안정된 추론을 위한 베이지안 네트워크 앙상블의 종분화 진화

유지오° 김경중 조성배
연세대학교 컴퓨터과학과
{taiji391°, uribyul}@sclab.yonsei.ac.kr, sbcho@cs.yonsei.ac.kr

Speciated evolution of Bayesian networks ensembles for robust inference

Ji-Oh Yoo°, Kyung-Joong Kim, Sung-Bae Cho
Dept. of Computer Science, Yonsei University

요 약

베이지안 네트워크는 불확실한 상황을 모델링하기 위한 확률 기반의 모델이다. 베이지안 네트워크의 구조를 자동 학습하기 위한 연구가 많이 있었고, 최근에는 진화 알고리즘을 이용한 연구가 많이 진행되고 있다. 그러나 대부분은 마지막 세대의 가장 좋은 개체만을 이용하고 있다. 시스템이 요구하는 다양한 요구조건을 하나의 적합도 평가 수식으로 나타내기 어렵기 때문에, 마지막 세대의 가장 좋은 개체는 종종 편향되거나 변화하는 환경에 덜 적응적일 수 있다. 본 논문에서는 적합도 공유 방법으로 다양한 베이지안 네트워크를 생성하고, 이를 베이즈 규칙을 통해 결합하여 변화하는 환경에 적응적인 추론 모델을 구축할 수 있는 방법을 제안한다. 성능 평가를 위해 ALARM 네트워크에서 인공적으로 생성한 데이터를 이용한 구조 학습 및 추론 실험을 수행하였다. 다양한 조건에서 학습된 네트워크를 실험한 결과, 제안한 방법이 변화하는 환경에서 더욱 강건하고 적응적인 모델을 생성할 수 있음을 확인할 수 있었다.

1. 서 론

인공지능 분야에서 베이지안 네트워크(Bayesian Network)는 불확실성을 처리하기 위한 중요한 방법의 하나로 부각되고 있다. 베이지안 네트워크는 현실 세계의 문제를 결합 확률 분포로 나타낸 모델로, 전문가의 지식을 쉽게 반영할 수 있는 장점이 있다.

이러한 모델을 학습시킬 때 최초에 강건한 모델을 찾는 것이 좋다. 환경이 변화할 때 그에 맞추어 모델을 적응 시키는 것은 비용이 많이 들거나 불가능한 경우도 있기 때문이다. 최적의 해 하나만을 찾는 일반적인 진화 알고리즘에서 강건한 해를 구하려면 실제 상황에서 있을지 모르는 노이즈까지 정확하게 반영할 수 있는 효과적인 평가 방법이 필요하다. 그러나 노이즈는 언제 나타날지 예측하기 힘들기 때문에 이러한 방법은 많은 어려움이 있다.

만약 다른 특성을 가진 여러 개의 해를 동시에 사용할 수 있다면, 최적 해 하나만을 사용하는 것보다 더 강건한 성능을 기대할 수 있을 것이다. 특히, 베이지안 네트워크는 사람이 쉽게 해석할 수 있고 전문가의 지식을 쉽게 반영할 수 있기 때문에, 환경이 변화했을 때 적은 비용으로 구조를 적응적으로 변화시킬 수 있다.

본 논문에서는 여러 개의 해를 찾는 종분화 진화 기법 중 하나인 적합도 공유를 이용하여 다양한 베이지안 네트워크를 생성하고, 이를 베이즈 규칙(Bayes Rule)을 통해 결합하는 방법을 제안한다. 동적인 환경에서 적합도 공유를 통해 생성된 여러 베이지안 네트워크 중 일부는 특정 상황에서 잘못된 추론을 할 가능성이 있지만, 다른 베이지안 네트워크와 상호 보완하여 올바른 결과로 수정될 수 있을 것이다. 제안한 방법의 유용성을 보이기 위해 대표적인 베이지안 네트워크 벤치마크 문제인 ALARM 네트워크에서 인공적으로 생성된 데이터를 사용하여 실험을 수행, 분석하였다.

2. 배경

2.1 베이지안 네트워크 (Bayesian Network, BN)

베이지안 네트워크는 각 환경 변수 간의 인과 관계를 나타낸 확률 기반의 그래프 모델이다. 각 환경 변수는 노드로 표현되고 각 환경 변수 간의 인과 관계는 노드와 노드 사이의 아크로 표현된다. 각 노드는 여러 가지 상태 값과 그에 대한 조건부 확률 테이블을 속성으로 가진다. 베이지안 네트워크는 방향성 비순환 그래프로 그 구조는 보통 전문가에 의해 설계되고, 각 노드의 조건부 확률 테이블은 전문가에 의해 설계되거나 표본 데이터로부터 계산된다. 네트워크를 학습한 후 어떤 상황에 대한 evidence가 관찰되면 이를 바탕으로 각 노드의 조건부 확률 테이블과 독립 조건을 이용, 베이지안 추론 알고리즘을 통해 각 노드의 상태에 대한 확률이 계산된다.

베이지안 네트워크의 구조를 자동 학습하기 위한 연구가 많이 진행되고 있는데, 진화 연산을 통한 학습 연구도 있다[1,2]. 베이지안 네트워크의 가정을 훼손시키지 않으면서 진화 연산을 하기 위해 순서(ordering) 가정을 두거나 회복(repair) 연산자를 이용하는 방법이 있다.

2.2 종분화 알고리즘과 앙상블

여러 개의 모델을 결합하기 위해서는 각 모델의 특성이 다를수록 좋다. 그러나 일반적인 진화 알고리즘은 다양성을 잘 유지하지 못하는 단점이 있어 결합에 적용하기 어려운 점이 있다. 몇몇 연구자들은 진화를 하는 도중에 다양성을 유지하기 위해 진화 알고리즘 중 선택 과정을 변형하는 종분화 알고리즘을 제안하였다. 대표적으로 적합도 공유, crowding, 제한적인 토너먼트 선택 등의 방법이 있다. 이렇게 해서 생성된 다양한 개체를 앙상블하면 각 개체간의 상호 작용을 통해 동적인 환경에 적응적인 모델을 만들 수 있다.

3. 종분화된 진화 베이지안 네트워크 앙상블

그림 1은 종분화된 진화 베이지안 네트워크 앙상블의 전체적인 순서도이다. 각각의 BN 개체를 임의의 구조로 초기화한 후, 베이지안 네트워크의 구조를 평가하는 DPSM(General Drichlet Prior Score Metric)을 사용해 각 베이지안 네트워크의 적합도를 계산한다. 적합도 공유에서 각 개체 간의 거리는 MDL을 통해 측정되고, 적합도에 따라 80%의 비율로 좋은 개체가 선택되고, 교차와 돌연변이 연산을 거친다. 미리 설정한 임계치보다 적합도가 높은 개체들이 일정 수가 탐색되거나, 최대 세대 수에 도달하면 진화는 멈추게 되고, 클러스터링을 통해 마지막 세대에서 결합할 개체들을 선택, 결합하게 된다.

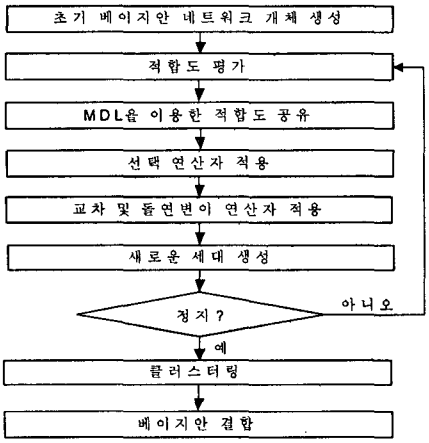

그림 1. 제안한 방법의 실행 순서도

3.1 염색체 표현

진화 알고리즘에서 각 개체의 표현 방법을 결정하는 것은 매우 중요한 문제이다. 본 논문에서는 베이지안 네트워크를 연결 행렬과 노드 순서를 통해 표현하는 방식을 제안한다. 베이지안 네트워크 구조는 변수의 순서를 나타내는 길이가 n인 L과 $n \times n$ 연결 행렬 C로 표현된다. C의 각각의 요소 c_{ij}는 다음과 같은 방법으로 표현되며 각 변수 사이에 아크가 존재하는지에 대한 여부를 표시한다.

$$c_{ij} = \begin{cases} 1 & i > j \text{ 이고 } x_i \text{와 } x_j \text{ 사이에 아크가 존재할 경우,} \\ 0 & \text{그렇지 않을 경우} \end{cases}$$

그림 2는 제안한 염색체 표현의 예를 보여준다.

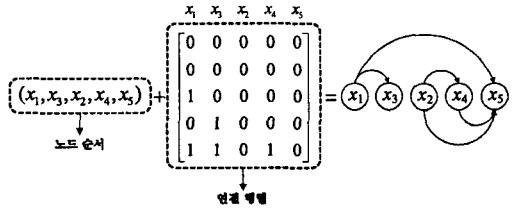

그림 2. 베이지안 네트워크의 염색체 표현의 예

3.2 유전 연산

각 개체의 적합도가 평가되면, 일정 비율로 좋은 개체가 선택되고, 교차와 돌연변이 연산을 거친다. 본 논문에서는 순위 기반 선택(Rank-based selection) 연산자를 사용하였다[1]. 이 연산자는 다음과 같이 계산된 개체 선택 확률 $p_{s,t}$를 통해 개체를 선택한다.

$$p_{s,t} = \frac{\lambda - rank(g(I_t^j)) + 1}{\lambda(\lambda+1)/2}$$

I_t^j는 세대 t에서 j번째 개체이고, $rank(g(I_t^j))$는 적합도에 따른 개체의 순위, λ는 개체의 수이다.

교차 연산자는 개체 집단 안에 있는 두 베이지안 네트워크의 구조를 교환하는 역할을 한다. 연결 행렬의 교차 연산은 1-점 교차 방식을 사용했고, 노드 순서의 교차 연산을 위해 순환 교차(Cycle crossover) 연산자를 사용하였다[2]. 순환 교차 연산자는 각 개체의 위치에 대응되는 상대 개체의 위치를 따라가며 사이클을 이루는 부분을 분리하고 이를 교환하는 방식이다. 그림 3은 순환 교차의 예이다.

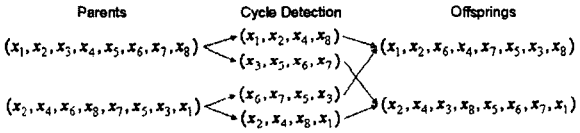

그림 3. 순환 교차의 예

연결 행렬의 돌연변이 연산은 bit-flip 연산을 사용하였다. 또한 노드 순서의 돌연변이 연산을 위해 전치 돌연변이(Displacement mutation) 연산자를 사용하였다[2]. 전치 돌연변이 연산은 먼저 무작위로 부분 문자열을 선택한 다음, 이 문자열을 제거해서 다른 임의의 위치에 삽입하는 방식이다. 그림 4는 전치 돌연변이의 예이다.

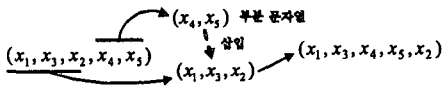

그림 4. 전치 돌연변이의 예

3.3 적합도 평가 및 개체간 거리 측정

적합도 평가를 위해서 데이터에 대해 베이지안 네트워크의 구조를 평가하는 score metric 중 성능이 가장 좋다고 평가되는 DPSM을 사용하였다[3]. DPSM은 데이터가 dirichlet 분포를 가지고 있다고 가정하고 구조를 평가하며, 다음과 같이 계산된다.

$$P(B,D) = P(B)\prod_{i=1}^{n}\prod_{j=1}^{q_i}\frac{\Gamma(N_{ij}')}{\Gamma(N_{ij}'+N_{ij})}\times\prod_{k=1}^{r_i}\frac{\Gamma(N_{ijk}'+N_{ijk})}{\Gamma(N_{ijk}')}$$

N_{ijk}는 주어진 데이터베이스 D에서 변수 x_i의 상태가 k이고, 그의 부모 $Pa(x_i)$의 상태는 j인 경우의 수인고, $N_{ij} = \sum_{k=1}^{r_i} N_{ijk}$이다. N_{ijk}'는 dirichlet 분포 차수로 일반적으로 10인 경우에 성능이 좋다. $N_{ij}' = \sum_{k=1}^{r_i} N_{ijk}'$이다.

종분화 기법으로 적합도 공유(fitness sharing) 방식을 사용하였다. 적합도 공유는 개체들이 밀집된 지역에 있는 개체들의 적합도를 떨어뜨려서 다른 개체 간 적합도를 공유하는 방법이다. i번째 개체의 적합도를 f_i, i번째와 j번째 개체 사이의 거리를 d_{ij}, 공유 반경을 σ_s라고 할 때 공유 적합도 fs_i는 다음과 같이 계산된다.

$$fs_i = \frac{f_i}{\sum_{j=1}^{popsize} sh(d_{ij})} \qquad sh(d_{ij}) = \begin{cases} 1 - \frac{d_{ij}}{\sigma_s}, & 0 < d_{ij} < \sigma_s \\ 0, & d_{ij} \geq \sigma_s \end{cases}$$

i번째와 j번째 개체 사이의 거리 d_{ij}는 베이지안 네트워크의 구조적 차이를 측정해 계산한다. 이를 위해 MDL 척도를 사용하여 측정한다. i번째 개체를 기준으로 j번째 개체의 역방향 아크 수를 r, 빠진 아크 수를 m, 추가된 아크 수를 a라고 하고, 총 노드의 수를 n이라 할 때, d_{ij}는 다음과 같이 계산된다.

$$d_{ij} = (r + m + a)\log_2[n(n-1)]$$

3.4 다중 베이지안 네트워크의 결합

마지막 세대로부터 개체를 선별하기 위해 단일 클러스터링 기법을 사용한다. 결합에 사용될 개체의 수는 사전에 정의된 거리 임계 치에 의해 자동으로 결정된다. 선별된 개체는 베이지안 결합 기법에 의해 결합된다. 베이지안 결합은 각 베이지안 네트워크의 오류 가능성을 염두에 두고, 오류에 대한 정보를 바탕으로 결합 베이지안 네트워크를 구성하고, 결합 결과에 반영하는 방법이다.

4.실험

4.1 실험

실험은 ALARM 네트워크에서 인공적으로 생성한 데이터를 사용하여 네트워크를 학습시킨 후 다양한 조건에서 성능을 비교 평가하였다. 그림 5는 ALARM 네트워크로 Beinlich 등이 설계하였고 병원에서 환자의 상태를 모니터링하기 위한 시스템이다[4].

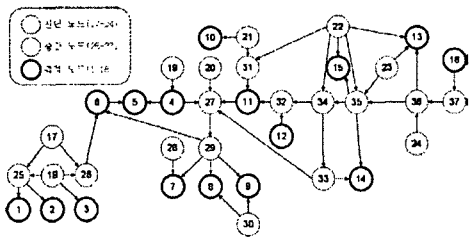

그림 5. ALARM 네트워크

베이지안 네트워크로부터 데이터를 생성하기 위해 확률적 논리 추출 기법을 사용하였고, 3000개의 케이스로 구성된 데이터를 생성하였다. 생성된 케이스 중 600개를 임의로 추출하여 테스트 데이터로 사용하고 나머지 2400개의 케이스는 학습 데이터로 사용하였다. 진화 알고리즘의 개체 수는 50으로 설정하였고, 최대 진화 수는 5000세대, 선택확률은 0.8, 교차 확률은 0.5, 돌연변이 확률은 0.01로 설정하였다.

4.2 실험 결과

실험 결과는 10번 수행한 결과를 평균한 것이다. 표 1은 분류 정확도를 비교한 것이다. 일반 진화 알고리즘으로 생성된 개체가 종분화 알고리즘의 개체보다 더 높은 정확도를 나타내고 있지만, 종분화 알고리즘을 통해 생성된 개체를 결합한 경우, 오히려 높은 정확도를 보이고 있음을 알 수 있다. 또한 결합한 모델이 표준 편차도 작게 나타나 안정적임을 알 수 있다.

표 1. "Hypovolemia" 노드에 대한 추론 정확도

일반 진화 알고리즘	종분화 알고리즘	종분화 개체의 결합
96.78±1.9	96.7±1.16	97.33±0.097

그림 6은 일반 진화 알고리즘과 종분화 알고리즘의 다양성 차이를 보여준다. x축은 개체를 y축은 개체 간 차이를 나타내는데, 종분화 알고리즘이 여러 개체에 걸쳐 차이가 고르게 나고 있어 더 다양한 개체가 진화되었음을 알 수 있다.

실제 현상에 대해 모델이 얼마나 신뢰성을 가지는가를 측정하기 위해 정보 보상(Information Reward) 정도를 사용할 수 있다[5].

$$IR = \sum_{i}^{n}[1 + log_2 P(x_i = v)]$$

여기서 IR은 정보 보상 정도를, i는 시험 데이터에 n개의 케이스 중 하나를 의미하고, v는 i번째 케이스에 대한 실제 값, $P(x_i)$는 학습 모델로부터 얻어진 실제 값에 대한 확률이다.

관찰 가능한 노드의 수가 줄어들면 그만큼 불확실성은 커지기 때문에 성능은 떨어질 수 있다. 그림 7은 관찰 가능한 노드의 수가 변할 때 정보 보상 정도가 어떻게 변화하는지 나타낸 그래프이다. 변화가 적을수록 성능이 좋다고 할 수 있다. 그림과 같이 제안한 방법이 다른 방법에 비해 비교적 높은 정보 보장 정도를 유지하면서 변화가 가장 작다. 이를 통해 결합한 개체들이 상호 작용을 통해 불확실한 상황에서도 적응적인 추론을 할 수 있음을 알 수 있다.

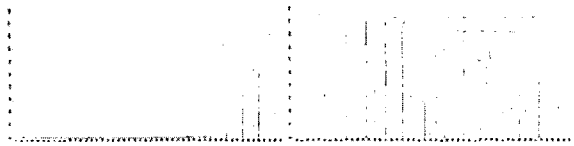

그림 6. 일반 진화 알고리즘과 종분화 진화 알고리즘

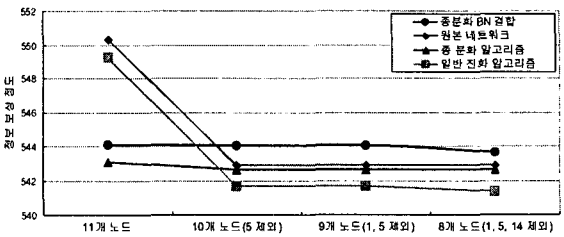

그림 7. "Anaphylaxis" 노드에 대한 정보 보상 정도 변화

5. 결론

본 논문에서는 불확실한 환경에 적응적인 베이지안 네트워크를 학습하기 위해 종분화를 통해 얻어진 베이지안 네트워크들을 베이지안 결합 방법을 사용하여 결합하는 모델을 제안하였다. ALARM 네트워크에서 생성한 데이터를 이용한 실험 결과에서 제안된 방법이 오직 하나의 해만을 찾는 일반 진화 알고리즘보다 좋은 결과를 내고 있음을 확인할 수 있었다. 결합에 사용한 각각의 베이지안 네트워크가 오류를 내는 경우에도, 결합을 통해 상호 보완하여 보다 정확한 결과를 내줄 수 있음을 확인했다. 또한 제안한 방법이 불확실한 상황에 좀더 적응적인 추론을 할 수 있음을 알 수 있었다.

향후 연구로는 제안한 방법을 실제 환경에서 일어날 수 있는 문제에 적용시켜보는 것이다. 또한 거리 측정 방법을 확률적인 기반의 KL distance로 바꾸어 적용시켜 보고 비교 분석을 할 예정이다.

참고 문헌

[1] P. Larranaga, M. Poza, Y. Yurramendi, R. H. Murga, and C.M.H. Kuijpers, "Structure learning of Bayesian networks by genetic algorithms: A performance analysis of control parameters," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 18, no. 9, pp. 912-926, 1996.

[2] P. Larranaga, C.M.H. Kuijpers, R. H. Murga, and Y. Yurramendi, "Learning Bayesian network structures by searching for the best ordering with genetic algorithm," *IEEE Trans. on Systems, Man and Cybernetics-Part A*, vol. 26, no. 4, pp. 487-493, 1996.

[3] S. Yang, and K.-C. Chang, "Comparison of score metrics for Bayesian network learning," *IEEE Trans. on Systems, Man and Cybernetics-Part A*, vol. 32, no. 3, pp. 419-428, 2002.

[4] I. A. Beinlinch, H. J. Suermondt, R. M. Chavez and G. F. Cooper, "The ALARM monitoring system: A case study with two probabilistic inference techniques for belief networks," *Proc. of the Second European Conf. on Artificial Intelligence in Medicine*, pp. 247-256, 1989.

[5] I. Good, "Rational decisions," *Journal of the Royal Statistical Society B*, vol. 14, pp. 107-114, 1952.