점진적 데이터 선택에 의한 Helmholtz machine 의 학습

Learning Helmholtz Machines by Incremental Data Selection

장정호, 장병탁, 서울대학교 컴퓨터공학부, {jhchang, btzhang}@bi.snu.ac.kr

초록: *Helmholtz machine 은 확률적 생성 모델의 일종으로서 주어진 데이터에 대한 density estimation 을 수행할 수 있다. 본 논문에서는 주어진 데이터로부터 점진적으로 데이터를 선택하여 모델을 학습하고 그 결과를 임의 선택에 의한 모델 학습과 비교한다.*

연구 배경: 기계 학습의 기본적인 목표는 주어진 데이터를 이용하여 어떤 작업에서의 성능을 향상시키고자 하는 것으로 볼 수 있다. 전통적으로 이러한 학습에서 구축하고자 하는 모델은 입력 공간에서 임의로 주어지는 데이터에 대해 학습을 수동적으로 학습을 진행한다. 이러한 접근법은 기본적으로 학습이 진행될수록 주어지는 데이터에 의해 추가될 수 있는 평균 정보의 양이 감소한다는 점이 단점이라 할 수 있다. 최근 감독 학습(supervised learning)에서는 이러한 문제점을 해결하려고 시도가 진행되고 있는데, 이러한 방법은 능동 학습(active learning) 또는 질의 학습(query learning)이라 일컬어진다 [1, 3]. 능동학습에서는 데이터를 데이터를 임의로 선택하지 않고 학습에 도움이 될 가능성이 큰 데이터를 신중하게 선택함으로써 비교적 적은 데이터만으로 학습이 진행될 수 있다.

본 논문에서는 비감독학습(unsupervised learning), 특히 은닉변수모델을 이용한 데이터 분포 추정을 위한 학습에서 능동학습의 개념을 적용하고자 한다. 만약 학습 대상이 되는 데이터의 양이 방대하지만, 데이터들에 특정 구조가 내재한다면 이러한 데이터들은 중복된 과잉정보를 포함한다고 볼 수 있다 [5]. 따라서 모든 데이터를 다 사용하지 않고 정보량이 많은 데이터만 선별하여 학습에 이용하는 것이 학습에 도움이 될 수 있을 것이다.

이러한 가정에 기반하여 점진적인 데이터 선택에 의한 Helmholtz machine 학습을 제시하고, 그 가능성을 검토해보고자 한다.

Helmholtz machine : Helmholtz machine [2]은 확률 생성 모델에서의 학습과 추론 과정을 보다 용이하게 하기 위한 근사적 방법의 일종이다. Helmholtz machine은 하나의 생성 네트워크(generative network)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 machine은 자기감독(self-supervised) 학습의 한 형태로 파악할 수 있다.

임의의 데이터 집합 $\mathbf{D}= \{d^{(1)}, d^{(2)}, \ldots, d^{(n)}\}$에 대해 생성된 모델의 적합성을 측정하기 위한 하나의 척도로 흔히 모델의 데이터에 대한 유사도(likelihood)가 사용된다. 모델의 파라미터 집합이 Θ 로 주어지고 집합 $\mathbf{D}$ 에 속한 각 데이터 $d^{(t)}$가 서로 독립이고 동일한 분포를 따른다면, 유사도에 로그를 취한 수치, 즉 로그 유사도는 다음과 같다.

$$\log(\mathbf{D}\,|\,\Theta) = \sum_{t=1}^{T} \log\left[\sum_{\alpha^{(t)}} P\left(d^{(t)}, \alpha^{(t)}\,|\,\Theta\right)\right] \qquad (1)$$

위 식에서 $\alpha^{(t)}$는 데이터 $d^{(t)}$를 생성하는 은닉 요인을 의미하는데 모든 α에 대한 분포 P를 계산하는 작업은 은닉 요인의 수에 따라 그 복잡도가 지수적으로 증가하기 때문에 위 식에 대한 로그 유사도 최대화 계산 역시 아주 어렵게 된다. 그래서 보다 쉬운 형태의 분포를 도입함으로써 위 문제를 근사적으로 해결할 필요가 있는데 Jensen의 부등식을 적용하면 아래와 같은 부등식을 얻을 수 있다.

$$\begin{aligned}\log(D\,|\,\Theta) &= \sum_{t=1}^{T} \log\left[\sum_{\alpha(t)} P\left(d^{(t)}, \alpha^{(t)}\,|\,\Theta\right)\right] \\ &= \sum_{t=1}^{T} \log\left[\sum_{\alpha(t)} Q\left(\alpha^{(t)}\right)\frac{P\left(d^{(t)}, \alpha^{(t)}\,|\,\Theta\right)}{Q\left(\alpha^{(t)}\right)}\right] \\ &\geq \sum_{t=1}^{T}\sum_{\alpha^{(t)}} Q\left(\alpha^{(t)}\right)\log\frac{P\left(d^{(t)}, \alpha^{(t)}\,|\,\Theta\right)}{Q\left(\alpha^{(t)}\right)}. \qquad (2)\end{aligned}$$

따라서 위 식 오른쪽의 하한값을 최대화시키면 로그 유사도에 대해서도 역시 최대화 작업을 수행할 수 있다. 이 방식은 *generalized EM*과 유사한 작업이라고 할 수 있다. 로그 유사도에 대한 하한값 추정을 가능하게 하는 Q를 학습하기 위한 것으로 인식네트워크가 사용된다.

데이터 밀도 기반의 데이터 샘플링에 의한 학습: 능동적 데이터 샘플링을 위한 방법은 크게 스트림 기반, 데이터 풀(pool)기반, 밀도 가중치 데이터 풀 기반으로 나눌 수 있다 [5]. 일반적인 학습의 경우 데이터의 에러 뿐 아니라 해당 데이터가 선택되었을 때, 얼마만큼 다른 데이터를 설명할 수 있는가도 중요하다는 것을 고려하여 본 논문에서는 밀도 가중치 데이터 풀에 기반하여 데이터를 선택한다. 특정 데이터에 대한 분포를 추정하기 위해 해당 데이터로부터 모든 다른 데이터에 대한 거리의 평균을 계산한다. Helmholtz machine은 이진 데이터를 그 입력 대상으로 하며, 이에 따라 두 데이터 간의 거리는 hamming distance에 의해 측정된다. 그리고 학습된 모델에 대한 데이터의 에러는 모델에 대한 데이터의 로그 유사도 값으로 취하였다.

Helmholtz machine의 학습에는 wake-sleep 알고리즘이 일반적으로 적용되는데, 이는 기본적으로 stochastic algorithm이다. 이러한 특성을 이용하여 후보 데이터 집합 $\boldsymbol{C}$로부터 학습 데이터 집합 $\boldsymbol{T}$에 데이터 d가 추가한 후의 모델의 학

습은 직전 모델 파라미터를 출발점으로 하여 미리 정해진 상수 c번 만큼 반복하여 모델을 학습한다.

표 1은 점진적 데이터 선택과 이를 통한 모델 학습을 위한 전체 절차이다.

$\boldsymbol{T}$: data set included for training a model $\boldsymbol{C}$: candidate data set for data sampling ● estimate parameters of the model using intial training set T. ● loop while adding data 1. Calculate likelihoods for each data in $\boldsymbol{C}$. 2. For appropriate number of data with lower likelihoods, select a sample d with highest data density and add this data to $\boldsymbol{T}$. $\boldsymbol{T}=\boldsymbol{T}\cup\{d\}$, $\boldsymbol{C}=\boldsymbol{C}-\{d\}$ 3. Estimate the parameter of the model with new data set $\boldsymbol{T}$.

표 1. 점진적 데이터 선택을 통한 Helmholtz machine 학습 과정

실험 결과: UCI hand-written digits 데이터를 사용하엿다. 0 부터 9 까지의 필기체를 포함하고 있으며, Helmholtz machine 의 입력으로 사용하기 위해 각 자질 값은 0 과 1 로 이산화하였다. 그림 1 은 데이터 선택에 따른 학습데이터와 검증데이터에 대한 로그 유사도 값의 변화를 보이는 그래프이다. 본 논문에서 제시한 데이터 선택에 의한 학습이 임의 선택에 의한 경우보다 학습 성능의 향상 속도면에서 보다 나은 결과를 보임을 알 수 있다.

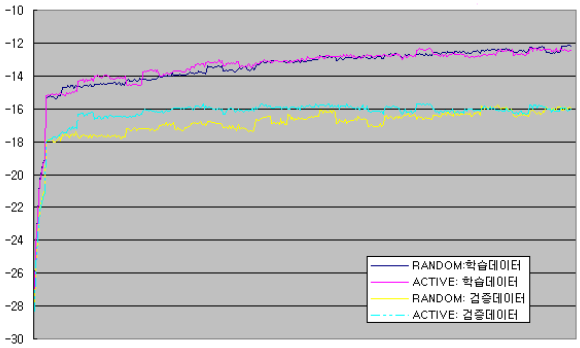

그림 1. 학습에 따른 로그 유사도 변화

[1] D. A. Chon and Z. Ghahramani. Active learning with statistical models. *Journal of Artificial Intelligence*, 4:129-145, 1996.

[2] B. J. Frey. Graphical Models for Machine Learning and Digital Coummnication. *The MIT Press*, 1998.

[3] M. Hasenager. Active Data Selection in Supervised and Unsupervised Learning. PhD thesis, Technische Fakultat der Universitat Bielefeld, 2000.

[4] D. Koller and S. Tong. Active learning for parameter estimation in Bayesian networks. In *Neural Information Processing Systes*, 2001.

[5] A. McCallum and K. Nigam. Employing EM in pool-based active learning for text classification. In *Proceedings of the 15th International Conference on Machine Learning* (ICML-98), 1998.