신경망 분류기를 이용한 암 관련 유전자 발현정보의 분류

권영준, 류중원, 조성배
연세대학교 컴퓨터과학과
{shining, rjungwon}@candy.yonsei.ac.kr, sbcho@csai.yonsei.ac.kr

Classification of Cancer-related Gene Expression Data Using Neural Network Classifiers

Youngjun Kwon° Jungwon Ryu Sung-Bae Cho
Dept. of Computer Science, Yonsei University

요 약

최근 생물 유전자 정보를 효과적으로 분석하기 위한 적절한 도구의 필요성이 대두되고 있다. 본 논문에서는 백혈병 환자의 골수로부터 얻어낸 DNA Microarray 유전 정보를 분류하여 환자가 가지고 있는 암의 종류를 예측하기 위한 최적의 특징추출방법과 분류 방법을 찾고자 한다. 이를 위해 피어슨 상관관계, 유클리디안 거리, 코사인 계수, 스피어맨 상관관계, 정보 이득, 상호 정보, 신호 대 잡음비의 7가지 특징 추출 방법을 사용하였으며, 역전파 신경망, 의사결정 트리, 구조 적응형 자기구성 지도, k-최근접 이웃 등 4가지의 기계학습 분류기를 이용하여 분류 실험을 하였다. 실험결과, 피어슨 상관관계와 역전파 신경망을 이용한 분류 방법이 97.1%의 인식률을 보임을 알 수 있었다.

1. 서론

생명 공학과 분석화학의 발달은 생물의 유전자 정보를 대량으로 얻어내는 것을 가능하게 하였다. 그러나 이렇게 얻어진 정보는 단순한 숫자의 나열이므로 이를 분석하여 의미 있는 정보를 뽑아내는 연구와 생물의 방대한 유전자 정보의 분석을 지원해 주는 효과적인 분석도구의 필요성이 대두되고 있다.

본 논문에서는 대량의 인간 유전자들 중 분류 작업과 큰 관련성을 갖는 유전자를 선별하기 위한 다양한 특징 추출 방법들과 몇 가지 신경망 분류기들을 이용하여 선별된 유전자들의 발현 정도를 보고 질병의 종류를 효과적으로 분류하는 방법에 대해 소개하고자 한다.

2. DNA Microarray

DNA Microarray는 용액이 투과하지 않는 고형 지지체 위에 고밀도로 DNA를 고정해 놓은 보합용 "probe array"이다. 기존에 유전공학에서 사용하던 방법들과는 달리 한 개의 DNA Microarray 위에는 최소한 수백 개 이상의 유전자가 놓여지므로, 다수의 유전자 발현변이 분석을 빠른 시간에 해결 할 수 있다.

두 개의 다른 환경에서 채집된 유전물질에 각각 다른 색깔의 형광물질(빨간색(Cy5)과 녹색(Cy3))을 합성한 것을 똑같은 양만큼 보합한 것들이 DNA Microarray의 각 셀을 이룬다. 그러므로 이것을 레이저 형광 스캐너로 읽어들이면 서로 다른 두 개의 환경 중 어떤 환경에서 유전물질이 많이 발현했느냐의 정도에 따라 빨간색에서부터 녹색 사이의 다양한 색상을 갖는 점들의 집합을 얻을 수 있으며, 이를 이용하여 시각적으로 수천 개의 유전 물질의 발현변이 정도를 한번에 확인 할 수 있다. 또한 두 가지의 색상이 발현되는 정도를 각각 양수와 음수로 하여 수치화 할 수 있으므로 컴퓨터를 이용한 분석이 가능하다.

3. 유전자 발현정보 분류시스템

본 논문에서 사용된 데이터는 72명의 백혈병 환자로부터 얻어진 골수 샘플로부터 제작된 DNA Microarray 데이터를 수치화 한 것이다. 급성 림프구성 백혈병(ALL, acute lymphoblastic leukemia)과 급성 골수성 백혈병(AML, acute myeloid leukemia)의 두 가지 클래스로 분류된다[1].

본 논문에서는 유전자 발현 정보를 이용하여 환자가 가지고 있는 병이 ALL, AML 중 어느 것인지를 분류하는 실험을 하였다. 이를 위한 시스템은 크게 특징 추출 단계와 패턴분류 단계의 두 단계로 나뉘어 진다(그림 1). 특징 추출 단계에서는 유전자 발현 정보를 분류하기 위해 7129개의 유전자로부터 분류에 관련되어있다고 생각되는 유전자 25개를 선택한다.

패턴분류 단계에서는 선택된 25개의 유전자들로 구성되어있는 38개의 학습데이터와 34개의 테스트 데이터를 몇 가지 분류기법에 적용하여 분류하였다.

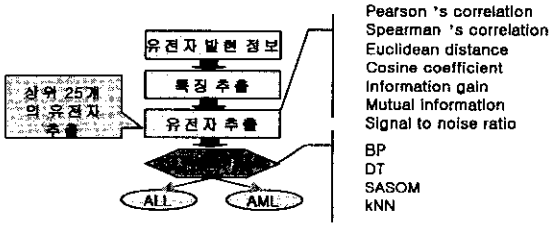

그림 1. 유전자 발현정보 분류시스템

3.1 특징 추출

(1) 상관관계법

상관관계분석을 이용하면 두 변수간의 선형적 관련성 정도와 관련 방향을 알 수 있다. 상관관계 분석의 의해 얻어지는 상관계수 r은 -1에서 +1사이의 값을 가진다. 즉, 측정된 자료가 좌표 상에 양(+)의 기울기를 갖는 직선에 분포한다면 그때의 상관계수는 양수가 되고, 음(-)의 기울기를 갖는 직선에 가깝게 분포하면 음의 상관계수를 갖는다. 본 논문에서는 학습 데이터의 각 유전자의 발현 수치와 각 샘플이 속하는 클래스를 반영한 클래스 패턴과의 상관관계를 계산하여 특징 추출의 척도로 사용하였다. 사용된 상관관계법은 피어슨 상관관계(Pearson's Correlation)와 스피어맨 상관관계 (Spearman's Correlation)이며, N개의 원소를 갖는 두 벡터 X와 Y사이의 각각의 상관 계수 $r_{pearson}$과 $r_{sparman}$은 다음과 같이 정의된다.

$$r_{pearson} = \frac{\sum XY - \frac{\sum X \sum Y}{N}}{\sqrt{\left(\sum X^2 - \frac{(\sum X)^2}{N}\right)\left(\sum Y^2 - \frac{(\sum Y)^2}{N}\right)}}$$

$$r_{spearman} = 1 - \frac{6\sum (Dx - Dy)^2}{N(N^2-1)}$$

(단, Dx, Dy는 각각 X와 Y의 순위배열)

(2) 유사도 측정법

두 개의 입력 벡터 X와 Y 사이의 유사성은 이들 변수간의 거리로 볼 수 있다. 거리는 두 대상이 얼마나 멀리 떨어져 있는가에 대한 척도이고, 유사성은 근접성의 척도이다. 군집분석에서는 이들 개념들에 의해 케이스들을 집단화할 수 있다. 본 논문에서 사용한 거리 지수는 유클리디안 거리방법($r_{euclidean}$)과 코사인 계수방법(r_{cosine})이며, 두 벡터 X와 Y 사이의 거리는 다음과 같이 정의된다.

$$r_{euclidean} = \sqrt{\sum (X-Y)^2}, \quad r_{cosine} = \frac{\sum XY}{\sqrt{\sum X^2 \sum Y^2}}$$

(3) 정보이론

전체 데이터로부터 의미 있는 정보를 뽑아내는 척도로 정보이론에서 사용하는 정보 이득(IG, Information Gain), 상호 정보(MI, Mutual Information), 신호 대 잡음 비(Signal to noise ratio) 등의 방법을 응용하였다. 즉, 정보 이득과 상호 정보의 경우는, 특정 유전자의 i번째 샘플이 특정 클래스 c에 속하는 가의 여부와 그 유전자가 발현했는가 여부(유전자 발현 정도가 양수인가 음수인가의 여부)의 두 가지 기준에 의해 네 가지 종류로 구분 짓고, 각 종류에 속하는 샘플의 수를 각각 A, B, C, D 라 했을 때, 주어진 유전자 g의 정보 이득과 상호 정보 각각의 계수는 다음과 같다[2].

$$IG = A \cdot \log \frac{A}{(A+B)\cdot(A+C)} + B \cdot \log \frac{B}{(A+B)\cdot(B+D)}$$

$$MI = \log \frac{A}{(A+B)\cdot(A+C)}$$

한 편, 38개의 샘플에 대해 주어진 유전자 g 를 클래스 c 에 속하는 것들과 그렇지 않은 것들로 분류한 후, 각각에 대해 정규분포 $(\mu_1(g), \sigma_1(g))$ 와 $(\mu_2(g), \sigma_2(g))$ 를 계산했을 때, 클래스 c 에 의해 분류되는 유전자 g 의 신호 대 잡음비 $P(g,c)$ 는 다음과 같이 계산된다[1].

$$P(g,c) \approx \frac{\mu_1(g) - \mu_2(g)}{\sigma_1(g) - \sigma_2(g)}$$

3.2 분류기

(1) 역전파 신경망

역전파(BP, Backpropagation) 신경망은 매우 다양한 분야에 활용되고 있는 주요한 순방향 다층 신경망이다. 이것은 입력 벡터가 나열되어있는 예제를 통해 반복적으로 출력을 각각의 입력에 따라 수정해가며 학습한다. 각각의 학습 입력을 거치는 것이 하나 주기로 하여 각 주기동안 신경망은 실제의 결과를 가지고 목적하는 결과와 비교해서, 오류를 산출한 뒤, 그 오류를 최소화하기 위해 가중치 값을 조절한다. 지도 학습이라 불리는 이 과정을 통해, 신경망은 올바른 결과와 입력 패턴을 연관시키는 것을 학습하게 된다.

(2) 의사결정 트리

의사결정 트리(DT, Decision Tree) 는 여러 단계의 복잡한 조건을 갖는 문제에서 각 조건과 그에 따른 해결방안을 트리 형태로 나타낸 것을 말한다. 가장 큰 조건이 트리의 뿌리를 만들고, 세부 조건이 트리의 각 가지를 만들며, 해결방안은 트리의 잎(leaf) 노드로 나타나게 된다.

(3) 구조적응 자기구성 지도

구조적응 자기구성 지도(SASOM, Structure Adaptive Self-organizing Map)는 일반적인 자기구성 지도의 구조가 초기에 결정되어 학습이 끝날 때까지 변하지 않는 다는 단점을 보완하기 위해 제안되었다[3]. 즉, 기존의 SOM 알고리즘을 이용하여 지도를 학습시킨 후, 학습된 지도의 노드들 중 서로 다른 클래스의 데이터가 섞여있는 노드를 반복적으로 분화하여 주어진 데이터에 대하여 최적의 위상을 갖는 지도를 생성한다.

(4) k-최근접 이웃

k-최근접 이웃(kNN, k-Nearest Neighbor) 기법은 기억기반추론의 가장 일반적인 기법으로 새로운 데이터에 대한 각각의 모든 기존 데이터와의 유클리디안 거리를 계산한 후 가장 가까이 있는 k개의 데이터에 기반 하여 새로운 데이터의 부류를 결정한다.

4. 실험 및 결과

4.1 실험 환경

72 개의 샘플 중 38개는 신경망을 학습시키기 위한 학습 데이터로 쓰였고, 다른 34개는 학습된 신경망의 성능을 평가하기 위한 테스트 데이터로 쓰였다.

38개의 학습 데이터 중 27개는 ALL을 앓고 있는 환자의 것이고, 11개는 AML을 앓고 있는 환자로부터 얻어진 것이다. 신경망을 학습시키기 위해 ALL에 속하는 샘플들을 클래스 0, AML 에 속하는 샘플들을 클래스 1로 표기하였다. 각 샘플은 수치화 된 7,129개 유전자들의 유전자 발현 정보로 구성되어 있다.

역전파 신경망 분류기를 이용한 분류 실험에서는 은닉층의 노드 수를 5에서 15사이의 값으로 취하였고, 학습률은 0.03~0.50으로, 모멘텀은 0.10~0.90으로 변화시켜가며 실험을 한 후 가장 좋은 결과를 정리하였다. k-최근접 이웃 방법을 이용한 분류 실험에서는 k 값을 1부터 38 까지 변화 시켜 가며 실험을 하였다.

4.2 결과 분석

분류 방법별 인식률에서는, 역전파 신경망이 평균적으로 가장 우수한 결과를 보여주었다. 의사결정 트리의 경우 여러 형태의 분류방법에 대해 고르지 못한 결과를 보였다. 구조 적응형 자기구성 지도를 이용한 분류 방법은 모든 특징 추출 방법들에 대해 비교적 고른 결과를 유지하였다. k-최근접 이웃 방법을 이용하여 분류 실험을 했을 때의 경우, 모든 특징 추출 방법에 대해 58.8% 의 동일한 최대 적중률을 보였다. 이는 34개의 테스트 데이터 모두에 대해 ALL 클래스에 속한다는 판단을 내렸을 때의 적중률이므로, 각 특징 추출 방법들간의 성능비교를 위해 k 값을 변화시켜가며 구한 결과의 평균값을 구하여 정리하였다.

특징 추출 방법별로는 피어슨 상관관계 법과 신호 대 잡음 비 방법이 가장 우수한 결과를 보였다(표 1).

표 1. 분류 및 특징 추출 방법별 인식률

분류방법 / 특징 추출방법	BP	DT	SASOM	kNN
피어슨 상관관계	97.1%	97.1%	88.2%	25.4%
스피어맨 상관관계	70.6%	82.4%	82.4%	44.0%
유클리디안 거리	97.1%	82.4%	82.4%	44.0%
코사인 계수	70.6%	73.5%	70.6%	35.7%
정보 이득	88.2%	47.1%	64.7%	58.5%
상호정보	67.7%	55.9%	64.7%	58.8%
신호 대 잡음비	94.1%	91.2%	94.1%	20.0%

그림 2는 피어슨 상관관계 계수에 의해 정렬된 전체 유전자 중, 계수 값이 가장 큰 쪽의 25개와 가장 작은 쪽의 25개의 유전자 표현 수준 값을 도표화 한 것이다. ALL 인 쪽과 AML 인 쪽의 색상이 분명하게 나뉨을 알 수 있다. 그림 3은 이 때 선택된 유전자 중 큰 쪽의 25개 유전자의 목록이다.

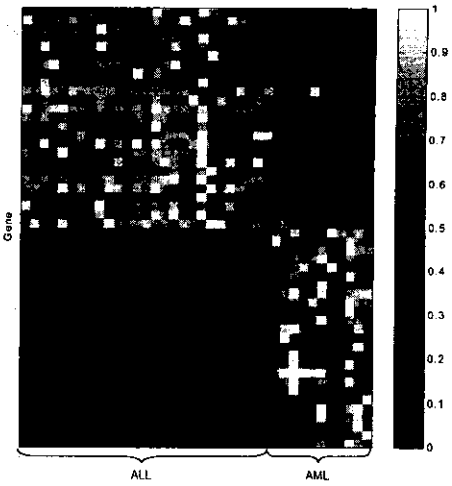

그림 2. $r_{pearson}$값에 의해 선택된 유전자의 표현 수준

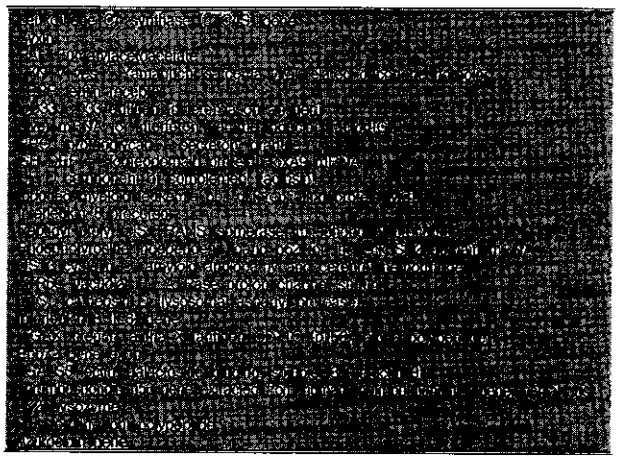

그림 3. $r_{pearson}$ 값에 의해 선택된 유전자들

참고 문헌

[1] T. R. Golub *et al,*. "Molecular classification of cancer: Class discovery and class prediction by gene-expression monitoring", *Science*, vol. 286, p. 531~537, 1999.

[2] F. Sebastiani, "Machine learning in automated text categorisation: A survey," *Technical Report IEI-B4-31-1999*, Istituto di Elaborazione dell'Informazione, C. N. R., Pisa, 1999.

[3] 김현돈, 조성배, "비교사 학습과 교사 학습 알고리즘을 결합한 구조 적응형 자기구성 지도", '99 *추계 정보과학회*, 서울, 1999.