논문 2004-41SP-3-9

전진선택법에 의해 선택된 부분 상관관계의 유전자들을 이용한 암 분류

(Classifying Cancer Using Partially Correlated Genes Selected by Forward Selection Method)

유 시 호*, 조 성 배**

(Si-Ho Yoo and Sung-Bae Cho)

요 약

유전 발현 데이터는 생명체의 특정 조직에서 채취한 샘플을 마이크로어레이상에서 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. 그러나 분류에 모든 유전자가 관여하지는 않으므로 효율적인 암의 분류를 위해서는, 관련성 있는 소수의 유전자만을 선별해내는 작업인 특징선택 방법이 필요하다. 본 논문에서는 회귀분석의 변수선택방법중 하나인 전진 선택법(forward selection method)을 사용하여 유전자들을 선택하고 분류하는 방법을 제안한다. 이 방법은 선택되는 유전자들의 중복된 정보를 최소화시켜 암의 분류에 있어 보다 효과적인 유전자 선택을 한다. 실험데이터는 대장암 데이터(Colon cancer dataset)를 사용하였고, 분류기는 k-최근접 이웃(KNN)을 사용하였다. 이 방법과 상관계수를 이용한 특징 선택 방법인 피어슨 상관계수와 스피어맨 상관계수방법과 비교해본 결과 전진 선택법에 의한 특징선택 방법이 암의 분류에 있어서 더 효과적인 유전자 선택을 한다는 사실을 확인하였다. 실험결과 90.3%의 높은 인식률을 보였다. 추가적으로 림프종 데이터에 대한 실험을 하였고, 그 결과 전진 선택법의 유용성을 확인할 수 있었다.

Abstract

Gene expression profile is numerical data of gene expression level from organism, measured on the microarray. Generally, each specific tissue indicates different expression levels in related genes, so that we can classify cancer with gene expression profile. Because not all the genes are related to classification, it is needed to select related genes that is called feature selection. This paper proposes a new gene selection method using forward selection method in regression analysis. This method reduces redundant information in the selected genes to have more efficient classification. We used k-nearest neighbor as a classifier and tested with colon cancer dataset. The results are compared with Pearson's coefficient and Spearman's coefficient methods and the proposed method showed better performance. It showed 90.3% accuracy in classification. The method also successfully applied to lymphoma cancer dataset.

Keywords: gene expression profile, feature selection, forward selection method, classification, regression analysis

I. 서 론

최근 몇 년간 암의 정확한 분류를 위한 연구가 활발하게 진행되어왔다. 하지만, 수 천개의 유전자로 이루어진 적은 샘플을 올바르게 분류하기는 쉽지 않다. 컴퓨터와 DNA 마이크로어레이 기술의 발달로 생명체에 관한 대량의 유전정보를 얻는 것은 가능하게 되었지만 이러한 대량의 유전정보가 암의 정확한 분류를 하는데 모두 필요한 것은 아니기 때문에 필요한 유전자들만을 선택하는 특징선택 방법이 필요하다[1].

상관계수를 이용한 방법, 유전자들 간의 유사도를 측정한 특징선택 방법, 전체 데이터로부터 의미있는 정보를 선택하는 정보이득이나 상호정보 등 매우 다양한 방법들이 암의 분류를 위한 유전자 선택에 이용되었다[2]. 하지만 이러한 연구들은 순위기반(rank-based) 방식으로서 선택되는 유전자와 목표 벡터간의 일대일 관계만

* 학생회원, ** 정회원, 연세대학교 컴퓨터과학과
(Department of Computer Science, Yonsei University)
※ 이 논문은 연세대학교 생체인식연구센터(BERC)를 통하여 한국과학재단(KOSEF)에 의해 지원받았음.
접수일자: 2003년7월21일, 수정완료일: 2004년3월17일

을 고려하였다. 목표 변수와 가장 비슷한 패턴을 가진 유전자를 먼저 선택하고 그 다음으로 목표변수를 가장 잘 설명하는 유전자를 선택하는 순서로 유전자들을 선택하였기 때문에, 선택된 유전자들끼리 중복된 정보를 가질 가능성이 있다.

본 논문에서는 회귀분석에 기반을 둔 전진 선택법을 사용하여 유전자들을 선택하는 새로운 방법을 제안한다. 유전자 선택에 있어, 각각의 유전자와 목표변수만의 일대일 관계가 아니라 선택되는 유전자들의 부분적인 상관관계를 고려하여 유전자들을 선택한다. 이러한 방법에 의해 선택된 유전자들은 서로 다른 정보를 가진 유전자가 먼저 선택되기 때문에 선택된 유전자들내의 중복된 정보를 최소화시킨다. 선택된 유전자들내의 중복된 정보를 최소화시킨 만큼, 그 유전자들의 조합은 단순한 순위기반에 의한 선택방법보다 암의 분류에 있어 더 많은 정보를 가질 수 있을 것이다[3].

선택된 유전자들은 분류기의 입력으로 들어가는데, 분류기는 학습 집단의 유전자들에 대하여 입력패턴이 최대한 올바른 출력을 내도록 학습된다. 학습된 분류기는 테스트 집단에 대하여 실제로 얼마나 정확한지 평가받는다. 기존 암의 분류에 사용된 분류기로는 다층신경망(MLP)[4], k-최근접 이웃(KNN)[5], SVM(Support vector machine)[6], 자기구성지도(Self-organizing map)[7] 등이 있다. 본 논문에서는 전진 선택법으로 선택된 유전자들을 k-최근접 이웃으로 분류한다. 유전발현 데이터는 대장암 데이터를 사용하였고, 성능을 평가하기 위해서 상관분석에 기반을 둔 피어슨 상관계수와 스피어맨 상관계수 방법과 비교하였다.

본 논문의 나머지 부분은 다음과 같다. Ⅱ장에서는 연구의 배경이 되는 DNA 마이크로어레이 기술과 여러 가지 특징선택 방법들에 대하여 알아본다. Ⅲ장에서는 제안하는 전진 선택법의 유전자 선택 알고리즘과 그 적용방법에 대해 기술하고 k-최근접 이웃에 대해 설명한다. Ⅳ장은 실험과정과 결과를 설명하고, Ⅴ장은 결론과 향후 연구에 대해 언급한다.

Ⅱ. 관련 연구

1. DNA 마이크로어레이

DNA 마이크로어레이는 용액이 투과되지 않는 딱딱한 지지체 위에 고밀도로 cDNA를 고정시켜 놓은 것으로 DNA 칩이라고도 한다. 마이크로어레이상의 각 셀은 두 개의 다른 환경에서 채집된 유전물질에 녹색의 Cy3와 빨간색의 Cy5라는 각각 다른 형광물질을 합성한 것을 동일한 양으로 보합한 것이다. 이것을 레이저 형광스캐너로 읽어 들이면 녹색부터 빨간색에 이르는 발현정도를 얻을 수 있는데, Cy5/Cy3의 비율에 밑이 2인 로그를 취한 값을 그 셀의 발현정보 값으로 얻게 된다.

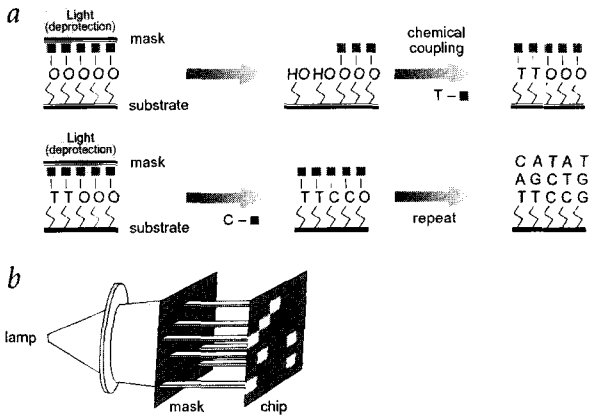


그림 1. 올리고뉴클리오타이드 마이크로어레이[8]
Fig 1. Oligonucleotide microarray[8]

$$\text{gene expression} = \log_2 \frac{Int(Cy5)}{Int(Cy3)} \quad (1)$$

본 논문에서 사용한 대장암 데이터를 만드는데 사용한 것은 올리고뉴클리오타이드 마이크로어레이(Oligo-nucleotide microarray)방식으로 그림 1과 같다. cDNA 방식과는 달리 빛의 투과성을 이용하여 유전 발현 데이터를 만든다. 빛에 불안정한 보호기(基)를 보유하는 링커 분자를 칩 표면에 입히고, 마스크를 이용하여 유전자 조각을 심을 부분에 부분적으로 빛을 투과함으로써 보호기를 제거한다. 보호기가 제거된 부분에 광활성 부위에서만 융합하는 광보호 뉴클리오타이드(nucleotide)에 빛을 투과시켜 줌으로써 nucleotide를 심는다. 이를 반복함으로써 유전자 조각의 길이가 대략 20~25 mers가 되도록 칩을 제작한다[8].

2. 특징 선택 방법

DNA 마이크로어레이의 데이터양은 매우 방대하기 때문에 효율적으로 필요한 유전자만을 선택하는 방법은 암의 분류에 있어서 매우 중요하다. 지금까지 유전자 선택에 사용된 특징 선택 방법들을 살펴보면 유전자들 간의 상관계수를 측정하여 암의 분류에 관여하는 유전

자들을 선택하는 피어슨상관계수나 스피어맨 상관계수 방법, 유사도 측정기반의 유클리디안 거리와 코사인계수를 사용한 방법, 전체 데이터로부터 의미 있는 정보를 선택하는 정보이득이나 상호정보[2]등이 있다. 또한 각 유전자를 두개씩 묶어서 암의 유무를 얼마나 잘 구별하는가에 따라 순위를 매겨서 순위가 높은 유전자 쌍을 선택하는 방법[9], 유전자 알고리즘과 KNN을 사용하여 암의 유무를 구별하는 유전자들을 찾아내는 방법[5] 등이 있다. 특징의 차원을 줄이는 방법으로 SVD (Single Value Decomposition)를 사용한 그룹도 있고[10] PCA (Principle Component Analysis)를 사용하여 유전자의 수를 줄인 그룹도 있다[11]. 또 좋은 정보를 가진 유전자들을 선택하기 위해서 베이지안 변수선택 방법을 사용한 연구도 있다[12].

이 중에서도 피어슨계수와 스피어맨계수 방법과 같은 상관분석방법은 회귀분석과 더불어 통계학에서 사용되는 대표적인 분석방법으로 두 변수간의 상호관계를 분석하는데 사용된다. 피어슨계수와 스피어맨계수방법은 변수들 간의 유사한 정도를 계산하여 목표변수와 가장 비슷한 패턴을 가진 변수들을, 가장 비슷한 정도가 높은 순위부터 차례대로 선택하는 방법이다.

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

식 (2)는 X와Y의 피어슨상관계수를 구하는 것으로, 여기서 상관계수 r은 [-1, 1]의 값을 가진다. 1의 값에 가까울수록 X와 Y는 양의 상관관계를 갖는 것이며, -1에 가까울수록 두 변수가 음의 상관관계를 갖는 것이다. 또한 r이 0에 가깝다면 두 변수 사이에 별로 관계가 없음을 의미한다[2].

$$r_{spearman} = 1 - \frac{6\sum(Dx - Dy)^2}{N(N^2-1)} \quad (3)$$

식 (3)은 변수의 순위배열을 사용하여 변수간의 상관관계를 분석하는 스피어맨 상관계수이다. 스피어맨 상관계수는 변수들의 값을 직접 이용하는 모수 분석과는 달리 변수들이 양적 변수가 아닌 경우에도 이용할 수 있는 비모수 분석방법이다. 여기서 상관계수 r은 스피어맨계수와 마찬가지로 [-1, 1]의 값을 가지며, X와 Y의 순위배열 D_x와 D_y를 사용하여 그 값을 구한다[2].

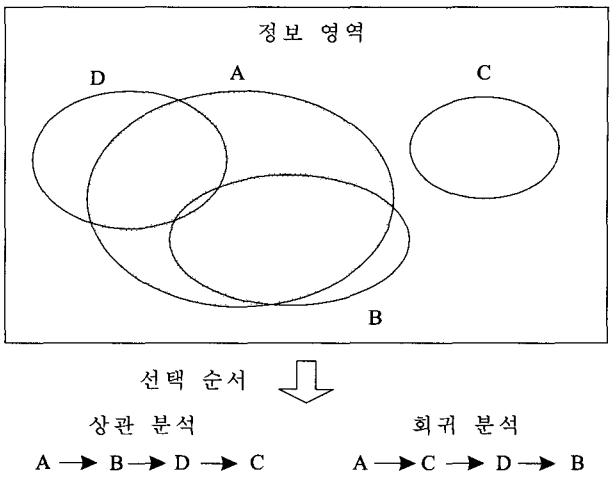


그림 2. 상관분석과 회귀분석의 비교
Fig. 2. Comparison between correlation and regression analysis

상관분석에 의한 특징선택 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 가능성이 있다.

III. 전진 선택법을 이용한 암 분류

1. 회귀모형

회귀모형이란 특정 변수와 이를 가장 잘 설명할 수 있는 변수들 사이의 관계를 분석하는 기법이다. 상관분석과는 달리 여러 개의 변수들이 다른 변수에 미치는 영향의 정도를 파악하거나 예측할 수 있다[13]. 하나의 목표 변수를 정하고, 그 변수에 영향을 미치는 독립변수들을 찾아낸다. 이때 모형을 설명하는 독립변수가 하나인 경우 단순회귀모형을 사용하고, 다수인 경우 다중회귀모형을 사용한다. 유전발현 데이터를 회귀모형에 적용시킬 경우, 유전자의 개수가 많기 때문에 다중회귀모형을 사용하며, 목표변수는 암의 유무를 나타낸다. 회귀분석의 기본적인 개념은 그림 2와 같다.

어떤 정보를 표현하는 사각형이라는 영역이 있을 때 이 정보의 영역을 얼마나 많이 표현하는가에 따라서 변수들을 선택하는 것이 상관분석이다. 그림 2에서 A가 가장 넓은 영역을 표현하므로 상관분석의 경우는 A를 선택하고, 그 다음으로 큰 영역의 타원인 B를 선택한다. 하지만 회귀분석의 경우, 처음에 가장 큰 타원인 A를 선택하는 것은 상관분석과 같지만 B는 상당부분 A가

나타내는 부분과 겹치기 때문에 오히려 C를 선택하게 된다. 즉 먼저 선택된 변수가 표현할 수 없는 영역의 크기를 가지고 새로운 변수를 선택한다. 그렇기 때문에 상관분석의 경우 (A→B→D→C)의 순으로 변수를 선택하고, 회귀분석은 (A→C→D→B)의 순으로 변수를 선택한다. 이러한 방법으로 회귀분석은 선택되는 변수들의 부분상관관계를 고려하여 중복되는 정보를 최소화한다.

$$y = \beta_0 + \beta_1 x_i + \varepsilon \quad i = 1,2,...,n \tag{4}$$

식 (4)는 단순회귀모형으로 목표변수 y와 이를 설명하는 변수 x로 표현된다. 절편 β_0와 기울기 β_1은 모수로 미지의 상수이다. 이들을 추정하기 위해서는 목표변수 y와 x의 관측 값들이 필요하다. ε은 평균이 0 이고 분산이 σ^2인 정규분포를 따른다. 회귀모형에서 목표변수를 설명하는 변수들을 선택하는 기준은 R^2값에 의해서 결정된다.

$$R^2 = \frac{SSR}{SSTO} \tag{5}$$

SSR은 모형에 의하여 설명될 수 있는 변동량이고 $SSTO$는 목표변수 y에 의한 총 변동량을 나타낸다. 그렇기 때문에 목표변수를 잘 설명하는 변수들은 R^2값이 크다. 이러한 방법으로 R^2값이 큰 순서대로 목표변수를 잘 설명할 수 있는 변수들이 선택된다. 회귀모형의 검증은 F-검정을 사용한다. 각 모형의 F-값을 구해 회귀모형의 적합성을 평가하고 유의수준을 기준으로 해당 모형의 선택 여부를 결정한다[13].

2. 유전자 선택 알고리즘

전진 선택법은 다중회귀분석에 기반을 둔 방법으로 목표변수에 대한 기여도에 따라 변수를 선택한다[13]. 가장 중요한 변수부터 하나씩 골라가면서 더 이상 중요한 변수가 없다고 판정될 때까지 변수들을 선택하는 방법이다. 따라서 이 방법을 유전자 선택에 이용하면 암의 유무를 나타내는 정보를 가진 변수를 목표변수로 잡고, 이 목표변수를 잘 설명하는 유전자들을 하나씩 증가시켜가면서 선택할 수 있다.

그림 3은 본 논문에서 제안하는 유전자 선택 알고리즘으로 $Model(x)$는 유전자 x에 대한 회귀모형을 뜻하고, x_G는 선택된 유전자들의 집합, $Max_R^2(x_i)$는 만들어진 회귀모형들의 R^2값 중 가장 큰 값을 뜻한다. 즉 유

```
procedure
 var       N: number of genes
           G: selected genes
 function  Model(x): make a regression model of x
 begin
   for i=1 to N
        Model(x_i)
   find Max_R^2(x_i) and put x_i into G
   do
        for i=1 to N
             Model(x_G, x_i),  x_G ≠ x_i
        find Max_R^2(x_i) and put x_i into G
   while Max_R^2(x_i)>0
 end
```

그림 3. 유전자 선택 알고리즘
Fig. 3. Gene selection algorithm

전자 x에 대하여 목표변수에 대한 회귀모형을 만들고 그 중에서 가장 목표변수를 잘 설명하는 유전자를 선택하는 것이다.

목표변수는 암인 샘플은 1, 정상인 샘플은 0으로 한 이상벡터로 설정한다. 처음에 총 N개의 유전자들에 대해 N개의 회귀모형을 만들고 각각의 R^2값을 모두 계산하여 가장 큰 R^2값을 가진 모형의 유전자를 선택한다. 그리고 두 번째 반복부터는 먼저 선택된 유전자들을 제외한 나머지 유전자들과 먼저 선택된 유전자들과 결합한 다중회귀모형을 만든다. 그리고 이 모형들 중에서 가장 큰 R^2값을 가지는 유전자를 추가로 선택하여 x_G에 포함시킨다. 이때 R^2값이 0보다 크면 선택을 하고 0보다 작거나 같으면 알고리즘은 멈춘다. 이 알고리즘에 의해 얻어진 유전자들의 집합 x_G는 상관분석방법과는 달리 유전자들 간의 부분상관관계까지도 고려하여 서로 중복되는 정보를 최소화시킨 유전자들의 집합이다.

3. k-최근접 이웃(k-Nearest Neighbor)

선택된 유전자의 집합을 메모리 기반 방식으로 가장 널리 쓰이는 분류기인 k-최근접 이웃으로 분류한다. KNN은 테스트 샘플이 입력되면 이것과 각 학습 샘플과의 유사도를 계산하고 그 중 k개의 가장 가까운 학습 샘플을 선택한다[5]. 이 k개중 많은 수를 차지하는 부류에 속하는 것으로 판정한다. 일반적으로 이진 분류에 있어서 애매함을 방지하기 위하여 k값은 홀수를 사용한다. 샘플간의 유사도 계산에는 피어슨 계수를 사용하였다.

$$P(X, c_j) = \sum_{d_i \in kNN} Sim(X, d_i) P(d_i, C_j) - b_j \tag{6}$$

식 (6)은 입력 X가 클래스 c_j로 분류될 확률 $P(X,c_j)$를 구하는 식이다. $Sim(X,d_i)$는 입력 X와 d_i간의 유사도 측정값이고, d_i는 학습 샘플들이다. 유사도는 코사인 계수로 측정하였다.

Ⅳ. 실험 및 결과

실험 데이터로는 대장암 데이터를 사용하였고 추가적으로 림프종 데이터에 대한 실험을 하였다. 대장암 데이터(http://www.sph.uth.tmc.edu:8052/hgc/default.asp)는 2000개의 유전자로 이루어져 있으며 62개의 샘플을 31개의 학습 샘플과 31개의 테스트 샘플로 나누어 실험하였다. 먼저, 31개의 학습 샘플을 가지고 전진 선택법을 이용하여 암의 유무를 잘 설명할 수 있는 유전자들을 선택하였다. 그림 3의 유전자 선택 알고리즘을 사용한 결과, $R^2>0$을 만족하는 유전자가 총 18개가 선택되었다. 이렇게 선택된 18개의 유전자를 가지고 KNN을 학습시켰다. KNN에서 k값의 범위는 1~10까지 변화시키면서 실험을 하였고, 그 중에서 가장 높은 인식률을 최종 결과 값으로 하였다.

림프종 데이터(http://llmpp.nih.gov/lymphoma/)는 총 4026개의 유전자로 이루어져 있으며 47개의 샘플을 22개의 학습 샘플과 25개의 테스트 샘플로 나누어 실험하였다. 이중 24개는 GC B-like DLBCL이고, 23개는 activated B-Like DLBCL이다. 림프종 데이터의 경우 그림 3의 유전자 선택 알고리즘에 의해 총 10개의 유전자가 $R^2>0$을 만족하여 선택되었다. 대장암 데이터와 마찬가지로 이렇게 선택된 10개의 유전자를 가지고 KNN을 학습시켰다.

1. 대장암 데이터 분석

먼저 피어슨계수와 스피어맨계수방법으로 대장암 데이터의 2000개 유전자중에서 상위 18개의 유전자를 선택하였다. 그리고 전진 선택법으로 선택된 18개의 유전자와 비교 실험하였다. 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 1은 전진 선택법에 의해 선택된 유전자들을 설명한 표로, 선택되는 순서에 따라서 유전자들을 나열해보았다. 총 18개의 유전자의 R^2값이 0보다 크게 나왔다. 표 2는 전진 선택법에 의해 선택된 18개 유전자들의 유

표 1. 전진선택법에 의해 선택된 유전자: 대장암
Table 1. Genes selected by forward selection method: Colon

순위	유전자ID	유전자 설명
1	R8712	MYOSIN HEAVY CHAIN, NONMUSCLE (Gallus gallus)
2	U0202	Human pre-B cell enhancing factor (PBEF) mRNA, complete cds.
3	U3662	Human Y-chromosome RNA recognition motif protein (YRRM) gene, exon 12, partial cds, subclone 7S2.
4	H6253	SPORE GERMINATION PROTEIN B2 (Bacillus subtilis)
5	T7102	Human (HUMAN)
6	H5607	GTP CYCLOHYDROLASE I (Homo sapiens)
7	T9947	GLUCOSE-6-PHOSPHATASE (Homo sapiens)
8	J0014	Human dihydrofolate reductase pseudogene (psi-hdl).
9	M2821	Homo sapiens low density lipoprotein receptor (FH 10 mutant causing familial hypercholesterolemia) mRNA, 3′ end.
10	H2475	FRUCTOSE-BISPHOSPHATE ALDOLASE A (HUMAN);.
11	R4985	COAGULATION FACTOR V PRECURSOR (Homo sapiens)
12	T9855	DNA-DIRECTED RNA POLYMERASES I AND III 16 KD POLYPEPTIDE (Saccharomyces cerevisiae)
13	T4964	MYRISTOYLATED ALANINE-RICH C-KINASE SUBSTRATE (Homo sapiens)
14	T6109	ENDOGLIN PRECURSOR (Homo sapiens)
15	M8473	Human autoantigen calreticulin mRNA, complete cds.
16	H6439	CALCINEURIN B SUBUNIT ISOFORM 1 (Homo sapiens)
17	T7258	GLUTAMATE RECEPTOR 5 PRECURSOR (Homo sapiens)
18	H1506	PROTEIN KINASE CLK (Mus musculus)

표 2. 전진선택법에 의해 선택된 유전자들: 대장암
Table 2. Selected genes by forward selection method: Colon

순위	유전자 번호	R^2값	F-값	Pr>F
1	gene493	0.6344	50.33	<.0001
2	gene1147	0.1549	20.58	<.0001
3	gene1927	0.0559	9.74	0.0043
4	gene1587	0.057	15.15	0.0006
5	gene66	0.0322	12.29	0.0017
6	gene1427	0.0218	11.99	0.002
7	gene597	0.0157	12.94	0.0015
8	gene1919	0.0133	19.93	0.0002
9	gene1584	0.0053	11.94	0.0024
10	gene55	0.0031	9.74	0.0054
11	gene459	0.0028	14.74	0.0011
12	gene1340	0.0019	19.84	0.0003
13	gene2000	0.0007	13.09	0.0021
14	gene955	0.0004	10.26	0.0055
15	gene287	0.0002	8.78	0.0097
16	gene92	0.0002	11.68	0.0042
17	gene332	0.0001	14.83	0.002
18	gene858	0.0001	20.94	0.0006

전자 번호(gene number), R^2값, F-값, 그리고 F-값의 유의수준을 나타낸 표이다.

맨 처음 선택된 유전자 R8712는 MYOSIN HEAVY CHAIN, NONMUSCLE (Gallus gallus)로, 표 2에서 보면 0.6344의 R^2값을 가진 것으로 보아 약 63%의 매우 큰 비중으로 암의 분류에 대한 정보를 가짐을 알 수 있다. 표 2에서 이 유전자를 보면, 493번째 유전자로 유의수준 또한 0.0001보다 적은 것으로 보아 신뢰도 또한 매우 높다고 할 수 있다. 두 번째 선택된 유전자 U0202는 Human pre-B cell enhancing factor (PBEF) mRNA, complete cds로 0.1549의 R^2값을 가진 것으로 보아 첫 번째 선택된 유전자보다는 가진 정보량이 적지만, 유의수준이 역시 높기 때문에 신뢰도 또한 높다. 하지만 15번째 이후로 선택된 유전자들은 가지고 있는 정보량(R2)이 매우 적고, 신뢰도 또한 매우 낮은 걸 알 수 있다.

그림 4는 전진선택법으로 선택되는 유전자들을 하나씩 증가시켜가면서 민감도, 특이도, 인식률의 변화과정을 나타낸 그림이다. 선택되는 유전자가 하나씩 증가하면서 세 가지 평가값들이 대체적으로 증가하는 추세를 보이다가, 15개 근처에서 모두 최고 수치를 보였다. 민감도의 경우 15와 17개에서 최고치를 보였고, 특이도와 인식률의 경우는 15개일 때가 가장 높은 결과를 보였다. 표 2에서 17, 18번째 선택된 유전자들의 경우, 매우

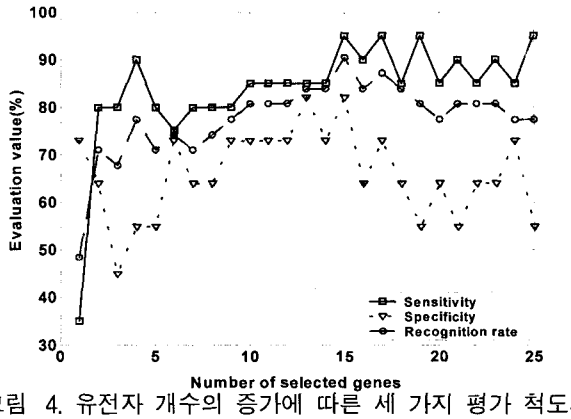


그림 4. 유전자 개수의 증가에 따른 세 가지 평가 척도의 변화

Fig. 4. Trace of three criteria, increasing the number of genes

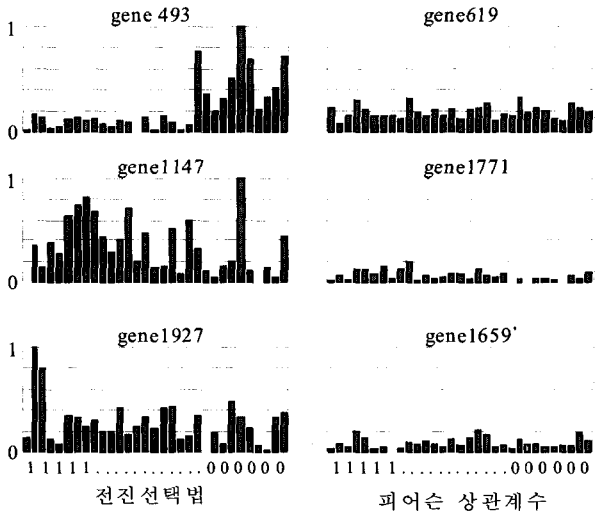


그림 5. 전진 선택법과 피어슨 상관계수에 의해 선택된 상위 3개의 유전자들의 발현 정도

Fig. 5. Expression levels of selected top three genes by forward selection method and Pearson's correlation coefficient

낮은 R^2값을 가지고 있는 것을 볼 때, 15개 정도의 유전자 집합이 암의 분류에 있어 가장 유용한 정보를 가지고 있는 것을 확인할 수 있다. 실제로 선택된 유전자들의 발현정도는 그림 5와 같다. 그림 5를 보면 전진선택법에 의해 선택된 유전자들의 부분상관관계를 쉽게 파악할 수 있다. 가로축은 각 샘플이 되고 1은 정상샘플, 0은 암인 샘플을 나타낸다. 세로축은 유전자들의 발현 정도를 나타내며 2000개의 유전자들을 정규화 시킨 값으로 표시하였다. 유전자들의 발현 정도를 0~1 사이로 정규화시켰을 때 전진선택법의 경우 첫 번째 선택된 유전자(gene493)와 두 번째 선택된 유전자(gene1147)는 서로 상반된 발현 분포를 보인다. 이것은 두 번째 선택

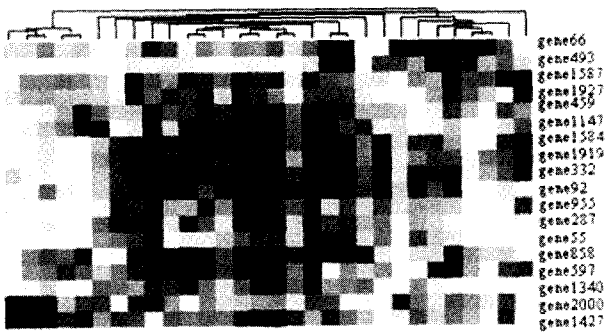


그림 6. 클러스터링 결과(전진선택법)
Fig. 6. Hierarchical clustering result (forward selection method)

표 3. 각 선택 방법에 대한 결과값(%): 대장암
Table 3. Results of three gene selection methods: Colon

	민감도	특이도	인식률
피어슨	75.0	82.0	77.4
스피어맨	100.0	9.0	67.7
전진선택법	**95.0**	**82.0**	**90.3**

과정에서 첫 번째 유전자가 표현할 수 없는 부분을 잘 표현하는 유전자를 우선으로 선택하였기 때문이다. 세 번째 유전자(gene1927) 역시 첫 번째 유전자나 두 번째 유전자가 표현하지 않는 부분을 잘 표현하는 것을 볼 수 있다.

그림 6은 대장암 데이터에서 선택한 18개 유전자들의 발현 정도를 클러스터링한 결과이다. 그림 5의 결과와 마찬가지로 그림 6의 클러스터링 결과도 전진 선택법에 의해 선택된 유전자들의 특성을 보여준다. 특정한 클러스터를 형성하지 않고 흩어진 분포를 보이는 것은 전진선택법에 의해 선택된 유전자들이 서로 다른 분포를 가지고 있기 때문이다. 그렇기 때문에 선택된 유전자들의 집합은 서로 중복되는 정보를 최소화시키며 암의 분류에 있어 보다 효율적인 기능을 보인다.

표 3을 보면, 세 가지 유전자 선택 방법에 의해 선택된 유전자들의 민감도, 특이도, 인식률의 최고치를 알 수 있다. 스피어맨 상관계수의 경우 민감도가 100%로 전진선택법의 경우(95%)보다 높지만, 특이도가 9%로 전진선택법의 82%에 훨씬 미치지 못한다. 즉 암인 샘플은 잘 분류하지만, 암이 아닌 샘플의 경우 거의 분류하지 못한다는 것을 알 수 있다. 피어슨 상관계수의 경우는 반대로 특이도가 민감도보다 정확하여 스피어맨 상관계수와는 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 나타낸다.

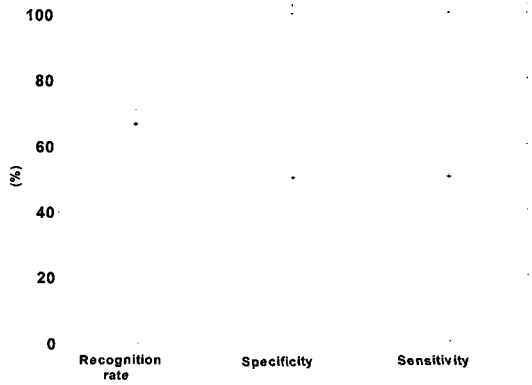


그림 7. 10-fold cross validation 결과
Fig. 7. 10-fold cross validation results

표 4. 선택된 유전자들의 혼동행렬: 대장암
Table 4. Confusion matrix of selected genes: Colon

	피어슨			스피어맨			전진선택법		
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 4는 선택된 유전자들의 혼동행렬로서 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. 그러나 위 실험의 경우, 그림 3의 알고리즘에 입력되는 학습 데이터에 따라 결과가 다를 수 있기 때문에 신뢰도를 높이기 위하여 10-교차검증(fold cross validation)실험을 하였다. 총 31개의 학습 샘플을 10개의 집단으로 나눈 후, 9개의 집단을 학습시키고 나머지 1개의 집단으로 테스트하였다. 10개의 집단으로 나누는 작업을 총 10번 반복하여 10번의 실험을 하여 평균을 구했다. 민감도는 88.0%, 특이도는 73.34%, 인식률은 82.37%의 평균값을 가졌고, 그림 7이 그 결과이다. 민감도가 평균적으로 가장 높은 값을 보이지만 인식률이 편차가 더 적기 때문에 보다 안정적인 수치를 보인다.

2. 림프종 데이터 분석

본 논문에서 제안하는 방법을 림프종 데이터에도 적

표 5. 전진선택법에 의해 선택된 유전자: 림프종
Table 5. Genes selected by forward selection method: Lymphoma

순위	유전자	유전자 설명
1	gene1268	*CD10=CALLA=Neprilysin=enkepalinase; Clone=2008
2	gene544	*DRADA2a=dsRNA adenosine deaminase DRADA2a=RNA editing enzyme;Clone=13269
3	gene824	(Unknown; Clone=137066)
4	gene2313	(Unknown UG Hs.29205 alpha integrin binding protein 63; Clone=135121)
5	gene3125	(Unknown UG Hs.137428 ESTs, Highly similar to (defline not available 3249713) [H.sapiens]; Clone=123429)
6	gene919	(Unknown UG Hs.117333 Homo sapiens mRNA for KIAA1093 protein, partial cds; Clone=133762)
7	gene667	(Unknown UG Hs.187585 ESTs; Clone=82539)
8	gene2406	(Unknown UG Hs.100914 ESTs; Clone=133502)
9	gene233	*Unknown UG Hs.136819 ESTs; Clone=12889
10	gene3207	*Similar to DNA polymerase beta=DNA alkylation repair protein; Clone=13581

표 6. 전진선택법에 의해 선택된 유전자들: 림프종
Table 6. Selected genes by forward selection method: Lymphoma

순위	유전자 번호	R^2값	F-값	Pr>F
1	gene1268	0.8169	89.25	<.0001
2	gene544	0.1045	25.25	<.0001
3	gene824	0.0577	49.82	<.0001
4	gene2313	0.0118	22.28	0.0002
5	gene3125	0.0042	14.07	0.0017
6	gene919	0.0026	17.23	0.0009
7	gene667	0.0013	21.23	0.0004
8	gene2406	0.0006	22.58	0.0004
9	gene233	0.0002	11.89	0.0048
10	gene3207	0.0001	61.90	<.0001

용시켜 보았다. 림프종의 경우 4026개의 유전자중에서 총 10개의 유전자가 그림 3의 알고리즘에 의해 선택되었다. 10개의 유전자만 R^2값이 0보다 컸다. 림프종 데이터에서 선택된 유전자들은 표 5와 같고 이러한 유전자들의 R^2값, F-값, 그리고 F-값의 유의수준은 표 6과 같다. 맨 처음에 선택된 유전자(gene1268)는 0.8169의 R^2값을 가진다. 81%가 넘는 큰 비중을 가지고 선택된 것을 알 수 있다. F-값 또한 89.25로 다른 유전자들보다 훨씬 높은 값을 가지고 있다. 대장암 데이터보다 적은 수의 유전자들이 선택된 것은 이처럼 맨 처음 선택된 유전자가 81%가 넘는 큰 비중을 차지하기 때문이다. 림프종의 경우도 피어슨 상관계수와 스피어맨 상관계수와 민감도, 특이도, 인식률을 가지고 비교하였다.

표 7은 각 유전자 선택 방법에 대한 세 가지 평가척도 중에서 가장 높은 결과들이고, 표 8은 이에 대한 혼동행렬이다. 특이도의 경우만 제외하고는 전진선택법에 의한 유전자 선택이 높은 수치를 보인다. 민감도 90.9%로 피어슨 상관계수(63.6%)나 스피어맨 상관계수(54.5%)보다 훨씬 정확하게 암을 분류하며, 인식률 역

표 7. 각 선택 방법에 대한 결과값(%): 림프종
Table 7. Results of three gene selection methods: Lymphoma

	민감도	특이도	인식률
피어슨	63.6	71.4	68.0
스피어맨	54.5	42.9	48.0
전진선택법	**90.9**	**57.2**	**72.0**

표 8. 선택된 유전자들의 혼동행렬: 림프종
Table 8. Confusion matrix of selected genes: Lymphoma

	피어슨			스피어맨			전진선택법		
P \ A	0	1		P \ A	0	1		P \ A	0	1
0	10	4		0	6	6		0	**8**	1
1	4	7		1	8	5		1	6	**10**

시 72.0%로 피어슨 상관계수(68.0%) 나 스피어맨 상관계수(48.0%)보다 높다. 하지만 특이도의 경우는 피어슨 상관계수가 더 높은 정확성을 보이는 것을 보아, 암이 아닌 샘플에서만은 피어슨 상관계수로 선택한 유전자들이 더 많은 정보를 제공하는 것을 알 수 있다.

V. 결론 및 향후 연구

본 논문에서는 전진선택법을 이용한 유전자선택 방법을 제안하였다. 단순한 일대일 상관관계보다는 선택된 다수 유전자들의 부분 상관관계를 고려한 유전자 선택이 암의 분류에 있어 보다 더 효율적이기 때문에 본 논문에서 제안하는 방법을 통해 이를 검증해 보았다. 대장암 데이터를 사용하여 실험해 본 결과 전진선택법에 의한 특징선택 방법이 상관분석기반의 특징선택 방법보다 암의 분류에 있어 효과적으로 유전자를 선택하는 것을 확인하였다. 또한 전진선택 방법은 중복된 정보를 최소화시키는 유전자들의 조합을 형성하여 암의

분류에 더 많은 정보를 제공한다는 사실도 실험을 통하여 확인하였다. 추가적으로 림프종 데이터에 대하여 제안하는 방법을 적용해 본 결과, 대장암 데이터의 결과와 마찬가지로 제안하는 방법의 우수성을 확인해 볼 수 있었다.

하지만, 선택된 유전자들에 대한 생물학적인 의미는 분석하지 못하였기 때문에, 이에 대한 추가적인 연구가 필요하다. 실제 유전자가 가진 기능과 역할을 안다면 보다 더 심도 있는 연구의 진행이 가능할 것이다. 또한 본 논문에서는 KNN하나의 분류기만을 가지고 실험을 하였기 때문에, 추가적으로 다른 분류기에 대한 실험을 통해 보다 검증된 결과를 얻는 것이 바람직하다.

참 고 문 헌

[1] C. A. Harrington, C. Rosenow, and J. Retief, "Monitoring gene expression using DNA microarrays," Curr. Opin. Microbiol, vol. 3, no. 3, pp. 285-291, 2000.

[2] S. B. Cho and J. W. 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] W. D. Shannon, M. A. Watson, A. Perry, and K. Rich, "Mantel statistics to correlate gene expression levels from microarrays with clinical covariates," Genetic Epidemiology, vol. 23, no. 1, pp 87-96, 2002.

[4] J. Khan, J. S. Wei, M. Ringnér, L. H. Saal, M. Ladanyi, F. Westermann, F. Berthold, M. Schwab, C. R. Antonescu, C. Peterson, and P. S. Meltzer, "Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks," Nature, vol. 7, no. 6, pp. 673-679, June 2001.

[5] L. Li, C. R. Weinberg, T. A. Darden, and L. G. Pedersen, "Gene selection for samlple 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, 2001.

[6] M. P. S. Brown, W. N. Grundy, D. Lin, N. Cristianini, C. Sugnet, M. Ares, Jr., and D. Haussler, "Support vector machine classification of microarray gene expression data," USCS-CRL-99-09, pp. 1-23, June 1999.

[7] P. Tamayo, "Interpreting patterns of gene expression with self-organizing map: Methods and application to hematopoietic differentiation," Proc. of National Academy of Sciences, vol. 96, pp. 2907-2912, 1999.

[8] R. J. Lipshutz, S. P. Fodor, T. R. Gingeras, and D. J. Lockhart, "High density synthetic oligonucleotide arrays," Nature Genetics, vol. 21, pp. 20-24, 1999.

[9] 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.

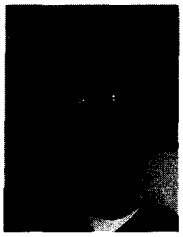
[10] S. Dudoit, J. Fridlyand, and T. P. Speed, "Comparison of discrimination methods for the classification of tumors using gene expression data," Technical Report 576, Department of Statistics, University of California, Berkeley, 2000.

[11] M. Xiong, L. Jin, W. Li, and E. Boerwinkle, "Computational methods for gene expression -based tumor classification," BioTechniques, vol. 29, no. 6, pp. 1264-1270, 2000.

[12] K. E. Lee, N. Sha, E. R. Dougherty, M. Vannucci, and B. K. Mallick, "Gene selection: A bayesian variable selection approach," Bioinformatics, vol. 19, no. 1, pp 90-97, 2003.

[13] J. Rawlings, "Applied regression analysis," Wadsworth Books, Belmont, CA, 1998.

저 자 소 개



유 시 호(학생회원)
1998년 연세대학교 컴퓨터과학과 학사 졸업
2002년~현재 연세대학교 컴퓨터과학과 석사 과정
<주관심분야: 바이오정보기술, 패턴인식>

조 성 배(정회원)
1988년 연세대학교 전산과학과 학사 졸업
1990년 한국과학기술원 전산과학과 석사 졸업
1993년 한국과학기술원 전산과학과 박사 졸업
1993년~1995년 일본 ATR 인간정보통신연구소 객원연구원
1998년 호주 Univ. of New South Wales 초청연구원
1995~현재 연세대학교 컴퓨터과학과 정교수
<주관심분야: 신경망, 패턴인식, 지능정보처리>