고차상관관계를 표현하는 랜덤 하이퍼그래프 모델 진화를 위한 베이지안 샘플링 알고리즘

(A Bayesian Sampling Algorithm for Evolving Random Hypergraph Models Representing Higher-Order Correlations)

이 시 은 † 이 인 희 †† 장 병 탁 †††
(Si Eun Lee) (In-Hee Lee) (Byoung-Tak Zhang)

요 약 유전자알고리즘의 교차나 돌연변이 연산을 직접적으로 사용하지 않고 개체군의 확률분포를 추정하여 보다 효율적인 탐색을 수행하려는 분포추정알고리즘이 여러 방법으로 제안되었다. 그러나 실제로 변수들간의 고차상관관계를 파악하는 일은 쉽지 않은 일이라 대부분의 경우 낮은 차수의 상관관계를 제한된 가정하에 추정하게 된다. 본 논문에서는 데이타의 고차상관관계를 표현할 수 있고 최적 해를 좀 더 효율적으로 찾을 수 있는 새로운 분포추정알고리즘을 제안한다. 제안된 알고리즘에서는 상관관계가 있을 것으로 추정되는 변수들의 집합으로 정의된 하이퍼에지로 구성된 랜덤 하이퍼그래프 모델을 구축하여 변수들 간의 고차상관관계를 표현하고, 베이지안 샘플링 알고리즘(Bayesian Sampling Algorithm)을 통해 다음 세대의 개체를 생성한다. 기만하는 빌딩블럭(deceptive building blocks)을 가진 분해가능(decomposable) 함수에 대하여 실험한 결과 성공적으로 최적해를 구할 수 있었으며 단순 유전자알고리즘과 BOA (Bayesian Optimization Algorithm)와 비교하여 좋은 성능을 얻을 수 있었다.

키워드 : 베이지안 샘플링 알고리즘, 베이지안 진화연산, 분포추정알고리즘, 랜덤 하이퍼그래프

Abstract A number of estimation of distribution algorithms have been proposed that do not use explicitly crossover and mutation of traditional genetic algorithms, but estimate the distribution of population for more efficient search. But because it is not easy to discover higher-order correlations of variables, lower-order correlations are estimated most cases under various constraints. In this paper, we propose a new estimation of distribution algorithm that represents higher-order correlations of the data and finds global optimum more efficiently. The proposed algorithm represents the higher-order correlations among variables by building random hypergraph model composed of hyperedges consisting of variables which are expected to be correlated, and generates the next population by Bayesian sampling algorithm.

Experimental results show that the proposed algorithm can find global optimum and outperforms the simple genetic algorithm and BOA(Bayesian Optimization Algorithm) on decomposable functions with deceptive building blocks.

Key words : Bayesian Sampling Algorithm, Bayesian evolutionary computation, EDA, random hypergraph

† 정 회 원 : 백석대학교 정보통신학부 교수
leese@bu.ac.kr
†† 학생회원 : 서울대학교 컴퓨터공학부
ihlee@bi.snu.ac.kr
††† 종신회원 : 서울대학교 컴퓨터공학부 교수
btzhang@bi.snu.ac.kr
(Corresponding author임)
논문접수 : 2008년 10월 28일
심사완료 : 2009년 1월 22일

정보과학회논문지: 소프트웨어 및 응용 제36권 제3호(2009.3)

1. 서 론

자연세계의 진화현상에 기반하여 교차와 돌연 변이 연산을 이용하는 유전자알고리즘은 적응적 탐색과 최적화를 통해 실세계의 문제 해결에 많이 응용되어 왔으나 문제의 복잡도가 커짐에 따라 최적해에 접근하지 못하는 결과를 가져올 수도 있다. 이러한 문제점을 해결하기 위하여 기존 유전자알고리즘의 교차와 변이 연산자들을 직접적으로 사용하지 않고 개체군의 확률분포를 추정하여 효율적인 탐색을 수행하려는 시도들이 있어 왔다. 대표적으로 분포추정 알고리즘(Estimation of Distribution Algorithms, EDA)[1]이 있으며 적합도가 좋은 후보 해들의 확률 분포를 학습하여 확률 모델을 생성하고 그 확률 모델로부터 새로운 후보 해들을 샘플링 해 나간다. 따라서 앞 단계로부터 키워온 빌딩블럭들을 분해하지 않고 변수들간에 상호작용이 있는 문제 해결에 있어 좋은 성능을 나타내어 왔다.

후보 해들로 이루어진 개체군의 확률 분포를 평가하는 것은 주어진 문제에 존재하는 변수들 사이의 관계를 파악하는 것으로 관계표현능력에 따라 여러 분포추정알고리즘이 있다[2]. 먼저 각 변수들이 독립이라고 가정하는 PBIL(Population Based Incremental Learning), cGA(compact Genetic Algorithm)과 UMDA(Univariate Marginal Distribution Algorithm) 등의 알고리즘이 있다[1,3,4]. 그러나 변수들간의 연관성이 있는 경우, 즉 변수들이 서로 독립이 아닌 경우에는 역시 올바른 해를 얻기에 어려움이 있다. 이에 대한 해결의 하나로 두 변수들 간의 상호작용을 고려한 알고리즘인 BMDA (Bivariate Marginal Distribution Algorithm) 등이 있다[5]. 또한 BOA(Bayesian Optimization Algorithm)를 통해 베이지안 네트워크로 개체군을 모델링함으로써 더 높은 차수의 상호작용이 있는 데이타를 표현할 수 있게 되었다[6].

그러나 변수들간에 상호작용이 있는 문제의 경우 그들의 결합 확률을 모델링하고 빌딩블럭을 깨지 않으며 관련있는 빌딩블럭이 무엇인지를 알아내는 일은 사전지식이 없으면 일반적인 알고리즘에서는 좋은 성능을 나타내기 어렵다. 따라서 주어진 문제에 존재하는 변수들간의 관계를 파악하기 위해 그들간의 관계를 확률그래프모델(probabilistic graphical model)로 구축하는 것이 필요하다. 여러 모델 중 그래프를 이용하여 변수들 간의 확률관계를 표현하는 것은 자연스러운 것으로 각 노드는 변수를, 각 에지는 변수들 간의 관계를 표현할 수 있어 그래프구조를 조사함으로써 변수들간의 연관성을 알아 낼 수 있다.

최근에 Zhang[7]은 하이퍼그래프 모델을 이용하여 다양한 기계학습분야의 문제를 해결하는데 좋은 방법론을 제시하였다. 또한 하이퍼그래프는 데이타 집합을 저장하는 확률적 메모리로 사용되어 얼굴, 디지털이미지 패턴 분류문제 등 거대한 문제공간에 대한 텍스트 분류 및 완성 문제, DNA 마이크로어레이 데이타 분석 문제 등에서 기존의 방법들과 견줄만한 좋은 결과를 가져왔다[8-11].

본 논문에서는 랜덤 하이퍼그래프 모델을 사용한 새로운 분포추정알고리즘을 제안하고자 한다. 주어진 문제의 변수들 간에 존재하는 임의의 차수의 상관성을 파악하기 위해 랜덤 하이퍼그래프모델을 구축하고 그로부터 하이퍼에지의 사후확률분포에 비례하여 다음 세대의 개체군을 베이지안 샘플링함으로써 변수들 간의 상관관계를 반영한 진화된 개체들을 생성하게 된다. 생성된 개체군에서 적합도가 우수한 개체들을 선택하고 그로부터 다시 하이퍼에지들을 생성하여 랜덤 하이퍼그래프 모델을 진화시켜 나가므로 원하는 최적해에 도달하는 방법을 제안하고자 한다. 본 논문의 구성은 다음과 같다. 2절에서는 분포추정 계열의 대표적 알고리즘 중의 하나인 BOA에 대해 살펴 본다. 3절에서는 랜덤 하이퍼그래프 모델을 구축하는 방법에 대해 기술하며 4절에서는 구축된 하이퍼그래프 모델로부터 베이지안 진화연산에 기반을 둔 베이지안 샘플링 알고리즘을 통해 다음 세대의 개체군을 구성하는 방법에 대해 제안한다. 5절에서는 변수들간의 상관관계 파악이 문제 해결에 핵심이 되는 기만적(deceptive) 문제에 대한 제안된 알고리즘의 실험방법과 결과를 보여주며 6절에서는 결론에 관해 언급한다.

2. 분포추정알고리즘

기존의 유전자알고리즘은 복잡한 문제에 대해서는 임의의 교차와 변이 연산에 의해 만들어진 새로운 탐색점들을 가지고 부모세대로부터 물려받은 좋은 부분 해, 즉 빌딩블럭을 키워나가는데 한계가 있다. 분포추정알고리즘(Estimation of Distribution Algorithms, EDA)은 이러한 결점을 극복하기 위해 개발되었으며 기존의 유전자알고리즘과는 다르게 교차와 변이 연산을 직접적으로 사용하지 않고 학습된 확률분포로부터 새로운 세대의 개체군을 구성하게 된다. 분포추정알고리즘의 개략적인 흐름도는 그림 1과 같다.

1. 개체군을 초기화한다.
2. 적합도가 우수한 개체들을 선택한다.
3. 선택된 개체들의 분포를 추정한다.
4. 추정된 분포로부터 다음 세대의 새로운 개체들을 샘플링한다.
   종료 조건 아니면 단계 2로 진행, 반복한다.

그림 1 분포추정알고리즘의 흐름도

그러므로 선택된 개체들의 확률분포를 잘 추정하는 것이 주요한 문제가 된다. 후보 해들로 이루어진 개체군의 확률 분포를 추정하는 것은 주어진 문제에 존재하는 변수들 사이의 관계를 파악하는 것으로 관계 표현 능력에 따라 여러 분포추정알고리즘이 있으며 다변수 데이타의 임의의 상관 관계를 나타내는 대표적인 알고리즘으로 BOA(Bayesian Optimization Algorithm)이 있다[6].

BOA에서는 선택된 개체들로부터 베이지안 네트워크를 구축하고 그로부터 새로운 개체군을 생성하게 된다. 비순환 방향 그래프(acyclic directed graph)인 베이지안 네트워크에서 각 노드는 하나의 변수에 해당한다. X_i는 변수 또는 변수에 해당하는 노드를 나타낸다고 하자. 변수들간의 의존관계는 방향이 있는 에지로서 표현한다. 베이지안 네트워크는 각 변수 X_i에 대한 지역확률분포의 곱으로 결합확률분포를 효율적으로 표현하는 확률그래프모델이다. $X = (X_1,...,X_n)$를 문제공간 변수들의 벡터라 할 때 결합확률분포는 다음과 같다.

$$p(X) = \prod_{i=1}^{n} p(X_i \mid \Pi_{X_i}),$$

여기서 각 변수 X_i에 대하여 Π_{X_i}는 베이지안 네트워크에서 X_i로 들어오는 방향의 에지를 가진 변수들의 집합을 말하며 Π_{X_i}집합의 각 구성원을 X_i의 부모노드라 부르고 X_i를 Π_{X_i} 집합의 구성원의 자식노드라 부른다. $p(X_i \mid \Pi_{X_i})$는 Π_{X_i}에 대한 X_i의 조건부확률이다.

선택된 개체들로부터 최적의 베이지안 네트워크를 구축하는 방법은 NP-complete한 문제로 한 노드로 들어오는 에지의 수를 제한하는 등의 제약조건을 두어 네트워크의 복잡도를 단순화하는 것이 필요하다. 네트워크가 데이타를 얼마나 잘 표현하느냐에 대한 측정값인 스코어링 메트릭(scoring metric)을 최대화하는 네트워크를 greedy한 방법으로 구축해나가는데 빈 네트워크로부터 출발하여 현재 네트워크의 스코어를 개선시키는 에지 추가(edge addition)와 같은 원시 그래프연산(graph primitive operation)을 적용해 나간다. 매 단계 네트워크는 사이클이 아니도록 하며 더 이상 스코어가 개선되지 않을 때 종료한다.

새로운 개체군이 생성되는 것은 베이지안 네트워크의 확률적 논리 샘플링을 통해 이루어진다. 먼저 ancestral 순서를 결정하는데 이는 부모노드들이 자식노드보다 앞서는 순서를 말하며 그 계산된 순서에 따라 새 후보 해를 구성하는 변수들의 값을 결정하게 된다. 한 변수의 값을 결정하려고 할 때 ancestral 순서가 앞선 그 변수의 부모노드들의 값은 이미 다 결정되어 있으므로 해당 변수값의 분포는 변수의 부모값들이 주어졌을 때 상응하는 조건부확률분포로부터 얻게 된다.

본 논문에서는 베이지안 네트워크와 비교하여 볼 때 변수들 간의 확률적 의존성(probabilistic dependence)은 표현하지 않지만, 하이퍼에지로 변수들간의 고차상관관계를 쉽게 표현하며 그를 반영하는 다음 세대의 개체를 바로 생성할 수 있는 하이퍼그래프 모델을 사용하여 새로운 분포추정알고리즘을 제안하고자 한다. 다음 3절에서는 랜덤 하이퍼그래프 모델에 대해 간략히 요약하고 이를 이용하여 어떻게 개체군의 변수들 간의 상관관계를 표현하는 하이퍼에지들을 구성할 지에 대해 살펴본다. 4절에서는 랜덤 하이퍼그래프 모델로부터 다음 세대의 개체군을 생성하는 베이지안 샘플링 방법에 대해 설명한다.

3. 하이퍼그래프를 이용한 분포추정

3.1 랜덤 하이퍼그래프 모델

그래프는 확률모델의 구조를 시각화하는데 적합한 구조 중 하나로 노드는 변수를, 에지는 변수들간의 관계를 나타내게 된다. 따라서 그래프 구조를 조사함으로써 변수들간의 상관성을 알아 낼 수 있다. 일반적인 그래프의 경우는 보통 2개의 노드를 하나의 에지로 연결하게 되는데 3개 이상의 노드를 하나의 에지로 연결하여 그들 간에 상관관계가 있음을 표시할 필요가 있다. 이러한 구조로 하이퍼그래프[12]가 있으며 노드들의 부분집합을 하이퍼에지로 갖게 된다.

하이퍼그래프 G는 n개의 노드들의 집합을 V, l개의 하이퍼에지들의 집합을 E라 할 때 $G=(V,E)$로 표시되며 $V = \{v_1,...,v_n\}$, $E = (E_1,...,E_l)$이다. 하이퍼에지의 차수(order)는 하이퍼에지를 구성하는 노드들의 개수를 말하며 차수 k의 하이퍼에지 $E_i = \{v_{i_1},...,v_{i_k}\}$를 k-하이퍼에지라고 하고, 아래첨자 $i_j \in \{1,...n\}$이다. 즉 차수 k의 하이퍼에지는 k개 변수들을 나타내는 노드들로 구성된 집합이라고 볼 수 있다. E 내에서 하이퍼에지는 중복하여 존재할 수 있으며 E_i의 개수를 $w(E_i)$라고 하자.

그림 2는 하이퍼그래프 $G = (V, E)$, $E = \{E_1,...,E_6\}$, $V = \{v_1,...,v_8\}$, $E = \{v_3,v_4,v_5\}$, $E_2 = \{v_5,v_8\}$, $E_3 = \{v_6,v_7,v_8\}$, $E_4 = \{v_2,v_3,v_7\}$, $E_5 = \{v_1,v_2\}$, $E_6 = \{v_7\}$인 경우의 예를 보인 것이다.

k-하이퍼에지의 가능한 수 $|E_k|$는 n개의 노드들의 집합에서 k개의 노드들을 선택하는 경우의 수가 된다.

$$|E_k| = {}_nC_k = \frac{n!}{k!(n-k)!}$$

상호작용이 있는 변수들을 같은 하이퍼에지에 있는 노드들로 나타냄으로써 그들간의 관계를 표현할 수 있다. 그러나 초기에는 변수들간의 상관성을 알지 못하므

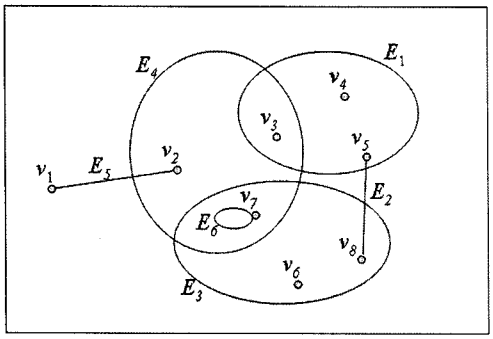

그림 2 하이퍼그래프의 표현[12]

로 모든 가능한 하이퍼에지들의 조합에 대해 고려해야 한다. 현실적으로 모든 경우의 변수들의 상관성을 고려하는 것은 불가능하므로 전체 $L \le \sum_{k=1}^{n} 2^k \cdot |E_k|$ 개의 하이퍼에지 중에서 실제로 우리가 고려하는 하이퍼에지들의 집합은 랜덤프로세스에 의해 선택한 작은 집합이라 할 수 있다. 전체 모든 가능한 하이퍼에지들 중에서 우리가 고려하는 랜덤 하이퍼그래프 모델을 구성하는 하이퍼에지들을 랜덤프로세스에 의해 생성하는 방법은 다음과 같다.

데이타 $\mathbf{x} = (x_1,...,x_n)$ 가 n 개의 변수들로 이루어져 있을 때 k 개의 변수들로 이루어지는 k -하이퍼에지 E_h를 그림 3과 같이 랜덤프로세스에 의해 구성한다.

$E_h = \phi$
for $j = 1$ *to* k
  $i = random(n)$
  $E_h = E_h \cup x_i$
end
return(E_h)
// *random*(n)은 동일한 확률로 이미 선택된 인덱스가 아닌
// 1부터 n 사이의 인덱스 중 하나를 선택

그림 3 하이퍼에지를 구성하는 랜덤프로세스

따라서 변수들의 부분집합으로 구성된 각 하이퍼에지들에 대하여 데이타(개체군의 개체들)와의 가능도(likelihood)를 계산하고, 사전확률분포(전 단계의 사후확률분포)를 반영하여 사후확률분포를 계산한 후 그에 따라 하이퍼에지들을 베이지안 샘플링해나감으로써 문제공간 변수 간의 고차상관관계를 표현하는 진화된 다음 세대의 개체들을 생성하게 된다.

베이지안 네트워크가 방향이 있는 에지로 변수들 간의 확률적 의존성을 표현하는데 반해 하이퍼그래프는 연관성이 있을 것으로 예측되는 변수들로 하이퍼에지를 구성하여 조건부확률의 표현 없이 구성 변수들의 집합이 빌딩블럭으로 유지되어야 함을 표시한다. 이 때 각 변수(노드로 표시됨)들은 그림 2와 같이 서로 다른 여러 하이퍼에지에 나타날 수도 있다. 또한 하나의 하이퍼에지를 구성하는 변수들은 적합도가 우수한 선택된 개체(데이타)들로부터 결정되므로 하이퍼에지를 구성하는 변수들의 상관성이 높을 경우에는 해당 하이퍼에지가 여러 번 중복하여 생성될 수 있다. 두 경우 모두, 최적의 모델을 구성하는 것은 NP-complete한 문제라 베이지안 네트워크는 데이타를 잘 표현하는 스코어를 개선시키는 방향으로 greedy한 방법으로 구축해나가며, 하이퍼그래프모델은 데이타로부터 랜덤하게 생성한 하이퍼에지들로 구축하게 된다. 본 논문에서 고려하는 랜덤 하이퍼그래프 모델은 조건부확률은 표현할 수 없지만, 베이지안 네트워크의 경우와 같은 차수의 제한을 두지 않고 별도의 계산 없이 최대 n의 고차상관관계를 하이퍼에지로 쉽게 표현할 수 있다. 또한 작은 차수의 하이퍼에지들을 결합하여 고차의 하이퍼에지를 구축할 수 있는 장점이 있다.

3.2 하이퍼그래프 분포

생성된 랜덤 하이퍼그래프 모델을 통한 변수들의 확률분포 추정을 위해 n 개의 노드, $\{v_1,v_2,...,v_n\}$을 가진 하이퍼그래프의 n 개 노드들간의 결합확률분포를 다음과 같이 나타내기로 한다.

$$p(v_1, v_2, \ldots, v_n) = \prod_{i=1}^{n} \sum_{j=1}^{l} 1_{(v_i \in E_j)} p(E_j) \ , \tag{1}$$

여기서 하이퍼에지 E_j는 노드들의 부분집합이며 l은 하이퍼그래프를 구성하는 하이퍼에지의 총 가지 수이다. 먼저 노드 v_i가 l개의 각 하이퍼에지에 포함되는 지의 여부, $v_i \in E_j$, 를 아래와 계산한다.

$$1_{(v_i \in E_j)} = \begin{cases} 1 & if \ \ v_i \in E_j \\ 0 & otherwise \end{cases} .$$

만약 노드 v_i가 E_j의 원소가 되는 경우에는, 식 (1)의 $p(E_j)$는 현재 하이퍼그래프를 구성하는 전체 하이퍼에지들의 개수 중 E_j가 몇 개나 존재하는 지의 비율로 다음과 같이 계산한다.

$$p(E_j) = \frac{w(E_j)}{\sum_{i=1}^{l} w(E_i)} ,$$

$$\sum_{j=1}^{l} p(E_j) = 1, \ p(E_j) \ge 0,$$

여기서 $w(E_j)$는 하이퍼에지 E_j의 개수를 말한다.

다음 4절에서는 진화연산을 베이지안 추론(Bayesian

inference)에 기반한 확률과정으로 모델링하는 베이지안 진화연산(Bayesian evolutionary computation)[13,14]에 기초하여 위와 같이 구성한 랜덤 하이퍼그래프 모델로부터 그를 구성하는 하이퍼에지들의 사후확률분포를 구하고 그 확률에 비례하여 다음 세대의 개체들을 생성하는 베이지안 샘플링 알고리즘에 대해 제안한다.

4. 랜덤 하이퍼그래프 모델의 진화를 위한 베이지안 샘플링 알고리즘

하이퍼그래프 모델의 진화과정을 베이지안 진화연산에 기반하여 다음과 같이 고려할 수 있다. k-하이퍼에지 $E_h = \{x_{h_1}, x_{h_2}, ..., x_{h_k}\}$에 대하여 $p(E_h)$를 사전(prior) 확률분포라 하고 훈련 데이타의 집합 $D = \{\mathbf{d}_1, \mathbf{d}_2, ..., \mathbf{d}_N\}$를 관찰한 후 이 데이타에 대한 가능도를 $p(D|E_h)$라 할 때 사후(posterior)확률분포 $p(E_h|D)$는 다음과 같다.

$$p(E_h \mid D) = \frac{p(D \mid E_h)p(E_h)}{p(D)}, \tag{2}$$

여기서 $p(D)$는 정규화 상수(normalizing constant)로 다음과 같이 정의된다.

$$p(D) = \sum_{E_j} p(D \mid E_j)p(E_j).$$

따라서 식 (2)는 다음과 같이 나타낼 수 있다.

$$p(E_h \mid D) \propto p(D \mid E_h)p(E_h). \tag{3}$$

식 (3)의 사전확률분포 $p(E_h)$는 문제에 대한 사전지식을 고려하여 정의되며 특별한 사전지식이 없는 경우에는 균등분포를 가정한다. 식 (3)의 가능도 $p(D|E_h)$는 하이퍼에지 E_h를 구성하는 변수값들로 이루어진 빌딩블럭을 갖는 데이타가 몇 개나 되는 지에 대한 값으로 빌딩블럭의 유용성에 대한 증거가 되며 다음과 같이 계산된다.

$$p(D \mid E_h) = \sum_{i=1}^{N} \prod_{j=1}^{k} \delta_j(\mathbf{d}_i), \tag{4}$$

$$\delta_j(\mathbf{d}_i) = \begin{cases} 1 & if \;\; x_{h_j} = d_{h_j}^{(i)} \\ 0 & otherwise \end{cases},$$

여기서 $h_j = \{1,...n\}$이며 $\delta_j(\mathbf{d}_i)$는 i 번째 데이타인 $\mathbf{d}_i = (d_1^{(i)},...,d_n^{(i)})$에 대하여 E_h를 구성하는 각 변수의 값과 데이타 $\mathbf{d}_i$의 같은 인덱스를 갖는 변수의 값이 일치하는지의 여부를 결정한다. 식 (4)의 $\prod_{j=1}^{k} \delta_j(d)$의 값이 1이면 하이퍼에지의 변수값들과 $\mathbf{d}_i$의 해당 인덱스를 갖는 변수의 값들이 모두 일치함을 나타낸다. 따라서 하이퍼에지를 구성하는 변수들의 상관성이 있다는 증거가 되므로 초기값이 0인 가능도를 1 증가시키게 된다. 이런 과정을 훈련 데이타의 집합(개체군) $D = \{\mathbf{d}_1, \mathbf{d}_2, ..., \mathbf{d}_N\}$의 모든 데이타에 대하여 반복하여 하이퍼에지 E_h의 최종 가능도를 얻게 된다.

현재 세대의 개체군으로부터 학습된 랜덤 하이퍼그래프의 사후확률분포는 다음 세대를 생성하기 위한 사전확률분포가 되어 알고리즘이 반복되며 매 세대 진화된 개체들을 생성하게 된다. 그림 4는 본 논문에서 새롭게 제안하는 분포추정알고리즘으로 랜덤 하이퍼그래프를 확률모델로 사용하여 분포추정을 수행하는 베이지안 샘플링 알고리즘(Bayesian Sampling Algorithm)의 개략적인 흐름도를 보인다.

(랜덤 하이퍼그래프를 이용한 베이지안 샘플링 알고리즘)

1. $Pop(1)$를 N개의 개체들로 구성된 초기 해 집단이라 하자. $t=1$.
2. N개의 개체들의 적합도를 평가한 후 우수한 상위 M개의 개체들을 선택한다.
3. 선택된 M개의 개체들로부터 랜덤 하이퍼그래프 모델을 구축한다.
4. $t=t+1$, 얻어진 랜덤 하이퍼그래프 모델로부터 베이지안 샘플링을 통해 다음 세대 $Pop(t)$를 생성한다.
5. 종료 조건이 아니면 단계 2로 진행, 반복한다.

그림 4 랜덤 하이퍼그래프 모델 진화를 이용한 베이지안 샘플링 알고리즘의 흐름도

4.1 개체군으로부터 랜덤 하이퍼그래프 모델 구축

개체군의 집합을 $X = \{X_1,...,X_N\}$, 여기서 각 개체를 $X_i = \{x_1^{(i)}, ..., x_n^{(i)}\}$라고 하자. 즉 개체군의 크기는 N, 각 개체는 n 개의 변수값들로 이루어져 있다. 먼저 주어진 절단 임계치(truncation threshold) τ가 $0<\tau<1$일 때 적합도가 우수한 $M = \tau \times N$ 개체들이 선택된다.

변수들의 상관관계를 나타내는 하이퍼에지로 개체군의 분포를 추정하기 위하여 선택된 M개의 개체들로부터 상관관계가 있을 것으로 예측되는 k개 변수들의 조합으로 구성되는 총 L개의 하이퍼에지들을 생성하여 그림 5와 같이 하이퍼그래프 G를 구축하게 된다.

```
G = φ
X = {X_1, X_2, ..., X_M}
for i = 1 to M
  choose X_i = (x_1^(i), ..., x_n^(i)) ∈ X
  for q = 1 to rate
    E_m = φ
    for j = 1 to k
      h = random(n)
      E_m = E_m ∪ x_h^(i)
    end
    G = G ∪ E_m
  end
end
return(G)

// 총 하이퍼에지의 수 L = M × rate
// k는 E_m의 차수로 1에서 n 사이에서 특정 값으로 고정,
   또는 매 반복 시 랜덤하게 지정가능
```

그림 5 랜덤 하이퍼그래프 모델을 구축하는 프로세스

4.2 베이지안 샘플링을 통한 새로운 개체군 생성

구축된 랜덤 하이퍼그래프 모델로부터 다음 세대의 개체군이 되는 N개의 개체들을 생성하기 위한 베이지안 샘플링 방법은 다음과 같다. 사후확률분포가 우수한 하이퍼에지에 대하여 우리는 해당 하이퍼에지를 구성하는 변수들간에 상관성이 높다는 것을 알 수 있으므로 그 변수 값들의 조합을 그대로 유지하는 개체를 생성한다. 이 때 E_h로부터 베이지안 샘플링되는 개체의 수 m은 E_h의 사후확률분포에 비례한다.

$$m = N \times \frac{P(E_h \mid D)}{\sum_{j=1}^{l} P(E_j \mid D)} .$$

3-하이퍼에지 $E_h = \{x_p, x_q, x_r\}$로부터 베이지안 샘플링되는 m개의 다음 세대 개체들은 그림 6과 같다.

$$X_1 = (x_1^{(1)}, \cdots, x_p, \cdots, x_q, \cdots, x_r, \cdots, x_n^{(1)})$$

$$\vdots$$

$$X_m = (x_1^{(m)}, \cdots, x_p, \cdots, x_q, \cdots, x_r, \cdots, x_n^{(m)})$$

그림 6 베이지안 샘플링 방법을 통한 개체 생성

E_h를 구성하는 변수의 값 x_p, x_q, x_r을 그대로 고정시켜 샘플링하므로써 학습된 변수들의 상관관계를 나타내는 빌딩블럭을 그대로 유지하게 한다. 이 때 E_h를 구성하지 않는 나머지 변수들의 값은 랜덤하게 샘플링하여 개체군의 다양성을 유지하도록 한다. 각 하이퍼에지 E_h에 대하여 E_h의 사후확률분포에 비례하는 개수의 개체들을 생성해 나감으로써 N개의 개체들을 구성하여 다음 세대의 개체군을 이룬다.

하이퍼에지의 사후확률분포에 비례하여 다음 세대의 개체들을 샘플링하므로, 만약 사전확률분포가 같을 경우 가능도가 큰 영역에서 많은 수의 개체들이 얻어진다. 베이지안 진화연산의 관점에서 보면 현재 세대에서 구축된 랜덤 하이퍼그래프 모델이 다음 세대의 모델을 구축할 때 시작점이 되어 사전확률분포를 나타내는 모델로서 사용된다. 실제로 우리가 고려하는 D는 N개의 개체들로 이루어진 개체군이 되며 세대가 진행됨에 따라 주어진 데이타(개체군)를 구성하는 변수들의 상관관계를 잘 반영하는 랜덤 하이퍼그래프 모델로 진화해 나간다.

따라서 진화된 모델로부터 베이지안 샘플링 방법을 통해 변수들의 상관관계를 반영한 다음 세대를 구성하는 진화된 개체들을 생성하게 되므로 개체군은 점차적으로 개선되고 다양성은 점차적으로 감소하여 마침내 최적해에 수렴하게 된다.

5. 실험 및 결과

5.1 함수최적화

본 논문에서 제안한 알고리즘의 성능을 평가하기 위하여 함수 최적화 문제로 잘 알려진 유니트함수(unitation)에 대하여 실험, 비교하였다. 유니트함수는 이진 입력스트링이 갖고 있는 1의 개수에 의존하는 함수 값을 갖게 되며 이러한 유니트함수들이 더해져서 좀 더 복잡한 함수가 될 수 있다. 즉 함수 f_k를 길이가 k인 스트링에서 정의된 유니트함수라고 할 때 이러한 함수 l개를 결합하여 다음과 같은 함수 f를 구성한다.

$$f(X) = \sum_{i=0}^{l-1} f_k(S_i) ,$$

여기서 X는 n개 변수들의 집합이며 S_i는 X의 k개 변수들로 이루어진 부분 집합이다. 또한 복잡한 함수를 더 작은 차수의, 즉 변수들의 부분 집합으로 구성되는 더 간단한 함수들의 합으로 표시할 수 있을 때 원래의 함수를 덧셈으로 분해가능(additively decomposable)하다고 하며 이 때 복잡한 문제를 더 작은 부분 문제들로 분할하여 풀 수 있게 된다.

실험은 단순 유니트함수가 아닌 기만(deceptive)함수에 대해 수행하였다. 기만현상(*deception*)[15]은 탐색에 의한 최적화 알고리즘의 성능에 대한 연구를 통해 잘 알려진 현상으로, 기만함수는 최적해(global optimum)가 아닌 그릇된 지역해(local optimum)로 알고리즘을 유도하게 되는데 그 이유는 최적해가 비 관심 지역에 위치한 고립된 최고점이기 때문이다.

차수 3을 갖는 기만함수인 *3-deceptive*는 다음과 같이 정의된다. 여기서 u는 입력 스트링에서 1의 개수이다.

$$f_{3-deceptive}(u) = \begin{cases} 0.9 & if\ u = 0 \\ 0.8 & if\ u = 1 \\ 0 & if\ u = 2 \\ 1 & if\ u = 3 \end{cases} .$$

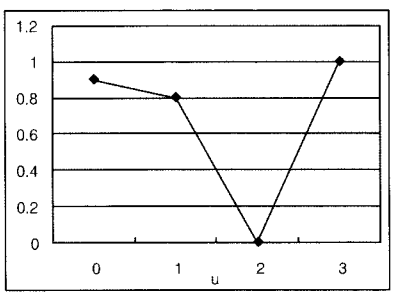

그림 7 *3-deceptive* 함수

3-deceptive 함수는 한 개체를 구성하는 3비트의 입력 스트링에 적용되며 전체 문제의 크기만큼 더해져 전체 적합도를 계산하게 된다. 따라서 크기 n의 문제에 대해 n비트 모두 1로 구성되는 하나의 최적해를 갖게 된다. 그러므로 올바른 해를 얻기 위하여 각 분할 내에의 위치들간의 상관성을 고려하여야 하며 그렇지 않고 비트를 독립적으로 고려할 경우 최적해가 아닌 지역해에 잘못 이르게 된다.

차수가 $k \geq 3$인 트랩(*trap*)함수는 완전 기만적(fully-deceptive)으로 차수 k보다 작은 어떠한 통계자료(statistics)도 최적해에 이르지 못하게 한다[16]. 따라서 각 트랩함수에 속하는 변수들끼리는 빌딩블럭을 형성하게 되고 확률모델을 구축할 때 함께 처리되어야 한다. 그러므로 문제에 대한 정보가 주어지지 않았을 때에는 변수들이 랜덤하게 분포되어 있거나 가깝게 밀집되어 있거나 똑같이 어려운 문제가 된다. 차수가 5인 트랩함수는 앞의 *3-deceptive*과 비교하여 5비트 그룹을 고려하게 된다. 각 그룹의 비트들은 함께 처리되며 그렇지 않을 경우 잘못된 결과를 얻게 된다. *trap-5*는 다음과 같이 정의된다.

$$f_{trap-5}(u) = \begin{cases} 4-u & if\ u < 5 \\ 5 & if\ u = 5 \end{cases}.$$

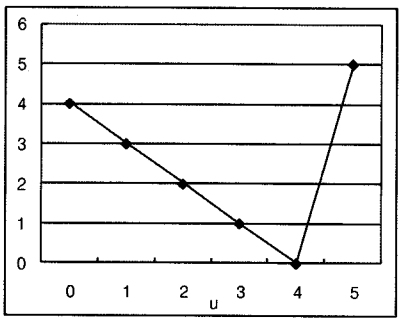

그림 8 *trap*-5 함수

5.2 실험 결과

모든 문제에 대해서 개체군이 수렴할 때까지 10번의 독립적인 실행의 평균값으로 적합도 평가 회수를 구하였다. 개체군은 각 비트위치에서 어떤 값이 전체 개체군 크기의 95%에 이를 때 수렴한다고 정하였다. 그리고 모든 실험에서 절단 임계치 $\tau = 50\%$로 하여 적합도가 우수한 순으로 절반의 개체들을 절단 선택(truncation selection)하였다[17]. 이전 세대의 하이퍼에지들을 갱신하기 위하여 사전확률분포가 높은 하이퍼에지 순으로 적합도가 높은 순서대로 선택된 각 개체로부터 임의의 변수값의 조합을 추가하여 하이퍼에지들을 갱신, 베이지안 샘플링을 위한 랜덤 하이퍼그래프 모델을 구축하였다. BOA(Bayesian Optimization Algorithm)와의 성능 비교를 위해 BOA와 단순 유전자알고리즘(Simple Genetic Algorithm, 이하 SGA)에 대한 실험 결과는 [6]으로부터 인용하였다.

deceptive-3 함수에 대하여 베이지안 샘플링 알고리즘(이하 BSA)의 경우, 문제의 크기, 즉 개체를 구성하는 변수의 개수가 $n = 15$에서 $n = 180$인 경우 개체군의 크기를 $N = 100$에서 $N = 3000$으로 실험하였으며 하이퍼에지의 차수는 랜덤하게 지정하였다. 개체군이 최적해에 수렴할 때까지의 평균 적합도 평가 회수에 대한 결과가 표 1과 그림 9에 있다.

표 2와 그림 10은 *trap*-5 함수에 대한 결과이다. 적합도 평가 회수는 세대 수와 개체 수의 곱이며 이 때 BSA는 앞에서와 같이 변수의 개수가 $n = 15$부터 $n = 180$인 경우 개체군의 크기를 $N = 100$에서 $N = 3000$으로 실험하였고 하이퍼에지의 차수는 랜덤하게 지정하였다.

실험을 통해 문제의 크기가 커짐에 따라 빌딩블럭이 기만적이므로 일점교차와 변이연산(변이율=1%)을 사용한 SGA의 복잡도가 지수적으로 증가하였으나 BSA는 선형의 완만한 증가도를 갖는다. 따라서 문제와는 독립되고 일점교차와 변이연산과 같은 고정된 변형(variation) 연산자를 사용하는 SGA와 비교하여 우리가 제안하는 알고리즘은 훨씬 낮은 복잡도를 갖고 기만적 덧셈 분할 가능한(deceptive additively decomposable)

표 1 *deceptive*-3 함수에 대한 BSA의 결과

변수의 개수 (n)	개체군의 크기 (N)	세대수 평균±표준편차	평균 적합도 평가회수
15	100	13.6±1.82	1360
30	100	16.8±1.92	1680
45	1000	22.6±2.07	22600
90	1500	31±3.39	46500
120	2000	31.2±3.63	62400
150	2500	33.2±3.71	83000
180	3000	33.8±3.87	101400

표 2 *trap*-5 함수에 대한 BSA의 결과

변수의 개수 (n)	개체군의 크기 (N)	세대수 평균±표준편차	평균 적합도 평가회수
15	100	12.6±1.52	1260
30	100	18.0±1.58	1800
45	1000	24.2±2.39	24200
90	1500	31.2±3.11	46800
120	2000	36.8±3.27	73600
150	2500	38.2±2.71	95500
180	3000	42.6±3.61	127800

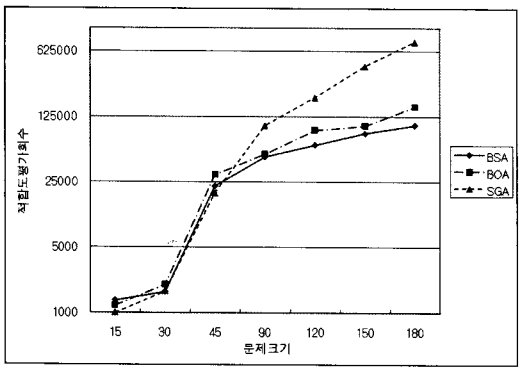

그림 9 *3-deceptive* 함수에 대한 결과

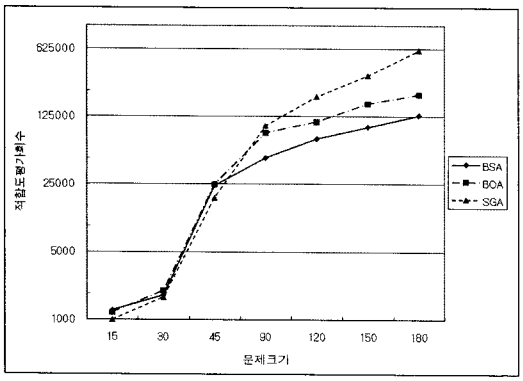

그림 10 *trap-5* 함수에 대한 결과

문제를 풀 수 있음을 알 수 있다. 또한 BOA와 비교하여 문제의 크기가 작은 경우에는 비슷한 성능을 가지나 문제의 크기가 커질수록 BOA보다 적은 적합도 평가회수를 얻을 수 있었다. 또한 랜덤 하이퍼그래프 모델을 구축할 때 하이퍼에지의 차수를 고정시켰을 때 보다 랜덤하게 결정하였을 경우가 더 좋은 성능을 보임을 알 수 있었는데 이는 좀 더 빠른 시간 안에 더 큰 빌딩블럭을 형성할 수 있기 때문으로 보인다.

이러한 결과들로부터 문제공간의 임의의 고차 상관관계를 랜덤 하이퍼그래프 모델을 통해 잘 학습할 수 있으며 이를 통해 올바른 최적해에 도달할 수 있음을 확인할 수 있었다.

6. 결 론

본 논문에서는 선택 과정을 통해 살아남은 개체들을 구성하는 변수들의 가능한 경우의 결합분포를 평가할 수 있도록 에지를 구성하는 노드의 개수를 2개 이하로만 제한하지 않는 하이퍼그래프 구조를 사용하였다. 그리하여 상관관계가 있을 것으로 추정되는 변수들을 하이퍼에지를 구성하는 노드들로 나타낼 수 있었고 주어진 문제의 변수들 간에 존재하는 임의의 차수의 상관성을 파악하기 위한 랜덤 하이퍼그래프 모델을 구축하여 베이지안 샘플링 방법을 통해 개체군을 진화시켜 나가며 최적해를 찾아가는 새로운 분포추정 알고리즘을 제안하였다.

BOA 등과 같은 경우, 모델을 구축하기 위해 greedy한 방법을 사용해도 적지 않은 시간이 소요되며 한 노드로 들어오는 에지의 수를 제한하는 등의 제약조건을 두어 네트워크의 복잡도를 단순화하게 된다. 본 논문에서 제안한 방법에서는 하이퍼에지를 구성하는 변수들을 랜덤하게 선택함으로써 임의의 상관관계를 학습할 수 있으며 우리가 고려할 거대한 전체 문제 공간과 비교하여 상당히 작은 집합인 랜덤 하이퍼그래프 만을 유지하면서도 랜덤하게 생성한 하이퍼에지들을 통해 개체군내에 어떠한 변수들의 상관관계가 유력한 지를 평가하고 그 모델로부터 베이지안 샘플링을 통해 결합확률분포를 반영한 개체들을 생성하게 된다. 실험을 통해 단순 유전자알고리즘이 해결하기 어려운 기만함수 문제에 대하여 올바른 결과를 얻을 수 있음을 보였고 BOA보다 좋은 성능을 나타내었다.

참 고 문 헌

[1] Mühlenbein, H. & Paaβ, G.., "From recombination of genes to the estimation of distributions I. Binary parameters," *Parallel Problem Solving from Nature*, pp. 178-187, 1996.

[2] Larranaga, P., "A review on estimation of distribution algorithms," *Estimation of Distribution Algoruthms: A New Tools for Evolutionary Computation*, Kluwer Academic Publisher, pp. 57-100, 2001.

[3] Baluja, S., "Population-based incremental learning: A method for integrating genetic search based function optimization and competitive learning," Tech. Rep. No.CMU_CS_94_163, Pittsburgh, PA: Carnegie Mellon University, 1994.

[4] Harik, G. R., Lobo, F. G., & Goldberg, D. E., "The compact genetic algorithm," IlliGAL Report No.97006, Urbana, IL: University Of Illinois at Urbana Champaign, 1997.

[5] Pelikan, M., & Mühlenbein, H., "The bivariate marginal distribution algorithm," *Advances in Soft Computing-Engineering Design and Manufacturing*, pp. 521-535, 1999.

[6] Pelikan, M., Goldberg, D. E., & Cantu-Paz, E., "BOA: The Bayesian optimization algorithm," *Genetic and Evolutionary Computation Conference GECCO-99*, Vol.1, pp. 525-532, 1999.

[7] B.-T. Zhang, "Random hypergraph models of learning and memory in biomolecular networks: shorter - term adaptability vs longer-term persis-

tency," *IEEE Symposium on Foundations of Computational Intelligence*, pp. 344-349, 2007.
[8] B.-T. Zhang & J.-K. Kim, "DNA hypernetworks for information storage and retrieval," *Lecture Notes in Computer Science*, DNA12, 4287:298-307, 2006.
[9] S. Kim, M.-O. Heo & B.-T. Zhang, "Text classifier evolved on a simulated DNA computer," *IEEE Congress on Evolutionary Computation(CEC2006)*, pp. 9196-9202, 2006.
[10] C.-H. Park, S.-J. Kim, S. Kim, D.-Y. Cho & B.-T. Zhang, "Finding cancer related gene combinations using a molecular evolutionary algorithm," *IEEE 7th international conference on Bioinformations & Bioengineering (BIBE2007)*, pp. 158-163, 2007.
[11] B.-T. Zhang, "Hypernetworks: A molecular evolutionary architecture for cognitive learning and memory," *IEEE Computational Intelligence Magazine*, 3(3):49-63, 2008.
[12] Berge, C., *Hypergraphs*, North-Holland. 1989.
[13] B.-T. Zhang, "A Bayesian framework for evolutionary computation," *IEEE Congress on Evolutionary Computation(CEC1999)*, pp. 722-728, 1999.
[14] B.-T. Zhang, "A unified Bayesian framework for evolutionary learning and optimization," *Advances in Evolutionary Computing*, Springer-Verlag, Chap15, pp. 393-412, 2003.
[15] Goldberg, D. E., "Simple genetic algorithms and the minimal deceptive problem," *Genetic Algorithms and Simulated Annealing*, Pitman, pp. 74-88, 1987.
[16] Deb, K. & Goldberg, D. E. "Analyzing deception in trap functions," *Proceedings of Foundations of Genetic Algorithms FOGA-II*, pp. 93-108, 1993.
[17] Claudio F. Lima, Pelikan, M., Goldberg, D. E., & Fernando G. Lobo, Kumara Sastry, and Mark Hauschild, "Influence of selection and replacement strategies on linkage learning in BOA.," *IEEE International Conference on Evolutionary Computation (CEC2007)*, pp. 1083-1090, 2007.

이 시 은

1991년 서울대학교 컴퓨터공학 학사. 1997년 서울대학교 컴퓨터공학 석사. 1999년~현재 서울대학교 컴퓨터공학부 박사과정. 1998년~2005년 백석문화대학 컴퓨터정보학부 조교수. 2006년~현재 백석대학교 정보통신학부 조교수. 관심분야는 진화연산, 기계학습, 생물정보학, 데이타마이닝

이 인 희

2001년 2월 서울대학교 컴퓨터공학부 학사. 2001년 3월~현재 서울대학교 컴퓨터공학부 석박사 통합과정. 관심분야는 진화연산, 생물정보학, 기계학습, DNA 컴퓨팅

장 병 탁

1986년 서울대학교 컴퓨터공학 학사. 1988년 서울대학교 컴퓨터공학 석사. 1992년 독일 Bonn대학교 컴퓨터공학 박사. 1992년~1995년 독일국립정보기술연구소(GMD) 연구원. 1995년~1997년 건국대학교 컴퓨터공학과 조교수. 1997년~현재 서울대학교 컴퓨터공학부 교수, 인지과학, 뇌과학, 생물정보학 협동과정 겸임. 2001년~현재 서울대학교 바이오정보기술연구센터 센터장. 관심분야는 Biointelligence, Probabilistic Models of Learning and Evolution, Molecular/DNA Computation