진화 신경망을 이용한 DNA Microarray 데이터 분석

김경중° 조성배
연세대학교 컴퓨터과학과
uribyul@candy.yonsei.ac.kr, sbcho@cs.yonsei.ac.kr

Analysis of DNA Microarray Data Using Evolutionary Neural Networks

Kyung-Joong Kim° Sung-Bae Cho
Dept. of Computer Science, Yonsei University

요 약

DNA Microarray 기술은 유전자의 발현여부를 매우 빠르게 검사할 수 있는 도구이며 각종 질병의 발생 여부를 예측하기 위한 정보를 제공한다. 유전자 발현 데이터로부터 암의 발생 여부를 예측하기 위해서는 기존의 접근방법과 다른 기계학습 기법이 요구된다. 일반적으로 샘플의 개수가 극히 적은 반면에 특징의 개수는 수천에서 수만 개가 존재하기 때문에 문제의 특성에 맞는 분류기의 구조를 결정하는 것이 매우 어려운 일이기 때문이다. 진화 신경망은 신경망의 구조와 가중치를 동시에 학습하며 사용자는 각 개체의 적합도를 평가할 수 있는 방법만 제공해 주면된다. 특히 신경망의 구조를 사전에 고정하지 않아도 되는 장점이 있기 때문에 전문적인 지식이 없는 사용자라도 이용가능하다. 대장암 데이터에 대한 실험결과 제안하는 분류기 모델이 다층 퍼셉트론, SVM (support vector machine), 최근접 이웃 방법에 비해 향상된 성능을 보였다.

1. 서 론

질병의 발생여부를 판단하기 위해서 의사들은 많은 종류의 척도를 사용한다. 유전자 발현정보도 그 중의 하나로 단백질의 생성이 정상적으로 이루어지고 있는지를 판단할 수 있는 중요한 정보이다. 최근 DNA 마이크로 어레이 기술의 등장은 매우 빠른 속도로 유전자 발현정도를 측정하는 것을 가능하게 했다[1]. 그러나 수천 개의 유전자 발현 정보로부터 암의 발생여부를 예측하는 것은 기계학습 기법의 도움 없이는 불가능한 작업이다.

의사는 기계학습의 도움을 받아 암의 발생여부 예측에 유용한 유전자를 구분해 낼 수 있으며 분류모델을 설계할 수 있다. 하지만 대부분의 분류기 모델은 많은 파라미터를 가지고 있으며 전문지식을 바탕으로 결정해 주어야 하는 부분이 많이 있다. 일반적으로 의사는 문제에 대한 충분한 지식을 지니고 있지만 분류기를 문제에 적용하는 방법에 대해서는 생소할 수밖에 없다.

자연은 문제에 대한 해를 찾기 위해 오랜 시간을 거쳐 변형을 만들어 내고 무수한 실험을 수행한다. 이러한 과정은 진화이론으로 설명이 되고 있으며 컴퓨터과학자들은 이것을 모방하여 실제적인 문제를 해결하는 방법을 개발했다. 분류기의 구조를 설계하는 문제에도 진화 알고리즘은 적용이 가능하며 사용자가 구조 결정에 대한 지식이 없더라도 해를 찾을 수 있다. 이러한 연구 흐름에 맞추어 가장 활발히 이루어지고 있는 연구가 진화신경망이다[2].

진화 알고리즘은 신경망의 가중치, 구조, 학습 규칙을 결정하기 위해 사용되며 한번에 두 가지 이상의 요소를 동시에 학습하기도 한다. 본 논문에서는 신경망의 구조와 가중치를 동시에 학습하는 방법을 사용하여 대장암 데이터의 분류 문제를 해결한다. 대장암은 서구사회에서 암으로 인한 사망의 두 번째 흔한 원인으로 알려져 있다.

2. DNA Microarray 기술

인간의 유전정보는 병의 원인을 이해하기 위해 매우 중요한 역할을 한다. 주변의 상황과 발병 상황에 따라 각 유전자의 발현정도가 달라지며 이러한 정보는 병의 진단 및 치료에 사용될 수 있다. 인간의 유전자 정보는 수천 개에 달하기 때문에 기존의 수작업을 통해 일일이 발현정도를 확인하기에는 한계가 있다. 그림 1은 DNA microarray 장비와 이미지 스캐너로 읽어 들인 발현정보를 보여준다. 환자 한명마다 7129개의 유전자에 대한 발현 정도를 색깔로 표현해 준다.

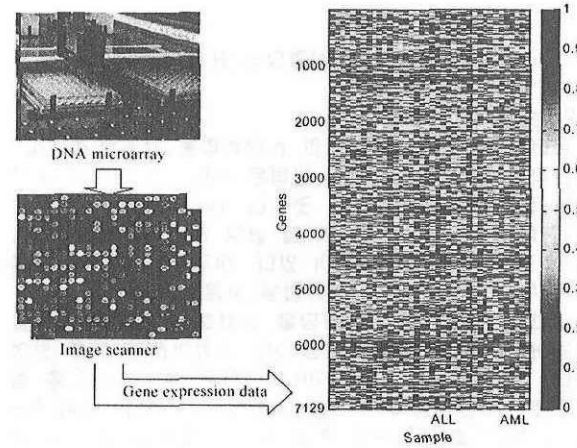


그림 1. DNA Microarray를 이용한 발현정보 획득

3. 진화 신경망

유전자 발현정보를 분석하기 위해서 다양한 종류의 분류기가 사용되어 왔다. 대표적인 것으로 신경망, 최근접 이웃 방법, SVM(support vector machine) 등이 있으며 적은 수의 샘플에서 좋은 성능을 내기 위해 다양한 기법이 제안되어 왔다. 이러한 방법의 공통점은 분류기의 구조에 대해 사용자가 미리 결정을 해주어야 한다는 점이다. 진화 신경망은 구조에 대한 제약을 두지 않기 때문에 이러한 문제점을 극복할 수 있다.

진화 신경망 알고리즘은 일반적인 유전자 알고리즘을 사용한다. 초기에 신경망 집단을 생성한 후 학습 데이터를 사용하여 부분학습을 한다. 검증 데이터를 사용하여 신경망의 적합도를 측정한 후 선택, 교차, 돌연변이를 수행한다. 새로운 집단에 만족스러운 해가 있을 경우나 일정한 세대를 반복한 후에 알고리즘을 종료한다. 만약 조건을 만족하지 못한다면 부분학습을 다시 수행하고 위의 과정을 반복한다. 부분학습을 위해서 오류역전파 알고리즘을 사용하며 일반적으로 사용하는 반복횟수보다 적은 설정을 사용한다. 부분학습을 함으로써 유전자 알고리즘의 탐색반경을 줄일 수 있다.

3.1 표현

그림 2는 신경망의 행렬기반 표현방법을 보여준다. 왼쪽 신경망은 4개의 노드를 지니고 있으며 4개의 연결을 가지고 있다. 노드의 개수를 N이라고 하면 $N \times N$의 행렬을 사용한다. 행렬 M은 신경망 구조의 가중치와 구조를 동시에 나타낸다.

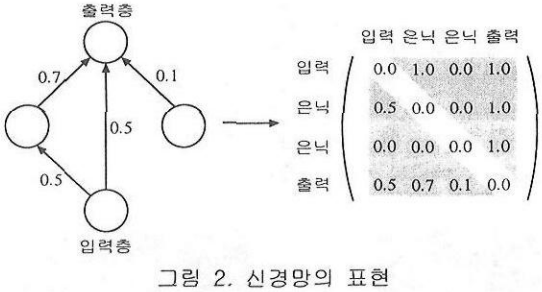


그림 2. 신경망의 표현

3.2 선택, 교차, 돌연변이

유전자 알고리즘은 다양한 선택방법을 제공하고 있다. 일반적으로 많이 사용되는 방법은 적합도의 크기에 비례하여 선택확률을 부여하는 것이다. 이 방법은 개체 사이의 적합도 차이가 지나치게 클 경우 하나의 해로만 빠르게 수렴해 버리는 문제점이 있다. 이러한 문제점을 해결하기 위해 순위기반 선택 방법을 사용했다.

교차연산은 두개의 신경망을 교차하여 새로운 자손을 만들어 내는 방법이다. 그림 3은 교차연산의 예를 보여준다. 임의의 은닉노드를 하나 선택한 후 그 노드를 중심으로 두 신경망의 구조를 교환하는 것이다. 그림 3에서 회색노드를 중심으로 연결 구조가 교환된 것을 볼 수 있다.

돌연변이 연산은 추가/삭제의 두 가지 형태로 이루어진다. 임의로 선택한 연결이 이미 존재하는 경우 그 연결은 삭제된다. 만약 연결이 존재하지 않는 경우 새로운 연결을 생성하고 가중치를 0과 1사이의 임의의 실수로 결정한다.

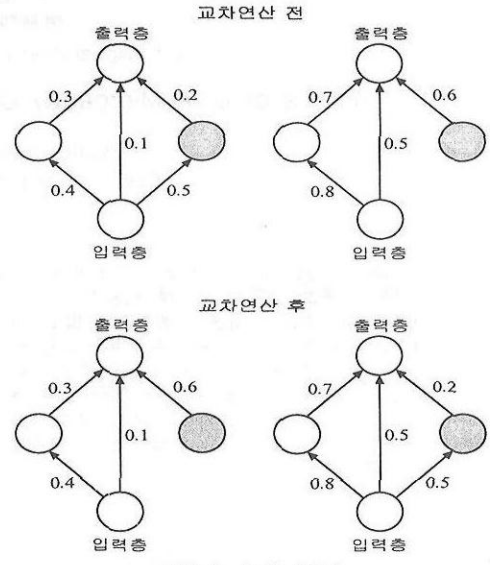


그림 3. 교차 연산

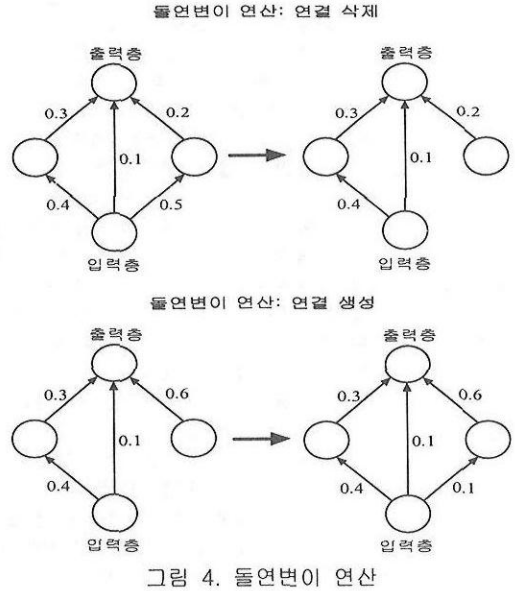


그림 4. 돌연변이 연산

4. 실험결과 및 분석

대장암 데이터는 62개의 샘플로 이루어져 있으며 각 샘플은 2000개의 유전자에 대한 발현정도를 가지고 있다. 62개의 샘플 중에서 40개는 대장암 세포이고 나머지는 정상 세포이다[3]. 31개의 샘플이 학습 데이터로 사용되었으며 나머지는 테스트 데이터로 사용되었다. 2000개의 유전자를 모두 사용할 경우 학습에 너무 많은 시간이 걸리고 성능도 낮아지기 때문에 중요한 유전자를 선

택하는 과정을 거친다. 몇 개의 유전자를 선택할 것인가는 어려운 문제인데 일반적으로 20개에서 40개의 유전자가 적절하다고 알려져 있다. 본 논문에서는 30개의 유전자를 사용하였다. 유전자의 선택을 위해서 정보이득 기준을 사용하였다. 아래의 수식은 정보이득을 계산하는 방법을 설명해 준다. C는 클래스의 개수를 의미하며 $P(G_i, C_j)$는 샘플의 클래스가 C_j인 경우에 유전자 G_i가 일정한 값을 넘을 확률이다.

$$IG(G_i) = \sum_{j=1}^{C}\left\{P(G_i, C_j)log\frac{P(G_i, C_j)}{P(C_j)P(G_i)} + P(\overline{G}_i, C_j)log\frac{P(\overline{G}_i, C_j)}{P(C_j)P(\overline{G}_i)}\right\}$$

표 1은 진화 알고리즘의 변수를 보여준다. 실험은 10번을 반복한 후 평균값을 내었다. 데이터의 수가 매우 적기 때문에 검증데이터를 따로 사용하는 것은 불가능하고 테스트 데이터를 이용하였다. 그림 5는 테스트 데이터에 대한 분류성능을 보여준다. 가장 좋은 성능을 보인 진화 신경망의 구조는 매우 복잡했다. 그림 6은 진화 신경망의 구조를 보여준다. 전체구조가 복잡하기 때문에 노드 종류별로 연결 구조를 나누어 보았다. 입력 노드(유전자)와 출력 사이의 연결 강도를 분석해 보면 각 유전자가 대장암에 미치는 영향을 분석해 볼 수 있다.

표 1. 진화 알고리즘의 변수

집단의 크기	20
선택 확률	0.5
교차 확률	0.1
돌연변이 확률	0.3
세대의 수	200

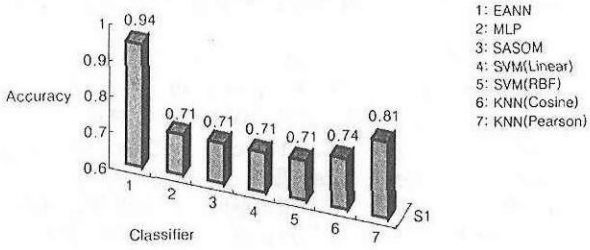
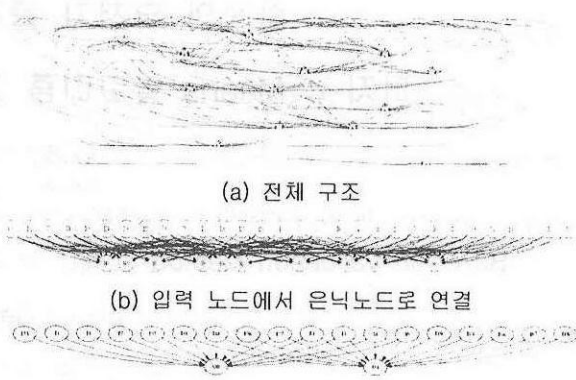


그림 5. 다양한 분류기와의 성능비교

5. 결론 및 향후연구

유전자 발현정보를 분석하기 위해서 진화 신경망을 제안하였다. 기존의 기계학습 기법들은 구조를 사전에 결정해야 하고 전문적인 지식을 요구한다는 단점을 가지고 있다. 이러한 문제를 해결하기 위해 진화 알고리즘을 이용하여 구조와 가중치를 동시에 결정하는 진화 신경망을 제안하였다. 대장암 데이터에 대한 실험결과 제안하는 방법이 사람이 설계하기 어려운 복잡한 구조를 생성해내었다. 향후연구는 신경망의 구조를 분석하는 방법을 통해 유전자가 암에 미치는 영향을 분석해 보는 것이다.

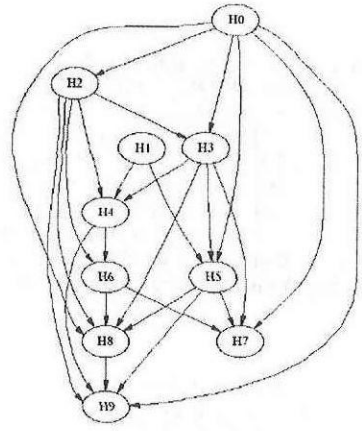
감사의 글
본 연구는 과학기술부의 뇌과학 과제의 지원을 받음.



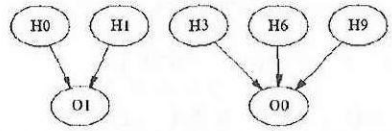
(a) 전체 구조

(b) 입력 노드에서 은닉노드로 연결

(c) 입력 노드에서 출력 노드로 연결



(d) 은닉 노드 사이의 연결



(e) 은닉 노드와 출력 노드 사이의 연결

그림 6. 가장 좋은 성능을 보인 진화 신경망 구조

참고문헌

[1] C. A. Harrington, C. Rosenow and J. Retief, "Monitoring gene expression using DNA microarrays," *Current Opinion in Microbiology*, vol. 3, pp. 285-291, 2000.

[2] X. Yao, "Evolving artificial neural networks," *Proceedings of the IEEE*, vol. 87, no. 9, pp. 1423-1447, Sep 1999.

[3] U. Alon, *et al.*, "Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays," *Proc. Natl. Acad. Sci. USA*, vol. 96, pp. 6745-6750, June 1999.