실수값 인자 데이터의 비지도 학습을 위한 에너지 기반 하이퍼네트워크 모델

김권일 [1O] 허민오 [2] 이상우 [2] 장병탁 [2]
서울대학교 협동과정 뇌과학전공 [1] 서울대학교 컴퓨터공학부 [2]
{kikim, moheo, swlee, btzhang}@bi.snu.ac.kr

Energy-based Hypernetworks Model for Unsupervised Learning on Real-valued Data

Kwonill Kim[1O] Min-Oh Heo[2] Sang Woo Lee[2] Byoung-Tak Zhang[2]
Interdisciplinary Program in Neuroscience, Seoul National University[1]
School of Computer Science & Engineering, Seoul National University[2]

요 약

하이퍼네트워크(Hypernetworks)는 하이퍼에지(hyperedge)들로 이루어진 생성 모델(generative model) 로서, 주로 이산(binary) 데이터에 적용되어왔다. 본 논문에서는 이산 데이터와 실수 데이터를 모두 다룰 수 있는 새로운 하이퍼네트워크 모델을 에너지 기반 모델(energy-based model)의 형태로 제시하고, 비지도 학습(unsupervised learning) 알고리즘으로 데이터를 성공적으로 학습함을 간단한 실험을 통해 보이겠다.

1. 서론

하이퍼네트워크(Hypernetworks)는, 임의의 개수의 vertex들로 이루어진 하이퍼에지(hyperedge)들과 그에 대응하는 가중치(weight)들로 구성된 하이퍼그래프(hypergraph)로 표현되는, graphical generative model 이다[1, 2]. 하이퍼네트워크는 데이터 인자 간의 관계를 해석하기 용이한 형태로 학습하는 특징을 가지는데, 이를 활용하여 언어 모델[3], 비디오 데이터 분석[4], 생물학 데이터 분석[5] 등 다양한 분야에 적용되고 있다.

그런데, 기존의 하이퍼네트워크 모델에서는 이산화 되거나 범주화된 데이터만을 사용해 왔다. 하이퍼네트워크는 학습 과정과 샘플링 과정에서 입력 값이 각 하이퍼에지가 표현하는 패턴과 일치하는지 여부를 판단하는데, 기존 모델은 실수 인자의 일치 여부를 판단할 수 없었고, 따라서 실수 데이터를 이산화하여 적용하였다. 그러나, 로봇 제어와 같이 이산화하기 알맞지 않은 데이터들이 대두됨에 따라, 이를 해결할 수 있는 새로운 하이퍼네트워크 모델과 학습 알고리즘이 요구되고 있다. [6]의 경우 이 문제를 해결하려 새로운 모델을 제시하였으나, 이는 분류 문제에만 적용될 수 있는 판별 모델(discriminative model)이라는 한계를 가진다.

또한 하이퍼네트워크는, 에너지 기반 모델(energy-based model)로 정의될 수 있음에도 불구하고, 대부분의 연구에서 진화 학습의 틀에서 학습되어왔다. 에너지 기반 모델은 RBM(Restricted Boltzmann Machine)이나 MRF(Markov Random Fields)와 같이 상태 변수들의 에너지 식으로 결합 확률분포를 정의하는 모델을 뜻하며, 에너지 식을 적절히 정의하여 원하는 특성을 추가하거나, 이 에너지 식으로부터 학습 알고리즘을 손쉽게 유도할 수 있는 이점이 있다[7, 8].

본 논문에서는 이산 데이터와 실수 데이터를 모두 다룰 수 있는 새로운 에너지 기반 하이퍼네트워크 모델을 제시한다. 2장에서는 새로운 에너지 식과 조건부 확률 식, 그리고 비지도 학습 알고리즘에 대해 설명하고, 3장에서는 새 하이퍼네트워크 모델이 이산 및 실수 데이터를 성공적으로 학습할 수 있음을 실험을 통해 보이겠다.

2. 에너지 기반 하이퍼네트워크 모델

2.1. 정의

에너지 기반 하이퍼네트워크 모델은 두 종류의 상태 변수, visible variable $\mathbf{x}=[x_d]_{d=1}^{D}$ 와 hidden variable $\mathbf{h}=[h_m]_{m=1}^{M}$ 를 가진다. x_d 는 -1 또는 1의 값을 가질 수 있는 이산 변수이거나, 임의의 실수 값을 가질 수 있는 연속 변수이고, h_m 는 0 또는 1의 값을 가질 수 있는 이산 변수이다. 상태 변수들이 어떤 값을 가질 때, 에너지와 결합 확률분포는 아래와 같이 정의된다.

$$\begin{aligned} E(\mathbf{x},\mathbf{h}) &= -\sum_{m}^{M} w_m c_m h_m - \sum_{m}^{M}\sum_{d}^{D} h_m w_m u_{md} x_d + \sum_{d}^{D} \frac{1}{2\sigma_d^2} x_d^2 \\ &= -\sum_{m}^{M} w_m \left(\varepsilon - \frac{1}{2}\sum_{d}^{D} u_{md}^2 \right) h_m - \sum_{m}^{M} h_m w_m \sum_{d}^{D} u_{md} x_d \\ &\quad + \sum_{d}^{D} \frac{1}{2}\left(1 + \sum_{m}^{M} h_m w_m \alpha_{md} \right) x_d^2 \end{aligned} \quad (0.1)$$

$$P(\mathbf{x},\mathbf{h}) = \frac{1}{Z}\exp\left(-E(\mathbf{x},\mathbf{h})\right) \quad (0.2)$$

여기서 $\mathbf{u}_m = [u_{md}]_{d=1}^{D}$ 는 m 번째 하이퍼에지의 패턴을 의미하며, 입력된 $\mathbf{x}$ 가 이 패턴과 유사할수록 h_m 이 1이 될 확률이 높아진다. w_m 은 m 번째 하이퍼에지의 가중치를 뜻하며, 0 이상의 값을 가지고, 본 모델의 유일한 학습 가능 매개 변수이다. ε 은, 0에서 2사이의 값을 가지는, $\mathbf{h}$ 의 발화 정도를 조절하는 매개 변수이며 α_{md} 는 u_{md} 가 0 일 때 0 이고, 0이 아니면 1의 값을 가진다.

2.2. 학습 방법

본 모델의 학습은 주어진 학습 데이터 집합(training dataset) $\mathcal{D} = \{\mathbf{x}^{(n)}\}_{n=1}^{N}$ 에 대해, 아래의 log-likelihood 함수를 최대화하는 가중치 조합을 찾는 것이다.

$$L(\mathbf{w}) = \frac{1}{N}\sum_{n=1}^{N} \ln P(\mathbf{x}^{(n)}) \tag{0.3}$$

이를 위해 gradient-ascent 기법을 적용하면, 아래와 같이 학습식을 유도할 수 있다.

$$w_{m,\text{new}} = w_{m,\text{old}} + \eta \frac{\partial L}{\partial w_m}$$

$$\frac{\partial L}{\partial w_m} = -\sum_{n}^{N}\sum_{\mathbf{h}} P(\mathbf{h}\mid\mathbf{x}^{(n)}) \frac{\partial E(\mathbf{x}^{(n)},\mathbf{h})}{\partial w_m} + \sum_{n}^{N}\sum_{\mathbf{x}',\mathbf{h}'} P(\mathbf{x}',\mathbf{h}') \frac{\partial E(\mathbf{x}',\mathbf{h}')}{\partial w_m}$$

$$\frac{\partial E(\mathbf{x},\mathbf{h})}{\partial w_m} = -h_m\left(\varepsilon - \frac{1}{2}\sum_{d}^{D}\alpha_{md}(x_d - u_{md})^2\right)$$

그런데, 위의 학습식을 계산하기 위해서는 결합 확률분포 $P(\mathbf{x},\mathbf{h})$ 로부터 $\mathbf{x}$ 와 $\mathbf{h}$ 를 sampling 해야 한다. 이를 Gibbs sampling 하기 위한 조건부 확률도 결합 확률분포 식에서 유도되며, 아래와 같다.

$$P(\mathbf{h}\mid\mathbf{x}) = \prod_{m}^{M} P(h_m\mid\mathbf{x}), \quad P(\mathbf{x}\mid\mathbf{h}) = \prod_{d}^{D} P(x_d\mid\mathbf{h})$$

$$P(h_m = 1\mid\mathbf{x}) = sigm\left(w_m\left(\varepsilon - \frac{1}{2}\sum_{d}^{D}\alpha_{md}(x_d - u_{md})^2\right)\right) \tag{0.4}$$

$$sigm(x) = 1/(1+\exp(-x))$$

$$P(x_d = 1\mid\mathbf{h}) = sigm\left(2\sum_{m}^{M} h_m w_m u_{md}\right), \quad \text{for } x_d \in \{-1,1\} \tag{0.5}$$

$$P(x_d\mid\mathbf{h}) = \mathcal{N}(x_d;\mu_d,\sigma_d^2), \qquad \text{for } x_d \in \mathbb{R}$$

$$\mu_d = \sigma_d^2\sum_{m}^{M} h_m w_m u_{md}, \quad \sigma_d^2 = 1/\left(1+\sum_{m}^{M} h_m w_m \alpha_{md}\right)$$

$\mathcal{N}(x_d;\mu_d,\sigma_d^2)$ 은 1차원 가우시안 확률 분포를 뜻한다. 또한, 학습 속도 향상을 위해 Gibbs sampling을 근사한 Persistent Contrastive Divergence를 사용한다[9].

그리고, 보다 적은 수의 하이퍼에지들로 이루어진 sparse한 모델을 구하기 위해, 목표함수 (2.3)에 L_1 Regularization 항 $-\lambda\sum_{m=1}^{M} w_m$ 을 추가할 수 있다($\lambda \geq 0$).

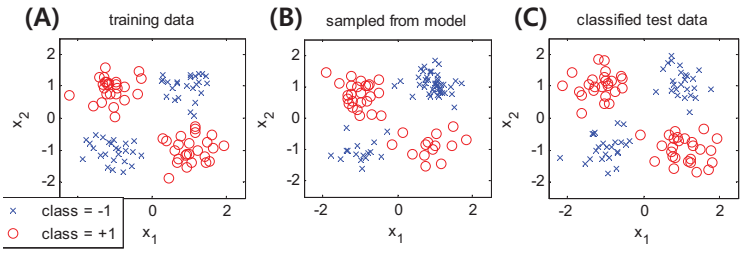

그림 1. XOR 데이터 학습 결과. (A) XOR 패턴의 training data. (B) 학습된 하이퍼네트워크 모델에서 샘플링 된 데이터. (C) 학습된 하이퍼네트워크 모델로 test data를 분류한 결과.

3. 실험

3.1. XOR 데이터

에너지 기반 하이퍼네트워크 모델이 데이터의 결합 확률분포를 학습할 수 있음을 확인하기 위하여, 그림 1. (A) 와 같은 XOR 패턴의 데이터 샘플 100개를 생성하여 학습하였다. 학습된 모델에서 샘플링 된 샘플들은 그림 1. (B)와 같이 (A)와 유사한 분포를 보였으며, 테스트 데이터에 대해서도 그림 1. (C)에서와 같이 XOR 패턴을 정확히 분류해내었다.

3.2. UCI Machine Learning Repository 데이터

에너지 기반 하이퍼네트워크 모델이 실제 데이터 역시 잘 학습함을 확인하기 위하여, UCI Machine Learning Repository[1]의 Iris dataset[2](이하 Iris), Wisconsin Diagnostic Breast Cancer dataset[3](이하 WDBC), Pima Indians Diabetes dataset[4](이하 Diabetes)으로 에너지 기반 하이퍼네트워크(이하 EHN)의 분류 성능을 실험하였다. 그리고, SVM, Naive Bayes(이하 NB), k-NN와의 분류성능 비교를 위해서 data mining software인 Weka [5] 를 사용하여 실험하였다.

모든 실험은 10-fold cross-validation을 10회 반복하여, 평균 테스트 분류 정확도를 비교하였다. EHN은 Iris와 WDBC 실험 시, order가 3 또는 4인 하이퍼에지 만을 사용하였고, learning rate는 1, λ 는 0.1로 하였다. Diabetes 실험에서는 order가 9인 하이퍼에지를 사용하고, λ 를 0으로 하였다. SVM, NB, k-NN의 경우에는 대개 Weka의 기본 설정 값을 사용하였으나, Diabetes 실험에서는 gamma=1인 RBFKernel을 SVM에 사용하였다. 또한 k-NN의 경우, k의 값을 Iris, WDBC, Diabetes 실험에 대하여 각각 1, 9, 7로 설정하고 실험하였다.

그 결과, 표 1에서와 같이 EHN이 다른 classifier들에

[1] http://archive.ics.uci.edu/ml/
[2] http://archive.ics.uci.edu/ml/datasets/Iris
[3] http://archive.ics.uci.edu/ml/datasets/Breast+ Cancer+ Wisconsin+ %28Diagnostic%29
[4] http://archive.ics.uci.edu/ml/datasets/Pima+ Indians+ Diabetes
[5] http://www.cs.waikato.ac.nz/ml/weka/

비해 크게 뒤떨어지지 않는 분류 성능을 가짐을 확인할 수 있었다.

표 1. UCI Machine Learning Repository 데이터 학습 결과.

10-fold CV Accuracy(%)	SVM	NB	k-NN	EHN
Iris	96.3±0.3	94.8±0.7	95.4±0.4	90.1±1.3
WDBC	97.6±0.2	93.3±0.2	97.1±0.3	92.7±0.6
Diabetes	77.0±0.5	75.5±0.4	73.9±0.4	74.1±0.9

4. 결론 및 향후 과제

본 논문에서는 이산 데이터와 실수 데이터를 모두 다룰 수 있는 새로운 에너지 기반 하이퍼네트워크 모델을 제시하였다. 향후 학습 속도와 성능을 향상시키기 위해, deep architecture로 확장하거나, 학습 알고리즘을 개선하는 연구가 이어질 것으로 기대된다.

감사의 글

이 논문은 2012년도 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구이며(No. 2012-0005643, Videome), 정부(지식경제부)의 재원으로 한국산업기술평가관리원의 지원(10035348, mLife) 및 교육과학기술부의 BK21-IT 프로그램에서 일부 지원되었음.

참고문헌

[1] B. Zhang, “Hypernetworks: A Molecular Evolutionary Architecture for Cognitive Learning and Memory,” *IEEE Computational Intelligence Magazine*, vol. 3, no. 3, pp. 49-63, Aug. 2008.

[2] B. Zhang and J. Kim, “DNA Hypernetworks for Information Storage and Retrieval,” *in Preliminary Proceedings of the Twelfth International Meeting on DNA Computing (DNA 12), Lecture Notes in Computer Science*, vol. 4287, pp. 283-292, 2006.

[3] M. Heo, M. Kang, and B. Zhang, “Visual Query Expansion via Incremental Hypernetwork Models of Image and Text,” *Proceedings of the Eleventh Pacific Rim International Conference on AI (PRICAI2010), Lecture Notes in Artificial Intelligence*, vol. 6230, pp. 88-99, 2010.

[4] 유준희, 석호식, 장병탁, “드라마 배경 변환 탐지를 위한 베이지안 필터링 방법,” *정보과학회논문지 : 컴퓨팅의 실제 및 레터*, vol. 18, no. 4, pp. 341-345, 2012.

[5] J. Ha, J. Eom, S. Kim, and B. Zhang, “Evolutionary hypernetwork models for aptamer-based cardiovascular disease diagnosis,” *GECCO ’07*, 2709-2716, 2007.

[6] 하정우, 장병탁, “이산화 과정을 배제한 실수 값 인자 데이터의 고차 패턴 분석을 위한 진화연산 기반 하이퍼네트워크 모델,” *정보과학회논문지 : 소프트웨어 및 응용*, vol. 37, no. 2, pp. 120-128, 2010.

[7] Y. Bengio, “Learning Deep Architectures for AI,” *Foundations and Trends in Machine Learning*, vol. 2, no. 1, pp. 1-127, 2009.

[8] Y. Bengio, P. Lamblin, and D. Popovici, “Greedy layer-wise training of deep networks,” *NIPS’06*, 153-160, 2007.

[9] T. Tieleman, “Training restricted Boltzmann machines using approximations to the likelihood gradient,” *ICML ’08*, 1064-1071, 2008.