효율적 구조 학습 알고리즘과 데이타 차원 축소를 통한 베이지안망 기반의 마이크로어레이 데이타 분석법

(A Method for Microarray Data Analysis based on Bayesian Networks using an Efficient Structural Learning Algorithm and Data Dimensionality Reduction)

황 규 백 † 장 정 호 † 장 병 탁 ††

(Kyu-Baek Hwang) (Jeong-Ho Chang) (Byoung-Tak Zhang)

요 약 DNA chip 기술에 의해 얻어지는 마이크로어레이(microarray) 데이타는 세포나 조직 내의 수천 개 유전자의 발현도(expression level)를 한번에 측정한 것으로, 유전자 발현 양상에 기반한 암의 진단, 유전자의 기능 예측 등에 이용되고 있다. 다양한 데이타 분석 기법들 중 베이지안망(Bayesian network)은 데이타의 각 속성들간의 관계를 그래프 형태로 표현할 수 있는 특징을 가지고 있다. 이는 마이크로어레이 데이타의 분석을 통해 여러 유전자와 조직의 특성(암의 종류 등) 사이의 관계를 밝히는데 유용하다. 하지만 대부분의 마이크로어레이 데이타는 sparse data로 베이지안망을 비롯한 각종 분석 기법의 적용을 어렵게 하고 있다. 본 논문에서는 베이지안망에 기반한 마이크로어레이 데이타 분석을 위해 효율적 구조 학습 알고리즘과 데이타 차원 축소를 이용한다. 제시되는 분석법은 실제 마이크로어레이 데이타인 NCI60 data set에 적용되었으며, 그 유용성은 데이타로부터 학습된 베이지안망이 실제 생물학적으로 알려진 사실들을 어느 정도 정확하게 표현하는지에 의해 평가되었다.

키워드 : 마이크로어레이 데이타 분석, 베이지안망, 데이타 차원 축소

Abstract *Microarray data, obtained from DNA chip technologies, is the measurement of the expression level of thousands of genes in cells or tissues. It is used for gene function prediction or cancer diagnosis based on gene expression patterns. Among diverse methods for data analysis, the Bayesian network represents the relationships among data attributes in the form of a graph structure. This property enables us to discover various relations among genes and the characteristics of the tissue (e.g., the cancer type) through microarray data analysis. However, most of the present microarray data sets are so sparse that it is difficult to apply general analysis methods, including Bayesian networks, directly. In this paper, we harness an efficient structural learning algorithm and data dimensionality reduction in order to analyze microarray data using Bayesian networks. The proposed method was applied to the analysis of real microarray data, i.e., the NCI60 data set. And its usefulness was evaluated based on the accuracy of the learned Bayesian networks on representing the known biological facts.*

Key words : microarray data analysis, Bayesian networks, data dimensionality reduction

---

· 이 연구는 과학기술부의 뇌신경정보학연구사업(BrainTech), IMT-2000 출연금 기술개발지원사업, 국가지정연구실사업(NRL) 및 서울대학교 간접연구경비에 의하여 지원되었음

† 비 회 원 : 서울대학교 컴퓨터공학부
kbhwang@bi.snu.ac.kr
jhchang@bi.snu.ac.kr

†† 종신회원 : 서울대학교 컴퓨터공학부 교수
btzhang@cse.snu.ac.kr

논문접수 : 2002년 4월 9일
심사완료 : 2002년 9월 2일

1. 서 론

DNA 마이크로어레이(microarray)는 세포나 조직 내의 수천 개 유전자의 발현 양상(gene expression pattern)을 한번에 볼 수 있게 하는 도구이다[1]. 유전자는 복잡한 생명 현상을 조절하는 역할을 하며, 그 발현 양상은 세포의 형질(phenotype)과 관련이 있다. 예를 들어, 간의 세포와 피부의 세포는 서로 다른 유전자 발현 양상을 보이며, 같은 조직의 세포라도 정상인 경우와 이상이 있는 경우(암과 같은)는 그 유전자 발현 양상이 다를 수 있다. 또한, 유전자의 발현은 세포의 주기, 다른 유전자의 발현, 외부 환경 등 수많은 요인의 영향을 받는다. 따라서, 유전자 발현 양상과 세포의 형질 사이의 관계나 여러 유전자의 발현간의 인과 관계에 대한 분석은 유전자 발현 양상에 기반한 질병 진단, 유전자 기능 예측, 유전자망(genetic network) 구성 등 생명과학 연구의 여러 분야에 응용된다. 한편, 여러 세포나 조직의 수천 개 유전자의 발현도(gene expression level)를 한번에 측정한 마이크로어레이 데이타는 high-throughput data analysis를 통한 유전자 발현 관련 연구를 가능케 하고있다.

지금까지 유전자 발현 관련 연구를 위한 마이크로어레이 데이타의 분석에는 통계학 및 기계학습 분야의 여러 기법들이 분석 목적에 따라 적용되어 왔다. 여러 분석 알고리즘들 중 대표적인 것들로는 계층적 클러스터링(hierarchical clustering)[2], PCA(Principal Component Analysis)[3], 신경망[4], 베이지안망(Bayesian network)[5, 6, 7, 8] 등이 있다. 각종 클러스터링 기법은 마이크로어레이 데이타 분석에 가장 널리 이용되고 있으며, 유전자 발현 양상에 기반한 질병 진단[9]이나 유전자 기능 예측[10] 등에 주로 사용된다. 신경망 계열의 패턴인식 기법은 유전자 발현 양상을 통한 질병 진단 및 연구 등에 이용되고 있다[4, 8]. 베이지안망은 유전자 발현 양상을 설명할 수 있는 확률모델을 제시할 수 있다는 특징이 있으며, 유전자 발현 양상에 기반한 암 종류의 구분[8], 유전자망의 구성[5, 6, 7] 등에 이용되어 왔다. 본 논문에서는 베이지안망 기반의 마이크로어레이 데이타 분석을 다룬다. 이 때 고려되어야 할 사항은 마이크로어레이 데이타가 sparse하다는 점이다. 베이지안망으로 sparse한 데이타를 분석하기 위한 기법으로 본 논문에서 제안하는 것은 효율적 베이지안망 구조 학습 알고리즘과 데이타 차원 축소이다.

논문의 구성은 다음과 같다. 우선 2절에서는 베이지안망에 기반한 마이크로어레이 데이타 분석에 관해 서술한다. 구체적으로, 베이지안망에 대한 기본적인 사항과 학습알고리즘, 베이지안망을 이용한 마이크로어레이 데이타 분석의 예, sparse한 데이타 분석의 어려움 및 관련 연구가 기술된다. 3절에서는 그러한 어려움을 해결하기 위해 본 논문에서 제시하는, 대규모 베이지안망의 효율적 구조 학습 알고리즘과 데이타 차원 축소 방법에 대해 서술한다. 4절에서는 실제 마이크로어레이 데이타인 NCI60 data set[11]을 베이지안망으로 분석한 결과를 제시하고, 생물학적으로 알려진 사실과의 비교를 통해 이를 검증한다. 마지막으로 5절에서 결론과 향후 연구 방향 등에 관해 서술한다.

2. 베이지안망 기반의 마이크로어레이 데이타 분석

2.1 베이지안망

베이지안망(Bayesian network)[12, 13]은 다수의 변수(random variable)들의 결합확률분포(joint probability distribution)를 효율적으로 표현하는 확률그래프모델(probabilistic graphical model)의 한 종류이다. 결합확률분포의 효율적인 표현을 위해서 변수들 사이의 조건부독립성(conditional independence)이 이용된다. 세 변수 집합 **X**, **Y**, **Z**[1]가 $P(\mathbf{y}, \mathbf{z}) > 0$을 만족하는 모든 **x**, **y**, **z**에 대해서 다음 수식을 만족할 때, '**X**는 **Z**의 값이 주어진 경우에 **Y**에 대해서 조건부독립'이라 한다.

$$P(\mathbf{x}|\mathbf{y},\mathbf{z}) = P(\mathbf{x}|\mathbf{z}) \quad (1)$$

확률그래프모델은 그래프의 형태를 띠고 있으며, 그래프의 노드는 변수와 일대일대응이 된다.[2] 이 그래프 구조는 변수들 사이의 조건부독립성들을 표현한다. 베이지안망은 간선에 방향이 있는, DAG(Directed-Acyclic Graph) 형태의 구조를 가진다. 베이지안망의 간선은 연결된 두 노드 사이에 확률적의존성(probabilistic dependence)이 있음을 나타낸다. 베이지안망의 구조가 표현하고 있는 변수들 사이의 조건부독립성은 다음과 같다. 모든 노드는 자신의 부모 노드들의 값이 주어진 경우에, 자신의 자손이 아닌 노드들과는 조건부독립이 된다. 이에 따르면, N개의 노드 $\{X_1, X_2, \cdots, X_N\}(=\mathbf{X})$을 가지는 베이지안망은 **X**의 결합확률분포를 다음과 같이 표현한다.

---

1) 본 논문에서는 대문자 알파벳(X,Y,Z 등)으로 변수를 표시한다. 굵은 대문자 알파벳(**X**,**Y**,**Z** 등)은 변수의 집합을 나타낸다. 소문자 알파벳(x,y,z 등)은 해당 변수에 값이 할당된 상태를 나타내고, 굵은 소문자 알파벳(**x**,**y**,**z** 등)은 해당 변수 집합의 모든 원소에 값이 할당된 상태를 가리킨다

2) 확률그래프모델의 노드는 변수와 일대일대응 관계에 있으므로, 본 논문에서 노드와 변수라는 용어는 같은 대상을 가리키며, 문맥에 따라 적절한 용어가 사용된다.

$$P(\mathbf{X}) = \prod_{i=1}^{N} P(X_i \mid \mathbf{Pa}(X_i)) \quad (2)$$

수식 (2)에서, $\mathbf{Pa}(X_i)$는 X_i의 부모 노드 집합을 나타낸다. $P(X_i|\mathbf{Pa}(X_i))$는 X_i에 대한 지역확률분포(local probability distribution)라 한다. 각 노드의 지역확률분포는 그 노드의 부모 노드의 값이 주어진 경우의 조건부확률분포이다. 다음 절에서는 베이지안망을 데이타에서 학습하는 방법에 대해 서술한다.

2.2 베이지안망의 학습

일반적으로 베이지안망의 학습은 두 부분으로 구성된다. 첫 단계는 베이지안망의 DAG 구조를 학습하는 것이다. 두번째는 고정된 베이지안망 구조에서 각 노드의 지역확률분포를 학습하는 것이다. 두번째 단계인 지역확률분포의 학습은 적절한 가정 하에서 간단한 계산으로 해결된다[13]. 여기서는 구조의 학습에 대해 기술한다. 구조 학습에는 점수기반탐색이 널리 이용된다. 이 기법에서 학습알고리즘은 주어진 학습데이타에 가장 적합한 망 구조를 탐색한다. 학습데이타에 대한 베이지안망 구조의 적합성은 점수의 형태로 측정된다. 널리 이용되는 망 구조의 점수에는 BD(Bayesian Dirichlet) 점수와 MDL(Minimum Description Length) 점수가 있으며, 둘은 점근적으로 서로 같은 값을 가진다는 사실이 알려져 있다[14]. 본 논문에서는 BD 점수[15]를 사용한다. BD 점수는 베이지안망의 지역확률분포가 다항분포인 경우에 사용되며, 학습데이타 D와 망 구조 G에 대한 결합확률로 다음과 같은 식으로 표현된다.

$$P(G,D) = P(G)\cdot P(D \mid G)$$
$$= P(G)\cdot \prod_{i=1}^{N}\prod_{j=1}^{q_i}\frac{\Gamma(\alpha_{ij})}{\Gamma(\alpha_{ij}+N_{ij})}\cdot\prod_{k=1}^{r_i}\frac{\Gamma(\alpha_{ijk}+N_{ijk})}{\Gamma(\alpha_{ijk})} \quad (3)$$

수식 (3)에서 $P(G)$는 망 구조 G의 사전확률(prior probability)이며, $\Gamma(\cdot)$는 $\Gamma(1)=1$, $\Gamma(x+1)=x\cdot\Gamma(x)$를 만족하는 감마함수이다. N은 노드의 개수이며 q_i는 $\mathbf{Pa}(X_i)$가 가질 수 있는 상태의 개수이다. r_i는 X_i가 가질 수 있는 상태의 개수이다. α_{ijk}는 지역확률분포모델의 파라미터에 대한 분포모델[3]의 hyper parameter로, 파라미터의 분포에 대한 사전지식(prior knowledge)에 해당한다. N_{ijk}는 학습데이타 D에서 X_i가 $\mathbf{Pa}(X_i)$의 j번째 상태 하에서 k번째 값을 가지는 경우의 횟수이다. α_{ij}와 N_{ij}는 각각 다음과 같이 계산된다.

$$\alpha_{ij} = \sum_{k=1}^{r_i}\alpha_{ijk} \qquad N_{ij} = \sum_{k=1}^{r_i} N_{ijk} \quad (4)$$

실제 구현에서는 계산상의 문제 때문에 BD 점수의 로그값이 이용된다.

점수기반탐색 기법의 핵심이 되는 부분은 탐색 전략이다. 가능한 망 구조의 개수는 노드 개수의 지수승 이상(super exponential)이며, 가장 점수가 높은 베이지안망을 찾는 문제는 NP-hard임이 알려져 있다[16]. 따라서 greedy 탐색 알고리즘이 일반적으로 이용된다. Greedy 탐색 알고리즘은 최적의 해를 찾는 것을 보장하지는 않지만, 실제 문제에 효율적으로 적용될 수 있음이 보여져 왔다[13, 14]. 베이지안망 구조 학습을 위한 greedy 탐색 알고리즘은 다음과 같다.

1. 초기 그래프 구조 G_0를 생성한다. (초기 구조로는 간선이 없는 빈 그래프 구조가 많이 이용된다.)
2. 다음의 과정을, $m=1, 2, 3, \cdots$에 대해서 수렴할 때까지 반복한다.
   2-1. G_{m-1}의 구조에서 가능한 모든 지역 변화(간선의 추가, 간선의 방향 전환, 간선의 삭제) 중 망 구조의 점수를 가장 크게 향상시키는 지역 변화를 찾는다.
   2-2. 2-1에서 찾은 지역 변화를 행하고, 변화된 그래프 구조를 G_m으로 한다.

이 알고리즘의 수렴 조건은 G_{m-1}의 점수가 G_m의 점수와 같은 경우이다.

2.3 마이크로어레이 데이타 분석을 위한 베이지안망

마이크로어레이 데이타 분석의 경우, 베이지안망의 노드는 유전자에 해당하며 노드가 가지는 값은 유전자의 발현도(gene expression level)가 된다. 그 외에 실험 조건, 조직의 종류, 세포 주기 등을 나타내는 노드를 둘 수 있다. 그림 1은 유전자와 암의 종류를 노드로 가지는 베이지안망의 한 예이다.

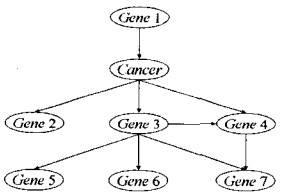


그림 1 암의 종류와 유전자를 노드로 가지는 베이지안망의 예. 'Cancer' 노드(맨 위에서 두번째 노드)는 암의 종류를 값으로 가지며, 유전자에 해당하는 노드('Gene 1' ~ 'Gene 7')는 유전자의 발현도를 값으로 가진다. 이 베이지안망은 유전자 발현과 암의 종류 사이의 다양한 조건부독립성을 표현하고 있다. 그림에 표시되어 있지는 않지만 각 노드는 자신의 지역확률분포를 가지고 있다.

3) 지역확률분포모델이 다항분포인 경우에는 디리슈레분포(Dirichlet distribution)가 이용된다[13].

베이지안망을 통한 마이크로어레이 데이타 분석은 크게 정성적 분석(qualitative analysis)과 정량적 분석(quantitative analysis)으로 나눌 수 있다. 정성적 분석은 베이지안망의 그래프 구조를 이용한 분석이다. 예를 들어, 그림 1의 베이지안망에서 '*Gene* 3'과 '*Gene* 4'의 발현이 공통으로 '*Gene* 7'의 발현에 영향을 준다는 사실을 알 수 있다. 또한, '*Gene* 1'은 '*Cancer*'와 '*Gene* 3'을 거쳐서 '*Gene* 6'에 영향을 준다는 사실도 알 수 있다. 정량적 분석은 각 노드 사이의 확률적의존성을, 조건부확률을 계산함으로써 정량화하는 분석이다. 예를 들어, 그림 1의 베이지안망에서 조건부확률 P('*Gene* 7'|'*Gene* 3', '*Gene* 4')를 계산해 봄으로써 '*Gene* 3', '*Gene* 4'의 발현이 '*Gene* 7'의 발현에 어떠한 영향을 주는지 확인할 수 있다.[4] 또, P('*Cancer*'| '*Gene* 1', '*Gene* 2', '*Gene* 3', '*Gene* 4')를 계산해서 '*Gene* 1', '*Gene* 2', '*Gene* 3', '*Gene* 4'의 발현도가 주어진 경우에 암의 종류를 예측할 수도 있다. 베이지안망에서 조건부확률을 계산하는 것은 확률적 추론(probabilistic inference)이라고 한다. 임의의 베이지안망에서 임의의 조건부확률을 추론하는 것은 NP-hard 문제이지만[17, 18], 대부분의 경우 베이지안망의 구조를 이용한 효율적 추론알고리즘의 적용이 가능하다[12, 19, 20].

2.4 마이크로어레이 데이타 분석의 어려움 및 관련 연구

마이크로어레이 데이타 분석의 어려움은 대부분의 마이크로어레이 데이타가 sparse하다는데 기인한다. 특정 연구를 위해 제작되는 마이크로어레이의 개수는 비용 문제와 표본 준비의 어려움 때문에 보통 수십 개에 불과하다. 반면 한장의 마이크로어레이에는 보통 수천 개의 유전자의 발현도가 측정된다. 마이크로어레이 데이타에서 하나의 마이크로어레이는 하나의 데이타 예제에 해당하며 각 유전자는 데이타 속성에 해당하므로 마이크로어레이 데이타의 분석은 수천 개의 속성과 수십 개의 예제를 가지는 sparse data의 분석이 된다.

마이크로어레이 데이타의 이러한 특성은 베이지안망의 적용에 다음과 같은 문제점들을 야기한다. 첫째는 베이지안망의 학습에 관한 문제이다. 마이크로어레이 데이타의 분석을 위해서 수백 개의 노드를 가지는 대규모 베이지안망을 학습해야 하는 경우가 있으며, 일반적인 베이지안망 구조 학습 알고리즘은 시간 및 공간 복잡도 때문에 이러한 경우에의 적용이 어렵다. 2.2절의 greedy 탐색 알고리즘은 노드의 개수가 N인 베이지안망을 학습할 때, 매 단계마다 $O(N^2)$의 지역 변화를 검사한다. 이는 N이 수백에 이르는 경우 막대한 학습 시간과 점수 계산에 필요한 공간을 요구한다. 두번째 문제는 수백 개의 노드를 가지는 베이지안망에서의 확률적 추론(probabilistic inference)의 어려움이다. 세번째 문제는 데이타 속성의 개수에 비해 데이타 예제의 개수가 적기 때문에 발생하는, 학습 결과에 대한 낮은 신뢰도이다.

이러한 문제들에도 불구하고 다수의 연구자들이 베이지안망을 이용한 마이크로어레이 데이타의 분석을 시도했다. [8]은 백혈병의 구분을 위해서, 통계적으로 선택된 4개의 유전자로 구성된 베이지안망 분류기를 구축했다. [6]은 마이크로어레이 데이타에서 10개 미만의 유전자만을 선택한 뒤, 은닉노드(hidden node)를 가지는 베이지안망을 통한 유전자 발현 양상 분석을 시도했다. 대규모 베이지안망 학습의 문제 해결을 위해, [21]은 sparse candidate 알고리즘을 제안했다. [22]는 학습 결과의 신뢰도 측정을 위해 bootstrap 기법의 이용을 제안했다. [5]는 이러한 방법에 기반해서 800개에 이르는 효모의 유전자로 구성된 베이지안망을 마이크로어레이 데이타로부터 학습했다. [7]은 학습 결과의 신뢰도를 높이기 위해 생물학적 사전지식과 베이지안망 학습을 결합한 마이크로어레이 데이타 분석법을 제시했다.

본 논문에서는 대규모 베이지안망의 구조를 효율적으로 학습할 수 있는 알고리즘과 클러스터링에 기반한 데이타 차원 축소를 이용한 베이지안망 기반의 마이크로어레이 데이타 분석법을 제시한다. 데이타 차원 축소는 확률적 추론과 신뢰할 수 있는 학습 결과를 위해서 데이타 속성의 개수를 줄이는 것이다. 특히, 확률적 추론을 위한 데이타 차원의 축소는 지금까지 제안된 기법과 구별되는 부분이다. 다음 절에서는 이러한 분석법에 대해서 서술한다.

3. 대규모 베이지안망의 효율적 구조 학습 알고리즘과 데이타 차원 축소 기법

3.1 대규모 베이지안망의 구조 학습 알고리즘

수백 개의 노드를 가지는 대규모 베이지안망의 학습 알고리즘에는 [21, 23]이 있다. 이 알고리즘들은 전체 탐색 공간 중 불필요한 부분을 미리 제거함으로써, 효율적인 greedy 탐색을 행한다. 불필요한 부분을 줄이는 전략이 각 알고리즘의 핵심이 된다. [21]은 greedy 탐색을 행하기 전에 각 노드의 부모 노드의 후보를 제한함으로써 전체 탐색공간을 줄이는 sparse candidate 알고리즘을 제안했다. 본 논문에서는 [23]의 local to

4) 한 유전자의 발현이 다른 유전자의 발현에 영향을 주는 경우는, 증가(up-regulation)와 감소(down-regulation)로 크게 나눌 수 있다.

global 탐색 알고리즘을 적용했다. Local to global 탐색 알고리즘은 각 노드의 Markov blanket[19] 구조를 미리 구성함으로써 전체 탐색 공간을 줄인다.

3.1.1 Markov blanket 구조

변수 집합 $\mathbf{X}(=\{X_1, X_2, \cdots, X_N\})$가 주어졌을 때, 변수 X_i의 Markov blanket, $\mathbf{MB}(X_i)$는 다음을 만족하는 $\mathbf{X}-\{X_i\}$의 부분집합이다.[5)]

$$P(X_i \mid \mathbf{X}-\{X_i\}) = P(X_i \mid \mathbf{MB}(X_i)) \quad (5)$$

즉, X_i는 $\mathbf{MB}(X_i)$의 모든 원소들의 값이 주어진 경우, $\mathbf{X}-\{X_i\}-\mathbf{MB}(X_i)$의 변수들과 조건부독립이 된다. 베이지안망 구조 G가 주어진 경우, 노드 X_i의 Markov blanket, $\mathbf{MB}_G(X_i)$는 X_i의 부모 노드들, 자식 노드들, 배우자 노드들로 구성된다. 본 논문에서는 X_i와 $\mathbf{MB}_G(X_i)$의 원소들을 노드로 가지는, G의 부분그래프를 X_i의 'Markov blanket 구조'로 정의한다.

3.1.2 Local to global 탐색 알고리즘

Local to global 탐색 알고리즘의 핵심은 각 노드의 Markov blanket 구조를 미리 구성함으로써 전체 greedy 탐색에 걸리는 시간을 줄이는 것이다. 알고리즘은 다음과 같이 진행된다.

1. 초기 그래프 구조 G_0를 생성한다. (G_0는 간선이 없는 빈 그래프 구조가 많이 이용된다.)
2. 다음의 과정을, $m=1, 2, 3, \cdots$에 대해서 수렴할 때까지 반복한다.
   2-1. G_{m-1}과 학습데이타 D에 기반해서 각 노드의 Markov blanket 구조를 구성한다.
   2-2. 모든 노드의 Markov blanket 구조를 합해서 그래프 H_m을 만든다. (H_m은 cycle을 가질 수 있다.)
   2-3. H_m에서 directed cycle을 구성하지 않는 간선들은 G_m의 간선으로 고정된다.
   2-4. H_m의 부분그래프 중 점수가 좋은 베이지안망 구조 G_m을 greedy 탐색을 통해서 찾는다.

위의 알고리즘에서는 각 노드의 Markov blanket 구조의 구성(알고리즘의 2-1)과 greedy 탐색(알고리즘의 2-4)이 교대로 수행된다. 수렴 조건은 G_{m-1}의 점수가 G_m의 점수와 같거나 큰 경우이며 최종 학습 결과는 G_{m-1}이 된다. 알고리즘의 2-1의 Markov blanket 구조 구성 방법은 지면 관계상 본 논문에서 서술하지는 않는다. 자세한 방법은 [23]에 서술되어 있다. 또한, [23]은 local to global 탐색 알고리즘과 [21]의 sparse candidate 알고리즘, 그리고 일반적인 greedy 탐색 알고리즘의 성능을 2개의 인공 데이타와 하나의 실제 데이타에 대한 실험을 통해 비교했다. 비교 결과에 대한 요약은 다음과 같다. Local to global 탐색 알고리즘과 sparse candidate 알고리즘 모두 수백 개의 노드를 가지는 베이지안망을 적절한 시간 내에 학습할 수 있다. 또한, 변수가 가지는 값의 개수가 크지 않은 경우의 학습 속도는 두 알고리즘이 비슷하다. 학습의 정확도 역시 대체로 비슷하다. 다만, 변수의 개수가 많은 경우, local to global 탐색 알고리즘이 BD 점수가 조금 더 높은 망 구조를 학습하는 경우가 있었다. 마이크로어레이 데이타의 경우 변수가 가지는 값의 개수가 크지 않으므로(대부분 2 ~ 3 정도), local to global 탐색 알고리즘의 적용이 적합하다고 할 수 있다.

3.2 클러스터링에 기반한 데이타 차원 축소

수백 개의 노드를 가지는 베이지안망에서의 확률적 추론은 거의 불가능하다. 본 논문에서는 이를 위해 마이크로어레이 데이타의 속성의 개수를 줄여서 분석하는 방법을 택했다. 이러한 데이타 차원 축소에는, 개개의 속성대신 비슷한 특성을 보이는 속성들의 대표값(prototype)을 이용하는 방법과 전체 속성 중 관심있는 부분만을 선택하는 방법이 있으며, 두 가지 방법 모두 데이타의 속성에 대한 클러스터링을 통해 이루어진다.

본 논문에서는 속성의 클러스터링을 위해 STVQ (Soft Topographic Vector Quantization) 방법[24]을 사용한다. STVQ는 통계물리학 이론에 기반을 둔 클러스터링 알고리즘으로 안정화된 클러스터링 결과와 더불어 SOM(Self-Organizing Map)과 같이 데이타에 대한 topographic map을 제공한다. STVQ에서 클러스터링을 위한 cost function은 수식 (6)과 같이 주어진다.

$$E = \sum_{i=1}^{N}\sum_{j=1}^{M} m_{ij} e_{ij} \quad (6)$$

N과 M은 각각 데이타 속성의 개수와 클러스터의 개수를 의미하며, m_{ij}는 속성 i가 클러스터 j에 속하는지의 여부를 나타내는 이진변수이다. e_{ij}는 속성 i가 클러스터 j에 할당될 때의 에러이며 수식 (7)과 같이 정의된다.

$$e_{ij} = \frac{1}{2}\sum_{k=1}^{M} h_{jk} \|\mathbf{x}_i - \mathbf{z}_k\|^2, \quad \sum_{k=1}^{M} h_{jk} = 1 \ (\forall j) \quad (7)$$

$\mathbf{x}_i$는 속성 i의, 모든 데이타 예제에서의 값을 나타내는 벡터이며, $\mathbf{z}_k$는 클러스터 k의 중심 벡터값으로 해당 클러스터에 할당된 속성들의 가중치 평균값(weighted average)에 의해 계산된다. h_{jk}는 클러스터 j와 클러스터

5) 변수 집합에서 한 원소의 Markov blanket은 여러 개 존재할 수 있으며, [19]는 그 중 가장 작은 집합을 Markov boundary라 했다. 본 논문에서 Markov blanket은 Markov boundary를 가리킨다.

k 사이의 neighborhood 함수이며, 이를 통해 SOM에서와 같이 데이타에 대한 2차원 또는 3차원 상의 가시화된 클러스터 구조를 제공한다. 학습의 초기화 과정에서 클러스터 개수는 사용자에 의해 미리 주어지며, deterministic annealing에 기반하여 EM 알고리즘[25]을 반복적으로 수행함으로써 클러스터링을 수행한다.

벡터 $\mathbf{x}_i$와 중심 벡터 $\mathbf{z}_k$ 사이의 거리 척도(distance measure)로는 피어슨 상관계수(Pearson correlation coefficient) r_{ik}를 사용하며, 표준화된(평균 0, 표준편차 1) 벡터들의 경우, 이는 유클리드 거리(Euclidean distance)와 밀접한 관련이 있다. 즉,

$$\|\mathbf{x}_i-\mathbf{z}_k\|^2=(\mathbf{x}_i^T\mathbf{x}_i+\mathbf{z}_k^T\mathbf{z}_k-2\mathbf{x}_i^T\mathbf{z}_k)$$
$$=\left(2d-2d\times\frac{\mathbf{x}_i^T\mathbf{z}_k}{d}\right)=2d(1-r_{ik}) \quad (8)$$

이다. 수식 (8)에서 d는 $\mathbf{x}_i$와 $\mathbf{z}_k$의 차원, 즉 데이타 예제의 개수를 의미한다.

데이타 속성에 대한 클러스터링 결과는 비슷한 속성들을 원소로 가지는 M개의 속성 클러스터들이다. 전체 속성 중 관심있는 부분의 속성만을 분석에 이용하는 경우는 이러한 클러스터들 중 일부 클러스터에 속하는 속성들만을 분석에 이용한다. 이 때는 분석하려는 데이타의 특징과 분석 목적을 고려해야 한다. 다른 방법은 속성들의 대표값을 이용하는 것이며, 차원이 축소된 데이타에서 l번째 데이타 예제의 k번째 속성 대표값 $\mathbf{v}_k(l)$은 k번째 클러스터 C_k에 속한 속성값들의 평균값으로 다음과 같이 계산된다($k=1, 2, \cdots, M$).

$$\mathbf{v}_k(l)=\frac{1}{|C_k|}\sum_{i\in C_k}\mathbf{x}_i(l) \quad (9)$$

$|C_k|$는 C_k에 속하는 데이타 속성들의 개수이며, $\mathbf{x}_i(l)$은 l번째 데이타 예제의 i번째 속성의 값을 의미한다($i=1, 2, \cdots, N$). 수식 (9)에 의해서 M개의 속성 대표값을 구하고 분석에 이용할 수 있다.

4. 실 험

4.1 NCI60 data set

NCI60 data set[11]은 미국의 국립암연구소(NCI: National Cancer Institute)에서 신약개발과정에 이용하기 위해 만든 데이타이다. 데이타는 60개의 암 조직으로 구성되어 있다. 데이타의 속성에는 암의 종류, 유전자, 약물이 있다. 암의 종류는 결장암, 신장암, 난소암, 유방암, 전립선암, 폐암, 중추신경계암, 백혈병, 피부암의 9개 중 하나이다. 유전자 속성은 각 암 조직의 9,703개의 유전자의 발현도로 cDNA 칩[1]을 이용한 마이크로어레이로 측정되었다. 약물 속성은 각 암 조직에 대한 1,400종의 약물의 활성이다. 실험에서 베이지안망 분석은 [11]과 마찬가지로 암의 종류에 따른 발현도 변화를 강하게 보이는 1,376개의 유전자와 실제 치료에 이용되고 있는 항암제 118개에 대해 행해졌다. 또한, 보다 정확한 분석을 위해 유전자와 약물 중 결측치를 4개 이상 가지는 것과 이름이 없는 유전자는 속성에서 제외했다. 결과적으로, 분석에 이용된 데이타는 890개의 속성(805개의 유전자, 84개의 약물, 암의 종류)을 가지는 60개의 데이타 예제로 구성된다.

4.2 NCI60 data set의 차원 축소

NCI60 data set의 속성 중 암의 종류는 그대로 실험에 이용되었다. 다른 속성들의 축소는 대표값을 이용하는 경우와 속성들 중 관심있는 부분을 선택하는 두 가지 방법을 모두 이용하였다. 유전자 및 약물 대표값은 각각 별도로 계산되었으며, 대표값의 개수는 다양한 경우를 실험하였다. 그림 2는 실험에 이용된 약물과 유전자의 대표값의 개수를 나타낸다.

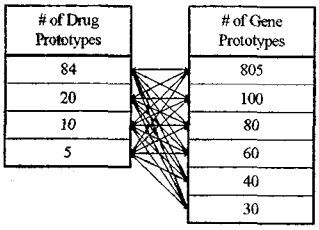


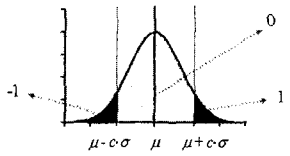
그림 2 실험에 이용된 약물 및 유전자 대표값의 개수. 약물 대표값의 개수로는 5, 10, 20의 3가지 경우를 실험하였으며, 유전자 대표값의 개수로는 30, 40, 60, 80, 100의 5가지 경우를 실험하였다. 그 밖에 개개의 속성을 그대로 사용한 실험(약물: 84개, 유전자: 805개)도 행했으며, 실험에 이용된 대표값 data set의 종류는 총 24(=4×6)가지이다.

속성의 부분집합 선택은 다음과 같이 이루어졌다. 약물 L-asparaginase 주변의 확률적 관계[11]를 분석하기 위해 유전자와 약물을 함께 클러스터링했다. 이 때, 약물의 반응은 ABCB1 유전자와 같은 약물 저항 유전자에 의해 비활성화되는 경우가 있다는 점에 착안하여 [11], 각 약물은 60개의 데이타 예제에서의 활성에 대한 음(negative)의 값의 벡터로 표현되도록 하였다. 이후 topographic map 형태의 STVQ 클러스터링 결과에서 L-asparaginase가 속한 클러스터와 그 인접 클러스

터들을 분석의 대상으로 삼았다. 이렇게 해서 12개의 유전자와 4개의 약물이 선택되었다.

4.3 속성값의 이산화

데이타에서 유전자의 발현도와 약물의 활성은 실수값이다. 이 값들은 베이지안망의 지역확률분포모델로 다항분포모델을 이용하기 위해 이산화(discretization)되었다. 이산화는 low(-1), normal(0), high(1)의 3구간으로 행해졌다. 이산화를 위한 경계값은 각 속성의 평균과 표준편차에 따라 결정했다. 구체적으로 $\mu-c\cdot\sigma$와 $\mu+c\cdot\sigma$가 경계값으로 이용되었다. 여기서 μ는 속성의 평균값이며 σ는 표준편차이다. c는 경계값의 위치를 결정하기 위한 상수로 0.43, 0.50, 0.60이 이용되었다. 그림 3은 각 속성값이 표준정규분포를 따른다는 가정 하에 low(-1), normal(0), high(1)의 c값에 따른 분포이다.



c	Distribution Ratio		
	-1	0	1
0.43	33.3%	33.3%	33.3%
0.50	30.8%	38.3%	30.8%
0.60	27.4%	45.1%	27.4%

그림 3 이산화 경계값 결정 상수 c에 따른 각 속성값의 분포도(각 속성값이 표준정규분포를 따른다고 가정).

4.4 분석 결과

여기서는 25가지의 data set(24종류의 대표값 data set과 부분집합 data set) 중 3종류의 data set에 대한 베이지안망 분석 결과를 서술한다. 3개의 data set은 데이타 차원 축소를 행하지 않은 data set(Data Set 1), 대표값 data set(Data Set 2)과 부분집합 data set (Data Set 3)이다. 표 1은 이 data set들의 특성과 적용된 베이지안망 학습알고리즘 및 파라미터, 학습 시간, 확률적 추론 알고리즘의 적용 가능성을 나타내고 있다.

4.4.1 Data Set 1에 대한 분석 결과

Data Set1에서 학습된 베이지안망은 890개의 노드로 구성되어 있다. 이 베이지안망은 확률적 추론 알고리즘의 적용이 불가능하므로, 각 노드에 연결된 간선의 개수만을 분석한다. 베이지안망의 간선은 노드 사이에 확률적의존성이 있음을 나타내기 때문에, 간선이 많이 연결된 노드는 많은 노드들에 영향을 주는 노드로 해석된다. 표 2는 간선이 많이 연결되어 있는 노드 10개를 나열한 것이다. 가장 많은 영향을 끼치는 노드는 암의 종류이며 나머지 9개는 모두 유전자 노드이다. 이 9개의 유전자들은 암의 종류에 따라 변이를 많이 보이는 유전자로 볼 수 있다.

표 1 각 data set의 특성 및 적용된 베이지안망 학습알고리즘, 학습 시간, 확률적 추론 알고리즘의 적용 가능성. Greedy 탐색 알고리즘의 괄호 안의 숫자는 random 초기화에 의한 greedy 탐색 알고리즘의 적용 횟수를 나타낸다. Local to global 탐색 알고리즘에서 괄호 안의 숫자는 알고리즘에 적용된 파라미터값을 나타낸다([23] 참고). 실험은 256MB의 RAM을 가지는 펜티엄 III 1GHz 기계에서 행해졌다.

	Data Set 1	Data Set 2	Data Set 3
유전자 속성 개수	805	40	12
약물 속성 개수	84	5	4
Greedy 탐색 알고리즘	적용 불가	○(20)	○ (100)
Local to global 탐색 알고리즘	○(5~8)	○(5~15)	적용할 필요 없음
평균학습시간(초)	3233.7	123.9	15.6
확률적 추론	불가능	가능	가능

4.4.2 Data Set 2에 대한 분석 결과

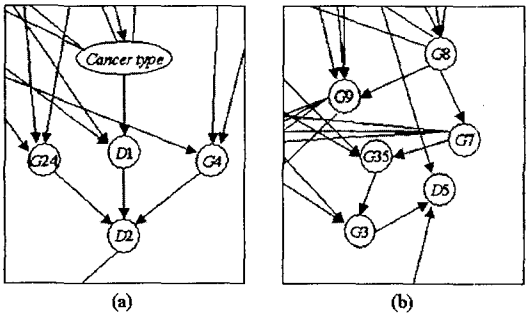
40개의 유전자 대표값과 5개의 약물 대표값으로 구성된 베이지안망에서는 생물학적으로 알려져 있는 관계인 ASNS 유전자와 L-asparaginase 약물간의 negative correlation과 DPYD 유전자와 5FU 약물(fluorouracil) 사이의 negative correlation이 각각 분석되었다. 그림 4는 학습된 베이지안망의 두 부분을 나타낸다. 그림 4(a)에서 $G4$는 ASNS를 포함하는 유전자 대표값이고 $D2$는 L-asparaginase를 포함하는 약물 대표값이다. 그림에서 $G4$와 $D2$는 직접적으로 의존관계에 있다. 그림 4(b)에서 $G8$은 DPYD를 포함하는 유전자 대표값이며 $D5$는 5FU를 포함하는 약물 대표값이다. $G8$과 $D5$는 직접적으로 연관 관계를 가지고 있지는 않음을 알 수 있다. 표 3은 이 베이지안망에서 추론된 조건부확률 $P(D2|G4)$를 나타낸다. 추론된 조건부확률은 $D2$와 $G4$사이의 negative correlation을 명확하게 보이지는 않고 있다. 예를 들어, $P(D2=\text{low}|G4=\text{high})$는 $P(D2=\text{high}|G4=\text{high})$보다는 커야 한다. 결과적으로 46개의 노드를 가지는 베이지안망은 알려져 있는 생물학적 사실을 명확하게 보이는데 실패했다고 할 수 있다. 이런 결과는 이산화 과정에서의 정보 손실이나 개개의 속성 대신 속성의 대표값을 사용하는데 기인하는 것으로 여겨진다.

4.4.3 Data Set 3에 대한 분석 결과

그림 5는 Data Set 3에서 학습된 17개의 노드를 가지는 베이지안망의 일부분을 나타낸다. 그림에서 암의 종류와 L-asparaginase의 활성 사이의 직접적인 의존관계를 파악할 수 있으며, L-asparaginase와 ASNS와

표 2 890개의 노드 중 다른 노드들에 영향을 많이 주는 노드 10개. 가장 많은 영향을 주는 노드는 암의 종류를 나타내는 'Cancer type' 노드이다. 나머지 9개는 모두 유전자 노드이다. 간선 개수는 이산화 경계값 결정 상수 c의 값(0.43, 0.50, 0.60)에 따른 3개의 베이지안망에 대한 평균값이다.

노드 이름	노드에 연결된 간선의 개수
Cancer type	125.0
SID W 487878, SPARC/osteonectin [5′:AA046533, 3′:AA045463]	25.0
Homo sapiens Cyr61 mRNA, complete cds Chr.1 [486700, (DIW), 5′:AA044451, 3′:AA044574]	18.3
SID W 162479, Homo sapiens epithelial-specific transcription factor ESE-1b (ESE-1) mRNA, complete cds [5′:H27938, 3′:H27939]	16.0
CDH2 Cadherin 2, N-cadherin (neuronal) Chr. [325182, (DIRW), 5′:W48793, 3′:W49619]	13.7
H.sapiens mitogen inducible gene mig-2, complete CDS Chr.14 [488643, (IW), 5′:AA045936, 3′:AA045821]	13.3
SID W 429623, Homo sapiens clone 24659 mRNA sequence [5′:AA011634, 3′:AA011635]	13.3
SID W 290871, Integrin alpha-3 subunit [5′:N99380, 3′:N71998]	13.0
COL4A1 Collagen, type IV, alpha 1 Chr.13 [145292, (EW), 5′:R78225, 3′:R78226]	12.7
COL4A1 Collagen, type IV, alpha 1 Chr.13 [489467, (IEW), 5′:AA054624, 3′:AA054564]	12.7



(a) (b)

그림 4 Data Set 2에서 학습된 46개의 노드를 가지는 베이지안망의 일부. G로 시작하는 노드는 유전자 대표값, D로 시작하는 노드는 약물 대표값을 나타낸다. 'Cancer type' 노드는 암의 종류를 나타낸다.

의 확률적의존성도 알 수 있다. 또한 ASNS는 P5CR 유전자에 직접적으로 의존함도 알 수 있다. 표 4, 5는 이 베이지안망에서의 확률적 추론의 몇몇 결과를 보여준다. 표 4의 조건부확률은 ASNS와 L-asparaginase 사이의 negative correlation과 일치한다. 게다가 암의 종류가 백혈병(leukemia)일 때에는 더 강한 negative correlation을 보인다 (표 4의 괄호안의 수치). 이러한 사실 역시 생물학적으로 알려진 사실과 일치한다[11].

표 3 그림 4의 베이지안망에서의 $P(D2|G4)$에 대한 확률적 추론 결과

	$D2$=low	$D2$=normal	$D2$=high
$G4$=low	0.32096	0.27086	0.40818
$G4$=normal	*0.31387*	*0.41247*	*0.27366*
$G4$=high	0.32167	0.34920	0.32913

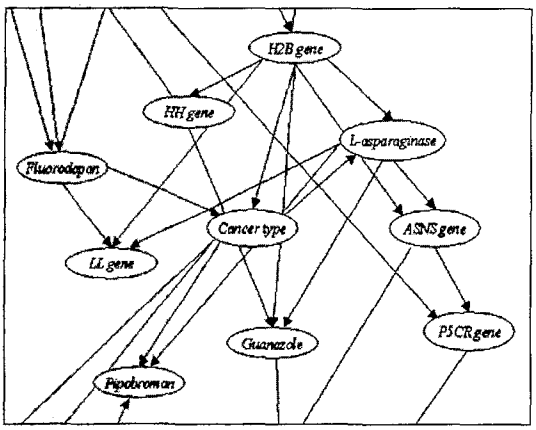


그림 5 Data Set 3에서 학습된 17개의 노드를 가지는 베이지안망의 일부

표 4 그림 5의 베이지안망에서의 $P(L\text{-}asparaginase|ASNS)$에 대한 확률적 추론 결과. 괄호 안의 숫자는 $P(L\text{-}asparaginase|ASNS$, *'Cancer type'* = 'Leukemia')의 확률적 추론 결과이다.

	L-asparaginase = low	*L-asparaginase* = normal	*L-asparaginase* = high
ASNS = low	0.19857 (0.17536)	0.27471 (0.22838)	0.52672 (0.59626)
ASNS = normal	0.31110 (0.27128)	0.49795 (0.53790)	0.19095 (0.19081)
ASNS = high	0.42159 (0.38500)	0.36279 (0.42437)	0.21561 (0.19063)

표 5 그림 5의 베이지안망에서의 $P(L\text{-}asparaginase|P5CR)$에 대한 확률적 추론 결과.

	L-asparaginase = low	*L-asparaginase* = normal	*L-asparaginase* = high
P5CR = low	0.27510	0.35226	0.37263
P5CR = normal	0.31621	0.41072	0.27307
P5CR = high	0.33837	0.39664	0.26499

또한 표 5에서는 P5CR과 L-asparaginase 사이의 negative correlation도 확인할 수 있다. 실제로, P5CR은 alanine과 aspartate metabolism에 관여하고 있는 유전자이다. 또한, ASNS는 arginine과 proline metabolism에 관계하고 있다. 이 두 metabolism은 KEGG(Kyoto Encyclopedia of Genes and Genomes)의 metabolic and regulatory pathway(http:// www.genome.ad.jp/kegg)에서 서로 상당히 인접한 위치에 있다. 따라서, L-asparaginase와의 negative correlation에 있어서 P5CR과 ASNS가 비슷한 양상을 보이는 것은 생물학적으로도 의미를 가질 가능성이 크다.

5. 결론 및 향후 연구

본 논문에서는 베이지안망을 이용해서 마이크로어레이 데이타를 분석하는 방법을 제시했다. 특히 대규모 베이지안망의 효율적 구조 학습 알고리즘을 적용했으며, 확률적 추론을 이용한 지식 추출을 위해서 클러스터링에 기반한 데이타 차원 축소를 행했다. 제시된 기법은 기존의 베이지안망을 이용한 마이크로어레이 분석[5, 6, 7]과는 달리 확률적 추론을 이용한 정량적(quantitative) 분석을 행했다는 특징이 있다. 제시된 분석법은 실제 마이크로어레이 데이타에 적용되었으며 생물학적으로 의미있는 다수의 사실을 데이타 분석만을 통해서 밝혀낼 수 있었다. 특히, 4.4.3절의 ASNS 유전자와 P5CR 유전자의 L-asparaginase 약물에 대한 저항성이 밀접하다는 발견은 실제 생물학적인 검증을 통해서 그 가능성이 매우 높음을 보였다. 이 사실은 제시된 기법이 생명공학 연구에 있어서 high-throughput data analysis를 통한 가설생성기(hypothesis generator)의 역할을 할 수 있음을 보인다. 또한, 마이크로어레이 데이타 분석에 있어서 속성의 대표값을 이용하는 것은 데이타의 정보를 모두 고려하지 못하기 때문에 잘못된 분석 결과를 낼 수 있다는 사실도 보였다.

앞으로의 연구방향은 다음과 같다. 우선, 대규모 베이지안망에서도 특수한 경우에 따라서 효율적으로 확률적 추론을 할 수 있는 알고리즘의 개발이다. 이는 데이타 차원 축소를 행하지 않고서도 확률적 추론을 통한 정량적 분석을 하기 위해서이다. 또한 클러스터링 결과와 베이지안망의 구조와의 관계에 대한 보다 정밀한 분석이 필요하다. 이를 통해 적절한 데이타 차원 축소의 정도를 결정할 수 있을 것이다. 마지막으로 sparse data에서 신뢰도 있는 결과를 얻기 위한 베이지안망 구조 학습법이 더욱 연구되어야 하며, eMCMC(evolutionary Markov chain Monte Carlo)[26]는 그러한 방법의 기반이 될 수 있을 것이다.

참 고 문 헌

[1] Schena, M. (ed.), *Microarray Biochip Technology*, Eaton Publishing, MA, 2000.

[2] Eisen, M.B., Spellman, P.T., Brown, P.O., and Botstein, D., Cluster analysis and display of genome-wide expression patterns, *Proceedings of the National Academy of Sciences of the United States of America*, vol. 95, no. 25, pp. 14863-14868, 1998.

[3] Raychaudhuri, S., Stuart, J.M., and Altman, R.B., Principal components analysis to summarize microarray experiments: application to sporulation time series, *Pacific Symposium on Biocomputing 5 (Proceedings of PSB'00)*, pp. 452-463, 1999.

[4] Khan, J., Wei, J.S., Ringnér, M., Saal, L.H., Ladanyi, M., Westermann, F., Berthold, F., Schwab, M., Antonescu, C.R., Peterson, C., and Meltzer, P.S., Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks, *Nature Medicine*, vol. 7, no. 6, pp. 673-679, 2001.

[5] Friedman, N., Linial, M., Nachman, I., and Pe'er, D., Using Bayesian networks to analyze expression data, In *Proceedings of the 4th Annual International Conference on Computational Molecular Biology (RECOMB'00)*, pp. 127-135, 2000.

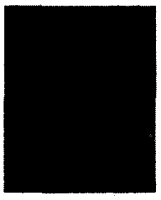
[6] Hartemink, A.J., Gifford, D.K., Jaakkola, T.S., and Young, R.A., Using graphical models and genomic expression data to statistically validate models of genetic regulatory networks, *Pacific Symposium on Biocomputing 6 (Proceedings of PSB'01)*, pp. 422-433, 2000.

[7] Hartemink, A.J., Gifford, D.K., Jaakkola, T.S., and Young, R.A., Combining location and expression data for principled discovery of genetic regulatory network models, *Pacific Symposium on Biocomputing 7 (Proceedings of PSB'02)*, pp. 437-449, 2001.

[8] Hwang, K.-B., Cho, D.-Y., Park, S.-W., Kim, S.-D., and Zhang, B.-T., Applying machine learning techniques to analysis of gene expression data: cancer diagnosis, Lin, S.M. and Johnson, K.F. (eds.), *Methods of Microarray Data Analysis (Proceedings of CAMDA'00)*, Kluwer Academic

Publishers, MA, pp. 167-182, 2002.
[9] Leping, L., Pedersen, L.G., Darden, T.A., and Weinberg, C.R., Computational analysis of leukemia microarray expression data using the GA/KNN method, Lin, S.M. and Johnson, K.F. (eds.), *Methods of Microarray Data Analysis (Proceedings of CAMDA'00)*, Kluwer Academic Publishers, MA, pp. 81-95, 2002.
[10] Spellman, P.T., Sherlock, G., Zhang, M.Q., Iyer, V.R., Anders, K., Eisen, M.B., Brown, P.O., Botstein, D., and Futcher, B., Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization, *Molecular Biology of the Cell*, vol. 9, no. 12, pp. 3273-3297, 1998.
[11] Scherf, U., Ross, D.T., Waltham, M., Smith, L.H., Lee, J.K., Tanabe, L., Kohn, K.W., Reinhold, W.C., Myers, T.G., Andrews, D.T., Scudiero, D.A., Eisen, M.B., Sausville, E.A., Pommier, Y., Botstein, D., Brown, P.O., and Weinstein, J.N., A gene expression database for the molecular pharmacology of cancer, *Nature Genetics*, vol. 24, no. 3, pp. 236-244, 2000.
[12] Jensen, F.V., *An Introduction to Bayesian Networks*, Springer-Verlag, NY, 1996.
[13] Heckerman, D., A tutorial on learning with Bayesian networks, Jordan, M.I. (ed.), *Learning in Graphical Models*, MIT Press, MA, pp. 301-354, 1999.
[14] Friedman, N. and Goldszmidt, M., Learning Bayesian networks with local structure, Jordan, M.I. (ed.), *Learning in Graphical Models*, MIT Press, MA, pp. 421-459, 1999.
[15] Heckerman, D., Geiger, D., and Chickering, D.M., Learning Bayesian networks: the combination of knowledge and statistical data, *Machine Learning*, vol. 20, no. 3, pp. 197-243, 1995.
[16] Chickering, D.M., Learning Bayesian networks is NP-complete, Fisher, D. and Lenz, H.-J. (eds.), *Learning from Data: Artificial Intelligence and Statistics V*, Springer-Verlag, NY, pp. 121-130, 1996.
[17] Cooper, G.F., Computational complexity of probabilistic inference using Bayesian belief networks, *Artificial Intelligence*, vol. 42, no. 2-3, pp. 393-405, 1990.
[18] Dagum, P. and Luby, M., Approximating probabilistic inference in Bayesian belief networks is NP-hard, *Artificial Intelligence*, vol. 60, no. 1, pp. 141-153, 1993.
[19] Pearl, J., *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*, Morgan Kaufmann Publishers, CA, 1988.
[20] Spirtes, P., Glymour, C., and Scheines, R., *Causation, Prediction, and Search*, 2nd edition, MIT Press, MA, 2000.
[21] Friedman, N., Nachman, I., and Pe'er, D., Learning Bayesian network structure from massive datasets: the "sparse candidate" algorithm, In *Proceedings of the 15th Conference on Uncertainty in Artificial Intelligence(UAI'99)*, pp. 206-215, 1999.
[22] Friedman, N., Goldszmidt, M., and Wyner, A., Data analysis with Bayesian networks: a bootstrap approach, In *Proceedings of the 15th Conference on Uncertainty in Artificial Intelligence(UAI'99)*, pp. 196-205, 1999.
[23] Hwang, K.-B., Lee, J.W., Chung, S.-W., and Zhang, B.-T., Construction of large-scale Bayesian networks by local to global search, *Lecture Notes in Artificial Intelligence (Proceedings of PRICAI'02)*, vol. 2417, pp. 375- 384, 2002.
[24] Graepel, T., Burger, M., and Obermayer, K., Self-organizing maps: generalizations and new optimization techniques, *Neurocomputing*, vol. 21, pp. 173-190, 1998.
[25] Dempster, A.P., Laird, N.M., and Rubin, D.B., Maximum likelihood from incomplete data via the EM algorithm(with discussion), *Journal of Royal Statistical Society B*, vol. 39, no. 1, pp. 1-38, 1977.
[26] Zhang. B.-T. and Cho, D-Y, System identification using evolutionary Markov chain Monte Carlo, *Journal of Systems Architecture*, vol. 47, no. 7, pp. 587-599, 2001.

황 규 백
1997년 서울대학교 공과대학 컴퓨터공학과 학사. 1999년 서울대학교 컴퓨터공학부 석사. 1999년 ~ 현재 서울대학교 컴퓨터공학부 박사과정. 관심분야는 기계학습, 확률그래프모델, 바이오인포매틱스, 기계번역



장 정 호
1995년 서울대학교 공과대학 컴퓨터공학과 학사. 1997년 서울대학교 공과대학 컴퓨터공학과 석사. 1997년 ~ 현재 서울대학교 컴퓨터공학부 박사과정. 관심분야는 기계학습, 은닉변수모델, 바이오인포매틱스, 텍스트마이닝

장 병 탁
정보과학회논문지 : 소프트웨어 및 응용 제 29 권 제 4 호 참조