제19회 한국정보처리학회 춘계학술발표대회 논문집 제10권 제1호 (2003. 5)

전진 선택법을 이용한 유전자 발현정보 기반의 암 분류

유시호, 조성배
연세대학교 컴퓨터과학과

bonanza@candy.yonsei.ac.kr, sbcho@csai.yonsei.ac.kr

Cancer Classification with Gene Expression Profiles using Forward Selection Method

Si-Ho Yoo, and Sung-Bae Cho
Dept of Computer Science, Yonsei University

요 약

유전 발현 데이터는 생명체의 특정 조직에서 채취한 샘플을 microarray상에서 측정한 것으로, 유전자들의 발현 정도가 수치로 나타난 데이터이다. 일반적으로 정상조직과 이상조직에서 관련 유전자들의 발현 정도는 차이를 보이기 때문에, 유전 발현 데이터를 통하여 암을 분류할 수 있다. 하지만 분류에 모든 유전자가 관여하지는 않으므로 관련성 있는 유전자만을 선별해내는 작업인 특징 선택 방법이 필요하다. 본 논문에서는 회귀분석의 변수선택방법중 하나인 전진 선택법(forward selection method)을 사용하여 유전자들을 선택하고 분류하는 방법을 제안한다. 실험데이터는 대장암 데이터를 사용하였고, 분류기는 KNN을 사용하였다. 이 방법과 상관계수를 이용한 특징 선택 방법인 피어슨 상관계수와 스피어맨 상관계수방법과 비교해본 결과 전진 선택법에 의한 특징 선택 방법이 암의 분류에 있어서 더 효과적인 유전자 선택을 한다는 사실을 확인하였다. 실험결과 90.3%의 높은 인식률을 보였다.

1. 서론

최근 몇 년간 암의 정확한 분류를 위한 연구가 활발하게 진행되어왔다. 하지만, 적은 샘플의 수와 수천개의 유전자를 가지고 올바르게 분류를 하기는 쉬운 일이 아니다. 컴퓨터의 발달과 DNA microarray 기술의 발달로 인하여 생명체에 관한 대량의 유전정보를 얻는 것은 가능하게 되었지만 이러한 대량의 유전정보가 암의 정확한 분류를 하는데 모두 필요한 것은 아니다. 그렇기 때문에 필요한 유전자만을 선택해내는 적절한 특징 선택 방법이 필요하다[1].

본 논문에서는 회귀분석에 기반을 둔 전진 선택법을 사용하여 유전자들을 선택하였다. 기존의 특징 선택 방법들은 선택되는 유전자들간의 관계를 고려하지 않고, 일정 기준에 의해서 순서대로 유전자들을 선택하는 순위기반(rank-based)의 선택방법들이었다[2]. 하지만 전진 선택법은 선택되는 유전자들간의 관계를 고려하여, 서로 중복된 정보가 최소화되도록 한다. 선택되는 유전자들끼리 서로 다른 정보를 가진 만큼, 그 유전자들의 조합은 순위기반에 의한 선택방법보다 더 많은 정보를 가질 수 있을 것이다[3].

본 논문에서는 분류기로는 K-최근접 이웃(KNN), 유전 발현 정보 데이터는 대장암 데이터를 사용하여 실험을 하였다. 마지막으로 상관분석에 기반을 둔 피어슨 상관계수와 스피어맨 상관계수 방법과 비교하여 전진 선택법에 의한 특징 선택방법의 우수성을 평가하였다.

2. 배경

2.1 DNA microarray

DNA microarray는 칩을 제작하는 방식에 따라 크게 hybridization에 의한 방법과 sequencing에 의

한 방법으로 나뉜다. Hybridization에 의한 대표적인 방법은 cDNA microarray, oligonucleotide microarray가 있는데 본 논문에서 사용한 대장암 데이터는 oligonucleotide microarray를 사용하여 얻은 유전 발현 데이터이다.

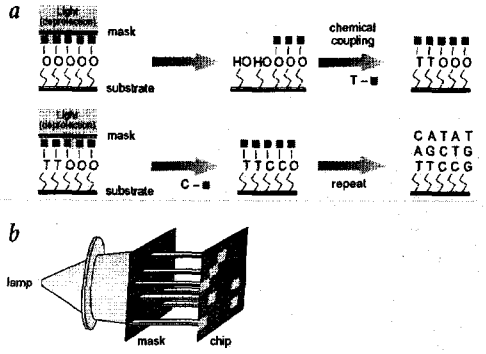

그림 1. Oglinonucleotide microarray

그림 1은 Photolithography를 이용한 DNA 칩 제작과정을 보여준다. 즉, 칩 표면에 빛에 불안정한 보호기(基)를 보유하는 링커 분자를 입히고, 유전자 조각을 심을 부분을 mask를 이용하여 부분적으로 빛을 투과함으로써 보호기를 제거한다. 보호기가 제거된 부분에 광활성 부위에서만 융합하는 광보호 nucleotide에 빛을 투과시켜 줌으로써 nucleotide를 심는다. 이를 반복함으로써 유전자 조각의 길이가 대략 20-25 mers가 되도록 칩을 제작한다.

2.2 관련 연구

2.2.1 특징 선택 방법

DNA miroarray의 데이터양은 매우 방대하기 때문에 암의 분류에 있어서 효율적으로 필요한 유전자만을 선택하는 방법은 매우 중요하다. 지금까지 유전자 선택에 사용된 특징 선택 방법들을 살펴보면 다음과 같다.

유전자들 간의 상관계수를 측정하여 암의 분류에 관여하는 유전자들을 선택하는 피어슨상관계수나 스피어맨 상관계수 방법, 유사도 측정기반의 유클리디안 거리와 코사인계수를 사용한 방법[3], SVD를 사용한 유전자들의 차원을 줄이는 방법[4], GA와 KNN을 사용하여 암의 유무를 구별하는 유전자들을 찾아내는 방법 등이 있다.

이 중에서도 상관분석에 기반을 둔 피어슨계수와 스피어맨계수 방법을 본 논문에서 사용한 회귀분석 방법과 비교하고자 한다. 피어슨계수와 스피어맨계수방법은 변수들 간의 유사한 정도를 계산하여 목표 변수와 가장 비슷한 패턴을 가진 변수들을, 가장 비슷한 정도가 높은 순위부터 차례대로 선택하는 방법이다.

$$r_{pearson}=\frac{\Sigma XY-\frac{\Sigma X\Sigma Y}{N}}{\sqrt{(\Sigma X^2-\frac{(\Sigma X)^2}{N})(\Sigma Y^2-\frac{(\Sigma Y)^2}{N})}} \quad (1)$$

식 1은 X와Y의 피어슨상관계수를 구하는 것으로, 여기서 상관계수 r은 [-1, 1]의 값을 가진다. 1의 값에 가까울수록 X와 Y는 양의 상관관계를 갖는 것이며, -1에 가까울수록 두 변수는 음의 상관관계를 갖는 것이다.

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

식 2는 변수의 순위배열을 사용하여 변수간의 상관관계를 분석하는 스피어맨 상관계수이다. 여기서 상관계수 r은 스피어맨계수와 마찬가지로 [-1, 1]의 값을 가지며, X와 Y의 순위배열 Dx와 Dy를 사용하여 그 값을 구한다.

상관분석에 의한 특징 선택 방법은 가장 널리 사용되고 있지만, 선택되는 변수들간의 관계를 고려하지 않는다는 단점이 있다. 상관계수가 매우 높은 변수들이 선택되더라도 실제로는 상당 부분 중복된 정보를 가진 변수들의 집합이 될 수 있는 가능성이 있다. 반면에 회귀분석에 기반을 둔 전진 선택법은 선택되는 변수들 간의 관계를 고려하여, 중복되는 정보를 최소화시킨다. 그러므로 암의 분류에 있어서 보다 유용한 정보를 가진 유전자들의 집합을 찾아낼 수 있다.

2.2.2 회귀모형분석

회귀모형분석이란, 특정 변수와 특정 변수를 가장 잘 설명할 수 있는 변수들 사이의 관계를 분석하는 기법이다. 하나의 목표 변수를 정하고, 그 변수에 영향을 미치는 독립변수들을 찾아낸다. 이때 모형을 설명하는 독립변수가 하나인 경우 단순회귀모형을 사용하고, 다수인 경우 다중회귀모형을 사용한다. 유전 발현 정보 데이터를 회귀모형에 적용시킬 경우, 유전자의 개수가 많기 때문에 다중회귀모형을 사용

$$y=\beta_0+\beta_1 x_i+\varepsilon, \quad i=1,\ldots,n \quad (3)$$

하며, 목표변수는 암의 유무를 나타낸다.

식 3은 단순회귀모형으로 목표변수 y와, 이를 설명하는 변수x로 표현된다. 절편 β_0와 기울기 β_1은 모수로 미지의 상수이다. 이들을 추정하기 위해서는

목표변수 y와 x의 관측 값들이 필요하다. ε은 평균이 0, 분산이 σ^2인 정규분포를 따른다.

회귀모형에서 목표변수를 설명하는 변수들을 선택하는 기준은 R^2값에 의해서 결정된다.

$$R^2 = \frac{SSR}{SSTO} \quad (4)$$

SSR은 모형에 의하여 설명될 수 있는 변동량이고 $SSTO$는 목표 변수 y에 의한 총 변동량을 나타낸다. 그렇기 때문에 목표변수를 잘 설명하는 변수들은 R^2의 값이 크다. 이러한 방법으로 R^2값이 큰 순서대로 목표변수를 잘 설명할 수 있는 독립변수들이 선택된다.

3. 방법

3.1 전진 선택법

전진 선택법은 다중회귀분석에 기반을 둔 방법으로 목표변수에 대한 기여도에 따라 변수를 선택한다[3]. 이 방법을 유전 발현 정보 데이터에 응용시켜서 암의 유무에 관련된 정보를 가진 유전자들을 선택하는 방법을 제안한다. 새로운 유전자가 선택될 때마다 먼저 선택된 유전자와의 관계를 파악하여 그 중요도에 따라 선택여부를 판단하는 방법으로 다음의 절차를 따른다.

표 1. 유전자 선택 알고리즘

단계1) n개의 유전자x에 대하여 회귀모형 y를 적합시켜 R^2값을 가장 크게 하는 유전자 x를 선택한다. $$y = \beta_0 + \beta_1 x_i + \epsilon, \quad i = 1, \ldots, n$$
단계2) 단계 1에서 선택한 유전자 x를 x_j에 넣는다.
단계3) 단계 2에서 선택한 유전자(x_j)와 나머지 모든 유전자(x_k)에 대하여 두 유전자 쌍(x_j,x_k)에 대한 회귀모형을 적합시켜 R^2값을 가장 크게 하는 x_k를 선택한다. $$y = \beta_0 + \beta_1 x_j + \beta_2 x_k + \epsilon, \quad j \neq k$$
단계4) 위의 과정을 $R^2>0$을 만족하는 x가 없을 때까지 반복한다.

표 1에서 x는 각각의 유전자를 뜻하고 y는 암인지 아닌지를 나타내는 목표변수이다. $R^2>0$으로 정한 것은 암의 유무를 설명할 수 있는 정보를 가진 유전자들을 선택하기 위해 정한 임계값이다. 위의 알고리즘은 상관분석방법과는 달리 유전자들 간의 관계까지도 고려하여 서로 중복되는 정보를 최소화시키는 유전자들의 집합을 만들어낸다.

3.2 K-최근접 이웃(KNN)

K-최근접 이웃은 메모리 기반 방식의 가장 널리 쓰이는 분류기이다. KNN의 동작원리는 간단하다. 테스트 샘플이 입력되면 이것과 각 학습 샘플과의 유사도를 계산하고 그 중 k개의 가장 가까운 학습 샘플을 선택한다[2]. 샘플간의 유사도 계산에는 피어슨 계수를 사용하였다.

$$P(X, c_j) = \sum_{d_i \in kNN} Sim(X, d_i) P(d_i, C_j) - b_j \quad (5)$$

식 5는 입력 X가 클래스 c_j로 분류될 확률, $P(X,c_j)$를 구하는 식이다. $Sim(X,d_i)$는 입력 X와 d_i 간의 유사도 측정값이고, d_i는 학습 샘플들이다.

4. 실험 및 결과

4.1 실험 환경

실험 데이터로는 대장암 데이터(Colon)를 사용하였다. 대장암 데이터는 2000개의 유전자로 이루어져 있으며 62개의 샘플(31개:학습, 31개:테스트)이 사용되었다. 먼저, 31개의 학습 샘플을 가지고 전진 선택법을 이용하여 암의 유무를 잘 설명할 수 있는 유전자들을 선택하였다. 표 1의 알고리즘을 사용한 결과, $R^2>0$을 만족하는 유전자가 총 18개가 선택되었다. 이렇게 선택된 18개의 유전자를 가지고 KNN을 학습시켰다. KNN에서 k값의 범위는 1~8까지 변화시키면서 8번씩 실험을 하였고, 그 중에서 가장 높은 인식률을 최종 결과 값으로 선택하였다. 유사도 측정 도구로는 코사인 계수를 사용하였다.

4.2 실험 결과 및 분석

먼저 피어슨계수, 스피어맨계수방법으로 대장암 데이터의 2000개의 유전자중에서 18개의 유전자를 선택하였다. 18개를 선택한 이유는 전진 선택법에서 선택된 18개의 유전자와 비교하기 위해서이다. 평가 척도로는 민감도(sensitivity), 특이도(specificity), 그리고 인식률(recognition rate)을 사용하였다. 민감도는 테스트 샘플 중에서 암인 샘플을 올바르게 암으로 분류한 샘플의 비율이고, 특이도는 테스트 샘플 중에서 암이 아닌 샘플을 올바르게 암이 아닌 샘플로 분류한 비율이다.

표 2. 각 선택 방법에 대한 결과값(%)

	민감도	특이도	인식률
Pearson	75	82	77.4
Spearman	100	9	67.7
Forward	**95**	**82**	**90.3**

표 2는 각 방법에 대한 세 가지 평가 척도값들을

표로 나타내었고, 그림 2는 이 값들을 그래프로 표현한 결과들이다. 이 결과들을 비교해보면 전진 선택법의 경우, 다른 방법들에 비해 세 가지 척도에서 모두 높은 값들을 나타내는 것을 볼 수 있다.

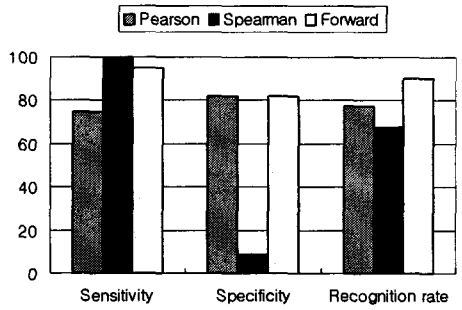

그림 2. 각 방법에 대한 결과값

스피어맨의 경우 민감도가 100으로 전진 선택법의 경우(95)보다 높지만, 특이도가 9로 전진 선택법의 82에 훨씬 미치지 못한다. 즉 암인 샘플은 잘 분류하지만, 암이 아닌 샘플의 경우 거의 분류하지 못한다는 것을 알 수 있다. 피어슨계수의 경우는 반대로 특이도가 민감도 보다 정확하여 스피어맨과는 달리 암이 아닌 샘플을 더 정확하게 분류하였다. 전진 선택법의 경우, 암인 샘플이나 암이 아닌 샘플 모두 대체적으로 잘 분류하며, 세 가지 평가 척도에서 모두 고르게 높은 수치를 나타내는 것을 볼 수 있다.

표 3. 선택된 유전자들의 혼돈행렬

Pearson			Spearman			Forward		
P \ A	0	1	P \ A	0	1	P \ A	0	1
0	9	5	0	1	0	0	**9**	1
1	2	15	1	10	20	1	2	**19**

표 3은 선택된 유전자들의 혼돈행렬로서 각 유전자 선택 방법에 대하여 실제 샘플이 암인 경우는 1로 표시하였고, 암이 아닌 경우는 0으로 표시하였다(열부분, A로 표시). 그리고 분류기에 의해 예측된 결과값은(행부분, P로 표시) 똑같이 암인 경우 1, 암이 아닌 경우 0으로 표시하여 행렬을 구성하였다. 전진 선택법의 경우, 암인 샘플을 암으로 제대로 예측한 샘플이 19개(19/20), 암이 아닌 샘플을 암이 아닌 경우로 예측한 샘플이 9개(9/11)로 암의 분류에 관련된 정보를 가진 유전자들을 잘 선택하였다는 것을 알 수 있다.

그림 3은 전진 선택법으로 선택되는 유전자들의 인식률을 나타낸 그래프이다. 하나의 유전자를 선택한 경우부터 시작하여, 전진 선택법으로 유전자를 한 개씩 추가시키는 과정을 살펴보았다. 맨 처음 하나의 유전자만을 선택한 경우는 매우 낮은 인식률(48.4%)을 보였지만, 선택되는 유전자의 개수가 증가하면서 인식률이 증가되는 것을 확인할 수 있다. 실제로 15~17개 사이의 유전자가 선택되었을 때 가장 높은 인식률(90.3%)을 보였다.

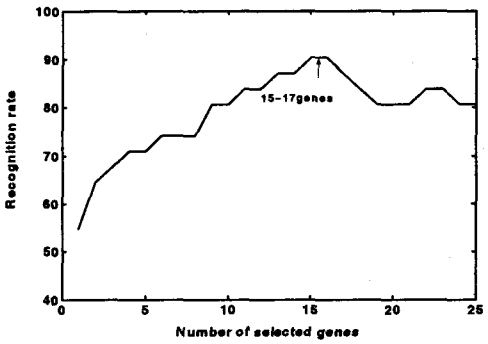

그림 3. 유전자의 개수에 따른 인식률 변화

전진 선택법에 의하여 선택된 유전자의 개수(18개)와 거의 같은 수를 가진 유전자들의 집합이 최고 인식률을 가지기 때문에 전진 선택법에 의한 유전자들의 선택은 암의 분류에 있어 중요하다고 볼 수 있다. 전진 선택법에 의한 특징 선택 방법이 상관분석 기반의 특징 선택 방법보다 암의 분류에 있어, 효과적으로 유전자를 선택하는 것을 확인하였다.

감사의 글

본 연구는 보건복지부 보건의료기술 진흥사업의 지원에 의하여 이루어진 것임.

참고문헌

[1] T. R. Golub, et al., "Molecular classification of cancer class discovery and class prediction by gene-expression monitoring," *Science*, vol. 286, no. 15, pp. 531-537, October 1999.

[2] S-B. Cho and J. 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.

[3] 성웅현, 이승천, *회귀분석*, 범문사, 2001.

[4] T. H. Bo and I. Jonassen, "New feature subset selection procedures for classification of expression profiles," *Genome Biology*, vol. 3, no. 4, research0017.1-0017.11, 2002.