통합 상관된 특징 집합을 이용한 림프종 데이터의 분류

박찬호*, 조성배
연세대학교 컴퓨터과학과

e-mail : cpark@candy.yonsei.ac.kr*, sbcho@csai.yonsei.ac.kr

Classification of Lymphoma Dataset with Combinatorially Correlated Feature Set

Chanho Park* and Sung-Bae Cho
Dept. of Computer Science, Yonsei University

요 약

근래, DNA microarray와 관련된 기술의 발달은 한번에 수천 개 이상의 유전자발현데이터를 얻을 수 있게 해주었고, 많은 연구기관에서 이를 이용한 질병 분류에 관하여 연구를 진행하고 있다. 하지만 수천 개의 유전자 모두가 암에 관계된 것은 아니기 때문에, 관련 유전자의 선별 작업을 먼저 수행하는 것이 필요하며, 이를 위하여 통계기반 방법, 정보이론기반 방법 등 다양한 방법이 사용되고 있다. 본 논문에서는 의미 있는 유전자를 선택하는 방법으로서, 일반적인 순위-기반 방법이 양의 상관관계만 이용한다는 점을 보완하여, 유전자와 학습데이터 사이의 음의 상관관계까지도 고려한 방법을 제시하였다. 제안한 방법의 성능을 검증하고자 잘 알려진 암 관련 유전자발현데이터인 림프종 데이터에 대하여, MLP와 KNN을 이용한 분류를 해 보았다. 실험 결과 통합 상관관계를 가지는 특징 집합이 일반적인 순위-기반 방식의 특징 집합에 비하여 높은 분류 인식률을 보여주었다.

1. 서론

병을 빨리 발견하고 치료하고자하는 사람들의 희망은 관련 학문과 기술의 많은 발전을 가져왔으며, 근래에는 한 번에 수천 개 유전자의 발현정보를 알 수 있게 될 정도에 이르렀다. 하지만 그 유전자들이 모두 특정 질병과 연관되어 있는 것은 아니기 때문에, 특정 질병과 관련된 유전자를 선별하는 작업이 필요하며, 이를 특징선택 혹은 유전자선택이라 한다. 특징선택과정을 거쳐서 선별된 유전자들은 분류기의 입력으로 들어가서 분류기를 학습시키게 되고, 이러한 일련의 과정은 하나의 분류시스템이 된다.

특징선택과정에서 좋은 유전자 집합을 선택할수록 분류시스템의 성능도 좋아지기 때문에, 특징선택은 암의 발견에 있어서 매우 중요한 역할을 한다[1]. 특징선택은 특징의 중요성을 측정하는 도구를 두어 그 값의 순위에 의하여 선택하는 순위-기반의 방법과 학습데이터 자체의 특성을 이용하여 특징을 선택하는 방법 등으로 구분할 수 있다. 유전자발현데이터에서의 순위-기반 방법은 주로 학습데이터의 표현형벡터(학습데이터의 암인지 정상인지 여부를 1과 0을 사용하여 표현한 것)와 유전자의 발현벡터(0에서 1사이로 정규화 된 값) 사이의 유사성을 다양한 방법으로 측정하여 순위에 따라 선택하게 되는데, 이 경우 발생하는 단점중 하나는 표현형벡터의 정확한 레이블을 알기 힘들다-암을 1로 정해야 할지, 정상을 1로 정해야 할지-는 것이고, 또 하나는 표현형벡터와 정확히 반대 양상을 보이는(발현되지 않음으로서 표현 양상이 나타나게 해주는) 유전자가 고려되지 않는다는 것이다.

본 논문에서 제안하는 특징들의 선택은 위에서 언급한 단점들을 보완해주는 역할을 한다. 즉, 학습데이터와 양의 상관관계뿐만 아니라, 음의 상관관계를 보이는 특징들까지 선택하였기 때문에 표현형 벡터의 레이블에 상관없이 특징을 선택하는 것이 가능하고, 질병에 크게 영향을 끼치거나 억제하는 유전자의 선택이 모두 가능하다.

제안한 방법의 타당성을 검증하기 위하여 잘 알려진 분류기인 다층신경망(MLP)과 *k*NN을 이용한 분류 시스템을 구성하고, 림프종 데이터를 분류하고자 한다.

2. DNA microarray

DNA microarray는 용액이 투과하지 않는 딱딱한 지지체 위에 고밀도로 cDNA를 고정시켜 놓은 것이다. Array를 구성하는 각 셀은 두 개의 다른 환경에서 채집된 유전물질에 녹색의 Cy3와 빨간색의 Cy5라는 형광물질을 합성한 것을 동일한 양으로 보합한 것이다. 이것을 스캐너로 읽어 들이면 녹색부터 빨간색에 이르는 발현정도를 나타내는 이미지를 얻을 수 있다[2].

발현정보 값으로는 식 (1)과 같이 Cy5/Cy3의 비율에 로그를 취한 값을 사용한다. 한편, 이상 샘플과 정상 샘플에서 관련 유전자는 서로 다른 발현정도를 보이므로, microarray상에서 얻은 유전자발현 정보는 질병의 분류에 이용할 수 있다.

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

3. 유전자선택-분류 시스템

본 논문에서 제안하는 분류시스템은 그림 1과 같다. DNA microarray를 통해서 나온 유전발현데이터는 특징선택기를 지나며 통합 상관관계를 갖는 특징집합이 선택되어지고, 분류기를 지나 분류결과를 내어놓는다.

3.1 유전자 선택방법

본 논문에서는 의미 있는 유전자를 선택하기 위하여 통계적 상관분석 기반의 피어슨 상관계수(PC)와 스피어맨 상관계수(SC), 유사도 기반의 유클리드 거리(ED)와 코사인 계수(CC), 그리고 정보이론 기반의 신호 대 잡음 비(SN)의 총 다섯 가지를 사용하였다[3].

먼저 피어슨 상관계수는 널리 알려진 상관계수 측정방법으로 표현형 벡터 X와 유전자 벡터 Y의 관계를 식 (2)와 같은 방법으로 측정한다. 이 계수는 -1에서 1까지의 값을 가지게 되고, 1에 가까울수록 X와 Y가 강한 양의 상관관계를 가진다는 의미이며, -1에 가까울수록 강한 음의 상관관계를 가진다는 것을 의미한다. 한편 두 벡터가 상관이 거의 없을 때에는 상관계수는 0에 가까운 값을 나타낸다.

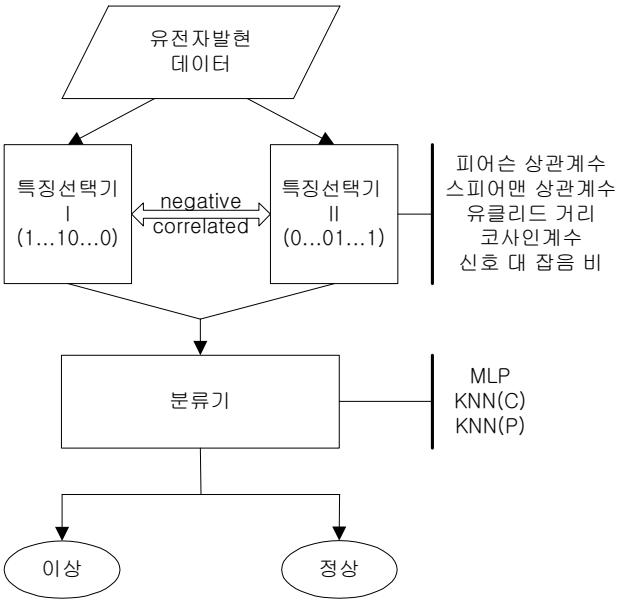

그림 1. 통합 상관된 특징을 이용한 분류 시스템

$$PC = \frac{\sum XY - \frac{\sum X \sum Y}{N}}{\sqrt{(\sum X^2 - \frac{(\sum X)^2}{N})(\sum Y^2 - \frac{(\sum Y)^2}{N})}} \qquad (2)$$

스피어맨 상관계수는 비모수분석 방법 중 하나로 벡터의 값이 아닌 순위를 이용하여 상관관계를 구한다는 점이 특징이다. 이것은 피어슨 상관계수와 마찬가지로 -1부터 1사이의 값을 가지게 된다. 두 벡터 X와 Y의 순위배열 D_x, D_y를 이용한 스피어맨 상관계수는 식 (3)과 같이 표현된다.

$$SC = 1 - \frac{6\sum (Dx - Dy)^2}{N(N^2 - 1)} \qquad (3)$$

한편, n차원상의 벡터 X와 Y의 유클리드 거리는 식 (4)와 같이 구할 수 있는데, 이 값은 거리이기 때문에 다른 계수들과는 달리 작을수록 서로 유사하다는 의미가 된다.

$$ED = \sqrt{\sum (X - Y)^2} \qquad (4)$$

코사인 계수는 n차원 위에 있는 두 벡터 사이의 방향을 측정하는 방법으로서, 서로 유사한 방향을 가리킬수록 두 벡터사이는 0에 가까워지고, 코사인 값은 1에 가까워진다. 두 벡터 X와 Y의 코사인 계수는 식 (5)와 같이 표현할 수 있다.

$$CC = \frac{\sum XY}{\sqrt{\sum X^2 \sum Y^2}} \qquad (5)$$

마지막으로, 학습데이터에 대하여 각 유전자 g를 클래스 c에 속하는 것들과 그렇지 않은 것들로 분류

한 후, 각각에 대하여 정규분포를 계산하였을 때, 클래스 c에 의하여 분류되는 유전자 g의 신호 대 잡음 비는 각 클래스에 속하는 유전자들의 평균과 표준편차를 이용하여 식 (6)과 같이 계산된다.

$$SN = \frac{\mu_1(g) - \mu_2(g)}{\sigma_1(g) + \sigma_2(g)} \quad (6)$$

3.2 통합 상관관계를 이용한 특징선택

어떤 하나의 변수를 설명할 때, 그것과 아주 비슷한 패턴을 이용하여 설명할 수 있는 것이 있는가 하면(양의 상관관계), 완전히 반대의 패턴을 이용하여 설명할 수 있는 것도 있고(음의상관관계), 둘을 통합시킨 것도 있다.

유전발현데이터의 경우 한 샘플을 설명하기 위한 변수(유전자)들이 수천 개에 달하고, 발현정도를 0과 1사이로 정규화 시켰다 하여도, 그 샘플의 표현형(암인지 정상인지의 여부)을 1로 표현해야 할지, 0으로 표현해야 하는지는 명확하지 않다. 따라서 표현형벡터를 오직 한가지로 정해서 그것에 의존한 순위-기반의 특징선택을 하는 것은 분류의 성능을 떨어뜨릴 수 있다. 그리고 어떤 유전자의 발현은 그 샘플의 표현양상을 크게 하는데 작용을 했을 수가 있고(+), 다른 어떤 유전자는 활동을 하지 않는 것이(-) 그 샘플의 표현양상을 크게 하는 것에 도움을 줄 수 있다. 따라서 본 논문에서는 이와 같은 것들을 고려하여 학습데이터와 유전자간의 양의 상관관계가 큰 것들뿐만 아니라 음의 상관관계가 큰 것들까지도 모두 가지는 특징 집합을 선택하였다. 이렇게 통합적인 상관관계를 보이는 특징들은 학습데이터의 표현형 벡터에 상관없이 일관성 있게 선택될 수 있고, 목적변수에 대하여 양의 작용을 하는 것과 음의 작용을 하는 것을 모두 포함한다.

본 논문에서는 통합 상관관계를 가지는 특징 집합을 선택하기 위하여 표현형벡터를 (1...10...0)으로 두고 특징의 반을 선택하였고, (0...01...1)로 두고 나머지 반을 선택하였다.

3.3 분류기

본 논문에서는 패턴인식분야에서 널리 쓰이는 분류기인 다층신경망(Multi-Layer Perceptron, MLP)과 kNN(k-nearest neighbor)을 사용하였다.

MLP는 노드간의 가중치를 조절하여 입력패턴이 목표출력에 최대한 가까운 값을 낼 수 있도록 해주는 분류기이다. 본 논문에서는 오류 역전파 알고리즘을 학습방법으로 하며, 하나의 은닉층을 가지는 MLP를 사용하였다.

KNN은 대표적인 기억기반 학습방법으로서 테스트 샘플과 학습데이터와의 거리를 계산하여 가장 가까운 k개에 대하여 어느 집단이 가까운지를 알아내는 분류기이다. KNN은 다양한 방식으로 거리를 측정할 수 있으며, 본 논문에서는 피어슨 상관계수(KNN(P))와 코사인 계수(KNN(C))의 두 가지를 거리측정 도구로 사용하였다. 각각은 3.1 유전자선택방법에서 표현된 식을 이용한다.

4. 실험 및 결과

4.1 실험환경

실험에 사용한 데이터는 잘 알려진 유전자발현 데이터인 림프종 암 데이터로서, 이 데이터는 4026개의 유전자로 구성되어 있으며 총 47개의 샘플이 사용되었다. 이중 24개는 GC B-like DLBCL이고, 23개는 activated B-Like DLBCL이다[4]. 총 데이터 중, 22개는 학습에 이용하고, 25개를 테스트에 이용하였다.

실험은 MLP, KNN(C) 및 KNN(P)에 대하여 진행하였다. 선택되는 특징의 개수는 모든 방법에서 30개로 맞췄으며, 신경망의 경우는 학습데이터의 목표인식률 98%, 최고 반복회수 500회, 은닉노드 수 18개에 대하여 다양한 학습률과 모멘텀을 적용시켰다. KNN의 경우는 k값을 1, 3, 5, 7, 9로 변화시켜가며 실험하였다.

4.2 실험결과

신경망 분류기를 사용한 경우의 인식률에 대한 실험 결과는 표 1과 같다. 표에서 알 수 있는 바와 같이 스피어맨 상관계수를 제외한 모든 특징선택 방법에서 통합 상관관계를 이용하여 선택된 특징들을 이용한 분류의 성능이 가장 좋았다.

표 1. 신경망 분류기에 대한 각 방법별 최고 인식률

특징선택법 \ 표현형벡터	(1...10...0)	(0...01..,1)	통합상관된 특징 집합
피어슨 상관계수	64	76	**92**
스피어맨 상관계수	60	**84**	72
유클리드 거리	60	76	**92**
코사인 계수	60	72	**92**
신호 대 잡음 비	60	76	**96**

한편 제안한 특징선택방법이 분류기 의존적이어서 MLP에 대해서만 좋은 특성을 보일 가능성도 있기 때문에 kNN 분류기에 대하여 동일한 데이터를

사용한 실험을 하였다. 이 결과는 표 2와 3에 있으며 모두 통합 상관관계를 가지는 특징 집합이 우수한 성능을 낸다는 것을 보여준다.

표 2. KNN(C) 분류기에 대한 각 방법별 최고 인식률(%)

특징선택법 \ 표현형벡터	(1...10...0)	(0...01..,1)	통합상관된 특징 집합
피어슨 상관계수	60	84	**92**
스피어맨 상관계수	68	80	**84**
유클리드 거리	64	76	**96**
코사인 계수	56	80	**92**
신호 대 잡음 비	60	76	**88**

표 3. KNN(P) 분류기에 대한 각 방법별 최고 인식률(%)

특징선택법 \ 표현형벡터	(1...10...0)	(0...01..,1)	통합상관된 특징 집합
피어슨 상관계수	76	80	**92**
스피어맨 상관계수	64	**84**	80
유클리드 거리	84	76	**92**
코사인 계수	72	76	**92**
신호 대 잡음 비	68	76	**88**

그림 2는 표 2의 결과를 확장시킨 것으로 여러 번 실험을 반복하여서 얻은 결과이다. 각 방법별로 평균 인식률이 직선으로 연결되어 있고, 네모박스 사이는 최고 인식률과 최저 인식률의 차이를 보여준다. 이 그림에서도 알 수 있듯이 통합 상관관계를 가진 유전자 집합을 사용한 분류가 최고 인식률, 최저 인식률, 평균인식률 모두에서 다른 방법들보다 뛰어난 성능을 보여주었으며, 편차 또한 적어 안정된 성능을 낸다는 것을 알 수 있다.

5. 결론

통합 상관관계를 가진 유전자 집합이 다른 일방향적인 순위-기반의 유전자 집합보다 분류에 있어 우수한 성능을 내는 것을 실험을 통해 확인할 수 있었다. 한편 스피어맨 상관계수 같은 경우는 통합 상관관계를 가진 유전자들을 선택하였을 때 일방향적 순위-기반 방법보다 조금 떨어지는 성능을 보이기도 하였는데 이는 스피어맨 방법이 다른 방법들과는 달리 벡터의 값을 직접 사용하는 것이 아니라 단지 순위에만 의존하기 때문에 그런 결과가 나왔을 가능성이 있다. 그래도 두 가지 표현형 벡터에 의한 일방향적 순위-기반 방법의 평균 성능 이상의 성능을 나타낸 것으로 보아서, 한 가지 방법만을 사용하였을 때에 비하여 위험부담을 줄일 수 있다.

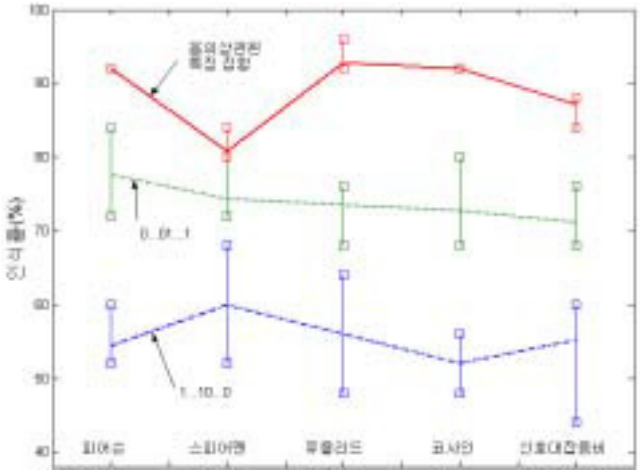

그림 2. KNN(C)에 대한 각 방법의 인식률 범위와 평균 인식률의 비교

한편, 본 논문에서 제안한 통합 상관관계를 지니는 특징 집합의 선택은 완전히 새로운 방법을 만드는 것이 아니고, 기존의 방법들에 적용시키기 용이하기 때문에 여러 가지 모수 분석에 이용하면 유용할 것이다.

감사의 글

본 연구는 보건복지부 보건의료기술 진흥사업의 지원에 의하여 이루어진 것임.

참고문헌

[1] Leping Li, et al., "Gene selection for sample classification based on gene expression data: Study of sensitivity to choice of parameters of the GA/KNN method," Bioinformatics, vol. 17, no. 12, pp. 1131-1142, June 2001.

[2] John Quackenbush, "Computational analysis of microarray data," Nature Reviews Genetics, vol. 2, pp. 418-427, June 2001.

[3] Sung-Bae Cho, and Joong-Won Ryu, "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.

[4] Ash A. Alizadeh, et al., "Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling," Nature, vol. 403, pp. 503-511, February 2000.