한국지능시스템학회 논문지 2008, Vol. 18, No. 4, pp. 572-578

R-CORE를 통한 베이지안 망 구조 학습의 탐색 공간 분석

Search Space Analysis of R-CORE Method for Bayesian Network Structure Learning and Its Effectiveness on Structural Quality

정성원* · 이도헌* · 이광형**

Sungwon Jung, Doheon Lee, Kwang H. Lee

* 한국과학기술원 바이오및뇌공학과

** 한국과학기술원 바이오및뇌공학과, AITrc

요 약

본 논문에서는 대규모 베이지안 망 구조 학습을 위해 제안되었던 R-CORE 방법의 탐색 공간의 크기에 대한 개략적인 분석과 실제 문제에 적용하였을 경우의 효과에 대한 실험적 결과를 제시한다. R-CORE 방법은 베이지안 망 구조 학습의 탐색 공간을 축소하기 위해 제안된 확률변수들의 재귀적 군집화와 오더 제한 방법이다. 알려진 벤치마크 베이지안 망을 이용한 분석을 통해, 제안되었던 R-CORE 방법이 worst case에는 기존의 방법과 유사한 탐색 공간을 가지나 평균적으로 기존 방법보다 훨씬 적은 탐색 공간만을 고려한다는 것을 보인다. 또한 평균적으로 훨씬 적은 탐색 공간만을 고려하는 결과, 구조 탐색에서 기존 방법에 비해 상대적으로 적은 overfitting이 일어남을 실험적으로 보인다.

키워드 : 베이지안 망 구조 학습, R-CORE, 탐색 공간 분석

Abstract

We analyze the search space considered by the previously proposed R-CORE method for learning Bayesian network structures of large scale. Experimental analysis on the search space of the method is also shown. The R-CORE method reduces the search space considered for Bayesian network structures by recursively clustering the random variables and restricting the orders between clusters. We show the R-CORE method has a similar search space with the previous method in the worst case but has a much less search space in the average case. By considering much less search space in the average case, the R-CORE method shows less tendency of overfitting in learning Bayesian network structures compared to the previous method.

Key Words : Bayesian network structure learning, R-CORE, search space analysis

1. 서 론

베이지안 망은 확률 변수들 사이의 확률적 의존성을 기술하기 위한 모델로 널리 사용되어져 왔다. 이러한 베이지안 망은 확률 변수들의 조건부 확률 분포를 기술하기 위한 매개변수를 포함하고 있는 directed acyclic graph (DAG) 형태의 모델이다. 그러한 조건부 의존성은 각 확률 변수에 연결되는 edge들에 의해 기술되게 된다. 베이지안 망 B는 $B=(G,\Theta)$로 표현되며, G는 $G=(V,E)$로 표현되는 DAG 이고 V는 G에서의 node에 해당하는 확률 변수들의 집합, E는 그 node들 사이의 방향성 있는 edge들의 집합이다. Θ는 V에 포함된 확률 변수들의 결합 확률 분포를 기술하는 매개변수들의 집합이다.

주어진 확률 변수들 사이의 확률적 의존성을 기술하는 모델을 구축하기 위한 방법으로, 확률 변수들에 대해 관찰된 값들을 이용한 베이지안 망 학습 방법이 있다. 베이지안 망 학습은 DAG 형태의 망 구조 G의 학습과 확률 매개변수 값의 집합 Θ의 학습 두 가지를 포함한다. 본 연구에서는 그 중 베이지안 망 구조에 대한 학습을 다루고자 한다. 일반적인 베이지안 망 구조 학습은 주어진 관찰 데이터에 가장 잘 들어맞는 최적의 DAG을 찾는 방법이다[1]. 최적의 DAG 구조를 찾는 일반적인 방법은 BDeu 점수[2]나 MDL 점수[3,4]와 같은 주어진 점수 척도 *Score*를 이용하여, 주어진 관찰 데이터의 집합 D에 대해 최대의 *Score*$(G|D)$를 갖는 DAG 구조 G를 찾는 방법이다. 주어진 확률 변수들의 수에 대해 가능한 DAG 구조의 수가 굉장히 많기 때문에 [5], 그러한 DAG을 찾는 데에는 greedy 탐색이나 유전자 알고리즘[6]과 같은 approximate 탐색 방법들이 일반적으로 사용되어지고 있다.

하지만 수백에서 수천 개의 많은 확률 변수를 포함한 대규모의 문제에 베이지안 망 학습을 적용하기 위해서는 보다 효과적인 학습 방법들이 요구된다. 기존 베이지안 망 학습에서 주로 다루었던 수십 개 정도의 확률 변수를 지닌 문제에 비하면 이러한 대규모 베이지안 망 학습은 훨씬 더 어려운 문제가 되기 때문이다. 예를 들어, 기존의 approximate 구조 학습에서 벤치마크로 많이 사용된 ALARM 베이지안

접수일자 : 2007년 5월 7일
완료일자 : 2008년 2월 10일
본 논문은 과학기술부 시스템생물학 연구사업(2005-00343)의 지원과 정보통신부 연구비 C1090060200001의 지원으로 수행되었음. 연구시설은 정문술 바이오정보전자센터의 도움을 받았음.

망은 37개의 node를 지닌다. 그러한 상황에서 사용되었던 기존의 approximate 탐색 방법들을 수백 개 정도의 node를 지닌 베이지안 망 구조 학습에 적용하기 위하여 여러 가지 탐색 공간 제한 방법들이 제안되어져 왔다[7,8,9]. 그러나 그러한 방법들 또한 수백 개 정도의 node를 지닌 베이지안 망 구조 학습이 한계였기 때문에, 수천 개 이상의 node를 지닌 보다 대규모의 베이지안 망 구조 학습을 위하여 R-CORE 방법[10]이 제안되었다. R-CORE 방법은 확률 변수들을 재귀적으로 군집화하고 군집 간의 오더를 제한하는 방법을 통해 훨씬 짧은 시간 내에 기존의 방법과 비슷한 수준의 베이지안 망 학습을 가능하게 한다는 것이 실험적으로 밝혀져 있다. 본 연구에서는 기 제안된 R-CORE 방법의 탐색 공간을 기존의 방법과 비교하여 분석하고, R-CORE의 탐색 공간 제한 효과가 베이지안 망 구조의 학습에 미치는 영향을 알려진 벤치마크 베이지안 망을 이용하여 실험적으로 분석하였다.

본 논문은 다음과 같이 구성되었다. 2장에서는 기 제안된 R-CORE 방법을 개략적으로 설명하고 그 효과를 기술한다. 3장에서는 R-CORE 방법과 기존 탐색 공간 제한 기법의 대표 방법으로서의 sparse candidate (SC) [7] 방법이 갖는 탐색 공간에 대한 비교 분석을 기술한다. 4장에서는 R-CORE의 탐색 공간 제한이 베이지안 망 구조 학습의 결과에 미치는 영향을 SC의 경우에 대비해 벤치마크 베이지안 망을 이용하여 실험적으로 비교하였다. 마지막으로 5장에서는 결론과 향후 과제를 기술하였다.

2. R-CORE 베이지안 망 구조 학습 방법

R-CORE 방법[10]은 베이지안 망 구조 학습에서 탐색 대상이 되는 후보 DAG을 제한하기 위해 모듈화 접근 방법을 사용한다. 모듈화 접근 방법의 기본적인 개념은 전체 망 구조가 '하부 망 구조의 망'으로 이루어져 있다고 가정하는 것이다. R-CORE 방법에서는 확률 변수들의 군집을 하부 망 구조 모듈로 간주하고, 군집간의 망 구조를 결정하여 그에 해당하는 DAG 구조만을 베이지안 망 구조 학습의 대상으로 삼는다. R-CORE 방법은 다음의 두 단계로 이루어져 있다.

- 구조 제한 단계
  1. 확률변수를 $c_{\max}(\ll n)$개의 군집으로 군집화한다. (n은 전체 확률 변수의 수)
  2. 군집간의 방향성을 cycle이 발생하지 않도록 하며 결정한다.
  3. 크기가 $c_{\max}$보다 큰 군집에 대해서는 이 과정을 반복적으로 수행한다.

- 구조 탐색 단계
  1. 구조 제한 단계에서 결정된 군집간 방향성을 준수하는 DAG들 중 주어진 관찰 데이터의 결합 조건부 확률 분포를 가장 잘 기술하는 DAG 구조를 탐색한다.

구조 제한 단계에서는, 확률변수들에 대한 계층적 군집화를 상위 계층으로부터 분할하는(divisive) 방식으로 수행하되, 이 때 군집 계층의 각 층에 존재하는 군집의 수가 $c_{\max}$보다 작고, 각 군집의 크기 또한 $c_{\max}$보다 작게 만들게 된다. 이 방법의 목적은 DAG 탐색 공간에서 cycle의 발생을 고려해야 하는 공간의 크기를 줄이는 것이다. 구조 제한 단계에서 군집간의 방향성을 acyclic하게 결정함에 따라 구조 탐색 단계에서 cycle을 고려하는 범위를 각 군집 내에서만으로 제한하는 것이 가능하다. 그러나 만약 어떤 군집의 크기가 특정 크기($c_{\max}$)보다 크게 된다면 우리가 원하는 수준으로 탐색 공간을 제한할 수 없게 된다. 이 문제의 해결을 위해 구조 제한 단계에서는 크기가 큰 군집에 대해서 군집화와 군집간 방향성 결정을 재귀적으로 수행하게 된다. 이러한 개념은 그림 1에 나타나 있다.

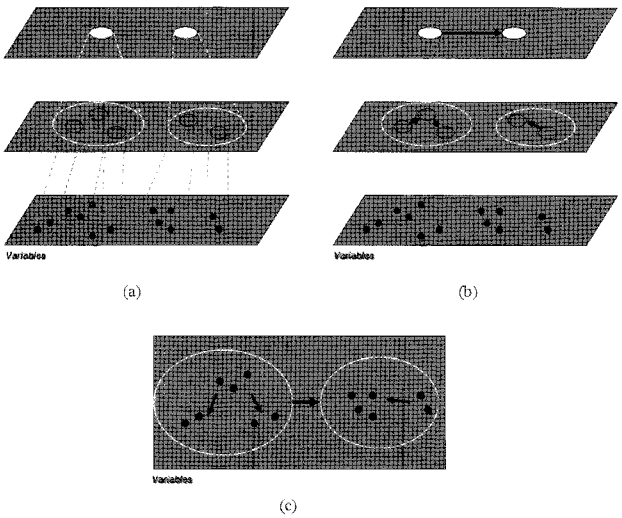


그림 1. R-CORE의 탐색 대상 DAG 구조 제한 방법의 개요. (a) 분할식 재귀적 군집화 (b) 군집간 방향성 결정 (c) 재귀적 군집화와 군집간 방향성 결정에 따라 제한된 DAG 구조.

Fig. 1. Conceptual diagram for DAG search space restriction approach of R-CORE method. (a) Recursive clustering (b) Determining intercluster directionality (c) The restricted candidate DAGs. Candidate DAGs should follow this intercluster directionality.

이와 같은 방법을 통해 구조 탐색 단계에서는 cycle을 고려하는 범위가 최대 $c_{\max}$개의 확률 변수들로 제한되게 된다. 그 결과 R-CORE방법은 기존의 방법이 탐색할 수 없는 대규모의 베이지안 망 구조를 빠른 시간 내에 학습할 수 있으며, 특히 망 구조의 규모에 비해 관찰된 학습 데이터의 수가 적은 경우 기존의 방법과 비슷한 성능의 베이지안 망 구조를 훨씬 짧은 시간 내에 학습할 수 있다는 점이 보고되어 있다[10].

3. R-CORE 방법과 기존 방법의 탐색 공간 분석

3.1 탐색 공간 분석의 조건

본 장에서는 기 제안된 R-CORE 방법이 고려하는 DAG 탐색 공간과 기존의 SC 방법이 고려하는 탐색 공간의 크기를 worst case와 평균적인 경우에 대해 대략적으로 비교한다. DAG 탐색을 위해서 사용하는 점수 척도는 $BDeu$ 점수

나 *MDL* 점수와 같은 분해 가능한 (decomposable) 점수를 사용하는 것으로 한다. 어떤 점수 척도 *Score*가 분해 가능함은 다음과 같은 성질을 지니고 있다는 것을 의미한다.

$$Score(G;D) = \sum_{i=1}^{n} Score(V_i|Pa_G(V_i);D)$$

이와 같이 어떤 점수 척도가 분해 가능하다는 것은 전체 망 구조에 대한 점수가 망에 포함된 각 노드와 그에 대한 부모 노드들의 집합으로부터 계산되는 점수의 합으로 계산될 수 있음을 의미한다.

3.2 모든 가능한 DAG 구조를 고려하는 경우의 탐색 공간

n개의 노드가 주어진 경우, 가능한 DAG의 수는 Robinson의 연구에 의해 알려져 있다[5]. 다양한 경우에 있어서 가능한 DAG의 수를 정확히 구하는 것은 본 연구의 범위에서 벗어난 일이므로, 본 연구에서는 개략적인 비교를 위해 Robinson의 연구와 같이 정확한 수의 비교가 아닌 가능한 DAG 수의 대략적인 비교로서 upper bound만을 비교하도록 한다.

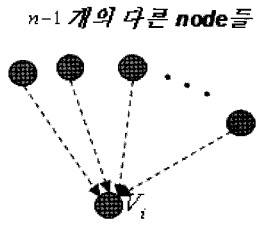


그림 2. 가능한 모든 DAG을 고려하는 경우, 한 node가 가질 수 있는 부모 node의 집합의 수와 DAG 수의 대략적인 upper bound.

Fig. 2. The case of considering all possible DAGs for n nodes. Implicit upper bound of possible DAGs without considering cycles.

n개의 node에 대해, 존재 가능한 모든 DAG의 수를 대략적으로 구하기 위해 각 node에 대해 가능한 부모 node 집합의 경우의 수를 구한 후 n개의 node 수의 조합으로서 전체의 upper bound를 추정하는 방법을 사용한다. 전체 n개의 node에 속한 한 node V_i가 있다고 하자. 가능한 모든 DAG의 구조를 고려하는 경우, cycle의 발생을 고려하지 않는 상황에서 V_i가 가질 수 있는 부모 node 집합의 경우의 수는 2^{n-1}이 된다. 그리고 이 조합이 모든 n개의 node에 대해 가능하기 때문에 cycle을 고려하지 않는 경우 가능한 모든 조합의 수는 $2^{n(n-1)}$이 된다. 여기에서 n개의 node에 대해 가능한 cycle을 포함하는 경우의 수 cycles[n]을 제외하면 실제 가능한 모든 DAG의 수가 되나, 대략적인 비교를 위해 upper bound인 $O(2^{n(n-1)})$로 표현하도록 한다(그림 2).

3.3 SC 방법의 탐색 공간

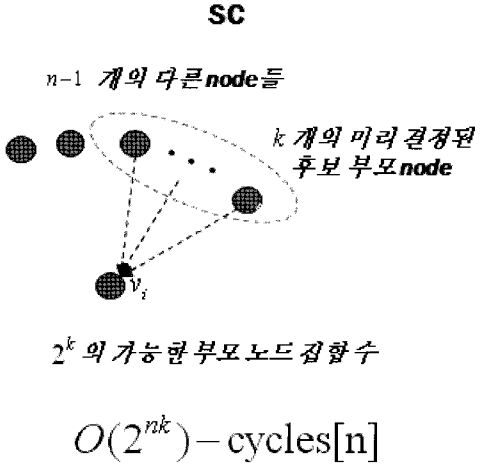


그림 3. SC 방법의 경우 한 node가 가질 수 있는 부모 node 집합의 경우의 수 및 전체 DAG 수의 대략적인 upper bound.

Fig. 3. The case of k predetermined candidate parents for each node in SC method. Implicit upper bound of possible DAGs without considering cycles.

SC방법은 수백 개 정도의 node를 지닌 베이지안 망 구조를 학습하기 위해 DAG 탐색 공간을 제한하는 방법 중 하나이다. 한 node V_i에 대해, 가능한 모든 DAG을 고려하는 경우에는 다른 $n-1$개의 모든 node가 V_i의 부모 node가 될 수 있다. SC 방법에서는 탐색 공간을 제약하기 위해 각 V_i에 대해 사전에 결정된 $k(<n)$개의 node만이 부모 node로 고려될 수 있도록 제한한다. 그 결과 한 node가 가질 수 있는 부모 node들의 집합의 경우의 수는 2^k가 된다. Cycle을 고려하지 않는 경우 모두 n개의 node에 대해 2^k가지의 부모 node 집합의 경우의 수가 있으므로 가능한 모든 조합의 수는 2^{nk}가 된다. 그 중 cycle이 발생하는 경우의 수 cycles[n]을 제외한 부분만을 고려해야 정확한 수가 되지만 대략적인 비교를 위해 upper bound인 $O(2^{nk})$로 표현하도록 한다(그림 3).

3.4 R-CORE 방법의 탐색 공간

R-CORE 방법에서는 그림 1(c)와 같이 결정된 군집간의 방향성에 따르는 DAG만을 탐색한다. 어떤 하나의 군집 C_j에 속한 한 node V_i가 있다고 하자. V_i의 부모가 될 수 있는 node는 C_j에 속한 V_i 이외의 다른 모든 node들과, 군집간의 방향성에서 보다 상위의 군집에 속한 다른 모든 node들이 된다. R-CORE 방법에서는 전체 node를 $c_{\max}(\ll n)$ 크기의 군집으로 재귀적으로 군집화하므로, 그러한 군집 계층 깊이의 worst case는 $O(\frac{n}{c_{\max}})$가 된다. 따라서 C_j보다 상위에 속하는 군집의 수는 worst case에 $O(\frac{n}{c_{\max}})$가 되므로, 각 군집의 최대 크기 $c_{\max}$를 감안하면 V_i가 가질 수 있

는 부모 node 집합의 경우의 수는 $O(2^{c_{\max}\frac{n}{c_{\max}}-1})$이 된다. 한 군집 내에 최대 $c_{\max}$개의 node가 존재할 수 있으므로, C_j에 속한 모든 node들의 경우의 수의 조합을 생각하면 하나의 군집에 속한 모든 node들의 부모 node 집합의 경우의 수에 대한 조합은 $O(2^{c_{\max}(c_{\max}\frac{n}{c_{\max}}-1)})$가 된다. n개의 node를 최대 크기 $c_{\max}$의 군집으로 계층적 군집화하였을 경우 군집의 최대 수는 $O(\frac{n}{c_{\max}})$이다. 일반적인 경우, 군집의 수에 대한 경우의 수의 조합을 고려하여 가능한 DAG 수의 대략적인 upper bound는 $O(2^{\frac{n}{c_{\max}}c_{\max}(c_{\max}\frac{n}{c_{\max}}-1)})$가 된다. 그러나 3.1절에서 언급한 분해 가능한 점수 척도를 사용하는 경우, 군집간의 방향성이 cycle이 존재하지 않도록 사전에 결정되어 있으므로 각 군집 내의 node들의 부모 node 결정을 군집 별로 서로 독립적으로 수행하는 것이 가능하다. 즉 $\frac{n}{c_{\max}}$개의 군집에 대한 조합을 고려하지 않고 각 군집에 대한 DAG 탐색을 독립적으로 $\frac{n}{c_{\max}}$회 수행하여 문제를 해결할 수 있음을 의미한다. 이 경우 DAG 탐색 공간의 대략적인 upper bound는 $O(\frac{n}{c_{\max}}2^{c_{\max}(c_{\max}\frac{n}{c_{\max}}-1)})$가 된다.

R-CORE의 upper bound는 기존의 SC와 비교하였을 때 worst case의 경우 $O(\frac{n}{c_{\max}}2^{c_{\max}(c_{\max}\frac{n}{c_{\max}}-1)})=O(\frac{n}{c_{\max}}2^{c_{\max}n})$가 되며, SC의 $O(2^{nk})$와 비교하였을 경우 $c_{\max}$의 값에 따라 SC 방법과 비슷하거나 높은 탐색 공간을 갖게 된다. 그러나 R-CORE 방법에서 최대 크기 $c_{\max}$인 군집들로 재귀적 군집화를 수행하는 과정에서 군집의 평균 크기 *a.c.*와 군집 계층의 평균 깊이 *a.d.*를 고려하는 평균적인 경우, R-CORE의 탐색 공간은 $O(\frac{n}{c_{\max}}2^{a.c.(a.c.\times a.d.-1)})$라 할 수 있으며 이것은 SC 방법에 비해 훨씩 작은 탐색 공간이다. 이후 4장에서의 실험을 통해, 실제의 경우 평균 군집 크기와 군집 계층의 평균 깊이가 worst case에 비해 훨씬 작고 그 결과 기존의 SC 방법에 비해 실제적으로 훨씬 작은 탐색 공간만을 고려하는 방법이라는 점을 보이도록 한다.

4. 실험 및 결과

4.1 실험 환경

본 실험에서는 기 제안되었던 R-CORE 방법이 평균적으로 고려하는 DAG 탐색 공간의 크기와 기존의 SC 방법이 고려하는 탐색 공간의 크기를 실험적으로 비교하였다. 실험을 위하여 ALARM, HAILFINDER, WIN95PTS, PATHFINDER, DIABETES 다섯 개의 알려진 베이지안 망을 벤치마크로 이용하여 각 5,000개씩의 학습 데이터를 샘플링하여 생성하였다. 다섯 개의 베이지안 망은 각기 37, 56, 76, 109, 413개의 node를 갖는다. 기 제안되었던 R-CORE 방법이 수천 개 이상의 node를 갖는 베이지안 망의 학습을 목표로 하고 있으므로 실제 수천 개 이상의 node를 갖는 베이지안 망을 벤치마크로서 활용하는 것이 바람직하다. 그러나 활용 가능한 기존의 베이지안 망 중에서는 그러한 대규모의 베이지안 망이 존재하지 않아 벤치마크로 사용하는 데 어려움이 있다. 또한 비교 대상인 SC 방법은 그러한 대규모의 망에 적용이 불가능하여 R-CORE와의 비교 또한 가능하지 않으므로 본 실험에서는 앞서 언급한 다섯 개의 베이지안 망 만을 벤치마크로서 활용하였다. 본 실험에서는 각 학습 데이터에 대해 R-CORE 방법의 최대 군집 크기 매개변수 $c_{\max}$를 5에서 40까지 변화시켜 가며 적용하였으며 SC 방법 또한 각 node의 부모 node의 수 매개변수 k를 5에서 15까지 변화시켜 가며 적용하였다. R-CORE의 적용 과정에서 평균 군집 크기와 평균 군집 계층 깊이를 조사하여 평균적인 경우에 R-CORE가 고려하는 탐색 공간의 크기를 기존의 SC 방법이 고려하는 탐색 공간과 비교하였다. 또한 R-CORE가 기존의 SC에 비해 훨씬 적은 탐색 공간을 실제적으로 탐색함에 따라 DAG 구조 탐색에서 얻는 효과를 알기 위해 각 실험에서 학습된 베이지안 망 구조의 원래의 베이지안 망 구조에 대한 오류 값을 보인다. 망 구조의 오류는 학습된 망 구조에 있는 두 node 사이의 edge 연결 상태가 원래의 망 구조에서와 다르면 오류가 1 증가하며, 모든 node 사이의 경우에 대한 오류의 합계를 구한 것이다. R-CORE 방법과 SC 방법의 매개변수 값의 변화에 따른 오류 변화를 보임으로서 R-CORE 방법이 SC 방법에 비해 갖는 효과를 보이고자 한다.

4.2 실험 결과

R-CORE 방법과 기존의 SC 방법이 각각의 매개변수 값 $c_{\max}$와 k에 따라 갖게 되는 탐색 공간의 크기는 표 1과 같다. R-CORE 방법이 고려하는 탐색 공간의 크기는 SC 방법이 고려하는 탐색 공간과 비교할 때, worst case에는 비슷하거나 더 큰 탐색 공간을 고려한다. 그러나 평균적인 경우의 대부분에 있어 R-CORE 방법이 기존의 SC 방법에 비해 훨씬 작은 DAG 탐색 공간만을 고려하는 것을 알 수 있다. 이것은 평균적 군집 크기 *a.c.*와 평균적 군집 계층의 깊이 *a.d.*가 각각 worst case의 군집 크기 $c_{\max}$, worst case의 군집 계층 깊이 $\frac{n}{c_{\max}}$에 비해 크게 작기 때문이다. 그 결과 실제 베이지안 망 구조를 학습하는 시간 또한 R-CORE 방법이 기존의 방법에 비해 일반적으로 훨씬 빠르다는 것이 밝혀져 있다[10].

탐색 공간의 대략적인 비교 결과에 더해, 매개변수 값을 바꾸어 가며 두 방법을 적용한 베이지안 망 구조 학습에서 얻은 결과들의 구조 에러는 그림 4와 같다. DIABETES 망을 학습함에 있어, SC 방법의 경우는 고려하는 탐색 공간이 지나치게 큼으로 인해 주어진 시간 내에 유의미한 학습 결과를 내는 데 실패하여 해당 결과를 보이지 않았다. 표 1에서 알 수 있듯이, R-CORE의 매개변수 $c_{\max}$와 SC의 매개변수 k가 보다 큰 값을 가질수록 해당 방법이 고려하는 DAG 탐색 공간은 더 커지게 된다. 그림 4의 결과에서 SC 방법은 가장 작은 베이지안 망인 ALARM을 학습할 때를 제외하면, 고려하는 탐색 공간이 커짐에 따라 구조 오류가 증가하는 현상을 보인다. 반면 기 제안된 R-CORE 방법은 고려하는 탐색 공간을 증가시킬 때 일반적으로 구조 오류가 증가하는 경우는 없으며, HAILFINDER와 DIABETES 망의 경우는 구조 오류가 감소하는 경향을 보이고 있다. 이

표 1. 각 매개변수 값에 따른 R-CORE와 SC의 탐색 공간의 대략적인 크기 비교 (a) ALARM (b) HAILFINDER (c) WIN95PTS (d) PATHFINDER (e) DIABETES

Table 1. Implicit comparison for search spaces of R-CORE and SC. (a) ALARM (b) HAILFINDER (c) WIN95PTS (d) PATHFINDER (e) DIABETES

(a) ALARM의 경우
(a) The case of ALARM

R-CORE

c_{max}	10	20	30
a.c	6.7	19.6	28.2
a.d	3.0	1.5	1.1
평균탐색공간	$O(3.7\times 2^{128})$	$O(1.9\times 2^{557})$	$O(1.2\times 2^{847})$
Worst case 탐색공간	$O(3.7\times 2^{370})$	$O(1.9\times 2^{740})$	$O(1.2\times 2^{1110})$

SC

k	5	10	15
탐색공간	$O(2^{185})$	$O(2^{370})$	$O(2^{555})$

(b) HAILFINDER의 경우
(b) The case of HAILFINDER

R-CORE

c_{max}	10	20	30	40
a.c	5.5	18.0	29.4	32.7
a.d	1.7	0.4	0.3	0.2
평균탐색공간	$O(5.6\times 2^{46})$	$O(2.8\times 2^{108})$	$O(1.9\times 2^{187})$	$O(1.4\times 2^{213})$
Worst case 탐색공간	$O(5.6\times 2^{560})$	$O(2.8\times 2^{1120})$	$O(1.9\times 2^{1680})$	$O(1.4\times 2^{2240})$

SC

k	5	10	15
탐색공간	$O(2^{280})$	$O(2^{560})$	$O(2^{840})$

(c) WIN95PTS의 경우
(c) The case of WIN95PTS

R-CORE

c_{max}	10	20	30	40
a.c	9.8	14.5	23.7	27
a.d	1.7	0.9	0.5	0.4
평균탐색공간	$O(7.6\times 2^{153})$	$O(3.8\times 2^{175})$	$O(2.5\times 2^{238})$	$O(1.9\times 2^{294})$
Worst case 탐색공간	$O(7.6\times 2^{760})$	$O(3.8\times 2^{1520})$	$O(2.5\times 2^{2280})$	$O(1.9\times 2^{3040})$

SC

k	5	10	15
탐색공간	$O(2^{380})$	$O(2^{760})$	$O(2^{1140})$

(d) PATHFINDER의 경우
(d) The case of PATHFINDER

R-CORE

c_{max}	10	20	30	40
a.c	7.5	19.2	23.9	34.8
a.d	5.7	2.7	1.9	1.2
평균탐색공간	$O(10.9\times 2^{313})$	$O(5.5\times 2^{976})$	$O(3.6\times 2^{1042})$	$O(2.7\times 2^{1449})$
Worst case 탐색공간	$O(10.9\times 2^{1090})$	$O(5.5\times 2^{2180})$	$O(3.6\times 2^{3270})$	$O(2.7\times 2^{4360})$

SC

k	5	10	15
탐색공간	$O(2^{545})$	$O(2^{1090})$	$O(2^{1635})$

(e) DIABETES의 경우
(e) The case of DIABETES

R-CORE

c_{max}	10	20	30	40
a.c	3.5	19.1	28.8	37.6
a.d	8.4	2.0	1.3	1.0
평균탐색공간	$O(41.3\times 2^{99})$	$O(20.7\times 2^{692})$	$O(13.8\times 2^{1049})$	$O(10.3\times 2^{1334})$
Worst case 탐색공간	$O(41.3\times 2^{4130})$	$O(20.7\times 2^{8260})$	$O(13.8\times 2^{12390})$	$O(10.3\times 2^{16520})$

SC

k	5	10	15
탐색공간	$O(2^{2065})$	$O(2^{4130})$	$O(2^{6195})$

결과는 기존의 SC 방법이 주어진 학습 데이터가 한정된 상황에서 지나치게 큰 탐색 공간을 탐색하는 경우 overfitting에 빠진다는 것을 의미한다. 반면 기 제안된 R-CORE 방법은 효과적으로 탐색 공간을 축소시킨 후 DAG 탐색을 수행한 결과 그러한 overfitting의 발생을 줄이며 보다 효과적인 구조 탐색이 가능하게 함을 알 수 있다.

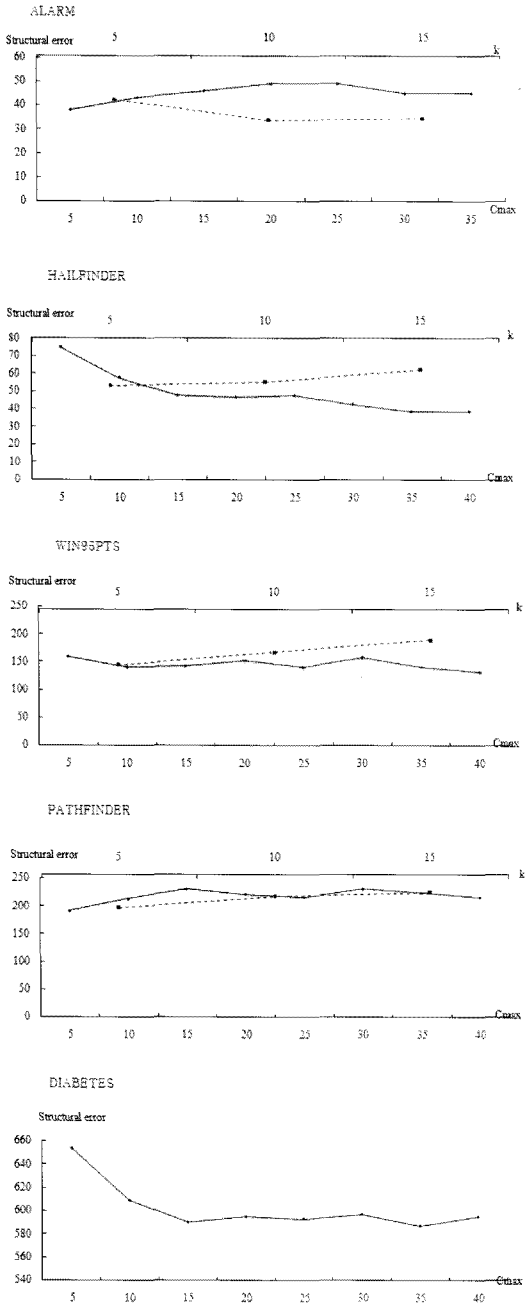


그림 4. R-CORE 방법과 SC 방법의 사용된 매개변수별 구조 에러의 변화. 실선은 R-CORE 방법의 구조 오류 변화. 점선은 SC 방법의 구조 오류 변화.

Fig. 4. Structural errors of R-CORE and SC with corresponding parameter values ($c_{\max}$ and k for each). The solid line is of R-CORE and the dotted line is of SC.

5. 결론 및 향후 과제

본 논문에서는 대규모 베이지안 망 구조의 학습을 위해 기 제안되었던 R-CORE 방법이 고려하는 탐색 공간의 대략적인 크기를 분석하여 기존의 방법과 비교하였다. 기존 연구[10]를 통해 R-CORE 방법이 실험적으로 이전의 방법들에 비해 훨씬 빠른 속도로 비슷한 베이지안 망 구조 학습 성능을 보임을 알 수 있었으나, 본 연구를 통해 R-CORE 방법이 고려하는 탐색 공간의 크기가 기존의 SC 방법이 고려하는 탐색 공간의 크기에 비해 평균적으로 훨씬 더 작음을 알 수 있었다. 또한 망 구조 학습을 위해 고려하는 탐색 공간의 크기를 효과적으로 크게 축소시킴으로서 overfitting의 발생을 억제하는 효과가 있음을 보였다.

실험 결과에서 R-CORE 방법의 매개변수 $c_{\max}$의 값을 증가시켜 고려하는 탐색 공간의 크기를 증가시키는 경우에, 구조 오류가 꾸준히 감소하는 경우와 그렇지 않은 경우가 존재하였다. 향후 과제로서 R-CORE 방법의 매개변수를 증가시키는 접근 방법을 통해 효과적으로 구조 오류를 감소시킬 수 있는 학습 대상 망 구조를 판단하는 validity 척도에 대한 연구가 필요할 수 있다.

참 고 문 헌

[1] R. E. Neapolitan, *Learning Bayesian Networks*, Pearson Prentice Hall, New Jersey, 2004.

[2] 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.

[3] P. Grunwald, "A Tutorial Introduction to the Minimum Description Length Principle," *Advances in Minimum Description Length: Theory and Applications*, MIT Press, 2004.

[4] J. Suzuki, "Learning Bayesian Belief Networks Based on the Minimum Description Length Principle: Basic Properties," *IEICE Transactions on Fundamentals*, Vol. E82-A, No. 9, pp. 1-9, 1999.

[5] R. W. Robinson, "Counting Labeled Acyclic Digraphs," *New Directions in the Theory of Graphs*, pp. 239-273, Academic Press, New York, 1973.

[6] R. Etxeberria, P. Larranaga and J. M. Picaza, "Analysis of the Behaviour of Genetic Algorithms when Learning Bayesian Network Structure from Data," *Pattern Recognition Letters*, Vol. 18, pp. 1269-1273, 1997.

[7] 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 Unceratainty in Artificial Intelligence*, pp. 206-215, 1999.

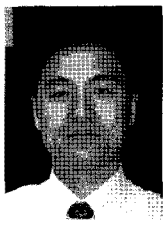
[8] K. Hwang, J. Lee, S. Chung and B. Zhang, "Construction of Large-Scale Bayesian Networks

by Local to Global Search," *PRICAI 2002*, pp. 375-384, 2002.
[9] L. E. Brown, I. Tsamardinos and C. F. Aliferis, "A Novel Algorithm for Scalable and Accurate Bayesian Network Learning," *MEDINFO*, 2004.
[10] S. Jung, K. Lee and D. Lee, "Enabling Large-Scale Bayesian Network Learning by Preserving Intercluster Directionality," *IEICE Transactions on Information and Systems*, Vol. E90-D, No. 7, 2007, to appear.

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 위촉연구원
1996년 ~ 2002년 : 전남대학교 전산학과, 의학과(겸임) 교수
2002년 ~ 현재 : 한국과학기술원 바이오및뇌공학과 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 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