Helmholtz Machine 학습에 기반한 문서 분류

장정호 [0]　　장병탁　　김영택

서 울 대 학 교　컴 퓨 터 공 학 부

jhchang@nova.snu.ac.kr　{btzhang,ytkim}@cse.snu.ac.kr

Text Categorization Using a Helmholtz Machine

Jeong-Ho Chang [0]　Byoung-Tak Zhang　Yung Taek Kim

School of Computer Science and Engineering, Seoul National University

요 약

이 논문에서는 Helmholtz machine 을 사용하여 데이터의 분포 추정을 함으로써 문서 분류기를 학습하는 방법 제안한다. Helmholtz machine 은 생성 모델과 인식 모델로 구성된 그래프 모델로서, 그래프 모델에서의 분포 추정을 보다 가능하게 하기 위한 근사 방법 중의 하나이다. Helmholtz machine 에서의 각 입력 노드는 문서를 구성하는 하나의 단어에 대응하는 이진 노드이다. 입력 노드의 개수가 많아지면 그만큼 학습 시간이 증가하기 때문에, 학습 시간을 줄이면서 적정 수준의 성능을 유지하기 위해 자질 선정이 필요하다. 이러한 요구 사항을 충족시키기 위해 정보획득량(information gain) 기준을 이용하였으며, 뉴스 그룹 데이터에 대해 그 성능을 측정하고 Naïve Bayes 를 이용한 것과 비교한다.

1. 서론

텍스트 문서 분류(text categorization)는 임의의 텍스트 문서를 이미 정해진 범주에 따라 분류하는 문제이다. 인터넷의 발전과 전산 기술의 발달에 따라 전산화된 문서의 양이 점점 더 증가하고 있고, 이에 따른 정보의 분류 문제 역시 중요한 문제로 제기되고 있다. 지금까지 이러한 목적을 위해 k-최근접 방법, Naïve Bayes, 신경망, support vector machine, PCA 등의 여러 분류 방법들이 제안되었으며 실제로 좋은 성능을 보이고 있다. 본 논문에서는 데이터에 대한 분포 밀도 추정을 통해 문서를 분류하는 실험을 하고 그 성능을 평가하고, 기존의 분류 모델에 비해 어떤 실험적 특성을 지니는지 살펴 보고자 한다.

지난 몇 년간 그래프 모델은 표현력과 계산 가능성 면에서 많은 발전을 하였으며 이에 따라 확률적 결정 모델링에 대한 많은 관심과 연구가 진행되어 왔다. 일반적으로 그래프 모델이 복잡할 경우, 모델 내에서의 정확한 확률적 추론은 너무나 많은 계산을 필요로 한다. 따라서 이에 대한 근사 방법에 대한 연구가 진행되었으며 MCMC (Markov Chain Monte Carlo), Variational inference 등이 제안되었다. 본 논문에서 이용하는 Helmholtz machine 역시 그래프 모델에서의 근사 추론을 위해 고안된 방법 중의 하나이다[3].

본 논문에서는 UseNet 뉴스 그룹에 대해 Helmholtz machine 을 이용하여 각 문서 범주에 대한 분포 밀도 추정을 하고 새로운 데이터를 분류할 때 각 범주에 해당하는 Helmholtz machine 으로부터 로그 유사도를 측정하여 문서를 분류하는 실험을 하고 그 결과를 Naïve Bayes 방법과 비교한다.

제 2 장, 3 장에서는 각각 Helmholtz machine 과 학습 알고리즘인 wake-sleep 알고리즘에 대해 설명하고, 제 4 장에서는 뉴스 그룹 데이터에 대한 실험 결과를 제시한다. 마지막으로 5 장에서는 결론과 앞으로의 향후 연구 방향에 대해 서술한다.

2. Helmholtz Machine

Helmholtz machine[2]은 베이지안 네트워크에서의 학습과 추론 과정을 보다 용이하게 하기 위한 근사적 방법의 일종이다. Helmholtz machine 은 하나의 생성 네트워크(generative model)와 인식 네트워크(recognition network)의 쌍으로 구성된다. 인식 모델은 하나의 데이터 또는 패턴이 주어질 때, 그 데이터에 내재된 특성들의 확률 분포를 추정하는데 이용되며, 생성 모델은 그 내부적 표현으로부터 입력 데이터를 추정함으로써 이러한 인식 모델을 학습시키는데 사용된다. 이러한 점에서 Helmholtz machine 은 자기감독(self-suervised) 학습의 형태로 파악할 수 있다.

하나의 패턴에 내재된 특성, 즉 그 패턴을 생성해 낸 내부 요소를 바로 파악하는 것이 항상 가능한 것은 아니며 또한 그 패턴을 생성할 수 있는 방법은 아주 다양할 수 있다. 따라서 EM 알고리즘과 같은 일반적인 최대 유사도 방법을 이용하여 이러한 모든 경우들을 고려한다는 것은 계산상 무척 어려운 작업이다. Helmholtz machine 에서는 보다 계산이 쉬운 근사 방법을 적용함으로써 이러한 문제를 해결한다.

파라미터가 θ인 어떤 모델에서 데이터 집합 $\boldsymbol{D}$ 를 생성할 확률에 대한 로그값, 즉 로그 유사도는 다음과 같다.

$$\log(D \mid \theta) = \sum_{t=1}^{T} \log\left[\sum_{\alpha^{(t)}} P\left(d^{(t)}, \alpha^{(t)} \mid \theta\right)\right]$$

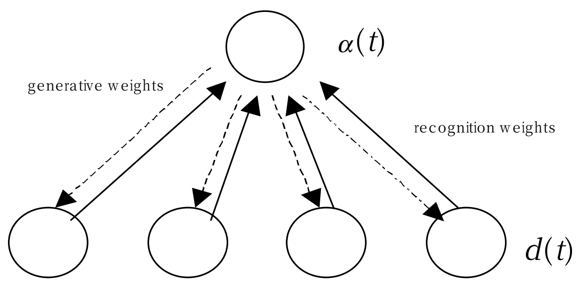
위 식에서 $\alpha^{(t)}$는 데이터 $d^{(t)}$를 생성하는 은닉 요인을 의미한다. 위 식에서 α에 대한 분포 P 를 계산하기가 아주 어렵기 때문에 위 식에 대한 로그 유사도 최대화 계산 역시 아주 어렵게 된다. 그래서 보다 쉬운 형태의 분포를 도입함으로써 위 문제를 근사적으로 해결할 필요가 있는데 Jensen 의 부등식을 적용하면 아래와 같은 부등식

을 얻을 수 있다.

$$\log(D\,|\,\theta)=\sum_{t=1}^{T}\log\left[\sum_{\alpha(t)}P\left(d^{(t)},\alpha^{(t)}\,|\,\theta\right)\right]$$
$$=\sum_{t=1}^{T}\log\left[\sum_{\alpha(t)}Q\left(\alpha^{(t)}\right)\frac{P\left(d^{(t)},\alpha^{(t)}\,|\,\theta\right)}{Q\left(\alpha^{(t)}\right)}\right]$$
$$\geq\sum_{t=1}^{T}\sum_{\alpha(t)}Q\left(\alpha^{(t)}\right)\log\frac{P\left(d^{(t)},\alpha^{(t)}\,|\,\theta\right)}{Q\left(\alpha^{(t)}\right)}$$

따라서 위 식 오른쪽의 하한값을 최대화시키면 로그 유사도에 대해서도 역시 최대화 작업을 수행할 수 있다. 이렇게 함으로써 *generalized EM* 과 유사한 작업이라고 할 수 있다.

Helmholtz machine 은 위와 같은 원리를 구현하기 위해서 생성 모델에 더하여 인식 모델을 도입하는데 이는 분포 Q 를 표현하기 위해서이다. 결과적으로 서로 다른 층에 속하는 노드들 사이에는 상향 연결뿐 아니라 하향 연결도 존재하게 된다. 다음 그림은 은닉층이 하나인 간단한 Helmholtz machine 이다.



네트워크를 구성하는 각 노드는 확률적 이진 노드이며, 그림에서 하향 연결들은 생성 모델을 구성하며, 상향 노드들은 인식 모델을 구성한다. 그리고 같은 층의 노드들은 그 부모 노드의 값이 주어졌을 경우 서로 독립적이다. 이 경우 노드 v 의 상태 s_v 는 다음과 같은 확률에 의해 결정된다.

$$P(s_v=1)=\frac{1}{1+\exp\left(-b_v-\sum_u s_u w_{uv}\right)}$$

위 식에서 b_v 는 노드 v 에 대한 편향값(bias)이며 w_{uv} 는 노드 u,v 를 연결하는 간선에 대한 가중치이다. 위와 같은 네트워크 구조상에서 생성 모델에 대한 파라미터와 인식 모델에 대한 파라미터를 반복적으로 조정함으로써 주어진 데이터 집합에 대한 로그 유사도를 최대화시키게 된다.

3. Wake-Sleep 알고리즘

Helmholtz machine 상에서 주어진 데이터 집합에 대한 로그 유사도를 최대화하기 위해 사용되는 학습 방법으로 흔히 사용되는 것이 *wake-sleep* 알고리즘이다[4][5].

Wake phase

1. 인식 모델을 통해 입력 층부터 최상위 층까지의 각 노드 값이 확률적으로 결정된다.
2. 생성 모델을 이용해 최상위 층부터 입력 층까지의 각 노드에서의 확률값을 계산해 나가면서 생성 모델의 파라미터 값을 수정하며, 다음과 같은 지역적 델타-규칙에 의해 이루어진다.

$$w_{uv}^{new}=w_{uv}^{old}+\Delta w_{uv}$$
$$\Delta w_{uv}=\gamma\, s_u(s_v-p(s_v=1))$$

위 식에서 γ는 학습율(learning rate)이다. 이 과정은 *generalized EM* 의 *M* (Maximization) 단계에 해당된다고 볼 수 있다.

Sleep phase

1. 생성 모델을 통해 최상위 층부터 입력 층까지 각 노드의 값이 확률적으로 결정된다. 즉 생성 모델에 의해 가상 데이터가 생성된다.
2. 인식 모델을 이용해 입력 층부터 최상위 층까지의 각 노드에서의 확률값을 계산해 나가면서 인식 모델의 파라미터 값을 수정하며, 간선의 가중치 값은 wake-phase 에서와 같은 식에 의해 수정된다.

이 과정은 *generalized EM* 의 ***E*** (expectation) 단계에 해당된다고 볼 수 있다.

Wake-sleep 알고리즘은 위의 두 단계를 반복적으로 실행함으로써 로그 유사도의 하한값을 최대화시키게 되며, 따라서 로그 유사도 역시 최대화한다.

Given data set $D=\{d_1,d_2,\cdots,d_n\}$

Initialize Helmholtz machine

Do

For all data set d_i

1. Do ***wake phase***
2. Do ***sleep phase***

Until (some likelihood criterion)

4. Helmholtz Machine 을 이용한 문서 분류

Helmholtz machine 을 통한 분포 추정을 통해 데이터에 대한 분류작업을 수행하고 그 결과를 Naïve Bayes 적용한 경우와 비교해 보았다..

4.1 문서 데이터

총 20 개의 범주를 포함하는 UseNet 뉴스그룹 데이터를 이용하였다. 불필요한 정보를 최소로 하기 위해 스테밍 알고리즘과 불용어 목록을 사용하였다. 또한 각 뉴스 본문 앞에 존재하는 헤더 정보도 모두 제거하였다. 모든 범주는 각각 1000 개의 문서를 포함하고 있다.

4.2 실험 방법

20 개의 범주 중에서 3 가지 범주에 대해서 실험을 하였다. 각 범주에 속하는 문서들 중 70%를 학습에 이용하고 나머지는 30%는 테스트에 이용하였다.

Helmholtz machine 의 경우 각 범주당 하나의 네트워크를 학습하였으며, 각 네트워크의 입력 노드는 단어 하나에 대응된다. 입력 노드의 수가 증가할수록 학습시간이

크게 증가하기 때문에 적절한 성능을 유지하면서 학습시간을 줄이기 위해 문서 벡터를 구성하는 단어의 수를 여과하였다. 각 단어마다 정보획득량(information gain)[1][6][7]을 계산하여 상위 2000 단어를 선정하였다. 또한 각 네트워크는 하나의 은닉층을 가지며, 입력노드는 2000 개, 은닉 노드는 10 개로 하였다.

각 범주에 대해 Helmholtz machine 을 학습시킨 후, 테스트 문서에 대해 분류 작업을 할 때는 문서를 각 네트워크에 입력으로 제공한 후, 가장 큰 로그 유사도 값을 출력하는 네트워크에 대응되는 범주를 해당 문서의 범주로 정하였다. 즉,

$$\hat{c} = \arg\max_{c \in C} P(d \mid c)$$

위 식에 의해 구한 범주 인덱스와 데이터 d의 실제 범주 인덱스가 일치할 때 올바르게 분류한 것으로 하였다.

4.3 결과

표 1 에 각 방법별로 각 범주에 대한 정확도로써 실험 결과를 제시하였다.

.

범주 집합	**NB-ALL**	**NB-2000**	**HM-2000**
talk.politics.guns	**93.67 %**	92.33 %	93.00 %
talk.politics.mideast	**93.67 %**	93.33 %	92.00 %
talk.politics.misc	82.00 %	79.00 %	**84.57 %**
Total	89.78 %	88.22 %	**89.89 %**

표 1. UseNet 뉴스그룹 문서에 대한 분류 실험

위 표에서 NB-ALL 은 28,000 여 개의 모든 단어를 이용한 Naïve Bayes 방법이며, NB-2000 은 정보 획득량에 의해 여과된 상위 2000 단어만 이용한 Naïve Bayes 방법이다. HM-2000 은 상위 2000 단어만 이용한 Helmholtz machine 이다. Helmholtz machine 의 경우 네트워크를 구성하는 노드들이 이진 노드이기 때문에 문서 내에 해당 단어의 유무로서 노드의 값을 설정하였다. 위의 실험 결과를 보면, 우선 Naïve Bayes 의 경우 28,000 여개의 모든 단어를 이용한 경우와 정보획득량에 따라 여과한 2000 개의 단어를 이용한 경우는 그 정확도(accuracy)면에서 많은 차이를 보이지 않는 것을 알 수 있다. 그리고 Helmholtz machine 의 경우 Naïve Bayes 방법과 비교해 볼 때 talk.politics.guns 와 talk.politics.mideast 의 경우에는 약간 정확도가 못 미치지만 talk.politics.misc 의 경우 그 정확도가 오히려 모든 단어를 이용했을 때보다도 더 높은 것을 알 수 있다. 이것은 아마도 그 이름에서도 알 수 있듯이 세 번째 범주의 경우 나머지 두 범주에 포함되지 않는 문서들을 포함하지만 전체적으로 정치(politics) 대담(talk)에 관한 문서이기 때문에 단순히 각 단어의 통계만으로 분류하면 오류를 포함할 가능성이 있지만 문서를 구성하는 단어들의 전체 분포를 추정함으로써 보다 더 나은 성능을 보일 수 있는 것 같다.

5. 결론

문서 분류 등의 분류 문제를 해결하기 위해서는 데이터에 대한 분포를 추정하여 간접적으로 분류 문제를 해결하는 것보다는 분류 문제를 위해 창안된 직접적 분류 모델을 이용하는 것이 더 좋다는 것이 일반적인 견해이다. 본 논문에서는 그러한 견해에 약간 배치되게 분포 밀도 추정을 통해 문서 분류 문제에 접근하였다. 이번 실험에서 이 방법은 그 성능 면에서 볼 때는 거의 비슷한 성능을 보이지만 학습 시간이나 계산의 어려움 측면에서 볼 때 기존 방법에 비해 훨씬 더 많은 비용을 요구한다. 하지만 결과에서도 보듯이 기존의 단순한 분류 모델에서는 분류 성능이 약간 미흡했던 범주에 대해서는 좀 더 나은 성능을 보임을 알 수 있다. 만약 그러한 부분이 아주 중요하다면 그래프 모델과 같은 분포 추정을 통한 문서 분류도 한 번 고려해 볼 만 할 것이다. 또한 이러한 그래프 모델을 통해 문서 속에 숨어 있는 은닉 정보를 획득함으로써 다른 분류 모델을 위한 선처리 단계로서 사용하는 것도 의미있을 것이다.

이번 실험에서 사용된 그래프 모델인 Helmholtz machine 의 경우 일반적으로 이진 노드를 사용하기 때문에 단어의 존재 여부만을 자질로 삼을 수 밖에 없었고 단어의 빈도나 TFIDF 등의 실수 정보를 이용할 수 없었는데, 앞으로 이러한 부분에 대해서 더 연구가 필요할 것 같다. 그리고 이러한 그래프 모델은 입력 노드의 수와 같은 그래프 구조 복잡도에 따라 계산 비용이 증가하기 때문에 입력 자질의 최적화나 구조의 최적화에 관한 연구도 필요하다.

감사의 글

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

6. 참고 문헌

[1] Cover, T. and Thomas, J., *Elements of Information Theory*, John Wiley & Sons, 1991.

[2] Dayan, P., Hinton, G. E., Neal, R. M., Zemel, R. S., “The Helmholtz Machine”, *Neural Computation*, vol. 7, pp. 889-904, 1995.

[3] Frey, B. J., *Graphcal Models for Machine Learning and Digital Communication*, The MIT Press, 1998.

[4] Hinton, G. E., Dayan, P., Frey, B. J., and Neal, R. M., “The Wake-Sleep Algorithm for Unsupervised Neural Networks”, *Science*, vol. 268, pp. 1158-1161, 1995.

[5] Neal, R. M., “Factor Analysis Using Delta-Rule Wake-Sleep Learning”, *Neural Computation*, vol. 9, pp. 1781-1803, 1997.

[6] Singh, M. “Learning Bayesian Networks for Solving Real-World Problems”, Ph.D thesis, University of Pennsylvania, 1998.

[7] Yang, Y., Pedersen, J. P., “A Comparative Study on Feature Selection in Text Categorization”, *Proceedings of the Fourteenth International Conference on Machine Learning*, pp. 412-420, Morgan Kaufman, 1997.