고차 데이터 분류를 위한 순차적 베이지안 샘플링을 기반으로 한 하이퍼네트워크 모델의 진화적 학습 기법

하정우O*, 김수진**, 장병탁*·**
*서울대학교 컴퓨터공학부 {jwha, btzhang}@bi.snu.ac.kr
**서울대학교 생물정보학 협동과정 sjkim@bi.snu.ac.kr

Evolutionary Learning of Hypernetwork Classifiers Based on Sequential Bayesian Sampling for High-dimensional Data

Jung-Woo HaO*, Soo-Jin Kim**, and Byoung-Tak Zhang*·**
*School of Computer Science and Engineering, Seoul National University
**Interdisciplinary Program in Bioinformatics, Seoul National University

요 약

본 연구에서는 고차 데이터 분류를 위해 순차적 베이지안 샘플링 기반의 진화연산 기법을 이용한 하이퍼네트워크 모델의 학습 알고리즘을 제시한다. 제시하는 방법에서는 모델의 조건부 확률의 사후(posterior) 분포를 최대화하도록 학습이 진행된다. 이를 위해 사전(prior) 분포를 문제와 관련된 사전지식(prior knowledge) 및 모델 복잡도(model complexity)로 정의하고, 측정된 모델의 분류성능을 우도(likelihood)로 사용하며, 측정된 사전분포와 우도를 이용하여 모델의 적합도(fitness)를 정의한다. 이를 통해 하이퍼네트워크 모델은 고차원 데이터를 효율적으로 학습 가능할 뿐이 아니라 모델의 학습시간 및 분류성능이 개선될 수 있다. 또한 학습 시에 파라미터로 주어지던 하이퍼에지의 구성 및 모델의 크기가 학습과정 중에 적응적으로 결정될 수 있다. 제안하는 학습방법의 검증을 위해 본 논문에서는 약 25,000개의 유전자 발현정보 데이터셋에 대한 분류문제에 모델을 적용한다. 실험 결과를 통해 제시하는 방법이 기존 하이퍼네트워크 학습 방법 뿐 아니라 다른 모델들에 비해 우수한 분류 성능을 보여주는 것을 확인할 수 있다. 또한 다양한 실험을 통해 사전분포로 사용된 사전지식이 모델 학습에 끼치는 영향을 분석한다.

1. 서 론

하이퍼네트워크(hypernetwork: HN)는[1-2] 하이퍼그래프(hypergraph) 구조를[3] 이용하여 데이터 변수들간의 고차(higher-order) 연관관계를 모델링하는 확률 그래프 모델이다. HN 모델은 패턴인식[4], 멀티미디어 데이터 분석[5] 등의 분야에서 널리 활용되었으며 특히 생물정보학[6]에서 중요인자의 조합 탐색이 가능한 분류모델로서 성공적으로 적용되었다. HN 모델의 학습은 샘플링 기반의 진화연산을 이용하여 하이퍼에지 (hyperedge)로 표현되는 고차의 변수 조합의 공간을 탐색하는 것으로서 학습을 통해 데이터 분포를 표현하는 최적의 부분패턴 집합이 생성되고 데이터 분류를 위한 하이퍼에지들의 가중치가 계산된다.

기존의 HN 분류 모델 학습방법에서는 데이터로부터 하이퍼에지를 생성할 때 데이터와 관련된 사전지식(prior knowledge)을 이용하지 않고 모든 변수를 동일한 확률로 선택하여 생성하였다[4-6]. 이 방법은 데이터를 구성하는 변수의 수가 수 만개를 초과하는 고차원 데이터를 분류하는데 있어서는 비효율성을 야기할 수 있다. 또한 학습에 있어서 하이퍼에지의 수로 표현되는 모델 크기와 하이퍼에지의 구성 등 중요한 파라미터들이 반복실험을 통해 경험적으로 획득된 값으로 결정되었다[4-6].

본 논문에서는 이러한 문제를 해결하기 위해 순차적 베이지안 샘플링 기법을[7] 기반으로 하여 HN 분류 모델을 학습하는 진화 알고리즘을 제안한다. 제안하는 학습방법에서는 학습효율 개선을 위해 데이터 변수들과 클래스 변수간의 상호정보량(mutual information: MI) 과 모델 복잡도를 경험적 사전분포(empirical prior distribution)로 정의하고 훈련데이터에 대한 분류성능 측정을 통해 우도(likelihood)를 계산한다. 이를 통해 산출된 사후분포(posterior distribution)를 모델 적합도(fitness)로 정의하고 이를 기반으로 매 세대마다 낮은 가중치를 갖는 하이퍼에지들을 새로운 하이퍼에지로 대체함으로써 다음 세대의 모델을 생성한다. 이를 통해 제안하는 학습방법에서는 분류성능이 개선될 뿐 아니라, 적합도로부터 모델의 중요 파라미터인 하이퍼에지의 구성과 모델 크기가 적응적으로 결정된다.

본 연구에서는 제안하는 학습방법의 평가를 위해 약 25000개의 유전자로 구성된 고차 데이터셋으로부터 암 유형을 분류하는 문제에 적용한다[8]. 실험결과를 통해 베이지안 샘플링 기반의 진화학습 방법이 분류성능과 학습 시간 면에서 기존 HN분류 모델의 학습기법을 크게 개선했을 뿐 아니라 다른 분류 모델들에 비해 더 좋은 분류성능을 보여주는 것을 확인 가능하다. 그리고 다양한 결과 분석을 통해 제안하는 학습방법이 어떠한 측면에서 기존의 학습 방법을 개선하는지를 확인한다.

2. 하이퍼네트워크 분류 모델

하이퍼네트워크(HN) 모델은 가중치가 부여된 하이퍼에지들의 집합으로 구성된 확률 그래프 모델이다.

HN모델에서는 정점(vertex)을 데이터 변수 값(variable value)으로 하이퍼에지(hyperedge)를 두 개 이상의 임의의 정점들로 구성된 집합으로 각각 정의한다. 그러므로 각각의 하이퍼에지는 복수 개의 데이터 변수들간의 고차 연관관계를 표현하는 데이터의 부분패턴을 의미하게 되며 HN모델은 이러한 부분패턴들로 구성된 집합이 된다. 수학적으로 HN모델은 정점의 집합 $V=\{v_1, v_2, .., v_{|V|}\}$과 하이퍼에지의 집합 $E=\{e_1, e_2, \ldots, e_{|E|}\}$를 이용하여 $H=(V,E)$로 정의한다. 이 때 하나의 하이퍼에지 내에 포함된 변수의 수를 하이퍼에지의 차수(degree of hyperedge: $|e|$)로 표현하고 정점의 차수(degree of vertex: $d(v)$)는 특정 정점을 포함하는 모든 하이퍼에지의 가중치의 합으로 표현된다[3]. HN분류모델은 HN모델을 분류문제에 적합하도록 정의한 모델로서 하나의 하이퍼에지가 복수 개의 데이터 변수 값과 클래스 값의 집합으로 표현되며 가중치는 에지가 클래스 변수를 잘 구분하는 데이터 변수의 조합일 수록 높게 정의된다. HN분류모델의 분류 과정을 설명하기 위해 다음 두 가지 함수를 정의한다.

$$f_i^{(n)} = f(\mathbf{x}^{(n)}, e_i) = \begin{cases} 1, \ \mathbf{x}^{(n)} \text{ 가 } e_i\text{-}\{y_i\}\text{와 매칭하는 경우} \\ 0, \ \text{그렇지 않은 경우} \end{cases} \quad (1)$$

$$\varphi_i^{(n)} = \varphi(y^{(n)}, y_i) = \begin{cases} 1, \ y^{(n)} = y_i \\ 0, \ y^{(n)} \neq y_i \end{cases} \quad (2)$$

위의 식에서 클래스 값을 제외한 e_i에 포함된 모든 변수 값이 $\mathbf{x}^{(n)}$의 변수 값과 같은 경우 매칭이라고 표현한다. 이 때 새로운 데이터 샘플($\mathbf{x}^{(n)}$, $y^{(n)}$)가 주어지면 다음과 같은 절차에 의해 분류가 진행된다.

1. 모든 하이퍼에지 집합 E에 대해 클래스 별 가중치 합 $c_{y'}$를 다음 식에 의해 산출

$$c_{y'} = \sum_{i=1}^{|E|} \left\{ w(e_i) f_i^{(n)} \delta_i^{y'} \right\}. \quad (3)$$

2. $\hat{y}^{(n)} = \arg\max_{y' \in Y} c_{y'}$ 인 $\hat{y}^{(n)}$를 $y^{(n)}$으로 예측

3. 베이지안 진화적 하이퍼네트워크 학습 알고리즘

본 연구에서는 진화연산을 이용하여 복잡한 조합의 공간을 탐색하는 HN분류모델의 학습을 순차적 베이지안 샘플링 프로세스로[7] 설명한다. t세대의 HN모델 H_t가 주어지면 베이지안 룰에 의해 조건부 사후분포는 다음과 같이 조건부 우도와 데이터가 고려된 경험적 사전분포의 곱에 비례한다.

$$p(H_t \mid X, Y) = \frac{p(Y \mid X, H_t) p(H_t \mid X)}{p(Y \mid X)} \propto p(Y \mid X, H_t) p(H_t \mid X) \quad (4)$$

그리고 t세대의 사후분포는 t+1세대의 경험적 사전분포로 활용됨으로써 모델의 학습이 진행된다.

기존 HN분류모델의 학습기법[4-6]과는 달리 본 논문에서는 경험적 사전분포 $p(H_t|X)$로서 균일분포(uniform distribution) 대신 각각의 데이터변수와 클래스 변수간의 상호정보량(MI)값을 사용한다. 즉 하이퍼에지의 생성시에 동일한 확률대신에 다음과 같은 확률로 하이퍼에지에 포함될 변수를 선택한다.

$$P_{MI}(X_i) = \left(MI(X_i, Y) + \eta \right) \Big/ \sum_{j=1}^{|X|} \left\{ MI(X_j, Y) + \eta \right\}. \quad (5)$$

위 식에서 η값을 통해 모델학습에서 MI가 끼치는 영향을 조절할 수 있으며 이를 통해 클래스와 연관성이 큰 변수가 하이퍼에지를 구성할 확률을 높일 수 있다. 또한 학습속도 향상 및 모델의 일반화 성능 강화를 위해 사전분포가 모델 크기에 반비례하도록 정의하며 모델 크기는 모든 정점의 차수의 합으로 표현한다. 우도는 조건부 확률로 정의되며 우도가 모델의 분류성능을 반영하도록 정의 하기 위해 분류성능은 클래스를 맞추면서 매칭되는 에지들과 클래스를 틀리면서 매칭되는 에지들의 가중치 합의 차이로 표현된다고 가정한다.

$$p(Y \mid X, H_t) \approx \prod_{n=1}^{N} p\left(y^{(n)} \mid \mathbf{x}^{(n)}, H_t \right), \quad (6)$$

$$p\left(y^{(n)} \mid \mathbf{x}^{(n)}, H_t \right) \equiv \sum_{i=1}^{|E_t|} w(e_i) \left\{ f_i^{(n)} \cdot \varphi_i^{(n)} - f_i^{(n)} \cdot \left(1 - \varphi_i^{(n)} \right) \right\} \Big/ \sum_{i=1}^{|E_t|} w(e_i), \quad (7)$$

$$p(Y \mid X, H_t) = \prod_{n=1}^{N} \left[\sum_{i=1}^{|E_t|} w(e_i) \left\{ f_i^{(n)} \cdot \left(2\varphi_i^{(n)} - 1 \right) \right\} \Big/ \sum_{i=1}^{|E_t|} w(e_i) \right]. \quad (8)$$

위의 식에서 가중치 $w(e)$는 다음과 같이 하이퍼에지 각각의 클래스 구분 능력과 일반화 성능을 위한 차수의 역수의 함수로 정의된다.

$$w(e_i) = \sum_{n=1}^{N} f_i^{(n)} \cdot \delta_i^{(n)} - (1-\alpha) \sum_{n=1}^{N} f_i^{(n)} + \frac{\beta}{|e_i|}. \quad (9)$$

그 결과 $|H_t|$를 모델 복잡도로 정의할 때, 모델의 적합도 F_t를 측정된 우도와 사전분포의 곱의 log로 정의하면 다음과 같다.

$$F_t = \log p(Y \mid X, H_t) + \log p(H_t \mid X) \quad (10)$$

$$\approx \sum_{n=1}^{N} \log \frac{\sum_{i=1}^{|E_t|} \left\{ w(e_i) f_i^{(n)} \left(2\varphi_i^{(n)} - 1 \right) \right\}}{\sum_{i=1}^{|E_t|} w(e_i)} + \lambda |H_t| + \gamma \sum_{e \in E_t} \log \sum_{x_i \in e} P_{MI}(x_i)$$

위의 적합도 식에서 λ는 모델 크기에 페널티를 주기 위한 0보다 작은 값이며 이 값의 크기는 다음 수식과 같이 매 세대별로 적합도 증가율에 따라 감소하는 하이퍼에지의 수에 사용된 상수인 τ 에 비례한다.

$$R_t = \frac{R_{max} - R_{min}}{\exp(t/\kappa)} + R_{min}, \ G_t = \gamma_t \cdot R_t, \ \gamma_t = \begin{cases} (F_{t-1}/F_t)^{\tau} & (F_t - F_{t-1} \geq 0) \\ F_{max}/F_t & (F_t - F_{t-1} < 0) \end{cases} \quad (11)$$

위의 식에서 R_t와 G_t는 각각 t세대에서 제거되는 낮은 가중치의 하이퍼에지 및 새로 샘플링되는 하이퍼에지의 수이다. HN분류모델은 적합도가 더 이상 증가하지 않을 때까지 학습이 진행되며 적합도의 정의에 의해 모델 크기와 하이퍼에지의 구성 및 진화 세대수가 적응적으로 결정된다.

4. 실험 결과 및 분석

본 연구에서는 제안하는 베이지안 샘플링 기반 HN분류모델 학습기법을 평가하기 위해 90명의 전립선 암 환자 환자로부터 획득한 24,591개의 유전자 발현데이터를 공격성 및 비공격성 암 유형을 분류하는 문제에 적용한다[8].. 실험을 위해 사용된 모델 파라미터로서 각 하이퍼에지의 차수는 3에서부터 10까지 다양하며 최초 하이퍼에지의 수는 1000으로 설정하였다. 그리고 식(9)에 사용된 α와 β는 각각 0.1 과

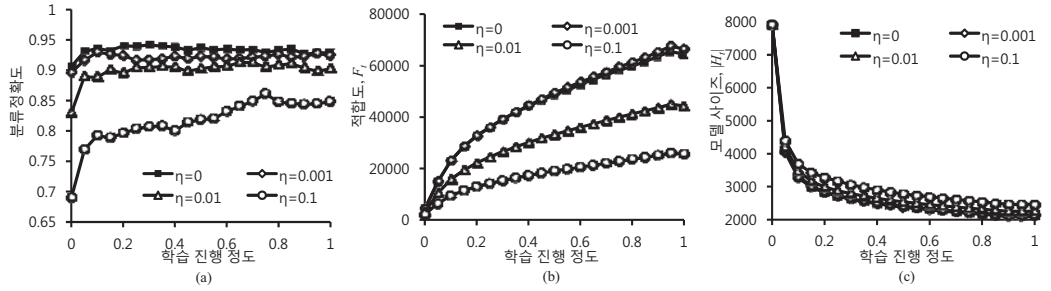

그림 2. η와 τ의 변화에 따른 분류성능(a), 모델적합도(b) 및 모델 복잡도(c)의 변화. τ=0.5로 고정.

표1. 분류성능 비교

모델	HNs	CHNs	SVMs	DTs	NBs
정확도	0.925	0.852	0.909	0.712	0.844
표준편차	0.012	0.087	0.012	0.025	0.020

CHNs는 기존 학습기법 이용한 HN모델, 기타 분류모델은 Weka에 구현된 알고리즘 사용. τ=0.5 η=0.001 사용

1이다. 표 1에서 실험결과로서 제시된 분류성능은 10-fold cross validation을 서로 다른 무작위 초기값을 이용해서 10회 시행한 후 평균을 구한 값이다.

표1을 통해서 제안한 베이지안 샘플링 기반의 학습방법을 적용한 HN분류모델이 기존 학습 방법을 이용한 HN모델뿐 아니라 다른 분류모델들에 비해 우월한 성능을 보여준다는 것을 알 수 있다. 또한 그림 1로부터 제시하는 방법이 τ 를 이용하여 모델의 학습시간을 크게 감소시킴을 알 수 있다. 그림 2는 MI 반영 정도 관련 변수인 η값의 변화에 따른 모델의 성능 변화를 표현한다. η가 커질수록 무작위 탐색에 가까워짐을 고려하면 그림 2을 통해 η값이 커질수록 분류성능 및 적합도가 감소하는 것을 알 수 있다. 그러나 작은 η값에 대해서는 η=0과 비슷한 성능을 보여줌을 알 수 있다. 그러므로 기존의 HN분류모델의 학습방법이 η=∞, τ=0 인 모델임을 고려하면, 그림 1과 2를 통해 본 연구에서 제안하는 학습모델이 기존 방법에 비해 사전분포의 도입을 통해 작은 모델 복잡도로 인한 학습시간 감소에도 불구하고 분류성능을 향상시켰음을 알 수 있다. 또한 표2는 학습된 모델에서 정점의 차수가 높은 상위10개 유전자들의 MI순위이다. 표2로부터 비록 하이퍼에지 생성시에 MI가 사용되지만 MI값이 작아도 고차적으로 연관성이 높은 유전자들은 학습을 통해서 모델을 구성하게 됨을 알 수 있으며 이는 제안하는

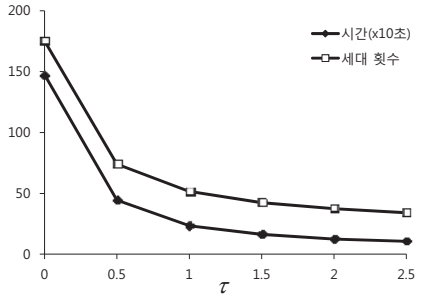

그림 1. τ 값의 증가에 따른 학습 속도 변화

표2. 차수순위가 상위10위에 속하는 유전자의 MI순위

순위	유전자	MI순위	순위	유전자	MI순위
1	ST7=L	91	2	MSH6=H	60
3	RBMS1=H	8	4	PEX26=L	26
5	FABP4=L	722	6	GRK5=L	61
7	LAG3=L	708	8	LSM10=L	2422
9	MGC35361=L	10196	10	MXD3=H	3058

방법이 인자선택(feature selection)을 전처리로 사용하는 경우와 달리 문제 공간을 줄이지 않고서도 효율적으로 학습가능하다는 것을 보여준다.

감사의 글

이 논문은 지식경제부 산업원천기술개발사업(10035348, mLife), 교육과학기술부 국가연구재단의 지원을 받아 수행된 연구(No. 2012-0005643) 및 BK21-IT사업에 의해 일부 지원되었음.

참고문헌

[1] B.-T. Zhang, “Hypernetworks: A molecular evolutionary architecture for cognitive learning and memory,” *IEEE Computational Intelligence Magazine*, 3(3), pp. 49–63, 2008.

[2] B.-T. Zhang and H.-Y. Jang, "A Bayesian algorithm for in vitro molecular evolution of pattern classifiers," In *Proc. of the Tenth International Meeting on DNA Computing* (*DNA*10), pp.294-303, 2004.

[3] Zhou, D., Huang, J., and Schoelkopf, B. “Learning with hypergraphs: Clustering, classification, and embedding,” *Advances in Neural Information Processing Systems* (*NIPS*) 19, 2007.

[4] J.-K. Kim and B.-T. Zhang, "Evolving hypernetworks for pattern classification," in *Proc. of IEEE Congress on Evolutionary Computation* (*CEC* 2007), pp.1856~1862, 2007.

[5] J.-W. Ha, B.-H. Kim, B. Lee, and B.-T. Zhang, "Layered hypernetwork models for cross-modal associative text and image keyword generation in multimodal information retrieval," *LNAI*, 6230, pp. 76-87, 2010.

[6] S. Kim, S.-J. Kim, and B.-T. Zhang, “Evolving hypernetwork classifiers for microRNA expression profile analysis,” in *Proc. of IEEE Congress on Evolutionary Computation* (*CEC* 2007), pp. 313-319, 2007.

[7] A. Doucet, S. Godsill, and C. Andrieu, "On sequential Monte Carlo sampling methods for Bayesian filtering," *Statistics and Computing* 10, pp.197-208, 2000.

[8] L. Wang *et al*. “Genome-wide transcriptional profiling reveals microRNA-correlated genes and biological processes in human lymphoblastoid cell lines,” *PLoS ONE*, 4(6) e5878, 2009.