# 다중 구조적응 자기구성지도의 퍼지결합을 이용한 웹 마이닝

## (Web Mining Using Fuzzy Integration of Multiple Structure Adaptive Self-Organizing Maps)

김 경 중 † 조 성 배 ††
(Kyung-Joong Kim) (Sung-Bae Cho)

**요 약** 폭발적으로 성장하고 있는 웹은 수백만 개의 웹 문서를 포함하고 있기 때문에, 적절한 웹 사이트를 찾기 어렵다. 사용자 프로파일을 사용하여 적절한 웹 사이트를 추천함으로써 웹의 탐색을 개인화할 수도 있지만 웹 컨텐츠에 대한 사용자의 평가는 사용자의 성격에 관한 다양한 측면을 표현하므로 사용자의 선호도를 예측하기 위해서는 보다 효과적인 방법이 필요하다. 사용자 프로파일은 비선형적인 특성을 가지고 있으므로 분류기를 사용하여 예측하여야 하며 다양한 특성을 예측하기 위해 분류기의 결합이 필요하다. 패턴분류와 시각화에 유용한 구조적응 자기구성지도(SASOM)는 개선된 SOM 모델로서 웹 마이닝에 적절하다. 퍼지 적분은 주관적으로 정의된 분류기의 중요도를 이용하여 결합하는 방법이다. 본 논문에서는 독립적으로 학습된 SASOM의 퍼지적분(fuzzy integral)기반 결합을 이용하여 사용자의 프로파일을 예측하고 UCI 벤치마크 데이타인 Syskill & Webert 데이타를 사용하여 그 성능을 평가한다. 실험결과 제안한 방법이 기존의 naive Bayes 분류기뿐만 아니라 SASOM의 투표결합보다 우수한 성능을 보였다.
**키워드** : 사용자 프로파일, 웹 마이닝, 구조적응 자기구성지도, 퍼지 적분, Syskill & Webert

***Abstract*** It is difficult to find an appropriate web site because exponentially growing web contains millions of web documents. Personalization of web search can be realized by recommending proper web sites using user profile but more efficient method is needed for estimating preference because user's evaluation on web contents presents many aspects of his characteristics. As user profile has a property of non-linearity, estimation by classifier is needed and combination of classifiers is necessary to anticipate diverse properties. Structure adaptive self-organizing map (SASOM) that is suitable for pattern classification and visualization is an enhanced model of SOM and might be useful for web mining. Fuzzy integral is a combination method using classifiers' relevance that is defined subjectively. In this paper, estimation of user profile is conducted by using ensemble of SASOM's learned independently based on fuzzy integral and evaluated by Syskill & Webert UCI benchmark data. Experimental results show that the proposed method performs better than previous naive Bayes classifier as well as voting of SASOM's.
**Key words** : User profile, Web Mining, SASOM, Fuzzy Integral, Syskill & Webert

## 1. 서 론

웹은 새로운 정보의 창고이지만 데이타베이스처럼 일관성을 유지하기 위해 중앙에서 관리할 수 없다. 누구나 웹 페이지를 개설할 수 있고 링크를 만들 수 있기 때문에, 웹은 매우 많은 수의 문서를 포함하고 있다. 사용자가 모든 정보를 검토할 수 없기 때문에 중요한 웹 문서만을 가려내는 작업이 필요하다. 본 논문에서는 HTML 문서와 사용자의 선호 기록으로부터 사용자 프로파일을 생성하기 위한 웹 컨텐츠 마이닝을 다룬다. 사용자 프로파일을 예측하는 것은 쉽게 한번에 추측되지 않는 특성을 지니고 있고 비선형적인 함수를 필요로 한다. 이러한 특성을 단일 분류기를 사용하여 예측하는 것은 어려운 일이며 서로 다른 전문성을 지니고 있는 상호보완적인 여러 개의 분류기를 결합하는 것이 필요하다.

· 이 연구는 과학기술부가 지원한 뇌과학 연구 프로그램에 의해 지원되었음
† 학생회원 : 연세대학교 컴퓨터학과
uribyul@candy.yonsei.ac.kr
†† 종신회원 : 연세대학교 컴퓨터학과
sbcho@cs.yonsei.ac.kr
논문접수 : 2003년 5월 19일
심사완료 : 2003년 9월 18일

자기구성지도(Self Organizing Map)는 지식 추출을 위해 고차원 데이타를 시각화하고 데이타를 클러스터링 하는데 유용한 신경망이다[1,2]. 몇몇 연구자들은 자기구성지도(SOM)을 패턴 분류에 적용하려고 시도해 왔다[3,4]. 신경망의 다른 모델처럼 SOM의 구조와 크기를 결정하는 것은 매우 어렵다. 이전 연구에서 하나 이상의 클래스를 표현하는 노드를 네 개의 노드로 구성되어 있는 부분노드로 분할하는 SOM을 위한 동적 노드분할 개념에 기초한 효과적인 패턴인식기를 제안했다[5]. 서로 다른 특징으로 독립적으로 학습한 구조적응 자기구성지도(SASOM)의 결합은 숫자 인식 문제에 높은 성능을 보였다[6].

SASOM은 효과적인 패턴 인식기로 사용될 수 있으며 또한 사용자 입력 벡터의 이차원 투영을 이해하기 위한 지도구조를 시각화할 수 있다. 본 논문에서, SASOM의 앙상블이 사용자 프로파일을 예측하기 위해 적용되었으며 각 SASOM은 서로 다른 특징을 사용하여 독립적으로 학습되었다. 정보이득, TFIDF, odds ratio의 세 가지 다른 특징추출 방법이 이 문제를 위해 사용되었다. 정보이득은 C4.5에서 결정 트리를 생성하기 위해 사용한 매우 효과적인 특징추출 방법이다[7]. TFIDF는 텍스트 추출에서 자주 사용되는 일반적인 방법이다. Odds ratio는 특징의 순위를 매기는 매우 간단한 방법이다[8]. 이러한 세 가지 방법은 텍스트를 위한 대표적인 특징 추출 방법이며 쉽게 구현이 가능하다.

투표, Bayesian 결합, BKS 방법, Borda 함수, Condorect 함수, 평균, 가중평균 등 많은 결합방법이 존재한다[9]. 그러나 이러한 방법들은 사용자의 주관적인 분류기에 대한 선호도를 결합과정에 포함시키지 못하며 융통성이 미흡하다. 퍼지적분은 퍼지 척도와 사용자의 주관적인 분류기에 대한 평가를 이용하여 다양한 자원으로부터의 신뢰도를 통합하는 결합방법이다[10]. 본 논문에서는 “좋아함” 또는 “싫어함”으로 표기되어 있는 HTML 문서로부터 사용자의 프로파일을 예측하기 위해 SASOM의 퍼지적분에 기초한 결합방법을 제안한다.

그림 1은 제안하는 방법의 개요를 보여준다. 테스트와 학습을 위해 사용되는 웹 텍스트 데이타는 입력 데이타로 사용되기 위해 전처리를 거친다. 각 특징 추출방법은 학습을 위한 중요한 특징 집합을 추출하고, 각 특징 집합은 하나의 SASOM을 학습하기 위해 사용된다. 학습 이후에 각 SASOM은 그림과 같이 서로 다른 위상을 지니며 퍼지 적분은 마지막에 여러 개의 분류기로부터 나온 신뢰도를 통합한다. 이 앙상블 분류기는 사용자 프로파일로서 알려지지 않은 웹 문서에 대한 사용자의 선호도를 예측하도록 사용될 수 있다. 제안하는 방법을 평가하기 위해 UCI KDD Syskill & Webert 데이타를 사용하였다[11]. 이 데이타는 네 개의 서로 다른 주제에 대해 관련된 웹 문서와 사용자의 선호도를 “hot”, “medium”, 또는 “cold” 중의 하나로 기록한 정보를 포함하고 있다. 알려지지 않은 웹 문서에 대해 사용자의 선호도를 “hot” 또는 “cold”로 예측하는 것이 본 논문에서 해결하려는 문제이다 (“medium”이 매우 적기 때문에 “medium”과 “cold”는 하나로 합쳐졌다). Pazzani는 naive Bayes 분류기가 위의 데이타에 대해 신경망, ID3

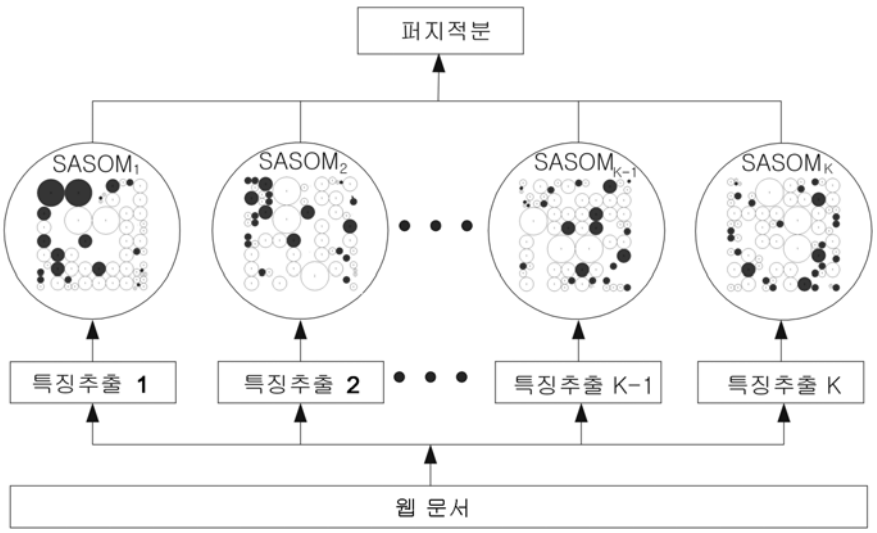



그림 1 제안하는 방법의 개요

등의 방법보다 우수한 결과를 냈다고 보고했다[12]. 본 논문에서는 제안하는 방법을 Pazzani의 실험결과와 다중 SASOM의 투표결합과 비교한다. 실험 결과는 퍼지 적분을 사용한 제안하는 방법이 naive Bayes 분류기를 포함한 기존의 연구결과와 SASOM의 투표결합보다 우수한 성능을 내는 것을 보여준다.

## 2. 관련연구

퍼지 적분은 분류기를 결합하는데 이용될 수 있으며 컴퓨터 시각, 얼굴 인식, 필기체 인식 등의 분야에서 응용되고 있다. Mirhosseini는 퍼지 적분을 사용하여 인식 작업에서 각 얼굴 요소의 중요도를 포함하도록 하는 새로운 얼굴 인식 시스템을 제안했다[13]. Cho는 온라인 필기 문자의 인식문제에 퍼지적분에 기초한 다중 네트워크의 결합이 유용하다는 것을 보였다[14]. Pham은 적응적인 퍼지적분을 사용하여 필기 숫자를 인식하는 분류기를 결합했다[15]. Kumar는 퍼지적분을 사용하여 다분광 데이타의 분류를 위한 신경망을 결합하였다[16]. 하지만 현재까지 다중 SASOM을 퍼지적분으로 결합한 사례는 없다.

데이타 마이닝의 가장 어려운 문제 중의 하나는 데이타베이스로부터 문서를 추출하고 재구성하는 것이다[17]. Kohonen은 500차원의 특징 벡터로 표현된 700만개의 특허 요약을 분류하기 위한 100만개이상의 노드를 지닌 자기구성지도의 유용성을 보였다[18]. 효과적인 패턴 분류기로서 기본 SOM의 변형이 많이 존재하는데, Cho는 가중치뿐만 아니라 구조도 적응할 수 있는 SA-SOM을 개발하였고, Bauer는 growing self-organizing map(GSOM) 알고리즘을 제안하였다. GSOM은 학습동안 적응되는 일반적인 하이퍼큐브형태의 모양을 하고 있다[19]. Suganthan은 계층적인 중첩 자기구성지도(HOSOM)을 제안하였는데, 이는 1차 수준의 SOM으로 구성되고 몇몇의 부분적으로 중첩된 이차 수준의 SOM들이 보완된다[20]. 계층적인 SOM이나 growing SOM은 패턴분류를 위해 개선되었지만 SASOM처럼 노드들이 서로 연속된 평면에 존재하는 것이 아니기 때문에 이웃하는 노드 사이에 불연속성이 발생할 수 있으며 위상보존에 문제가 있다.

Syskill & Webert 시스템은 관심 없는 웹 문서들로부터 흥미로운 웹 문서를 구분한다[12]. 이 시스템은 naive Bayes 분류기를 사용자 프로파일 예측을 위해 사용하였다. 각 사용자는 시작 페이지로부터 출발하며 웹 문서에 대해 아이콘을 사용하여 “좋아함” 또는 “싫어함”을 표시하고 표기한 웹 문서의 개수가 10개일 때까지 계속한다. 10개의 표기된 웹 문서가 모이면 Bayes 분류기 학습 알고리즘이 실행된다. 기존 실험결과에서 Bayes 분류기는 작은 학습 데이타에 대해서도 잘 작동하였다. 이 시스템은 웹 문서를 분석하고 각 사용자별로 학습한 Bayes 분류기의 선호도 값에 의해 링크를 추천한다.

Mladenic은 정보 이득, cross entropy, 상호 이득, weight of evidence, odds ratio 등의 대규모 텍스트 데이타에서의 특징 집합 추출에 적합한 새로운 특징 평가 방법을 제안했다[8]. Lewis는 Reuters와 MUC-3 텍스트 분류 데이타에 대한 분류 소속도 예측 문제에서 특징의 종류와 개수를 변화하는 것이 미치는 영향을 조사했다. 이 실험에서 단어에 기반한 인덱싱을 위한 최적의 특징 집합 크기는 큰 학습 데이타에도 불구하고 매우 낮은 것으로 알려졌다 (10개에서 15개의 특징)[21]. Pal은 퍼지 논리, 신경망, 유전자 알고리즘 등을 이용하여 구조를 보존하는 차원 축소와 함께 특징 평가, 선택, 추출하는 방법을 제안했다[22].

## 3. SASOM의 퍼지결합

본 논문에서는 기계학습 기법의 앙상블을 사용한 사용자 프로파일의 예측을 위한 일반적인 프레임워크를 제안한다. 분류기의 결합은 각 분류기가 상호 독립적이면 보다 높은 성능을 낼 수 있다. 이러한 목적을 위해 보통 각 분류기를 서로 다른 특징을 사용하여 독립적으로 학습하는 방법을 사용한다[6,20]. 서로 다른 특징 추출방법을 사용하여 각 SASOM은 독립적으로 학습되며 퍼지적분을 사용하여 결합된다. 각 분류기의 중요도는 결합단계에서 주관적으로 결정된다.

### 3.1 특징 추출

특징 선택은 단어의 빈도나 의존관계 등의 정보에 기초하여 특징을 순위 매기는 과정이다. 텍스트 분류에서 특징은 텍스트의 단어이며 이진 값을 가진다(“텍스트에 있다”와 “텍스트에 없다”). 20개의 웹 문서 집합에는 5000개 이상의 특징이 존재하기 때문에 특징 선택과정이 필요하다. 많은 특징은 성능을 향상시키는 데 유용하지 않으며 분류기의 학습도 어렵게 만든다.

본 논문에서는 서로 다른 특성을 지니는 세 가지 특징 선택 방법을 사용한다. TFIDF는 단어의 빈도와 역문서 빈도를 곱한 것이다. 이 기준은 텍스트 추출에서 자주 사용되며 매우 단순하다. 예를 들어, 20개의 웹 문서가 있고 그 중에서 10개의 문서가 “안녕”이라는 단어를 포함하고 있으면 (DF=0.5)이고 1000개의 단어가 20개의 문서에 있고 “안녕”이라는 단어가 120번 나왔다면 (TF=0.12)이다. TFIDF는 이 두 가지를 다음의 수식으로 계산한 것이다(TFIDF=0.036).

$$TFIDF = TF \times \log \frac{1}{DF} \quad (1)$$

TFIDF는 특징의 중요도를 계산하기 위해 학습 데이타의 클래스 정보를 사용하지 않기 때문에 분류의 성능이 낮을 수 있다. 정보이득은 정보이론에 기초한 방법인데, $S$는 웹 텍스트의 집합이며 $E$는 기대 정보이득일 때, $E(W,S)$는 단어 $W$의 문서집합 $S$에 대한 기대값을 나타낸다.

$$E(W,S) = I(S) - P(W = present)\, I(S_{w=present}) + P(W = absent)\, I(S_{w=absent}) \quad (2)$$

$$I(S) = \sum_{c \in \{hot, cold\}} -p(S_c)\log_2(p(S_c))$$

마지막 특징 추출방법은 여러 개의 클래스 중에서 하나를 위해 좋은 성능을 내도록 하는 것을 목적으로 사용되어진 odds ratio이다[8].

$$OddsRatio\ (W) = \log \frac{odds\ (W = present\ |\ C_1)}{odds\ (W = present\ |\ C_2)} \quad (3)$$

$C_1$과 $C_2$는 이진분류 문제의 클래스 레이블이다. $odds(X_i)$는 다음과 같이 정의된다. $X_i$는 확률변수이고 $n$은 데이타의 개수이다.

$$odds(X_i) = \begin{cases} \dfrac{\frac{1}{n^2}}{1-\frac{1}{n^2}} & P(X_i) = 0 \\ \dfrac{1-\frac{1}{n^2}}{\frac{1}{n^2}} & P(X_i) = 1 \\ \dfrac{P(X_i)}{1-P(X_i)} & P(X_i) \neq 0 \wedge P(X_i) \neq 1 \end{cases} \quad (4)$$

이러한 세 가지 특징 선택 방법은 서로 다른 특성을 지니고 있다. TFIDF는 특징의 중요도를 계산할 때 문서의 클래스 레이블을 고려하지 않는다. 반면, 정보이득은 문서의 클래스 정보를 사용한다. Odds ratio는 클래스 레이블을 사용하지만 하나의 특정 클래스에 유용한 특징만을 선택한다.

### 3.2 SASOM

SOM은 지도의 위상을 보존하는 특성을 가지고 있는 신경망 모델이고 고차원의 데이타를 저차원의 공간에 시각화하는데 자주 이용된다. 기본 SOM은 지도의 구조를 고정하여 각 노드가 하나 이상의 다른 클래스 레이블을 가진 데이타를 포함하기 때문에 분류에서 낮은 성능을 보인다. 이러한 특성은 비교사 학습에는 매우 유용하지만 분류에는 약점이 된다. 만약 노드가 다른 클래스 레이블을 가지고 있으면, SASOM은 노드를 4개의 노드로 이루어진 부분지도로 분할한다. 동적인 노드 분할 개념은 동시에 SOM의 적절한 노드 개수와 입력과 출력 노드 사이의 연결 가중치를 결정한다.

SASOM의 학습을 위한 기본과정은 다음에 나와 있으며 SOM과 유사하다.

1. 기본 SOM으로부터 시작한다(각 노드가 모든 입력 노드와 완전 연결되어 있는 4×4 맵)
2. 현재 네트워크를 Kohonen의 알고리즘으로 학습한다[2].
3. 다음 사항을 결정하기 위해 알려진 입출력 패턴을 사용하여 네트워크를 조정한다.
   (a) 어떠한 노드가 여러 개의 노드를 지닌 부분 맵(2×2 맵)으로 교체될 것인가?
   (b) 어떠한 노드가 삭제될 것인가?
4. 모든 노드가 하나의 유일한 클래스만을 표현하면 종료하고 아니면 2번으로 간다.

기본 SOM의 학습 알고리즘은 다음과 같다.

$$\| x - w_c \| = \min_i \{\| x - w_i \|\} \quad (5)$$

네트워크의 초기지도는 4×4 노드로 구성되어져 있고, 노드 $i$의 가중치 벡터는 $w_i \in R^n$이다. $x$와 $w_i$의 유클리디안 거리는 승리자 노드라고 불리우는 wc에서 최소가 된다. 승리자 노드와 거리가 $N_c$보다 작은 모든 노드는 다음 규칙에 따라 가중치를 갱신한다 ($\alpha(t)$는 학습률이며 $0<\alpha(t)<1$).

$$w_i(t+1) = \begin{cases} w_i(t) + \alpha(t)[x(t) - w_i(t)] & \text{if } i \in N_c(t) \\ w_i(t) & \text{if } i \notin N_c(t) \end{cases} \quad (6)$$

하나 이상의 클래스를 표현하는 노드는 2×2의 부분지도로 교체된다. 자식 노드의 가중치는 부모노드와 이웃노드의 가중치를 기초로 다음의 수식으로 결정된다.

$$C = \frac{(P \times 2) + \sum N_c}{S} \quad (7)$$

$N_c$ : 자식의 이웃노드

$S$ : $N_c$ + 2

그림 2는 노드 분할의 예를 보여준다. 이 경우에, $C_0$의 가중치는 다음과 같이 결정된다.

$$C_0 = \frac{(P_4 \times 2) + P_0 + P_1}{4} \quad (8)$$

그림 3은 다른 특징 집합을 사용하여 독립적인 분류기를 학습하는 전체과정을 보여준다. 특징 집합은 웹 문서의 집합에 속한 단어들의 집합이며, 특징선택 방법을 사용하여 각각의 분류기를 위한 가장 중요한 특징을 선택할 수 있다. 예를 들어, 첫 번째 특징 부분집합은 “Initial” “Distance” “Node”와 “Several”을 가장 중요한 특징으로 선택했다. 이 특징을 사용하여 입력 벡터가 생성된다. 예를 들어, 첫 번째 웹 문서는 첫 번째 특징

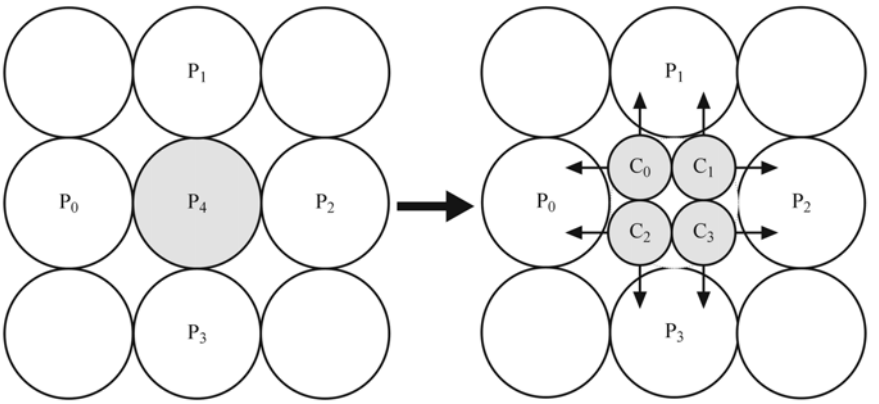



그림 2 노드 분할의 예

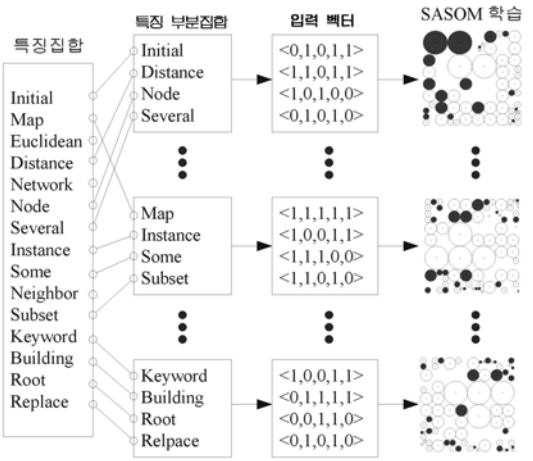



그림 3 서로 다른 특징 집합을 사용한 SASOM의 학습

집합을 사용하여 <0,1,0,1,1>로 표현되며 “Initial”은 존재하지 않고, “Distance”는 존재하며, “Node”는 존재하지 않고 “Several”은 존재하며 이 문서의 클래스 레이블은 “1”이다. 이러한 입력 벡터를 사용하여 SASOM을 학습한다.

### 3.3 퍼지 적분

다중 분류기의 결과를 결합하는 방법은 많이 있으며, 각 분류기는 문제에 대해 동일한 중요도를 지니고 있다고 가정한다. 가중 평균은 각 분류기의 중요도를 학습 데이타에 대한 분류 성능을 사용하여 객관적으로 결정하도록 한다. 하지만 퍼지적분은 주관적으로 평가된 각 분류기의 중요도를 이용하며 최종 결정을 사용자에 의해 정의된 분류기의 중요도와 각 클래스에 대한 분류기의 신뢰도를 통합하여 내린다.

퍼지 적분은 Sugeno[23]에 의해 소개되었으며 관련된 퍼지 척도[24,25]는 정보를 통합하기 위한 유용한 방법을 제공한다. 퍼지 척도는 다음과 같이 정의된다.

**정의 1** : $X$는 원소들의 유한집합이다. 함수 $g:2^X \rightarrow [0,1]$을 퍼지 척도라고 한다.

1) $g(\varnothing)=0$
2) $g(X)=1$
3) $g(A) \le g(B)$ if $A \subset B$

퍼지 척도는 0과 1 사이의 실수 값을 $X$의 각 부분집합에 할당한다. 퍼지 척도 $g$의 정의로부터 Sugeno는 다음과 같은 부가적인 속성을 만족하는 $g_\lambda$-퍼지 척도를 제안했다. 모든 $A,B \subset X$, $A \cap B=\varnothing$에 대하여,

$$g(AB)=g(A)+g(B)+\lambda g(A)g(B),\ \lambda>-1. \quad (9)$$

두 개의 중복되지 않은 부분집합의 합집합에 대한 척도 값은 각 구성요소의 척도 값으로부터 직접 계산할 수 있다. Sugeno는 퍼지 척도, 특히 $g_\lambda$-퍼지 척도의 관점에서 정의된 비선형적인 함수인 퍼지적분의 개념을 제안했다.

**정의 2** : $X$는 유한 집합이고, $h : X \rightarrow [0,1]$는 $X$의 퍼지 부분집합일 때, 퍼지 척도 $g$와 $X$에 대한 함수 $h$의 퍼지적분은 다음과 같이 정의된다.

$$h(x) \circ g(\cdot) = \max_{E \subseteq X} \left[ \min \left( \min_{x \in E} h(x),\ g(E) \right) \right] \quad (10)$$

다음 퍼지적분의 속성은 쉽게 증명될 수 있다[20].

1) 모든 $x \in X$에 대해 만약 $h(x)=c$, $0 \le c \le 1$이면 $h(x) \circ g(\cdot)=c$.
2) 모든 $x \in X$에 대해 만약 $h_1(x) \le h_2(x)$이면 $h_1(x) \circ g(\cdot) \le h_2(x) \circ g(\cdot)$
3) 만약 $\{A_i | i=1, \cdots, n\}$이 집합 $X$의 분할이라면

$$h(x) \circ g(\cdot) \geq \max_{i=1}^{n} e_i, \quad (11)$$

$e_i$는 $A_i$의 $g$에 대한 $h$의 퍼지적분이다.

퍼지적분의 계산은 다음과 같다. $Y=\{y_1, y_2, \cdots, y_n\}$은 유한집합이며 $h : Y \rightarrow [0,1]$은 함수이다. $h(y_1) \geq h(y_2) \geq h(y_3) \geq \cdots \geq h(y_n)$이라고 가정하면 $Y$에 관한 퍼지 척도 $g$에 대한 퍼지적분 $e$는 다음과 같이 계산된다.

$$e = \max_{i=1}^{n} [\min(h(y_i), g(A_i))] \quad (12)$$

$A_i=\{y_1, y_2, \cdots, y_i\}$이다. $\lambda$는 다음 수식으로 계산된다.

$$\lambda + 1 = \prod_{i=1}^{n} (1 + \lambda g^i) \qquad \lambda \in (-1, +\infty) \text{ and } \lambda \neq 0. \quad (13)$$

이 수식은 다음 재귀 계산으로부터 도출되었는데, $\lambda$는 $(n-1)$차 다항식으로부터 쉽게 계산될 수 있다.

$$g(A_1) = g(\{y_1\}) = g^1$$
$$g(A_i) = g^i + g(A_{i-1}) + \lambda g^i g(A_{i-1}), \quad \text{for } 1 < i \leq n. \quad (14)$$

$C=\{c_1, c_2, c_3, \cdots, c_N\}$는 클래스의 집합이며 이진 분류 문제의 경우 $|C|=2$이다. $Y=\{y_1, y_2, \cdots, y_n\}$는 분류기의 집합이고, $h_k : Y \rightarrow [0,1]$는 객체 $A$(분류의 대상)의 클래스 $c_k$에 대한 부분 평가이다. $h_k(y_i)$는 객체 $A$를 클래스 $c_k$라고 분류기 $y_i$가 확신하는 정도이다. $Y$ 집합은 각 클래스의 $h_k(y_i)$ 값에 따라 내림차순으로 정렬되어 있다. $A_{ki}$는 클래스 $c_k$를 위한 $Y$의 최초 $i$개 원소들의 집합이다.

$$\text{Final class} = \underset{c_k \in C}{\text{argmax}} \left[ \max_{i=1}^{n} [\min(h_k(y_i), g(A_{ki}))] \right] \quad (15)$$

각 SASOM은 알려지지 않은 문서의 클래스 레이블을 "0" 또는 "1"로 결정한다(이진 분류 문제). 만약 $SASOM_1$이 문서를 "0"으로 분류했다면 $h_0(SASOM_1)=1.0$이고 $h_1(SASOM_1)=0.0$이다. 만약 세 개의 SASOM 분류기가 있고 사용자가 각 분류기를 $g^1$, $g^2$, $g^3$로 평가했다면 $\lambda$는 $g^1$, $g^2$, $g^3$로부터 계산된다. 수식 (13)을 사용하면 2차원 방정식으로 쉽게 계산된다. 각 클래스 ck에 대해 분류기는 $h_k(SASOM_i)$값에 따라 정렬된다. 정렬된 순서대로 분류기는 $y_1$, $y_2$, $y_3$로 이름지어진다. $g(y_1)$, $g(y_1,y_2)$, $g(y_1,y_2,y_3)$를 사용하여 알려지지 않은 문서의 클래스를 수식 (15)를 사용하여 결정한다.

## 4. 실험 결과

제안한 앙상블은 사용자 프로파일을 만들기 위해 웹 문서 데이타와 사용자 선호도 데이타를 이용하여 학습한다. UCI KDD 데이타베이스로부터 웹 문서와 사용자의 선호도 값("hot" 또는 "cold")을 가지고 있는 Syskill & Webert 데이타를 구할 수 있다. Syskill & Webert 데이타는 "Bands" "Biomedical" "Goats" "Sheep"의 네 가지 주제를 가지고 있으며 본 논문에서는 "Bands"와 "Goats" 데이타를 사용했다.

"Bands" 데이타는 61개의 HTML 문서를 "Goats" 데이타는 70개의 HTML 문서를 각각 가지고 있다. 각 문서는 "hot" 또는 "cold"의 클래스를 가진다. 그림 4는 HTML 파일과 인덱스 데이타를 보여준다. 각 HTML 파일은 주제와 관련된 텍스트를 포함하고 있다. 인덱스 파일은 이름, 평가, URL, 평가날짜, 제목을 순서대로 포함하고 있다. 웹 문서의 전처리는 특징을 선택하고 클래스 정보를 이용하여 입력 벡터를 생성하는 과정이다. 학습 데이타로부터 $k$개의 중요한 특징을 각각 세 가지 다른 특징 추출 방법을 사용하여 뽑는다. 각 방법은 특징을 다른 방식으로 순위를 매긴다. 그림 5는 각 방법에 대한 특징들의 다른 순위를 보여준다. Bands 데이타에 대해 10개의 학습 데이타를 사용하여 1200개의 단어를 모았다. 이 그림에서 단어의 순위는 각 방법마다 다르다. 문서 $D=<v_1, v_2, v_3, \cdots, v_{128}, c>$는 SASOM을 학습시킬 수 있는 서로 다른 세 개의 입력 벡터로 표현된다. HTML 문서의 전처리 과정은 다음과 같다.

1. Non-letter를 제거한다.
2. 대문자를 소문자로 바꾼다. Stop-list를 아래의 조건을 만족하는 특징으로 선택
   (a) 단어를 빈도순으로 정렬한다.
   (b) 가장 높은 순위로 결정된 600개의 단어를 stop-list로 결정
3. Stop-list를 제거한다.
4. 인덱스를 생성한다. <특징, 특징을 포함하는 문서의 리스트>
5. TFIDF, 정보이득, Odds ratio 등을 사용하여 특징의 중요도를 계산한다.
6. TFIDF를 사용하여 특징을 정렬하고 $k$개의 특징을 선택한다. 정보이득과 Odds ratio 방법에 대해서도 $k$개의 특징을 선택한다. $k$값은 128로 선택한다[12].
7. 학습 데이타와 테스트 데이타에 대해 입력 벡터를 생성한다.

그림 5는 서로 다르게 작동하는 세 가지 특징추출 방법에 대해 보여준다. 그림에서 보듯이 단어의 순위는 세 가지 경우에 대해 매우 다르게 나타난다. 이 그림에서 막대의 높이는 단어의 순위를 나타낸다. 만약 단어가 중요하면 순위가 작은 수일 것이므로 막대의 높이는 낮아질 것이다.

해결하려는 문제는 생성한 입력 벡터를 사용하여 학습한 서로 다른 세 개의 분류기의 앙상블을 사용하여 알려지지 않은 문서의 클래스를 예측하는 것이다. 각 주제에 대해 8번의 다른 실험을 했다(각 실험은 학습 데이타와

```
<A NAME="EL_SOB"></A>
<TITLE>EL SOB</TITLE>
<CENTER>
<H1>
EL SOB
</H1>
<A HREF="/IUMA-2.0/ftp/volume2/EL_SOB/EL_SOB.jpg">
<IMG WIDTH=101 HEIGHT=124 BORDER=2 SRC="/IUMA-2.0/ftp/volume2/EL_SOB/sm-EL_SOB.gif"></A>
<P><BR></P>
</CENTER>
<CENTER>
<I><FONT SIZE=5>Skin a Cat</FONT></I><BR>
```

Excerpt of HTML text (File name "1")

```
1|cold|http://www.iuma.com/IUMA-2.0/ftp/volume2/EL_SOB/|Fri Oct 13 15:21:56 PDT 1995|EL SOB
2|hot|http://www.iuma.com/IUMA-2.0/ftp/volume3/Lead_Pipe_Cinch/|Tue Oct 17 09:01:56 PDT 1995|Lead Pipe Cinch
3|hot|http://www.iuma.com/IUMA-2.0/ftp/volume2/Porter,_JL/|Tue Oct 17 09:05:01 PDT 1995|Porter, JL
4|cold|http://www.iuma.com/IUMA-2.0/ftp/volume3/Dr._Octojoculus/|Tue Oct 17 09:11:23 PDT 1995|Dr. Octojoculus
5|cold|http://www.iuma.com/IUMA-2.0/ftp/volume7/Adam_Bomb/|Tue Oct 17 09:12:24 PDT 1995|Adam Bomb
6|cold|http://www.iuma.com/IUMA-2.0/ftp/volume1/Russlee/|Tue Oct 17 09:15:45 PDT 1995|Russlee
```

Syskill & Webert ratings

그림 4 UCI Syskill & Webert 데이타

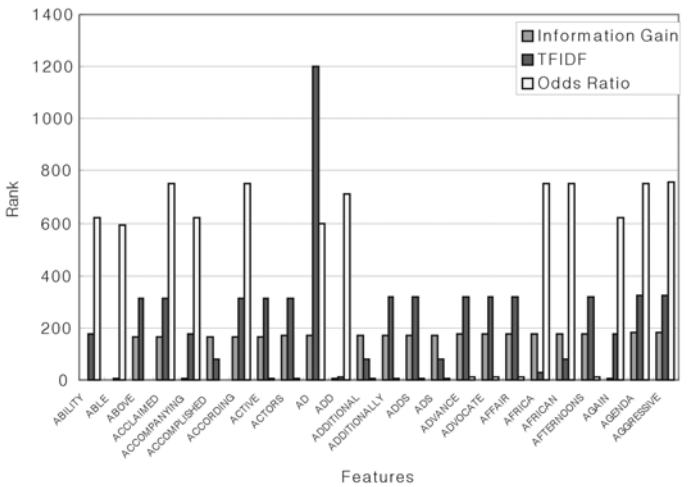
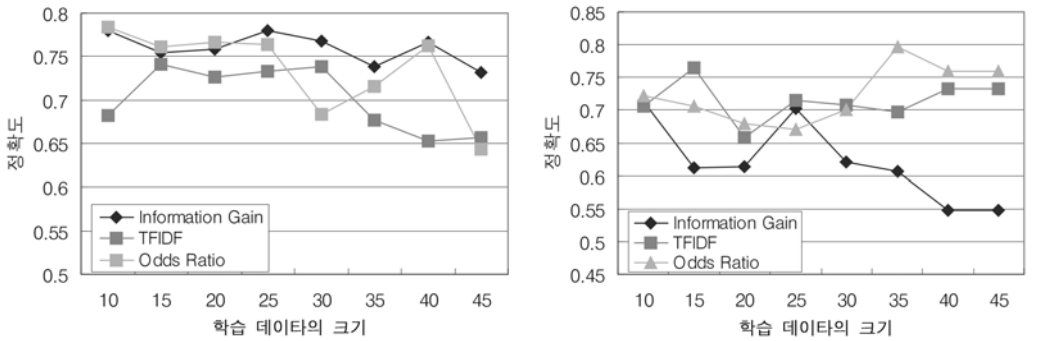



그림 5 각 특징 추출 방법에 대한 특징의 순위

테스트 데이타를 다르게 했다). 학습 데이타는 10부터 45까지 5씩 증가했으며 나머지는 테스트 데이타이다. 실험은 8 가지의 다른 조건에 대해 10번 반복했으며 결과는 그들의 평균이다. 비교를 위해 Pazzani의 naive Bayes 분류기, $k$ 근접 이웃 분류, ID3, perceptron, 오류 역전파, PEBLS, Rocchio 등이 사용되었다[12].

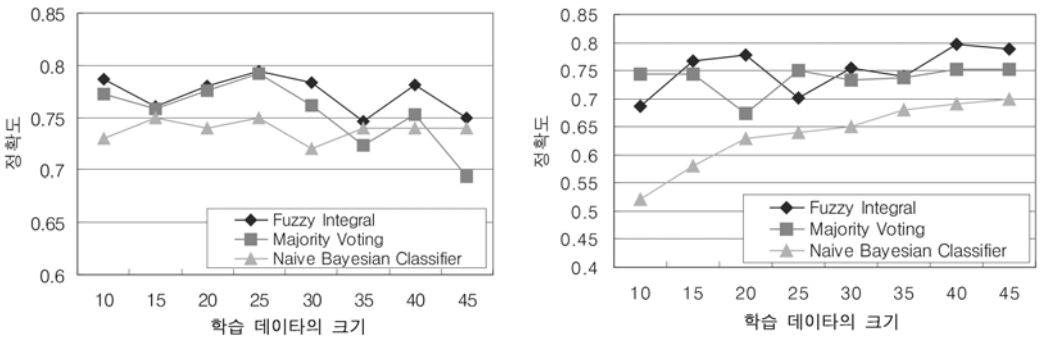
그림 6은 앞의 세 가지 특징 집합을 사용하여 학습된 각 분류기의 성능을 보여준다. 각 분류기는 다른 성능을 보여준다. Bands 데이타에서는 정보 이득이 가장 뛰어난 성능을 보였고 Goats 데이타에서는 odds ratio와 TFIDF가 좋은 성능을 보였다. 정확도는 테스트 데이타에 대한 올바른 예측의 비율을 의미한다.

주관적인 할당은 제안하는 방법의 강점이며 검증 데이타의 성능에 기초한 몇몇 전략들이 이전에 제안되어 졌지만[10] 이번에는 검증 데이타의 부족이라는 문제로 사용하지 못했다. 본 논문에서는 학습 데이타의 성능을 바탕으로 주관적인 평가 값을 할당했다. 그림 7은 퍼지 적분을 사용한 SASOM 앙상블의 성능과 투표를 사용




(a) Bands (b) Goats

그림 6 다른 특징 집합을 사용하여 학습한 단일 SASOM 분류기의 성능




(a) Bands (b) Goats

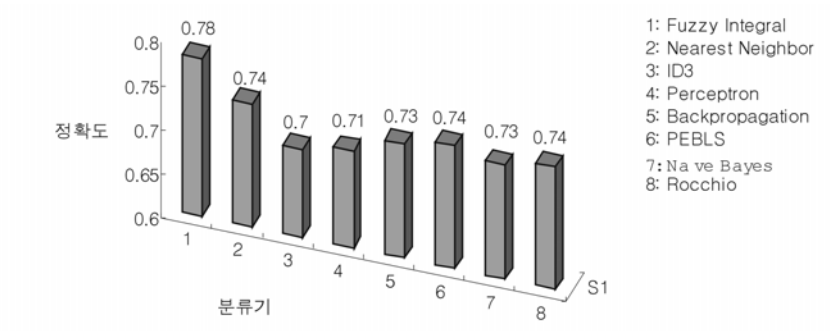
그림 7 퍼지적분을 사용한 SASOM 분류기의 결합

한 결합의 성능을 보여준다. 비교를 위해 Pazzani의 naive Bayes 분류기가 사용된다. Bands 데이타에서 퍼지적분은 naive Bayes 분류기와 투표 결합보다 우수한 성능을 보였다. Goats 데이타에서 퍼지적분은 다른 두 가지 방법보다 우수한 성능을 보였다. SASOM의 투표 결합은 naive Bayes 분류기 보다 우수한 성능을 보였지만 퍼지적분 보다는 낮은 성능을 보였다. 그림 8은 다른 7개 분류기와의 비교를 보여준다. 20개의 데이타가 적절한 중간 학습 데이타의 수로 선택되어졌다. 학습 데이타의 크기는 20이며 나머지는 테스트 데이타로 사용되어졌다. 퍼지적분이 다른 분류기보다 우수한 성능을 냄을 확인할 수 있었다. 그림 9는 다른 결합방법과의 비교를 보여준다. 각 결합방법별로 학습 데이타의 크기를 5부터 45까지 5씩 증가시키면서 8가지 경우를 생각하였다. 각 8가지 경우에 대해 10번 반복 실험을 한 후 80번에 대한 평균을 구하였다. 비교실험 결과 퍼지적분이 다른 결합방법에 비해 우수한 성능을 보였다.

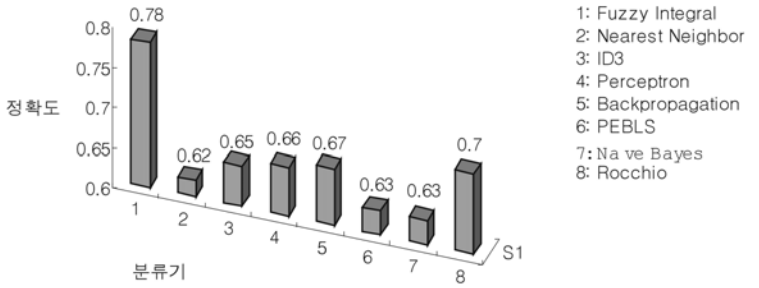
## 5. 결론 및 향후연구

본 논문에서는 퍼지적분을 사용한 다중 SASOM의 앙상블을 이용하여 사용자의 선호도에 기초한 웹 문서 분류를 수행하였다. 실험 결과는 제안하는 방법이 기존의 연구결과 뿐만 아니라 SASOM의 투표 결합보다 우수한 성능을 내는 것을 보여주었다. 퍼지적분은 분류기의 중요도를 주관적으로 결정할 수 있는 방법을 제공한다. SASOM은 문서를 높은 성능으로 분류하고 내부 작동 과정을 이해하기 쉽도록 시각화할 수 있는 방법을 제공한다. 제안하는 방법은 사용자 프로파일을 효과적으로 생성하여 적은 노력만으로도 효과적인 웹 마이닝을 수행하도록 도울 수 있다.

향후연구는 SASOM의 구조가 지니고 있는 의미를 체계적으로 분석하고 Bagging 또는 Boosting 등의 다양한 분류기를 생성하는 방법과 비교할 것이다.

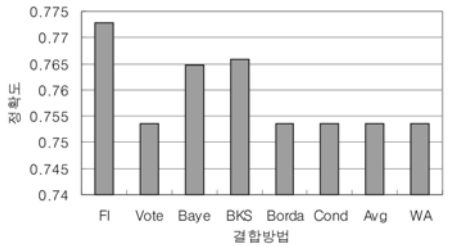



(a) Bands

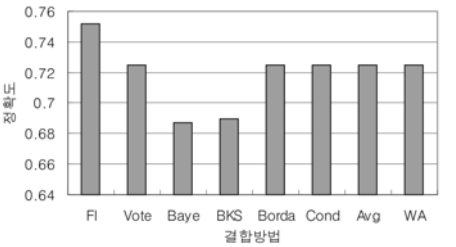



(b) Goats

그림 8 다른 분류기와의 성능비교




(a) Bands




(b) Goats

그림 9 다른 결합방법과의 비교 (결합방법은 순서대로 퍼지적분, 투표, Bayesian 방법, BKS, Borda 함수, Condorect 함수, 평균, 가중 평균이다.)

## 참고문헌

[1] J. Vesanto, "SOM-based data visualization methods," *Intelligent Data Analysis*, vol. 3, no. 2, pp. 111~126, August 1999.

[2] T. Kohonen, "The self-organizing map," *Proceedings of the IEEE*, vol. 78, no. 9, pp. 1464~1480, 1990.

[3] P. N. Suganthan, "Pattern classification using multiple hierarchical overlapped self-organising maps," *Pattern Recognition*, vol. 34, no. 11, pp. 2173~2179, Nov 2001.

[4] S.-B. Cho, "Neural-network classifiers for recognizing totally unconstrained handwritten numerals," *IEEE Transactions on Neural Networks*, vol. 8, no. 1, pp. 43~53, Jan 1997.

[5] S.-B. Cho, "Self-organizing map with dynamical node splitting: Application to handwritten digit recognition," *Neural Computation*, vol. 9, no. 6, pp. 1343~1353, 1997.

[6] S.-B. Cho, "Ensemble of structure-adaptive self-organizing maps for high performance classifica-

tion," *Information Sciences*, vol. 123, no. 1~2, pp. 103~114, March 2000.
[7] J. R. Quinlan, "Induction of decision trees," *Machine Learning*, vol. 1, pp. 81~106, 1986.
[8] D. Mladenic and M. Grobelnik, "Feature selection on hierarchy of web documents," *Decision Support Systems*, vol. 35, pp. 45~87, 2003.
[9] A. Verikas, A. Lipnickas, K. Malmqvist, M. Bacauskiene, and A. Gelzinis, "Soft combination of neural classifiers: A comparative study," *Pattern Recognition Letters*, vol. 20, pp. 429~444, 1999.
[10] S.-B. Cho, and J.-H. Kim, "Combining multiple neural networks by fuzzy integral for robust classification," *IEEE Transactions on Systems, Man and Cybernetics*, vol. 25, no. 2, pp. 380~384, February 1995.
[11] S. Hettich and S. D. Bay, The UCI KDD Archive, http://kdd.ics.uci.edu.
[12] M. Pazzani and D. Billsus, "Learning and revising user profiles: The identification of interesting web sites," *Machine Learning*, vol. 27, pp. 313~331, 1997.
[13] A. R. Mirhosseini, H. Yan, K.-M. Lam, and T. Pham, "Human face image recognition: An evidence aggregation approach," *Computer Vision and Image Understanding*, vol. 71, no. 2, pp. 213~230, 1998.
[14] S.-B. Cho and J.-H. Kim, "Multiple network fusion using fuzzy logic," *IEEE Transactions on Neural Networks*, vol. 6, no. 2, pp. 497~501, 1995.
[15] T. D. Pham, "Combination of multiple classifiers using adaptive fuzzy integral," *Proceedings of the 2002 IEEE International Conference on Artificial Intelligence Systems (ICAIS'02)*, pp. 50~55, 2002.
[16] A. S. Kumar, S. K. Basu and K. L. Majumdar, "Robust classification of multispectral data using multiple neural networks and fuzzy integral," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 35, no. 3, pp. 787~790, May 1997.
[17] S. Mitra, S. K. Pal, and P. Mitra, "Data mining in soft computing framework: A survey," *IEEE Trans. on Neural Networks*, vol. 13, no. 1, pp. 3~14, January 2002.
[18] T. Kohonen, S. Kaski, K. Lagus, J. Salojarvi, J. Honkela, V. Paatero and A. Saarela, "Self organization of a massive document collection," *IEEE Transactions on Neural Networks*, vol. 11, pp. 574~585, 2000.
[19] H.-U. Bauer and T. Villmann, "Growing a hypercubical output space in a self-organising feature map," *IEEE Transactions on Neural Networks*, vol. 8, no. 2, pp. 218~226, 1997.
[20] P. N. Sugathan, "Hierarchical overlapped SOM-based multiple classifiers combination," *In the 5th International Conference on Control, Automation, Robotics & Vision (ICARCV'98)*, pp. 924~927, 1998.
[21] D. Lewis, "Feature selection and feature extraction for text categorization," *Proceedings of the DARPA Workshop on Speech and Natural Language*, pp. 212~217, 1992.
[22] N. R. Pal, "Soft computing for feature analysis," *Fuzzy Sets and Systems*, vol. 103, pp. 201~221, 1999.
[23] M. Sugeno, "Fuzzy measures and fuzzy integrals: A survey," *Fuzzy Automata and Decision Processes*, Amsterdam: North Holland, pp. 89~102, 1977.
[24] K. Leszeynski, P. Penczek and W. Grochulskki, "Sugeno's fuzzy measures and fuzzy clustering," *Fuzzy Sets and Systems*, vol. 15, pp. 147~158, 1985.
[25] R. R. Yager, "Element selection from a fuzzy subset using the fuzzy integral," *IEEE Transactions on Systems, Man and Cybernetics*, vol. 23, pp. 467~477, 1993.

김 경 중
2000년 2월 연세대학교 컴퓨터과학과 졸업(학사). 2002년 2월 연세대학교 컴퓨터과학과 석사과정 졸업(석사). 2002년 3월~연세대학교 컴퓨터과학과 박사과정 재학중. 관심분야는 인공지능, 진화연산, 검색엔진, 이동로봇제어

조 성 배
1988년 연세대학교 전산과학과(학사) 1990년 한국과학기술원 전산학과(석사) 1993년 한국과학기술원 전산학과(박사) 1993년~1995년 일본 ATR 인간정보통신연구소 객원연구원. 1998년 호주 Univ. of New South Wales 초청연구원. 1995년~현재 연세대학교 컴퓨터과학과 정교수. 관심분야는 신경망, 패턴인식, 지능정보처리