□기술해설□

구조적 물체의 점진적 연역학습을 위한 속성 랜덤 그래프의 이용

경기대학교 성동수*
한국과학기술원 박규호**

목 차

1. 서 론

본 논문에서, 속성 랜덤 그래프의 개념을 이용하여 구조적 물체를 점진적으로 집단화하는 알고리즘을 소개한다. 또한 이 알고리즘에서 사용하는 평가함수를 속성 랜덤 그래프의 엔트로피 개념을 이용하여 유도하고, 이를 위하여, 먼저 속성 랜덤 그래프를 정의한다. 이 개념은 랜덤 버택스와 엣지에 단일의 속성을 가지고 있는 기존 랜덤 그래프 정의의 일반화 된 형태이다. 그다음, 속성 그래프들의 점진적인 집단화를 위한 엔트로피 최소화에 근거한 새로운 평가함수를 소개하고, 이 평가함수를 이용한 점진적인 알고리즘을 소개한다. 속성 그래프는 구조적 물체를 표현하기 위하여 사용되어 진다. 즉, 속성 그래프에서, 그래프 성분은 주어진 물체의 전반적인 구조를 나타내며, 버택스와 엣지에 있는 정보들은 각 요소 및 요소사이의 특성들을 나타낸다. 이것의 응용분야는 패턴인식, 인공 지능 등 다양한 분야에서 사용되며, 속성 그래프 정합, 속성 그래프들의 집단화 등 속성 그래프에 관련된 많은 흥미 있는 문제들이 있다. 특히 이중 집단화는 패턴 인식 및 기계학습의 중요한 분야중의 하나이다.

집단화는 주어진 입력들을 유사성의 정도에 따라 분류시키는 것이며, 이중 계층적 집단화는 입력들을 계층적으로 분류 및 집단화하여 개념 트리의 형태로 구성하는 것이며, 이는 궁극적으로 지식처리시스템이나 전문가 시스템의 지식 획득 문제를 쉽게 해준다. Fisher[6]는 평가함수로 category utility를 이용하여 속성 그래프의 점진적 집단화 알고리즘을 제시하였다. 그러나 여기에서 사용한 속성그래프는 엣지속성을 사용하지 않았다. Wong[3]은 랜덤 그래프사이의 거리측정 방법을 제시하고 이를 이용하여 집단화 알고리즘을 제시하였다. 그러나 이 방식의 단점은 점진 능력이 없다는 것이다. 집단화 문제는 엔트로피를 최소화하는 문제로 생각할 수 있으며[7], Wallance[8]는 집단화를 위하여 엔트로피 최소화에 근거한 평가함수를 제시하고 이를 이용하였다. 그러나 그의 방법은 점진 능력도 없으며, 속성 그래프들을 입력으로 하지도 않았다. 그뒤, 속성 그래프를 점진적으로 집단화하는 알고리즘들[9][10]이 제시되었으며,

*정회원
**종신회원

본 논문에서는 이를 소개할 것이다.

본 논문의 특징을 살펴보면 다음과 같다. 첫번째, 본 알고리즘은 엔트로피를 최소화하는 방법으로 구조적 물체들을 집단화하였다는 점이며, 둘째, 이러한 작업이 점진적으로 수행된다는 점이다. 이뜻은 학습과 분류가 동시에 이루어질 수 있음을 의미한다. 즉 분류작업이 모든 입력물체들을 이용하여 학습이 끝난뒤, 입력물체를 분류하는 것이 아니라, 학습진행 중에 언제라도 분류도 진행되어 질 수 있음을 의미한다. 이것은 실생활에서 학습은 점진적으로 일어나며, 분류는 모든 것을 학습한 뒤 이루어지지 않는다는 사실로부터, 이 사실은 중요한 특징임을 알 수 있다. 세번째 특징은 임의의 물체의 빠른 분류를 위하여 지식들을 계층 개념의 형태로 구성한다는 점이며, 네번째로, 구조적 물체들을 다룰 수 있다는 점이다. 이러한 특징들은 복잡한 실생활에서 적용할 수 있는 학습 시스템의 중요한 특성들임을 알 수 있다. 이전의 학습 시스템은 이러한 특성들의 일부분을 다루었을뿐 전체적으로 모든 부분을 다루지는 못하였다.

서론뒤, 본 논문의 구성은 다음과 같다. 2장에서는 속성 랜덤그래프를 정의하고, 이 랜덤그래프의 엔트로피를 계산하는 식을 유도할 것이다. 3장에서는 점진적 학습 알고리즘에서 여러가지 집단화된 결과들을 평가하기 위하여 필요한 평가 함수를 엔트로피 최소화에 근거하여 정의할 것이다. 4장에서는 점진적 학습 알고리즘을 소개하고, 5장에서는 간단하게 실험결과들을 소개한뒤, 마지막으로 6장에서 결론을 맺는다.

2. 속성 랜덤 그래프의 정의 및 엔트로피 계산

이 장에서는, 속성 랜덤 그래프를 정의하고, 이 랜덤그래프의 엔트로피를 구하는 식을 유도할 것이다.

정의 1:

속성그래프 $G_a=(V_a, E_a)$는 노드와 엣지에 속성값을 가지는 그래프로 정의된다.

여기에서 $V_a=\{v_1,...,v_p,...,v_q,...,v_n\}$는 속성 버텍스들의 집합이며, $E_a=\{...,e_{pq},...\}$는 속성 엣지들의 집합이며, 여기에서 엣지 e_{pq}는 속성을 가짐과 동시에 버텍스 v_p와 v_q를 연결하는 역활을 한다.

정의 2:

속성 랜덤 쌍은 속성이름과 랜덤 변수 x로 구성되어 있다. 속성 랜덤 버텍스 α와 속성 랜덤 엣지 β는 속성 랜덤 쌍들의 집합으로 구성되어 있다.

정의 3:

속성 랜덤 그래프 $R_a=(A_a, B_a)$는 랜덤 버텍스와 랜덤 엣지에 속성값을 가지고 있는 랜덤 그래프로 정의되며[11], 여기에서, $A_a=\{\alpha_1,...,\alpha_p,...,\alpha_q,...,\alpha_n\}$는 속성 랜덤 버텍스들의 집합이며, $B_a=\{...,\beta_{pq}\}$는 랜덤 엣지들의 집합이다.

정의 4:

속성 랜덤 그래프 $R_a=(A_a, B_a)$의 확률 공간에서 출력은 속성 그래프 $G_a=(V_a, E_a)$이다.

정의 5:

랜덤 그래프 R_a의 확률공간은 그래프 모노모피즘 $M:G_a\rightarrow R_a$과 함께, 출력 그래프 G_a들로 구성되어 있다. 여기에서 M은 다음의 두개의 사상들로 표현된다. (μ, γ), $\mu: v\rightarrow\alpha$과 $\gamma: e\rightarrow\beta$. $\mu(v)$과 $\gamma(e)$는 랜덤 버텍스 α와 랜덤엣지 β를 나타내며, $\mu^{-1}(v)=\alpha$이고 $\gamma^{-1}(\beta)=e$이다. 속성 그래프 G_a가 속성 랜덤그래프 R_a로부터의 출력일 확률 $P(G_a, M)$는 다음을 만족한다.

$$P(G_a, M)\geq 0 \text{ for all } G_a\in\Gamma, \qquad (1)$$

$$\sum_{G_a\in\Gamma} P(G_a, M)=1,$$

여기에서 Γ는 R_a의 치역이다.

속성 랜덤 그래프의 모든 엣지들과 버텍스들로 이루어진 확률 공간은 가능한 부분집합과 불가능 부분집합으로 나누어질 수 있다. 불가능한 부분집합은 출력된 결과에서 적어도 하나의 엣지가 하나의 버텍스에만 연결된 경우이다. 따

라서 불가능집합에 속한 출력은 (1)에 의하면 속성 랜덤그래프의 출력이 될수없다. 따라서 출력그래프의 확률은 불가능 부분집합의 확률공간이 0인, 랜덤그래프의 모든 랜덤 엣지들과 랜덤 버택스들로 이루어진 공간내의 대응되는 출력의 확률과 같다고 정의된다.

$$P(G_a, M) = \text{Prob}\{(\alpha = \mu^{-1}(\alpha), \quad \text{for all } \alpha \in A_a), (\beta = \gamma^{-1}(\beta) \quad \text{for all} \quad \beta \in B_a)\}. \quad (2)$$

실생활에서의 응용에서 출력그래프의 확률을 계산하기 위하여 확률장의 차원을 낮출 필요가 있으며, 이를 위하여 다음의 가정들이 이루어졌다. 첫째, 랜덤 버택스들은 α, $\alpha \in A_a$, 상호 독립적이며, 랜덤 엣지들과도 독립적이다. 둘째, 랜덤 엣지 β, $\beta \in B_a$, 들은 상호 독립적이며, 각 엣지와 연결된 랜덤 버택스 $\sigma(\beta)$와 $\tau(\beta)$를 제외한 나머지 랜덤 버택스와는 독립적이다. 셋째, 랜덤 버택스와 랜덤 엣지내의 속성 랜덤 쌍들은 상호 독립적이다. 끝으로, 만일 랜덤 엣지의 양쪽의 랜덤 버택스의 출력이 실재로 존재한다면, 랜덤 엣지로부터의 출력의 확률은 양쪽의 실제값에 관계없이 없다. 위의 가정들로부터, 출력 그래프 $P(G_a, M)$의 확률식은 다음과 같이 표현된다.

$$P(G_a, M) = \prod_{\alpha \in A_a} \text{Prob}\{\alpha = \mu^{-1}(\alpha)\} \prod_{\beta \in B_a} \text{Prob}\{\beta = \gamma^{-1}(\beta) : \sigma(\beta) \neq \phi, \quad \tau(\beta) \neq \phi\}, \quad (3)$$

여기에서 B_a는 각 랜덤 엣지의 끝점의 출력이 출력 그래프 C_a안에 존재하는 랜덤 엣지의 집합이다. 위의 수식을 이용하기 쉬운 형태로 표현하기 위하여, 다음의 기호들이 사용되었다.

$$P_\alpha = \text{Prob}\{\alpha = \phi\}, \quad (4)$$
$$Q_\alpha = 1 - P_\alpha, \quad (5)$$
$$P_\beta = \text{Prob}\{\beta = \phi : \sigma(\beta) \neq \phi, \tau(\beta) \neq \phi\}, \quad (6)$$
$$Q_\beta = 1 - P_\beta, \quad (7)$$
$$C_\beta = \text{Prob}\{\sigma(\beta) \neq \phi, \tau(\beta) \neq \phi\}, \quad (8)$$
$$P_\alpha(v) = \text{Prob}\{\alpha = v\} = Q_\alpha \prod_{x_i \in \alpha} p_{x_i}(X_i) \quad (9)$$
$$P_\beta(e) = \text{Prob}\{\beta = e : \sigma(\beta) \neq \phi, \tau(\beta) \neq \phi\}, \quad (10)$$
$$= Q_\beta \prod_{x_i \in \beta} P_{x_i}(X_i \mid \sigma(\beta) \neq \phi, \tau(\beta) \neq \phi), \quad (11)$$
$$= Q_\beta \prod_{x_i \in \beta} P^*_{x_i}(X_i) \quad (12)$$

식 (3)에서 (12)까지를 이용하여, 출력 그래프의 확률을 아래와 같이 나타낼 수 있다.

$$P(G_a, M) = \prod_{\alpha \in A_a} P_\alpha(\mu^{-1}(\alpha)) \prod_{\beta \in B_a} P_\beta(\gamma^{-1}(\beta)). \quad (13)$$

속성 랜덤그래프의 변화성을 반영하기 위하여, 엔트로피의 특성이 사용될 수 있다. 엔트로피의 정의로부터, 속성 랜덤그래프의 엔트로피는 다음과 같이 정의된다.

$$H(R_a) = -\sum_{(G_a, M) \in \Gamma} P(G_a, M) \log P(G_a, M), \quad (14)$$

여기에서 Γ는 R_a의 치역이다. 만일 랜덤그래프의 출력 그래프들 사이에 많은 유사성을 가지고 있다면, $H(R_a)$의 값은 작으며, 그렇지 않을 경우 그 값은 크다. 식 (13)과 식 (14)로부터 아래의 식이 얻어진다.

$$H(R_a) = \sum_{\alpha \in A_a} H_\alpha + \sum_{\beta \in B_a} H_\beta. \quad (15)$$

식 (15)의 속성 랜덤 버택스 α의 엔트로피는 다음과 같이 표현된다.

$$H_\alpha = -\sum_v P_\alpha(v) \log P_\alpha(v) \quad (16)$$

만일 식 (9)를 식 (16)에 대입하면, 다음과 같은 식이 유도된다.

$$H_\alpha = -Q_\alpha \sum_{x_i \in \alpha} \sum_{X_i \in x_i} P_{x_i}(X_i) \log P_{x_i}(X_i) - P_\alpha \log P_\alpha - Q_\alpha \log Q_\alpha, \quad (17)$$

(15)에 있는 랜덤 엣지 β의 조건 엔트로피는 다음과 같이 계산된다.

$$H_\beta = -C_\beta \sum_e P_\beta(e) \log P_\beta(e). \quad (18)$$

이 식은 식 (12)에 의하여 아래와 같이 계산된다.

$$H_\beta = -C_\beta Q_\beta \sum_{x_i \in \alpha} \sum_{X_i \in x_i} P^*_{x_i}(X_i) \log P^*_{x_i}(X_i)$$

$$-C_\beta P_\beta \log P_\beta - C_\beta Q_\beta \log Q_\beta. \quad (19)$$

3. 엔트로피 최소화에 근거한 평가 함수의 정의

집단화 문제는 일종의 에너지 최적화 문제로 생각할 수 있으며, 엔트로피는 집단화를 위한 평가함수로 사용될 수 있다. 왜냐하면, 랜덤그래프 R의 출력 그래프들간의 상당한 유사성이 있으며, 랜덤 그래프의 엔트로피 $H(R)$이 적어지기 때문이다. 따라서 집단화 문제는 엔트로피를 최소화 하는 문제로 생각할 수 있다.

만일 우리가 속성 그래프로 표현되는 물체들을 입력으로하여, 이를 N개의 집단 C_1, C_2,..., C_N으로 분류한다고 가정했을때, 각 집단은 속성 랜덤 그래프로 생각할 수 있고, 엔트로피 최소화에 근거한 평가함수는 아래와 같이 정의될 수 있다.

$$CF_1 = \sum_{k=1}^{N} P(C_k) H(R(C_k)) \quad (20)$$

, 여기에서 $P(C_k) = f_k / f$는 모든 입력물체에 대한 집단 C_k의 상대빈도이며, f_k는 집단 C_k의 빈도이다. 그리고 f는 전체빈도이다. $R(C_k)$는 집단 C_k를 표현하는 속성 랜덤 그래프이며, $H(R(C_k))$은 $R(C_k)$의 엔트로피 값이다.

모든 입력 그래프로 이루어진 랜덤 그래프 $H(R)$의 엔트로피 값은 각 군의 엔트로피값의 무게화합 보다 작다는 사실로부터, 엔트로피 최소화에 근거한 다른 평가함수를 아래와 같이 정의할 수 있다.

$$CF_2 = H(R) - \sum_{k=1}^{N} P(C_k) H(R(C_k)). \quad (21)$$

이 식은 평가함수를 최대화 한다는 점만 제외하면 그 의미는 식 (20)과 동일하다. 만일 입력물체들을 임의의 군으로 집단화 한다면, 위식은 군의 수가 많으면 많을수록 그 값이 커지기 때문에 변경되어야 함을 알 수 있다. 따라서 군의 수가 정해지지 않은 집단화의 경우의 평가함수는 아래와 같이 정의될 수 있음을 알 수 있다.

$$CF_3 = [H(R) - \sum_{k=1}^{N} P(C_k) H(R(C_k))] / N \quad (22)$$

, 여기에서 N은 정해지지 않은 군의 수이다.

만일 모든 입력 예들을 한번에 모두 가지고 있다면, 계층 개념은 이 입력들을 주어진 평가함수에 따라 재귀적으로 분리함에 의하여 얻을 수 있다. 점진적인 시스템의 정의[8]에 의하면 한번에 한개의 입력만을 받을 수 있으며, 이 입력을 이전의 계층 개념에 포함하여 새로운 계층 개념을 만들때에 이전의 모든 입력들로부터 만들어진 계층 개념을 모두 이용하면 안된다. 따라서, 점진적 집단화는 자연히 학습과 분류가 동시에 이루어져야 한다는 사실로 귀결된다. 사실상 대부분의 점진적 집단화 알고리즘들[5], [6]은 언덕오르기 전략과 함께 나무 탐색 방법으로 계층 개념을 형성한다. 즉 입력은 주어진 평가함수와 함께 분류에 의해서 계층 개념을 탐색하며, 동시에 탐색 경로에 있는 개념들은 이 입력을 포함하는 새로운 개념들로 재구성된다. 본 논문에섬 제시한 시스템도 입력으로 속성그래프를 가지는 점진적인 시스템이며, 따라서 시스템을 평가하기 위한 평가함수를 제외하고 일반적인 점진 시스템이 사용하고 있는 전략을 대부분 그대로 사용하였다. 그리고 계층 개념을 탐색하는 경로에 있는 개념들을 재구성함에 의하여 만들어지는 여러가지의 집단들을 평가하기 위한 평가함수를 식 (22)와 같이 정의하였다. 우리는 계층 개념 내에서 부모노드로 C를 가지고 자식들을 집단화한 경우의 평가함수를 아래와 같이 정의하였다.

$$CF(C, C') = [H(R(C))$$

$$- \sum_{k=1}^{N_c} P(C_k) H(R(C_k))] / N_C, \quad (23)$$

여기에서 C'는 $[C_1,...,C_{N_c}]$이며 이는 개념 C의 자식 개념들이다.

우리는 식 (23)을 우리의 집단화 알고리즘의

평가함수로 사용하였으며, 이는 자식 개념들로 이루어진 집단의 엔트로피를 최소화 하도록 집단화함을 알 수 있다.

4. 점진적 계층 집단화 알고리즘의 소개

이장에서는 점진적 집단화 알고리즘을 소개할 것이다. 집단화 알고리즘(IL)은 입력 예를 가장 잘 수용하는 개념을 찾아 계층 개념 트리를 탐색하고, 그 경로에 있는 개념들을 개선함에의해 점진적으로 계층 개념트리를 변화 시킨다. 알고리즘은 입력예와 함께 개념 트리를 탐색할때, 언덕오르기 탐색 방법을 사용하며, 이때 사용하는 평가함수는 3장에서 정의한 평가함수를 이용한다. 일반적으로, 알고리즘은 입력예를 가능한한 개념 트리의 가장 아래에 두기 위하여 개념 트리를 탐색한다. 탐색 과정을 돕기위하여 네 가지의 점수들이 계산되며, 이 점수로부터 다섯가지의 연산자중 하나가 적용된다. 이 연산자에 의하여 탐색 과정에 있는 경로에 해당하는 개념들이 변화되며, 이는 계층 개념트리의 변화를 의미한다.

학습 알고리즘은 표 1에 있는 것처럼 IL (N_{root}, I)로 시작하며, 여기에서 N_{root}는 계층 개념의 최상위 개념이며, I는 입력 예 이다. 아래의 과정에서 보면 알 수 있듯이, IL 입력 예가 계층 개념을 탐색도중 계속 탐색을 진행할 것인가 현재 상태에서 정지할 것인가를 결정하기위하여, 탐색도중 선택된 개념에 입력 예가 충분히 유사한가를 판단한다. 만일 입력 예와 선택된 개념사이의 유사도가 미리 정해둔 한계값보다 클 경우, 여러 연산자는 가해지지 않고, 단순히 현재 선택된 개념이 입력 예와의 합성에의하여 개선되며, 알고리즘은 여기에서 끝나게 된다. 일반적으로, 계층 개념은 생성 연산자 또는 분리 연산자에 의하여 좌우로 넓어지게 되며, 결합 연산자 및 종료 연산자와 함께 길어지게 된다. 그리고 $N^* = Synthesis(N, I)$은 개념 N^*가 개념 N과 I를 합성함에의하여 생성되

표 1 점진적 집단화 알고리즘

Procedure IL(N, I)
1. **if** Similarity(N, I) ≥ Threshold,
   1.1 N^* = Synthesis(N, I).
   1.2 **stop.**
2. if N is a terminal concept,
   Termination(N, I).
3. **else**
   3.1 N^* = Synthesis(N, I).
   3.2 compute S_{search}, $S_{creation}$, $S_{merging}$, $S_{splitting}$.
   3.3 **case** max_score = max(S_{search}, $S_{creation}$, $S_{merging}$, $S_{splitting}$),
   S_{search} : IL(C_j, I). /* 그림 1(b) */
   $S_{creation}$: Creation(N^*, I). /* 그림 1(c) */
   $S_{merging}$: IL(Merging(N, I), I). /* 그림 1(d) */
   $S_{splitting}$: IL(Splitting(N, I), I). /* 그림 1(e) */

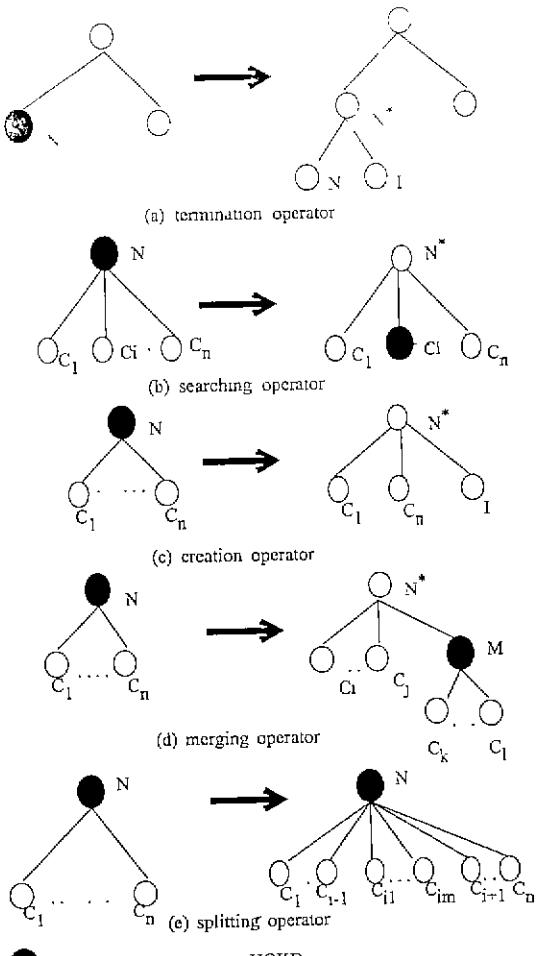


그림 1 IL에서 사용한 다섯가지 연산자들

었음을 의미한다. 위의 학습 알고리즘에서, $S_{searching}$, $S_{creation}$, $S_{merging}$, $S_{splitting}$은 각각 다음과 같이 계산된다.

탐색 점수:

입력 예 I와 가장 비슷한 개념을 결정하기 위하여, IL은 각각의 개념 C_i에 입력 예를 포함시켜 집단을 형성시켜 본다. 이러한 과정에의해, 공통의 부모 개념 N을 가지는 다수의 집단들이 형성되며, 각각의 집단들은 평가함수에의하여 평가된다. 그리고 그중 최고의 평가 값을 가지는 집단을 선택하고, 이값이 $S_{searching}$이되며, 아래와 같이 수식화 될 수 있다.

$$S_{searching} = \max_{1 \le i \le n} [CF(N^*, C^i)]$$

여기에서 C^i는 $[C_1, C_2, ..., C_i^*, ..., C_n]$이며, C_i^*는 C_i와 I를 합성함에의해 얻어진다. 또한 N^*는 N과 I를 합성함에의해 얻어진다.

생성 점수:

계층 개념에서 개념 C 아래로 탐색하는 것 대신에, 개념 C 아래에 새로운 개념을 생성한다. 이 새로운 개념은 단지 입력 예를 이용하여 생성되며, 이것에 의하여 만들어진 집단을 평가함수를 이용하여 평가한다. 평가에의하여 계산된 이값을 $S_{creation}$라하고 아래와 같이 수식화 된다.

$$S_{creation} = CF(N^*, C + I).$$

결합 점수:

결합 점수를 계산하기 위하여, 입력 예 I를 잘 수용할 수 있는 두개의 개념들을 선택한다. 그다음, 이 두개의 개념을 결합하여 얻어지는 이 집단을 평가한다. 이 평가점수를 $S_{merging}$이라 하고 이는 다음과 같이 수식화 된다.

$$M = [C_i \mid \max_{C_i \in C} CF(N^*, C^i)] \cup [C_j \mid \max_{C_j \in C - C_i} CF(N^*, C^j)]$$

$$S_{merging} = CF(N^*, C - C_i - C_j + M)$$

분리 점수:

분리 점수는 입력예를 가장 잘 수용하는 개념을 선택한후 이를 제거하고 그대신 이 개념의 자식 개념들을 이 개념의 위치에 놓고, 이것에 의하여 만들어진 개념 트리에서 탐색 점수를 계산하는 방법에의하여 얻어진다. 만일 선택된 개념이 m개의 자식 개념들을 가지고 있다면, 개념 트리의 현 노드아래의 n개의 개념들은 $n+m-1$개의 개념들로 확장된다.

$$C = C \cup H_j - [C_j \mid \max_{1 \le j \le n} CF(N^*, C^j)]$$

$$S_{splitting} = \max_{C_i \in C} CF(N, C^i)$$

여기에서 $H_j = [h_1, ..., h_m]$는 최고 개념 C_j의 자식들의 집합이다.

위의 점수를 계산하기 위하여, 계층 계념이 바뀌지는 않고, 다음에 설명할 연산자들에 의하여 바뀌게 된다. 즉 위의 점수들을 계산하는 이유는 어떤 연산자를 이용하여 계층 개념을 바꿀것인가를 결정하기 위함이다. 본 논문의 집단화 알고리즘에서 이용한 연산자의 종류는 다섯가지이며, 이를 각각 설명하면 아래와 같다.

탐색 연산자:

이 연산자는 선택된 개념과 입력예를 합성함에의해 선택된 개념을 변화시키고 계층 개념에서 선택된 개념 아래로 이 입력 예를 탐색 시킨다.

생성 연산자:

이 연산자는 그림 1(c)에 보는것처럼, 입력예를 이용하여 새로운 개념을 생성하고 그 개념을 개념 N의 자식 개념으로 둔다. 이 경우, 탐색은 여기에서 끝나게 된다.

합성 연산자:

이연산자는 그림 1(d)에서 보는 것 처럼, 선택된 두개의 개념을 합성함에의해 n개의 개념들을 n−1개의 개념들로 변환함에의해 계층 개념을 변환한다. 이때, 두 개념의 합성은 각각의 개념이 속성 랜덤그래프에 의하여 표현되기 때문에 두개의 속성랜덤그래프 사이의 정합문제가 되며 이는 seong[12]의 방법을 이용하였다.

분리 연산자:

이 연산자는 그림 1(a)에서 보는 것 처럼, 선택된 개념을 제거하고 그대신 그개념의 자식들을 그 위치에 두어 계층 개념을 변화 시키는 연산자이다.

종료 연산자:

만일 현재 시험중인 개념 N가 계층 개념의 끝 개념이라면, 종료 연산자가 적용된다. 이 연산자는 개념 N과 입력 예 I를 합성함에의해 내부 개념 N^*를 생성한다. 그리고 또한 입력 예 I와 함께 새로운 개념도 끝 개념으로 생성한다. 이렇게하여, N과 I는 그림 1(a)에 보는것처럼 개념 N^*의 자식 개념들이 됨을 알 수 있다.

5. 실험 결과들

알고리즘은 C 언어와 Common Lisp 언어로 구성되어 있으며 SUN workstation에서 수행되었다. 제시한 알고리즘의 유용성을 입증하기 위하여 2차원 형태의 장난감 및 비행기들의 형태를 입력으로하여 실험하였다. 전체 시스템은 그림 2에 보는바와 같이 3개의 모듈로 이루어져 있다. 입력 이진 형태의 구조적 서술을 위하여, 입력 물체를 여러 형태로 분리하는 것이 필요하며, 이를 위하여 Kim[4]이 제시한 인지적 다각형 형태 알고리즘을 이용하였다. 입력 이진 영상의 분리는 두단계에 걸쳐서 일어나며, 첫째, 물체의 외각이 엣지 찾기 및 외각 찾기 알고리즘에 의하여 찾아진다. 이를근거로하여, 외각선은 여러개의 세그먼트로 근사화되며, 근사화된 형체는 형체 분리 알고리즘을 이용하여 구조적으로 분리되어 진다. 그림 3(a)는 입력 이진 영상이며, 그림 3(b)는 엣지 찾기 알고리즘 및 근사화 알고리즘을 이용하여 얻어진 결과 영상이며 그림 3(c)는 분리알고리즘을 이용하여 분리된 영상을 도시한 것이다. 이 분리된 형태는 다음과 같이 구조적 서술로 표현된다. 첫째, 물체의 골격이 물체를 구조적으로 서술하기 위하여 추출된뒤, 각 요소 및 그들 사이의 속성들이 서술 된다. 이 구조적 서술은 분리된 형태의 요소를 버텍스에, 요소들 사이의 관계를 엣지에 할당함으로써, 속성 그래프로 표현된다.

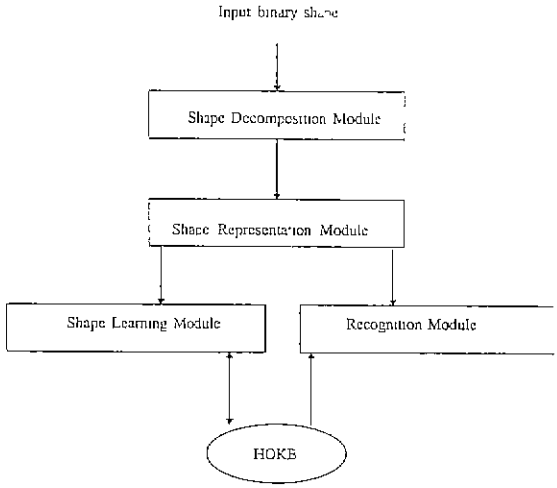


그림 2 전체적인 시스템 구성도

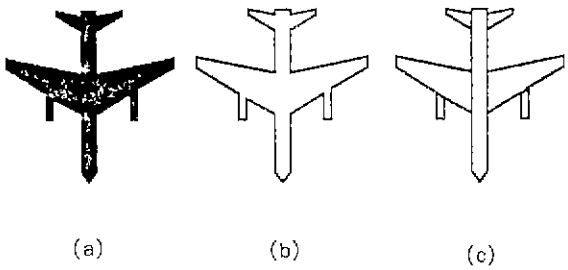
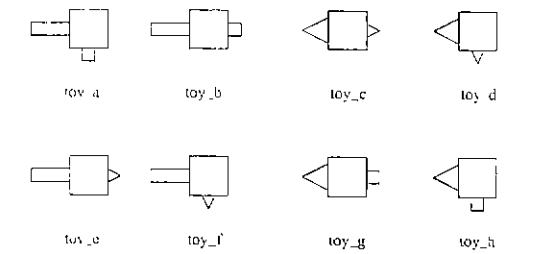


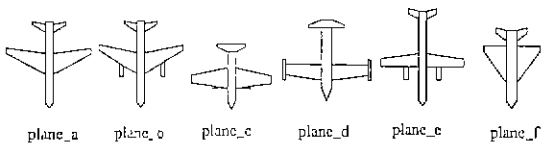
그림 3 (a) 이차원 이진 영상
(b) 근사화된 영상
(c) 분리된 영상

결과로서, 분할된 형태의 입력예가 그림 4에 있으며, 구성된 계층 개념이 그림 5에 나타나 있다.

이를 입력으로 하여 알고리즘을 수행한 결과, 주어진 물체가 소속된 군을 입력으로 주지 않음에도 불구하고, 모든 입력예들은 자신의 군으로 집단화 하였으며, 각 군들도 그들의 유사도에따라 계층적으로 집단화 되었음을 결과로 부터 알 수 있다. 계층적 개념을 형성하는 연산자의 하나인 종료 연산자에 정의된 한계값을 증가할수록, 계층 개념내의 끝 노드들의 수가 상

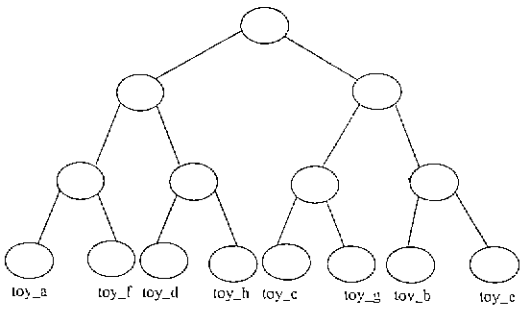


(a)

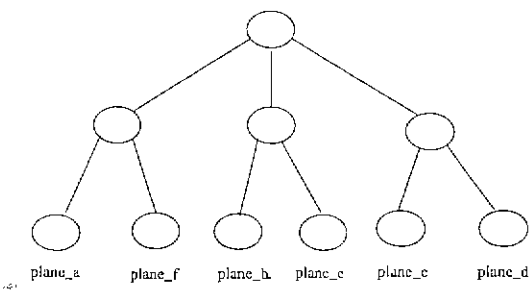


(b)

그림 4 (a)장난감 군
(b)비행기 군



(a)



(b)

그림 5 (a)장난감 군들에 대한 계층 개념
(b)비행기 군들에 대한 계층 개념

대적으로 줄어듬을 알 수 있었다. 즉 이 값이 증가할 수록 끝노드의 개념이 좀더 일반적인 개념이되며, 이로부터 전체의 계층 개념이 간단해짐을 알 수 있다. 이 값은 실제응용에서 집단화 알고리즘이 너무 큰 계층 개념을 만드는것을 방지해준다.

모르는 물체를 분류함에의해 만들어진 계층 개념의 질을 평가하기 위하여, 집단화 과정에서 사용되지 않은 나머지 물체들을 사용하였다. 결과로서, 거의 100%의 분류율을 얻었다. 왜냐하면, 군들사이에 뚜렷한 구조적 차이가 있으며, 평가함수 및 집단화 알고리즘의 우수성 때문이다. 그다음으로, 계층 개념에 저장된 가장 유사한 개념으로 분류되도록, 변질된 요소를 가지고 있는 입력을 이용하였다. 즉 가장큰 요소를 10%정도 늘리거나 줄임에의해 왜곡된 입력을 만들었다. 비행기 입력의 경우에, 알고리즘은 98%의 분류율을 보였다. 마지막으로, 학습한 물체중 일부가 가려짐에의하여 왜곡된 경우를 실험하기위하여, 제일작은 요소를 제외한 물체를 입력으로 사용하였다. 비행기의 입력을 가지고 이경우를 실험한 결과, 약 95%의 분류율을 보였다.

현존하는 계층 개념으로부터 입력 예를 인식하는데 드는 비용을 계산하기위하여, B를 계층 개념내 각 개념의 평균 가지수라 하고, n을 이전에 분류되어진 물체의 수라고하고, k_1을 정합을 위한 평균 비용이라하고, k_2를 평가를 위한 평균 비용이라 할때, 전체 비용은 대략 $O(K_1 B \log_B n)$이다. 이 비용은 개념들이 계층적으로 구성되어 있지 않은 시스템의 비용인 $O(K_1 n)$보다 적음을 알 수 있다. 계층 개념으로 새로운 입력을 포함시키는데 필요한 시간을 현재의 입력이전에 시스템에 주어진 입력수에 대하여 $K_2(4B-2)\log_B n$로 계산되며 따라서 이는 $O(K_2 B \log_B n)$이다. 안정한 계층 개념을 만드는데 필요한 입력의 수는 주어진 입력들의 입력순서에따라 약간 달라지게된다. 그러나 일반적으로 군 내부의 요소들의 구조적 유사도가 다른 군내의 요소과의 유사도보다 상당히 크다면, 입력들의 주어진 순서에 관계없이 유사한 계층 개념이 만들어짐을 실험들을 통하여 알 수 있었다.

6. 결 론

본 논문에서는, 구조적 물체들의 점진적으로 집단화하기 위하여 속성 랜덤 그래프의 엔트로피 최소화에 근거한 평가 함수를 정의 하였다. 또한, 정의된 평가함수를 이용하여 속성 그래프들로 표현되는 구조적 물체들을 점진적으로 집단화하는 알고리즘도 소개하였다. 속성그래프들을 입력으로하여, 집단화 알고리즘은 언덕오르기 전략을 이용하여 개념 트리를 점진적으로 형성한다. 즉 알고리즘은 개념 트리를 아래 방향으로 탐색하면서 각 층마다 모든 가능한 상태들로 개념들을 형성하고 정의된 평가함수를 이용하여 이중 최선의 상태를 찾아내어 개념 트리를 형성한다. 이런 과정을 개념 트리의 제일 아래까지 수행한다. 이러한 과정을 거쳐, 계층 개념이 점진적으로 만들어 지며, 이는 인간의 학습과정과 매우 유사함을 알 수 있다.

참 고 문 헌

[1] W. H. Tsai and K. S. Fu, "Error-correcting isomorphisms of attributed relational graphs for pattern analysis," IEEE Trans. Syst. Man. Cybern., vol. SMC-9, no. 12, pp. 757-768, Dec., 1979.

[2] A. K. C. Wong and H. E. Ghahraman, "Random graph : Structural-contextual dichotomy," IEEE Trans. Pattern Anal. Machine Intell., vol. PAMI-2, no. 4, pp. 341-348, July, 1980.

[3] A. K. C. Wong and M. You, "Entropy and distance of random graphs with application to structural pattern recognition," IEEE Trans. Pattern Anal. Machine Intell., vol. PAMI-7, no. 5, pp. 599-609, Sep., 1985.

[4] H. S. Kim, K. H. Park, and M. Kim, "Shape decomposition by collinearity," Pattern Recognition Letters 6, pp. 335-340, 1987.

[5] D. H. Fisher, "Knowledge acquisition via incremental conceptual clustering," Machine Learning, vol. 2, no. 2, pp. 139-172, 1987.

[6] J. H. Gennari and P. Langley and D. H. Fisher, "Models of incremental concept formation," Artificial Intell., vol. 40, pp. 11-61, 1989.

[7] S. Watanabe, "Pattern recognition as a quest for minimum entropy," Pattern Recognition, vol. 13, no. 5, pp. 381-387, 1981.

[8] R. S. Wallance and T. Kanade, "Finding natural clusters having minimum description length," in proc. Pattern Recognition, pp. 438-442, 1990.

[9] 성동수, 김호성, 박규호, "엔트로피 평가에 근거한 속성 그래프들의 점진적인 집단화," 정보 과학회 논문지 Vol. 21, No. 9, pp. 1692-1701, 1994.

[10] Dong Su Seong, Ho Sung Kim, Kyu Ho Park "Incremental Clustering of Attributed Graphs" IEEE Trans. Syst. Man. Cybern. Vol. 23, No. 5, 1993.

[11] Dong Su Seong, Ho Sung Kim, Kyu Ho Park "Definition of Attributed Random Graph and Proposal of its applications" IEICE Tran. on Information and Systems. Vol. E76-D, No. 8, 1993.

[12] Dong Su Seong, Young Kyu Choi, Ho Sung Kim, Kyu Ho Park "Optimal Graph isomorphism between Two Random Graphs" Pattern Recognition Letter Vol. 15, pp. 321-327, 1994.

성 동 수

1987 한양대학교 공피대학 전자공학과 졸업(학사)
1989 한국과학기술원 전기 및 전지공학과 졸업(석사)
1992 한국과학기술원 전기 및 전지공학과 졸업(박사)
1992~1993 한국과학기술원 정보전자연구소 연구원
1993~현재 경기대학교 공과대학 전자공학과 조교수
관심분야 : 지능컴퓨터, 멀티미디어, 빔렬처리

박 규 호

1973 서울대학교 전자공학과 졸업(학사)
1975 한국과학기술원 전기 및 전자공학과 졸업(석사)
1975~1978 동양정밀공업주식회사 개발과장
1977 Philips Research Lab. Eindhoven Netherlands 연구원
1983 프랑스 Parix XI 대학교 전자공학과 전공(박사)
1983~현재 한국과학기술원 전기 및 전자공학과 교수
관심분야 . 지능컴퓨터, 병렬처리, 컴퓨터비전

● '95 정보문화의 달 기념강연회 · 전시회 ●

- 일 자 : 1995년 6월 23일(금)
- 장 소 : 경남대학교 대회의실
- 주 최 : 영남지부
- 주 관 : 경남대 전자계산소, 전자계산학과, 전산통계학과
- 문 의 : 경남대 전자계산학과 박규석 교수
  T. 0551-49-2650
  F. 0551-46-6184