# 표준화 기반 표지 유전자를 이용한 난소암 마이크로어레이 데이타 분류 시스템

박수영* · 정채영**

# Ovarian Cancer Microarray Data Classification System Using Marker Genes Based on Normalization

Su-Young Park* · Chai-Yeoung Jung**

이 논문은 2011년도 조선대학교 학술연구비의 지원을 받아 연구되었음


## 요 약

표지 유전자는 특정한 실험 조건의 특성을 나타내주는 발현수준의 유전자를 의미한다. 이 유전자들은 여러 집단 간의 발현수준에서 유의한 차이를 보여주며, 실제로 집단 간의 차이를 유발하는 유전자일 확률이 높아 특정 생물학적 현상과 관련 있는 표지 유전자를 찾는 연구에 이용될 수 있다.

본 논문에서는, 먼저 그 동안 제안된 여러 표준화 방법들 중에서 가장 널리 사용되고 있는 방법들을 이용하여 데이터를 표준화 한 후 통계에 따라 유전자의 우선순위를 정함으로써 표지유전자를 추출할 수 있는 시스템을 제안하였다. 다층퍼셉트론 신경망 분류기를 이용하여 각 표준화 방법들의 성능을 비교분석하였다. 그 결과 Lowess 표준화 후 ANOVA를 이용하여 선택된 8개의 표지 유전자를 포함하는 마이크로어레이 데이터 셋에 MLP 알고리즘을 적용한 결과 99.32%의 가장 높은 분류 정확도와 가장 낮은 예측 에러 추정치를 나타내었다.

## ABSTRACT

Marker genes are defined as genes in which the expression level characterizes a specific experimental condition. Such genes in which the expression levels differ significantly between different groups are highly informative relevant to the studied phenomenon.

In this paper, first the system can detect marker genes that are selected by ranking genes according to statistics after normalizing data with methods that are the most widely used among several normalization methods proposed the while, And it compare and analyze a performance of each of normalization methods with mult-perceptron neural network layer. The Result that apply Multi-Layer perceptron algorithm at Microarray data set including eight of marker gene that are selected using ANOVA method after Lowess normalization represent the highest classification accuracy of 99.32% and the lowest prediction error estimate.



키워드

마이크로어레이,표준화, 표지 유전자, 다층퍼셉트론

Key word

microarray, Normalization, marker genes, multi-layer perceptron



* 정회원 : 조선대학교

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



접수일자 : 2011. 01. 18

심사완료일자 : 2011. 02. 09

## Ⅰ. 서 론

DNA 마이크로어레이(또는 microchip)에서 얻어진 자료를 간단히 마이크로어레이 자료라고 한다. 이러한 자료는 잡음(noise)이 많이 포함되어 있다. 잡음이 추가될수록 마이크로어레이의 품질은 떨어지기 마련이며 특히 일정한 패턴을 지닌 잡음은 분석결과에 큰 영향을 미칠 수 있다. 따라서 마이크로어레이를 분석하는 초기 단계에서 잡음을 제거하는 과정을 거친다. 이런 과정을 표준화(normalization)라고 한다[1].

최근에, 마이크로어레이 데이터로부터 정보력 있는 유전자를 선택하기 위해 특징 선택, 상관관계 방법, 비모수적 득점 접근, 그리고 베이지안 변수 선택 접근처럼 많은 방법들이 제안되었다[2]. 이러한 방법들은 암조직에서 특별한 마커로써 행동할 수 있는 유전자 생성물을 확인하는 것이지만 전체적인 유전자 발현 분석에 대해 더 좋은 통찰력을 성취하기 위한 어떠한 체계적인 접근이 없고 확신하는 새로운 마커도 전혀 확인되어 오지 않았다.

본 논문의 2장에서 표준화 방법에 대해 소개하고, 3장에서 표적 유전자와 표지 유전자를 선택하는 방법들은 설명한다. 4장에서는 본 논문이 수행한 시스템 설계 및 구현과정을 설명하고 결과를 비교분석 한다. 5장에서는 결론을 도출한다.

## Ⅱ. 표준화

### 2.1 표준화 방법

DNA 마이크로어레이 실험에서 얻어진 자료에서 Cy3의 발현 값을 G, Cy5의 발현 값을 R이라고 하자. 실험 대상의 전체 유전자 수를 p라고 하고 각각의 유전자를 j로 나타내자. 발현 값의 비(ratio) M과 intensity A는 다음과 같이 정의된다[1].

$$M = \log\frac{R}{G} = \log R - \log G, \qquad (1)$$
$$A = \log\sqrt{GR} = \frac{1}{2}(\log G + \log R)$$

표준화 방법은 global(G) 표준화 방법과 A를 고려하는 intensity dependent(ID)표준화 방법으로 구분한다. G 표준화 방법은 각 유전자별로 M을 다음과 같이 표준화 한다.

$$M_j^{Global} = M_j - \hat{c} \qquad (2)$$

여기서 $\hat{c}$는 M의 중앙값을 이용하여 추정할 수 있다. ID 표준화 방법에는 선형관계를 가정하는 경우와 비선형 관계를 가정하여 표준화하는 방법으로 나눌 수 있다. ID 비선형 표준화 경우에는 LOWESS와 같은 비선형 모형을 이용하여 다음과 같이 표준화 한다.

$$M_j^{LOWESS} = M_j - \hat{c}(A_j) \qquad (3)$$

여기서 $\hat{c}$가 적합한 비선형함수이다[1].

## Ⅲ. 유전자 선택 방법

### 3.1 t-test

유전자 i 의 t-score(TS)는 다음처럼 정의된다[2].

$$TS_i = \left\{ \left| \frac{\overline{x_{ik}} - \overline{x_i}}{d_k} \right|, k = 1, 2, ..., K \right\} \qquad (4)$$

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.

$$S_i^2 = \frac{1}{n-K} \sum_k \sum_{j \in C_k} (x_{ij} - \overline{x_{ik}})^2 \qquad (5)$$

$$d_k = \sqrt{1/n_k + 1/n} \qquad (6)$$

K 개의 클래스가 있고, max{yk, k = 1,2, ..., K}는 모두 yk, k= 1,2, ..., K의 최대값이다. Ck 는 nk 개의 샘플을 포함한 클래스 k를 참조하고, xij는 샘플 j안에 있는 유전자

i의 발현 값이다. $\overline{x_{ik}}$는 클래스 k안에 유전자 i의 평균 발현 값이고, $\overline{x_i}$ 는 유전자 i에 대한 전체 평균 발현 값이다. Si는 유전자 i에 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.

### 3.2 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 pn 개의 시뮬레이션 실행을 요청하기 때문이다[3].

본 논문에서 ANOVA는 표지 유전자를 선택하기 위해 수행되어졌다.

### 3.3 Multi-Layer Perceptron(이하: MLP)

인공 신경망의 대표적인 기계 학습 알고리즘인 다층 퍼셉트론은 대부분의 패턴 인식 문제에 대해 안정적인 성능을 보이며, 일단 학습이 끝나면 응용 단계에서는 매우 빠르게 결과를 출력한다. 다층퍼셉트론은 백프로퍼게이션(back propagation)알고리즘을 사용하는데 이것은 출력 층의 오차 신호를 이용하여 은닉 층과 출력 층 사이의 연결 강도를 변경하고 출력 층의 오차 신호를 은닉 층에 역전파하여 입력 층과 은닉 층 사이의 연결 강도를 변경하는 학습법이다[4].

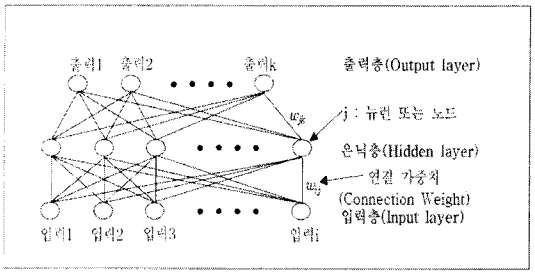


그림 1. MLP 신경망 구조
Fig. 1. Neural network structure of MLP

## Ⅳ. 실험 및 결과 고찰

제안하는 시스템의 흐름은 다음과 같다. 먼저 초기 난소암 마이크로어레이로부터 유전자 발현 데이터를 획득한다. DNA 칩을 이용한 마이크로어레이 실험에서 얻어진 자료에는 보톤 실험 자료에 비해 잡음이 많이 포함되어 있으며 일정한 패턴을 보이는 경우 분석결과는 치명적인 오류를 범할 수 있으므로 잡음을 제거하기 위해 각각 Global, Lowess 표준화를 거쳐 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. 기계 학습 기반 분류기로 MLP를 사용하여 표준화의 분류 성능을 비교하는 구조이다.

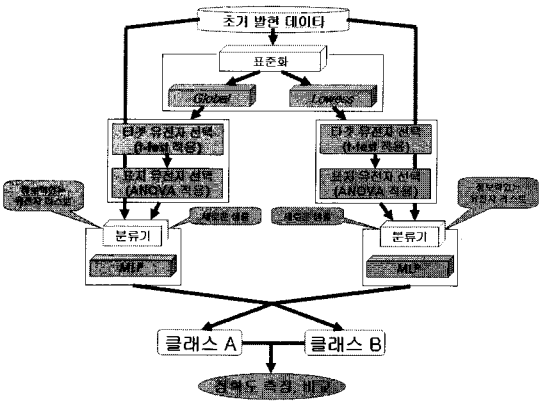


그림 2. 제안하는 분류 시스템
Fig. 2. proposing classification system

### 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 3은 실험에서 획득한 가공하지 않은 마이크로어레이 데이터 산점도와 표준화 후 마이크로어레이 데이터 산점도의 일부분이다.

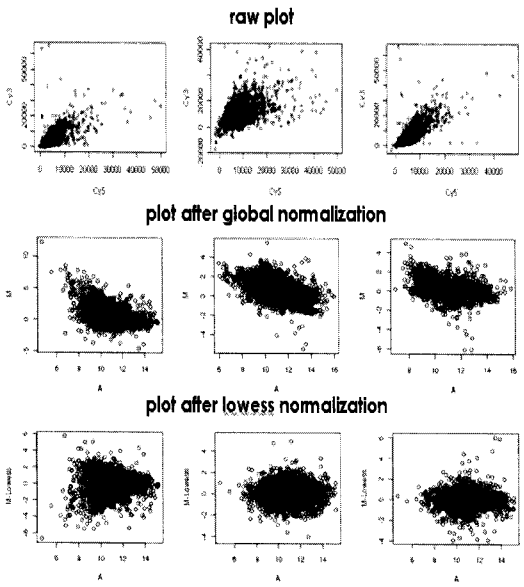


그림 3. 마이크로어레이 데이터의 산점도
Fig. 3. plot of microarray data

그림 4는 통계 프로그램 R을 사용하여 t-test 결과 획득한 유의수준 0.05%에 속하는 표적 유전자 산점도의 일부분이다.

그림 5는 t-test 후 선택된 표적 유전자들을 사용하여 ANOVA 방법에 의해 선택된 표지 유전자 산점도의 일부분이다.

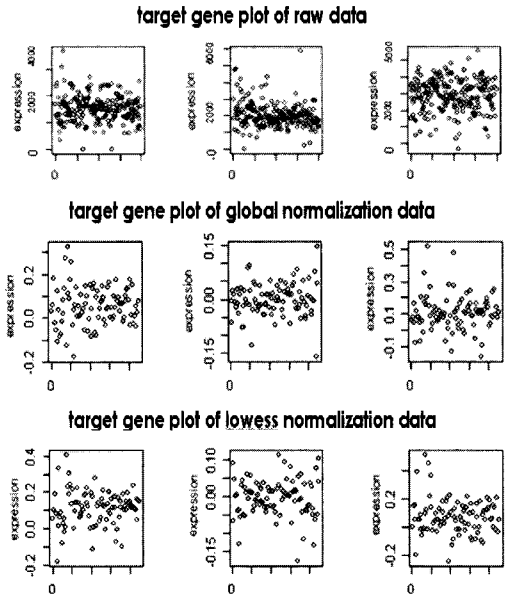


그림 4 t-test 후 표적 유전자 산점도
Fig. 4. Target gene plot after t-test

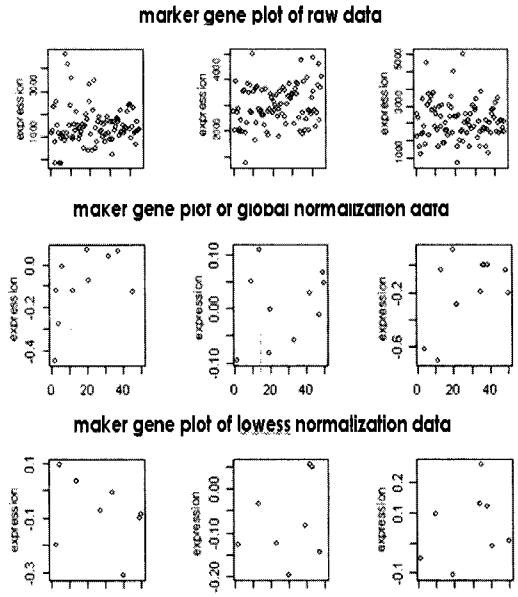


그림 5. ANOVA 후 표지 유전자 산점도
Fig. 5. Marker gene plot after ANOVA

### 4.2. 분석 결과

기계 학습 툴 WEKA를 이용하여 분류 성능을 평가하기 위해 MLP 신경망을 구현하고 모멘텀은 0.09로, 총 레이어수는 3으로 고정한 후, 학습률을 0.01에서 0.05로 변화시켜가며 실험하였으며 10-fold cross validation을 이용하여 정확도를 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 (7)와 (8)와 같다.

$$MSE = \frac{1}{r}\sum_{r=1}^{R}(\hat{\theta}_{n,r} - \tilde{\theta}_{n,r})^2 \quad (7)$$

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

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.

가공하지 않은 마이크로어레이 데이터 셋에는 20280개의 유전자가 사용되었다. 가공하지 않은 마이크로어레이 데이터 셋에는 3795개의 표적 유전자가 사용되었고, 표준화 후 마이크로어레이 데이터 셋에는 각각 200개의 표적 유전자가 사용되었다. 가공하지 않은 마이크로어레이 데이터 셋에는 90개의 표지 유전자가 사용되었고, 표준화 후 마이크로어레이 데이터 셋에는 각각 10개와 8개의 표지 유전자가 사용되었다.

가공하지 않은 마이크로어레이 데이터 셋과 표준화 후 마이크로어레이 데이터 셋 각각에서 선택된 표적 유전자와 표지 유전자를 포함하는 마이크로어레이 데이터 셋에 MLP 적용하여 측정한 분류 정확도와 예측 에러 추정치 결과는 표 1과 같다. 가공하지 않은 마이크로어레이 데이터 셋을 표준화 후 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

| gene selection | raw microarry Data | | |
|---|---|---|---|
| | Accuracy | Bias | MSE |
| target gene | 87.12 | -0.021 | 0.008 |
| marker gene | 88.05 | -0.019 | 0.016 |
| | microarray Data After global normalization | | |
| | Accuracy | Bias | MSE |
| target gene | 93.72 | -0.012 | 0.004 |
| marker gene | 96.14 | -0.010 | 0.002 |
| | microarray Data After Lowess normalization | | |
| | Accuracy | Bias | MSE |
| target gene | 98.08 | -0.001 | 0.002 |
| marker gene | 99.32 | -0.001 | 0.001 |

Lowess 표준화 후 표지 유전자가 선택된 마이크로어레이 데이터 셋에 MLP 알고리즘을 적용한 결과 98.08%와 99.32%의 가장 높은 정확도와 가장 낮은 MSE와 bias를 보였다. 반면, 기존의 표준화를 하지 않고 표적 유전자와 표지 유전자를 선택한 마이크로어레이 데이터 셋에 MLP 알고리즘을 적용한 결과에서는 87.12%와 88.05%의 가장 낮은 정확도와 가장 높은 MSE와 bias를 보였으며, global 표준화 후 표적 유전자와 표지 유전자가 선택된 마이크로어레이 데이터 셋에 MLP 알고리즘을 적용한 결과 표준화를 하지 않은 마이크로어레이 데이터 셋에 MLP 알고리즘을 적용한 결과보다 높은 정확도와 낮은 MSE와 bias를 보였다.

## V. 결 론

마이크로어레이 실험에서 얻어진 원자료에는 다양한 종류의 잡음이 포함되어 있다. 표준화 과정은 본격적인 마이크로어레이 자료의 통계적 분석 이전에 실시되는 가장 주요한 전처리 과정 분석이다. 또한, 암 연구에 있어 민감하고 특별한 표지 유전자를 발견한다는 것은 어려운 일이다. 본 논문에서는 표준화 방법을 적용한 후

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 MLP 알고리즘을 적용하여 분류 정확도를 비교 분석하는 시스템을 고안하고 결과를 비교분석하였다.

그 결과 Lowess 표준화 후 ANOVA를 이용하여 선택된 표지 유전자를 포함하는 마이크로어레이 데이터 셋에 MLP 알고리즘을 적용한 결과 99.32%의 가장 높은 분류 정확도와 가장 낮은 예측 에러 추정치를 나타내었다.

제안한 시스템은 Lowess 표준화 후ANOVA 방법에 의해 선택된 표지 유전자가 포함된 마이크로어레이 데이터 셋에 MLP 알고리즘을 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] M. Brown, W. Grundy, D. Lin, N. Christianini, C. Sugnet, M. Ares Fr., and D. Haussler, "Support vector machine classification of microarray gene expression data", UCSC-CRL 99-09, Department of Computer Science, University California Santa Cruz, Santa Cruz, CA, June, 1999

[2] 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] Golub, T.R., Slonim, D.K, Tamayo, P., Huard, D., Gaasenbeek, M., Mesiirov, J.P., Collrt, H., Loh, M.L.Dowing, J.R, Caligiuri, M.A., Bloomfield, D.D., and Lander, E.S., "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring", Science, vol. 286, no. 5739, pp. 531-537, 1999.


## 저자소개


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