생물학적 데이터의 베이지안 네트워크 학습에서의 효과적인 스코어링 척도 비교

황성철*, 이일병*
*연세대학교 컴퓨터과학과
e-mail:franz82@csai.yonsei.ac.kr

Comparison of Efficient Scoring Metrics for Bayesian Network Learning in Biological Domain

Sungchul Hwang*, Yillbyung Lee*
*Department of Computer Science
Yonsei University

요 약

본 논문에서는 베이지안 네트워크 학습 방법을 이용한 비교적 적은 양의 샘플 데이터에서 현실적인 네트워크 모델 추론을 위한 효율적인 스코어링 척도를 찾는 것을 목표로 하였다. UPSM, CUPSM, DPSM, BDe(Bayesian Dirichlet) 등을 각각 적용시켜본 결과를 통해 어떤 방법이 가장 적은 샘플의 데이터, 특히 생물학적 데이터에적합한지 알아보았다.

1. 서론

베이지안 네트워크는 수많은 데이터로부터 데이터가 어떠한 이벤트에 의해서 일어나는지, 또는 여러 이벤트 간의 관련성을 파악할 수 있도록 도와준다. 또한, 실세계에서 발생되는 데이터에 대해 쉽게 그 관계를 추론하고 표현할 수 있게 해 준다. 이러한 베이지안 네트워크를 추론하기 위해 K2 같은 많은 알고리즘들이 개발되어 왔다. 주어진 데이터와 베이지안 네트워크 간의 결합 확률 분포를 최대화하는 구조를 찾아내는 것이 이들 알고리즘의 목표이다. 이 과정에서 결합 확률 분포가 최대가 되는 그래프 구조를 찾기 위해 후보 네트워크에 대한 실제 네트워크의 가능성을 계산하는 것을 평가 척도라고 부른다. 서로 다른 평가 척도는 데이터 사전 확률 분포에 대해 조금씩 다른 가정에서 출발하게 되며, 최종적으로 네트워크를 찾아내는 과정에서 이러한 사전 확률 분포의 가정과 스코어링은 결과에 큰 영향을 미치는 중요한 역할을 담당하게 된다. 따라서 최대한 실제 추론하고자 하는 네트워크와 가까운 구조를 찾기 위해서는 데이터의 특성에 맞는 효율적인 스코어링 방법이 요구된다.

본 논문에서 다루고자 하는 데이터의 영역은 생물학 데이터, 특히 마이크로 어레이를 통해 얻어진 유전자 데이터에 대한 것이다. 마이크로어레이(Microarray) 기술은 기존의 분자생물학에서는 불가능했던 엄청난 양의 유전자 분석을 실현할 수 있도록 도와준다. 분석하고자 하는 샘플에 있는 유전자들의 발현 정도가 마이크로어레이 칩에 그대로 나타나게 되고, 이미지 프로세싱을 통해 칩에 나타난 유전자들의 발현 정도를 그대로 수치 데이터로 표현할 수 있다. 이렇게 유전자들의 발현 정보는 수치적으로 나타나게 되며, 이 데이터를 가지고 유전자들의 특성이나 수많은 유전자들을 포함하고 있는 샘플 간

"본 연구는 과기부 뇌신경정보학사업으로터 부분적인 지원을 받아 수행되었음."

의 분석이 최근의 영역에서 주로 이루어지고 있는 연구이다. 그 중에서도 본 논문에서는 특정 세포에 속한 수많은 유전자들이 서로 어떠한 상관관계를 가지고 발현되는지 나타내는 유전자 조절 네트워크를 구성하는 과정에서의 문제를 다루고자 한다.

마이크로어레이로부터 얻어진 데이터는 기본적으로 각각의 데이터에 대해 그것이 어떠한 유전자에 해당하는 것인지만 알 수 있고, 각 데이터간의 구체적인 관계에 대해서는 알 수 없는 상태에 있다. 이처럼 데이터의 불확실성(Uncertainty)에서의 각 개체(유전자) 간의 관계 추론을 위해 사용되는 방법이 베이지안 접근법(Bayesian Approach)이다. 이것은 제약기반 접근법(Constraint-based Approach)과 마찬가지로 마르코프 가정(Markov Assumption)을 기반으로 한다. 하지만, 제약기반 접근법이 범주형 정보를 분석의 결과로 사용하는 반면, 베이지안 접근법은 데이터를 확률적 추론을 위해 사용한다[1].

2. 베이지안 네트워크(Bayesian Network)

베이지안 네트워크(Beyesian Network)는 $B=(G,\theta)$ 로 나타낼 수 있는 확률 그래프 모델(Probabilistic Graphical Model)이다[3]. 여기서, $G=(\nu,\epsilon)$ 는 이러한 베이지안 네트워크의 구조를 표현하는 DAG(directed acyclic graph)이다. ν는 그래프에서의 노드 집합을 나타내며, ϵ은 간선 집합을 나타낸다. ν에 속하는 노드 X에 대해, X로의 직접적인 링크가 있는 것을 부모노드라고 하며, 이를 $pa(X)$로 표기한다. 전체 베이지안 네트워크(BN)에는 N개의 변수 $X_i(1 \le i \le N)$가 존재한다. 조건부 확률 $\theta_{ijk}= P(X_i = k \mid pa(X_i)= j)$ 는 X_i의 부모노드 $pa(X_i)$가 j 의 값을 가질 때, X_i가 k의 값을 가질 확률을 나타낸다. 한편, X_i가 루트 노드일 경우에는 θ_{ijk}가 X_i의 marginal probability에 해당한다. θ는 모든 모수 θ_{ijk}의 집합을 나타낸다.

2.1 모수 추정(Parameter Estimation)

MLE가 베이지안 네트워크를 구성하는 데이터에 잘 맞는 것처럼 보이지만. MLE의 단점은 그것이 특정 데이터에만 오버피팅되는 경향이 있다는 것이다. 적은 양의 데이터에 대해서만 학습을 시키고 추론을 하게 될 경우에 MLE접근법으로는 보다 현실적인 모델과는 다를 수 있다는 것이다. 이러한 문제점을 해결하기 위해서는 모수 추정에 있어서 베이지안 접근법을 사용하는 것이 보다 효율적이다. 베이지안 접근법에서는 기존의 측정된 데이터에 초기 사전분포 P(θ)를 가정하여 모수 추정을 시행하기 때문에 MLE의 문제점을 보완할 수 있는 장점이 있다. 새로운 데이터가 있을 때, 기존의 P(θ)와의 조합으로 사전 분포를 업데이트시켜가면서 초기의 분포를 점차 실제 모델에 맞게 변화되는 과정을 거치게 된다. 이렇게 업데이트 된 분포를 사후분포 P(θ|D)라고 한다[3].

$$P(\theta|D) = \frac{P(D|\theta)P(\theta)}{P(D)}$$

여기에서 P(D)는 주변우도(marginal likelihood)라고 하며, 데이터에서 모든 가능한 모수들의 평균을 나타낸다. 하지만 이것은 정규화 상수(Normalizing Constant)로, θ와는 독립적이기 때문에 네트워크 점수를 계산하는 데에는 반영하지 않는다.

사전 분포 P(Θ)는 경우에 따라 균등분포(Uniform Distribution) 또는 디리클레 사전분포(Dirichlet Prior)를 사용할 수 있다. 균등분포는 모든 경우의 확률적 분포가 동등하게 가정되는 것이며, 디리클레 분포의 경우 사전 분포를 다음과 같이 나타낼 수 있다[3].

$$P(\theta) = \text{Dirichlet}(\alpha_{x^1|u},\ldots,\alpha_{x^K|u}) \sim \prod_j \theta_{x^j|u}^{\alpha_{x^j|u}-1}$$

위의 식에서 사용되는 모수(Hyperparameter) $\alpha_{x^j|u}$는 각각의 $x^j \in Val(x)$ 에 대응되는 것이다. 이는 실험 데이터 D를 적용하기 이전에 가상의 모수로 여기는 것으로, 부모 노드 $Pa_x = u$ 를 가지는 $X = x$가 있을 경우의 수(Pseudo Count)에 해당하는 것이다.

디리클레 사전분포는 전역적 모수 독립성(Global parameter independence)과 지역적 모수 독립성(Local parameter independence)을 만족하는 성질을 가지며, 이 두 가지 독립성이 모두 만족될 때, 사전분포 P(θ)가 모수 독립성을 만족한다고 할 수 있다.

만약 이처럼 P(θ)가 디리클레 분포를 따른다면, 사후분포 P(θ|D)는 Sufficient Statistics M[x,u]에 의해 다음과 같이 표현될 수 있다.

$$P(\theta|D) = Dirichlet(\alpha_{x^1|u} + M[x^1,u],\ldots,\alpha_{x^K|u} + M[x^K,u])$$

$\alpha^* = \sum_j \alpha_{x^j|u}$는 효율적인 샘플 크기를 나타내며, 이것이 클수록 사전분포에 대해 강한 확신을 가지고 있고, 반대로 D가 클수록 사전분포에 대한 확신이 적다는 것을 나타낸다 .M[x,u]는 D에서 X_i가 x의 값을 가지고 $pa(X_i)$가 u의 값을 가지는 경우의 수를 나타낸다.

2.2 구조 학습(Structure Learning)

데이터의 실제적인 모델을 찾기 위해 구조 학습에서는 주어진 데이터에 기반하여 모델을 선택하는 과정을 필요로 한다. 이러한 구조 학습에 있어서의 방법론은 크게 두 가지로 구분할 수 있다. 첫 번째로, 구조 학습을 CSP(Constraint Satisfaction Problem)의 형태로 다루는 제약기반접근법(Constraint-Based Approach)이 있다. 이것은 데이터간의 상호 의존성을 찾아내고자 하는 알고리즘으로 주로 통계학적 가설검증 방법을 사용한다. 두 번째 방법은 구조 학습을 최적화 문제(Optimization Problem)의 측면에서 다루는 탐색과 스코어링에 기반한(Searching and Scoring-Based Approach) 방법이다. 여기서의 구조 학습은 최적화의 문제로서 스코어를 최대로 하는 구조 S_{opt}를 찾는 것이 문제이다. 이 알고리즘에서의 스코어링은 각각의 모델이 주어진 데이터에 대해 얼마나 적합한 형태를 갖는지에 따라 높은 점수를 부여한다. 한편, 데이터의 크기가 큰 경우 이러한 상황에서 쉽게 가장 최적화된 네트워크 구조를 찾는 것은 NP-Hard에 속하는 문제이다[5]. 따라서 스코어링 척도는 각각의 노드에 대해 분할된 구조를 가지는 것이 이러한 문제를 해결할 수 있는 방법이다. 스코어링 척도는 다음과 같이 정의된다[6].

2.2.1 Uniform Prior Score Metric(UPSM)

네트워크의 사전 분포는 균일하게 정의되며. UPSM은 다음과 같이 정의된다.

$$P(B_n, D) = P(B_s)\prod_{i=1}^{n}\prod_{j=1}^{q_i}\frac{(r_i-1)!}{(N_{ij}+r_i-1)!}\prod_{k=1}^{r_i}(N_{ijk})!$$

2.2.2 Conditional Uniform Prior Score Metric(CUPSM)

조건부 균등분포가 사전 확률로 가정되어 있다면, CUPSM은 다음과 같은 식을 가지게 된다. 이는 Beta 함수를 사용하여 계산되며, 가장 간단한 계산 과정을 가진다.

$$P(B_s, D) = P(B_s)\prod_{i=1}^{n}\prod_{j=1}^{q_i}\prod_{k=1}^{r_i-1}B \times (N_{ijk}+1, 1+\sum_{m=0}^{r_i-k-1}N_{ij(r_i-m)})$$

2.2.3 General Dirichlet Prior Score Metric(DPSM)

사전 확률 분포를 디리클레 분포로 가정할 때의 스코어링 척도는 다음과 같은 형태를 가진다.

$$Score_B(G,D) = \sum_i FamScore_B(X_i, pa(X_i) : D)$$

여기서 G는 전체 베이지안 네트워크를 나타내며, D는 주어진 모든 데이터를 나타낸다. 위의 식에서 나타난 것처럼 전체 네트워크의 스코어가 각각의 노드와 관련된(Family)부분에 대한 스코어로 분할하여 계산 가능한 것은 최적의 네트워크 모델을 찾는 데 데 있어서 매우 효율적이다. FamScore는 각각의 변수 X_i에 대한 스코어를 나타내며, 다음과 같이 나타낸다.

$$FamScore_B(X_i, Pa(X_i) : D) = \log\left[\prod_{u\in Pa(X_i)}\frac{\Gamma(\alpha_{x_i|u})}{\Gamma(\alpha_{x_i|u}+M[u])}\prod_{x_{ij}\in X_i}\frac{\Gamma(\alpha_{x_{ij}|u}+M[x_{ij},u])}{\Gamma(\alpha_{x_{ij}|u})}\right]$$

여기서 Γ는 감마 함수를 나타내고, $\alpha_{u_i} = \sum x_i$, $M[u_i] = \sum_{x_i} M[x_i, u_i]$ 를 만족한다.

2.2.4 Bayesian Dirichlet equivalence(BDe)

위의 경우에서 디리클레 사전 분포를 배제하고 대신 Likelihood equivalence 가정을 추가한 것이 BDe 스코어 이다. 여기서의 모수(Hyperparameter)는 Equivalent Structure간의 스코어가 다르지 않도록 하기 위해 다음과 같이 계산된다. BDe는 다른 메트릭들과 다르게 샘플의 수를 결정하기 위해 네트워크 구조에 대한 어느 정도의 사전 지식을 필요로 한다.

$$\alpha_{x_i|u_i} = M' P'(x_i, u_i)$$

한편, 선형 가우시안 변수들에 대해서는 *BGe(Bayesian Gaussian equivalence)* 스코어를 사

용하며, 사전 모수로는 normal-Wishart parameter를 사용한다[4]. 다변량 변수와 가우시안 변수가 동시에 사용되는 경우에는 MAP 또는 ML 모수를 사용한 BIC(Bayesian Information Criterion) 스코어를 사용한다. 그와 같은 경우에 대해 본 논문에서는 다루지 않는다.

3.척도의 비교 및 분석

위에서 설명한 각각의 척도에 대해 비교하기 위하여, 3개의 노드를 가지는 가능한 네트워크 구조 6개를 생성하였다. 이 중의 하나는 실제 데이터를 표현할 수 있는 네트워크의 구조이고, 나머지 다섯 개는 실제 네크워크 구조와 다른 형태를 가지고 있다. 각각의 척도를 사용하여, 실제 네트워크와 가장 가까운 값을 갖는 나머지 네트워크에 대한 평가를 진행하였고, 그 비율을 다음 그래프와 같이 나타내었다. 본 실험에서 사용된 샘플의 개수는 300개이다.

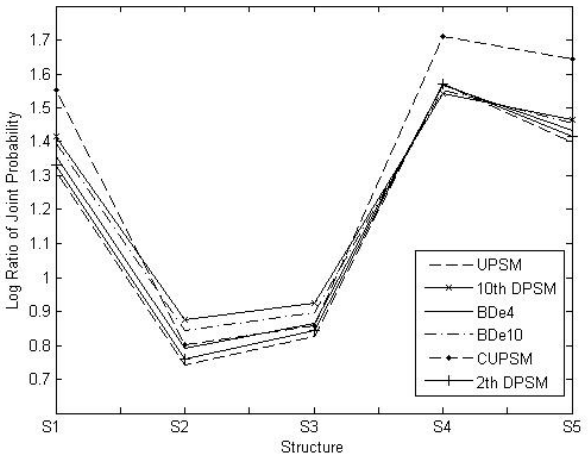

그림 2. 평가 척도 간의 실제 네트워크와의 평가 비율 비교

각각의 구조를 테스트한 결과 모든 평가 척도에서 S3구조가 실제 네트워크와의 평가 점수 비율이 1에 가까운 비슷한 구조로 나타났다. 6개의 사용된 척도 중 BDe4, BDe10이 평가 척도를 사용한 것이 적은 샘플의 수에서도 가장 실제 네트워크에 가까운 구조에 대해 비교적 높은 평가 성능을 가지는 것으로 나타났다.

4. 결 론

본 논문에서는 베이지안 네트워크를 평가하는 방법에 있어서 사용되는 몇 가지 척도들을 마이크로어레이 데이터의 상대적으로 적은 샘플의 수를 고려하여 비교해 보았다. 전체적인 베이지안 네트워크를 학습하는 것은 본 논문에서 다룬 척도를 기반으로 가장 실제 네트워크에 가까운 모델을 찾아가는 탐색 알고리즘이 추가 되어야할 것이다. 앞으로의 과제는 이처럼 사용된 척도와 탐색 알고리즘을 결합하여 효과적인 네트워크 구성 시스템을 구축하는 것이다. 또한 생물학적 데이터의 특성상 완전한 데이터란 존재하지 않으므로 이러한 점도 고려하여, 은닉 데이터에 대한 추정 및 추가 작업을 통해 보다 현실적인 형태로 접근해야 할 것이다.

참고문헌

[1] D. Heckerman, C. Meek, and G. Cooper A Bayesian Approach to Causal Discovery. In C. Glymour and G. Cooper, editors, Computation, Causation, and Discovery, pages 141-165. MIT Press, Cambridge, MA, 1999.

[2] Heckerman, D., Geiger, D., and Chickering, D., Learning Bayesian networks: The combination of knowledge and statistical data, Machine Learning, pp. 197-245, 1995

[3] K. Sivakumar, R. Chen, and H. Kargupta, Learning Bayesian Network Structure from Distributed Data. In Proceedings of the 3rd SIAM International Data Mining Conference, pp. 284-288, 2003

[4] D. Geiger, D. Heckerman, Learnging gaussian networks. In UAI '94, pp 235-243, 1994

[5] D.M. Chickering, Learning Bayesian networks is NP-Complete. In Learning from Data, Artificial Intelligence and Statistics, 1996

[6] G.F.Cooper and E.Herskovits, "A Bayesian method for the induction of probabilistic networks from data",Machine Learning, vol. 7, pp.299-347, 1992