력 변수 공간을 분할하는 방법에 따라 달라진다. 입력과 출력 변수의 연관성 즉, 일관성 척도를 최대화하기 위해 본 논문에서는 다음과 같은 정사영 최적 분할법을 제안하여 전건부 공간을 분할한다. 이미 분할된 후건부 변수별로 클래스를 형성하여 전건부 공간에 정사영시켰을 때 분할된 각 영역 $X_i \times Y_j$에 대해 $\prod_{i=1}^{l} \prod_{j=1}^{m} occup_{X_i \times Y_j}(Z_k)$이 최대가 되도록 하는 점선과 같은 분할 영역을 신경망(Neural Network)이나 유전자 알고리즘(Genetic Algorithm)과 같은 최적해 탐색 알고리즘을 이용하여 구한다.

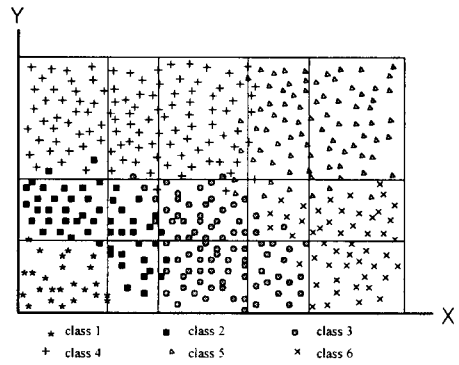

그림2 클래스별 정사영의 최적 분할

이와 같이 최적 분할이 이루어짐은 전건부 변수에 대한 언어 변수가 결정됨과 동시에 전건부 변수와 후건부 변수 사이의 사상 관계 f가 결정되어지므로 퍼

지 규칙이 생성됨을 의미한다.

4. modeling and results

본 논문에서 제안한 점유도와 정사영 최적 분할법을 이용한 데이터쌍의 모델을 위해 임의로 정의된 2입력 1출력 퍼지 시스템으로부터 랜덤하게 25×25 쌍의 데이터를 추출해서 사용하였으며 추출된 데이터는 그림3과 같다.

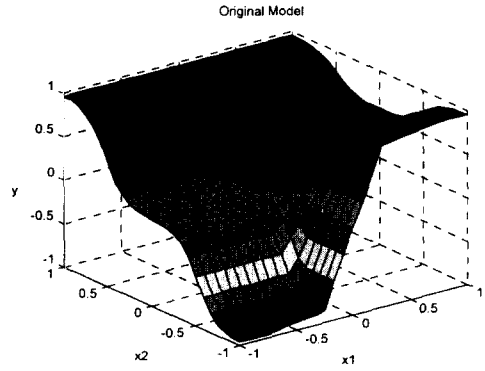

그림3 모델링에 사용된 데이터의 분포

모델링을 위한 첫 번째 단계로서 후건부 변수의 분할을 수행한다. 이는 fuzzy C-means 알고리즘과 같은 클러스터링 기법 등에 기반하여 할 수 있으나 본 논문에서는 동일한 5개의 영역으로 분할한다. 주어진 데이터쌍들을 분할된 영역에 따라 후건부 변수를 기준으로 크리스프하게 분류하고 각 클래스들을 입력 공간 $X\times Y$ 상에 정사영시키면 그림4와 같다.

입력 변수의 분할과 동시에 최적의 퍼지 규칙의 추출을 위해 $\prod_{i=1}^{l}\prod_{j=1}^{m} occup_{X_i\times Y_j}(Z_k)$을 최대화하도록 입력 공간을 분할하며 분할된 각 공간은 하나의 퍼지 규칙을 형성하게 된다.. 본 논문에서는 GA를 사용하였으며 크리스프하게 분할된 각 영역에 대해 식(2)에서 정의된 바와 같은 방법으로 퍼지 집합을 구성하게 하는 멤버쉽 함수를 구성할 수 있다. 멤버쉽 함수의 결정 과정에서 이루어진 최적 분할에 의해 각 입력 공간에 해당하는 출력 변수의 대응은 각 분할 공간에서의 후건부 클래스로 결정되며 그 결과는 표2와 같다. 그림6은 모델링 결과를 나타내고 그림7은 주어진 데이터와 모델과의 차를 나타낸다.

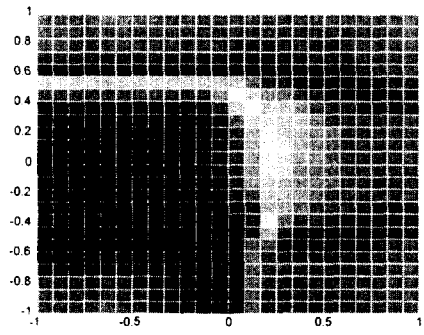

그림4 클래스별 정사영

Y \ X	NB	NS	Z	PS	PB
NB	NB	NS	Z	PS	PB
NS	NS	NS	Z	PS	PB
Z	PS	PS	PS	PS	PS
PS	PB	PS	PS	PS	PB
PB	PB	PB	PS	PB	PB

표2 modeling을 위한 rule table

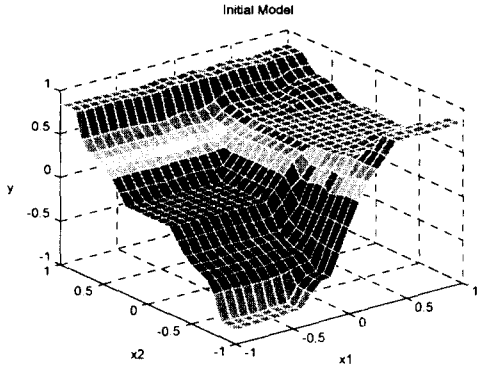

그림6 모델링 결과

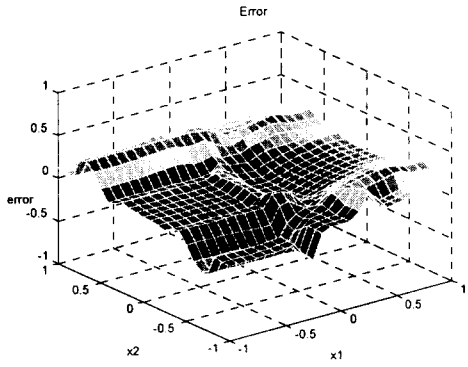

그림7 모델 에러

그림3, 그림6, 그림7에서 알 수 있듯이 본 논문에서 제안한 점유도와 정사영 최적 분할법을 이용한 퍼지 모델링은 그 모델링 알고리즘이 간단함에도 불구하고 모델링 에러가 거의 없는 상당히 근사화된 모델을 제공하는 초기의 퍼지 규칙과 입·출력 변수의 멤버쉽 함수를 생성할 수 있음을 알 수 있다. 제안한 모델링 알고리즘은 결과에서 보여주듯 모델의 구조를 파악하는데 적합한 특성을 나타냄을 알 수 있고 후건부 변수의 분할 클래스가 여러개 겹치는 부분에서 약간의 모델 에러가 발생하고 있는데 이를 줄이기 위한 보다 체계적인 정사영 분할법이 필요하다고 생각되며 경사 하강법 등의 미세 조정 과정을 거치면 에러는 상당히 줄 것으로 기대된다.

5. 결론

본 논문에서는 입·출력 속성의 관련성을 나타내는 Rough Set 기법을 이용하여 퍼지 규칙표에서 조건부 속성과 결론부 속성의 연관성을 나타내는 척도인 점유도를 제안하고 이로부터 언어 정보에 대한 사전 지식이 없는 입·출력 데이터쌍을 묘사할 수 있는 최적의 퍼지 규칙과 멤버쉽 함수를 찾아낼 수 있는 정사영 분할법을 제안하였고 이와 같이 구성된 퍼지 규칙이 주어진 데이터를 얼마나 잘 묘사할 수 있는지를 확인하였다. 본 논문에서 제안한 모델링 방법은 그 모델링 알고리즘이 간단함에도 불구하고 모델링 에러가 거의 없는 상당히 근사화된 모델을 제공하는 초기의 퍼지 규칙과 입·출력 변수의 멤버쉽 함수를 생성할 수 있음을 알 수 있다. 또한, 제안한 모델링 알고리즘은 모델의 구조를 파악하는데 적합한 특성을 나타냄을 알 수 있었다. 모델 에러를 줄이기 위해서는 보다 체계적인 정사영 분할법이 필요하다고 생각되며 경사 하강법 등의 미세 조정 과정을 거치면 에러는 상당히 줄 것으로 기대된다.

참고문헌

[1] J. C. Bezdek, "Some Recent Applications of Fuzzy C-means in Pattern Recognition and Image Processing," *IEEE Workshop on Lang. Autom.*, pp. 247-252, 1983

[2] Euntai Kim, Mignon Park, "Simply Identified Sugeno-Type Fuzzy Modeling," *Proc. of IIZUKA'96*, pp. 444-447, 1996

[3] T. Takagi, M. Sugeno, Fuzzy Identification of Systems and its Application to Modeling and Control," *IEEE Trans. Systems, Man, Cybernetics 15(1)*, 116-132, 1985

[4] M. Sugeno, G. T. Kang, "Structure Identification of Fuzzy Model," *Fuzzy Sets and Systems 28*, pp. 15-33, 1988

[5] M. Sugeno, T. Yasukawa, "A Fuzzy-Logic -Based Approach to Qualitative Modeling," *IEEE Trans. on Fuzzy Systems*, Vol.1, No.1, pp. 7-31, 1993

[6] H. Hayashi, H. Nomura, H. Yamasaki, N. Wakami, "Construction of Fuzzy Inference Rules by NDF and NDFL," *Int. J. of Approximate Reasoning*, Vol.6, No.2, pp. 241-266

[7] C. L. Karr, L. M. Freeman, D. L. Meredith, "Improved Fuzzy Process Control of Spacecraft Autonomous Rendezvous Using Genetic Algorithm," *SPIE Conf. on Intelligent Control and Adaptive Systems*, pp. 274-288, 1989

[8] Z. Pawlak, "Why Rough Sets?," *Proc. of FUZZ-IEEE'96*, pp. 738-743, 1996

[9] Z. Pawlak, "Rough Sets," *Int. J. Inform. Comput. Sci. 11*, pp. 341-356, 1982