퍼지 및 지능시스템학회 논문지 2007, Vol. 17, No. 5, pp. 684-690

클러스터간 조건부 확률적 의존의 방향성 결정에 대한 연구

Determining Direction of Conditional Probabilistic Dependencies between Clusters

정성원* · 이도헌* · 이광형**

Sungwon Jung, Doheon Lee, Kwang H. Lee

* 한국과학기술원 바이오 및 뇌 공학과

** 한국과학기술원 바이오 및 뇌 공학과, AITrc

요 약

본 논문은 확률변수들로 이루어진 클러스터의 집합과 확률변수들에 대해 관찰된 데이터가 주어진 상황에서, 클러스터 사이에 존재하는 조건부 확률적 의존의 방향성(directional tendency of conditional dependence in the Bayesian probabilistic graphical model)을 결정하는 방법을 기술한다. 클러스터 사이에 존재하는 조건부 확률적 의존의 방향성을 추정하기 위해 한 클러스터에서 다른 각 클러스터에 가장 가까운 확률변수를 해당 클러스터의 외부연결변수로 결정한다. 외부연결변수들 사이에서의 가장 확률이 높은 조건부 확률적 의존성을 나타내는 방향성 비순환 그래프(directed acyclic graph(DAG))를 찾음으로써, 주어진 클러스터들 사이에 존재하는 조건부 확률적 의존의 방향성을 결정한다. 사용된 방법이 클러스터 사이에 존재하는 조건부 확률적 의존의 방향성을 유의미하게 추정할 수 있음을 실험적으로 보인다.

키워드 : 오더 추정, 베이지안 망, 조건부 확률적 의존성

Abstract

We describe our method to predict the direction of conditional probabilistic dependencies between clusters of random variables. Selected variables called 'gateway variables' are used to predict the conditional probabilistic dependency relations between clusters. The direction of conditional probabilistic dependencies between clusters are predicted by finding directed acyclic graph (DAG)-shaped dependency structure between the gateway variables. We show that our method shows meaningful prediction results in determining directions of conditional probabilistic dependencies between clusters.

Key Words : order prediction, Bayesian network, conditional probabilistic dependency

1. 서 론

어떤 시스템을 분석하고자 하는 경우, 시스템을 구성하고 있는 각 요소들 사이의 상호 연관성을 분석하기 위해 다양한 방법들이 사용되어지고 있다. 그러한 방법들 중, 베이지안 망을 이용한 방법은 각 요소들을 확률변수로 간주하고 확률변수들 사이의 조건부 확률적 의존성(또는 독립성)을 표현하기 위한 모델로서 널리 사용되어지고 있다.[1,2,3] 베이지안 망 응용의 초기 단계에서는 주로 전문가가 직접 데이터를 분석한 이후, 분석 결과에서 추정되는 상호 연관 관계를 표현하기 위한 그래프 형태의 모델로 베이지안 망을 사용하였다.[4,5,6,7] 따라서 이미 만들어진 베이지안 망을 이용해 특정 사건의 확률을 추론하는 방법에 대한 연구가 널리 진행되었다.[8,9,10] 그러나 베이지안 망이 가진 표현력이 주목받으면서, 주어진 학습 데이터로부터 자동적으로 최적의 베이지안 망을 학습하기 위한 연구가 널리 수행되었다.[11,12,13,14]

베이지안 망 B는 확률변수 간의 조건부 확률적 의존 관계를 나타내는 방향성 비순환 그래프 (DAG) G와, 주어진 조건부 의존 관계에 대한 확률값을 기술하기 위한 매개변수의 집합 Θ로 다음과 같이 구성된다.

$$B=(G,\Theta)$$

주어진 학습 데이터의 집합 D로부터 최적의 베이지안 망을 학습하는 과정은 최적의 그래프 구조 G와 매개변수 집합 Θ를 학습하는 부분으로 나뉘어진다. 만약 그래프 구조 G가 주어져 있다면 매개변수의 집합 Θ를 학습하는 것은 상대적으로 용이한 일이 된다. 사전에 주어진 조건부 의존 관계 G에 따라, 관찰된 데이터 D로부터 해당 조건부 관계들의 확률값을 추정하면 되기 때문이다. 따라서 주어진 학습 데이터로부터 최적의 베이지안 망을 학습하는 주된 부분은 최적의 그래프 구조 G를 학습하는 과정이다. 주어진 학습 데이터 D로부터 최적의 그래프 구조 G를 학습하는 다양한 방법이 제

접수일자 : 2007년 4월 16일
완료일자 : 2007년 10월 1일
본 논문은 과학기술부 시스템생물학 연구사업(2005-00343)의 지원과 과학기술부 국가지정연구실 사업(2005-01450)의 지원으로 수행되었음. 연구시설은 정문술 바이오정보전자센터의 도움을 받았음.

안되어 왔으나, 일반적으로 $P(G|D)$를 최대로 하는 G를 구하는 과정이 널리 사용되어지고 있다.

대상 시스템의 요소, 즉 G의 각 노드를 구성하는 확률변수의 수가 작은 경우 최적의 구조를 학습하는 것은 일반적인 탐색 알고리즘으로 수행이 가능한 작은 규모의 문제가 된다. 그러나 확률변수의 수가 늘어남에 따라 가능한 그래프 구조의 수가 큰 폭으로 증가하기 때문에, 대부분의 구조 학습에서는 휴리스틱 탐색을 사용한다. 그러한 휴리스틱 방법들 중 하나로 1)확률변수들 사이의 오더가 주어진 것으로 가정한 방법들이 있다.[11,15,16,17] 확률변수들 사이의 오더가 주어지는 경우 가능한 방향성 비순환 그래프(DAG)의 수는 큰 폭으로 줄어들게 되는 장점이 있다. 그러나 확률변수들 사이의 오더가 주어지는 경우가 실제 문제에서는 거의 없다는 단점이 있다.

본 논문에서는 확률변수들 사이의 오더를 추정하는 데에 응용되어질 수 있는 방법으로, 확률변수 클러스터 사이에 조건부 확률적 의존 관계의 방향성을 추정하는 방법을 사용한다. 확률변수 클러스터 사이에 존재하는 조건부 확률적 의존 관계의 방향성 추정을 위해, 각 클러스터를 다른 클러스터들과 의존적으로 연결하는 외부연결변수를 결정한다. 그리고 외부연결변수들 사이의 조건부 확률적 의존성 구조를 주어진 학습 데이터를 이용해 학습함으로서 클러스터간의 조건부 확률적 의존성 구조를 추정한다. 본 연구의 방법은 엄밀히 말해 클러스터간의 인과적 오더를 추정하지는 않으나, 조건부 확률적 의존의 방향성 구조를 인과적 오더로 해석하는 경우가 일반적이므로 오더 추정을 위한 방법으로 사용되어질 수 있다.

본 논문은 다음과 같이 구성된다. 2장에서는 확률변수 클러스터간의 조건부 확률적 의존성 추정 문제를 정의하고, 클러스터간 조건부 확률적 의존의 방향성을 추정하기 위한 휴리스틱 방법을 기술한다. 3장에서는 다양한 벤치마크 베이지안 망을 이용하여, 휴리스틱 방법이 클러스터간 조건부 확률적 의존 관계의 방향성을 결정하는 데에 유용하게 사용되어질 수 있음을 실험적으로 보인다. 마지막으로 4장에서 본 연구의 결론과 향후 과제를 언급한다.

2. 클러스터간 조건부 확률적 의존성

2.1 접근 방법

주어진 확률변수들의 집합 V에 속한 확률변수들 v_i로 이루어진, 다음을 만족시키는 확률변수 클러스터들의 집합 C가 있다고 하자.

$$C=\{C_1, C_2, ..., C_\alpha\}$$
$$C_i \cap C_j = \varnothing \quad (i \neq j)$$
$$\bigcup_{i=1}^{\alpha} C_i = V$$

마르코프(Markov) 조건은 다음과 같이 정의된다.

정의 1. 어떤 집합 V에 속한 확률변수들의 결합확률분포 P와 방향성 비순환 그래프 (DAG) $G=(V,E)$가 있다고 가정하자. 주어진 G와 G상에서 P를 기술하기 위한 확률매개변수 Θ의 쌍 (G,Θ)에 대해, 만약 각 변수 $v_i \in V$가 G에서 자신의 모든 부모 변수들(parent random variables)이 주어졌을 때 다른 모든 비 하위 변수들(nondescendent variables)에 대해 조건부 독립이라면 (G,Θ)는 마르코프(Markov) 조건을 만족한다.

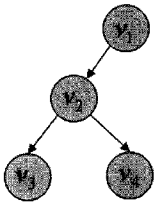


그림 1. 마르코프(Markov) 조건을 따르는 망 구조
Fig. 1. A DAG structure which follows the Markov condition.

그림 1의 그래프 구조를 예로 들면, 그림의 그래프 구조와 임의의 확률매개변수 집합 Θ가 마르코프(Markov) 조건을 만족시킨다고 가정하자. 이 때 마르코프(Markov) 조건의 정의에 의해, v_3는 v_2조건부로 v_1과 v_4에 대해 독립이 된다.

여기에서 주어진 확률변수의 클러스터들이 다음과 같은 조건을 만족시킨다고 가정한다.

가정 1. 주어진 클러스터들의 결합 확률 분포가 마르코프(Markov) 조건을 만족시키는 결합 조건부 확률 분포로 기술될 수 있다고 가정한다.

이러한 경우, 본 연구의 목표는 주어진 C와 D에 대해 그러한 결합 조건부 확률 분포를 기술하는 DAG 구조 중 가장 높은 확률, 즉 가장 높은 $P(G_C|D)$를 갖는 G_C를 찾는 것이다.

확률 변수의 집합 V에 속한 v_i들에 대한 방향성 비순환 그래프 (DAG) 구조 G_V의 경우, $P(G_V|D)$를 구하기 위한 방법으로는 Heckerman의 방법 등이 있다.[14] 그러나 본 연구에서 고려하는 문제는 확률변수 클러스터들을 대상으로 삼고 있으므로 기존의 방법을 그대로 사용할 수 없다. 따라서, 본 연구에서는 $P(G_C|D)$를 구하기 위한 방법으로 각 클러스터를 선택된 몇 개의 확률변수로서 표현하는 방법을 사용한다. 이를 위해 각 클러스터를 다른 클러스터와 연결해 주는 외부연결변수를 클러스터별로 사전에 결정하고, 선택된 외부연결변수들의 확률분포로서 클러스터의 확률분포를 대신하게 한다. 그리고 클러스터간 오더 (order between clusters in the graph where nodes correspond to clusters) 추정을 위해, 외부연결변수들 간의 결합 조건부 확률 분포로써 클러스터간의 결합 조건부 확률 분포를 추정하는 방법을 사용한다(그림 2).

클러스터들의 확률 분포가 마르코프(Markov) 조건을 만족시키는 방향성 비순환 그래프 (DAG) 구조 G_C를 통해 기술될 수 있다고 가정할 때 (가정 1), 클러스터들의 결합 확률분포 $P(C_1, C_2, ..., C_\alpha)$는 다음과 같이 각 클러스터들의 조건부 확률 분포를 통해 기술되어질 수 있다.

1) 확률변수들을 서로 방향성 연결선(directed edge)으로 연결하였을 때의 그래프에서 존재하는 확률변수 사이의 오더(order)를 의미한다.

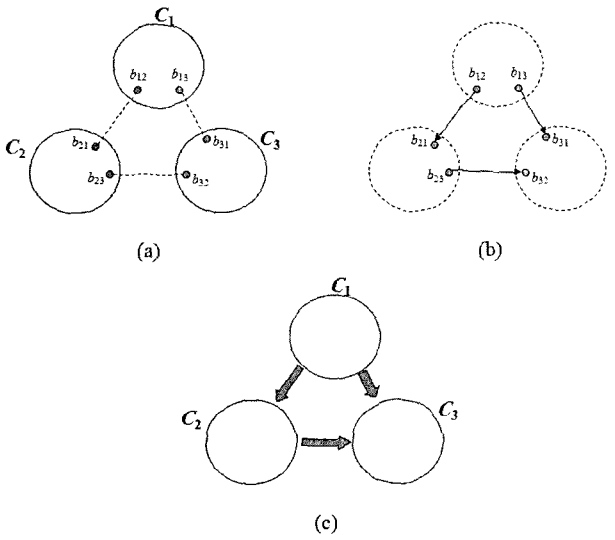


그림 2. (a) 각 클러스터 내부에서의 외부연결변수 결정 (b) 외부연결변수 사이에서, 동일 클러스터 내에 edge를 연결하지 않으며 베이지안 망 구조 학습 (c) 외부연결변수 사이의 베이지안 망 구조를 클러스터 사이의 베이지안 망 구조로 사상

Fig. 2. (a) Determining gateway variables in each cluster (b) Learning Bayesian network structure between gateway variables with no edge in the same cluster (c) Transferring the DAG structure between gateway variables to the DAG structure between clusters

$$P(C_1, C_2, ..., C_\alpha) = \prod_{i=1}^{\alpha} P(C_i | Pa(C_i))$$

$Pa(C_i)$는 G_C에서 C_i가 갖는 부모 클러스터들의 집합을 나타낸다. 따라서 각 클러스터가 갖는 조건부 확률 분포는 해당 클러스터의 부모 클러스터들이 주어진 경우 다른 클러스터들에 대해 독립이다. 또한 그러한 각 클러스터의 조건부 확률 분포는 해당 클러스터와 부모 클러스터간의 의존성 연결(G_C에서 두 클러스터를 연결하는 연결선)에 의해 결정된다.

클러스터들의 확률 분포가 마르코프(Markov) 조건을 따르는 방향성 비순환 그래프 (DAG) 구조에 의해 기술될 수 있다는 가정에 더해, 각 클러스터들이 다음과 같은 조건들을 만족시킨다고 가정한다.

가정 2. 주어진 확률변수들의 클러스터들은 각 클러스터 내부에 존재하는 확률변수들 사이에서의 확률적 의존성을 최대화하는 동시에, 서로 다른 클러스터에 속한 확률변수들 사이에서의 확률적 의존성을 최소화하는 방향으로 클러스터링한 결과이다.

가정 3. 확률변수들 사이의 결합 조건부 확률 분포를 최적으로 기술하는 방향성 비순환 그래프 (DAG) 구조 G_V는 성긴 (sparse) 구조를 갖는다.

가정 2는 두 확률변수 사이의 확률적 의존성을 어느 정도 측정할 수 있는 여러 가지 척도(e.g., 상호 정보 (mutual information) 척도, 상호 연관 (correlation) 척도, X-square 테스트 등)를 유사도 척도로 사용한 클러스터링 방법을 수행함으로써 충족시킬 수 있다. 또한 우리가 다루는 대부분의 대상 시스템이 갖는 이상적인 구조는 각 확률변수들이 의존하는 부모 확률변수의 수가 전체 확률변수 수에 비해 상대적으로 작은 성긴 구조를 갖는 경우가 일반적이므로 가정 3도 충족되는 경우가 대부분이다. 이러한 경우, 클러스터들 간의 조건부 확률적 의존성은 서로 다른 클러스터간의 확률적 의존성이 최소화되어 있게 된다. 또한 전체 망 구조가 성긴 형태라는 가정을 통해 클러스터간 확률적 의존성이 서로 가까운 외부 연결 변수들을 통해 결정될 것이라는 점을 추정할 수 있다. 본 연구에서는 각 클러스터로부터 그러한 외부 연결 변수들을 결정한 후, 외부연결 변수간의 결합 조건부 확률 분포를 최적으로 기술하는 망 구조를 학습하는 방법을 통해 클러스터간 조건부 확률적 의존성의 방향을 추정하는 방법을 사용한다.

2.2 클러스터간 확률적 의존성의 방향을 추정하는 휴리스틱 방법

확률변수간의 결합 조건부 확률분포를 최적으로 기술하는 방향성 비순환 그래프 (DAG) 구조 G_V가 있다고 가정하자. 우리가 다루는 문제가 앞의 2.1절에서 언급한 세 가지 가정을 만족시킨다고 할 때, 주어진 확률변수 클러스터들의 집합 C에 속한 서로 다른 클러스터 C_i와 C_j 사이의 조건부 확률적 의존성은 G_V에서 각 클러스터를 연결하고 있는 소수의 확률변수들인 '외부연결 변수'들에 의해 주로 결정될 거라 생각할 수 있다. 따라서 본 연구에서 사용하는 방법은 클러스터간의 조건부 확률적 의존성이 그러한 서로 다른 클러스터에 속한 외부연결 변수들 사이에서의 조건부 확률적 의존성에 의해 결정된다고 가정한다.

각 클러스터 C_i에 대해, 다른 클러스터 C_j $(i \neq j)$에 가장 높은 의존성을 보이는 확률변수 v_i가 C_j를 향한 대표 외부연결 변수로 결정되고 b_{ij}로 표기한다. 이를 위해 두 클러스터 C_i와 C_j에서 서로를 향해 가장 가까운 b_{ij}와 b_{ji}를 상호 정보 (mutual information) 척도를 이용하여 찾는다. 즉 C_i에 속한 모든 확률변수들과 C_j에 속한 모든 확률변수들 사이의 쌍의 조합 중 가장 높은 상호 정보 (mutual information) 값 $I(v_i;v_j)$을 갖는 v_i와 v_j를 찾은 뒤 각각을 b_{ij}와 b_{ji}로 결정한다(그림 2(a)). 본 연구에서 다루는 문제는 주어진 D에 대해 $P(G_C|D)$를 최적화하는 G_C를 찾는 것이며, 이를 위해 클러스터간의 베이지안 망 구조 학습을 수행한다. 단, 클러스터간의 베이지안 망 구조 학습 과정에서 클러스터 C_i와 C_j 사이의 방향성 있는 연결선(edge) e_{ij}를 고려할 때 두 클러스터에 속한 외부연결 변수들인 b_{ij}와 b_{ji} 사이의 연결선(edge)이 대신 고려된다. 이 때 한 클러스터에 속한 외부연결 변수들 b_{ij} $(\forall j)$ 사이에서의 연결선(edge)은 고려하지 않는다. 이와 같은 과정을 통해 외부연결 변수들의 집합 B에 대해 높은 $P(G_B|D)$를 보이는 방향성 비순환 그래프 (DAG) 구조 G_B를 찾은 뒤(그림 2(b)), G_B에 속한 한 연결선(edge) $b_{ij} \rightarrow b_{ji}$를 G_C에서의 한 연결선(edge) e_{ij}로 사상함으로서 클러스터 사이에 존재하는 조건부 확률적 의존의 방향을 결정하게 된다(그림 2(c)). 이러한 과정을 수행하는 알고리즘은 다음과 같다.

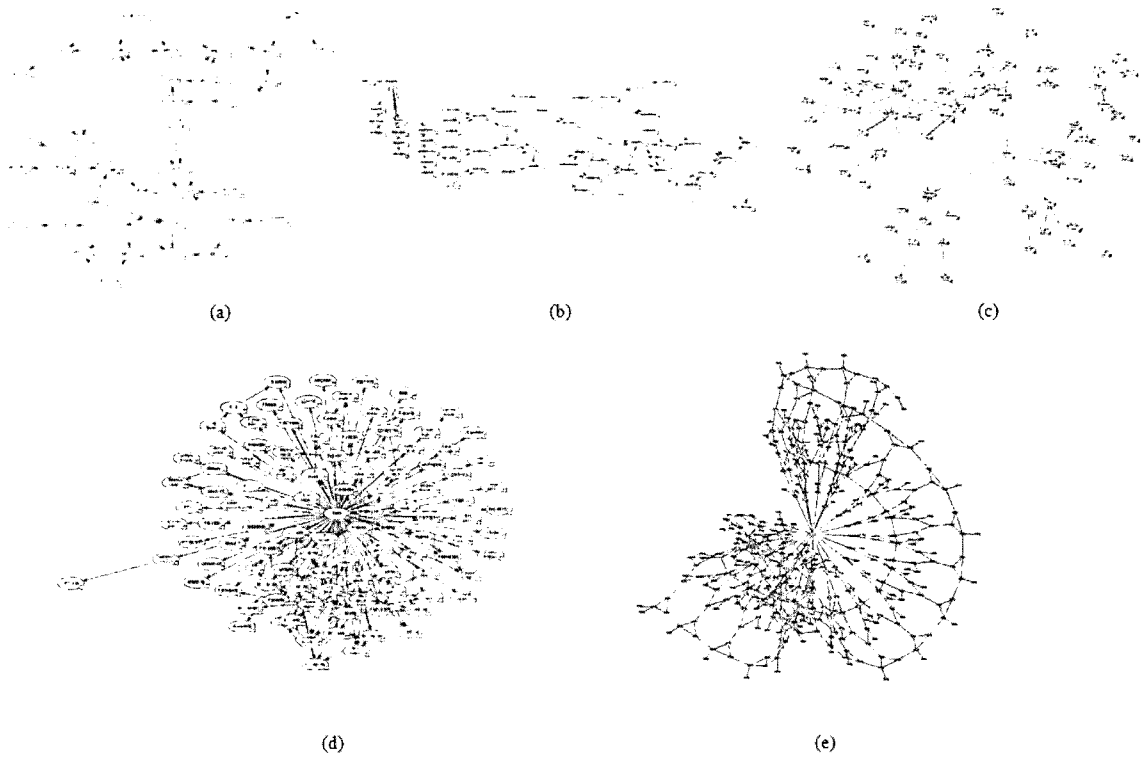


그림 3. LINK를 제외한 다섯 개의 벤치마크 베이지안 망. LINK는 망의 크기로 인해 표시하지 않음. (a) ALARM (b) HAILFINDER (c) WIN95PTS (d) PATHFINDER (e) DIABETES

Fig. 3. Five benchmark Bayesian networks except LINK. LINK is omitted because of its huge size. (a) ALARM (b) HAILFINDER (c) WIN95PTS (d) PATHFINDER (e) DIABETES

알고리즘 1.

(1) 외부연결 변수들의 집합 $\boldsymbol{B}$를 공집합으로 초기화한다.

(2) 집합 $\boldsymbol{C}$에 속한 클러스터들의 모든 조합 $(\boldsymbol{C}_i;\boldsymbol{C}_j)$에 대해,

(2-1) 가장 높은 $I(v_i;v_j)$를 보이는 $v_i(\in \boldsymbol{C}_i)$와 $v_j(\in \boldsymbol{C}_j)$를 각각 b_{ij}와 b_{ji}로 결정한다.

(2-2) $\boldsymbol{B}=\boldsymbol{B}\cup\{b_{ij},b_{ji}\}$

(3) $P(G_B|D)$를 최대화하는 방향성 비순환 그래프 (DAG) G_B를 다음과 같은 조건 하에서 탐색한다.

조건 1. 연결선(Edge)은 b_{ij}와 b_{ji} $(i\neq j)$사이에서만 존재할 수 있다.

조건 2. G_B를 G_C에 사상했을 때, 클러스터들 사이에 순환(cycle)이 존재하지 않는다.

(4) 찾은 G_B의 구조를 G_C에 사상한다.

알고리즘 1을 통해, 외부연결 변수들 사이의 방향성 비순환 그래프 (DAG) 구조를 클러스터간에 순환(cycle)이 허용되지 않는 조건하에 학습함으로서 클러스터간의 조건부 확률적 의존성을 반영하는 방향성 비순환 그래프 (DAG) 구조를 탐색하게 된다.

3. 실험 및 결과

3.1 실험 환경

표 1. 벤치마크 베이지안 망

Table 1. Benchmark Bayesian networks

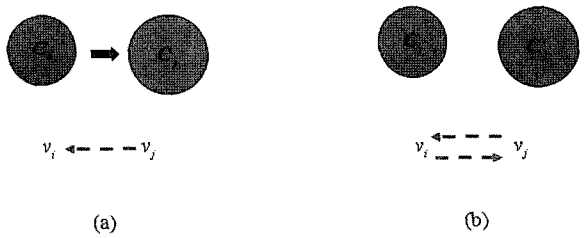
	확률변수 수	연결선 (Edge) 수	확률변수 부모 수의 평균	확률변수 부모 수의 최대값
ALARM	37	46	1.24	4
HAILFINDER	56	66	1.18	4
WIN95PTS	76	112	1.47	7
PATHFINDER	109	195	1.79	5
DIABETES	413	602	1.46	2
LINK	724	1125	1.55	3

앞에서 기술한 알고리즘 1의 성능을 평가하기 위해 기존의 알려진 베이지안 망을 벤치마크로서 이용하였다. 실험에서는 표 1에 기술된 여섯 개의 벤치마크 베이지안 망을 사용하였으며, 크기가 큰 LINK 베이지안 망을 제외한 나머지 다섯 개의 벤치마크 베이지안 망의 구조는 그림 3에 나타나 있다. 본 연구의 알고리즘은 주어진 클러스터들 사이에 존재하는 조건부 확률적 의존의 방향성을 추정하는 방법이다. 사용된 방법의 성능 평가를 위하여, 각 베이지안 망에 속한 확률변수들의 값을 5,000개씩 샘플링하여 관찰된 학습 데이터를 생성하였다. 그리고 각 베이지안 망 구조로부터 약한 확률적

의존성을 갖는 연결선(edge)을 하나씩 제거해 나가는 방법을 통해 특정 개수의 연결 요소(connected component)를 구성하였고, 각 연결 요소(connected component)에 속한 확률변수들의 집합이 하나의 클러스터를 이루도록 하였다. 이 때 각 연결선(edge)의 의존성 정도는 상호 정보 (mutual information) 척도를 이용하여 계산하였다. 이러한 방법으로 ALARM, HAILFINDER, WIN95PTS, PATHFINDER, DIABETES 그리고 LINK 베이지안 망에 속한 확률변수들을 각각 20, 35, 40, 40, 40 그리고 80개의 클러스터들로 사전에 클러스터링하였다. 본 실험에서는 각 벤치마크 베이지안 망으로부터 주어진 학습 데이터와 확률변수의 클러스터들에 대해 제안된 방법을 적용하여, 보다 높은 $P(G_B|D)$를 갖는 G_C를 탐색하는 과정에서 관찰되는 클러스터간 조건부 확률적 의존 방향성 추정의 오류 변화를 관찰하였다.

추정된 클러스터간 조건부 확률적 의존의 방향성을 기술하는 G_C가 있을 때, 추정 결과의 오류는 원래의 벤치마크 베이지안 망의 구조 G_V로부터 다음과 같이 계산하였다.

순서를 무시한 모든 클러스터의 쌍 C_i와 C_j에 대해,
모든 확률변수의 조합 $v_i(\in C_i)$와 $v_j(\in C_j)$에 대해,
  IF G_C가 e_{ij} ($C_i \rightarrow C_j$)를 포함하면,
    IF G_V가 $v_i \leftarrow v_j$를 포함하면 오류 1 증가
    ENDIF
  ELSE
    IF G_V가 $v_i \rightarrow v_j$ 혹은 $v_i \leftarrow v_j$를 포함하면 오류 1 증가



(a) (b)

그림 4. 클러스터 사이의 방향성 추정에 따른 오류의 경우. (a) 두 클러스터 사이에 의존 방향성이 있는 것으로 추정된 경우 (b) 두 클러스터 사이에 의존 방향성이 없는 것으로 추정된 경우

Fig. 4. Error cases in approximating the direction of dependencies between clusters. (a) When a direction is approximated. (b) When no directional dependency is approximated.

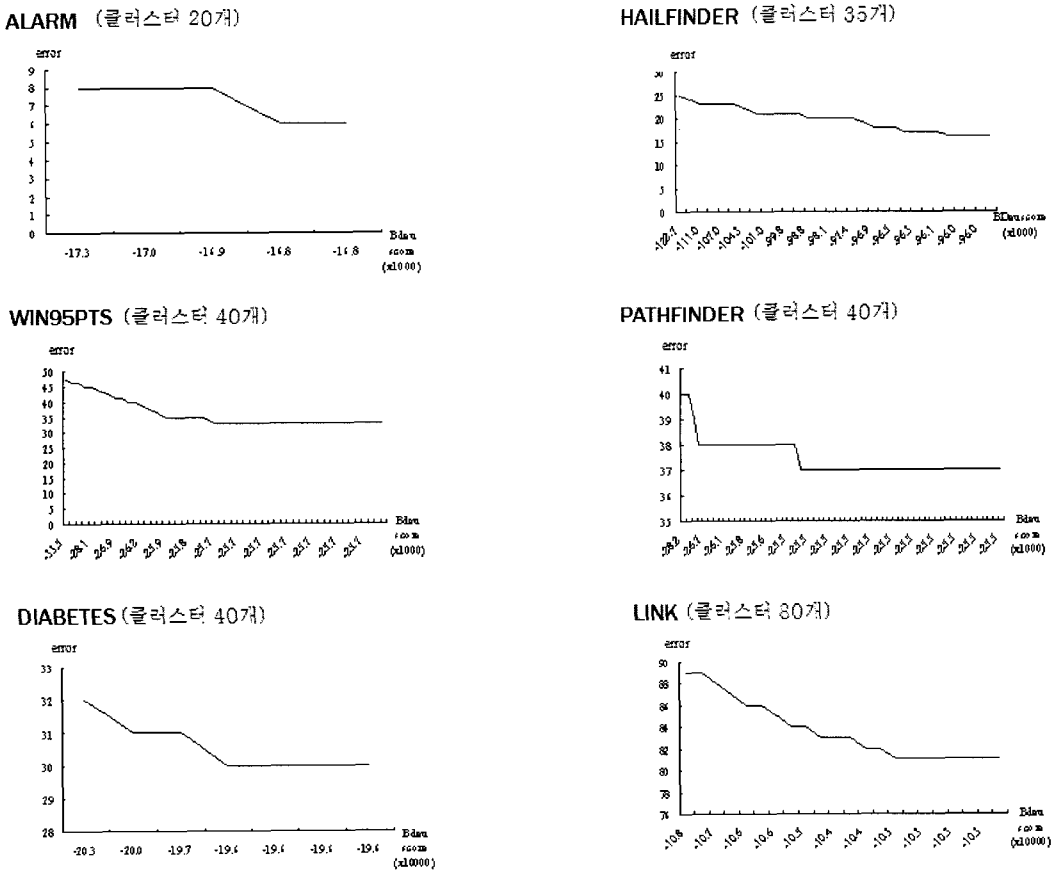


그림 5. 클러스터간 조건부 확률적 의존의 방향성 추정 과정에서 $P(G_B|D)$값이 높아짐에 따른 추정 결과의 오류 변화

Fig. 5. Graphs of error and $P(G_B|D)$ in the process of approximating the direction of conditional probabilistic dependencies between clusters

클러스터간 조건부 확률적 의존 방향성의 추정 결과에서 오류를 위와 같이 계산하는 것은 다음과 같은 의미를 지닌다. 두 클러스터간 방향성 추정의 역방향으로 실제 존재했던 확률변수 사이의 의존성이 오류로 간주되며, 또한 두 클러스터 사이에서의 의존성을 추정하지 못한 경우 실제 존재했던 확률변수 사이의 의존성들을 오류로서 계산하게 된다(그림 4).

알고리즘 1의 단계 (3)에 존재하는 방향성 비순환 그래프 (DAG) 탐색 과정에서는 알고리즘의 특성 파악을 보다 쉽게 하기 위해 간단한 그리디(greedy) 탐색 방법을 사용하여 $P(G_B|D)$ 탐색에 따른 G_C에서의 추정 오류 관찰을 수행하였다.

3.2 실험 결과

여섯 개의 벤치마크 베이지안 망 각각에 대해 클러스터간 조건부 확률적 의존성의 방향을 추정해 나가면서 발생하는 오류의 변화를 관찰한 그래프는 그림5와 같다. 각 그래프의 가로축은 $P(G_B|D)$를 의미하는 BDeu 점수의 로그 값을 나타내며 세로축은 해당 점수의 G_B가 반영하는 클러스터간 조건부 확률적 의존성의 방향 관계 G_C에서의 오류를 나타낸다. 그림에서 나타난 결과와 같이 모든 망 구조의 경우에서 보다 높은 $P(G_B|D)$ 값을 갖는 G_C가 항상 보다 적은 클러스터간 조건부 확률적 의존성의 방향 오류를 보임을 알 수 있다.

이 실험에서는 그리디(greedy) 탐색 알고리즘을 통해 방향성 비순환 그래프 (DAG) 구조를 탐색하였으므로 모든 그래프 구조의 분포에 대해 BDeu 점수와 오류를 확인한 것은 아니다. 그러나 비교적 간단한 탐색 알고리즘인 그리디(greedy) 탐색 알고리즘을 적용했을 때, BDeu 점수가 국지해(local optimum)로 수렴하는 과정에서 오류가 단조 감소하는 현상을 보인 것은 본 연구의 방법을 이용하여 클러스터간 조건부 확률적 의존성의 방향 관계를 유의미하게 추정할 수 있음을 의미한다. 본 실험에서 방향성 비순환 그래프 (DAG) 구조의 탐색에 사용된 그리디(greedy) 탐색 알고리즘 외에 보다 높은 성능의 탐색 알고리즘을 사용할 경우 본 실험에서의 결과보다 더 큰 폭으로 조건부 확률적 의존성 방향 추정의 오류를 줄이는 것이 가능하다.

4. 결론 및 향후 과제

본 연구에서는 확률변수 클러스터들간 조건부 확률적 의존성의 방향성을 추정하기 위해, 클러스터의 외부 연결 변수를 결정하여 조건부 결합 확률 분포를 나타내는 베이지안 망 구조를 학습하는 방법을 사용하였다. 벤치마크 베이지안 망 여러 개를 이용한 실험 결과, 본 연구에서 사용한 방법이 클러스터간 의존성의 방향을 추정하는 데에 유의미하게 사용될 수 있음을 알 수 있었다.

본 연구에서 사용된 클러스터간 조건부 확률적 의존성의 방향 추정 방법은 베이지안 망 학습을 위한 탐색 공간 축소, 확률변수간의 오더 결정 등의 문제에 다양하게 응용되어질 수 있다. 향후 과제로는 그러한 다양한 응용 문제에의 적용과 응용 대상에 적합한 최적화 등이 있을 수 있다. 또한 보다 좋은 외부 연결 변수를 결정하기 위해, 오더 추정 과정에서 외부 연결 변수를 동적으로 변화시키는 방법을 향후 과제로서 수행할 수 있다. 이러한 방법으로 클러스터간 확률적 의존성의 추정 결과에서의 오류를 보다 더 줄일 수 있을 것으로 생각된다.

참 고 문 헌

[1] L. Sun and P. P. Shenoy, "Using Bayesian Networks for Bankruptcy Prediction: Some Methodological Issues," *European Journal of Operational Research*, Vol. 180, pp. 738-753, 2007.

[2] L. Uusitalo, "Advantages and Challenges of Bayesian Networks in Environmental modelling," *Ecological Modelling*, Vol. 203, pp. 312-318, 2007.

[3] H. Zhou and S. Sakane, "Mobile Robot Localization Using Active Sensing Based on Bayesian Network Inference," *Robotics and Autonomous Systems*, Vol. 55, pp. 292-305, 2007.

[4] B. Abramson, J. Brown, W. Edwards, A. Murphy and R. L. Winkler, "Hailfinder: A Bayesian system for Forecasting Severe Weather," *International Journal of Forecasting*, Vol. 12, pp. 57-71, 1996.

[5] S. Andreassen, R. Hovorka, J. Benn, K. G. Olesen and E. R. Carson, "A Model-Based Approach to Insulin Adjustment," *Proceedings of the Third Conference on Artificial Intelligence in Medicine*, pp. 239-248, 1991.

[6] I. A. Beinlich, H. J. Suermondt, R. M. Chavez and G. F. Cooper, "The ALARM Monitoring System: A Case Study with Two Probabilistic Inference Techniques for Belief Networks," *Proceedings of the Second European Conference on Artificial Intelligence in Medicine*, 1989.

[7] D. E. Heckerman, E. J. Horvitz and B. N. Nathwani, "Toward Normative Expert Systems: Part I The Pathfinder Project," *Methods of Information in Medicine*, Vol. 31, pp. 90-105, 1992.

[8] H. J. Suermondt and G. F. Cooper, "A Combination of Exact Algorithms for Inference on Bayesian Belief Networks," *International Journal of Approximate Reasoning*, Vol. 5, pp. 521-542, 1991.

[9] B. R. Cobb and P. P. Shenoy, "Operations for Inference in Continuous Bayesian Networks with Linear Deterministic Variables," *International Journal of Approximate Reasoning*, Vol. 42, pp. 21-36, 2006.

[10] B. R. Cobb and P. P. Shenoy, "Inference in Hybrid Bayesian Networks with Mixtures of Truncated Exponentials," *International Journal ofApproximate Reasoning*, Vol. 41, pp. 257-286, 2006.

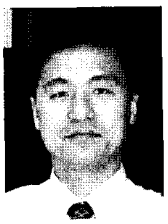
[11] S. Acid and L. M. Campos, "BENEDICT: An Algorithm for Learning Probabilistic Belief Networks," *In Proceedings of 6th International Conference IPMU'96*, Grenade, pp. 979-984, 1996.

[12] L. E. Brown, I. Tsamardinos and C. F. Aliferis, "A Novel Algorithm for Scalable and Accurate Bayesian Network Learning," *MEDINFO*, 2004.
[13] N. Friedman, I. Nachman and D. Pe'er, "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.
[14] D. Heckerman, D. Gerger and D. M. Chickering, "Learning Bayesian Networks: The Combination of Knowledge and Statistical Data," *Machine Learning*, Vol. 20, pp. 197-243, 1995.
[15] G. F. Cooper and E. Herskovitz, "A Bayesian Method for the Induction of Probabilistic Networks from Data," *Machine Learning*, Vol. 9, pp. 309-347, 1992.
[16] E. Herskovitz, G. Cooper, "Kutato: An Entropy-Driven System for Construction of Probabilistic Expert Systems from Databases," *In Proceedings of 6th International Conference on Uncertainty in Artificial Intelligence*, Cambridge, MA, pp. 54-62, 1990.
[17] J. Suzuki, "Learning Bayesian Belief Networks Based on the MDL Principle: An Efficient Algorithm using the Branch and Bound Technique," *IEICE Trans. Information and Systems*, Vol. E81-D, pp. 356-367, 1999.

저 자 소 개

정성원(Sungwon Jung)
1998년 : 한국과학기술원 전산학과 졸업
2000년 : 한국과학기술원 전산학과 졸업 (공학석사)
2007년 : 한국과학기술원 전산학과 졸업 (공학박사)
2007년 ~ 현재 : 한국과학기술원 IBM-KAIST 바이오컴퓨팅 연구센터 연구원

관심분야 : 기계학습, 바이오정보학, 인공지능, 의료정보학
Phone : 042-869-5356
Fax : 042-869-8680
E-mail : swjung@biosoft.kaist.ac.kr



이도헌(Doheon Lee)
1990년 : 한국과학기술원 전산학과 졸업
1992년 : 한국과학기술원 전산학과 졸업 (공학석사)
1995년 : 한국과학기술원 전산학과 졸업 (공학박사)
1994년~1995년 : 한국전자통신연구원 위촉 연구원
1996년 ~ 2002년 : 전남대학교 교수
2002년 ~ 현재 : 한국과학기술원 바이오 및 뇌 공학과 교수

관심분야 : 바이오 데이터 마이닝, 바이오 시스템 모델링, 바이오정보학
Phone : 042-869-4316
Fax : 042-869-8680
E-mail : dhlee@biosoft.kaist.ac.kr

이광형(Kwang H. Lee)
1978년 : 서울대학교 산업공학과 졸업
1980년 : 한국과학기술원 산업공학과 졸업 (공학석사)
1985년 : INSA de Lyon 전산학과 졸업 (공학박사)
1985년~현재 : 한국과학기술원 바이오 및 뇌 공학과 교수
2000년~현재 : 한국과학기술원 미래산업 석좌교수

관심분야 : 바이오정보학, 인공지능, 퍼지 시스템
Phone : 042-869-4313
Fax : 042-869-8680
E-mail : khlee@biosoft.kaist.ac.kr