

베이지안 GTM을 이용한 패턴 분류

Pattern Classification by Using Bayesian GTM

최준혁*, 김종배**, 김대수***, 임기욱****

김포대학 컴퓨터계열, 한국전자통신연구원 컴퓨터·소프트웨어기술 연구소,
한신대학교 컴퓨터학과, 선문대학교 산업공학과

JunHyeog Choi*, JoongBae Kim**, DaeSu Kim*** · KeeWook Rim****

Department of Computer Science, Kimpo College*,

Computer & Software Technology Lab., ETRI**,

Department of Computer Science, Hanshin Univ.***,

Department of Industrial Engineering, Sunmoon Univ.****

E-mail : jhchoi@kimpo.ac.kr, jjkim@etri.re.kr, daekim@hucc.hanshin.ac.kr,
rim@omega.sunmoon.ac.kr

ABSTRACT

Bishop이 제안한 Generative Topographic Mapping(GTM)은 Kohonen이 제안한 자을 학습 신경망인 Self Organizing Maps(SOM)의 확률적 버전이다. 본 논문에서는 이러한 GTM 모형에 베이지안 추론을 결합하여 작은 오분류율을 가지는 분류 알고리즘인 베이지안 GTM(Bayesian GTM)을 제안한다. 이 방법은 기존의 GTM의 빠른 계산 처리 능력과 베이지안 추론을 이용하여 기존의 분류 알고리즘보다 우수한 결과가 나타남을 실험을 통하여 확인하였다.

Key word : Generative Topographic Mapping, Bayesian Inference, Slice Sampling

1. 서론

신경망 분야에서 Kohonen은 주어진 입력에 대한 정확한 출력을 모르는 상황에서 스스로 학습할 수 있는 Self Organizing Maps(SOM)[4][6]을 제안하였다. 이 모형은 다른 신경망과 마찬가지로 모형에 대한 정확한 해석이 어렵다. Bishop은 이러한 SOM 모형에 이론적 규칙을 뒷받침할 수 있는 SOM의 확률적 버전인 Generative Topographic Mapping(GTM)[2]을 제안하였다.

본 논문에서는 이러한 GTM 모형에 베이지안 추론(Bayesian inference)을 적용하여 패턴을 분류하는 모형을 제안하였다. 이러한 베이지안 추론에 의한 베이지안 GTM 모형은 기존의 신경망의 과대적합(over-fitting)과 모형 설명의 어려움[1][7]을 극복하고 분류의 정확성을 향상시킨다. 본 논문의 2장에서 일반적인 GTM 모형에 대해 알아보고, 3장에서는 본 논문에서 제안하는 베이지안 GTM에 대한

구조를 설명하고, 4장에서는 베이지안 GTM과 기존의 분류 알고리즘과의 분류 성능에 대해서 비교한 실험 결과를 보이고, 5장에서 베이지안 GTM의 향후 연구 계획에 대해서 알아본다.

2. GTM 모형

많은 변수들 간에는 상당한 상관 관계가 존재한다. 이러한 것들을 모형화하는 한 가지 방법은 잠재 변수(latent variable)를 이용하여 자료의 분포를 모형화하는 것이다. GTM은 잠재 변수에 의해 모형을 표현한다. 잠재 변수 모형은 고차원의 공간을 잠재 변수를 이용하여 저차원의 공간으로 변환시킨다. 자료 공간에서 잠재 공간으로 선형 변환을 통해 차원을 축소시키는 인자 분석 등은 잠재 변수 모형의 한 예가 될 수 있다. 잠재 변수 모형의 중요한 적용은 바로 자료의 시각화를 통한 군집 분석이다. 많은 시각화 모형은 D-차원의 자료 공간으로부터 2차원의

시각 공간으로의 투영으로 정의된다. 반면에 GTM 모형은 잠재 공간에서 자료 공간으로의 사상으로 정의된다. GTM은 자료 공간에서 잠재 공간으로 차원을 축소해 비선형 변환을 사용하는 비선형 잠재 변수 모형이다. 이 때, 모수의 추정은 EM (Expectation Maximization) 알고리즘을 사용한다. 결국 GTM은 모형에 대한 설명력이 부족한 SOM의 이론적 한계를 극복하고 좀 더 정확한 데이터의 군집화를 수행하게 된다.

3. Bayesian GTM

3.1 모형의 구조

D-차원 자료 벡터, d 는 L차원의 잠재 변수 벡터, l 과 잡음(noise) 벡터, e 를 이용하여 식 (1)과 같이 표현된다[2].

$$d = y(l, m) + e \quad (1)$$

여기서, w 는 모수 행렬(parameter matrix)이고 y 는 잠재 변수와 가중치의 함수로 표현된다. 각 자료는 베이즈 법칙(Bayes' rule)[3]에 의해 식 (2)와 같은 사후 확률식을 갖는다.

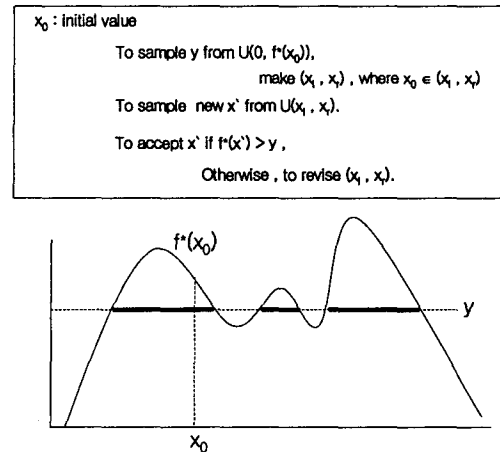
$$p(l|d, w) = \frac{p(d|l, w)p(l)}{p(d|w)} \quad (2)$$

이 사후 확률식으로부터 구체적인 확률값은 Slicing sampling[5]에 의해 추출된 표본을 이용한 Markov Chain Monte Carlo(MCMC) 적분[8]에 의해 구해진다.

3.2 Slice Sampling

Slice sampling은 Gibbs sampling, Metropolis algorithm과 유사한 MCMC 방법이다. 이 알고리즘은 균등 분포(uniform distribution)를 사용함으로써 Gibbs sampling이나 Metropolis algorithm에 보다 원하는 분포로부터 간편하게 표본들을 뽑을 수 있다. Slice sampling은 [그림 1]의 각 단계를 반복 수행한다.

Gibbs sampling이나 Metropolis algorithm과 같은 기존 방법들의 문제점은 임의 보행 현상(random walk behaviors)이 발생한다는 것이다. Slice sampling도 마찬가지로 임의 보행 현상을 피할 수는 없지만 기존의 방법들보다는 임의 보행 현



[그림 1] Slice sampling 알고리즘

상에서 많이 벗어날 수 있고 적용이 쉽다. 그러나 단계의 길이가 너무 작으면 슬라이스를 만들 때까지 불필요하게 많은 시간이 소요되고, 너무 크면 슬라이스를 수정하는데 불필요한 작업을 반복해야 하기 때문에 슬라이스의 크기 결정에 신중해야 한다. 본 논문의 베이저안 GTM에서 분류함수로 사용되는 사후 확률 분포의 최종 확률 계산을 위한 몬테 칼로 시뮬레이션에 사용되는 표본의 추출에 이 Slice sampling을 사용한다.

3.3 제안하는 Bayesian GTM 알고리즘

1) Initialize

1.1) Set data vectors.

$$d = (d_1, d_2, \dots, d_D)$$

1.2) Prior of weights.

$$w \sim f(\theta)$$

2) Determine Posterior probability

2.1) Mapping to Latent variable.

$$d = y(l, w) + e$$

2.2) Compute posterior probability distribution.

$$p(l|d, w) = \frac{p(d|l, w)p(l)}{p(d|w)}$$

3) Classification

3.1) Classification by posterior probability distribution.

this phase needs the slice sampling for MCMC simulation.

3.2) Replace until converge.

4. 실험 및 결과

본 논문의 Bayesian GTM의 패턴분류 실험을 위하여 균집의 결과가 알려져 있는 Fisher의 붓꽃 자료와 digit data를 이용하여 기존의 분류 기법으로 주로 사용되고 있는 CHAID(Chi), CART(Gini)와 알고리즘의 성능을 오분류율 측도로 비교하였다. 4장의 실험에서 사용한 데이터의 분포는 가우시안 분포(Gaussian distribution)를 사용하였다. 이는 많은 데이터가 가우시안 분포를 따르고 있고, 베이지안 추론에 대한 적용이 용이하면서도 정확도가 높기 때문이다.

4.1 Fisher의 붓꽃 자료를 이용한 실험

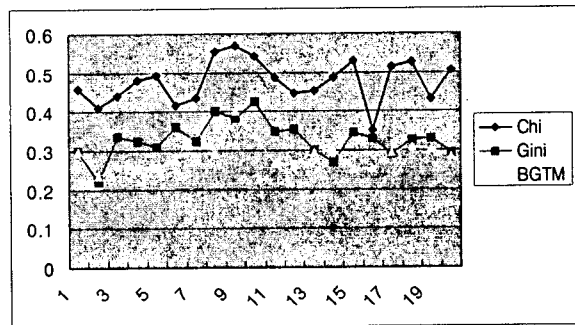
Bayesian GTM과 기존의 대표적인 분류 알고리즘들을 비교하기 위하여 Fisher의 Iris 자료 [9]를 이용하였다. 이 자료는 붓꽃의 특성치에 관한 자료로서 4개의 입력 변수를 갖는다.

<표 1> Iris data의 분류 결과

분류 방법 횟수	Chi	Gini	BGTM
1	0.088	0.074	0.049
2	0.074	0.056	0.027
3	0.069	0.052	0.007
4	0.066	0.066	0.072
5	0.085	0.051	0.024
6	0.029	0.029	0.005
7	0.183	0.050	0.040
8	0.083	0.067	0.001
9	0.015	0.045	0.005
10	0.094	0.063	0.053
11	0.086	0.052	0.007
12	0.046	0.062	0.021
13	0.070	0.018	0.008
14	0.098	0.066	0.039
15	0.095	0.079	0.022
16	0.062	0.015	0.005
17	0.167	0.074	0.064
18	0.051	0.051	0.001
19	0.095	0.054	0.051
20	0.136	0.061	0.017
평균	0.0846	0.0541	0.0257

입력 변수가 되는 꽃받침의 길이, 꽃받침의 넓이, 꽃잎의 길이, 꽃잎의 넓이에 따라 붓꽃의

종류를 Sectosa, Versicolor, Verginica의 세 집단으로 분류하였다. 이는 모델이 4개의 입력 변수와 1개의 목표 변수로 이루어짐을 의미한다. 초기 가중치는 가우시안 사전 분포로 하고, 베이즈의 규칙에 의해 갱신되는 사후 확률 분포 또한 사후 확률 분포가 된다. 그리고 각 개체의 분류를 위해 사후 확률 분포로부터 얻게 되는 최종 분류 예측 분포의 확률 계산은 Slice sampling을 사용하였으며, 이에 대한 실험 결과가 <표 1>과 [그림 2]에 나타나 있다.



[그림 2] Iris data 오분류율 비교

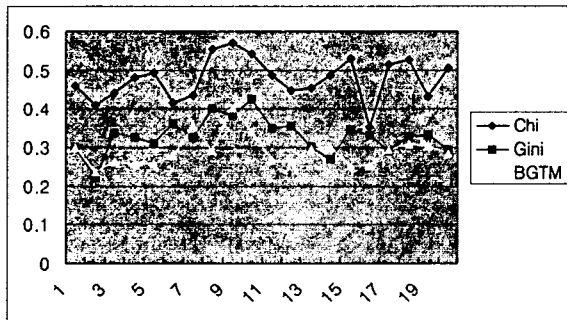
20번의 Iris data의 반복 실험 결과를 통해서 각 알고리즘의 오분류율의 평균을 비교해 보면 CHAID, CART, Bayesian GTM이 각각 0.0846, 0.0541, 0.0257로서 기존의 분류 알고리즘에 비해 본 논문에서 제한한 Bayesian GTM의 오분류율이 훨씬 작음을 알 수 있다. 이는 본 논문에서 설계한 베이지안 추론을 적용한 Bayesian GTM 모형이 기존의 다른 분류 알고리즘에 비해 좋은 분류 결과를 보임을 나타낸다.

4.2 Digit data를 이용한 실험

Digit data는 고장난 전자계산기로부터 얻어진 자료[9]로서 디지털 숫자의 각 위치를 나타내는 7개의 입력 변수와 0에서 9까지의 숫자를 인식하는 1개의 목표 변수로 이루어진다. 이 실험을 위한 Bayesian GTM 구조도 앞의 Iris data와 같은 구조를 갖는다. 이 digit data의 Bayesian GTM 실험 결과는 <표 2>와 [그림 3]에 나타나 있다. 앞의 Iris data의 결과와 마찬가지로 digit data의 실험에서도 Bayesian GTM 모형이 기존의 분류 알고리즘인 CHAID나 CART에 비해 평균 오분류율이 훨씬 작음을 알 수 있다. 즉, CHAID, CART, Bayesian GTM의 경우 각각 20번의 반복 실험에서 평균 오분류율은 0.4768, 0.3289, 0.2901로 나타났다.

<표 2> Digit data의 분류 결과

분류 방법 횟수	Chi	Gini	BGTM
1	0.460	0.299	0.300
2	0.411	0.219	0.237
3	0.442	0.338	0.263
4	0.482	0.325	0.258
5	0.493	0.310	0.256
6	0.415	0.362	0.279
7	0.434	0.325	0.355
8	0.553	0.400	0.292
9	0.570	0.384	0.402
10	0.541	0.425	0.340
11	0.488	0.350	0.266
12	0.448	0.356	0.263
13	0.455	0.299	0.298
14	0.487	0.269	0.300
15	0.529	0.345	0.251
16	0.352	0.330	0.253
17	0.513	0.290	0.293
18	0.526	0.329	0.310
19	0.433	0.330	0.303
20	0.506	0.294	0.285
평균	0.476	0.328	0.290



[그림 3] Digit data 오분류율 비교

5. 결론

1980년대 중반 Kohonen이 제안한 SOM 모형은 Unsupervised Learning에서 매우 많이 쓰이는 신경망 모형이다. 하지만 SOM은 모형을 정의하는 이론적 근거가 부족하였다. 이러한 단점을 보완하기 위해 Bishop은 1990년대 후반에 SOM의 확률 버전(probabilistic version)인 GTM을 제안하였다. 이 모형은 잠재변수를 사용하여 다차원의 자료를 저차원으로 사상

(mapping)시켜 자료를 군집화하는 알고리즘이다.

본 논문에서는 이러한 GTM 모형에 Bayesian 추론을 적용하여 분류 알고리즘으로 사용하였다. 실험을 통해 본 논문에서 제안한 베이저안 GTM모형은 기존의 다른 알고리즘에 비해 오분류율이 낮고 비교적 쉽게 적용할 수 있는 장점이 있음을 알 수 있었다.

향후, 본 논문의 Bayesian GTM을 범주형 데이터를 포함하는 모든 데이터 영역으로 확장, 발전시킨다면 각종 생체인식 시스템에도 적용되어 작은 오분류율로 패턴 인식 시스템을 관리할 수 있으리라 기대한다.

6. 참고문헌

- [1] Bishop, C. M. "Neural Networks for Pattern Recognition", Oxford University Press, 1995.
- [2] Bishop, C. M., Svensen, M., Williams, C. K. I. "GTM : a Principled Alternative to the Self Organizing Map", Proceedings 1996 International Conference on Artificial Neural Networks, ICANN 96, Volume1112, pages 165-170, Bochum, Germany, 1996.
- [3] Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. "Bayesian Data Analysis", Chapman & Hall, 1995.
- [4] Teuvo Kohonen, "Self-Organizing Maps", Springer, 1997.
- [5] Neal, R. M. "Bayesian Learning for Neural Networks", Springer, 1996.
- [6] Abhijit S. Pandya, Robert B. Macy. "Pattern Recognition with Neural Networks in C++", IEEE Press, 1995.
- [7] Ripley, B. D. "Pattern Recognition and Neural Networks", Cambridge University Press, 1996.
- [8] Martin A. Tanner. "Tools for Statistical inference", Springer, 1996.
- [9] UCI Machine Learning Repository, <http://www1.ics.uci.edu/~mllearn>