MDL Principle을 적용한 점수 기반 베이지안 네트워크 학습 방법

황성철° 이일병
연세대학교 컴퓨터과학과
{franz82°, yblee}@csai.yonsei.ac.kr

A Score-Based bayesian network learning method by adopting Minimum Description Length principle

Sungchul Hwang°, Yillbyung Lee
Dept. of Computer Science, Yonsei University

요 약

본 논문에서는 파라미터에 대한 정보가 없는 데이터, 즉, 각각의 이벤트 발생에 불확실성이 존재하는 데이터들에 대한 인과 관계의 학습을 위해 그래픽 모델인 베이지안 네트워크를 사용하였다. 이를 위해 기존에는 주로 네트워크 학습에 K2, Sparse Candidate 등의 방법이 사용되었다. 학습 및 추론에 있어서 어떻게 하면 기존의 방법보다 정확하고 빠르게 처리할 수 있을지에 대한 개선된 알고리즘을 제시하고 다른 알고리즘들과의 성능 비교를 통해 제시한 방법론이 보다 좋은 성능을 가짐을 보였다.

1. 서 론

기계학습(Machine Learning)을 통해 받아들이고 분석하는 실세계에서 일어나는 이벤트 중에는 항상 불확실성(Uncertainty)이 존재한다. 어떤 이벤트들이 어떻게 발생하게 되는지에 대한 아무런 정보가 없는 이러한 불확실성이 존재하는 상황에서의 학습은 경험적, 통계학적 추론을 통해 이루어질 수 있다. 인간이 학습하는 과정을 모방하기 위해 컴퓨터에서는 여러 가지 이벤트들에 대한 원인 관계 분석과 이벤트 간의 확률을 활용하여 모든 이벤트들의 상관관계를 학습한다. 실제로 이러한 모든 데이터들은 각각의 확률과 그에 대한 의존성을 따른다. 특히 다른 이벤트와의 조건부 의존성(conditional dependencies)은 이벤트 간의 가장 기본적인 연관성을 나타내주기 때문에 학습에 매우 중요한 부분으로 사용된다[3]. 이러한 특성을 학습에 적용하고 표현하기 위해 사용하는 방법이 확률 그래픽 모델(Probabilistic Graphical Model)이며, 여러 이벤트들에 대한 데이터 간의 관계분석이 요구되는 의사결정 시스템, 의료진단, 패턴 인식, 에이전트 시스템, 바이오인포매틱스(Bioinformatics) 등의 여러 분야에서 활용되고 있다.

"본 연구는 산자부 뇌신경정보학 사업으로부터 지원을 받아 수행되었음."

기존에 베이지안 네트워크 학습을 위해 사용되는 네트워크 탐색 방법으로는 Greedy Hill-climbing, Simulated Annealing, k2, Sparse Candidate 등의 방법이 있다. 하지만, 탐색의 효율성이나, 복잡도(complexity) 측면, 정확성의 측면에서 많은 문제점을 가지고 있는 것이 사실이다. 따라서 본 논문에서는 그와 같은 문제점들을 해결하기 위해 탐색공간을 줄이는 데에 Information theory를 사용한 이벤트들 간의 연관성 랭크 척도를 사용하였고, 그에 추가로 MDL(Minimum Description Length) Principle을 사용하여 효율적이고 보다 정확한 결과 네트워크를 생성하였다.

2. 베이지안 네트워크

베이지안 네트워크는 두 가지의 중요한 요소로 이루어진다. 네트워크를 구성하는 노드들간의 상호의존성을 표현하는 DAG(Directed Acyclic Graph) 구조와 부모 셋(Parents set)이 주어졌을 때의 각 노드에 대한 조건부 확률 테이블이다. 데이터의 조건부 확률 분포에 대한 정보가 없는 경우의 이를 알아내기 위한 베이지안 네트워크 학습에서 중요한 요소는 각 이벤트에 대한 실제 샘플 데이터로, 이를 바탕으로 네트워크에 대한 파라미터 학습을 수행하게 되고, 구조 학습(Structure Learning)을 통해 추론을 할 수 있는 능력을 갖게 된다[4]. 이처럼 구

조학습을 하기 위해서는 평가 척도(Scoring Function)와 최대 평가 점수를 가지는 네트워크를 찾기 위한 탐색 기법(Search Method)을 필요로 한다[5]. 이처럼 본 논문에서 적용할 학습 방법의 전체적인 형태는 점수기반(Score-based) 학습 방법이다. 이 방법은 생성된 네트워크와 실제 받아들여진 데이터와의 매칭 정도에 따른 점수를 최적화 하는 것이다. 즉, 학습 과정 중 실제 데이터와 생성한 네트워크와 얼마나 잘 매치되는지에 따라 구조를 평가하게 되고, 그에 따라 점수가 반영되는 방식이다. 최종적인 점수기반 학습 방법의 목적은 가장 높은 점수를 얻는 네트워크 구조를 찾아내는 것이라고 할 수 있다. 네트워크의 평가는 각 노드와 해당 노드의 부모노드를 포함하는 Family Set의 점수를 모두 합하여 나타낼 수 있다. 한편, 각 노드에서의 Family set과의 Family score를 계산하는 것은 부모 노드에서 해당노드로의 ADD, DELETE, REVERSE 연산을 모두 수행하면서 가장 높은 스코어를 갖는 것을 실제 Family Score에 반영하게 된다.

2.1 네트워크 평가 척도(Network Scoring Metrics)

네트워크를 평가하고 평가된 네트워크를 선택하기 위해 사용되는 기존의 평가 척도들은 다음과 같은 것들이 존재한다. 실제 네트워크 탐색에 있어서 이러한 평가 척도를 사용하여, 실제 데이터와 학습되는 네트워크가 얼마나 잘 매치될 수 있는지 평가하는 기준이 된다. 본 논문에서는 BDe, AIC, BIC, KL distance 평가척도만을 다루기로 한다.

2.1.1 Bayesian Dirichlet equivalence(BDe)

노드의 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 데에 있어서 매우 효율적이다. FamScore는 각각의 변수 X_i에 대한 스코어를 나타내며, 다음과 같이 나타낸다.

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

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

여기서 Γ는 감마 함수를 나타내고, $\alpha_{x_i|u_i} = M'P'(x_i, u_i)$를 만족한다[6].

2.1.2 Information criterion

information criterion을 적용한 평가 척도에는 AIC(Akaike's information criterion)와 BIC(Bayesian information criterion)가 있다.

$$AIC = -\log l(\hat{\theta}|x) + K$$
$$BIC = -\log l(\hat{\theta}|x) + \frac{1}{2}K\log m$$

여기서 $\hat{\theta}$는 파라미터 벡터에 대한 추정치를 나타내고 $l(\bullet)$은 샘플 x가 주어졌을 때의 likelihood function $P(D|\hat{G})$을 나타낸다. 또한 AIC와 BIC의 차이를 주는 m은 샘플의 크기로써 BIC에서는 이처럼 샘플의 수를 반영하여 페널티를 적용하는 척도로 나타내기 때문에 샘플의 수에 따른 Overfitting을 방지할 수 있지만, AIC는 그러한 점을 반영하지 않는다.

2.1.3 Kullback-Leibler Distance(cross-entropy)

확률이론에서 Kullback-Leibler Distance는 실제의 확률분포 P와 임의의 확률분포 Q의 distance를 나타내는 척도로 사용된다. 보통 P는 측정된 데이터를 나타내고 Q는 그러한 데이터를 추정하기 위해 만들어진 모델을 나타낸다고 볼 수 있다. 정수값을 가지는 확률 분포에서는 다음과 같이 KL-Distance를 나타낼 수 있다[7].

$$D_{KL}(P\|Q) = \sum_i P(i)\log\frac{P(i)}{Q(i)}$$

2.2 기존의 구조 학습 알고리즘

2.2.1 K2

k2는 점수기반 greedy search 알고리즘으로 역시 베이지안 메소드를 사용하는 학습 방법이다. k2는 다른 구조들을 평가하기 위해 베이지안 스코어 $P(B_s, D)$를 사용하며, 이를 최대화 하는 알고리즘으로 Greedy search를 사용한다. 한편, K2의 중요한 요소로 작용되는 것이 노드들의 순서인데, 부모노드는 자식노드보다 반드시 앞의 순서에 위치하고 있어야 한다는 것이다. (예를 들자면, 노드 X_i가 X_j의 앞에 온다면 X_j는 X_i 의 부모노드가 될 수 없다.) 그리고 알고리즘에서는 각 노드를 돌면서 $score_B(d, X_i, PA)$를 최대화 하는 노드를 찾아 부모노드 셋에 포함시키는 과정을 반복한다. 하지만 K2 알고리즘의 단점은 노드 순서(ordering)에 따라 결과의 차이가 크며, 가장 최적화 된 순서라는 것이 알려져 있지 않기 때문에 그러한 순서를 찾아내는 것이 쉽지 않다는 것이다 [2].

2.2.2 Sparse Candidate 알고리즘

Sparse Candidate 알고리즘은 기본적으로 전체 탐색 공간을 줄여나가면서 학습을 진행하는 것이 요점이다. 모든 노드에 대한 탐색과 학습을 진행하는 것은 이미 NP-Hard[1]의 문제로 여겨지고 있기 때문에 탐색공간을 줄여가면서 학습을 진행하는 것은 매우 효율적인 방법이라 할 수 있다. 각 노드마다 Mutual Information과 같은 척도를 사용하여, 후보 부모 집합(Candidate parents set)을 제한된 수로 저장하여, 이들 후보 노드만을 대상으로 하여 네트워크 평가 점수에 반영하는 방식이다.

2.2.3 개선된 알고리즘

다른 기존의 알고리즘들의 정확성과 시간복잡도를 개선하기 위하여 Sparse Candidate 알고리즘에 MDL(Minimum Description Length) 개념을 추가하였다. Sparse Candidate 알고리즘의 Restrict Step은 각 노드별로 Mutual Information 또는 다른 척도를 적용하여, 해당 노드의 부모노드가 될 후보 노드를 K개를 설정하는데, 이 부분에서 MDL 척도를 추가하여 가장 높은 값을 갖는 노드부터 순서대로 저장하여 이전에 사용된 척도와 MDL 척도 값의 합이 가장 큰 노드를 부모노드로 선택하고 이후에 업데이트하는데 반영하였다. MDL척도를 사용할 경우, 네트워크 모델의 복잡성과 데이터에 대한 적합성을 만족하는 측면에서 상호작용을 기대할 수 있다[8].

$$L_M = \Gamma \sum_{i=1}^{n} \left\{ d_i |Pa(x_i)| + d_f q^{|Pa(x_i)|}(q-1) \right\}$$

$$L_D = \sum_{i=1}^{m-1} H(x_i | pa(x_i))$$

여기서 MDL criterion L은 $L_M + L_D$로 나타낼 수 있다.

3. 실 험

실험은 앞서 소개된 네트워크 평가 척도들과 기존의 네트워크 탐색 알고리즘들을 비교 실험하였다. Initial Network는 탐색 알고리즘이 시작되기 이전의 초기 형태로 노드 사이에 어떠한 간선(edge)도 추가(add)되지 않은 상태를 나타낸다. 그리고 k2알고리즘, Sparse Candidate Algorithm을 제안한 탐색기법과 비교하여 BDeu, AIC, BIC, KL Distance 평가척도를 각각 적용하였다. 실험 데이터로는 10000개의 Sample을 가지는 Alarm Network Data를 사용하였다.

Score	BDe	AIC	BIC	KL
Initial Network	-162032	-162454	-163271	0.00
K2	-108888	-108905	-110957	2.326
Sparse Candidate	-103021	-106203	-108032	2.483
Proposed Algorithm	-102990	-105933	-107801	2.491

<표. 1> 각 알고리즘에 의해 생성된 결과 네트워크의 스코어 비교

<표 .1>은 각 알고리즘에 의해 생성된 결과 네트워크의 최종 스코어를 보여준다. 결과에서는 제안한 알고리즘 Sparse Candidate 알고리즘과 큰 차이는 없으나 이 4가지 모든 척도에서 좀 더 나은 성능을 보임을 알 수 있다.

Time(sec)	BDeu	AIC	BIC	KL
K2	79.3	86.3	77.3	76.2
Sparse Candidate	60.5	71.8	59.9	52.3
Proposed Algorithm	48.2	59.2	45.3	47.2

<표. 2> 각 알고리즘의 학습 시간 비교

<표. 2>에서는 각 알고리즘과 4가지 척도를 적용한 최종 학습시간을 나타낸다. 제안한 알고리즘에 BIC 척도를 적용했을 경우 가장 빠른 학습시간을 보여주었다.

Accuracy	False Alarm	Miss Error	Hamming Distance
K2	12	8	20
Sparse Candidate	6	2	8
Proposed Algorithm	4	2	6

<표. 3> 각 학습 알고리즘과 BIC 척도를 적용하여 생성된 결과 네트워크와 실제 네트워크와의 비교를 통한 정확성 측정

<표. 3>에서는 BIC 척도에 각 알고리즘을 적용하여, 생성된 네트워크의 정확성을 측정하기 위하여 False Alarm과 Miss Error를 사용하였고, 이 두가지 요소를 합하여 보다 간단하게 처리할 수 있도록 Hamming Distance로 나타내었다. Hamming Distance는 False Alarm과 Miss Error의 합을 나타낸다.

4. 결론 및 향후 과제

위의 실험결과 Sparse Candidate 알고리즘의 개선된 형태인 본 논문에서 제안한 알고리즘이 좀 더 나은 성능을 보임을 알 수 있었다. 전체 학습 시간의 측면에서나 정확성에서 큰 차이는 아니지만 나은 성능을 보여주었고, 다른 네트워크 척도를 비교 실험한 결과에서도 가장 나은 스코어를 기록함을 보였다. 하지만, 보다 정확하고 빠른 알고리즘을 구현하기 위해서는 많은 개선점이 필요할 것으로 보인다. 또한 특정 데이터에 국한된 것이 아닌 Microarray data나 다른 도메인에서도 적용이 가능한지에 대한 실험이 좀 더 필요할 것으로 보인다. 그러가 위해서는 연속형 데이터를 처리하기 위한 네트워크 평가 척도의 변형이 필요할 것이며, 좀 더 큰 범위의 데이터를 다루기 위한 방법도 모색해야 할 것이다. 또한 데이터에 대한 사전 지식이 있는 경우 이를 적용한다면 정확성을 더욱 높일 수 있을 것으로 보인다.

참고문헌

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

[2] 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

[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. 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.

[5] D. Heckerman. A tutorial on learning with Bayesian networks. In M. I. Jordan, editor, Learning in Graphical Models, 1998

[6] Shulin Yang, Kuo-Chu Chang, Comparison of Score Metrics for Bayesian Network Learning, IEEE Transactions on Systems, Man and Cybernetics-Part A:Systems and Humans, VOL.32, No.3, 2002

[7] S. Kullback and R. A. Leibler. On information and sufficiency. *Annals of Mathematical Statistics*, 22(1):79-86, March 1951.

[8] Feder, M. Maximum entropy as a special case of the minimum description length criterion. IEEE Transactions on Information Theory 32 (6), 847{849.. 1986