# 선별 시스템 기반 표지 유전자를 포함한 난소암 마이크로어레이 데이터 분류

박수영* · 정채영**

## Classification of Ovarian Cancer Microarray Data based on Intelligent Systems with Marker gene

Su-Young Park* · Chai-Yeoung Jung**


### 요 약

마이크로어레이 분류는 전형적으로 분류기 디자인과 에러 추정이 현저하게 작은 샘플에 기반한다는 것과 교차검증 에러 추정이 대다수의 논문에 사용된다는 주목할 만한 두 가지 특징을 소유한다. 마이크로어레이 난소 암 데이터는 수 만개의 유전자 발현으로 구성되어 있고, 이러한 정보를 동시에 분석하기 위한 어떤 체계적인 절차도 없다. 본 논문에서는, 통계에 따라 유전자의 우선순위를 정함으로써 표지유전자를 선택하였고, 널리 보급되어 있는 분류 규칙인 선형 분류 분석, 3-nearest-neighbor와 결정 트리 알고리즘은 표지 유전자를 선택한 데이터와 선택하지 않는 데이터의 분류 정확도 비교를 위해 사용되어졌다.

ANOVA를 이용하여 선택된 표지 유전자를 포함하는 마이크로어레이 데이터 셋에 선영 분류분석 규칙을 적용한 결과 97.78%의 가장 높은 분류 정확도와 가장 낮은 예측 에러 추정치를 나타내었다.

### ABSTRACT

Microarray classification typically possesses two striking attributes: (1) classifier design and error estimation are based on remarkably small samples and (2) cross-validation error estimation is employed in the majority of the papers. A Microarray data of ovarian cancer consists of the expressions of thens of thousands of genes, and there is no systematic procedure to analyze this information instantaneously. In this paper, gene markers are selected by ranking genes according to statistics, popular classification rules - linear discriminant analysis, k-nearest-neighbor and decision trees - has been performed comparing classification accuracy of data selecting gene markers and not selecting gene markers.

The Result that apply linear classification analysis at Microarray data set including marker gene that are selected using ANOVA method represent the highest classification accuracy of 97.78% and the lowest prediction error estimate.



키워드

마이크로어레이, statistics, gene markers, classification rules

Key word

microarray, statistics, gene markers, classification rules



* 정회원 : 조선대학교

** 정회원 : 조선대학교 (교신저자, cyjung@chosun.ac.kr)



접수일자 : 2010. 09. 16

심사완료일자 : 2010. 10. 05

## Ⅰ. 서 론

마이크로어레이 응용에서 주요 관심은 다른 발현 패턴 예를 들어 암 분류를 거쳐 분류를 수행하는 것이다. 이러한 수행은 마이크로어레이를 가지고 다른 조직으로부터 획득한 RNA에서 발현 수준을 평가하는 것을 요청하고, 발현 수준은 분류기 변수를 사용 될 수 있는 유전자를 결정하며, 이 유전자는 규칙을 샘플로부터 분류기를 디자인 위해 적용된 다음 에러 추정 처리에 적용된다.

최근에, 마이크로어레이 데이터로부터 정보력 있는 유전자를 선택하기 위해 특징 선택, 상관관계 방법, 비모수적 득점 접근, 그리고 베이지안 변수 선택 접근처럼 많은 방법들이 제안되었다[1].

본 논문에서, 마이크로어레이 데이터에 서 표지유전자 선택과 분류를 위한 시스템을 제안하였다. 제안된 시스템에서, t-test는 제안된 시스템에서 타겟 유전자를 발견하기 위해 처음으로 적용되었고, 타겟 유전자로부터 표지 유전자를 발견하기 위해 ANOVA 방법은 실시되었다. 분류 규칙 방법(linear discriminant analysis, 3-nearest-neighborhood and decision trees (CART))은 선택된 표지 유전자의 암 조직을 검증하는데 적용되었고, 기존의 가공하지 않은 데이터 셋과 표적 유전자 데이터 셋을 사용하는 마이크로어레이 시스템과 분류 정확도 비교를 위해 사용되어졌다.

본 논문의 2장에서는 표적 유전자와 표지 유전자를 선택하는 방법 대해 설명하고, 3장에서 시스템을 제안하고 시스템을 위해 필요한 분류 규칙 방법들은 설명한다. 4장에서는 3장에서 제안한 시스템이 실제로 어떤 성능을 보이는지에 대해 실험한 과정과 그 결과를 제시하고 이를 분석한다. 5장에서는 결론을 도출한다.

## Ⅱ. 유전자 선택 방법

### 2.1 t-test

유전자 i 의 t-score(TS)는 다음처럼 정의된다[2].

$$TS_i = \left\{ \left| \frac{\overline{x_{ik}} - \overline{x_i}}{d_{k^{s_i}}} \right|, k = 1, 2, ..., K \right\} \quad (1)$$

여기에서

$\overline{x_{ik}} = \sum_{j \in C_k} \overline{x_{ij}} / n_k$ 이고, $\overline{x_i} = \sum_{j=1}^{n} x_{ij} / n$ 이다.

$$S_i^2 = \frac{1}{n-K} \sum_k \sum_{j \in C_k} (x_{ij} - \overline{x_{ik}})^2 \quad (2)$$

$$d_k = \sqrt{1/n_k + 1/n} \quad (3)$$

$K$ 개의 클래스가 있고, $\max\{y_k, k = 1,2, ..., K\}$는 모두 $y_k, k= 1,2, ..., K$의 최대값이다. $C_k$ 는 $n_k$ 개의 샘플을 포함한 클래스 $k$를 참조하고, $x_{ij}$는 샘플 $j$안에 있는 유전자 $i$의 발현 값이다. $\overline{x_{ik}}$는 클래스 $k$안에 유전자 i의 평균 발현 값이고, $\overline{x_i}$ 는 유전자 $i$에 대한 전체 평균 발현 값이다. $S_i$ 는 유전자 $i$에 대해 모여진 클래스 내 표준 편차이다. 실제로 여기에서 사용된 t-score 는 특별한 클래스와 모든 클래스의 전체 중심 사이에 t-통계량이다.

본 논문에서는 t-test 기반 특징 서열 측정을 사용하여 각 유전자의 중요성 순위를 계산하였고, 그 다음 단계에서 분류를 위해 유의수준 0.05%에 속하는 중요한 유전자만을 유지하였다.

### 2.2 Analysis Of Variance(ANOVA)

ANOVA 는 전체 결과 분산에서 각 입력 요소 (파라미터)의 평균 기여(주요 효과)를 평가하고 요소들 사이에 상호작용을 또한 평가 할 수 있다. 다른 포괄적인 방법들(즉, Sobol과 multiple regression)이 넓은 범위의 양적인 요소를 견본으로 조사하는 반면, ANOVA 에서 각 요소는 제한된 수의 전혀 다른 값(수준)에서 취해진다.

ANOVA 결과는 시뮬레이션 디자인이 잘 안정된다면 특히 직교한다면 해석하기가 더 쉽고 완전히 같은 반복 인수를 갖는 ANOVA 모델 디자인은 훌륭한 통계 특징을 갖지만 시뮬레이션의 수는 빨리 증가한다. 왜냐하면 $p$ 개의 수준을 갖는 n개 요소의 완전한 디자인은 $p^n$ 개의 시뮬레이션 실행을 요청하기 때문이다[3].

본 논문에서 ANOVA는 표지 유전자를 선택하기 위해 수행되어졌다.

## Ⅲ. 제안된 시스템

본 논문에서는 제안된 시스템의 성능을 테스트하기 위해 China Medical University Hospital에서 수집된 난소암 마이크로어레이 데이터를 사용하였다.

### 3.1 제안하는 시스템 구조도

제안된 시스템에서는 초기 난소암 마이크로어레이 데이터에서 난소 종양과 난소암 클래스를 발견한 후 두 클래스와 밀접하게 관련된 표적 유전자를 발견하기 위해 t-test는 처음으로 적용되었고, 표적 유전자로부터 표지 유전자를 발견하기 위해 ANOVA 방법은 실시되었다.

linear discriminant analysis(이하 LDA), k-nearest-neighbor(이하 KNN) and decision trees(이하 DT) 방법은 선택된 표지 유전자의 암 조직을 검증하는데 적용되었고, 기존의 마이크로어레이 시스템과 분류 정확도 비교를 위해 사용되어졌다. 제안된 시스템 구성도는 그림 1과 같다.

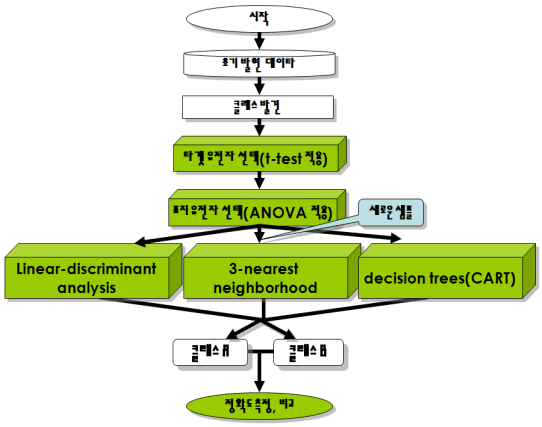


그림 1. 제안하는 분류 시스템
Fig. 1 proposing classification system

### 3.2 LDA

LDA 는 패턴을 부분공간으로 사상하였을 때 사상된 패턴들의 클래스별 분리도를 최대로 하는 축들로 구성된 공간을 탐색하는 것이다. 이를 위해 Fisher의 판변함수에서 정의하는 클래스-내 분산과 클래스-간 분산은 각각 식(4), 식(5)과 같다.

$$S_w = \sum_{i=1}^{c} \sum_{x \in C_i} (x - m_i)(x - m_i)^T \quad (4)$$

$$S_b = \sum_{i=1}^{c} n_i (m_i - m)(m_i - m)^T \quad (5)$$

$c$ 는 클래스의 수, $Ci$ 는 $i$ 번째 클래스에 속하는 패턴 집합, $n_i$ 는 $i$ 번째 클래스에 속하는 샘플 수, $m$ 은 전체 학습피턴의 평균이다.

### 3.3 KNN

비모수적인 기법들의 대표적인 알고리즘인 KNN 분류 알고리즘은 훈련데이터 $X_1, ..., X_m$ 은 모집단 $X$ 에서 그리고 $Y_1, ..., Y_n$ 은 모집단 $Y$ 에서 추출된 표본이고 $X_i$'s는 모집단 X의 i 번째 표본이다. 새로운 관측 값이 $Z$ 일 때 $KNN$ 분류자 $\Theta(Z)$ 는 식 (6)처럼 정의된다.

$$\Theta(Z) = \text{type X id the number of Xi's} \in \text{Nk(z)} \geq \frac{k}{2} = \text{type Y other wise} \quad (6)$$

여기에서 Nk(Z) 는 Z 의 k-nearest neighborhood 로써 X1, ... Xm 와 Y1 ... Yn 중 Z 와의 거리가 제일 가까운 것부터 k 번째로 가까운 훈련자료를 포함하는 것이다. 즉, 새로 관측된 Z 의 type 은 K-nearest neighborhood 에 포함된 훈련자료 중 더 많이 포함된 type 으로 예측된다. KNN 분류방법에서 k 선택은 매우 중요하다. 이는 k 가 너무 작으면 과 적합 (overfitting) 되고 k 가 너무 커지면 오 분류가 매우 증가하기 때문이다.

### 3.4 DT

의사 결정 트리는 수집된 데이터의 레코드들은 분석하여 이들 집합의 분류를 위한 규칙을 생성해내어 트리

모양의 구조로 의사 결정의 집합을 나타낸다.

가장 보편적으로 사용되는 것으로는 CART, CHAID, C4.5 등의 의사 결정 트리 알고리즘이 있으며, 그 중 본 연구에서는 C4.5를 이용하였다. C4.5 알고리즘은 기계 학습알고리즘분야의 효력 있는 ID3 알고리즘을 기반으로 한다. C4.5 알고리즘과 ID3 알고리즘은 엔트로피(Entropy)를 최소화하는 방향으로 가지치기를 수행하는 방법이다. 식 (7)은 엔트로피를 계산하는 방법이다 [4].

$$Entropy(D) = \sum_{i=1}^{c} -p_i \log p_i \qquad (7)$$

여기에서, $p_i = \frac{freq(C_i, S)}{|S|}$ 이다. S는 주어진 데이터들의 집합이고, C는 클래스값들의 집합이다. 따라서, $freq(C_i, S)$ 은 S에서 클래스 Ci에 속하는 레코드 수이다. $|S|$는 주어진 데이터들의 집합의 데이터 개수이다.

## Ⅳ. 실험 및 결과 고찰

### 4.1. 실험 결과 및 고찰

실험에 사용된 샘플은 China Medical University Hospital에서 수집된 5개의 난소 종양과 난소암 샘플이 포함된 난소암 마이크로어레이 데이터를 사용하였다. 데이터는 샘플들에서 획득한 유전자를 각각 Cy5, Cy3로 염색한 다음, 2400개 이상의 알려진 유전체와 7070여개의 새로운 유전체가 찍힌 유리칩을 이용한 cDNA 마이크로어레이 실험에서 획득한 마이크로 어레이 데이터를 사용하였다. 그림 2는 실험에서 획득한 가공하지 않은 마이크로어레이 데이터 산점도이다.

그림 3은 통계 프로그램 R을 사용하여 t-테스트 결과 획득한 유의수준 0.05%에 속하는 3795개의 표적 유전자 산점도이다.

선택된 타겟 유전자들은 표지 유전자 선택을 위해 ANOVA 방법에 사용되어졌다. ANOVA 방법을 적용한 결과 GeneNumber 'HG2538-HT2636, D78129 HG3914-HT4186, D83669, D14689, D38462, HG3264-HT3444, AD000092, HG3342-HT3523'등 10개의 표지 유전자들이 추출되었다.

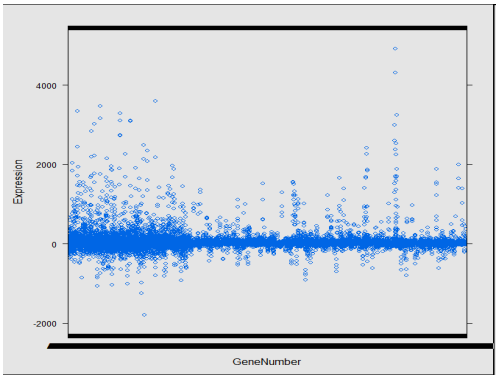


그림 2. 원자료의 산점도
Fig. 2 plot of original data

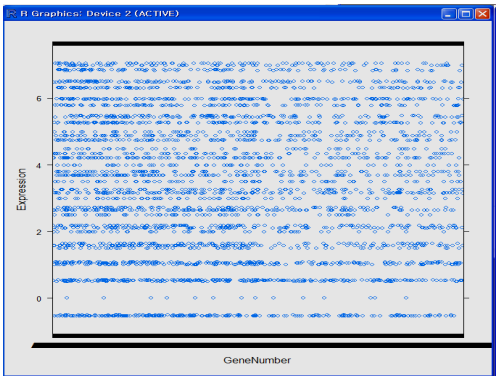


그림 3. 타겟 유전자의 산점도
Fig. 3 plot of target gene

### 4.2 분석 결과

본 논문의 목적은 표지 유전자의 분류 정확도를 측정하기 위해 Raw 마이크로어레이 데이터와 표적 유전자가 선택된 마이크로어레이 데이터를 사용하는 기존의 시스템과 표지 유전자가 선택된 마이크로어레이 데이터를 사용하는 제안된 시스템에 각각 LDA, KNN, DT를 적용하여 분류 정확도와 예측 에러 차이를 조사하는 것이다.

기계 학습 툴 WEKA를 이용하여 각각의 데이터 셋의 분류 정확도와 예측 에러 차이를 측정하였다.

각 데이터 셋의 분류 정확도를 비교하기 위해 조건 위험 추정치는 계산 되었고 서로 그리고 실제 조건 위험과 비교되었다. 이것에 대한 평가는 평균제곱 오차를 나타내며, 실제 클래스와 예측한 클래스 차이를 제곱한 결과를 나타내는 MSE(the Mean Squared Error)와 치우침을 나타내는 bias로 수행되었고, 이 값이 작을수록 좋은 분류를 나타낸다. MSE와 bias는 다음 식 (8)와 (9)와 같다.

$$MSE = \frac{1}{r}\sum_{r=1}^{R}(\hat{\theta}_{n,r} - \tilde{\theta}_{n,r})^2 \quad (8)$$

$$Bias = \frac{1}{r}\sum_{r=1}^{R}(\hat{\theta}_{n,r} - \tilde{\theta}_{n,r}) \quad (9)$$

여기에서, $(\hat{\theta}_{n,r})$은 리 샘플링 조건 위험이고 $(\tilde{\theta}_{n,r})$는 r 번째 반복의 조건 위험이다. 모든 결과에 있어, 전체 반복 수 R = 100 으로 조정되었다.

Raw 마이크로어레이 데이터 셋에는 20280개의 유전자가 사용되었고, 타겟 유전자가 선택된 마이크로어레이 데이터 셋에는 3795개의 타겟 유전자가 사용되었고, 표지 유전자가 선택된 마이크로어레이 데이터 셋에는 10개의 유전자가 사용되었다.

마이크로어레이 데이터 셋에 LDA, KNN, DT를 적용하여 측정한 분류 정확도와 예측 에러 추정치 결과는 표 1과 같다. 이를 Raw 마이크로어레이 데이터 셋을 표적 유전자와 표지 유전자 선택에 따른 분류 정확도와 예측 에러를 비교하기 위한 실험의 대조군으로 하였다. 단위는 퍼센트(%) 이다.

표 1. 마이크로어레이 데이터 셋에 대한 분류정확도와 예측 에러 추정치
table 1. Classification Accuracy and Prediction error estimate on Microarray Data

| Classification method | Law Data | | |
|---|---|---|---|
| | Accuracy | Bias | MSE |
| LDA | 90.12 | -0.019 | 0.008 |
| KNN | 90.05 | -0.018 | 0.016 |
| DT | 89.88 | -0.2 | 0.02 |

| | Data Selected Target gene | | |
|---|---|---|---|
| | Accuracy | Bias | MSE |
| LDA | 95.42 | -0.013 | 0.004 |
| KNN | 92.84 | -0.014 | 0.006 |
| DT | 94.02 | -0.013 | 0.004 |
| | Data Selected Marker gene | | |
| | Accuracy | Bias | MSE |
| LDA | 97.78 | -0.003 | 0.003 |
| KNN | 96.92 | -0.004 | 0.004 |
| DT | 96.89 | -0.004 | 0.004 |

표지 유전자가 선택된 마이크로어레이 데이터 셋에 선형분류 규칙 알고리즘을 적용한 결과 97.78%의 가장 높은 정확도와 가장 낮은 0.003%의 MSE, -0.003의 bias를 보였다. 반면, 기존의 표지 유전자를 선택하지 않고 가공하지 않은 데이터 셋을 사용한 마이크로어레이 데이터셋에 DT알고리즘을 적용한 결과에서는 89.88%의 가장 낮은 정확도와 0.02%, -0.2%의 가장 높은 MSE와 bias를 보였으며, 표적 유전자가 선택된 마이크로어레이 데이터 셋에 분류 규칙 알고리즘을 적용한 결과 가공하지 않은 마이크로어레이 데이터 셋에 분류 규칙 알고리즘을 적용한 결과보다 높은 정확도와 낮은 MSE와 bias를 보였다.

## V. 결 론

암 연구에 있어, 민감하고 특별한 표지 유전자를 발견한다는 것은 어려운 일이다. 본 논문에서는 분류 규칙 알고리즘을 이용하여 가공하지 않은 마이크로어레이 데이터 셋, t-test를 사용하여 선택된 표적 유전자를 포함하는 마이크로어레이 데이터 셋 그리고 ANOVA를 사용하여 선택된 표지 유전자를 포함하는 마이크로어레이 데이터 셋에 널리 보급되어 있는 분류 규칙인 LDA, KNN와 DT 를 이용하여 분류 정확도를 비교 분석하는 시스템을 고안하고 결과를 비교분석하였다.

ANOVA를 이용하여 선택된 표지 유전자를 포함하는 마이크로어레이 데이터 셋에 LDA 규칙을 적용한 결과 97.78%의 가장 높은 분류 정확도와 가장 낮은 예측 에러 추정치를 나타내었다.

제안한 시스템은 ANOVA 방법에 의해 선택된 표지 유전자 포함된 마이크로어레이 데이터셋에 LDA 규칙을 적용한 결과 난소암을 가장 잘 분류한다는 것을 증명하였다. 따라서, 본 논문에서 제안한 시스템은 난소 암 마이크로어레이 데이터에서 유전자 선택과 분류를 하는데 있어 뛰어난 성능을 보였기 때문에 암 진단을 위한 다른 연구에 또한 사용될 수 있을 것으로 기대된다.

## 참고문헌

[1] Jeng J-T, Lee T-T, Lee Y-C. Classification of ovarian cancer based on intelligent systems with microarray data. In: IEEE international conference on systems, man and cybernetics. New York: IEEE Systems, Man and Cybernetics Society; 2005. p.1053-8

[2] J.Devore, and R. Peck, Statistic: the Exploration and Analysis of Data, 3rd ed. Pacific Grove, CA.:Duxbury Press, 1977.

[3] Kobilinsky, A., 1997. Les plans factoriels. In: Droesbeke, J.-J., Fine, J., Saporta, G. (Eds.), Plans d'Experiences, Applications a l'Entreprise. Editions Technip, Paris, pp. 69-209.

[4] Tom Michael, Machine Learning, McGraw-Hill Companies International Editions, 1997.

**박수영(Su-Young Park)**

2001년 조선대학교 컴퓨터통계학과 이학사
2003년 조선대학교 컴퓨터통계학과 이학석사
2007년 조선대학교 컴퓨터통계학과 이학박사
※관심분야 : 신경망, 인공지능, 정보보호, 멀티미디어, 멀티미디어 컨텐츠, Bioinformatics

**정채영(Chai-Yeoung Jung)**

1987년 조선대학교 컴퓨터공학과 공학석사
1989년 조선대학교 컴퓨터공학과 공학박사
1986년~현재 조선대학교 컴퓨터통계학과 교수
※관심분야 : 신경망, 인공지능, 정보보호, 멀티미디어, 멀티미디어 컨텐츠, Bioinformatics