제18회 한국정보처리학회 추계학술발표대회 논문집 제9권 제2호 (2002.11)

공통 선택된 특징을 이용한 유전 발현 데이터의 분류

박찬호*, 조성배
연세대학교 컴퓨터과학과
e-mail:cpark@candy.yonsei.ac.kr*, sbcho@cs.yonsei.ac.kr

Classification of Gene Expression Profiles Using Common Features Selected

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

요 약

최근 생명공학 기술과 분석화학 기술의 발달로 생물 유전 데이터를 대량으로 얻는 것이 가능하게 되었다. 아울러 이렇게 얻어진 데이터를 적절하게 처리하고 분석하는 방법들도 여러 가지가 소개되어 왔다. 본 논문에서는 DNA 마이크로어레이 정보를 분류하기 위하여 세 가지 데이터에 대하여 여러 가지 특징 선택 방법으로 선택된 유전자들을 사용하여 신경망 분류기에 적용시켜 보았다. 실험 결과 백혈병 데이터의 경우 피어슨 상관계수를 이용한 분류가 97.1%로 가장 높은 인식률을 보여주었다. 한편 여러 가지 특징 선택 방법에 의하여 공통적으로 선택된 유전자를 사용하여 분류하면 더 높은 인식률이 나올 것 같았지만 실제로는 기대에 못 미치는 성과를 보여주었다. 따라서 무조건 여러 번 선택된 특징을 선택하기 보다는 특징들끼리의 상관관계를 고려하여 선택하는 방법이 필요할 것이다.

1. 서론

최근 DNA 마이크로어레이 기술의 발달로 예전에는 얻기 힘들었던 유전 정보를 대량으로 얻어내는 것이 가능하게 되었다. 그러나 이렇게 얻어진 정보는 단순한 숫자의 나열이므로 그 상태로는 어디에도 이용될 수가 없다. 따라서 이 숫자의 나열을 분석하여 의미있는 정보를 뽑아내는 연구가 필요한데, 수년 전부터 그러한 연구 방법들이 제시되어 왔다.

본 논문에서는 대량의 인간 유전자들 중에서 표현형과 큰 관련성을 갖는 유전자를 선택하기 위한 다양한 특징 선택 방법들과 신경망 분류기를 이용하여 질병의 종류를 효과적으로 분류하는 방법에 대해 소개하고자 한다. 또한 여러 가지 특징 선택법에서 동시에 뽑힌 특징들을 이용하여 분류를 해 봄으로써 분류의 성능을 개선시켜보고자 한다.

2. 실험 데이터

실험에는 백혈병 데이터, 림프종 데이터, 대장암 데이터의 세 가지가 이용되었다. 모든 데이터는 필터링 과정에 의하여 전처리 되었고, 밑이 2인 로그 값으로 변환된 값을 사용하였다.

(1) 백혈병 데이터(Leukemia dataset)

백혈병 데이터는 72명의 백혈병 환자로부터 얻어진 골수 샘플로부터 제작되어졌다. 이 중 38개가 학습에 이용되었고, 34개는 테스트에 이용되었다. 각 샘플은 7129개의 유전자로 구성되어 있고, 급성 림프구성 백혈병(ALL: Acute Lymphoblastic Leukemia)과 급성 골수성 백혈병(AML: Acute Myeloid Leukemia)의 두 가지 클래스로 분류된다[1].

(2) 림프종 데이터(Lymphoma dataset)

림프종 데이터(http://llmpp.nih.gov/lymphoma/)는 4026개의 유전자로 구성되어 있으며 총 47개의 샘플이 사용되었다. 이중 24개는 GC B-like DLBCL이고, 23개는 activated B-Like DLBCL이며, 22개는 학습에 이용하고, 25개를 테스트에 이용하였다[2].

(3) 대장암 데이터(Colon dataset)

대장암데이터는 2000개의 유전자로 이루어져 있으며 62개의 샘플이 사용되었다. 이중 40개는 암 조직이고, 22개는 정상 조직이다. 본 논문에서는 31개의

학습 샘플과 31개의 테스트 샘플로 나누어 실험하였다[3].

3. 특징 선택 방법

본 논문에서는 데이터의 특징을 선택하기 위하여 총 7가지의 특징선택 방법을 사용하였다[4].

(1) 피어슨 상관 계수

상관계수분석이란 변수간의 관련성을 분석하기 위해 사용하는 방법으로서, 하나의 변수가 다른 변수와 관련성이 있는지를 알 수 있고, 또 관련 정도가 어느 정도인지 알아보기 위한 방법이다. 피어슨 상관계수는 상관계수분석에서 자주 이용되는 계수이며 상관계수 r은 [-1, 1]의 값을 갖는다. r의 값이 1에 가까울수록 두 변수는 양의 상관관계를 나타내게 되고, 서로 유사하다는 것을 의미한다. 반면 r이 -1에 가깝다면 두 변수의 관계는 음의 상관관계가 되며 서로 반대방향으로 관계가 있다는 것을 의미한다. r이 0에 가깝다면 두 변수 사이에 별로 관계가 없음을 의미한다. N개의 원소를 갖는 두 벡터 X와 Y 사이의 피어슨 상관계수는 다음과 같이 정의된다.

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

(2) 스피어맨 상관계수

비모수분석은 변수들이 양적 변수가 아니어도 될 때 이용할 수 있는 상관계수분석 방법으로, 스피어맨 상관계수와 같은 방법들이 있다. 스피어맨 상관계수는 변수의 순위배열을 사용하여 변수간의 상관관계를 분석하는 방법으로 피어슨 상관계수와 마찬가지로 상관계수는 [-1, 1]의 값을 갖는다. 한편 스피어맨 상관계수는 X와 Y의 순위배열 Dx와 Dy를 사용하여 다음 식으로 나타낼 수 있다.

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

(3) 유클리드 거리(Euclidean Distance)

유클리드 거리는 두 변수간의 유사성을 측정하기 위한 방법이다. 두 변수간의 유사성은 거리로 나타낼 수 있는데, 거리가 가까울수록 유사성이 높다. 유클리드 거리는 두 변수간의 기하학적 공간에서의 거리를 나타내며, 거리 값이 크게 나올수록 유사한 정도가 낮은 것이기 때문에 사실 비유사성 정도를 나타낸다고 볼 수 있다. 두 벡터 X와 Y의 유클리드 거리는 다음과 같이 비교적 쉬운 식으로 나타낼 수 있다.

$$r_{euclidean} = \sqrt{\Sigma(X - Y)^2}$$

(4) 코사인계수

두 변수간의 유사성 측정을 위한 다른 방법으로 코사인계수가 있다. 코사인계수 방법에서 계수값은 [-1, 1]의 범위를 갖게 되는데, 두 변수간의 유사성은 계수값이 클수록 높게 된다. 왜냐하면 코사인계수는 두 변수 사이의 각을 측정해서 코사인 값으로 나타내어 주는데 유사성이 높을수록 각이 작고, 코사인값은 1에 가까워지기 때문이다. X와 Y의 코사인 계수는 다음 식에 의하여 구할 수 있다.

$$r_{cosine} = \frac{\sum XY}{\sqrt{\sum X^2 \sum Y^2}}$$

(5) 정보이론

전체 데이터로부터 의미 있는 정보를 뽑아내는 척도로 정보이론에서 사용하는 정보이득(Information Gain), 상보정보(Mutual Information), 신호 대 잡음비(Signal to Noise Ratio) 등의 방법을 이용하였다. 정보이득과 상호정보의 경우는 특정 유전자의 i번째 샘플이 특정 클래스 c에 속하는 가의 여부와 그 유전자가 발현했는가 여부의 두 가지 기준에 의하여 네 가지로 구분 짓고, 각 종류에 속하는 샘플의 수를 각각 A, B, C, D라 했을 때, 주어진 유전자 g의 정보 이득과 상호 정보 계수는 각각 다음과 같다.

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

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

한편, 학습 샘플에 대해 주어진 유전자 g를 클래스 c에 속하는 것들과 그렇지 않은 것들로 분류한 후, 각각에 대하여 정규분포를 계산하였을 때, 클래스 c에 의하여 분류되는 유전자 g의 신호 대 잡음비는 다음과 같이 계산된다.

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

4. 신경망 분류기

본 논문에서는 위의 일곱 가지 특징 선택 방법에 의하여 뽑힌 데이터에 대하여 역전파 다층 신경망을 사용하여 분류하였다[5]. 이 방법은 학습 벡터들을

반복적으로 입력 받으면서 가중치를 조절하여 원하는 기대출력이 나오도록 학습시키는 방법으로 패턴 분류 문제에 있어서 많이 사용되는 방법이다. 가중치의 조절은 실제의 출력과 기대출력의 차이를 줄이는 방향으로 진행되며, 이 과정을 통하여 신경망은 올바른 결과와 입력패턴을 연관시키는 학습을 하게 된다. 그림 1은 N개의 입력과 M개의 출력을 갖는 신경망 분류기의 구조를 간략하게 보여준다. 본 논문에서 사용한 신경망은 25개의 입력노드와 8개의 은닉노드, 2개의 출력노드를 갖는다. 이진분류의 경우 일반적으로 신경망의 출력 노드는 두 개가 된다. 가중치의 경우 입력 노드와 은닉노드 사이, 은닉노드와 출력노드 사이의 것이 모두 조절된다.

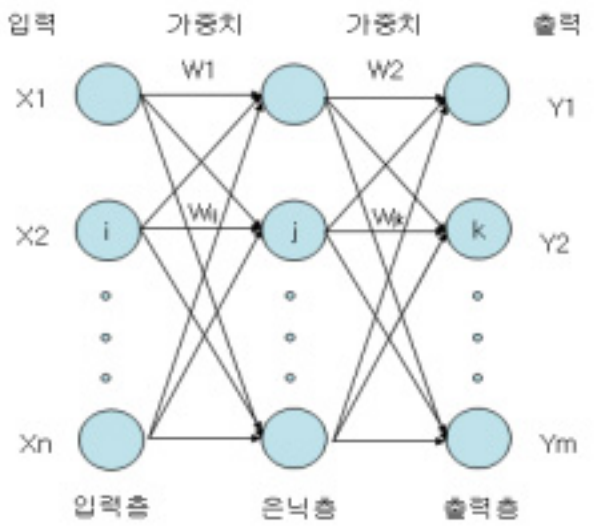

그림 1. 신경망 분류기의 구조

5. 실험결과

5.1 단순특징-분류

위에서 설명한 7가지의 특징 선택 방법을 사용하여 선택된 특징을 신경망 분류기의 입력으로 사용한 결과는 다음과 같다.

표 1. 각 방법별 특징선택-분류 인식률(%)

	백혈병	림프종	대장암
Pearson	97.1	64.0	77.4
Spearman	94.1	60.0	67.7
Euclidean	91.2	60.0	77.4
Cosine	94.1	68.0	83.9
IG	94.1	92.0	80.6
MI	64.7	76.0	77.4
SN	67.6	76.0	71.0
평균	86.1	71.0	76.5
표준편차	13.78	11.48	5.50

백혈병 데이터의 경우 피어슨 상관계수로 뽑힌 특징들이 97.1%의 가장 높은 인식률을 보여주었고, 림프종 데이터의 경우는 정보이득 방법이 92.0%의 인식률을 보여주었다. 한편 대장암 데이터의 경우 코사인 계수로 뽑힌 유전자들이 83.9%의 가장 높은 인식률을 보여주었다. 한편 표준편차를 살펴보면 백혈병 데이터가 13.78로 가장 높았고, 대장암 데이터가 5.50으로 가장 낮은 수치를 보여주었데, 이는 백혈병 데이터가 특징 선택 방법에 따라 그 인식률이 크게 좌우되고, 대장암 데이터는 상대적으로 작게 영향을 받는다는 것을 나타낸다.

5.2 특징의 중복

각 특징 선택 방법에 의하여 고른 25개씩의 유전자들 중에서 피어슨 상관계수와 유클리드 거리, 코사인 계수로 뽑힌 유전자들은 서로 중복되는 경향이 컸다. 아래의 표는 그 결과이다.

표 2. 백혈병 데이터의 공통 특징

	Cos.	Euc.	IG	MI	Peas.	SN	Spe.
Cos.		19	1	0	21	0	0
Euc.			1	0	19	0	0
IG				0	1	0	0
MI					0	0	0
Peas.						0	0
SN							0
Spe							

표 3. 림프종 데이터의 공통 특징

	Cos.	Euc.	IG	MI	Peas.	SN	Spe.
Cos.		20	7	3	23	1	0
Euc.			5	2	20	1	0
IG				0	7	0	0
MI					2	0	0
Peas.						0	0
SN							0
Spe							

표 4. 대장암 데이터의 공통 특징

	Cos.	Euc.	IG	MI	Peas.	SN	Spe.
Cos.		13	0	0	16	0	0
Euc.			0	0	9	0	0
IG				0	0	1	0
MI					0	1	0
Peas.						0	0
SN							0
Spe							

이 결과는 클래스를 분류하는데 있어 공통으로 뽑힌 유전자들이 매우 중요한 역할을 한다고 추측해 볼 수 있다. 따라서 공통되는 유전자들을 사용하여 신경망 분류기를 사용한다면 더 좋은 결과를 기대할

수 있다. 특히 피어슨 계수, 코사인 계수, 유클리드 거리의 세 가지 방법에 동시에 뽑힌 유전자들의 개수는 매우 고무적인 수치를 보여주었다. 아래의 표는 백혈병 데이터에서 이 세 가지 방법에 의하여 동시에 선택된 17개 유전자들의 정보를 보여준다. 한편 림프종 데이터에서는 19개의 유전자들이 동시에 선택되었고, 대장암 데이터에서는 9개의 유전자들이 선택되었다.

Leukotriene C4 synthase (LTC4S) gene
Zyxin
LYN V-yes-1 Yamaguchi sarcoma viral related oncogene homolog
CD33 CD33 antigen (differentiation antigen)
LEPR Leptin receptor
PRG1 Proteoglycan 1, secretory granule
Liver mRNA for interferon-gamma inducing factor(IGIF)
DF D component of complement (adipsin)
Induced myeloid leukemia cell differentiation protein MCL1
Phosphotyrosine independent ligand p62 for the Lck SH2 domain mRNA
INTERLEUKIN-8 PRECURSOR
FAH Fumarylacetoacetate
ITGAX Integrin, alpha X (antigen CD11C (p150), alpha polypeptide)
LYZ Lysozyme
Peptidyl-prolyl CIS-TRANS isomerase, mitochondrial precursor
Interleukin 8 (IL8) gene
LGALS3 Lectin, galactoside-binding, soluble, 3 (galectin 3)

그림 2. 세 가지 특징선택방법으로 겹친 백혈병 유전자들

표 5는 세 가지 특징선택 방법에서 공통적으로 선택된 유전자들을 신경망 분류기의 입력으로 사용하여 분류시킨 결과와 일곱 가지를 따로 사용했을 때의 평균적인 인식률이다. 그러나 성능이 개선되리라는 기대와는 달리 전혀 개선된 성능을 보여주지 못하였다.

표 5 공통된 특징을 사용하였을 때의 인식률(%)

	백혈병	림프종	대장암
공통	91.2	68.0	71.0
평균	86.1	71.0	76.5

이 표에서 볼 수 있듯이 백혈병 데이터를 제외하고는 오히려 평균치보다도 낮은 성능을 보여주었다. 이렇게 된 원인으로는 몇 가지를 추측해 볼 수가 있다. 그 중 하나는 공통되어 뽑힌 특징의 수가 적기 때문인 것 같다. 단순 특징 선택 방법으로는 25개씩 뽑았다가, 그 중에서 공통으로 선택된 것을 고르려니 자연히 25개보다 적을 수밖에 없었다. 성능이 좋지 않은 다른 원인으로는 특징들끼리의 의존관계를 고려하지 않았기 때문인 것 같다. 만약 유전자1, 2, 3이 모두 높은 순위로 선택되었는데 모두 의존적인 관련이 있다면, 그 중 하나만 의미가 있고 나머지는 사실 별 의미가 없는 것이다. 더군다나 공통된 특징을 뽑다 보니 더더욱 유의한 특징들이 줄어들었을 것이다. 따라서 정확한 이유를 알기 위해서는 특징들끼리의 의존 관계 분석 및 공통된 환경에서의 실험이 필요할 것이다.

6. 결론

각 특징선택 방법으로 뽑힌 특징들 중 공통적으로 선택된 유전자들은 분명 다른 유전자들보다 분류 성능에 있어서 더 좋은 특징일 것이라 예상되었다. 하지만 실험 결과에서 알 수 있듯이, 오히려 좋지 않은 결과가 나오기도 하였다. 이는 특징을 선택할 때 무조건 여러 번 뽑힌 유전자가 항상 좋은 성능을 내는 것이 아니라는 것을 보여준다. 달리 말한다면 그런 유전자들 개개는 표현형 벡터와 매우 유사한 벡터라고 볼 수 있지만, 그들의 조합만으로는 전체 해 공간을 탐색하는 것에 무리가 있다고 할 수 있다. 본 논문에서 택한 방법은 유전자들끼리의 상관관계는 고려하지 않았는데, 더 좋은 분류 성과를 내기 위해서는 유전자들끼리의 상관관계를 고려하여 그 조합이 전체공간을 해 공간과 그렇지 않은 공간으로 잘 분류할 수 있도록 해야 할 것이다.

참고문헌

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

[2] A. 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.

[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

[4] 권영준, 류중원, 조성배, "신경망 분류기를 이용한 암 관련 유전자 발현정보의 분류", 한국정보과학회 춘계 학술발표논문집(B), vol. 28, no. 1, pp. 295-297, 2001.

[5] R. P. Lippmann, "Pattern Classification Using Neural Networks," *IEEE Communications Magazine,* pp. 47-64, November, 1989.