진화 알고리즘과 MDL을 이용한 베이지안 네트워크 갱신

김경중, 조성배
연세대학교 컴퓨터과학과
{kjkim, sbcho}@cs.yonsei.ac.kr

Refinement of Bayesian Networks Using Minimum Description Length and Evolutionary Algorithm

Kyung-Joong Kim, Sung-Bae Cho
Dept. of Computer Science, Yonsei University

요 약

베이지안 네트워크는 확률이론에 기초해 불확실성이 존재하는 실세계 문제를 해결하는데 많은 기여를 하고 있다. 최근, 네트워크 구조를 데이터로부터 자동으로 학습하는 많은 연구가 이루어져 보다 손쉽게 많은 사람들이 사용할 수 있게 되었다. 하지만 한번 학습하여 고정된 네트워크의 구조는 새롭게 수집되는 데이터의 특성을 잘 반영하지 못하는 문제를 지니고 있다. 환경의 변화에 맞게 지속적으로 네트워크 구조를 갱신하기 위한 연구가 진행되고 있으며 본 연구에서는 Lam이 제안한 MDL기반 평가함수를 이용한 진화적 갱신 방법을 제안하여 갱신 성능을 향상시키고자 한다. 벤치마크 네트워크인 ASIA에 대한 실험 결과 제안한 방법이 기존의 지역적 탐색 방법에 비해 향상된 성능을 제공함을 확인하였다.

1. 서 론

베이지안 네트워크는 불확실성을 다룰 수 있도록 고안된 그래프 기반 확률 모델이다[1]. 각 노드는 시스템의 변수를 의미하고 에지는 변수들 사이의 확률적 의존관계를 나타낸다. 유연한 추론이 가능하고 데이터로부터 모델 학습도 가능하기 때문에 불확실성을 지니고 있는 실세계 문제를 해결하는데 유용하다. 로봇을 포함한 이동장비의 환경인식, 이미지 인식, 움직임 인식, 문서 분류, 유전자 제어 네트워크 구축 등 다양한 분야에서 응용이 이루어지고 있다. 최근, 데이터로부터 네트워크 구조를 학습하는 연구가 활발히 진행되고 있으나 한번 구조를 고정시킨 후 지속적으로 수집되는 데이터를 활용하지 못하는 문제로 인해 적응성이 떨어지는 단점을 지니고 있다. 즉 기존의 지식도 활용하면서 새롭게 들어오는 데이터의 특성도 지속적으로 반영할 수 있는 갱신기법에 대한 연구가 필요하다.

갱신과정의 입력은 새롭게 수집한 데이터와 기존 네트워크 구조이며 출력은 갱신한 네트워크이다. 갱신한 네트워크는 도메인을 기존 네트워크 보다 잘 설명할 수 있는 향상된 모델이어야 한다. 즉 갱신 네트워크는 새로운 데이터의 분포를 가능한 정확하게 표현해야 하며 기존의 네트워크 구조의 특성을 가능한 많이 보존해야 한다. 또한 갱신 네트워크의 구조를 가능하면 단순하게 유지해야 한다. 대부분의 상황에서 이러한 요구조건들은 동시에 만족되지 못한다. 즉 갱신 네트워크는 세 가지 요구 조건 사이에서 균형을 유지해야 한다. 이러한 문제를 해결하기 위해 Lam은 기계학습 분야에서 충분한 기초를 갖추고 있는 Minimum Description Length (MDL) 원칙을 갱신문제에 적용하였다[2]. 그는 MDL에 기초한 갱신 평가 함수를 제안하고 이를 이용한 탐색 알고리즘으로 MDL 수치가 높은 local graph를 결합하는 방법을 사용하였다. 이 방법은 일종의 greedy 탐색 방법으로 지역해(Local minimum)에 빠질 수 있다는 단점이 있다. 본 논문에서는 전역해 탐색에 유리한 진화 알고리즘을 베이지안 네트워크 갱신 문제에 적용하여 성능향상을 시도하였다.

2. 진화 알고리즘을 이용한 베이지안 네트워크 갱신

2.1 MDL 평가 함수

MDL 원칙은 데이터를 설명하는 가장 좋은 모델은 다음 두 가지 항목의 합을 최소화 하는 것이라는 아이디어에 기초한다.

1) 모델을 표현하는데 필요한 정보의 양
2) 주어진 모델을 이용하여 입력원를 표현하는데 필요한 정보의 양

MDL 원칙을 베이지안 네트워크의 갱신 문제에 적용할 경우 입력원은 새롭게 들어온 데이터와 기존의 네트워크 구조이다. 즉 이 두 종류의 입력으로부터 새로운 네트워크 구조 H_p를 학습하는 것이 목표이다. 탐색 알고리즘은 다음 세 항목의 합을 최소화하는 H_p를 찾아야 한다.

1) 네트워크 자체의 표현 길이 (H_p의 표현 길이)
2) 주어진 네트워크 H_p에 대한 새로운 데이터 표현길이
3) 주어진 네트워크 H_p에 대한 기존 네트워크 구조에 대한 표현 길이

베이지안 네트워크와 관련한 도메인이 N개의 변수로 이루어져 있으며 $\overrightarrow{X_N}=\{X_1,\ldots,X_N\}$으로 정의한다. Y_{X_i}는 변수 X_i의 부모 노드들의 집합을 나타낸다. 첫 번째 항목인 네트워크 표현 길이는 네트워크의 구조가 복잡할수록 표현길이가 길어지도록 되어 있다. 표현 길이는 다음과 같이 정의된다.

$$\sum_{X_i\in \overrightarrow{X_N}}[k_i\log_2(N)+d(s_i-1)\prod_{j\in Y_{X_i}} s_j]$$

k_i 는 부모 노드의 개수이며, s_i는 변수가 취할 수 있는 상태 값의 개수이다. d는 실수 값을 저장하는데 필요한 비트수이다.

두 번째 항목인 주어진 네트워크에 대한 데이터 표현 길이는 다음과 같이 정의된다. 데이터베이스는 M개의 데이터를 포함하고 있으며 각 독자적 사건 e_i의 상대빈도는 r_i이다. q_i는 각 사건 e_i에 대해 주어진 네트워크 모델이 부여한 확률 값이다. 데이터 표현 길이는 다음과 같이 정의된다.

$$-M\sum_i r_i \log_2(q_i)$$

세 번째 항목인 주어진 네트워크 구조에 대한 기존 네트워크 구조의 표현길이는 다음과 같이 정의된다. 주어진 네트워크를 B_N으로 정의하고 기존의 네트워크는 B_E로 정의한다.

$$(r+m+a)\log_2[N(N-1)]$$

이때 $r\ m\ a$는 각각 B_E에 있는 에지 중에서 B_N과 비교했을 때 방향이 반대인 에지, 삭제된 에지, 추가된 에지의 개수를 나타낸다.

MDL 평가함수는 여러 항목의 합으로 분할이 가능하다. 이러한 특성은 탐색 알고리즘을 효율적으로 설계하는데 유용하다. 세 가지 항목의 합인 총 표현 거리는 다음과 같은 식으로 정의된다.

$$T=\sum_{X_i\in X_N}[k_i\log_2(N)+d(s_i-1)\left(\prod_{j\in Y_{X_i}} s_j\right)-MW(X_i,Y_{X_i}) +(r_i+m_i+a_i)\log_2[N(N-1)]]$$

$R()$은 데이터베이스에서 일치하는 데이터의 상대빈도를 의미한다. 만약 $Y_{X_i}=\varnothing$이면 $W(X_i,Y_{X_i})=0$이다.

$$W(X_i,Y_{X_i})=\sum_{X_i,Y_{X_i}} R(X_i,Y_{X_i})\log_2\frac{R(X_i,Y_{X_i})}{R(X_i)R(Y_{X_i})}$$

2.2 진화 탐색

진화 알고리즘은 여러 개의 개체를 동시에 이용하는 집단 기반 탐색 방법이며 전역해의 탐색에 유리하다. Bayesian 네트워크를 표현하는 방법은 연결행렬을 사용하는 방법과 변수 순서를 이용하는 방법이 있다[3]. 전자는 유전자 연산에 의해 사이클이 발생할 수 있는 문제가 있지만 수정 연산자를 사용하여 보완하면 손쉽게 이용할 수가 있다. 후자는 탐색 공간을 줄이고 수정 연산자를 사용하지 않아도 되는 장점이 있지만 유전 연산에 의한 해의 변화폭이 매우 크다는 문제점이 있다. 본 논문에서는 연결행렬을 사용하여 베이지안 네트워크를 표현하였다. N개의 변수를 가진 베이지안 네트워크는 $N\times N$ 행렬에 의해 표현되며 행렬의 (i,j) 항목은 j번째 노드가 i번째 노드의 부모일 경우에 1이 되며 그렇지 않은 경우 0이다.

총 표현 거리 T값은 작을수록 좋은 것이기 때문에 적합도는 다음과 같이 정의하였다.

$$f_i=\frac{T_{\max}-T_i}{T_{\max}-T_{\min}}$$

개체의 선택은 Roulette Wheel 방법을 사용하였다. 교차는 일점 교차를 사용하였으며 한 노드를 교차점으로 정한 후 그 노드보다 순서가 뒤인 노드들의 정보를 모두 교환하도록 했다. 돌연변이 일정한 확률로 에지가 있는 경우 에지를 삭제하고 에지가 없는 경우 에지를 추가하도록 했다.

교차 연산 혹은 돌연변이 연산을 수행하게 되면 사이클이 발생할 수 있기 때문에 이를 정정해 주는 수정연산자를 정의하였다. 수정연산자는 사이클을 탐지한 후 사이클을 형성하고 있는 에지들 중에서 하나를 임의로 선택하여 삭제하도록 하였다.

초기 집단의 개체는 임의로 생성하였으며 사이클이 생기지 않도록 확인하는 과정을 거쳤다. 각 세대의 최고 적합도 개체는 다음 세대에도 지속적으로 살아남도록 허용하였다.

3. 실험 및 결과

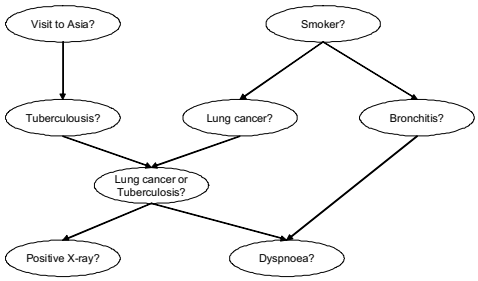

그림 1. ASIA 네트워크의 구조

제안한 방법의 유용성을 보이기 위해 널리 사용되는 ASIA 네트워크(그림 1)를 사용하였다. 매 실험마다 원본 네트워크로부터 모든 변수를 포함하는 완전한 데이터를 생성하였다. 기존의 네트워크 구조는 원본 네트워크 구조를 훼손율 0.1~0.9 사이의 값을 가지는 δ의 비율로 변형하여 생성하였다. 원본 네트워크에 C개의 에지가 존재한다면 생성된 기존 네트워크는 대략 $C\delta/3$개의 새로운 에지를 추가하고, 대략 $C\delta/3$개의 에지의 방향을 바꾸고, 대략 $C\delta/3$개의 에지를 삭제하여 생성한다. 훼손 네트워크를 생성하는 과정에서 사이클이 생기는 지를 확인하여 훼손 네트워크가 정상적인 베이지안 네트워크가 되도록 하였다. 훼손율이 크면 클수록 원본 네트워크와 생성된 기존 네트워크는 많은 차이가 생긴다. 원본 네트워크의 사용 목적은 오직 새로운 데이터를 생성하기 위한 것이기 때문에 갱신 알고리즘은 이를 알지 못한다. 표 1은 갱신 실험을 위한 파라미터 설정을 보여준다. 파라미터는 반복적인 실험을 통해 결정되었으며 그림 2는 세대별 갱신 성능을 보여준다. 정확도는 10번의 실험을 반복하여 얻은 결과의 평균이다. 정확도는 다음과 같이 정의한다.

$$accuracy=(1-SD/(N\times(N-1)/2))\times 100$$
$$SD=R+A+D$$

N: 노드의 개수
R: 원본 네트워크와 방향이 반대인 $edge$의 수
A: 원본 네트워크에 없지만 추가된 $edge$의 수
D: 원본 네트워크에 있지만 삭제된 $edge$의 수

그림 3은 갱신 네트워크를 평가하는 방법을 보여준다. 갱신 네트워크는 원본 네트워크와의 유사성으로 평가한다 .

표 1. 실험을 위한 파라미터 설정

집단의 크기	50	진화세대수	5000
돌연변이율	0.1	새로운 데이터의 수	1000~2000
교차율	0.8	훼손 비율	0.1~0.9

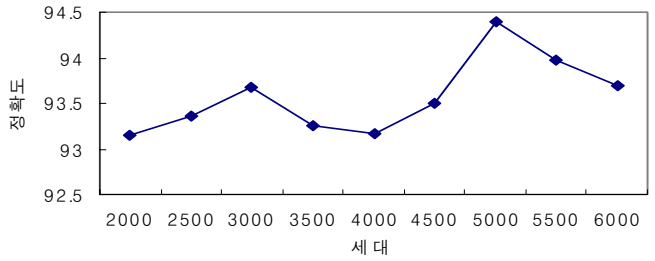

그림 2. 세대별 진화탐색의 갱신 성능

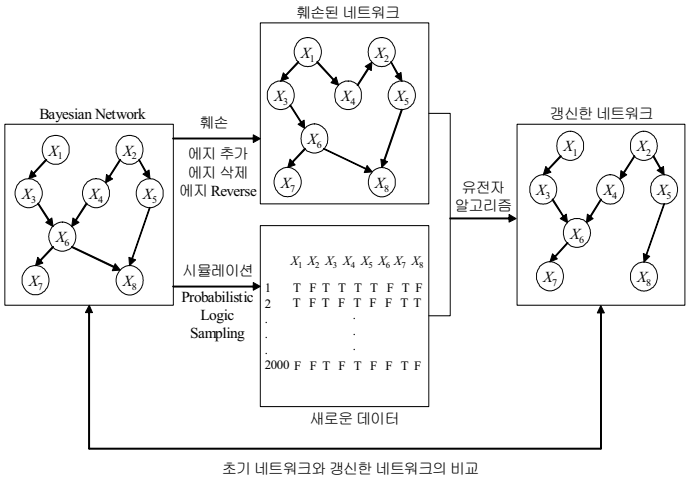

그림 3. 제안한 방법의 평가

제안한 방법의 성능을 비교하기 위해 대표적인 지역 탐색 방법인 greedy 탐색이 사용되었다. Greedy 탐색 방법은 에지가 없는 네트워크 구조로부터 시작하여 가장 높은 성능 향상을 보이는 에지 1개를 추가하고 구조를 고정시킨다. 고정된 구조를 시작점으로 하여 아직까지 추가 되지 않은 에지들 중에서 가장 높은 성능 향상을 보이는 것을 추가하고 구조를 고정한다. 이러한 작업은 더 이상 성능향상이 없거나 더 이상 추가할 에지가 없을 때까지 반복한다. Greedy 탐색은 빠른 탐색 성능을 보이지만 지역해에 빠질 수 있는 단점이 있다.

그림 4와 5는 각각 새로운 데이터의 수와 훼손율의 변화에 따른 성능의 차이를 보여준다. 제안한 진화 방법이 greedy 방법에 비해 높은 성능을 보여주었다. 새로운 데이터의 수가 증가할수록 갱신의 성능은 향상되고 훼손율이 커질수록 갱신의 성능이 하락되는 것을 확인할 수 있다.

4. 결론 및 향후연구

본 논문에서는 베이지안 네트워크의 구조를 갱신하기 위해 MDL 평가함수와 진화 알고리즘을 사용하는 방법을 제안하였다. 대표적인 지역 탐색 방법의 하나인 greedy 방법과 비교했을 때 제안하는 방법이 향상된 성능을 보임을 확인하였다. 향후 연구로 보다 규모가 큰 벤치마크 데이터에 대한 평가 수행과 Evolutionary Programming을 이용한 진화 속도향상을 진행할 것이다.

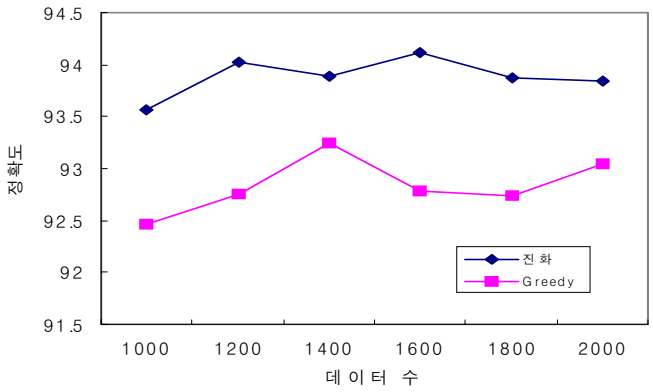

그림 4. 새로운 데이터의 수에 따른 성능 비교

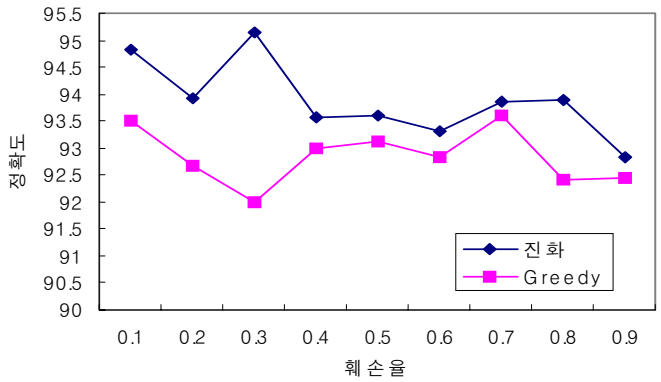

그림 5. 훼손 비율에 따른 성능 비교

감사의 글

본 논문은 정보통신부가 지원하는 21세기 프론티어 유비쿼터스 컴퓨팅 사업에 의해 지원되었음

참고문헌

[1] J. Pearl, *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*, Morgan Kaufmann, 1988.

[2] W. Lam, "Bayesian network refinement via machine learning approach," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 20, no. 3, pp. 240-251, 1998.

[3] 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.