암의 분류를 위한 음의 상관관계 유전자의 신경망 쌍

원홍희° 조성배
연세대학교 컴퓨터과학과
cool@candy.yonsei.ac.kr° sbcho@cs.yonsei.ac.kr

Neural Network Pair with Negatively Correlated Genes for Cancer Classification

Hong-Hee Won° and Sung-Bae Cho
Dept. of Computer Science, Yonsei University

요 약

정확한 암의 분류는 암의 진단 및 치료에 있어 매우 중요하지만, 암을 진단하기 위한 기존의 여러 방법들은 종종 불완전한 결과를 도출한다. 최근의 마이크로어레이 기술에 기반한 분자 수준의 진단은 정확하고 객관적이며 체계적인 암의 분류를 위한 방법론을 제시해준다. 유전자 발현 데이터는 일반적으로 수천 개 이상의 유전자를 포함하는데, 유전자 발현 데이터의 모든 유전자가 암과 관련이 있는 것이 아니므로 정확한 암을 분류하기 위하여 중요한 유전자만을 추출하는 것이 바람직하다. 본 논문에서 음의 상관관계를 갖는 두 개의 이상적인 유전자 벡터를 정의한 후 이와 유사한 정도를 기준으로 중요한 유전자 집단을 추출하고, 각각을 신경망으로 학습하여 결합하는 신경망 쌍을 제안한다. 실험 결과는 음의 상관관계를 갖는 두 개의 유전자 집단이 암의 클래스를 잘 구분할 수 있음을 보여주었다. 이 유전자 집단을 특징으로 하여 각각 학습한 신경망을 베이시안 방법으로 결합한 결과, 벤치마크 데이터에 대하여 신경망 쌍이 개별 분류기에 비해 우수한 성능을 보임을 확인하였다.

1. 서 론

DNA 마이크로어레이 기술은 특정한 실험 환경에서 수천 개 이상의 유전자의 발현정도를 동시에 측정할 수 있고 분자 수준에서의 생명 현상의 이해를 가능하게 하였다[1]. DNA 마이크로어레이 기술은 암의 예측 및 진단 분야에 적용되어 많은 도움을 줄 것으로 기대된다. 이 기술은 대량의 유전자 데이터를 생성하기 때문에 기존의 많은 유전자 관련 문제들이 통계적이고 분석적인 이슈로 대두되었다.

암의 정확한 분류는 암의 치료에 있어 매우 중요하지만, 조직병리학 등의 방법에 의존하는 임상학적 암 분류는 종종 불완전하여 오진할 수 있는 가능성이 있다. 유전자 발현 정보에 근거한 분자 수준의 암 분류는 정확하고 객관적이며 체계적인 암의 분류를 위한 방법론을 제시해준다. 하지만 유전자 발현 데이터는 일반적으로 매우 많은 양의 유전자 정보를 포함하고 있고 모든 유전자가 암과 관련이 있는 것은 아니므로 암과 관련이 있는 중요한 유전자만을 추출하여 이를 기준으로 암을 분류하는 것이 바람직하다.

본 논문에서는 암의 클래스를 대표하는 두 개의 이상적인 유전자 벡터를 정의하고, 각 이상적인 유전자 벡터와 각 유전자의 유사도에 따라 의미있는 유전자를 추출하였다. 두 개의 이상적인 유전자 벡터는 클래스 A에서는 높이 발현되고, 클래스 B에서는 낮게 발현되는 벡터와 클래스 A에서는 낮게 발현되고, 클래스 B에서는 높게 발현되는 벡터로 구성된다. 두 벡터는 음의 상관관계를 가지므로 이를 기준으로 선택한 두 유전자 집단 역시 음의 상관관계를 갖는다. 음의 상관관계 특징은 암을 분류하기 위한 두 개의 다른 영역을 대표하기 때문에 이러한 특징을 결합함으로써 보다 정확하게 암을 분류할 수 있다. 본 논문에서는 제안한 방법을 벤치마크 암 데이터에 대해 적용하고, 그 유용성을 체계적으로 분석하였다.

2. DNA 마이크로어레이

DNA 마이크로어레이는 유전자 발현정보를 얻기 위해 고형 지지체(substrate) 위에 실험하고자 하는 대량의 유전자를 고정해 놓은 것이다. 한번에 수천 개 이상의 유전자가 DNA 마이크로어레이 위에 고정되어 분석되기 때문에 대량의 유전자 데이터를 신속하게 제공한다. DNA 마이크로어레이 기술은 고형 지지체 위에 유전자를 고정하는 방법에 따라 cDNA 마이크로어레이 기술과 oligonucleotide 마이크로어레이 기술로 나뉜다. 각 방법으로 실험하고자 하는 수천 개의 유전자 DNA 서열을 고정시키기 때문에 정확하고 효율적으로 유전자 발현정보를 제공한다.

두 개의 샘플로부터 추출한 실험 시료와 참조 시료의 DNA나 RNA 서열을 RT-PCR(reverse transcription polymerase chain reaction)방법을 통하여 역전사(reverse-transcription)시키는 과정에서 각각 다른 형광물질(빨간색의 형광물질 Cy5와 녹색의 형광물질 Cy3)로 염색하여cDNA를 생성한다. 이를 마이크로어레이에 첨가하면 유전자의 상보적(complementary) 결합에 의해 시료에 발현된 유전자는 각 지점에 고정되고, 각 지점은 고정된 시료에 따라 빨간색 혹은 녹색을 띄게 된다. 레이져 형광 스캐너를 이용하여 마이크로어레이의 각 지점의 형광정도를 읽어 낸다. 각 유전자의 형광정도는 그 유전자의 발현정도를 나타내며 유전자 발현정보 데이터로 사용하기 위하여 Cy3와 Cy5 형광정도의 상대적 강도를 다음과 같이 구한다.

$$gene\_expression = \log_2 \frac{\mathrm{Int(Cy5)}}{\mathrm{Int(Cy3)}} \quad (1)$$

3. 음의 상관관계 유전자의 신경망 쌍

본 논문은 이상적인 유전자 벡터를 그림 1과 같이 클래스 A에서 높은 값을 갖고, 클래스 B에서는 낮은 값을 갖는 벡터와 클래스 A에서는 낮은 값을 갖고, 클래스 B에서는 높은 값을 갖는 두 개의 벡터로 정의하였다. 각각의 이상적인 유전자 벡터와 유사한 두 개의 유전자 집단을 추출하여 분류를 위한 특징으로 사용하고, 이를 신경망으로 학습하여 결합하고자 한다.

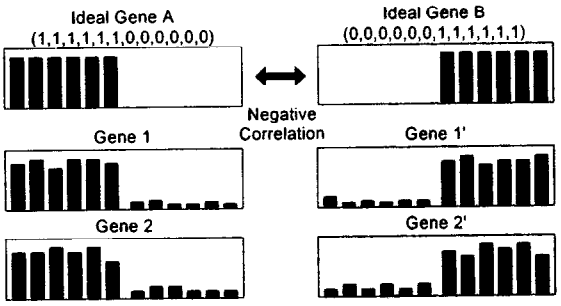

그림 1. 음의 상관관계 특징에 의해 선택된 의미 있는 유전자

3.1 음의 상관관계

M개의 샘플과 N개의 유전자를 갖는 $M \times N$ 행렬의 유전자 발현 데이터가 있고, M개의 샘플은 클래스 A와 클래스 B로 나뉜다고 하면 각 유전자 데이터 g_i는 수식 (2)와 같은 벡터로 표현할 수 있다.

$$g_i = (e_1, e_2, e_3, \cdots, e_M),\ i=1\sim N \qquad (2)$$

클래스 간의 뚜렷한 패턴의 차이가 존재하는 이상적인 유전자 벡터를 g_{ideal}이라 하고 수식 (3)와 같은 벡터로 표현했을 때, 분류에 사용하고자 하는 의미있는 유전자들은 이상적인 유전자 벡터 g_{ideal}과의 유사도가 큰 벡터라 정의할 수 있다.

$$g_{ideal} = (e_1', e_2', e_3', \cdots, e_M') \qquad (3)$$

본 논문에서는 수식 (3)의 이상적인 유전자 벡터를 Ideal Gene A (1,1,...,1,0,0,...,0)와 Ideal Gene B (0,0,...,0,1,1,...,1)로 정의하였다. 두 개의 이상적인 벡터 간의 Pearson 상관계수는 -1이므로 완전한 음의 상관관계를 갖는다. 따라서 이를 바탕으로 추출한 두 유전자 벡터 집단(feature subset)도 음의 상관 정도가 큰 특징 관계가 된다. 음의 상관관계를 갖는 특징들은 학습 데이터의 두 가지 다른 측면을 대표하기 때문에 이러한 특징들을 결합함으로써 보다 넓은 해공간을 탐색할 수 있다[2]. SGS I (significant gene subset I)은 Ideal Gene A에 기반하여 선택된 유전자의 집단으로 정의하였고, SGS II는 Ideal Gene B에 기반하여 선택된 유전자 집단으로 정의하였다. SGS는 다음 수식과 같이 정의된다.

$$\mathrm{SGS} = \arg\max\{Sim(gene_i, Ideal\ gene\ vector)\} \qquad (4)$$

$Sim(X, Y)$는 벡터 X와 Y의 유사도를 의미하며, 본 논문에서는 $Sim(X, Y)$을 계산하기 위하여 표 1과 같이 네 개의 유사도 척도를 사용하였다. Pearson 상관계수(PR)와 Spearman 상관계수(SR)는 벡터 간의 유사도를 측정하기 위한 통계적 상관관계 분석 방법이며, Euclidean 거리(ED)와 cosine 계수(CC)는 벡터 간의 거리를 측정함으로써 그 유사도를 측정하는 방법이다.

표 1. 이상적인 유전자 벡터와 각 유전자와의 유사도 척도

$PR(g_i, g_{ideal}) = \dfrac{\sum g_i g_{ideal} - \dfrac{\sum g_i \sum g_{ideal}}{N}}{\sqrt{\left(\sum g_i^2 - \dfrac{(\sum g_i)^2}{N}\right)\left(\sum g_{ideal}^2 - \dfrac{(\sum g_{ideal})^2}{N}\right)}}$
$SP(g_i, g_{ideal}) = 1 - \dfrac{6\sum (D_g - D_{ideal})^2}{N(N^2-1)}$
$ED(g_i, g_{ideal}) = \sqrt{\sum (g_i - g_{ideal})^2}$
$CC(g_i, g_{ideal}) = \dfrac{\sum g_i g_{ideal}}{\sqrt{\sum g_i^2 \sum g_{ideal}^2}}$

3.2 신경망 쌍

본 논문에서 앙상블을 위한 분류기로 다층신경망(multi-layer perceptron, MLP)을 사용하였다. MLP는 인공신경망의 대표적인 기계 학습 알고리즘으로서, 일반적인 패턴 인식 문제에서 강하고 안정적인 성능을 보인다. MLP는 오류 역전파(error back-propagation) 알고리즘을 사용하여 신경망의 결과 값이 분류 목표치에 가까워지도록 연결 강도를 조절해나감으로써 주어진 패턴을 학습한다.

본 논문에서는 여러 결합 방법 중에서 짝수 개의 분류기 결합에서 각 클래스가 동일한 선택을 받는 경우 결과의 선택에 모호함이 발생하는 것을 막기 위해서 각 분류기의 사전 정보를 이용하는 베이시안 결합 방법을 사용하였다. 투표 방법은 각 분류기의 결과만으로 결합하는 반면, 베이시안 결합 방법은 각 분류기의 오류 가능성이 최종 결과에 영향을 미치도록 한다. 결과적으로 결합하는 분류기에 대한 사전 지식을 이용함으로써 분류기의 가중치를 달리하는 결합 방식이다. k개의 개별 분류기 결합에서 $c_i(i=1,...,m)$는 샘플의 실제 클래스이고 $c(classifier_j)$는 j번째 분류기가 낸 클래스이며 η는 각 클래스 c_i의 사전 확률일 때, 베이시안 결합 방법은 수식 (5)과 같이 구한다.

$$c_{ensemble} = \arg\max_{1\le i\le m}\{\eta \prod_{j=1}^{k} P(c_i \mid c(classifier_j))\} \qquad (5)$$

4. 실험 및 결과

4.1 실험 환경

실험 데이터로 사용한 백혈병 데이터는 72개의 샘플 데이터로 구성되어 있으며, 백혈병의 두 가지 종류인 급성 골수성 백혈병(acute myeloid leukemia, AML) 환자 25명과 급성 림프성 백혈병(acute lymphoblastic

leukemia, ALL) 환자 47명으로부터 얻어진 데이터이다 [3]. 72개의 샘플 중에서 38개를 학습 데이터로 사용하였고, 나머지 34개를 실험 데이터로 사용하였는데, 각 샘플은 7129개의 유전자 발현 정보를 갖고 있다.

특징 추출 단계에서 표 1의 유사도 척도를 사용하여 각 이상적인 유전자 벡터와 가장 유사한 25개의 유전자로 각 유전자 집합(SGS I과 SGS II)을 구성하였다. 학습 단계에서 MLP의 모멘텀은 0.9로 정하였고 총 레이어의 수는 3으로 고정한 후에 학습률을 0.01에서 0.50으로 변화시키며 실험하였다. 또한 학습 데이터에 과적합되는 것을 막기 위하여 학습과정의 최대 반복은 100으로 고정하였다. SGS I로 학습한 신경망을 MLP I이라 정의하고, SGS II로 학습한 신경망을 MLP II라 정의하였다.

4.2 실험 결과

그림 2는 SGS I과 SGS II의 유전자 발현정도를 보여준다. SGS I과 SGS II는 각각 Ideal Gene A와 Ideal Gene B의 유사도를 Pearson 상관계수로 측정하고 이를 기준으로 선택된 유전자 집합이다. 50개의 유전자(SGS I과 SGS II)의 발현 정도는 ALL과 AML을 분명하게 구분해주는 것을 알 수 있다. SGS I은 ALL에서 과소발현(underexpressed)되었고, AML에서는 과다발현(overexpressed)되었다. SGS II는 ALL에서 과다발현되었고, AML에서는 과소발현되었다. 이러한 현상은 다른 유사도 척도에 의하여 유전자를 선택하였을 때도 유사하였다.

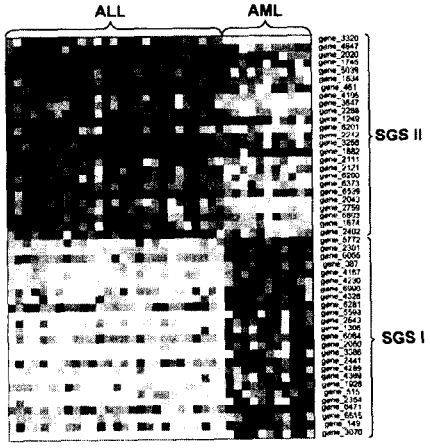

그림 2. Pearson 상관계수에 의해 선택된 SGS I과 SGS II의 유전자 발현 정도

그림 3은 Ideal Gene A와 Pearson 상관계수에 의해 추출된 유전자 수에 따른 MLP의 분류 인식률을 보여준다. 25-30개의 유전자로 구성된 유전자 집합에 대해서 가장 우수한 성능을 보였으며, 이 구간에서 그 성능은 안정적이었다. 25-30개의 유전자로 구성된 유전자 집합으로 학습한 MLP의 분류 인식률은 97.1%였다. 이러한 실험을 바탕으로 안정적인 분류 성능을 보장하고 가능하면 적은 수의 특징을 분류에 사용하고자 SGS를 25개의 유전자로 구성하였다. 표 2는 각 유사도 척도에 따른 MLP I과 MLP II 개별 분류기의 분류 인식률과 신경망 쌍(MLP I + MLP II)의 앙상블 분류 인식률을 보여준다. MLP I의 개별 분류기의 성능이 MLP II의 개별 분류기의 성능보다 우수하였으며, 신경망 쌍의 성능이 가장 우수하였다.

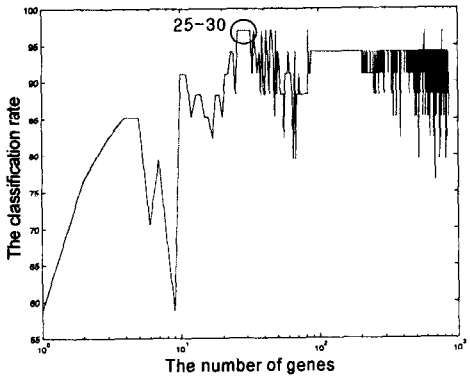

그림 3. 유전자 수에 따른 MLP의 분류 인식률(%)

표 2. 개별 분류기와 앙상블 분류기의 분류 인식률(%)

개별 분류기				
	PR	SP	ED	CC
MLP I	97.1	82.4	91.2	94.1
MLP II	79.4	79.4	61.8	76.5
앙상블 분류기				
MLP I + MLP II	97.1			

5. 결 론

본 논문은 음의 상관관계를 갖는 이상적인 유전자 벡터를 이용하여 추출된 두 개의 유전자 집합이 암을 구분하는데 있어 유의한 정보를 줄 수 있음을 확인하였다. 또한 두 유전자 집합을 특징으로 하여 각각 학습한 신경망을 결합한 결과 신경망 쌍의 분류 성능이 개별 분류기의 분류 성능과 비교하여 가장 우수하였다.

감사의 글

본 연구는 보건복지부 보건의료기술 진흥사업의 지원에 의하여 이루어진 것임.

참고 문헌

[1] C. A. Harrington, *et al.*, "Monitoring gene expression using DNA microarrays," *Curr. Opin. Microbiol.*, vol. 3, pp. 285-291, 2000.

[2] S.-B. Cho, *et al.*, "Classifying gene expression data of cancer using classifier ensemble with mutually exclusive features," *Proc. of the IEEE*, vol. 90, no. 11, pp. 1744-1753, 2002.

[3] A. Ben-Dor, *et al.*, "Tissue classification with gene expression profiles," *Journal of Computational Biology*, vol. 7, pp. 559-584, 2000.