클러스터링을 이용한 효율적인 대규모 베이지안 망 학습 방법

정성원° 이광형* 이도헌*
한국과학기술원 전산학과°
한국과학기술원 바이오시스템학과*
{swjung°, khlee, dhlee}@biosoft.kaist.ac.kr

An Efficient Learning Method for Large Bayesian Networks using Clustering

Sungwon Jung° Kwang H. Lee* Doheon Lee*
Department of Electrical Engineering & Computer Science, KAIST°
Department of BioSystems, KAIST*

요 약

본 논문에서는 대규모 베이지안 망을 빠른 시간 안에 학습하기 위한 방법으로, 클러스터링을 이용한 방법을 제안한다. 제안하는 방법은 베이지안 구조 학습에 있어서 DAG(Directed Acyclic Graph)를 탐색하는 영역을 제한하기 위해 클러스터링을 사용한다. 기존의 베이지안 구조 학습 방법들이 고려하는 후보 DAG의 수가 전체 노드 수에 의해 제한되는 데 반해, 제안되는 방법에서는 미리 정해진 클러스터의 최대 크기에 의해 제한된다. 실험 결과를 통해, 제안하는 방법이 기존의 대규모 베이지안 망 학습에 활용되었던 SC(Sparse Candidate) 방법 보다 훨씬 적은 수의 후보 DAG만을 고려하였음에도 불구하고, 비슷한 정도의 정확도를 나타냄을 보인다.

1. 서 론

베이지안 망 학습은, 주어진 학습 데이터로부터 확률적 의존 관계를 나타내는 DAG 구조와, 확률 분포를 기술하는 확률 파라미터를 학습하는 것이라 할 수 있다. 그 중 본 논문에서는 DAG 구조를 학습하기 위한 방법을 제시한다.

DAG 구조 학습은, 주어진 학습 데이터로부터 데이터 요소 간의 확률적 의존 관계를 가장 잘 나타내는 최적의 DAG를 찾아내는 것이다. 이러한 DAG를 학습하는 방법에는 여러 가지가 있으나, 그 중 하나는 각각의 후보 DAG를 베이지안 점수[1] 등의 방법을 이용하여 평가하고 그 중 가장 높은 점수를 보이는 DAG를 찾는 점수 기반 탐색 방법이다.

이러한 DAG 구조 학습에 있어서 고려할 수 있는 계산 비용에는 여러 가지가 있으나, 그 중 큰 영향을 미치는 부분 중 하나는 고려되는 가능한 모든 DAG의 수이다. n개의 노드가 있는 경우, 가능한 모든 DAG의 수는 다음과 같음이 알려져 있다[2].

$$f(n) = \sum_{i=1}^{n} (-1)^{i+1} \binom{n}{i} 2^{i(n-i)} f(n-i), \qquad n > 2$$
$$f(1) = 1$$
$$f(0) = 0$$

예를 들어, 노드 수가 2, 3, 5 그리고 10으로 증가함에 따라 가능한 DAG의 수는 3, 25, 29,281 그리고 약 4.2×10^{18}개로 증가하게 된다. 이러한 환경에서 가장 높은 점수를 갖는 DAG 구조를 찾는 문제는 NP-complete 문제임이 알려져 있다[3]. 따라서 일반적인 베이지안 구조 학습 방법은 최적 DAG에 근사하는 답을 찾는 휴리스틱 탐색 방법을 많이 사용하고 있다.

그러나 최근 바이오 정보학 등의 분야에서 보다 큰 규모의 베이지안 망을 효율적으로 학습할 수 있는 방법이 요구되어지고 있다. 기존의 일반적인 베이지안 망 학습의 규모가 주로 수십개 정도의 노드로 구성된 망 학습이었던 데 반해 이러한 분야에서 요구되는 베이지안 망은 수백에서 수천개에 이르는 노드를 갖고 있는 훨씬 큰 규모인 것이 특징이다. 이러한 대규모의 베이지안 망 학습은 기존의 휴리스틱 기반 탐색 방법으로도 결과를 얻는 데에는 많은 시간이 걸리게 된다. 이러한 문제점을 해결하기 위해, 탐색 대상으로 고려하는 후보 DAG의 범위를 줄이는 방법들이 제안되어 왔다[4][5].

기존의 탐색 범위를 좁히는 방법들은 각 노드들의 부모 노드 및 자식 노드의 대상 범위를 좁히는 접근 방법을 택하고 있다. 그러나 고려되는 후보 DAG의 수는 여전히 DAG 탐색이 일어나는 전체 노드 수에 의해 제한되어지고 있으며, 수백개 이상의 노드를 갖는 경우 고려해야 할 DAG 후보들의 수가 급격히 증가함으로 인해 여전히 많은 시간이 걸리는 문제점을 갖고 있다.

본 논문에서는 이러한 후보 DAG 탐색이 일어나는 탐색 공간을 전체 n개의 노드에 의한 탐색 공간으로부터

미리 정해진 c_{max}개의 노드 수에 의한 공간만큼으로 제한하는 방법을 제안한다.

2. 제안된 탐색 공간 제한 기법

본 논문에서 제안하는 방법은 다음과 같은 과정으로 구성된다.

- 노드의 계층적 클러스터 구성
- 상위 클러스터 내부에서의 베이지안 망 학습
- 상위 레벨에서의 서열에 따른 하위 레벨 베이지안 망 학습

제안하는 방법에서 DAG 탐색이 일어나는 부분은 각 클러스터 내부에서의 베이지안 망 학습이며, 이 때 탐색 공간은 한 클러스터 내부에 있는 노드 혹은 하위 레벨 클러스터 수에 의해 제한된다. 이 때 클러스터 계층 구조 안에 존재하는 모든 클러스터의 수는 한 클러스터의 최대 크기를 c_{max}로 제한하는 경우 $O(\frac{n}{c_{\max}})$가 된다. 따라서 제안하는 방법은 $O(c_{\max})$의 노드를 갖는 DAG 탐색을 $O(\frac{n}{c_{\max}})$번 수행하게 되며, 이것은 $O(n)$의 노드를 갖는 DAG의 탐색 공간에 비해 크게 제한된 규모의 탐색 공간이다.

2.1 노드의 계층적 클러스터링

최대 c_{max}개의 구성원을 갖는 클러스터의 계층을 구성하기 위해, 다음과 같은 클러스터간 유사도와 클러스터 결합 함수를 사용한다.

- $Similarity(C_i, C_j) = \dfrac{\sum_{l=1}^{|C_i|}\sum_{m=1}^{|C_j|} MI(N_l, N_m)}{|C_i| \times |C_j|}$

- $MERGE(C_i, C_j)$

  $if\ |C_i \cup C_j| \le c_{\max},\ return\ C_i \cup C_j$
  $if\ |C_i \cup C_j| > c_{\max},\ return\ \{C_i, C_j\}$

2.2 상위 레벨에서의 클러스터간 베이지안 구조 학습

클러스터 사이에 DAG형태의 서열을 추정하기 위하여, 클러스터들간의 베이지안 구조 학습을 수행한다. 이 때 각 클러스터에서 하나의 대표 노드를 선택하여 베이지안 구조 학습에 사용한다. 각 클러스터에서 선택되는 대표 노드는 다음과 같은 식으로 표현되는 외부와의 연결이 가장 많을 것으로 추정되는 노드로 한다.

$$Similarity(N_i, U - C_{N_i})$$

2.3 상위 레벨에서의 서열이 적용된 하위 레벨 클러스터에서의 베이지안 학습

하위 레벨 클러스터 C_i 내에 있는 노드들 사이의 베이지안 학습을 수행할 때, C_i의 부모 클러스터의 대표 노드를 C_i에 임시로 포함시킨 상태에서 베이지안 학습을 수행한다. 이 때 포함된 C_i의 부모 클러스터의 대표 노드를 이하에서 '추상 노드' 라 칭한다.

C_i 내부에서의 베이지안 학습 결과, 한 노드 N_i의 부모 노드들 중 추상 노드 A_k가 있는 경우, 다른 부모 노드들은 그대로 유지한 채 추상 노드 A_k를 해당 클러스터 C_k에 포함되어 있는 노드들로 분해한다. 이 과정은 다음과 같이 요약된다.

$$N_i \leftarrow \{N_a, N_b, \ldots, A_k\}$$
$$\rightarrow_{Decomposition} \quad N_i \leftarrow \{N_a, N_b, \ldots, N_l, N_m\} \quad (N_l \sim N_m \in C_k)$$

3. 실험 결과

제안된 방법을 평가하기 위하여, 많은 수의 노드를 가진 베이지안 망 학습에 적합한 방법 중 하나인 SC(Sparse Candidate) 방법을 비교 대상으로 하였다. SC 방법은 각 노드별로 부모 노드가 될 수 있는 후보들을 미리 결정하는 휴리스틱 접근 방법을 택하고 있으며, 바이오정보학 분야에서 대규모 베이지안 망 학습에 많이 사용되어지고 있다. 제안된 방법을 SC 방법과 비교하기 위해, 6개의 실험용 베이지안 망(표 1)으로부터 각 1,000개의 학습 데이터를 샘플링하여 생성하였다.

베이지안 망	노드 수	엣지 수
ALARM	37	46
BARLEY	48	84
HAILFINDER	56	66
WIN95PTS	76	112
PATHFINDER	109	195
DIABETES	413	602

표 1 실험용 베이지안 망

각 1,000개의 학습 데이터를 이용하여, 제안된 방법과 SC 방법을 통해 베이지안 망을 학습한 후 얻어진 결과를 본래의 베이지안 망과 비교하여 노드간 엣지 연결이 틀린 수를 서로 비교하였다. 또한 양 방법을 사용하였을 경우에 대해, DAG 탐색 과정에서 찾아 본 후보 DAG의 수를 비교함으로써, 고려된 탐색 공간의 차이를 비교하였다.

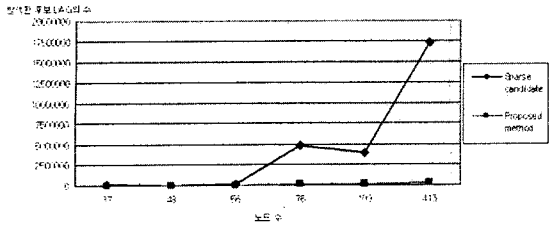

그림 1 탐색한 후보 DAG의 수

그림 1은, 각각의 방법을 사용하여 제한된 공간에서 최적의 DAG를 근사하는 하나의 DAG를 찾을 때 까지 탐색한 후보 DAG의 수를 기록한 것이다. 노드의 수가 늘어남에 따라 SC 방법은 탐색하게 되는 후보 DAG의 수가 급격하게 늘어남에 비해, 제안된 방법의 경우에는 그 증가의 폭이 극히 적음을 알 수 있다. 탐색하는 후보 DAG의 수는 곧 알고리즘의 실행 시간에 반영되므로, 제안된 방법은 SC 방법에 비해 훨씬 짧은 시간에 베이지안 구조를 학습하게 됨을 알 수 있다.

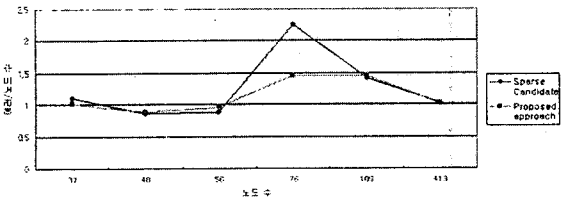

그림 2 에러/노드의 비교

그림 2는, 각 방법을 사용하였을 경우에 얻은 베이지안 망을 원래의 베이지안 망과 비교하여 노드 사이의 엣지 연결이 잘못된 수의 비율을 비교한 그래프이다. 비록 제안하는 방법이 SC 방법에 비해 훨씬 적은 규모의 탐색 공간만을 고려하였음에도 불구하고, 결과로 얻은 베이지안 망 구조의 정확성에는 큰 차이가 없는 것을 알 수 있다.

3. 결론 및 향후 과제

본 논문에서는 대규모 베이지안 망의 효율적인 학습을 위한 탐색 공간 제한 기법을 제안하였다. 제안된 방법은 후보 DAG 탐색 공간의 규모를 제한하는 휴리스틱이며, 전체 노드 수를 대상으로 가능한 모든 후보 DAG들을 고려할 때에 비해 훨씬 적은 규모의 후보 DAG만을 고려하는 방법이다.

상대적으로 훨씬 적은 규모의 탐색 공간만을 고려하도록 하는 방법임에도, 실험 결과 기존의 대규모 베이지안 망 학습을 위한 SC 방법과 비슷한 정확도를 보이는 결과를 얻을 수 있음을 알 수 있다.

향후 과제로는, 보다 효율적인 클러스터의 구축 및 클러스터간 DAG 구조 학습 방법이 있을 수 있다. 또한 제안된 방법은 기존 방법들에 비해 훨씬 적은 탐색 공간만을 고려하므로, 반복 학습 혹은 진화 탐색 방법과 같이 보다 정확한 학습을 위한 추가적인 방법의 적용이 가능할 수 있다.

참고 문헌

[1] David Heckerman, Dan Gerger and David M. Chickering, "Learning Bayesian Networks: The Combination of Knowledge and Statistical Data", Machine Learning, vol. 20, pp. 197-243, 1995

[2] Robinson, R. W., "Counting unlabeled Acyclic digraphs", Lecture Notes in Mathematics, Vol. 622, Combinatorial Mathematics V, 1977

[3] D. M. Chickering, "Learning Bayesian networks is NP-complete", AI&STAT V, 1996

[4] Laura E. Brown, Ioannis Tsamardinos, Constantin F. Aliferis, "A Novel Algorithm for Scalable and Accurate Bayesian Network Learning", MEDINFO, 2004

[5] Nir Friedman, Iftach Nachman and Dana Peer, "Learning Bayesian Network Structure from Massive Datasets: The "Sparse Candidate" Algorithm", Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence, pp. 206-215, 1999