앙상블 베이지안망에 의한 유전자발현데이터 분류

황규백 장정호 장병탁

서울대학교 컴퓨터공학부

{kbhwang, jhchang}@bi.snu.ac.kr btzhang@cse.snu.ac.kr

Classification of Gene Expression Data by Ensemble of Bayesian Networks

Kyu-Baek Hwang, Jeong-Ho Chang, and Byoung-Tak Zhang

School of Computer Science and Engineering, Seoul National University

요 약

DNA칩 기술로 얻어지는 유전자발현데이터(gene expression data)는 생체 조직이나 세포의 수천개에 달하는 유전자의 발현량(expression level)을 측정한 것으로, 유전자발현양상(gene expression pattern)에 기반한 암 종류의 분류 등에 유용하다. 본 논문에서는 확률그래프모델(probabilistic graphical model)의 하나인 베이지안망(Bayesian network)을 발현데이터의 분류에 적용하며, 분류 성능을 높이기 위해 베이지안망의 앙상블(ensemble of Bayesian networks)을 구성한다. 실험은 실제 암 조직에서 추출된 유전자발현데이터에 대해 행해졌다. 실험 결과, 앙상블 베이지안망의 분류 정확도는 단일 베이지안망보다 높았으며, naïve Bayes 분류기, 신경망, support vector machine (SVM) 등과 대등한 성능을 보였다.

1. 서론

DNA칩 기술로 얻어지는 유전자발현데이터(gene expression data)는 생체 조직이나 세포의 수천개의 유전자의 발현량(expression level)을 한번에 측정한 것이다. 한편, 유전자발현양상(gene expression pattern)은 많은 경우에 암의 종류를 구분할 수 있는 표지로 이용될 수 있음이 확인되었으며[4], 이는 기계학습에서 다뤄온 분류 문제와 일치하므로 신경망 등의 기법이 이러한 문제에 적용되어 왔다.

한편, 확률그래프모델(probabilistic graphical model)의 하나인 베이지안망(Bayesian network)[6]은 다수의 변수(random variable)들의 결합확률분포(joint probability distribution)를 변수들 사이의 조건부독립성(conditional independency)에 기반해 효율적으로 표현한다. 베이지안망이 표현하는 결합확률분포에서 조건부확률(conditional probability)이 계산될 수 있기 때문에, 베이지안망은 분류 문제에 바로 적용될 수 있다. 또한, 베이지안망의 구조는 각 입력자질(input feature) 및 class 변수 사이의 연관 관계를 나타내며 이는 데이터마이닝 및 지식 획득에 이용될 수 있다. 하지만, 베이지안망의 학습은 어려운 문제이며, 일반적인 베이지안망이 실제의 분류 문제에 적용되는 경우, 데이터의 부족 등 여러 원인으로 인해 그 학습 정확도가 다른 분류기에 비해 떨어지는 경우가 많이 있다 [2].

본 논문에서는 베이지안망 분류기의 분류 정확도를 높이기 위해 앙상블 베이지안망(ensemble of Bayesian networks)을 제안한다. 제안하는 앙상블 베이지안망은 선택적 Bayesian model averaging에 기반하고 있으며, 앙상블 머신의 원리에 기반해 이를 구축하는 간략한 알고리즘을 이용한다.

실험에서는 [4]의 백혈병에 대한 분류 데이터를 이용했으며, 앙상블 베이지안망의 분류 정확도가 단일 베이지안망을 이용하는 경우보다 높음을 보였다. 또한, 다른 분류기들과의 성능 비교를 통해 앙상블 베이지안망이 이들과 대등한 분류 정확도를 가짐을 보였다.

2. 베이지안망 (Bayesian Networks)

2.1 베이지안망의 정의

변수 집합 $\mathbf{X} = \{X_1, X_2, \ldots, X_n\}$에 대한 베이지안망은 망 구조 G와 각 변수들의 지역확률분포(local probability distribution)의 파라미터 집합 $\boldsymbol{\theta}$로 이루어진다. G는 DAG (directed acyclic graph) 형태로, 여기서 각 노드는 $\mathbf{X}$의 변수들과 일대일대응이 된다. G가 나타내는 조건부독립성(conditional independence assertions)에 의하면 $\mathbf{X}$의 결합확률분포는 아래와 같이 표현된다 [6].

$$P(\mathbf{X}) = \prod_{i=1}^{n} P(X_i \mid \mathbf{Pa}_G(X_i)) \qquad (1)$$

여기서 $\mathbf{Pa}_G(X_i)$는 망 구조 G에서 X_i의 부모노드의 집합을 나타낸다. 각 노드(변수)의 지역확률분포는 수식 (1)의 Π 안의 각 항에 해당한다.

2.2 베이지안망의 학습

베이지안망의 학습은 망 구조 G의 학습과 지역확률분포 파라미터 집합 $\boldsymbol{\theta}$의 학습의 두 단계로 이루어진다. Complete data를 포함한 몇가지 가정하에 $\boldsymbol{\theta}$는 maximum likelihood estimation 등의 방법으로 간단하게 학습될 수 있다 [6]. 망 구조의 학습은 보통 주어진 데이터 $D = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_M\}$에 가장 적합한 G의 탐색으로 이루어진다. 이 때, G의 D에 대한 적합도 $Score(G;D)$의 log 값은 다음과 같이 구성된다.

$$\begin{aligned} \log Score(G;D) &= penalizing\ term + log\ likelihood \\ &= penalizing\ term + \sum_{t=1}^{M} \log P_G(\mathbf{x}_t) \\ &= penalizing\ term \\ &\quad + \sum_{t=1}^{M} \sum_{i=1}^{n} \log P(X_i = x_{ti} \mid \mathbf{Pa}_G(X_i) = \mathbf{pa}_{ti}) \end{aligned} \qquad (2)$$

여기서 $P_G(\mathbf{x}_t)$는 G에 의한 $\mathbf{x}_t$의 결합확률이고, x_{ti}, $\mathbf{pa}_{ti}$는 각각 $\mathbf{x}_t$에서의 X_i, $\mathbf{Pa}_G(X_i)$의 값이다. *penalizing term*은 망 구조의 복잡도에 대한 것으로, 베이지안 학습의 경우 G에 대한 사전확률(prior probability) $P(G)$와 연관지어 생각할

수 있다.

2.3 베이지안망 분류기

일반적인 베이지안망을 분류 문제에 적용하는 경우 신경망과 같은 분류기에 비해 성능이 떨어지는 경우가 많다 [2]. 이는 원칙적으로는 학습에 이용되는 데이터가 유한하기 때문이다. 베이지안망 학습 알고리즘들은 점근적으로 정확한 결과를 내기 때문에, 유한한 데이터에서 학습된 베이지안망은 가장 정확한 결과가 아닐 가능성이 크다. 이렇게 학습된 베이지안망의 분류 정확도가 낮을 수 있는 것은 다음과 같은 이유에서이다. 분류 문제의 경우, 베이지안망을 구성하는 n개의 변수를 class 변수 C와 입력 자질 A_1, A_2, ..., A_{n-1}로 생각할 수 있다. 이 경우, 수식 (2)의 *log likelihood*를 다음과 같이 쓸 수 있다.

$$\begin{aligned} log\ likelihood &= \sum_{t=1}^{M} \log P_G(A_1 = a_{t1}, A_2 = a_{t2}, ..., A_{M-1} = a_{(M-1)t}) \\ &+ \sum_{t=1}^{M} \log P_G(C = c_t \mid A_1 = a_{t1}, A_2 = a_{t2}, ..., A_{M-1} = a_{(M-1)t}) \end{aligned} \quad (3)$$

수식 (3)에서 두번째 term만 분류 성능과 연관이 있기 때문에 수식 (2)를 최대화하는 것은 분류 성능을 낮추는 결과를 가져올 수도 있다 [2].

이와 같은 문제에 대한 손쉬운 해결책은 베이지안망의 구조를 분류에 적합하게 고정하는 것이다. naïve Bayes 분류기는 그러한 대표적인 예이며, 이 구조가 표현하는 조건부독립성이 실제와 다른 경우에도 분류 정확도는 높을 수 있음이 알려져 있다 [1]. 하지만 이렇게 구조에 제한을 두는 것은 베이지안망 학습을 데이터마이닝 및 지식 획득에 이용할 수 없게 한다. 구체적으로, 학습을 통해 입력 자질 사이의 관계를 밝힐 수 없다. 본 논문에서는 구조에 제한을 두지 않으면서 분류 정확도를 높이기 위한 방안을 다룬다.

3. 앙상블 베이지안망 분류기

3.1 베이지안망의 베이지안 학습 (Bayesian learning for Bayesian networks)

베이지안 학습(Bayesian learning)은 학습의 대상(예: 파라미터)에 대한 사전확률(prior probability)과 데이터가 주어진 경우의 사후확률(posterior probability)을 설정하고, 이들에 기반해서 학습 및 예측을 하는 방법이다. 특히, 하나의 학습결과를 선택하는 maximum likelihood (ML)나 maximum a posteriori (MAP) 기법과는 달리 가능한 모든 결과를 사후확률 분포에 기반해서 이용함으로써 over-fitting을 막고 일반화 성능을 높일 수 있는 기법이다.

베이지안망의 파라미터 $\boldsymbol{\theta}$ 에 대한 베이지안 학습은 다음과 같이 구현된다.

$$\begin{aligned} P_G(C \mid A_1, ...A_{n-1}) &= \alpha \cdot P_G(C, A_1, ...A_{n-1}) \\ &= \alpha \cdot \int \{P(C \mid \mathbf{Pa}_G(C)) \cdot \prod_{i=1}^{n-1} P_B(A_i \mid \mathbf{Pa}_G(A_i))\} P(\boldsymbol{\Theta} \mid D) d\boldsymbol{\Theta} \end{aligned} \quad (4)$$

여기서 a 는 normalization constant이다. 또한, 망 구조 G에 대한 베이지안 학습은 다음과 같이 구현된다.

$$\begin{aligned} P_G(C \mid A_1, ...A_{n-1}) &= \alpha \cdot P_G(C, A_1, ...A_{n-1}) \\ &= \alpha \cdot \sum_{G \in \mathbf{g}} P_G(C, A_1, ...A_{n-1}) \cdot P(G \mid D) \\ &= \frac{\alpha}{P(D)} \cdot \sum_{G \in \mathbf{g}} P_G(C, A_1, ...A_{n-1}) \cdot P(G) \cdot P(D \mid G) \end{aligned} \quad (5)$$

이 식에서 $\mathbf{g}$는 가능한 모든 G의 집합이다. 그러나 수식 (5)의 계산은 가능한 망 구조의 개수가 $O(2^{n^2})$에 해당하기 때문에 n이 6을 넘어가는 경우 거의 불가능하다. 본 논문에서는 G의 부분집합의 일부에 대해서만 수식 (5)의 계산을 행하는 선택적 베이지안 학습을 앙상블 베이지안망으로 구현한다.

3.2 앙상블 베이지안망

앙상블 베이지안망은 수식 (5)를 다음과 같이 근사한 것이다.

$$\begin{aligned} P_G(C \mid A_1, ...A_{n-1}) &= \alpha \cdot P_G(C, A_1, ...A_{n-1}) \\ &\approx \alpha \cdot \sum_{G \in \mathbf{g}_{high}} P_G(C, A_1, ...A_{n-1}) \cdot P(G \mid D) \\ &= \frac{\alpha}{P(D)} \cdot \sum_{G \in \mathbf{g}_{high}} P_G(C, A_1, ...A_{n-1}) \cdot Score(G; D) \end{aligned} \quad (6)$$

여기서 $\mathbf{g}_{high}$는 높은 $P(G|D)$값을 가지는 $\mathbf{g}$의 부분집합이다. 수식 (5)를 잘 근사하기 위해서는 $\mathbf{g}_{high}$를 잘 찾아야 한다. 하지만 이는 베이지안망 구조 공간이 거대하며 multimodality가 높기 때문에 쉽지 않다. 본 논문에서는 앙상블 머신의 관점에서 이 문제를 접근한다.

앙상블 머신을 만드는 경우, 앙상블의 각 구성원이 서로 다를 때에 좋은 일반화 성능을 얻을 수 있다 [5]. 본 논문에서는 이를 위해 입력 자질 공간을 분할한다. 그리고 각 멤버 베이지안망은 각 분할된 공간의 자질에 기반해서 분류를 행하게 된다. 부분 집합의 크기는 앙상블 머신의 구성원의 개수에 따라 결정된다. 즉, K개의 구성원을 가지는 앙상블 베이지안망의 경우, 하나의 베이지안망이 $(n-1)/K$ 개의 입력 자질을 담당하게 된다. 선택된 자질들은 class 변수의 자식으로 고정되며, 베이지안망 구조 학습 알고리즘은 이 그래프를 부분 그래프로 포함하며 학습되도록 설정된다. 알고리즘의 개요는 아래과 같다.

For $k = 1, 2, ..., K$
- $n-1$ 개의 자질 중에서, $(n-1)/K$ 개의 자질을 임의로 선택한다. (이전의 iteration 에서 선택된 자질들은 선택되지 않는다.)
- 선택된 자질들을 class 변수의 자식으로 고정시킨다.
- 이 그래프를 포함하는 베이지안망을 greedy search algorithm 으로 학습한다.

4. 실험

실험은 [4]의 백혈병 데이터를 이용했다. 이 데이터는 백혈병 환자에서 추출된 조직에 대한 72개의 마이크로어레이로 구성되어 있다. 이 중 47개는 acute lymphoblastic leukemia (ALL)이며 25개는 acute myeloid leukemia (AML)로, 이진 분류 문제이다. 각 마이크로어레이는 7,129개의 유전자에 대한 발현량을 측정한 것이다. 실험은 [4]의 방법에 따라 선택된 암 종류와 관계가 깊은 50개의 유전자에 대해서 행해졌다. 또한 각 발현값은 베이지안망의 지역 확률분포모델로 다항 분포를 이용하기 위해 평균값을 기준으로 0과 1로 이산화되었다. 베이지안망의 점수 *Score*(*G*;*D*)로는 BD (Bayesian Dirichlet) 점수[3]가 이용되었다. 앙상블 머신의 크기 K는 5, 7, 10, 15, 20에 대해 실험을 했다.

4.1 실험 결과

실험 데이터의 크기가 작기 때문에 분류 정확도 평가는 leave-one-out cross validation (LOOCV)을 이용했다. 아래의

그래프는 앙상블의 크기에 따른 성능의 변화이다.

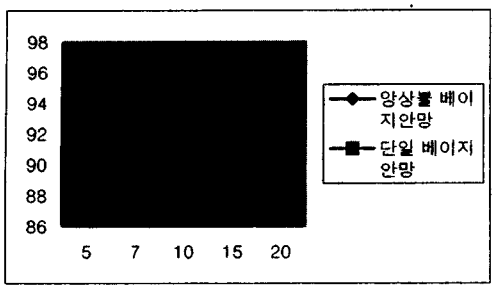

그림 1. 앙상블 머신의 크기 K 에 따른 분류 성능 변화 (LOOCV, %)

그림 1에서 단일 베이지안망은 앙상블 머신의 구성원 중 가장 점수가 좋은 것만을 가지고 분류한 경우이다. K가 5, 7, 10인 경우는 앙상블 머신의 성능이 더 뛰어나며, 전체적으로 분류 성능이 가장 좋은 경우는 7개의 구성원을 가지는 앙상블 베이지안망의 97.22%이다. 앙상블 머신의 크기가 큰 경우(K = 15, 20)에는 단일 베이지안망의 성능이 더 좋다. 원인 분석을 위해 각 경우의 class 변수의 Markov blanket[7]의 크기를 구해 보았다. Class 변수의 Markov blanket은 입력 자질들 중 분류에 직접 영향을 주는 변수들의 집합이다.

표 1. K 값에 따른 앙상블 머신(Total #)과 개개의 베이지안망 (Average #)에서의 class 변수의 Markov blanket 의 크기.

앙상블 머신의 크기	5	7	10	15	20
Total #	35.00	37.24	34.50	37.51	36.72
Average #	14.14	12.53	9.85	8.99	8.11

위의 표를 보면 K값에 관계없이 Markov blanket에 포함되는 자질의 수는 거의 36 정도로 항상 일정함을 알 수 있다. 하지만 각 멤버 베이지안망에 포함되는 자질의 개수는 앙상블 베이지안망의 크기 K가 커질수록 작아짐을 알 수 있다. 이는 각 구성원의 분류 성능이 뛰어나지 못할 수 있을 가능성을 크게 만들며, 이러한 이유로 그림 1의 결과가 나왔다고 생각할 수 있다.

4.2 다른 분류기들과의 비교

제시된 방법의 성능을 다른 분류기들과 비교했다. 비교의 대상은 [4]의 weighted voting scheme, c4.5 결정트리, naïve Bayes 분류기, 신경망, support vector machine (SVM)이다. weighted voting은 실제 발현값을 입력으로 사용했으며, 결정트리는 이산값을 사용했다. 다른 기법들에는 두가지 경우를 다 실험했다. 표 2는 정확도 비교 결과이다. 비교 결과를 보면 앙상블 베이지안망의 성능이 다른 방법들에 비해 떨어지지 않음을 알 수 있다. 우선 가장 성능이 좋은 경우는 SVM에 원래 자질값을 사용한 경우의 98.61%이다. 그러나 이산화된 자질값만을 이용하는 경우에는 가장 높은 정확도인 97.22%의 성능을 SVM과 앙상블 베이지안망이 같이 보이고 있다는 것을 알 수 있다.

표 2. 앙상블 베이지안망과 다른 분류 기법과의 정확도 비교

분류기		분류정확도 (LOOCV)
Weighted voting		95.83%
결정트리 (c4.5)		95.83%
naïve Bayes 분류기	이산값	95.83%
	실수값	97.22%
신경망	이산값	95.83%
	실수값	97.22%
Support vector machine	이산값	97.22%
	실수값	**98.61%**
베이지안망	단일 베이지안망	95.83%
	앙상블 베이지안망	**97.22%**

5. 결론

본 논문에서는 베이지안망의 구조에 대한 제약을 이용하지 않고 분류 정확도를 높이기 위한 방법으로 앙상블 베이지안망을 제안했다. 이는 베이지안 학습을 간략하게 사용한 것으로 볼 수 있으며, 단일 베이지안망만을 이용하는 경우보다 높은 정확도를 낼 수 있음을 보였다. 또한, 다른 분류기법들과의 비교를 통해 이들과 대등한 성능을 앙상블 베이지안망이 낼 수 있음을 보였다. 향후 연구 방향은 다음과 같다. 우선, 베이지안 학습을 MCMC(Markov chain Monte Carlo) 등의 기법을 이용해서 더 정확하게 근사하는 기법을 생각할 수 있다. 이러한 기법의 적용은 분류 성능을 더 높일 수 있을 것으로 기대된다.

감사의 글

이 논문은 교육인적자원부의 BK21 사업과 과학기술부의 IMT-2000, NRL 및 BrainTech 사업에 의하여 지원되었음.

참고 문헌

[1] Domingos, P. and Pazzani, M., On the optimality of the simple Bayesian classifier under zero-one loss, *Machine Learning*, vol. 29, no. 2, pp. 103-130, 1997.

[2] Friedman, N., Geiger, D., and Goldszmidt, M., Bayesian network classifiers, *Machine Learning*, vol. 29, no. 2, pp. 131-163, 1997.

[3] Friedman, N. and Goldszmidt, M., Learning Bayesian networks with local structure, *Learning in Graphical Models*, Jordan, M.I., (ed.), pp. 421-459, Cambridge, MA: MIT Press, 1999.

[4] Golub, T.R., Slonim, D.K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J.P., Coller, H., Loh, M.L., Downing, J.R., Caligiuri, M.A., Bloomfield, C.D., and Lander, E.S., Molecular classification of cancer: class discovery and class prediction by gene expression monitoring, *Science*, vol. 286, no. 5439, pp. 531-537, 1999.

[5] Haykin, S., *Neural Networks: A Comprehensive Foundation*, Upper Saddle River, NJ: Prentice-Hall, 1999.

[6] Heckerman, D., A tutorial on learning with Bayesian networks, *Learning in Graphical Models*, Jordan, M.I., (ed.), pp. 301-354, Cambridge, MA: MIT Press, 1999.

[7] Pearl, J., *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*, San Francisco, CA: Morgan Kaufmann Publishers, 1988.