# 텍스트 문서 분류를 위한 베이지안망 학습

황규백 장병탁 김영택
서 울 대 학 교 컴 퓨 터 공 학 부
kbhwang@scai.snu.ac.kr {btzhang,ytkim}@cse.snu.ac.kr

# Learning Bayesian Networks for Text Documents Classification

Kyubaek Hwang Byoung-Tak Zhang Yung Taek Kim
School of Computer Science and Engineering, Seoul National University

요 약

텍스트 문서 분류는 텍스트 형태로 주어진 문서를 종류별로 구분하는 작업으로 웹페이지 검색, 뉴스그룹 검색, 메일 필터링 등의 분야에 응용될 수 있는 기반 작업이다. 지금까지 문서를 분류하는데는 k-NN, 신경망 등 여러 가지 기계학습 기법이 이용되어 왔다. 이 논문에서는 베이지안망을 이용해서 텍스트 문서 분류를 행한다. 베이지안망은 다수의 변수들간의 확률적 관계를 표현하는 그래프 모델로 DAG 형태인 망 구조와 각 노드에 연관된 지역확률분포로 구성된다. 그래프 모델을 사용할 경우 학습에 이용되는 각 속성들간의 관계를 사람이 알아보기 쉬운 형태로 학습할 수 있다는 장점이 있다. 실험 데이터로는 Reuters-21578 문서분류데이터를 이용했으며 베이지안망의 성능은 나이브 베이즈 분류기와 비슷했다.

## 1. 서론

베이지안망(Bayesian network)은 다수의 변수들간의 확률적 관계를 표현하는 그래프 모델이다. 베이지안망의 장점은 다음과 같다[2]. 첫째, 모든 변수들간의 의존관계(dependency)를 표현하기 때문에 불완전한 데이터를 가지는 상황에 적절히 대처할 수 있다. 둘째, 변수들간의 인과관계(causal relationship)를 학습하는데 사용될 수 있기 때문에 응용분야에 대한 이해를 도울 수 있다. 셋째, 모델 자체가 원인(causality)과 확률적의미(probabilistic semantics)를 표현하고 있기 때문에 사전지식(prior knowledge)과 학습 데이터를 결합하는데 적합하다. 넷째, 베이지안망에 베이즈 통계기법을 적용함으로써 데이터 과대적합(data overfitting)을 막을 수 있다.

망의 구조는 변수들간의 관계를 표현하며 이는 해당 분야의 전문가가 구성할 수 있다. 정해진 망 구조와 완전 데이타(complete data)를 가지고 있으면 베이지안망의 지역확률분포의 갱신은 닫힌 형식으로 간단히 계산된다. 따라서 전문가가 원하는 변수들간의 관계를 정확하게 기술할 수 있으면 학습된 베이지안망은 원하는 변수들간의 결합확률분포를 완벽하게 표현한 것이 된다.

본 논문에서는 미리 주어진 망 구조를 가지고 각 노드들의 지역확률분포를 학습했을 때 베이지안망이 보이는 학습행태와 성능을 평가하고자 한다. 평가를 위한 척도로는 나이브 베이즈 분류기를 사용한다. 나이브 베이즈 분류기가 표현하는 변수들간의 독립가정을 이용했을 때 베이지안망의 성능은 나이브 베이즈 분류기와 비슷함 확인되었다.

## 2. 베이지안망(Bayesian network)

베이지안망은 변수에 해당하는 노드와 그 노드(변수)들간의 인과관계를 나타내는 간선들로 구성된 DAG 이며 변수들간의 결합확률분포(joint probability distribution)를 효율적으로 표현할 수 있는 그래프 모델이다. 변수집합 $\mathbf{X} = \{X_1, \ldots, X_n\}$에 대한 베이지안망은 다음의 2 가지 부분으로 구성된다.

(1) 집합 $\mathbf{X}$ 의 변수들간의 조건부독립성(conditional independence assertion)을 표현하고 있는 망 구조 $S$

(2) 각 변수들과 연관되어 있는 지역확률분포(local probability distribution) 집합 $P$

망의 구조 $S$ 는 DAG 이다. $S$ 의 각 노드는 $\mathbf{X}$ 의 변수들과 일대일대응이 된다. $X_i$ 는 변수와 그 변수에 해당하는 노드를 동시에 가리킨다. $\mathbf{Pa}_i$ 는 그래프 $S$ 에서 $X_i$의 부모노드(변수)의 집합을 나타낸다. 부모노드는 0 개 이상이다. $S$ 에서 간선으로 연결되지 않은 노드들은 서로 조건부독립관계에 있다. 망 구조가 나타내는 조건부독립성에 의하면 주어진 구조 $S$ 에 대한 $\mathbf{X}$ 의 결합확률분포는 다음과 같이 주어진다.

$$p(\mathbf{x}) = \prod_{i=1}^{n} p(x_i \mid \mathbf{pa}_i)$$

지역확률분포 $P$ 는 위 수식의 $\Pi$ 안의 각 항에 대응된다. 따라서 확률분포 $p(\mathbf{x})$를 나타내기 위해서는 $(S, P)$가 필요하다.

### 2.1 베이지안망에서의 확률의 학습

망 구조가 확정되면 주어진 망 구조가 변수집합 $\mathbf{X}$ 의 실제 확률분포를 표현한다고 생각한다. 그러면 다음과 같은 식을 쓸 수 있다.

$$p(\mathbf{x} \mid \boldsymbol{\theta}_s, S^h) = \prod_{i=1}^{n} p(x_i \mid \mathbf{pa}_i, \boldsymbol{\theta}_i, S^h)$$

위의 식에서 $\boldsymbol{\theta}_i$ 는 하나의 노드 $X_i$ 에 대한 확률분포 $p(x_i \mid \mathbf{pa}_i, \boldsymbol{\theta}_i, S^h)$의 파라미터이며 $S^h$ 는 망 구조에 대한 가설이다. 베이지안망에서 확률의 학습은 결국 데이터를 이용해서 각 파라미터의 사후분포를 계산하는 것이다. 각

노드의 파라미터의 분포가 지수족(exponential family)이며 파라미터독립성(parameter independence)[8]을 가정한 경우 각 지역확률분포는 닫힌 형식으로 계산된다[2].

예를 들어, 각 노드가 모두 자유다항분포(unrestricted multinomial distribution)이며 노드 $X_i$ 는 $r_i$ 개의 값을 가진다고 하자. 그러면 각 노드는 파라미터벡터 $\boldsymbol{\theta}_i = (\boldsymbol{\theta}_{i1}, \boldsymbol{\theta}_{i2}, \ldots, \boldsymbol{\theta}_{iqi})$, $\boldsymbol{\theta}_{ij} = (\theta_{ij2}, \theta_{ij3}, \ldots, \theta_{ijri})$를 가진다(단, $q_i = \Pi_{Xi\in \boldsymbol{Pa}i} r_i$, $\theta_{ij1} = 1 - (\theta_{ij2} + \theta_{ij3} + \ldots + \theta_{ijri})$). 그러면 각 파라미터 $\theta_{ijk}$ 는 모두 디리슈레분포(Dirichlet distribution)를 따르며 각 파라미터의 초기분포는 다음과 같다.

$$p(\boldsymbol{\theta}_{ij} \mid S^h) = \mathrm{Dir}(\boldsymbol{\theta}_{ij} \mid \alpha_{ij1}, \ldots, \alpha_{ijr_i})$$

파라미터독립성에 의하면 완전데이터 $D$ 가 주어졌을 때 파라미터 $\boldsymbol{\theta}_{ij}$ 는 독립적으로 학습될 수 있고 다음과 같이 갱신된다.

$$p(\boldsymbol{\theta}_{ij} \mid D, S^h) = \mathrm{Dir}(\boldsymbol{\theta}_{ij} \mid \alpha_{ij1} + N_{ij1}, \ldots, \alpha_{ijr_i} + N_{ijr_i})$$

여기서 $N_{ijk}(k = 1, \ldots, r_i)$는 데이터 $D$ 에서 노드 $X_i$ 가 $\mathbf{pa}_{ij}$ 하에서 $k$ 번째 값을 가진 경우의 횟수이다.

### 2.2 베이지안망에서의 추론

베이지안망이 구성된 다음에는 주어진 조건에 대한 특정 변수의 확률분포를 알 수 있다. 이 확률은 모델에 직접 나타나 있는 것이 아니기 때문에 계산(추론)되어야 한다. 베이지안망은 모든 변수들간의 결합확률분포를 표현하기 때문에 원칙적으로는 주어진 모든 사례의 확률을 계산할 수가 있다. 노드의 파라미터 $\boldsymbol{\theta}_i$ 가 확률분포를 가지기 때문에 $N + 1$ 번째에 임의의 사례 $\mathbf{x}_{N+1}$ 이 나올 확률은 다음과 같이 파라미터에 대한 기대값을 구함으로써 얻어진다.

$$p(\mathbf{x}_{N+1} \mid D, S^h) = \int \prod_{i=1}^{n} \theta_{ijk}\, p(\boldsymbol{\theta}_s \mid D, S^h)\, d\boldsymbol{\theta}_s$$

$$= \prod_{i=1}^{n} \int \theta_{ijk}\, p(\boldsymbol{\theta}_{ij} \mid D, S^h)\, d\boldsymbol{\theta}_{ij}$$

$\boldsymbol{\theta}_{ij}$ 는 디리슈레분포를 따르기 때문에 위의 식은 다음과 같이 계산된다.

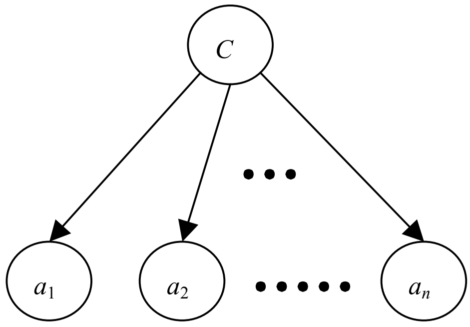
$$p(\mathbf{x}_{N+1} \mid D, S^h) = \prod_{i=1}^{n} \frac{\alpha_{ijk} + N_{ijk}}{\alpha_{ij} + N_{ij}}$$

$$\alpha_{ij} = \sum_{k=1}^{r_i} \alpha_{ijk}, \quad N_{ij} = \sum_{k=1}^{r_i} N_{ijk}$$

## 3. 베이지안망을 이용한 분류기

베이지안망을 이용한 분류기는 다음과 같이 구성된다. 우선 분류에 필요한 각 속성(attribute)들이 각각 하나씩의 노드를 구성한다. 그리고 클래스를 나타내는 노드가 하나 필요하다. 그리고 각 노드들간의 조건부독립성을 표현하는 그래프의 구조가 필요하다. 가장 간단한 가정은 클래스의 값이 주어진 경우, 모든 속성들은 서로 독립이라는 가정이다. 예를 들어 $n$ 개의 속성을 가지는 문제의 경우 그림 1 과 같은 베이지안망을 구성할 수 있다.

그림 1 의 베이지안망에서 노드 $C$ 는 클래스를 나타내며 분류기가 분류하는 클래스의 개수에 해당하는 값들을 가질 수 있는 이산노드이다. 노드 $a_1, \ldots, a_n$ 은 각 속성에 해당하는 노드들로 역시 이산노드이다.




**[그림 1]** $n$ 개의 속성을 가지는 분류 문제에 해당하는 베이지안망

### 3.1 베이지안망 분류기의 학습

베이지안망 분류기의 학습은 다음과 같은 알고리즘으로 이루어진다.

**[학습알고리즘]**

```
while there is a case l in dataset D, do
    i ← the number of nodes in the graph
    j ← the number of configurations of Pa_i
    k ← the number of configurations of X_i
    for all i, j, k, do
        α_ijk ← some value
        N_ijk ← 0
    endfor
    for all the nodes in the graph S, do
        pick a node X_i
        for all the configurations of Pa_i, do
            pick a configuration pa_ij
            for all the values of X_i, do
                pick a value of x_ijk
                update N_ijk
            endfor
        endfor
    endfor
endwhile
```

데이터의 각 사례(case)를 가지고 $N_{ijk}$ 의 값을 갱신해 나가는 것이 학습의 과정이다. 위의 알고리즘에서 $i$ 는 그래프의 노드의 개수이며, $j$ 는 노드 $X_i$ 의 부모 $\mathbf{Pa}_i$가 가지는 상태의 개수, $k$ 는 $X_i$가 가지는 상태의 개수이다. $\alpha_{ijk}$ 는 사전 디리슈레분포(prior Dirichlet distribution)의 파라미터로 여러가지 방법을 이용해서 정해줄 수 있다.

## 4. 실험

### 4.1 데이터

실험 데이터는 문서분류에 이용되는 Reuters-21578 데이터를 이용하였다. 실험에 이용한 데이터의 특성은 표 1 에 나와 있다. Reuters-21578 데이터에서 acq 클래스에 대한 분류데이터를 사용했으며 학습데이터는 총 8762 개이다. 이중 정례(positive example)는 1483 개, 반례(negative example)는 7279 개이다. 또한 분류에 이용된 속성의 개수는 8754 개이며 이는 학습 데이터에 나타난 모든 단어의 *tfidf* 값을 계산해서 선정되었다. 분류에 이용된 모든

속성은 이진값을 가지며 클래스 노드는 해당 클래스에 속하는가 아닌가를 나타내는 이진값을 가진다.

| 클래스 | 속성의 개수 | 학습 데이터 | 테스트 데이터 |
|---|---|---|---|
| acq | 8754 | 8762 | 3009 |

**[표 1]** 실험에 이용된 Reuters-21578 데이터의 특성

### 4.2 실험 결과

Reuters-21578 데이터에서 acq 클래스에 대한 이진분류를 했으며 결과는 표 2 와 같다. 이를 보면 정례를 구분하는 성능이 반례를 구분하는 성능보다 떨어짐을 알 수 있다. 이는 학습데이터에서의 정례의 개수가 1483 개로 반례의 개수 7279 개보다 훨씬 적기 때문으로 보인다.

| | | 학습 데이터(%) | 테스트 데이터(%) |
|---|---|---|---|
| 정례 (positive example) | recall | 98.11 | 77.81 |
| | precision | 78.69 | 56.08 |
| 반례 (negative example) | recall | 94.59 | 83.54 |
| | precision | 99.59 | 93.31 |

**[표 2]** 베이지안망 분류기의 성능

표 3 에서는 나이브 베이즈 분류기와의 비교를 보이고 있다. 표 3 을 보면 나이브 베이즈 분류기와 베이지안망 분류기의 성능이 동일함을 알 수 있다 이는 실험에 이용된 베이지안망이 표현하는 변수들간의 조건부독립성이 나이브 베이즈 분류기에서의 가정과 동일하고 사전 디리슈레분포의 파라미터값을 작게 주었기 때문이다. 학습데이터의 분량에 비해 사전분포의 파라미터값이 작았기 때문에 사전분포가 사후분포에 미치는 영향이 적었고 결과적으로 나이브 베이즈 분류기와 같은 값을 학습하게 된 것이다.

| 분류기 | 학습 정확도 (%) | 테스트 정확도 (%) |
|---|---|---|
| 베이지안망 분류기 | 95.18 | 82.32 |
| 나이브 베이즈 분류기 | 95.18 | 82.32 |

**[표 3]** Reuters-21578 acq 에 대한 분류 성능

### 5. 결론

본 논문에서는 베이지안망 분류기(Bayesian network classifier)를 구성 및 학습하고 이를 나이브 베이즈 분류기(naïve Bayes classifier)와 비교해 보았다. 비교 결과는 비슷한 성능을 내는 것으로 나타났다. 이는 본 논문에서 사용한 베이지안망이 나이브 베이즈 분류기와 동일한 조건부독립성을 표현하고 있고 사전분포가 학습에 큰 영향을 주지 못했기 때문이다.

사전분포의 영향을 크게해서 학습을 하면 나이브 베이즈 분류기와는 다른 성능을 보일 것이다. 사전분포는 사람이 줄 수 있는 것으로 사람이 가지고 있는 사전지식이라고 할 수 있다. 이를 이용해서 학습을 하는 것은 결국 사전지식과 데이터의 결합으로 이는 그래프 모델이 가지고 있는 장점 중 하나이다.

그래프 모델의 또 다른 장점은 학습 결과를 사람이 이해하는 것이 쉽다는 점이다. 즉 본 논문에서 사용된 베이지안망의 경우 주어진 클래스가 각 속성에 어떤 영향을 주는지 쉽게 알 수 있다.

한편, 베이지안망은 같은 데이터에 대해서 헬름홀츠 머신(Helmholtz machine)이나 시그모이드 빌리프망(sigmoid belief network)과의 성능비교에서는 조금 떨어지는 것으로 나타났다. 이는 망의 구조가 변수들간의 실제 관계(true relationship)를 표현하고 있지 못하기 때문이다. 따라서 데이터를 가지고 자동적으로 망의 구조를 학습할 수 있으면 더욱 좋은 성능을 낼 수 있을 것이며 이에 대한 연구가 필요하다.

### 감사의 글

본 연구는 과학기술부 뇌연구개발사업(BR-2-1-G-06)에 의하여 일부 지원되었음.

### 6. 참고 문헌

[1] Charniak, E., Bayesian Networks without Tears, *AI Magazine*, Vol. 12, No. 4, pp.50-63, 1991.

[2] Heckerman, D., A Tutorial on Learning with Bayesian Networks, Technical Report MSR-TR-95-06, Microsoft Research, Redmond, WA, 1995.

[3] Jensen, F., Lauritzen, S. and Olsen, K., Bayesian updating in recursive graphical models by local computations, *Computational Statistics Quarterly*, Vol. 4, pp.269-282, 1990.

[4] Jensen, F., *An Introduction to Bayesian Networks*, Springer, 1996.

[5] Lauritzen, S. and Spiegelhalter, D., Local computation with probabilities on graphical structures and their application to expert systems, *J. Royal Statistical Society B*, Vol. 50, pp.157-224, 1988.

[6] Pearl, J., *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*, San Mateo, CA: Morgan Kaufmann, 1987.

[7] Singh, M. and Provan, G., Efficient Learning of selective Bayesian network classifier, Technical Report MS-CIS-95-36, Computer and Information Science Department, University of Pennsylvania, Philadelphia, PA, 1995.

[8] Spiegelhalter, D. and Lauritzen, S., Sequential updating of conditional probabilities on directed graphical structure, *Networks*, Vol. 20, pp.579-605, 1990.