대규모 데이터 분석을 위한 계층적 베이지안망 학습

황규백° 김병희 장병탁

서울대학교 컴퓨터공학부

{kbhwang°, bhkim}@bi.snu.ac.kr, btzhang@cse.snu.ac.kr

Hierarchical Bayesian Network Learning for Large-scale Data Analysis

Kyu-Baek Hwang°, Byoung-Hee Kim, and Byoung-Tak Zhang

School of Computer Science and Engineering, Seoul National University

요 약

베이지안망(Bayesian network)은 다수의 변수들 사이의 확률적 관계(조건부독립성: conditional independence)를 그래프 구조로 표현하는 모델이다. 이러한 베이지안망은 비감독학습(unsupervised learning)을 통한 데이터마이닝에 적합하다. 이를 위해 데이터로부터 베이지안망의 구조와 파라미터를 학습하게 된다. 주어진 데이터의 likelihood를 최대로 하는 베이지안망 구조를 찾는 문제는 NP-hard임이 알려져 있으므로, greedy search를 통한 근사해(approximate solution)를 구하는 방법이 주로 이용된다. 하지만, 이러한 근사적 학습방법들도 데이터를 구성하는 변수들이 수천 ~ 수만에 이르는 경우, 방대한 계산량으로 인해 그 적용이 실질적으로 불가능하게 된다. 본 논문에서는 그러한 대규모 데이터에서 학습될 수 있는 계층적 베이지안망(hierarchical Bayesian network) 모델 및 그 학습방법을 제안하고, 그 가능성을 실험을 통해 보인다.

1. 서론

베이지안망(Bayesian network)은 다수의 확률변수(random variable)들의 결합확률분포(joint probability distribution)를 변수들 사이의 조건부독립성(conditional independency)에 기반해 효율적으로 표현하는 확률그래프모델(probabilistic graphical model)이다. 변수집합 $\mathbf{X} = \{X_1, X_2, \ldots, X_n\}$에 대한 베이지안망은 DAG(directed acyclic graph) 형태의 망 구조 G에 기반해 결합확률분포 $P(\mathbf{X})$를 다음과 같이 표현한다.

$$P(\mathbf{X}) = \prod_{i=1}^{n} P(X_i \mid \mathbf{Pa}_G(X_i)) \qquad \text{(수식 1)}$$

위의 식에서 $\mathbf{Pa}_G(X_i)$는 망 구조 G에서 변수 X_i의 부모 노드들의 집합을 나타낸다.

이러한 베이지안망은 기술적(descriptive)인 면과 예측적(predictive)인 면을 모두 가지고 있다. 즉, 각 변수들 사이의 의존관계(probabilistic dependency)를 그래프 구조로 표현함으로써 데이터 영역에 대한 지식을 기술할 수 있으며, 동시에 주어진 결합확률분포로부터 필요한 조건부확률을 추론함으로써 새로운 예제에 대한 예측을 할 수도 있다. 따라서, 베이지안망을 데이터로부터 학습하는 것은 데이터마이닝과 같은 작업에 상당히 유용할 수 있다.

이를 위해서, 베이지안망의 구조와 파라미터를 주어진 데이터로부터 학습하게 된다. 구조의 학습은 NP-hard 문제임이 알려져 있으며[1], likelihood를 목적함수로 하는 greedy search가 주로 이용된다. 하지만, 데이터의 특성(attribute)의 개수가 수천 이상이 되는 경우 이러한 greedy search 방법의 적용도 현실적으로는 불가능하다. 이는 greedy search의 계산량이 변수 개수의 제곱에 비례하며, 탐색 공간의 복잡도도 변수 개수에 따라 엄청나게 증가하기 때문이다.

이러한 문제를 극복하기 위해 greedy search의 탐색공간을 줄이는 기법들이 제시되어 왔다[2]. 하지만 이러한 방법들은 각 변수의 이웃이 될 가능성이 높은 변수들을 특정 기준에[1] 기반해 미리 선정함으로써 지역적인 구조 파악에는 적합하나, 그래프의 전체적인 구조를 파악하기는 어렵다는 단점을 가지고 있다. 또한, 변수의 개수가 수천, 수만에 이르는 경우, 학습된 결과를 가시화하는 것도 어려운 문제가 된다. 본 논문에서는 이러한 단점을 극복하기 위한 계층적 베이지안망(hierarchical Bayesian network) 모델과 그 학습방법을 제안한다. 본 논문에서 제시하는 기법은 실세계의 복잡한 망들이 규칙적인 구조들을 가지고 있다는 사실에 기반한다. 예를 들어 [3]은 재귀적 모듈 구조를 많은 실세계의 복잡한 망들이 가지고 있다는 사실을 밝혀냈다. 이에 기반해 원래 베이지안망에서 서로 가까이에 위치한 변수들을 하나의 모듈로 묶고 이를 대표하는 변수를 둠으로써, 대규모 베이지안망을 모듈 사이의 계층적 관계로 표현하는 방법을 제시한다.

본 논문은 다음과 같이 구성되어 있다. 2장에서는 본 논문에서 제시하는 계층적 베이지안망 모델 및 그 학습기법을 기술한다. 3장에서는 실제 망 구조에서 생성된 데이터를 이용해서 제시한 기법의 성능을 실험적으로 평가하며, 마지막으로 4장에서 결론을 제시한다.

2. 계층적 베이지안망(Hierarchical Bayesian Network)

2.1 모델 정의

문제 영역이 n개의 변수 $\mathbf{X} = \{X_1, X_2, \ldots, X_n\}$으로 이루어져 있다고 가정하자. n개의 변수를 두개씩 묶어서 각 묶음을 하나의 은닉변수(hidden variable)로 표현할 수 있다. 이렇게 만들어진 은닉변수들은 $n/2$개가 된다.[2] 이렇게 만들어진 은닉변수들로 구성된

[1] 상호정보량(mutual information)등이 이용될 수 있다.

[2] n이 홀수인 경우는 하나의 남는 변수는 그대로 위의 계층으로 이동된다.

층을 제1은닉층(hidden layer 1)이라 정의한다. 다시 제1은닉층에서 변수들을 두개씩 묶어서 새로운 은닉변수들로 표현할 수 있다. 이러한 은닉변수들로 이루어진 층을 제2은닉층(hidden layer 2)이라 정의한다. 이러한 과정을 $\log_2 n$번 반복하면 최종적으로 하나의 은닉변수로 구성된 층이 존재하게 되며, 그림 1은 이러한 과정을 예시하고 있다.

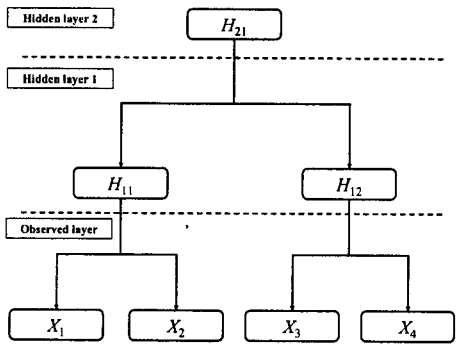

그림 1. 관찰변수(observed variable) 4 개로 이루어진 계층적 베이지안망

그림 1을 보면 은닉층이 거듭될수록 전체 영역을 표현하는 변수의 개수가 작아짐을 알 수 있다. 최종 계층의 하나의 변수는 전체 영역에 대한 확률분포 $P(\mathbf{X})$를 근사화해서 표현하게 된다.

2.2 계층 학습 방법

그림 1의 베이지안망을 M개의 예제로 구성된 완전 데이터(complete data) $D_{\mathbf{X}} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_M\}$으로부터 학습하는 경우를 가정하자. 또한, 각 변수는 이진(binary)이라고 가정한다. 학습은 망의 구조가 고정되어 있기 때문에 다음의 log likelihood 식을 최대화하는 파라미터를 구하는 것이 목표가 된다.

$$L(\Theta) = \sum_{l=1}^{M} \log P(\mathbf{x}_l) = \sum_{l=1}^{M} \log \sum_{\mathbf{H}} P(\mathbf{H}, \mathbf{x}_l) \qquad \text{(수식 2)}$$

위의 식에서 Θ는 계층적 베이지안망을 구성하는 파라미터들의 집합을 나타내며, 그림 1의 예제의 경우 H_{21}에 대한 하나의 사전확률(prior probability)과 나머지 변수들에 대한 조건부확률에 대한 파라미터들로 구성된다. 모든 변수들은 이진이므로 파라미터의 개수는 1 + 2 * 6 = 13개가 된다. (수식 2)의 목적함수를 최대화하기 위해서는 은닉변수들의 값이 주어져 있지 않기 때문에, EM(expectation-maximization) 알고리즘[4]에 의존하게 된다. EM 알고리즘은 아래의 간단한 과정으로 이루어진다.

E-step: 필요한 충분통계량(sufficient statistics)들을 현재 파라미터에 기반하여 추정

M-step: 추정된 충분통계량에 기반한 maximum likelihood 파라미터들의 추정

필요한 충분 통계량은 다음의 식으로 계산된다.

$$E_{\Theta}(N_{ijk}) = \sum_{l=1}^{M} P(x_i^k, \mathbf{pa}_G^j(X_i) \mid \mathbf{x}_l) \qquad \text{(수식 3)}$$

N_{ijk}는 변수 X_i의 부모가 j번째 값을 갖는 경우 X_i의 값이 k가 되는 경우의 회수이다. 위의 식에서 x_i^k와 $\mathbf{pa}_G^j(X_i)$들이 모두 $\mathbf{x}_l$에 주어져 있는 경우는 확률이 0이나 1이 되며, 그렇지 않은 경우는 그림 1의 베이지안망에서의 추론에 의지하게 된다. 계산된 충분통계량을 이용한 maximum likelihood 파라미터들의 추정은 다음과 같이 이루어진다.

$$\theta_{ijk} = \frac{E_{\Theta}(N_{ijk})}{\sum_k E_{\Theta}(N_{ijk})} \qquad \text{(수식 4)}$$

θ_{ijk}는 변수 X_i 부모가 j번째 값을 갖는 경우 X_i의 값이 k가 되는 경우의 확률이다.

그러나 계층적 베이지안망을 학습하는데 위에서 기술한 EM 알고리즘을 그대로 사용하는 경우, 은닉변수가 너무 많기 때문에 지역해(local maxima)에 매우 쉽게 빠지게 된다. 따라서 본 논문에서는 데이터 증가(data augmentation)에 기반해 계층적으로 파라미터들을 학습해 나가는 방법을 이용한다. 이는 아래와 같은 알고리즘으로 이루어진다.

- **For** $i = 1$ **to** 계층의 개수
  - 계층 i에서의 두개의 변수와 계층 $i + 1$에서의 은닉변수로 구성된 각 베이지안망의 파라미터를 학습
  - 학습된 파라미터를 기반으로 계층 $i + 1$의 은닉변수들의 값을 생성하여 M개의 예제로 구성된 새로운 데이터 생성

위의 알고리즘은 각 계층에서는 각각의 베이지안망을 학습하므로 EM 알고리즘으로 학습하는 파라미터의 개수가 적으며, 따라서 그만큼 지역해에 빠질 가능성은 적어진다. 각 계층에서는 각각의 변수쌍에 대한 log likelihood를 최대화하는 방식으로 파라미터들을 학습하게 되며, 이 경우 각 은닉변수들이 자신의 자식인 아래 계층의 변수들의 값을 잘 표현하게 된다. 이는 $P(X_i, X_j)$에 대한 likelihood를 최대화하는 것에 해당한다.

2.3 은닉층에서의 구조학습

2.2절의 계층 학습 과정을 거치게 되면, 각 은닉층의 은닉변수들에 대한 데이터들이 생성되게 된다. 이러한 데이터들과 은닉변수들은 원래 데이터 영역의 변수들에 대한 근사 표현이 된다. 이러한 근사 영역에서의 베이지안망 학습을 통해 우리는 원래 데이터 영역에서의 전체적인 베이지안망 구조를 근사화할 수 있다.

하나의 은닉층에서의 베이지안망 학습에는 기존의 베이지안망 구조학습 알고리즘들을 그대로 이용할 수 있다. 이 과정은 아래의 목적함수에 대한 최적해를 구하는 과정이 된다.

$$L(S_{\mathbf{H}i}, \Theta_{\mathbf{H}i}) = \sum_{l=1}^{M} \log P(\mathbf{h}_l \mid S_{\mathbf{H}i}, \Theta_{\mathbf{H}i}) \qquad \text{(수식 5)}$$

위의 수식에서 $S_{\mathbf{H}i}$와 $\Theta_{\mathbf{H}i}$는 i번째 계층에서의 베이지안망 구조와 파라미터들을 가리킨다. $\mathbf{h}_l$은 계층 $i-1$의 변수들로부터 생성된 데이터의 예제를 가리킨다.

2.4 각 계층에서 변수들을 묶는 방법

각 계층을 올라가면서 효율적 표현을 위해 희생되는 정확도는 $P(H)$와 $P(X_i, X_j)$사이의 괴리이며, 이는 다음과 같이 표현된다.

$$D(P(H)\|P(X_i))+D(P(H)\|P(X_j)) = \sum P(H)\log\left(\frac{(P(H))^2}{P(X_i)P(X_j)}\right) \quad \text{(수식 6)}$$

위의 수식에서 $D(\cdot\|\cdot)$는 relative entropy를 의미하며 이는 확률 분포 사이의 차이에 해당한다. 위 식의 합 계산은 각 변수들이 가지는 값들의 조합에 대해 행해진다. 위의 식을 최소화 하기 위해서는 $P(X_i)$, $P(X_j)$가 서로 비슷해야 한다. 만일 $P(X_i)$와 $P(X_j)$가 동일하다면 (수식 6)에서 하나의 relative entropy를 최소화하는 것으로 $P(X_i, X_j)$의 likelihood를 최대화할 수 있다. 따라서 각 층에서 변수들을 묶을 때 서로의 mutual information이 가까운 순서대로 묶는 것이 정보량의 손실을 최소화할 수 있는 방법이다.

3. 실험

3.1 실험데이터

실험데이터는 scale free 특성[3]을 가지도록 생성된 50개의 노드를 가지는 베이지안망에서 생성했다.

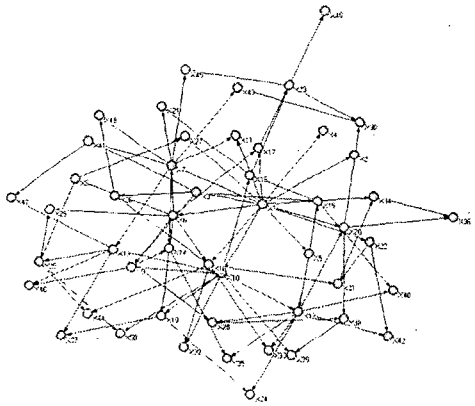

그림 2. 데이터를 생성한 베이지안망

그림 2의 베이지안망에서 probabilistic logic sampling의 방법으로 1000개의 데이터를 생성했다.

3.2 실험결과

우선 묶이는 변수 사이의 상호정보량(mutual information)에 따른 은닉변수에서의 정보손실 정도를 계산해 보았다. 그림 3은 앞에서 생성한 데이터의 변수를 순서대로 묶은 경우{(X_1, X_2), (X_3, X_4), ..., (X_{49}, X_{50})}의 결과를 표시하고 있다. 그림에서 보면 묶이는 두 변수 사이의 상호정보량과 정보손실의 정도는 반비례 관계에 있음을 알 수 있다. 그림 3의 y축은 은닉변수와 관찰변수들 사이의 상호정보량으로 이는 정보손실에 반비례한다. 그림 3의 데이터들 사이에는 양의 상관관계가 존재한다 (상관계수 값은 0.176). 그림 4는 제1은닉층에서 학습된 베이지안망 구조를 보여주고 있다.

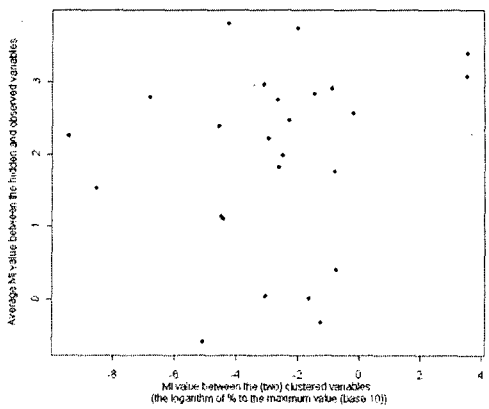

그림 3. 두 변수사이의 상호정보량에 따른 정보량의 손실정도

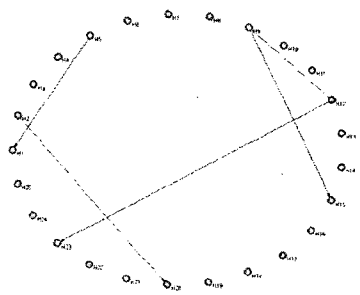

그림 4. 제 1 은닉층에서 학습된 베이지안망 구조

4. 결론

본 논문에서는 변수가 수천에 이르는 대규모 데이터에서 학습할 수 있는 계층적 베이지안망 모델을 제안하고 그 학습방법을 제시하였다. 가상의 베이지안망에서 생성된 데이터를 이용한 결과, 은닉층에서 효율적으로 베이지안망을 구성할 수 있음을 보였다.

감사의 글

이 논문은 교육부 BK21사업, 과학기술부 국가지정연구실사업(NRL)에 의하여 지원되었음. 김병희는 서울과학장학생 사업에 의해 지원받았음.

참고 문헌

[1] Chickering, D. M., Learning Bayesian networks is NP-complete, In D. Fisher and H.-J. Lenz, editors, *Learning from Data: Artificial Intelligence and Statistics V*, pp. 121-130, Springer-Verlag, 1996.

[2] Goldenberg, A. and Moore, A., Tractable learning of large Bayesian net structures from sparse data, In *Proceedings of ICML 2004*, 2004.

[3] Song, C., Havlin, S., and Makse, H. A., Self-similarity of complex networks, Nature, vol. 433, pp. 392-395, 2005.

[4] McLachlan G. J. and Krishnan, T., *The EM Algorithm and Extensions*, *John Wiley & Sons*, 1997.