베이지안 신경망을 이용한 유전자 발현 데이터에서의 피처 추출 기법

이상근[0] 장병탁
서 울 대 학 교 컴 퓨 터 공 학 부
sklee[0]@bi.snu.ac.kr btzhang@cse.snu.ac.kr

Feature Extraction Method for Gene Expression Data using Bayesian Neural Network

Sang-Kyun Lee[0] and Byoung-Tak Zhang
School of Computer Science and Engineering, Seoul National University

요 약

Microarray 로 표현되는 유전자 발현 데이터는 일반적으로 샘플(sample) 수에 비해 많은 수의 유전자를 포함한다. 피처 추출은 이러한 데이터에 기계학습 방법론을 효과적으로 적용하기 위한 방법 중 하나로, 학습 성능을 향상시키고 계산 시간을 줄일 수 있을 뿐만 아니라 중요한 피처들을 발견할 수 있다는 점에서 큰 의미를 갖는다. 본 연구에서는 베이지안 신경망(Bayesian Neural Network)에 기반한 자동유효성탐지(Automatic Relevance Detection, ARD) 기법을 사용하여 유전자 발현 데이터에서 학습 오류를 줄이는 동시에 학습에 필요한 최소한의 유전자 집합을 추출할 수 있는 방법을 제시했다. CAMDA 2003 에서 제시된 폐종양 환자의 유전자 발현 데이터에 대해 실험한 결과, 12600 개의 유전자 중에서 가장 중요하다고 여겨지는 187 개의 유전자를 발견했으며, 높은 학습 성능을 달성했다.

1. 서 론

DNA칩 기술로 얻어지는 유전자 발현 데이터(gene expression data)는 생체 조직에 포함되어 있는 유전자의 발현량을 측정한 결과이다. 병리 현상을 나타내는 환자의 특정 생체 조직에 대한 유전자 발현 데이터를 이용하여 위험 환자군을 분류하는데 이용하려는 연구가 진행되어 왔으며, 나름대로의 성과를 거두고 있다[1]. 이러한 패턴 분류의 문제는 기계학습에서 다루는 문제와 일치하며, 따라서 신경망 등의 기계학습 기법이 이러한 문제에 적용되어 왔다.

그러나 일반적으로 유전자 발현 데이터는 많은 수의 유전자를 포함하는 반면 매우 적은 수의 샘플(sample)만을 포함하므로, 이른바 차원의 저주(curse of dimensionality)라는 문제를 안고 있다. 즉, 표본의 분류를 위한 탐색 공간(search space)의 차원이 매우 커서 효과적인 기계학습이 어려운 것이다.

피처 선택이란 이러한 탐색 공간에서 특정 샘플 패턴의 분류를 위해 유효한 피처만을 골라내어 탐색 공간의 차원을 줄여나가는 기법을 의미한다. 이 방법의 장점은 (1)계산 시간을 줄일 수 있고, (2)유효하지 않은 피처들에 의한 기계학습성능 저하를 방지할 수 있다는 점이다. 피처 선택을 위한 기법에는 (1)특정 기준에 따라 각 피처들을 정렬한 다음 일정한 범위 이내의 것들만을 선택하는 필터링(filtering)기법, (2)적용할 기계 학습 방법론의 오차를 최소로 하는 피처 집합을 찾아내는 랩퍼(wrapper) 기법으로 나눌 수 있다. 필터 기법의 경우 그 수행 시간은 빠르지만 유의미한 피처들을 무시하거나 반대로 무의미한 피처를 포함시킬 위험성이 있고, 랩퍼 기법의 경우에는 선택된 피처 집합에 의한 기계학습 성능은 어느 정도 보장되지만, 피처 선택에 요구되는 시간이 입력의 개수에 따라 지수적으로 증가하므로 실제 유전자 발현 데이터의 분석에 적용하기는 어렵다는 단점이 있다.

본 논문에서는 기계 학습 방법론중의 하나인 베이지안 신경망(Bayesian Neural Network, BNN)에 기반한 자동유효성탐지(Automatic Relevance Detection, ARD)기법을 이용하여, 2003년도 CAMDA(Critical Assessment of Microarray Data Analysis) 학회에서 제시된 선암종(adenocarcinoma) 폐종양 유전자 발현 데이터에 대해 피처 추출을 실시했다. 실험 결과, 총 12,600개의 유전자 중에서 기계 학습을 통해 위험 환자군을 성공적으로 식별하는데 필요한 187개의 유전자를 발견할 수 있었다.

2. 피처 선택 방법

2.1 베이지안 신경망(Bayesian Neural Network)

신경망(neural network)은 공학 응용 분야에서 자주 사용되는 모델로, 특히 본 연구에서는 다계층 퍼셉트론(multi-layer perceptron)[3]을 사용한다. 신경망은 실수 입력의 집합 x_i 를 받아들여 하나 이상의 출력값 $f_k(x)$ 를 계산하는 네트워크이며, 다계층 퍼셉트론의 경우 하나 이상의 히든 유닛(hidden unit)을 포함하는 여러 히든 레이어(hidden layer)를 포함하기도 한다. 일반적으로 그림1과 같이 하나의 히든 레이어를 포함하는 신경망이 주로 사용된다.

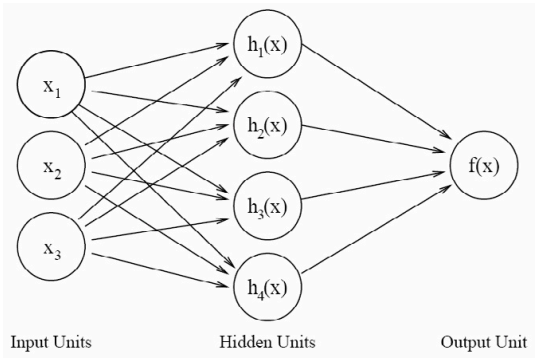

그림 1. 3개의 입력 유닛, 4개의 히든 유닛, 하나의 출력 유닛을 포함하는 다계층 퍼셉트론.

그림1의 네트워크에서 출력은 다음과 같은 수식으로 계산된다.

$$f_k = b_k + \sum_j v_{jk} h_j(x) \quad (1)$$

$$h_j = \tanh(a_j + \sum_i u_{ij} x_i) \quad (2)$$

u_{ij} - 입력 유닛 i로부터 히든 유닛 j로의 *weight*

v_{jk} - 히든 유닛 j로부터 출력 유닛 k로의 weight

a_j, b_k - 히든 유닛과 출력 유닛의 bias

즉, 각각의 출력 유닛은 bias와 히든 유닛 출력의 가중합(weighted sum)의 합을 표현하고, 각 히든 유닛은 bias와 입력 유닛값의 가중합의 합을 비선형 활성화 함수(activation function)에 통과시킨 값을 표현한다. 본 연구에서 사용한 활성화 함수는 hyperbolic tangent(tanh)로서, 이 함수는 차원이 높은 문제 공간에서 sigmoid에 부드러운(smooth)한 특성을 갖는다[3].

한편 관찰된 실험 데이터 $x^{(i)}$를 미지의 parameter θ를 갖는 모델이 얼마나 잘 표현하는가를 나타내는 likelihood function $L(\theta)$을 다음과 같이 정의할 수 있다.

$$L(\theta) = L(\theta \mid x^{(1)},...,x^{(n)}) \propto P(x^{(1)},...,x^{(n)} \mid \theta) \quad (3)$$

이제 관측치 $x^{(1)},...,x^{(n)}$에 대한 parameter θ의 사후확률분포(*posterior distribution*)를 다음과 같이 계산한다.

$$P(\theta \mid x^{(1)},...,x^{(n)}) = \frac{P(x^{(1)},...,x^{(n)} \mid \theta) P(\theta)}{P(x^{(1)},...,x^{(n)})} \propto L(\theta \mid x^{(1)},...,x^{(n)}) P(\theta) \quad (4)$$

이때 미지의 값 $x^{(n+1)}$의 예측치에 대한 예측확률분포(*predictive distribution*)는 다음과 같이 주어진다.

$$P(x^{(n+1)} \mid x^{(1)},...,x^{(n)}) = \int P(x^{(n+1)} \mid \theta) P(\theta \mid x^{(1)},...,x^{(n)}) d\theta \quad (5)$$

여기서 한가지 고려할 점은 식(4)에서 보듯 모델 parameter에 대한 사전확률분포(prior distribution), $P(\theta)$를 계산해야 한다는 점이다. 일반적으로 학습하고자 하는 문제의 구조에 대한 사전 정보가 없는 경우 적절한 $P(\theta)$를 정하는 일은 쉽지 않다. 이러한 문제점 때문에 신경망에 베이지안 추론을 적용하는 것이 부적절하다고 여겨져 왔으나, weight와 bias parameter에 대한 사전확률분포 $P(\theta)$의 표준편차를 hyperparameter로 대체하여 고려하는 모델을 적용하여 이 문제점을 극복하는 방안이 제시된 바 있다[3]. 새로운 모델에서는 각 parameter의 형식, 즉 (1)입력 유닛의 weight, (2)히든 유닛의 bias, (3)출력 유닛의 bias에 각각 hyperparameter를 설정하고, 예측 오차를 최소화하는 parameter를 탐색한다.

2.2 자동연관성탐지(Automatic Relevance Detection)

자동유효성탐지(ARD) 기법은 학습에 주어진 목표값을 예측하기 위해 많은 입력 가운데서 가장 유효한 입력만을 자동적으로 선택하기 위한 기법이다. ARD는 신경망의 한 입력 유닛에 연결된 링크들의 weight의 표준편차가 입력과 히든유닛 사이의 링크 weight의 hyperparameter와 어떤 분포를 이룬다는 가정으로부터 출발한다. 즉 입력 hyperparameter가 입력레이어와 히든레이어간의 high-level 사전 확률 분포를 표현한다면, ARD hyperparameter는 하나의 입력유닛과 히든레이어간의 low-level 사전 확률 분포를 표현한다. 본 연구에서는 MCMC (Markov chain monte carlo) 기법을 이용하여 이러한 2단계의 hyperparameter를 추정했다.

추정된 ARD hyperparameter를 보면 각 입력유닛에 연결된 링크 weight의 분포를 알 수 있으며, 입력유닛의 표준편차가 작은 경우 이 유닛에 연결된 링크의 weight값들도 작게 되므로 이러한 입력은 신경망 계산에 큰 영향을 주지 못할 것임을 예측할 수 있다. 반면 큰 표준편차를 갖는 입력은 신경망에 많은 영향을 주므로, 이러한 입력만을 골라내어 유효한 입력 집합을 구할 수 있게 된다.

3. 실 험

실험에는 CAMDA 2003에서 제시된 4개의 데이터셋 중 Harvard 연구팀의 것을 사용했다. 그 이유는 이 데이터셋이 Affymetrix 칩으로부터 얻어진 것으로 cDNA칩을 사용한 다른 데이터셋보다 포함된 노이즈가 작을 것으로 예상되었고, 6만여 개 정도의 많은 probe gene을 포함하는 HG-U95칩을 사용하여 그 신뢰성이 가장 높다고 판단되었기 때문이다.

이 데이터셋은 폐종양의 일종인 선암종(adenocarcinoma) 중 육안으로 식별이 어렵지만 매우 위험하다고 알려진 전이성 (metastatic) 선암종 환자를 구별해 내기 위해 제작된 마이크로어레이 데이터로서,

총 156명의 환자들의 폐 조직샘플에 대한 12600개의 유전자의 발현 데이터로 구성되어 있다.

3.1 실험 절차

실험 절차는 크게 다음의 3단계로 이루어진다. 우선, 전처리단계에서 각 발현량의 값이 0보다 큰 값을 갖도록 전체 데이터를 평행이동하고, 유전자 발현량이 너무 큰 값을 갖는 것을 보정하기 위해 모든 값을 log scale로 변환하였다. 다음 각 표본의 유전자 발현량이 평균이 0, 표준편차가 1이 되도록 Centering과 Normalization을 실시하였다.

다음으로, 후보피처선택 단계에서는 χ^2 통계량과 PFA (principal feature analysis)[2]를 사용하여 얻어진 두 개의 피처 집합을 union하여 1,738개의 최종 후보피처집합을 생성했다.

마지막으로 BNN 과 ARD 기법을 사용하여 최종피처 집합을 선택하였다.

3.2 실험 결과

앞에서 생성한 1,738개의 최종후보피처들을 입력으로 사용하여 10개의 히든유닛과 1개의 출력유닛을 갖는 베이지안 신경망을 구성하였다. 또한 입력과 히든레이어 사이에는 high-level과 low-level hyperparameter를 두어 ARD의 적용이 가능하도록 하였다. 실험에는 토론토 대학의 베이지안 모델링 소프트웨어를 사용했다[4]. 사용한 파라미터는 다음과 같다.

Parameter	Value
Hyperparameter sampling method	MCMC
Length of Markov chain	1000
MCMC iteration	100

그림 2에 100회의 hypertransition에 대한 high-level hyperparameter, train error, test error, 그리고 low-level hyperparameter의 평균값의 추이가 나타나 있다. Test error를 보면, 76번째 transition에서 가장 작은 에러를 갖는 신경망을 찾았음을 알 수 있다.

76번째 신경망이 피처선택을 위해 가장 좋은 신경망인가를 판별하기 위해서는 high-level과 low-level hyperparameter의 추이를 함께 살펴보아야 한다. 즉, 두 hyperparameter가 높을수록 특정 입력에 weight가 몰려 있을 가능성이 높으므로, low-level hyperparameter의 값은 크지만 high-level은 그렇지 않은 76번째 신경망이 꼭 좋은 결과라는 보장은 없는 것이다.

감사의 글

본 연구는 과학기술부의 국가지정연구실 사업(NRL)과 Systems Biology 사업에 의해 지원되었음.

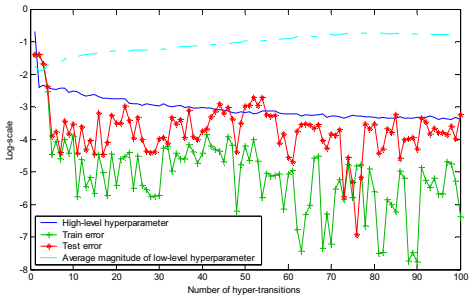

그림 2. Log-scale plot of high-level hyperparameter, train error, test error, and average squared magnitude of low-level(ARD) hyperparameters (100 iterations)

따라서 50번째 transition을 기준으로 그래프의 앞쪽과 뒤쪽에서 각각 test error가 최소인 17번과 76번 신경망을 후보로 추출하였다. Hyperparameter 값의 크기 순서대로 피처집합의 크기를 바꾸어 가며 신경망으로 실험한 결과, 17번 베이지안 신경망에서 추출한 187개의 유전자를 입력으로 사용하여 15개의 히든유닛을 가진 신경망을 구성할 경우, 모든 유전자 발현 샘플에 대해 test error가 1% 미만이 됨을 확인하였다. 이와는 대조적으로 모든 피처를 사용하는 경우에는 80개 이상의 히든유닛을 사용하는 신경망을 구성해야만 비슷한 성능을 얻을 수 있었다.

4. 결 론

본 논문에서는 베이지안 신경망과 MCMC 샘플링 기법, ARD 기법을 사용하여 총 12,600개의 유전자 중에서 선암종 폐종양 위험군 환자를 정확히 구분해 낼 수 있는 187개의 중요한 유전자를 추출할 수 있었다. 이렇게 추출된 유전자들은 통계학적으로 의미가 있을 뿐만 아니라, 앞으로 중요한 생물학적 연구 대상으로도 활용될 수 있을 것으로 기대된다.

참고문헌

[1] A. Bhattacharjee, W. G. Richards, J. Staunton J, C. Li , S. Monti , P. Vasa, C. Ladd, J. Beheshti, R. Bueno, M. Gillette, M. Loda, G. Weber, E. J. Mark, E. S. Lander, W. Wong, B. E. Johnson, T. R. Golub, D. J. Sugarbaker, M. Meyerson, "Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses", *In Proc Natl Acad Sci USA*, 98(24):13790-5, Nov, 2001.

[2] I. Cohen, Q. Tian, X. S. Zhou, and T. S. Huang, "Feature Selection Using Principal Feature Analysis", *ICIP'02*, 2002.

[3] R. M. Neal, Bayesian Learning for Neural Networks, Springer-Verlag, 1996.

[4] R. M. Neal, Software for Flexible Bayesian Modeling and Markov Chain Sampling, http://www.cs.toronto.edu/~radford/fbm.software.html