# 유전자 발현 데이터를 이용한 암의 유형 분류 기법

# Cancer-Subtype Classification Based on Gene Expression Data

조 지 훈, 이 동 권, 이 민 영, 이 인 범*
(Ji-Hoon Cho, Dongkwon Lee, Min-Young Lee, and In-Beum Lee)

**Abstract :** Recently, the gene expression data, product of high-throughput technology, appeared in earnest and the studies related with it (so-called bioinformatics) occupied an important position in the field of biological and medical research. The microarray is a revolutionary technology which enables us to monitor several thousands of genes simultaneously and thus to gain an insight into the phenomena in the human body (e.g. the mechanism of cancer progression) at the molecular level. To obtain useful information from such gene expression measurements, it is essential to analyze the data with appropriate techniques. However the high-dimensionality of the data can bring about some problems such as curse of dimensionality and singularity problem of matrix computation, and hence makes it difficult to apply conventional data analysis methods. Therefore, the development of method which can effectively treat the data becomes a challenging issue in the field of computational biology. This research focuses on the gene selection and classification for cancer subtype discrimination based on gene expression (microarray) data.

**Keywords :** gene expression, gene selection, classification, multivariate statistical method, machine learning

## I. 서론

인간 및 다양한 생물체를 대상으로 한 게놈 프로젝트의 발전으로 유전자 시퀀스에 관한 정보가 날로 증가하고 있다. 이러한 막대한 양의 데이터에서 가치 있는 정보는 추출하기 위해 SAGE (serial analysis of gene expressio)나 수 천 개의 유전자 발현 패턴을 동시에 탐색하는 마이크로어레이 (microarray)등 다양한 방법이 개발되어 왔다[1]. 유전자 발현 패턴은 유전자의 기능, 조절 및 상호작용을 분석하는 단서를 제공한다. 유전자 발현 패턴에 관한 대부분의 연구는 수 천 개의 유전자 발현 정도를 세포 응답, 표현형, 각종 조건 하에서 기록한 데이터를 분석하는 것이다. 특히 유전자 발현 패턴을 이용해 암의 전이 단계 메커니즘을 분석하는 작업은 활발히 연구되는 분야이다[2-10].

유전자 발현 패턴을 분석하는 작업은 크게 두 단계-저단계 분석(low-level analysis)과 고단계 분석(high-level analysis)로 나뉘어 진다[11]. 저단계 분석은 스캐너에서 얻은 DNA 마이크로어레이 이미지 분석과 마이크로어레이 데이터의 에러를 보정하는 작업을 말한다[12]. 즉, 저단계 분석은 일종의 데이터 전처리 과정이라 할 수 있으며 마이크로어레이 데이터의 품질을 향상시키는 중요한 역할을 한다. 데이터 분석의 정확도와 신뢰도는 저단계 분석이 없다면 보장할 수 없다. 데이터를 적절하게 전처리한 후에는 가치 있는 정보, 예를 들면 특정 암에서만 발현되는 유전자나 병의 진단에 관련된 유전자 또는 병의 유형을 찾기 위해 고단계 분석을 수행한다. 고단계 분석은 비감독 학습(unsupervised learning)과 감독 학습(supervised learning)으로 나눌 수 있다. 비감독(unsupervised)이라는 말은 일반적으로 각 샘플이 어떤 범주에 속하는지 모를 경우 유전자 발현 패턴에 내재된 패턴을 찾는 작업의 경우에 사용된다. 반면 이미 각 샘플이 속하는 범주 (예를 들면 환자의 상태)가 주어져있으면 감독 학습을 채택하여 샘플을 분류하는 시스템을 만든다. 본 논문은 유전자 발현 데이터 (주로 암 데이터)를 감독 학습으로 분류하는 작업에 초점을 둔다. 임상적인 관점에서 보면 특정 질병에서만 발현되는 유전자를 찾거나 새로운 환자를 적절히 진단하고 처방하기 위해서는 어떤 범주에 속하는지 판별하는 작업이 중요하다[2-10,13-15]. 예를 들어 두 가지 유형의 백혈병 (급성 림프구성 백혈병, ALL 과 급성 골수성 백혈병, AML)은 유사한 조직화학적 특징을 갖지만 각 유형에 대한 처방은 완전히 다르다. 각 유형은 유전자 발현 패턴이 분자 수준에서 세포의 상태를 나타낸다는 것을 이용해서 만들어진 분류 시스템으로 구분할 수 있다. 마이크로어레이 데이터를 위한 감독 학습 시스템이 효과적으로 작동하려면 적절한 유전자 선별 방법이 함께해야 한다[5,15-18].

유전자 발현 데이터에는 수천 개의 유전자 발현 정보가 있기 때문에 통상적인 데이터 분석 방법은 차원의 저주(curse of dimensionality), 과적합 (over fitting) 혹은 역행렬의 계산 불가 때문에 직접적으로 적용될 수 없다[13,19-21]. 따라서 데이터 분석을 용이하게 하고 해석을 돕기 위해서는 적절한 개수의 유전자를 선별해야 한다. 또한 유전자 선별은 마이크로어레이를 만드는데 필요한 유전자 개수를 줄여주기 때문에 이후의 연구를 좀 더 경제적이고 빠르게 만든다. 또 특정 유전자와 질병간의 관계를 확신하기 위해 필요한 실험 횟수를 줄일 수 도 있다. 결국, 적절한 유전자 선별 방법과 분류 알고리즘을 동시에 개발하는 것은 감독 학습 방법을 이용한 데이터 분석에 있어서 가장 핵심적인 문제라고 할 수 있다. 유전자 선별은 기계 학습 분야에서 변수 선별과 비슷한 개념으로 볼 수 있으며 이는 적절한 기준 (예를 들면, 최소자승오류 (minimum squared error))에 기초해 가장 많은 정보를 포함하고

* 책임저자(Corresponding Author)
논문접수 : 2004. 9. 20., 채택확정 : 2004. 10. 26.
조지훈, 이민영, 이인범 : 포항공과대학교 화학공학과
(cjhjhj@postech.ac.kr/orange@postech.ac.kr/iblee@postech.ac.kr)
이동권 : LG화학(miruya@lgchem.com)
※ 본 연구는 Brain Korea 21 프로젝트의 지원을 받아 수행되었습니다.

있는 변수 (즉, 유전자)를 고르는 과정이라 간주할 수 있다. 전산 생물학(computational biology)분야의 이전 연구에서, 많은 연구진이 배수 기준 (fold-difference criterion)을 사용해 관련된 유전자를 선별했다[3,7]. 즉, 만약 하나의 범주에 속하는 유전자들의 평균 발현 양이 다른 범주의 유전자들의 평균 발현 양보다 두 배 (혹은 세 배, 네 배) 높으면 유용한 유전자라고 선택하는 방법이다. 그러나 이 방법은 정보를 과도하게 단순화하기 때문에 유전자 발현 데이터에서 모든 정보를 추출 할 수 없어서 상당히 비효율적이다. 또 이상치 (outlier)에 민감하다[22]. 이러한 문제를 극복하고 좀 더 타당한 유전자 선별 기준을 개발하기 위해, 많은 연구진이 위에 언급된 바와 같이 다양한 연구를 수행하였다. 본 논문에서는 피셔의 판별분석법 (Fisher discriminant analysis, FDA), 판별 부분최소 자승법 (discriminant partial least squares, DPLS), 서포트 벡터 머신 (support vector machine, SVM), 커널 판별 분석법 (kernel Fisher discriminant analysis, KFDA)와 같은 분류 시스템과 연결된 유전자 선별 방법을 제시한다.

## II. 본론

### 1. 다변량 통계론적 접근법

패턴 인식 분야에서, 다변량 통계적 방법 (multivariate statistical method, MVS)은 결과를 저차원으로 제공하기 때문에 의사 결정 경계 (decision boundary)를 쉽게 식별할 수 있도록 시각화 해주고, 항상 통계적 중요도를 고려한다는 이점이 있다. 본 단락에서는 다변량 통계 방법 중 FDA와 DPLS를 암의 유형 분류 방법으로 제안한다. 제안된 방법은 통계적 중요도에 기반하여 좋은 분류 성능을 보이고 분류 성능의 손실을 최소로 하면서 분류에 필요한 유전자의 개수를 현격히 줄여준다. 또 분류에 참여하는 유전자 개수의 감소는 질병의 유형을 구별하는데 최적의 접근법이라는 것을 증명했다. 본 논문에서 제안된 방법은 백혈병 (acute leukemia) 데이터 (2 개의 범주, ALL과AML)에 적용되었다[5].

FDA의 기본적인 아이디어는 다변량 샘플 $\mathbf{x}$를 단변량 샘플 $\mathbf{y}$로 정사영(프로젝션) 시켰을 때 최대한으로 $\mathbf{y}$가 분리되도록 만드는 것이다[21,23]. FDA는 각 범주간의 분산은 최대로 하고 동시에 범주내의 분산은 최소로 하는 프로젝션 벡터를 찾는다. 만약 $n$ 차원의 샘플 $\mathbf{x}_1,\ldots,\mathbf{x}_n$이 있고 $n_1$, $n_2$는 각각 범주 1과 범주 2에 속하는 샘플의 개수라고 하자. 각각의 범주 $c$에 대한 분산 행렬 $\mathbf{S}_c$와 범주 내 분산 행렬 $\mathbf{S}_W$는 다음과 같다.

$$\mathbf{S}_c = \sum_{\mathbf{x}\in class\ c} (\mathbf{x}-\bar{\mathbf{x}}_c)(\mathbf{x}-\bar{\mathbf{x}}_c)^T \tag{1}$$

$$\mathbf{S}_w = \sum_{c=1}^{C} \mathbf{S}_c \tag{2}$$

여기서 $\bar{\mathbf{x}}_c$ 는 범주 $c$의 평균 벡터이고 $C$는 범주의 개수이다. 위첨자 T는 행렬의 전치(transpose)를 의미한다. 범주간 분산 행렬 $\mathbf{S}_B$는 다음과 같다.

$$\mathbf{S}_B = \sum_{c=1}^{C} n_c (\bar{\mathbf{x}}_c - \bar{\mathbf{x}})(\bar{\mathbf{x}}_c - \bar{\mathbf{x}})^T \tag{3}$$

여기서 $\bar{\mathbf{x}}$ 는 전체 데이터의 평균 벡터이다. 범주 간 분산 행렬과 범주 내 분산 행렬의 합은 전체 분산 행렬 $\mathbf{S}_T$와 같고, 이는 다음과 같이 정의 된다.

$$\mathbf{S}_T = \sum_{i=1}^{n} (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^T = \mathbf{S}_W + \mathbf{S}_B \tag{4}$$

여기서 $\mathbf{S}_T, \mathbf{S}_W, \mathbf{S}_B \in R^{d\times d}$ $\mathbf{S}_W$의 역행렬이 존재한다는 가정하에 프로젝션 벡터 (loading)는 다음과 같은 최적화 문제를 풀어서 얻을 수 있다.

$$\max_{\mathbf{w}\neq 0} \frac{\mathbf{w}^T\mathbf{S}_B\mathbf{w}}{\mathbf{w}^T\mathbf{S}_w\mathbf{w}} \tag{5}$$

위 식을 최대화하는 벡터 $\mathbf{w}$가 $\mathbf{S}_B\mathbf{w} = \lambda\mathbf{S}_w\mathbf{w}$ 의 일반화된 고유벡터(eigenvector)라는 것은 쉽게 증명할 수 있다[24]. 여기서 람다는 범주간의 분리 정도를 나타내는 $\mathbf{S}_w^{-1}\mathbf{S}_B$ (위첨자 -1은 역행렬을 의미한다)의 고유값 (eigenvalue)이다.

백혈병의 유형을 분류하는 작업은 다음과 같은 과정으로 성취된다. 로딩(loading)벡터 $\mathbf{w}$로 스코어 벡터(score vector) $\mathbf{y} = \mathbf{w}^T\mathbf{x}$를 얻을 수 있다. 그리고 각 샘플 $\mathbf{x}$를 $\mathbf{y}$값에 따라 적당한 범주로 나눈다.

DPLS는 독립변수 $\mathbf{X}$와 의존변수 $\mathbf{Y}$ 사이의 공분산, 즉 상호관계를 최대화하는 차원 감소 방법이라고 할 수 있다. DPLS는 $\mathbf{X}$와 $\mathbf{Y}$ 간의 관계를 부분 최소자승법 (local least-square fits)을 사용해 모델링한다. 일반적인 PLS와는 달리 DPLS 모델에서 의존 변수 $\mathbf{Y}$는 각 샘플이 범주 소속과 관련된 정보를 갖는다. $\mathbf{Y}\in R^{n\times C}$ 행렬의 각 행은 다음과 같은 구조를 갖는다.

$$y_{ic} = \begin{cases} 1, & i\text{번째 샘플이 범주 } c\text{에 속할 경우} \\ 0, & \text{그 이외의 경우} \end{cases}$$

여기서 $\mathbf{y}_c$는 $\mathbf{Y}$의 $c$번째 열이고 $c = 1, 2, \ldots, C$이다.
결과적으로 $\mathbf{Y}$ 행렬은 다음과 같은 이진 변수가 된다.

$$\mathbf{Y} = [\mathbf{y}_1 \quad \mathbf{y}_2 \quad \cdots \quad \mathbf{y}_C] = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \cdots & \vdots \\ 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \cdots & \vdots \\ 0 & 1 & \cdots & 0 \\ 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & \cdots & 1 \\ \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix} \tag{6}$$

$\mathbf{X}$와 $\mathbf{Y}$ 행렬은 평균은 0이고 분산은 1이 되도록 조정되어야 한다. 독립변수 $\mathbf{X}$는 스코어 행렬 $\mathbf{T}\in R^{n\times lv}$ 와 로딩 행렬 $\mathbf{P}\in R^{d\times lv}$, 잔차 행렬 $\mathbf{E}$ ($\mathbf{X}$와 같은 크기의 행렬)로 다음과 같

이 분해된다. 여기서 PLS 모델의 차원을 결정하는 $h$는 내재변수 (latent vaiable)의 개수이다.

$$\mathbf{X} = \mathbf{TP}^T + \mathbf{E} \tag{7}$$

게다가 **Y**는 **X**처럼 스코어 행렬 **U**와 로딩 행렬 **Q**, 잔차 행렬 **F**로 분해된다.

$$\mathbf{Y} = \mathbf{UQ}^T + \mathbf{F} \tag{8}$$

NIPALS 알고리즘 [25]을 이용하여 **X**와 **Y**의 공분산을 최대화하는 회귀모델 행렬 **B**가 결정되고 DPLS 모델은 다음과 같이 주어진다.

$$\mathbf{Y} = \mathbf{TBQ}^T + \mathbf{F}^* \tag{9}$$

여기서 **F**는 회귀모델 행렬 **B**에 의해 오류가 최소가 되는 예측 잔차 행렬이다.

최종 DPLS 모델이 예측에 사용되었을 때 예측된 $\hat{\mathbf{Y}}$ 행렬은 이진수 (0 혹은1)가 아니라 실수 (real number)이다. 이러한 실수를 범주 인덱스로 바꾸는 방법은 [25]의 논문에 제안되어 있다. DPLS 방법의 장점은 독립변수 (즉, 유전자)의 직교성에 있다. DPLS 모델링을 이용해 독립변수를 직교화하는 것은 각 유전자를 독립으로 만들기 때문에 유전자간의 상관관계를 더 이상 고려할 필요가 없게 된다. 여기서 DPLS는 PLS 회귀방법의 변형이기 때문에 이는 패턴 분류에 회귀분석방법을 적용한 것으로 생각할 수 있다.

FDA에서는 분류 효율의 손실 없이 변수의 차원을 줄이고 유용한 정보를 지닌 유전자를 선별하기 위해 윌크의 람다 (Wilks' lambda)와 $F$-test를 이용해 단계적 판별 분석(stepwise discriminant analysis)을 수행하였다. DPLS에서는 VIP (정사영에서의 변수 가중치, variable importance in projection)를 이용하여 **Y**에 대한 모든 변수들의 영향력을 나타내었다. VIP는 DPLS 모델의 가중치 벡터와 모델의 차원에 의해 설명되는 정도로부터 계산할 수 있다. 만약 어떤 변수가 1 보다 큰 VIP 값을 갖는다면 이 변수는 **Y**를 잘 설명한다고 할 수 있다. 유용한 정보를 지닌 유전자를 선별하는 작업을 반복하면서 예측 오차를 최소화하는 최적의 유전자 집합을 찾을 수 있다.

판별 분석에서는 윌크의 람다값이 계산되고 이 값에 따라 변수들이 오름차순으로 정렬된다. 그 다음, 위에서부터 300개의 변수가 선택되고 단계적 판별 분석과정으로 가장 분리 효율이 큰 변수들 만 고르게 된다. 다음 그림은 최종적으로 선별된 유전자 집합 (6개)을 나타낸다.

DPLS 모델에서 VIP 값으로 변수를 선택하는 것은 위의 방법보다 더 단순하다. DPLS 모델을 7번 반복한 후 5개의 유전자가 선택되었고 이 모델은 오분류 된 샘플이 3개였다. 다음 그림은 이 결과를 보여준다.

두 가지 방법은 모두 만족할 만한 분류 결과와 성능 (95% 이상의 정확도)을 갖는다. 이 분류 성능은 모두 7129개의 유전자에서 10개 이하로 추출한 유전자로 성취된 결과이다. 이전 연구들에서는 임의적으로 50개나 그 이상의 유전자를 선택했지만 이것은 샘플의 개수를 고려하면 상당히 많은 숫자이다. 유전자를 더 적게 선별해주는 것은 통계적으로 중요하고 과적합이 적고 복잡도가 낮아지기 때문에 주목할 만한 가치가 있다. 제안된 방법에 의해 선별된 유전자의 대부분은 발현수준이 상당히 높고 ($10^3$ 이상) ALL 과 AML 샘플 사이에 큰 차이를 보인다. 이것은 선별된 유전자들이 노이즈나 이상치로 인한 오염에 큰 영향을 받지 않아서 백혈병 유형의 구분을 위한 결정적인 후보로 결정되기 쉽다는 것을 의미한다.

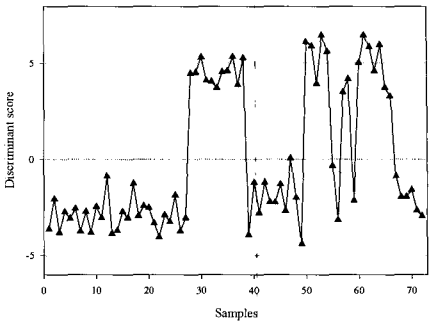



그림 1. 백혈병 데이터의 판별 분석 결과. 음수값은ALL환자를 나타내며 양수값은 AML환자를 나타낸다.

Fig. 1. Classification result of leukemia dataset using FDA (including gene selection), negative score indicates ALL and positive one indicates AML.

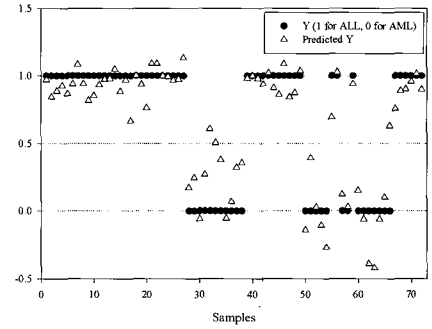



그림 2. DPLS를 이용한 백혈병 환자의 분류.

Fig. 2. Classification result of DPLS (including gene selection). There are 3 mislcassifications.

## 2. 기계학습법적인 접근법

일반적으로 기계학습법에서는 통계적인 중요도 (significance)보다는 판별오류 (misclassification error)를 중요시하여 데이터 분석기를 구현하게 된다. 따라서 비선형성을 갖거나 복잡한 형태의 데이터 분석에 용이하도록 그 구조를 변환시킬 수 있다. 최근에는 커널 머신(kernel machine)이라는 최신 기법이 소개되고 있으며, 성능의 우수성 또한 여러 문헌에서 입증되고 있다[27-28]. 커널 머신은 Cover의 정리 [29]에 입각하고 있는데, 요약하면, 복잡한 패턴을 갖는 데이터는 저차원의 공간 (low-dimensional space)보다 고차원의 공간 (high-dimensional space)에서 선형적으로 패턴이 분리될 가능성이 높다는 내용이다. 커널 머신은 내재적인 비선형 매핑

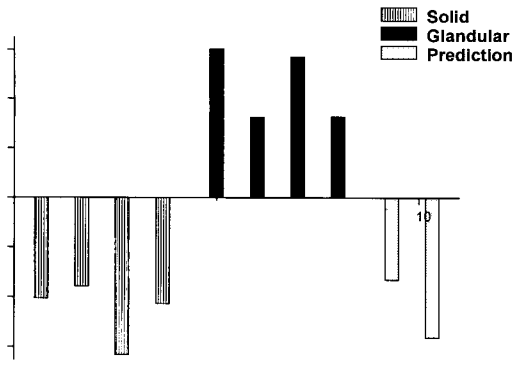



그림 3. 간암 환자의 판별 분석 결과.
Fig. 3. Classification result of HCC patients.

(mapping)을 이용해서 주어진 데이터를 가상의 고차원 공간, 즉 feature space로 이동시킨 후 일련의 데이터 분석을 수행하게 된다. 고차원 상에서의 데이터 분석이 복잡하게 보일 수 있지만, 실제로는 고차원 데이터를 다루지는 않는다. 이것은 커널 트릭 (kernel trick)이라 불리는 기술을 통해서 수행된다 [27-28]. 다음의 웹사이트에서는 커널 머신에 대한 방대한 양의 참고문헌과 교육자료를 제공하고 있다(http://www.kernel-machine.org).

서포트 벡터 머신 (support vector machine, SVM)은 위에서 소개한 커널 머신 중에서 대표적인 하나의 방법으로 판별 분석에 있어서 그 우수성을 자랑하고 있는 데이터 분석 기법의 하나이다. 본 단락에서는 SVM을 이용하여 간암 (hepatocellular carcinoma, HCC)의 유형을 분류한 연구에 대해서 간단히 논해보겠다. 경북대학교에서 수집한 3.1k cDNA 칩 데이터를 이용해서, 간암의 두 가지 유형 (고형암(solid), 선암(pseudo glandular))을 분류하는 SVM판별 분석기를 설계하였다. 8명의 환자 데이터를 학습 데이터로 사용하고 2명의 환자 데이터를 예측해 봄으로써 그 성능을 판별해보도록 하였다. 유전자 선택은 앞서 설명한 윌크의 람다 (Wilk's lambda)값을 사용해서 수행되었다. 판별 분석의 결과는 그림 3과 같다.

데이터 분석 결과, 예측된 환자는 고형암인 것으로 판별되었으며, 확인 결과 실제로 고형암 환자인 것으로 판명되었다. 결과적으로 통계론적 유전자 선택법과 맞물린 SVM판별 분석기가 효과적으로 간암의 유형을 분류할 수 있음을 보여주었다. 또한 선택된 유전자들은 생물학적, 병리학적으로 간암과 관련지을 수 있었다. 일반적으로 간암은 다양한 내, 외부적 자극과 염증, 세포외 간질 (extracellular matrix, ECM) 생성 세포의 활성화 등과 관련된 세포 손상에서 비롯되는 경우가 많은데, 본 연구에서 선택된 유전자 중 상당수가 ECM과 연관된 것이었음을 확인할 수 있었다[30].

앞서 언급한 SVM과는 별도로, 커널 FDA라는 또 다른 커널 머신과 새로이 개발한 유전자 선택법을 이용하여 암의 유형분류에 적용해보았다. 데이터 분석적인 관점에서, 특정 범주의 중심과 샘플간의 거리를 이용하여 유전자의 중요도를 판단할 수 있는 새로운 측정치를 개발하였다. 제안된 새로운 측정치의 장점은 특정 범주 내부에서의 샘플들간의 편차가 적다는 점, 그리고 두 개의 범주를 갖는 데이터뿐만 아니라 여러 개의 범주를 갖는 데이터에도 쉽게 확장이 가능하다는 점을 들 수 있다. 예를 들어, $p \times n$ 크기의 데이터 행렬 $\mathbf{X}$가 있다고 가정한다. 이 경우 $x_{ij}$는 $i$번째 유전자, $j$번째 샘플의 유전자 발현량을 나타내게 된다 ($i = 1, 2,\ldots, p$ 이고 $j = 1, 2,\ldots, n$). 또한 전체 데이터 상에서 $K$개의 범주가 존재하고 $k$ ($k = 1, 2,\ldots, K$)번째 범주는 $n_k$개의 샘플을 갖는다고 가정한다. $C_k$는 $k$번째 범주에 속하는 샘플들의 집합을 나타내는 지수 (index)라고 생각해 보면, $k$번째 범주의 중심 (the centroid of class $k$)은 다음과 같은 식으로 표현된다.

$$\bar{x}_{ik} = 1/n_k \sum_{j \in C_k} x_{ij} \tag{10}$$

위 식에서 $\bar{x}_{ik}$는 $k$번째 범주에서의 $i$번째 유전자의 평균 발현량을 의미한다고 할 수 있다. 평균 발현량을 구한 뒤에는 거리 행렬 $\mathbf{Z}$를 새로이 정의할 수 있는데, 거리 행렬 $\mathbf{Z}$의 각각의 요소는 다음과 같이 구해진다.

$$z_{ij} = \sqrt{(x_{ij} - \bar{x}_{ik})^2}, \; j \in C_k \tag{11}$$

$z_{ij}$로 이루어진 $1 \times n$ 크기의 열 벡터 (row vector) $\mathbf{z}_i$는 범주 1의 중심으로부터의 $n_1$개의 거리, 범주 2의 중심으로부터의 $n_2$개의 거리,..., 범주 $K$의 중심으로부터의 $n_K$개의 거리 등, 도합 $n$개의 샘플의 각각의 범주 중심으로부터의 거리들로 이루어져 있다. 이 열 벡터 $\mathbf{z}_i$를 내부 범주 거리 벡터 (within-class distance vector)라 정의한다. 만일 $i$번째 유전자가 판별 분석에 유용하다면 (즉, 서로 다른 범주 간에는 발현 패턴의 차이가 크고, 같은 범주 안에서는 발현량 편차가 적다면), 위에서 정의한 내부 범주 거리 벡터의 평균 (mean)과 표준 편차 (standard deviation)는 작은 값을 갖게 될 것이다. 주의할 점은 평균과 표준 편차라는 통계치는 각 범주에 속하는 샘플의 개수에 영향을 받을 수 있기 때문에, 범주 개수가 많은 경우에는 내부 범주 거리 벡터의 평균과 표준편차가 왜곡될 우려가 있다는 점이다. 이러한 상황을 방지하기 위하여, 다음과 같은 가중평균 (weighted mean)과 가중 표준 편차 (weighted standard deviation)를 사용함으로써 샘플 개수에 대한 의존도를 제거할 수 있다.

유전자 $i$에 대해서,

$$\text{mean}_w(\mathbf{z}_i) = \sum_{j=1}^{n} \frac{w_j}{W} z_{ij} \tag{12}$$

$$\text{std}_w(\mathbf{z}_i) = \sqrt{\frac{\sum_{j=1}^{n} (z_{ij} - \text{mean}_w(\mathbf{z}_i))^2}{(n - 1/n)\sum_{j=1}^{n} w_j}} \tag{13}$$

여기서, $W = \sum_{j=1}^{n} w_j, w_j = \frac{1}{n_k} (j \in C_k)$

다음과 같은 측정치 $r_i$를 정의함으로써, 유전자 $i$의 판별분석에 있어서의 중요도를 나타낼 수 있다.

$$r_i = \mathrm{mean}_w(\mathbf{z}_i)\cdot \mathrm{std}_w(\mathbf{z}_i) \tag{14}$$

측정치 $r_i$가 작은 값을 갖는다는 것은 유전자 $i$의 발현 패턴이 각각의 범주 중심에서 크게 벗어나지 않으며, 동시에 특정 범주 안에서 작은 편차를 보인다는 것을 의미한다. 따라서 유전자 $i$가 암의 특정한 유형(즉, 범주)에 유의한 발현 패턴을 보인다고 결론 지을 수 있는 것이다. 그러나 한 가지 문제점은, 위의 측정치에 의해서 균등분포 (uniform distribution)를 따르며 발현 패턴의 변화가 거의 없는(판별분석에 있어서 도움이 되지 않는) 유전자가 선택이 될 가능성이 있다는 것이다. 이러한 문제점을 해결하기 위하여, 각 범주 중심 $(\bar{x}_{i1},\bar{x}_{i2},\ldots,\bar{x}_{iK})$의 편차를 포함시켜 다음과 같은 최종적인 측정치를 정의하였다[31].

$$R_i = \frac{\mathrm{mean}_w(\mathbf{z}_i)\cdot \mathrm{std}_w(\mathbf{z}_i)}{\mathrm{std}(\bar{\mathbf{x}}_i)} \tag{15}$$

$\bar{\mathbf{x}}_i = [\bar{x}_{i1} \quad \bar{x}_{i2} \quad \cdots \quad \bar{x}_{iK}]$는 $1 \times K$ 크기를 갖는 열 벡터로서 유전자 $i$에 대한 각각의 범주 중심값으로 이루어져 있다. 유전자 선택은 위와 같은 새로운 측정치를 이용해서 수행되었으며, 직접적인 판별분석인 커널 FDA (KFDA)를 이용해서 수행되었다 [28]. 일반적인 피셔의 판별분석 (FDA)가 많은 수의 변수(즉, 유전자)를 갖는 데이터에 사용될 수 없는데에 반해서 KFDA는 변수의 개수에 영향을 받지 않고, 커널 머신이 갖는 장점, 즉 고차원 데이터 처리를 요구하지 않는다는 장점이 있으므로 유전자 발현 데이터의 분석에 유리하다 [27-29]. KFDA의 자세한 알고리즘은 다음의 참고문헌에 잘 나타나있으므로 본 논문에서는 생략하기로 하겠다[28].

위의 유전자 선택법과 판별 분석법을 백혈병 데이터[5]와 small round blue cell tumor (SRBCT) 데이터[32]에 적용하여 암의 유형 분류 성능을 살펴보았다. 데이터 분석은 다음과 같은 절차로 이루어졌다. 전체 데이터를 이용해서 모든 유전자의 $R_i$값을 측정한 뒤, 상위 20% ($R_i$값이 큰)의 유전자를 제외하고 유전자 군 (gene subset)을 만들고 KFDA판별 분석기를 이용하여 판별 분석을 수행, 오류값 (misclassification error)을 계산하였다. 더 이상 제외할 유전자가 없을 때까지 이러한 단계를 반복한 뒤, 최소 오류값을 나타내는 유전자 군을 선택하게 된다. 두 개의 데이터에 대한 판별 분석 오류값과 최소 오류값에서의 유전자 군의 히트맵 (heat map)은 각각 그림 4, 5와 같다.

백혈병 데이터에 대한 많은 연구결과[5,33-34]와 비교해 볼 때 본 연구에서 선택된 유전자들이 대부분 기존 연구결과와 합치함을 확인할 수 있었다. 또한 SRBCT 데이터로부터 얻은 결과도 기존의 연구들과 많은 부분 일치하는 결과를 보였다 [15,32]. 지금까지 언급한 내용들은 소위 필터(filter)라고 불리우는 방법으로서 각각의 유전자들의 유의성 순위를 매기고, 그 순위에 따라서 가장 타당한 유전자 군을 선택하는 방법이다. 본 논문의 나머지 부분에서는 래퍼(wrapper)라고 불리우는 유전자 선택-판별 분석법을 이용한 암의 유형 분류법에 대해서 논하고자 한다. 간단히 래퍼는 판별 분석기 그 자체

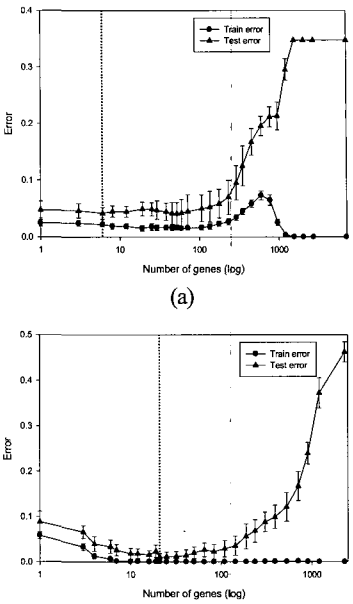



그림 4. 판별 분석 오류값 그래프 (a) 백혈병 데이터, 최소 오류값에서 6개의 유전자 선택 (b) SRBCT 데이터, 최소 오류값에서 21개의 유전자 선택.

Fig. 4. Cross-validation result (a) leukemia data, at the minimum error point, 6 genes are selected (b) SRBCT data, 21 genes are selected.

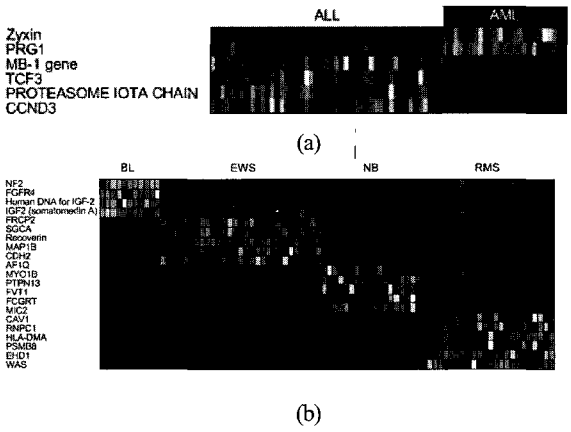



그림 5. 선택된 유전자 군의 히트맵(heat map). 밝을수록발현량이 높음을 의미하고, 어두울수록 발현량이 적음을 나타낸다.

Fig. 5. Heatmap of selected genes, black color indicates high-expression value and white does low-expression value.

를 의미하는 것이다. 그러나 래퍼는 판별 분석과 동시에 유전자 선택을 수행하는 것으로써, 일반적으로 필터보다 좋은 성능을 보인다고 알려져 있다.

본 단락에서는 유전자 발현량 데이터 분석을 위한새로운 래퍼의 개발에 대해서 논해보도록 하겠다. 앞서 언급했던 KFDA를 기반으로 하여, 새로운 유전자 선택 기준 (criterion)을 제안함으로써 암의 유형 분류에 유용하게 쓰일 수 있는

래퍼를 개발하고, 이를 백혈병 데이터[5]와 유방암 데이터[35]에 적용해 봄으로써 그 성능을 판별하였다. 새로운 유전자 선택 기준은 최근에 Rakotomamonjy [36]가 커널 머신에 대해서 제안한 미분법을 응용해서 개발하였다. 또한 직접적으로 판별 오류에 대한 영향을 수식화하기 위하여 다음과 같은 최소자승오류 기반 KFDA수식 [37]을 이용하였다. 두 개의 범주를 갖는 경우 (binary classification)에 한하여 최소자승오류 기반 KFDA의 유도를 간략하게 살펴보도록 하겠다.

$\mathbf{X}=\{\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_{n_1},\mathbf{x}_{n_1+1},\mathbf{x}_{n_1+2},\ldots,\mathbf{x}_{n_1+n_2}\}$ 를 주어진 유전자 발현 데이터라고 가정하자. 이때 $\mathbf{x}_j\in R^p$ 는 j번째 샘플의 발현량을 나타내는 벡터이고 $n_1$ 과 $n_2$ ($n_1+n_2=n$) 는 각각 범주 1과 2에 속하는 샘플의 개수를 의미한다. KFDA 판별 분석기는 다음과 같이 표현할 수 있다.

$$f(\mathbf{x})=\mathbf{K}^T\boldsymbol{\alpha}+b \tag{16}$$

여기서 $\mathbf{K}_{jk}=k(\mathbf{x}_j,\mathbf{x}_k)=\Phi(\mathbf{x}_j)\cdot\Phi(\mathbf{x}_k)$, $\boldsymbol{\alpha}$와 $b$ 는 상수값을 나타낸다. $\Phi(\mathbf{x})$ 는 커널 머신에서 쓰이는 내재적인 비선형 매핑이다. Xu 등은 최소자승법적인 측면에서 KFDA판별 분석기를 다음과 같은 식으로 표현할 수 있다는 것을 증명한 바 있다[37].

$$\begin{bmatrix}\mathbf{K}\mathbf{K}^T+\mu\mathbf{I} & \mathbf{K}\mathbf{U}\\ (\mathbf{K}\mathbf{U})^T & n\end{bmatrix}\begin{bmatrix}\boldsymbol{\alpha}\\ b\end{bmatrix}=\begin{bmatrix}\mathbf{K}\mathbf{y}\\ \mathbf{U}^T\mathbf{y}\end{bmatrix} \tag{17}$$

위 식에서 $j$ 와 $k$ 는 1,..., $n$ 의 값을 가지며, $\mathbf{U}$는 $n$개의 1이라는 값으로 이루어진 행 벡터 (column vector)이다. 또한 $\mu$ 는 정규화 상수값을 의미하고 $\mathbf{y}$는 사용자 정의된 출력값이다. 최소자승법적인 측면에서의 KFDA 판별 분석기의 해는 사용자 정의된 출력값에 영향을 받는다고 Xu 등은 밝히고 있다[37]. 위의 식과 출력값을 이용해서 상수값 $\boldsymbol{\alpha}$와 $b$를 다음과 같이 구할 수 있다.

$$\boldsymbol{\alpha}=\left(\mathbf{K}\mathbf{K}^T+\mu\mathbf{I}-n^{-1}\mathbf{K}\mathbf{U}\mathbf{U}^T\mathbf{K}^T\right)^{-1}\left(\mathbf{K}\mathbf{Y}-n^{-1}\mathbf{K}\mathbf{U}\mathbf{U}^T\mathbf{y}\right) \tag{18}$$

$$b=n^{-1}\left(\mathbf{U}^T\mathbf{y}-(\mathbf{K}\mathbf{U})^T\boldsymbol{\alpha}\right) \tag{19}$$

최종적으로 실제 KFDA 판별 분석기의 출력값은 다음과 같이 표현된다.

$$\hat{\mathbf{Y}}=\mathbf{K}^T\boldsymbol{\alpha}+\mathbf{U}b \tag{20}$$

위의 출력값은 앞서 밝혔듯이 최소자승법 측면에서 유도된 식이므로 실제 출력값과 사용자 정의 출력값의 차이, 즉 다음과 같은 오류 $E$를 최소화하게 된다.

$$E=\frac{1}{2}\|\mathbf{y}-\hat{\mathbf{y}}\|^2=\frac{1}{2}\left(\mathbf{y}-\mathbf{K}^T\boldsymbol{\alpha}-\mathbf{U}b\right)^T\left(\mathbf{y}-\mathbf{K}^T\boldsymbol{\alpha}-\mathbf{U}b\right) \tag{21}$$

본 연구에서는 Xu 등이 제안한 사용자 정의 출력값을 그대로 사용하였는데, 다음과 같다.

$$y_j=\begin{cases}n/n_1, & j\text{번째 샘플이 클래스 1에 속할 경우}\\ -n/n_2, & j\text{번째 샘플이 클래스 2에 속할 경우}\end{cases}$$

위의 출력값을 대입하면 상수값 $\boldsymbol{\alpha}$와 $b$는 다음과 같이 계산 가능한 식으로 표현된다.

$$b=n^{-1}\left(\mathbf{U}^T\mathbf{Y}-(\mathbf{K}\mathbf{U})^T\boldsymbol{\alpha}\right)=-n^{-1}\left(n_1\mathbf{M}_1+n_2\mathbf{M}_2\right)^T\boldsymbol{\alpha} \tag{22}$$

$$\boldsymbol{\alpha}=n\left(1-\frac{n_1n_2}{n^2}\gamma\right)\mathbf{N}^{-1}(\mathbf{M}_1-\mathbf{M}_2) \tag{23}$$

여기서의 $\mathbf{N}$과 $\mathbf{M}_i$ 는 앞서 FDA에서 밝힌 바와 같은 의미의 식으로서 계산된다. 자세한 계산법은 참고문헌 [28]에 상세히 표현되어 있다.

최근 Rakotomamonjy는 편미분법에 기반, SVM에 적용 가능한 새로운 변수 선택법을 제안하였다[36]. 이 방법은 편미분을 사용하기 때문에, 기존의 방법 [17]과 달리 그람 행렬 (Gram matrix) $\mathbf{K}$를 한번만 계산해도 된다는 이점이 있다. 이러한 복잡도에서의 이점은 유전자 발현량 데이터와 같은 고차원 데이터의 분석에 상당한 시간, 비용 절감 효과를 가져올 수 있다. Rakotomamonjy의 방법은 다음과 같은 개념에서 출발한다. 각각의 변수, 즉 유전자에 대해서 가상의 비례 인자 $\mathbf{v}=[v_1\,v_2\cdots v_n]$를 도입하고, 어떤 기준값 (예컨대 판별 분석 오류)을 비례 인자에 대해서 편미분한 값을 계산해서 가장 큰 값을 보이는 변수가 기준값에 가장 큰 영향을 미쳤다고 간주하는 것이다. 실제로 커널 머신에서 위와 같은 가상 비례 인자를 도입하면 각각의 변수에 대해서 다음과 같이 미분이 가능하다.

$$\begin{aligned}\frac{\partial k}{\partial v_i}&=-\frac{2}{\sigma}v_i(x_{ij}-x_{ik})^2k(\mathbf{v}\otimes\mathbf{x}_j,\mathbf{v}\otimes\mathbf{x}_k)\\ &=-\frac{2}{\sigma}(x_{ij}-x_{ik})^2k(\mathbf{x}_j,\mathbf{x}_k)\end{aligned} \tag{24}$$

$$k(\mathbf{x}_j,\mathbf{x}_k)=\exp\left(-\frac{\|\mathbf{x}_j-\mathbf{x}_k\|^2}{\sigma}\right) \tag{25}$$

$$k(\mathbf{v}\otimes\mathbf{x}_j,\mathbf{v}\otimes\mathbf{x}_k)=\exp\left(-\frac{\|\mathbf{v}\otimes\mathbf{x}_j-\mathbf{v}\otimes\mathbf{x}_k\|^2}{\sigma}\right) \tag{26}$$

위 식들에서 미분대상인 $i$번째 변수의 가상 비례 인자 $v_i$ 만 1의 값을 갖고, 나머지 인자들은 모두 0을 갖는다고 간주한다. 또한 $\otimes$는 벡터의 요소끼리의 곱을 의미하는 연산자이다. 위와 같은 편미분법을 응용해서 최소자승오류 기반 KFDA 판별 분석기에 대하여, 판별 오류에 가장 큰 영향을 끼치는 (즉, 가장 유용한) 변수 (유전자)를 선택할 수 있는 측정치를 다음과 같이 정의할 수 있다.

$$R_i=\left|\frac{\partial E}{\partial v_i}\right|=\frac{1}{2}\left|\frac{\partial\left(\mathbf{y}-\mathbf{K}^T\boldsymbol{\alpha}-\mathbf{U}b\right)^T\left(\mathbf{y}-\mathbf{K}^T\boldsymbol{\alpha}-\mathbf{U}b\right)}{\partial v_i}\right| \tag{27}$$

만약 유전자 $i$가 판별분석에 유용하게 사용될만한 것이라면, 그것은 판별 오류에 심각하게 영향을 끼치게 될 것이고 자연히 $R_i$라는 측정치는 큰 값을 갖게 될 것이다. 위 수식의 계산이 복잡해보이지만, 실제로는 그람 행렬 $\mathbf{K}$와 관련된 항들만이 실제 미분에 영향을 받게 되고 나머지 항들은 상수값으로 간주된다. $R_i$에 관한 식을 전개해 보면 다음과 같다.

$$R_i = \frac{1}{2}\left|\frac{\partial\left(\boldsymbol{\alpha}^T\mathbf{K}\mathbf{K}^T\boldsymbol{\alpha}\right)}{\partial \nu_i} - 2\frac{\partial\left(\boldsymbol{\alpha}^T\mathbf{K}\mathbf{Y}\right)}{\partial \nu_i} - 2\frac{\partial\left(\boldsymbol{\alpha}^T\mathbf{K}\mathbf{U}\mathbf{b}\right)}{\partial \nu_i}\right| \tag{28}$$

$\mathbf{K}\mathbf{K}^T$ 의 미분값은 다음과 같이 구할 수 있다.

$$\frac{\partial\left(\mathbf{K}\mathbf{K}^T{}_{jk}\right)}{\partial \nu_i} = -\sum_{m=1}^{n}\frac{2}{\sigma}\left\{\left(x_{ij}-x_{im}\right)^2 + \left(x_{ik}-x_{im}\right)^2\right\}k\left(\mathbf{x}_j,\mathbf{x}_m\right)k\left(\mathbf{x}_k,\mathbf{x}_m\right)\Big|_{\nu_i=1} \tag{29}$$

또한 $\mathbf{K}$의 미분값은 앞서 (24) ~ (26)에서 보인 대로 쉽게 구할 수 있다. 실제로 새로운 유전자 선택 측정치 $R_i$ 는 다음 식에 의해서 계산된다

$$R_i = \frac{1}{2}\left|\begin{array}{l}\boldsymbol{\alpha}^T\left\{\left(\mathbf{D}_i\otimes\mathbf{K}\right)\mathbf{K}^T + \mathbf{K}\left(\mathbf{D}_i\otimes\mathbf{K}\right)^T\right\}\boldsymbol{\alpha} \\ -2\boldsymbol{\alpha}^T\left(\mathbf{D}_i\otimes\mathbf{K}\right)\mathbf{Y} + 2\boldsymbol{\alpha}^T\left(\mathbf{D}_i\otimes\mathbf{K}\right)\mathbf{U}\mathbf{b}\end{array}\right| \tag{30}$$

$\mathbf{D}_i$ 는 샘플 거리 행렬이라고 정의한 것으로서, 수학적으로는 다음과 같이 표현된다.

$$\mathbf{D}_i = \begin{bmatrix}\left(x_{i1}-x_{i1}\right)^2 & \left(x_{i1}-x_{i2}\right)^2 & \cdots & \left(x_{i1}-x_{in}\right)^2 \\ \vdots & \vdots & \ddots & \vdots \\ \left(x_{in}-x_{i1}\right)^2 & \left(x_{in}-x_{i1}\right)^2 & \cdots & \left(x_{in}-x_{in}\right)^2\end{bmatrix} \in R^{n\times n} \tag{31}$$

본 연구에서 제안한 방법은 계산 속도나 부담을 현격히 감소시켰다는 장점을 갖고 있다. 기존의 방법이 반복적으로 그람 행렬 $\mathbf{K}$를 구해야하는데 비하여 제안한 방법은 단 한번의 계산만으로 모든 절차를 마칠 수 있다. 또한 Rakotomamonjy 의 방법은 $\mathbf{K}$의 미분을 위해서 복잡한 계산을 요구하는 데에 비해서 본 방법은 간단히 계산될 수 있는 $\mathbf{D}_i$만을 필요로 한다. 지금까지 설명한 방법을 토대로 하여 두 종류의 암에 대해서 유형을 분류한 결과는 표 1과 같다.

결과에서 볼 수 있듯이 기존의 방법들과 비교했을 때 손색없는 판별 분석 결과를 보임을 알 수 있다. 표 1(a)에서는 제안한 방법이 기존의 방법보다 더 많은 오류를 범하고 있는 것으로 나타나있다. 그러나 이것은 데이터 자체에 오류를 포함하고 있는 가능성이 크다고 알려져 있다 [32]. 따라서 본 연구에서 2개의 오류를 범한 것이 더 정확한 분석 결과라고 결론 지을 수 있겠다. 본 논문에서 수행된 모든 데이터 분석은 2.0 GHz Pentium PC (1 GB RAM) 환경에서 이루어졌으며, Windows XP 플랫폼을 기반으로 Matlab®software를 사용하였다.

표 1. 판별 분석 오류와 선택된 유전자 군의 크기 (a) 백혈병 데이터 (b) 유방암 데이터.

Table 1. Number of misclassifications and the size of selected gene subset (a) leukemia (b) breast cancer.

| Methods | Misclassifications | Size of subset |
|---|---|---|
| Guyon et al. (2002) | 0 | 4 |
| Cho et al. (2003) | 9 | 1 |
| Proposed | 2 | 17 |

(a)

| Methods | Misclassifications | Size of subset |
|---|---|---|
| Lee et al. (2002) | 0 | 27, 17, 10 |
| Cho et al. (2003) | 5 | 3 |
| Proposed | 0 | 21 |

(b)

## III. 결론

분자 생물학의 급속한 발전과 인간 게놈 프로젝트의 완료, 그리고 고효율 어레이 기술 (high-throughput array technology)의 도래는 분자 수준에서 생명현상을 바라보게 해 주었고, 엄청난 양의 데이터를 생산하면서 생물학적인 지식을 확장할 수 있는 기회를 부여하였다. 방대한 데이터에 담겨있는 생명 현상을 이해하기 위해서는 유용한 정보를 추출해 줄 수 있는 데이터 마이닝 기술이 필요하게 되었고, 이러한 필요성은 본 논문에서 다룬 바와 같이 생물정보학(bioinformatics), 전산 생물학(computational biology)적인 데이터 분석 기술의 개발에 의해서 충족되고 있다.

데이터 분석의 여러 측면 중에서도, 본 논문에서는 특히 암 환자의 유전자 발현 데이터로부터 암의 유형분류에 유용한 유전자의 선택법과 판별 분류 시스템 구축에 관한 내용에 초점을 맞추었다. 첫째로, 선형적인 다변량 통계적 기법을 사용한 데이터 분석 체계의 구축에 대해서 살펴보았다. 다변량 통계적 기법론적인 체계에서는 유전자 선택과 관련된 의사결정 (예를 들어 최적 유전자 군의 결정)을 통계적 중요도에 의해서 수행함으로써 기존 방법들이 간과해 온 통계적 개념을 강화하였다. 또한 통계론 적인 유전자 선택법과 SVM을 접목하여 성공적으로 간암 환자의 유형 분류 시스템을 구축하였다. 반면 기계학습법 적인 측면에서 기존 통계적 방법이 갖고 있는 단점을 보완할 수 있는 새로운 측정치를 개발하고, 최신 기술이라고 할 수 있는 커널 머신을 응용함으로써 만족할만한 성능을 보이는 암의 유형 분류 시스템을 제안하였다. 끝으로, 다변수 적인 측면에서 변수를 선택할 수 있는 래퍼를 커널 머신을 이용해서 구현함으로써, 소위 필터가 갖고 있는 단점을 극복하는 동시에 만족할만한 성능을 보이는 암의 유형 판별 분석기를 제안하였다.

결론적으로 본 연구는 암 환자의 유전자 발현 데이터로부터 유용한 정보를 추출해내고, 암과 관련된 생물학적 기작을 데이터 분석적인 측면에서 해석할 수 있는 방법론에 대해서 논하였다. 본 논문에서 제안한 방법들은 수학적이고 논리적인 방법론을 제시함으로써 치료용 약물의 목표 발견 (therapeutic target discovery)이나 질병의 메커니즘을 밝히는 일

등에 기여할 수 있을 것이다. 실제로 유전자를 선택하는 방법이나 판별 분석을 수행하는 방법에 의해서 기준, 표준이 되는 방법은 없다. 따라서 연구의 목적이 특정 질병에 대해서 전반적인 메커니즘이나 임상적 특성을 알아보는 것이라면, 가장 보편적인 방법이 사용되어야 할 것이다. 유전자 선택이나 판별 분석은 선험적인 생물학적 지식이 배제된 상태에서 이루어지는 경우가 많으므로, 가장 간단한 방법에서부터 복잡한 방법에 이르기까지 가능한 모든 방법에 의해서 데이터 분석을 시행한 후 지속적으로 선택되는 유전자와 가장 타당한 결론을 기반으로 연구하는 것이 최선의 안전한 방법이라고 할 수 있겠다.

## 참고문헌


[1] J. Quackenbush, "Computational genetics: computational analysis of microarray data," *Nature. Rev. Geneitics*, vol. 2, pp. 418-427, 2001.

[2] A. A. Alizadeh, M. B. Eisen, R. E. Davis, C. Ma, S. Lossos, A. Rosenwald, J. C. Boldrick, H. Sabet, T. Tran, X. Yu, J. I. Powell, L. Yang, G. E. Marti, T. Moore, J. Hudson Jr., L. Lu, D. B. Lewis, R. Tibshirani, G. Sherlock, W. C. Chan, T. C. Greiner, D. D. Weisenburger, J. O. Armitage, R. Warnke, R. Levy, W. Wilson, M. R. Grever, J. C. Byrd, D. Botstein, P. O. Brown, and L. M. Staudt, "Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling," *Nature*, vol. 403, pp. 503-511, 2000.

[3] U. Alon, N. Barkai, D. A. Notterman, K. Gish, Y. Barra, D. Mack and A. J. Levine, "Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays," *Proceedings of National Academy of Science USA*, vol. 96, pp. 6745-6750, 1999.

[4] S. A. Armstrong, J. E. Staunton, L. B. Silverman, R. Rieters, M. L. den Boer, M. D. Minden, S. E. Sallan, E. S. Lander, T. R. Golub and S. J. Korsmeyer, "MLL translocations specify a distinct gene expression profile that distinguishes a unique leukemia," *Nature Genetics*, vol. 30, pp. 41-47, 2002.

[5] T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield and E. S. Lander, "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring," *Science*, vol. 286, pp. 531-537, 1999.

[6] X. Chen, S. T. Cheung, S. So, S. T. Fan, C. Barry, J. Higgins, K.-M. Lai, S. Dudoit, I. O. L. Ng, M. van de Rijn, D. Botstein and P. O. Brown, "Gene expression patterns in human liver cancers," *Molecular Biology of the Cell*, vol. 13, pp. 1929-1939, 2002.

[7] D. A. Notterman, U. Alon, A. J. Sierk and A. J. Levine, "Transcriptional gene expression profiles of colorectal adenoma, adenocarcinoma and normal tissue examined by oligonucleotide array," *Cancer Research*, vol. 61, pp. 3124-3130, 2001.

[8] M. A. Shipp, K. N. Ross, P. Tamayo, A. P. Weng, J. L. Kutok, R. C. T. Aguiar, M. Gaasenbeek, M. Angelo, M. Reich, G. S. Pinkus, T. S. Ray, M. A. Koval, K. W. Last, A. Norton, A. Lister, J. Mesirov, D. S. Neuberg, E. S. Lander, J. C. Aster and T. R. Golub, "Diffuse large B-cell lymphoma outcome prediction by gene expression profiling and supervised machine learning," *Nature Medicine*, vol. 8, pp. 68-74, 2002.

[9] D. Singh, P. G. Febbo, K. Ross, D. G. Jackson, J. Manola, C. Ladd, P. Tamayo, A. A. Renshaw, A. V. D'Amico, J. P. Richie, E. S. Lander, M. Loda, P. W. Kantoff, T. R. Golub and W. R. Sellers, "Gene expression correlates of clinical prostate cancer behavior," *Cancer Cell*, vol. 1, pp. 203-209, 2002.

[10] L. van't Veer, H. Dai, M. J. van de Vijver, Y. D. He, A. A. M. Hart, M. Mao, H. L. Peterse, K. van der Kooy, M. J. Marton, A. T. Witteveen, G. J. Schreiber, R. M. Kerkhoven, C. Roberts, P. S. Linsley, R. Bernards and S. H. Friend, "Gene expression profiling predicts clinical outcome of breast cancer," *Nature*, vol. 415, pp. 530-536, 2002.

[11] B. M. Bolstad, R. A. Irizarry, M. Åstrand and T. P. Speed, "A comparison of normalization methods for high density oligonucleotide array data based on variance and bias," *Bioinformatics*, vol. 19, pp. 185-193, 2003.

[12] Y. H. Yang, S. Dudoit, P. Luu, D. M. Lin, V. Peng, J. Ngai and T. P. Speed, "Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation," *Nucleic Acids Research*, vol. 30, pp. e15, 2002.

[13] G. Sherlock, "Analysis of large-scale gene expression data," *Current Opinion in Immunology*, vol. 12, pp. 201-205, 2000.

[14] R. Tibshirani, T. Hastie, B. Narasimhan and G. Chu, "Diagnosis of multiple cancer types by shrunken centroids of gene expression," *Proceedings of National Academy of Science USA*, vol. 99, pp. 6567-6572, 2002.

[15] V. G. Tusher, R. Tibshirani and G. Chu, "Significance analysis of microarrays applied to the ionizing radiation response," *Proceedings of National Academy of Science USA*, vol. 98, pp. 5116-5121, 2001.

[16] Y. Lu and J. Han, "Cancer classification using gene expression data," *Information Systems*, vol. 28, pp. 243-268, 2003.

[17] I. Guyon, J. Weston, S. Barnhill and V. Vapnik, "Gene Selection for Cancer Classification using Support Vector Machines," *Machine Learning*, vol. 46, pp. 389-422, 2002.

[18] Y. Lee and C. Lee, "Classification of multiple cancer types by multicategory support vector machines using gene expression data," *Bioinformatics*, vol. 19, pp. 1132-1139, 2003.

[19] M. Defernez and E. K. Kemsley, "The use and misuse of chemometrics for treating classification problems," *Trends in Analytical Chemistry*, vol. 16, pp. 216-221, 1997.

[20] A. Brazma and J. Vilo, "Gene expression data analysis," *FEBS Letters*, vol. 480, pp. 17-24, 2000.

[21] S. Sharma, *Applied Multivarate Techniques*, John Wiley and Sons, New York, 1996.

[22] S. Dudoit, Y. H. Yang, T. P. Speed and M. J. Callow, "Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments," *Statistica Sinica*, vol. 12, pp. 111-139, 2002.

[23] R. A. Johnson and D. W. Wichern, *Applied Multivariate Statistical Analysis, Third Edition*, Prentice Hall, 1992.

[24] G. H. Golub and C. F. van Loan, *Matrix Computations*, The Johns Hopkins University Press, 1983.

[25] L. H. Chiang, E. Russell and R. D. Braatz, "Fault diagnosis in chemical processes using Fisher discriminant analysis, discriminant partial least squares, and principal component

analysis," *Chemometrics and Intelligent Laboratory Systems*, vol. 50, pp. 243-252, 2000.

[26] J.-H. Cho, D. Lee, J. H. Park, K. Kim and I.-B. Lee, "Optimal approach for classification of acute leukemia subtypes based on gene expression data," *Biotechnology Progress*, vol. 18, pp. 847-854, 2002.

[27] B. Schölkopf, A. Smola and K.-R. Müller, "Nonlinear component analysis as a kernel eigenvalue problem," *Neural Computation*, vol. 10, pp. 1299-1319, 1998.

[28] S. Mika, G. Rätsch, J. Weston, B. Schölkopf and K.-R. Müller, "Fisher discriminant analysis with kernels," *Proc. IEEE Neural Networks for Signal Processing Workshop*, pp. 41-48, 1999.

[29] S. Haykin, *Neural Networks : a comprehensive foundation, Second edition*, Prentice Hall, 1999.

[30] D. Lee, S. W. Choi, M. Kim, J. H. Park, M. Kim, J. Kim and I.-B. Lee, "Discovery of differentially expressed genes related to histological subtype of hepatocellular carcinoma," Biotechnology Progress, vol. 19, pp. 1011-1015, 2003.

[31] J. Khan, J. S. Wei, M. Ringner, 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 Medicine*, vol. 7, pp. 673-679, 2001.

[32] J. Fridlyand, S. Dudoit and T. P. Speed, "Comparison of discrimination methods for the classification of tumors using gene expression data," *Journal of the American Statistical Association*, vol. 97, pp. 77-87, 2002.

[33] A. Ben-Dor, L. Bruhn, N. Friedman, I. Nachman, M. Schummer and Z. Yakhini, "Tissue classification with gene expression profiles," *Journal of Computational Biology*, vol. 7, pp. 559-583, 2000.

[34] I. Hedenfalk, D. Duggan, Y. Chen, M. Radmacher, M. Bittner, R. Simon, P. Meltzer, B. Gusterson, M. Esteller, O.-P. Kallioniemi, B. Wilfond, A. Borg. and J. Trent, "Gene-expression profiles in hereditary breast cancer," *New England Journal of Medicine*, vol. 344, pp. 539-548, 2001.

[35] A. Rakotomamonjy, "Variable selection using SVM-based criteria," *Journal of Machine Learning Research*, vol. 3, pp. 1357-1370, 2003.

[36] J. Xu, X. Zhang and Y. Li, "Kernel MSE algorithm: A unified framework for KFD, LS-SVM and KRR," *Proceeding of International Joint Conference on Neural Networks 2001*, pp. 1486-1491, 2001.

[37] R. O. Duda, P. E. Hart and D. G. Stork, *Pattern Classification, Second Edition*, John Wiley & Sons, 2001.

[38] J.-H. Cho, D. Lee, J. H. Park and I.-B. Lee, "New gene selection method for classification of cancer subtypes considering within-class variation," *FEBS Letters*, vol. 551, pp. 3-7, 2003.

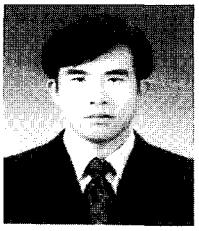
[39] J.-H. Cho, D. Lee, J. H. Park and I.-B. Lee, "Gene selection and classification from microarray data using kernel machine," *FEBS Letters*, vol. 571, pp. 93-98, 2004.

**조 지 훈**

1999년 연세대학교 화학공학과 졸업. 2004년 포항공과대학교 화학공학과 (공학박사). 관심분야는 데이터 마이닝, 생물정보학.

**이 민 영**

2003년 포항공과대학교 화학공학과 졸업. 2003년~현재 포항공과대학교 화학공학과 석사과정 재학중.



**이 동 권**

1995년 영남대학교 화학공학과 졸업. 1997년~2000년 공정산업의 지능자동화 연구센터 연구원. 2004년 포항공과대학교 화학공학과 (공학박사). 2004년~현재 LG화학.

**이 인 범**

1977년 연세대학교 화학공학과 졸업. 1987년 Purdue University 화학공학과 (공학박사). 1988년~현재 포항공과대학교 화학공학과 교수. 1998년~현재 공정산업의 지능자동화 연구센터 소장. 2002년~현재 제어자동화시스템공학회 교육이사. 2004년~현재 한국공학한림원 정회원.