MDL원리에 기반한 모델 선택을 포함한 분자 wDNF기계에서의 분자 EDA

이시은° 장병탁
백석대학교 정보통신학부 서울대학교 컴퓨터공학부
selee@bu.ac.kr btzhang@snu.ac.kr

Molecular EDA with model selection based on MDL principle in molecular wDNF machine

Si Eun Lee°, Byoung-Tak Zhang
Baekseok University Seoul National University

요 약

분자 wDNF기계를 통해 해 집단을 병렬적으로 탐색하여 유망한 텀들을 선택한 후 그를 구성하는 변수들의 분포를 평가, 확률 모델을 확립하고 그로부터 다음 세대의 해 집단을 구성함으로써 진화 알고리즘의 확장인 EDA을 DNA컴퓨팅으로 모델링한다. 또한 희박한(sparse) 해 집단에서 간략한 (parsimonious) wDNF 모델을 함께 찾으므로 단순히 해 집단의 분포만을 진화시켜 나가는 것이 아니라 모델의 구조도 같이 최적화 시켜 나가는 방안을 제시한다.

1. 서 론

진화알고리즘은 인공적인 선택과 유전적 재결합, 변이 연산자에 기반한 최적화 알고리즘으로 문제의 부분적 해답이 될 수 있는 빌딩 블럭들을 잘 키워나가 해에 도달하게 한다. 그런데 이러한 빌딩 블럭들이 재결합 등의 연산자에 의해 유지되지 않는 경우의 문제 해결을 위해 Estimation of Distribution Algorithm(이하 EDA)계열의 알고리즘들이 제안되었다[1]. 즉 직접적인 교차나 변이 연산이 없이 유망한 해 집단의 분포를 파악하여 평가된 분포에 따라 다음 세대의 해 집단을 생성한다. 그런데 실제로 해 집단의 분포를 평가하는 것은 쉽지 않은 일이라 각 변수들이 독립이라는 가정을 가지고 진행하는 경우가 많다. PBIL, cGA 와 UMDA 등의 알고리즘이 그 것이다 [2,3]. 그러나 변수들간의 연관성이 있는 경우, 즉 변수들이 서로 독립이 아닌 경우에는 역시 올바른 해를 얻기에 어려움이 있다. 이에 대한 해결로 두 변수들 간의 상호작용을 고려하는 알고리즘으로 BMDA 등이 있다[4]. 또한 BOA를 통해 베이지안 네트워크로 데이터를 모델링하므로써 더 높은 차수의 상호작용이 있는 데이터를 표현할 수 있게 되었다[5]. 그러나 변수들간에 상호작용이 있는 문제의 경우 그들의 결합 확률을 모델링하고 올바른 답을 찾아내는 문제 해결에 아직도 계산 시간 및 저장능력 등에 있어서 상당한 어려움이 있다고 할 수 있다.

따라서 본 논문에서는 EDA의 구현을 기존의 이진수 0,1에 기반하는 실리콘 컴퓨팅에 반해 바이오 분자들의 정보 저장 및 처리 특성을 이용한 차세대 컴퓨팅 기술 중 하나인 DNA컴퓨팅 관점에서 살펴보고자 한다. 본 논문의 구성은 다음과 같다. 2장에서는 이전 연구들을 통해 분자 wDNF (molecular weighted Disjunctive Normal Form) 기계에 대해 정리해 본다. 3장에서는 분자 wDNF기계로 해 집단의 진화를 표현하여 이를 EDA로 모델링한다. 4장에서는 단순히 wDNF로 표현된 해 집단 만을 진화시켜 나가는 것이 아니라 MDL원리에 기반하여 적은 개수의 변수들로 구성된 텀을 선호하는 요소를 포함하는 적합도 함수를 고려하여 정확도와 함께 모델의 구조도 같이 최적화시켜 나가는 방안에 대해 제시한다. 5장에서는 결론 및 향후 과제에 관해 언급한다.

2. wDNF 학습을 위한 DNA컴퓨팅

DNA 컴퓨팅은 A, T, G, C의 네 개의 뉴클레오티드로 정보를 표현한다. 이들은 용액상에서의 화학 반응에 의해 초 병렬적인 정보처리를 가능하게 하여 기존의 실리콘컴퓨팅에서는 처리하기 어려운 방대한 양의 계산도 가능하게 한다. DNA컴퓨팅의 기본 원리는 다음과 같다. 풀고자 하는 문제에 대한 가능한 해답을 DNA코드로 표현한 후 이들 코드를 올리고 합성 기술을 이용하여 다량 합성한다. 각각의 성분 분자들을 합성기에 넣고 화학반응을 시킴으로써 가능한 모든 해를 생성한다. 생성된 분자들 중에 찾는 해가 포함이 되었는지를 검사하여 답을 제시하게 된다. 이러한 DNA컴퓨팅을 wDNF기계의 학습에 사용할 수 있었다[6].

x_i 를 0 또는 1의 값을 갖는 애트리뷰트 또는 부울 변수라 할 때 먼저 텀(term) C_i 를 이들 변수들의 곱(conjunction) 이라고 하자.

$$C_i = (x_{i_1}, x_{i_2}, \cdots, x_{i_k}, \cdots, x_{i_n}) = x_{i_1} x_{i_2} \cdots x_{i_k} \cdots x_{i_n} ,$$

여기서 $x_{ik} \in \{x_1, x_2, \cdots, x_n\}$.

또한 DNF (disjunctive normal form) 을 텀들의 합 (disjunction)이라고 정의하자.

$$DNF = \{C_1, C_2, \cdots, C_j, \cdots, C_N\} = C_1 + C_2 + \cdots + C_j + \cdots + C_N$$

wDNF는 텀 C_i 에서 그를 구성하는 임의의 변수 x_i 의 r 승을 허용하고 또한 C_i 가 w_i 개 있을 수 있음을 허용하여 DNF를 일반화한 것이다.

$$wDNF = \{w_1 C_1, w_2 C_2, \cdots, w_j C_j, \cdots, w_N C_N\}$$
$$= w_1 C_1 + w_2 C_2 + \cdots + w_j C_j + \cdots + w_N C_N$$

훈련 데이터 집합 D 를 가지고 wDNF를 학습해 나가는 과정은 다음과 같다.

$$D = \{(\mathrm{x}_i, y_i)\}_{i=1}^{K},\ \mathrm{x}_i = (x_{i_1}, x_{i_2}, \ldots, x_{i_n}) \in \{0,1\}^n,$$
$$y_i \in \{0,1\}$$

먼저 변수들의 임의의 곱으로 구성된 텀들을 DNA 코드로 표현하여 DNA 분자 라이브러리를 초기화한다. 주어진 쿼리 데이터 x_q 에 대하여 라이브러리 내에서 그와 일치하는 분자들을 검출한다. 즉 쿼리 데이터 x_q 를 그를 구성하는 각 변수들로 나눈 후 각각 다수의 복사본들을 생성한다. 라이브러리 내의 wDNF와 변수 복사본들 간의 상보결합에 의한 이중가닥 형성을 시도하여 데이터의 y 값과 일치되는 텀들을 검출하여 증폭하고 그렇지 않은 텀들은 제거함으로써 라이브러리가 데이터 집합을 적합하게 표현하며 학습이 되어간다.

3. 분자 wDNF 기계를 이용한 분자 EDA

$X = \{x_1, x_2, \cdots, x_n\}$ 을 진화알고리즘에서 고려하는 문제 공간 내 변수들의 집합이라고 하자. 본 장에서는 분자 wDNF기계를 통해 해 공간을 병렬적으로 탐색하여 유망한 텀들을 선택한 후 그를 구성하는 변수들의 분포를 평가, 확률 모델을 확립하고 그로부터 다음 세대의 해 공간을 구성함으로써 진화알고리즘의 확장인 EDA을 다음의 알고리즘으로 모델링한다.

(알고리즘 3.1)

1. **(초기화)** 문제공간의 변수들의 임의의 조합으로 구성되는 N개의 텀들을 생성하여 초기 해 집단 $P(0)$라 하자. $t=1$.
2. **(D단계)** 데이터 $(\mathrm{x}, y) = (x_1, x_2, \ldots, x_n, y)$ 를 얻는다. 얻어진 데이터 x와 해 집단의 텀들간의 상보 결합 연산을 통해 가능도(likelihood)를 계산한다.
3. **(S단계)** 텀의 가능도에 따라 M 개의 텀들을 선택하여 $P(t)$ 를 구성한다.
4. **(P단계)** 부모 해 집단 $P(t)$ 의 분포를 평가한다.
5. **(V단계)** 얻어진 부모 해 집단의 분포로부터 샘플링하여 다음 세대 $P(t+1)$ 을 구성한다.
6. $t=t+1$, 종료 조건이 아니면 2단계로 진행, 반복한다.

문제 공간의 변수들의 임의의 조합으로 구성된 텀들로 구성된 초기 해 집단으로부터 시작하여 D단계에서는 새로운 데이터 x 를 관찰한 후 각 텀의 가능도를 계산한다. 즉 데이터 x 를 구성하는 x_i 들의 여러 복사본들과 상보결합이 되어 이중가닥을 형성하고 y 값이 같은 경우에는 1을, 그렇지 않은 경우는 0을 할당한다. S단계에서는 완전한 이중가닥을 형성한 텀들 중에서 M 개만을 선택한다. P단계에서 선택된 텀들로 이루어진 해 집단의 분포를 평가하는 것은 다음과 같다. 선택된 M 개의 텀들을 구성하는 각 변수들의 확률을 평가하기 위하여 해당 변수와 상보결합을 시도, 이중가닥이 형성된 것의 농도를 검출하여 확률 분포 $p(x_i)$ 를 계산한다.

$$p(x_i) = \frac{\sum_{j=1}^{M} \delta(C_j, x_i)}{M},\ i = 1, \ldots, n \ \ and$$
$$\delta(C_j, x_i) = 1 \ \ if \ C_j \ contains \ x_i, \ 0 \ \ if \ not.$$

이로부터 변수들이 독립일 경우 해 집단의 분포를 다음과 같이 평가한다.

$$p(X) = \prod_{i=1}^{n} p(x_i)$$

얻어진 분포로부터 변수들의 확률분포를 반영한 샘플링을 통해 생성된 N 개의 텀들로 다음 세대 $P(t+1)$ 를 구성하게 된다.

4. MDL원리를 반영한 적합도 함수

차수 k 의 wDNF의 경우, 최대 k 개의 변수들의 조합으로 구성되는 텀을 고려한다. 이렇게 정해진 차수의 wDNF의 경우에도 변수의 개수가 증가함에 따라 가능한 텀의 수가 폭발적으로 증가하게 된다. 이러한 점은 이상적인 DNA연산의 경우에는 문제가 되지 않으나 올리고 합성 기술의 고비용 등으로 인해 현실적인 실험에 어려움이 많다. 또한 실제로 임의의 텀을 구성하는 변수들의 대부분은 널 값이 많아 wDNF로 구축한 해 집단은 상당히 희박(sparse)하게 되어 검소한 (parsimonious) wDNF모델을 찾는 것이 중요한 문제로 대두된다. 따라서 적은 개수의

변수들로 동일한 정보를 표현할 수 있는 유용한 텀들을 찾아 내는 과정이 필요하게 된다.

학습의 정확도와 검소한 모델 구축간의 트레이드오프에 의한 최적의 모델 선택을 위해 코딩 이론에 따른 모델 복잡도를 다음과 같이 고려한다[7]. 즉 주어진 $p(x)$에 대하여 코딩 이론에 의해 그를 표현하는 코드 길이는 $L(p(x))=-\log(p(x))$ 이다. 그러므로 베이지안 추론에서 데이터 D에 대한 모델 A 의 사후확률분포 $p(A|D)$ 를 최대화하는 것은 코딩이론에 의하면 그를 표현하는 코드 길이를 최소화하는 것과 같게 된다.

$$L(p(A|D))=L(p(D|A)p(A))=-\log(p(D|A)p(A))$$
$$=-\log(p(D|A)-\log(p(A))=L(p(D|A))+L(p(A))$$

따라서 MDL(Minimum Description Length)원리에 기반한 각 텀 C_i 의 적합도 함수 F 를 다음과 같이 정의한다.

$$F(C_i)=\alpha E(i) \ +(1-\alpha)Comp(i)\ ,\ \alpha\in[0,1]$$

여기서 $E(i)$ 는 $p(D|A)$에 의존하는 값으로 텀 C_i 의 정확도를 반영한다. 알고리즘 3.1의 D단계에서의 가능도와 같으며 훈련데이터 x의 y 값과 일치하면 1을, 그렇지 않으면 0으로 정의하였다. 또한 $Comp(i)$ 는 $p(A)$에 의존하는 값으로 모델 복잡도의 폭발적 증가를 방지하기 위해 적은 개수의 변수들로 구성된 텀의 선택을 위한 바이어스로 다음과 같이 정의된다. 이 때 α 는 정확도와 모델 복잡도 간의 트레이드오프를 반영한다.

$$Comp\ (i)=\frac{n-\sum_{j=1}^{n}\delta(C_i,x_j)}{n},$$
$$\delta(C_i,x_j)=1\ if\ C_i\ contains\ x_j,\ 0\ if\ not.$$

DNA컴퓨팅의 특성으로 인해 더 높은 차수의 텀은 그보다 낮은 차수의 텀에 비해 완전 이중가닥을 형성하여 S단계에서 선택될 확률은 작아지게 되므로 자연스럽게 MDL 원리가 반영되어 단계가 진행됨에 따라 작은 차수의 텀들이 남게 되는 경향이 있다. 또한 알고리즘 3.1의 P단계에서는 변수들이 독립일 경우로 가정하여 해 집단의 분포를 평가하였으나 UMDA 등과 구별되는 점은 분자 wDNF기계의 텀들은 그를 구성하는 모든 변수들이 완전 이중가닥을 형성하여야 선택될 수 있으므로 내포적으로는 변수들간의 연관성도 반영하게 된다고 볼 수 있다.

5. 결론 및 향후 과제

본 논문에서는 분자 wDNF기계로 이루어진 분자 라이브러리의 진화를 통해 EDA를 DNA컴퓨팅으로 구현하였다. 또한 희박한 해 집단의 경우에는 간략한 wDNF 모델을 함께 찾으므로 단순히 해 집단의 분포만을 진화시켜 나가는 것이 아니라 모델의 구조도 같이 최적화 시켜 나가는 방안에 대해 살펴보았다. 본 논문에서는 변수들이 독립일 경우의 EDA를 모델링하였으나 향 후 연구를 통해 두 변수들간에 연관성이 있는 경우와 여러 변수들간에 연관성이 있는 경우의 EDA를 DNA 컴퓨팅을 통해 모델링하려고 한다. 또한 미리 차수 k 를 정하여 문제 공간에서 고려할 변수들을 선택한 후 라이브러리를 구성하는 제한 없이 문제공간의 크기를 제약하지 않은 상태로 유용한 텀들을 찾아 나가므로 모델의 구조까지 최적화 하는 문제의 실험을 진행하여 주어진 문제의 변수의 개수가 많아 지는 것에 비례하여 필요한 DNA분자 수는 증가하나 DNA컴퓨팅의 병렬성을 이용하여 같은 계산시간을 보장받게 됨을 보이려 한다.

감사의 글

이 논문은 과학기술부 국가 지정 연구실사업 (NRL) 에 의하여 지원되었음

참고문헌

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

[2]Baluja, S. (1994). 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.

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

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

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

[6]Zhang, B.-T., Jang, H.-Y. (2006). Molecular Learning of wDNF Formulae. *Lecture Notes in Computer Science, DNA11*, 3892:427-437.

[7]Zhang, B.-T., Ohm, P., & Mühlenbein, H. (1997). Evolutionary Induction of Sparse Neural Trees. *Evolutionary Computation*, Vol. 5, No. 2, 213-236.