퍼지 및 지능시스템학회 논문지 2006, Vol. 16, No. 6, pp. 716-720

신경망과 k-means 클러스터링을 이용한 사용자의 퍼지값 선호도 학습 방법

A method for learning users' preference on fuzzy values using neural networks and k-means clustering

윤태복, 나현종, 박두경, 이지형
Tae Bok Yoon, Hyun Jong Na, Doo Kyung Park and Jee Hyong Lee

성균관대학교 컴퓨터공학과
Email: {tbyoon, hanbando, haderme}@skku.edu, jhlee@ece.skku.ac.kr

요 약

퍼지 이론을 이용하면 여러 정보를 통합 요약하기에 수월하여, 웹 상에서 사용자에게 제공할 정보를 가공하는 방법으로 많이 사용되고 있다. 하지만 퍼지의 애매모호한 특성 때문에 사용자에게 맞게 퍼지 집합으로 표현된 같은 정보라 하여도 사용자마다 자신의 퍼지값 선호도에 따라 다른 선택을 할 수 있다. 따라서 애매한 퍼지값을 선택함에 있어 사용자의 퍼지값에 대한 선호도를 반영할 필요가 있다. 그러나 기존의 방법들은 정해진 기준을 획일적으로 적용하여, 사용자의 개인적인 선택 기준을 반영하지 못하는 문제가 있다.

본 논문에서는 사용자의 선호도를 학습하여, 사용자의 선호도에 맞는 정보를 선택하는 방법을 제안한다. 사용자의 선호도를 학습하기 위해서 학습 데이터가 필요한데, 이 데이터는 사용자에게 직접 물어 사용자의 선호도를 얻는데 사용된다. 이 때, 사용자에게 너무 많은 데이터로 질문을 한다면, 사용자에게 부담을 줄 수 있고, 또 너무 적은 데이터를 사용한다면, 학습을 잘 못하는 경향이 생길 수 있다. 이러한 문제에 대처하기 위해서 10개 정도의 데이터를 이용하여 사용자의 선호도를 학습하는 방법을 제안한다. 제안하는 방법은 먼저 두 퍼지값이 서로 겹칠 수 있는 모든 경우의 상대적 위치를 조사한 후 클러스터링을 이용하여 몇 가지 그룹으로 나누고, 나누어진 그룹을 이용하여 학습하였다. 이렇게 학습된 모델은 새로운 애매하게 겹치는 퍼지값에 대해 사용자를 대신해 어느 것을 어느 정도 선호하는지 추론하게 된다.

키워드 : 퍼지값 선호도, 신경망, k-means 클러스터링, 사용자 선호도, 선호도 학습

Abstract

Fuzzy sets are good for abstracting and unifying information using natural language like terms. However, fuzzy sets embody vagueness and users may have different attitude to the vagueness, each user may choose difference one as the best among several fuzzy values.

In this paper, we develop a method learning a user's preference on fuzzy values and select one which fits to his preference. Users' preferences are modeled with artificial neural networks. We gather learning data from users by asking to choose the best from two fuzzy values in several representative cases of comparing two fuzzy sets. In order to establish the representative comparing cases, we enumerate more than 600 cases and cluster them into several groups. Neural networks are trained with the users' answer and the given two fuzzy values in each case. Experiments show that the proposed method produces outputs closer to users' preference than other methods.

Keywords : preference on fuzzy values, artificial neural networks. k-mean clustering, user preference, learning preference

1. 서 론

인터넷에는 정형화되지 않은 많은 양의 정보가 존재한다. 인터넷 정보 서비스를 제공하는 시스템에서는 가공되지 않은 많은 정보를 그대로 보여주기 어려울뿐더러, 사용자 입장에서도 정확히 매칭 되는 정보를 찾기도 어려운 것이 현실이다. 이런 문제점을 이유로, 정보에 대하여 근사추론을 수행하여, 정보를 가공처리 할 수 있는 퍼지이론이 인터넷에 이용되기 시작하였다. 그러나 지금까지 응용 적용된 퍼지 이론은 사용자의 의견이나 선호도를 반영하는 측면에서는 부족한 점이 많이 발견되고 있다. 같은 정보라 하더라도 사용자의 관심과 취향에 따라 그 가치는 다르게 평가 될 것이다.

예를 들면, 서비스를 제공하는 시스템에서 사용자로부터 얻은 정보를 여러 가지 기법으로 처리하여 다음 그림 1과 같이 두 개의 퍼지값 A, B로 표현했다고 하자.

접수일자 : 2006년 11월 20일
완료일자 : 2006년 11월 30일
감사의 글 : 본 연구는 학술진흥재단의 지원을 받았습니다. (KRF-2002-041-D00468)

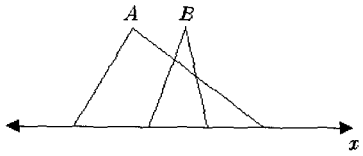


그림 1. 두 퍼지값의 비교

여기서, A, B는 어떤 사용자에게 맞는 정보를 평가한 값이라 하고, x축에서 오른쪽으로 갈수록 큰 값이고 큰 값의 평가를 가진 정보가 사용자에게 조금 더 맞는 정보라 한다면, 과연 시스템에서는 어떤 정보를 제공할 것인가?

이와 같이 두 퍼지값에서 어느 것을 선택할 것인가는 단순한 문제가 아니다. 만약 사용자가 작은 값보다는 큰 값을 원한다면 A를, 작은 값을 피하기를 원한다면 B를 선택하여 제공해 주면 될 것이다. 이와 같이 사용자의 성향을 반영한다면 조금 더 사용자에게 맞는 정보를 제공해 줄 수 있을 것이다.

그러나 지금까지의 퍼지값 비교방법에서는 사전에 정해진 기준에 의해서 나온 값이 모든 사용자에게 일괄적으로 똑같이 제공되어 온 게 현실이다. 그 유사한 방법 및 연구로는 산술연산, 대소비교연산, 퍼지숫자의 정렬 같은 방법들의 의해 두 퍼지값 비교나 여러 퍼지값 중에서 큰 것을 선택하는 방법들이 있다[1-6]. 하지만 이러한 방법들도 사용자의 선호도와는 관계없이 연구자 독자적으로 해석하여, 비교결과를 산출하는 것이 대부분이었다.

본 논문에서는 사용자의 선호도를 반영하여 학습하는 방법을 제안하였다. 제안하는 방법은 두 퍼지값이 겹치는 여러 경우에 대해서 어느 것을 어느 정도 선호하는지 사용자에게 직접 물어보고, 사용자로부터 얻은 데이터를 학습 기능이 있는 인공신경망을 이용하여 사용자 모델을 생성하는 것이다. 이렇게 생성된 모델은 새로운 데이터에 대해 어느 것을 어느 정도 선호하는지 사용자를 대신하여 답변을 한다.

본 논문의 구성은 다음과 같다. 2장에서는 사용자에게 물어볼 데이터를 선정함에 있어 가정 3가지와 클러스터링에 대해 설명하고, 3장에서는 사용자로부터 얻은 데이터를 인공신경망으로 학습하는 방법과 본 연구의 실험을 보이며, 4장에서는 본 논문의 결론을 맺는다.

2. 데이터 선정 조건 및 클러스터링

본 장에서는 인공신경망의 학습데이터 및 사용자에게 물어볼 질의데이터들을 선정하는 데 필요한 가정 3가지와 클러스터링에 대해 설명한다.

2.1 데이터 선정 조건

사용자의 선호도를 학습하기 위해서는 사용자의 선호도를 반영할 질의데이터가 필요하다. 질의데이터를 선정함에 있어서 가장 좋은 방법은 두 퍼지값이 겹칠 때 마다 사용자에게 직접 묻거나, 최대한 많은 경우에 대해 사용자에게 물어, 인공신경망을 학습하는 방법 일 것이다. 하지만 너무 많은 수에 대해 물어 본 다는 것은 사용자에게 부담을 주므로, 사용자의 선호도를 파악하기 위한 최소한의 데이터를 선정하는 것이 중요한 부분이다.

일반적으로 퍼지값들이 나올 수 있는 경우의 수를 살펴보면, 그림 2에서 삼각퍼지값 하나가 각 좌표가 a1, a2, a3라 하고, a1<a2<a3과 같은 조건을 갖는다면, 점 6개로 표현 될 수 있는 경우의 수는 ${}_6C_3=20$이다. 여기서, 두 삼각 퍼지값 A, B가 상대위치만 고려했을 때, 동시에 표현 될 수 있는 개수는 $20 \times 20 = 400$ 이라는 경우의 수가 나온다. 이렇듯 [0.0, 1.0]구간과 점 6개라는 작은 조건에서도 400이라는 개수가 나온 것은 절대위치까지 포함한다면 무수히 많은 수가 나온다는 것을 알 수 있다.

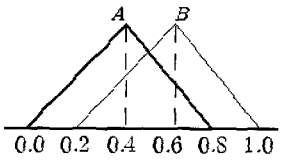


그림 2. 점 6개로 표현한 두 퍼지값

이와 같이 무수히 많은 경우에 대해서 사용자에게 물어본다는 것은 불가능 할 것이다. 그래서 본 논문에서는 다음과 같이 몇 가지 가정을 세워 그 수를 줄였다.

가정1. 모든 퍼지값들은 [0.0, 1.0] 에서만 표현한다. 단, 그 크기를 확대, 축소, 이동하였을 때에도 선호되는 값은 같다라고 가정한다.

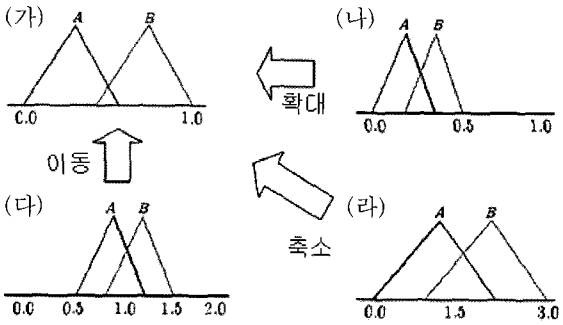


그림 3. 다른 좌표에서 표현된 퍼지값들

퍼지값들은 그 크기와 좌표와 위치에 따라 표현되는 모양이 다르다. 그림 3-(가)의 그림이 그림 3-(나)의 그림처럼 [0.0, 0.5]에서 표현 될 수도 있고, 그림 3-(다)처럼 [0.5, 1.5]에서 표현 될 수도 있다. 이때, 모든 구간에서 표현되는 퍼지값들은 그 크기를 [0.0, 1.0]으로 구간을 조절했을 때에도, 기존 구간에서 선호됐던 값들은 같다라고 가정한다.

가정2. f(A, B)=1−f(B, A) (f(a, b)는 어떤 함수 f에 의해 b에 대한 a의 선호도)

선호도 값을 [0.0, 1.0]에서 표현 했을 때, 그림 2와 같이 두 퍼지값이 겹치는 경우에 대해 퍼지값 B에 대한 퍼지값 A가 선호되는 정도가 0.3정도라 한다면, 상대적으로 A에 대한 B의 선호도는 1−0.3=0.7 이라고 가정한다. 즉, 어느 한 쪽의 선호도를 안다면 다른 쪽의 선호도는 f(A, B)=1−f(B, A)를 통해 얻는 다고 가정한다.

가정3. 두 개의 삼각 퍼지값이 겹치는 경우만을 사용 한다.

퍼지값을 표현하는 모양으로는 사다리꼴 모양, 종형 모양 등 여러 가지가 있지만, 본 논문에서는 이러한 모양들은 제외하고 두 개의 삼각 퍼지값이 겹치는 경우에 대해서만 사용자에게 질문을 하고, 학습을 하여 사용자의 선호도 모델을

생성하였다.

이와 같이 가정 3가지를 통해 [0.0, 1.0]에서 점 6개로 표현되는 경우의 수로 101가지를 생성하였다. 이렇게 생성된 101개의 데이터는 인공신경망의 학습데이터로 사용되는 데, 사용자에게 101개 모두를 물어본 다는 것은 다소 부담이 될 수 도 있을 것이다. 하지만 너무 적은 데이터를 사용한다면, 사용자에게는 부담은 줄겠지만, 인공신경망이 학습을 제대로 못하는 경우가 생길 수 있다. 따라서 본 논문에서는 학습데이터로는 101개 모두를 사용하는 대신, 사용자에게는 101개에 대해 클러스터링을 이용하여 몇 가지 유사한 데이터는 제외하고 특정한 몇 개의 데이터만을 선발하여 사용자의 질의 데이터로 사용하였다.

2.2 질의 데이터 선정을 위한 클러스터링

클러스터링은 주어진 데이터 집합에 유사성을 가지는 클러스터들로 분류하는 것이다. 하나의 클러스터에 속하는 데이터들은 다른 클러스터내의 데이터와 구분되는 유사성을 갖게 된다. 이러한 클러스터링에는 여러 종류가 있으나 본 논문에서는 간단하면서도 우수한 성능을 나타내는 K-means를 사용했다. K-means는 K개의 분할 영역을 결정해 나가는 방법으로 유클리드 거리 측정법에 기반을 두었다.

본 논문에서는 101개 경우들에 대해서 클러스터링의 입력값으로 아래와 같은 함수를 사용하여 10개의 클러스터로 분류를 하였다.

$$f(A,B)=\frac{F(A)}{F(A)+F(B)}$$

단, A와 B는 서로 비교 대상이 되는 퍼지숫자이며 F(A)는 아래와 같이 정의 되는 퍼지 숫자A의 무게 중심이다.

$$F(A)=\frac{\int g(x)\ \mu_A(x)dx}{\int \mu_A(x)dx}$$

이와 같은 방식으로 101가지 경우를 10개의 클러스터로 구분하였고, 각 클러스터에서 대표적인 경우를 선택하여, 이 10개의 경우만 사용자에게 질의하기 위하여 사용하였다. 그림 4는 선택된 경우의 일부분이다.

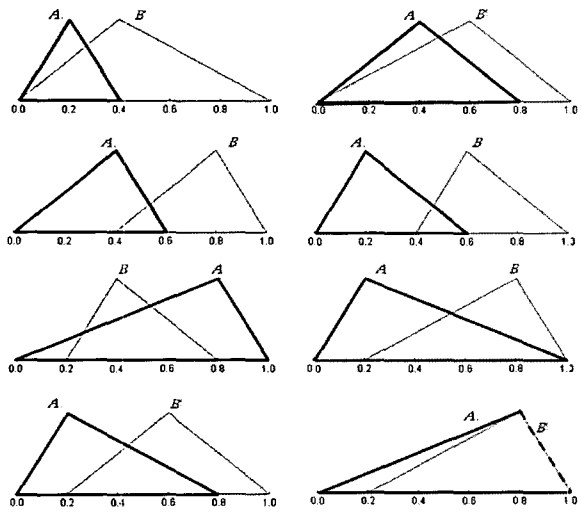


그림 4. k-means(k=10) 클러스터링의 일부

즉, k-mean 클러스터링을 통해 총 101개의 경우에서 질의 데이터는 10개를 추출하였다.

즉, 같은 클러스터 안에 있다는 것은 같은 유사성을 갖고 있다라고 가정하여, 최종적으로 사용자에겐 10개의 질의 데이터만을 보여준다. 그리고 10개의 질의 데이터로 사용자의 선호도를 얻은 후, 자기가 속한 클러스터의 나머지 다른 데이터들에 대해 동일한 선호도 값을 입력한다. 이 과정이 끝나면 최종 101개 데이터는 인공신경망의 학습데이터로 사용되고, 학습이 끝나면 사용자의 선호도 모델을 생성하게 된다.

3. 학습 및 실험

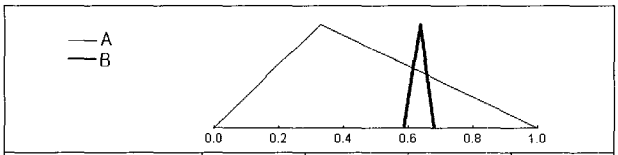
본 장에서는 2장에서 설명한 학습데이터를 이용하여 인공신경망을 학습시키는 과정에 대해 설명한다.

사용자는 2장에서 선택한 10개의 경우에 대해서만 선호도를 답변하면 된다. 이렇게 사용자로부터 얻은 데이터는 101개의 다른 데이터들의 선호도를 유추하게 된다. 본 장에서는 이렇게 얻은 101개의 데이터들을 학습 기능이 있는 인공신경망을 통해 사용자에 대한 선호도를 학습 시킨 후 사용자 선호도 모델을 생성하게 된다. 이렇게 생성된 모델은 새로운 데이터에 대해 어느 것을 어느 정도 선호하는지를 답변하게 된다. 본 논문에서는 사용하는 인공신경망은 여러 가지가 있지만 그 중에서 3-Layer feedforward을 사용하고, 학습 알고리즘으로는 오차역전파를 이용한다[9-11].

인공신경망에서 사용되는 입력 값으로는 두 퍼지값의 각 좌표 6개를 목표 값으로는 사용자의 선호도 값을 사용하였다.

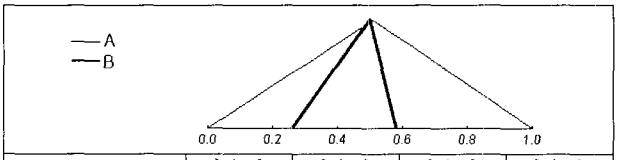
제안하는 비교방법이 기존의 방법과 여러 애매한 경우에 대하여 어떤 결과를 보이는지를 알아보기 위해, 제안하는 방법을 표1~표5와 같이 5가지 경우에 대하여 적용하였다. 그리고 사용자로부터 직접 값을 얻어 비교해 보았다.

표 1. 예제 1



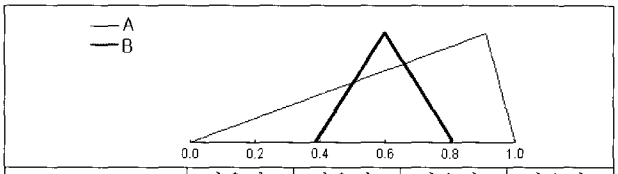
	사용자1	사용자2	사용자3	사용자4
사용자 입력값	0.7	0.65	0.5	0.5
학습결과	0.529	0.514	0.461	0.413
무게중심법	0.41049	0.41049	0.41049	0.41049
Yager의 방법	0.3943	0.3943	0.3943	0.3943
Cheng의 방법	0.43561	0.43561	0.43561	0.43561

표 2. 예제 2



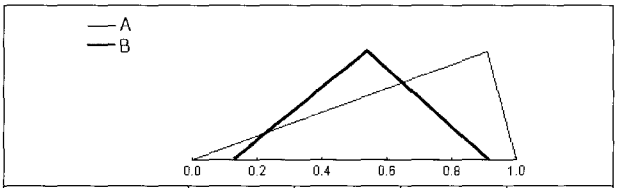
	사용자1	사용자2	사용자3	사용자4
사용자 입력값	0.5	0.7	0.3	0.5
학습결과	0.526	0.542	0.439	0.623
무게중심법	0.52982	0.52982	0.52982	0.52982
Yager의 방법	0.52332	0.52332	0.52332	0.52332
Cheng의 방법	0.51997	0.51997	0.51997	0.51997

표 3. 예제 3



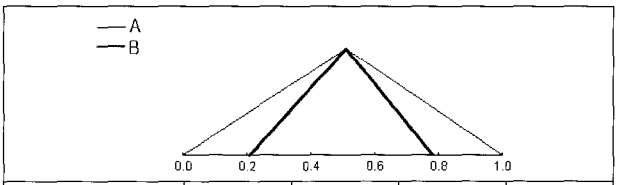
	사용자1	사용자2	사용자3	사용자4
사용자 입력값	0.6	0.75	0.45	0.55
학습결과	0.531	0.514	0.473	0.676
무게중심법	0.51482	0.51482	0.51482	0.51482
Yager의 방법	0.54023	0.54023	0.54023	0.54023
Cheng의 방법	0.51148	0.51148	0.51148	0.51148

표 4. 예제 4



	사용자1	사용자2	사용자3	사용자4
사용자 입력값	0.55	0.7	0.6	0.75
학습결과	0.531	0.514	0.433	0.773
무게중심법	0.54728	0.54728	0.54728	0.54728
Yager의 방법	0.57085	0.57085	0.57085	0.57085
Cheng의 방법	0.53553	0.53553	0.53553	0.53553

표 5. 예제 5



	사용자1	사용자2	사용자3	사용자4
사용자 입력값	0.55	0.6	0.35	0.5
학습결과	0.531	0.487	0.415	0.601
무게중심법	0.50166	0.50166	0.50166	0.50166
Yager의 방법	0.50124	0.50124	0.50124	0.50124
Cheng의 방법	0.50115	0.50115	0.50115	0.50115

표1～표5은 사용자마다 원하는 값이 조금씩 차이가 있는 경우만을 선택한 예제이다. 그리고 기존의 방법으로는 무게중심법과 Yager의 넓이를 이용한 방법, Cheng의 원점을 이용한 방법을 이용하였다.

Yager의 방법[그림 5]은 다음과 같이 정의된다[6,12].

$$F(A)=\frac{R+L}{2}$$

여기서, $R=\int f_A^R(y)dy$, $L=\int f_A^L(y)dy$이고, $f_A^R(y)$와 $f_A^L(y)$는 $\mu_A(x)$의 증가 및 감소하는 함수이다. Cheng의 방법[그림 6]은 $\mu_A(x)$에 대한 무게 중심이, c_A^x, c_A^y일 때, $F(A)=\sqrt{(c_A^x)^2+(c_A^y)^2}$을 통해 A의 function 값을 구할 수 있다[12].

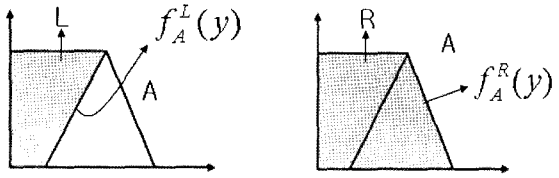


그림 5. Yager의 방법

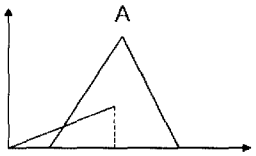


그림 6. Cheng의 방법

표 1～표 5까지를 보면 각 사용자마다 원하는 선호도가 다른 것을 알 수 있다. 그에 비해 기존의 방법을 적용하면 각 사용자에게 똑같은 결과값만을 보여줘야 하는 상황이 생긴다. 예를 들면, 표2의 사용자3의 경우 실제로 원하는 값은 A가 0.3정도로 선호를 한다는 것에 반해 기존의 방법은 동일하게 A가 0.5 이상으로 선호를 한다고 결론이 나온다. 하지만 제안한 방법을 적용했을 때, 각각의 사용자에 맞게 적용값이 틀려지고, 사용자3의 경우 0.439라는 값으로 적용할 수가 있어 개인에 맞게 적용할 수가 있다. 표5의 사용자3의 경우도 이와 마찬가지이다. 하지만, 표4의 사용자3의 경우나, 표5의 사용자2와 같은 경우는 이와 반대의 상황이 생기는 문제점이 발생할 수 있다. 이는 사용자는 A가 실제로 0.6정도로 선호한다는 것에 반해 추론값은 0.43~0.48정도가 나온 것을 알 수 있다. 이와 같은 문제점은 여러 가지 요소가 있을 수 있겠지만, 가장 큰 요인은 사용자가 선호도를 결정할 때 세운 기준에 일관성이 부족해서가 아닐까라는 판단을 한다. 만약 사용자가 동일한 판단 기준을 적용한다면 좋은 결과가 나오리라 생각한다.

4. 결 론

본 논문에서는 두 퍼지값의 비교 방법으로, 10개라는 아주 적은 데이터수를 이용하여 사용자의 선호도를 파악하는 방법을 제안하였다. 제안한 방법은, 조건 3가지를 이용하여, 인공신경망의 학습데이터로 사용될 101개 데이터들을 선정하였고, 이 데이터들은 클러스터링을 통해 다시 10개의 질의 데이터로 추출하였다. 따라서 10개의 데이터로 사용자는 원하는 퍼지값에 어느 정도 선호하는지를 표현하면 된다. 그리고, 사용자로부터 얻은 선호도는 나머지 데이터들의 선호도값을 유추하여 인공신경망을 통해 사용자의 선호도 모델을 생성하게 된다. 이렇게 생성된 모델은 두 퍼지값이 애매한 상황을 사용자에 맞게 추론하여 제공할 것이다. 응용분야로는 특히 사용자와 인터렉티브한 환경을 요구하는 분야에 이 제안이 적용한다면, 사용자 중심적인 정보 환경 구축에 좋은 결과가 예상된다. 향후 연구로는 사용자의 선호도 조사에서 일관성 있게 답변을 유도하는 연구가 필요할 것이다.

참 고 문 헌

[1] K. P. Yoon, "A probabilistic approach to rank complex fuzzy numbers", Fuzzy Sets and Systems, vol.80, pp.167-176, 1996.

[2] K. H. Lee, J. H. Lee, "A method for ranking fuzzy numbers and its application to decision making", IEEE Trans. Fuzzy Systems, vol.7, pp.677-685, 1999.

[3] J. H. Lee, K. H. Lee, "Comparison of fuzzy values on a continuous domain", Fuzzy Sets and Systems, vol.118, pp.419-428, 2001.

[4] 이지형, 이광형, "퍼지 비교 기반 퍼지 숫자의 등급과 방법", 정보과학회논문지, vol.28, n0.12, pp.930-937, 2001.

[5] 이지형, 이광형, "퍼지집합을 이용한 퍼지숫자의 순위 결정 방법", 정보과학회논문지, vol.27, no.7, 2000.

[6] C. H. Cheng, "A new approach for ranking fuzzy nubers by distance method", Fuzzy Sets and Systems, vol.95, pp.307-317, 1998.

[7] M. H. Dunham, Data Mining Introductory and Advanced Topics, Prentice Hall, pp.140-141, 2003.

[8] V. Peneva, I. Popchev, "Comparison of clusters from fuzzy numbers", Fuzzy Sets and Systems, vol.97, pp.75-81, 1998.

[9] J. Dunak, D. Wunsch, "Fuzzy number neural networks", Fuzzy Sets and Systems,vol.108, pp.49-58, 1999.

[10] S. Abe, D. Eng, Pattern Classification : Neuro-fuzzy methods and their comparison, Springer, pp.21-35, 2001

[11] D. W. Patterson, Artificial Neural Networks : theory and applications, Prentice Hall, pp.141-155, 1995.

[12] J. H. Lee, K. H. Lee, "How to determine which is larger: the utility function based methods for ranking fuzzy numbers", International Workshop on Advanced Intelligent Systems, Vol.1, pp.465-468, Taejon, Korea, Aug. 2000.

저 자 소 개

윤태복(TaeBok Yoon)

2001년 : 공주대학교 전자계산학과(학사)
2003년 : (주)디지탈솔루션
2005년 : 성균관대학교 컴퓨터공학(석사)
2005년~현재 : 성균관대학교 컴퓨터공학 박사과정

관심분야 : Game AI, ITS(Intelligent Tutoring System), User Modeling
Phone : +82-31-290-7987
E-mail : tbyoon@skku.edu

나현종(HyunJong Na)

2001년 : 관동대학교 컴퓨터공학(학사)
2004년 : 성균관대학교 컴퓨터공학(석사)

관심분야 : Fuzzy theory, Neural Network
Phone : +82-31-290-7987
E-mail : hanbando@skku.edu

박두경(DooKyung Park)

2005년 : 성균관대학교 컴퓨터공학(학사)
2006년~현재 : 성균관대학교 컴퓨터공학 석사과정

관심분야 : Ubiquitous computing (context-aware system), Intelligent Middleware, Learning
Phone : +82-31-290-7987
E-mail : haderme@skku.edu

이지형(JeeHyong Lee)

1993년 : 한국과학기술원 전산학과 (학사)
1995년 : 한국과학기술원 전산학과 (석사)
1999년 : 한국과학기술원 전산학과 (박사)

2000년 : 미국 SRI International, International Fellow
2002년~현재 : 성균관대학교 정보통신공학부 조교수

관심분야 : 지능시스템, 기계학습, 온톨로지
Phone : +82-31-290-7154
E-mail : jhlee@ece.skku.ac.kr