표준화 기반 유의한 유전자 선택 방법 조합을 이용한 마이크로어레이 분류 시스템 설계

박수영*, 정채영**

The Design Of Microarray Classification System Using Combination Of Significant Gene Selection Method Based On Normalization.

Su-Young Park*, Chai-Yeoung Jung**

본 연구는 문화관광부 및 한국문화콘텐츠진흥원의 문화콘텐츠기술연구소(CT) 육성사업의 연구결과로 수행되었음

요 약

정보력 있는 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 정보적 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 제안한 유사성 척도 조합 방법으로 정보력 있는 유전자들을 추출할 수 있는 시스템을 고안하였다. 다층퍼셉트론 신경망 분류기를 이용하여 각 표준화 방법들의 성능을 비교분석하였다. 그 결과 Lowess 표준화 후 피어슨 적률 상관 계수와 유클리디안 거리 계수 조합을 이용하여 선택된 200 유전자들을 멀티퍼셉트론 신경망 분류기로 분류한 결과 98.84%의 향상된 분류 성능을 보였다.

ABSTRACT

Significant 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 informative genes by similarity scale combination method being proposed in this paper 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 multi-perceptron neural network layer. The Result classifying in Multi-Perceptron neural network classifier for selected 200 genes using combination of PC(Pearson correlation coefficient) and ED(Euclidean distance coefficient) after Lowess normalization represented the improved classification performance of 98.84%.

키워드

Lowess normalization, PC-ED combination method, MLP(multi-Layer perceptron)

* 조선대학교 컴퓨터통계학과

** 교신저자

접수일자 2008. 09. 02

Ⅰ. 서 론

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

본 논문의 구성은 다음과 같다. 2장에서 표준화에 대해 소개하고, 3장에서 유의한 유전자 선택 방법과 본 논문에서 제안한 조합 방법을 소개한다. 4장에서는 본 논문이 수행한 시스템 설계 및 구현과정을 설명하고 결과를 비교분석 한다. 5장에서는 결론을 도출한다.

Ⅱ. 표준화

마이크로어레이 자료에는 많은 잡음이 포함되어 있다. 예를 들어, cDNA 마이크로어레이 실험에서는 녹색 Cy3와 적색 Cy5 염료간의 형광 물리적 차이에 의해서 잡음이 발생할 수 있으며 형광염료의 혼합비율의 차이에 의해서도 잡음이 발생할 수 있다. 또한 이미지 분석에서 스캐너의 레이저 강도 등의 다양한 요인에 의해서도 역시 잡음이 발생할 수 있다. 표준화는 유전자 발현 값에 영향을 미치는 다양한 형태의 잡음을 찾아내어 제거하는 과정이라고 할 수 있다.

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[2].

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

표준화 방법은 global(G) 표준화 방법과 A를 고려하는 intensity dependent(ID) 표준화 방법으로 구분한다. G 표준화방법은 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 $LOWESS$와 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.

Ⅲ. 유의한 유전자 선택 방법

3.1. 유의한 유전자 선택

각 클래스에 대한 특징을 극단적으로 뚜렷하게 나타내면서 이상적으로 발현하는 유전자를 G_{ideal}이라고 하면, 종양 세포의 특징을 1로 정의하고 나머지 정상세포 혹은 다른 종양 세포의 특징을 0으로 정의하여 식 (1)과 같은 벡터로 표현할 수 있다. G_{ideal}은 이상 유전자 모델과 같은 의미이다.

$$G_{ideal} = (1, 1, 1, \ldots, 1, 0, 0, 0, \ldots, 0) \qquad (1)$$

이제 여러 개의 유사성 척도를 각각 사용하여 식 (1)과 각 유전자 사이의 유사성 여부를 측정한다. 각 유사성 척도별로 이상 유전자 모델과 유사도가 높은 유전자들을 순차 정렬하고 상위의 유전자 일부를 선택하여 분류기의 학습 데이터로 사용한다. 이 때 선택해야 하는 상위 유전자의 수는 20에서 200개가 안정적인 분류 결과를 나타내는 것으로 알려져 있다[3].

본 논문에서는 서열척도로 스피어만의 서열 상관계수와 유클리디안 거리척도와 피어스만 상관관계 척도를 사용하였다.

- Pearson correlation Coefficient(PC)

$$PC(G_i, G_{ideal}) = \frac{\sum G_i G_{ideal} - \frac{\sum G_i \sum G_{ideal}}{N}}{\sqrt{\left(\sum G_i^2 - \frac{(\sum G_i)^2}{N}\right)\left(\sum G_{ideal}^2 - \frac{(\sum G_{ideal})^2}{N}\right)}}$$

- Spearman correlation Coefficient(SC)

$$SC(G_i, G_{ideal}) = 1 - \frac{6\sum (D_G - D_{ideal})^2}{N(N^2-1)}$$

- Euclidean distance(ED)

$$ED(G_i, G_{ideal}) = \sqrt{\sum (G_i - G_{ideal})^2}$$

그림 1. 유전자 선택을 위한 유사성 척도
Fig. 1. the similar scale for gene selection

3.2. 조합 방법

기존 방법과 같이 각 유사성 척도를 개별적으로 사용하여 유용한 유전자 목록을 만들게 되면, 중요한 정보를 내포하고 있다고 판단된 유전자 목록이 각 유사성 척도를 달리할 때마다 상이하게 나타난다. 따라서 제안된 시스템에서는 유사성 척도 한 가지를 사용해서 얻게 되는 유전자 목록의 일관성과 신뢰성의 결여를 보안하기 위해, 여러 개의 유사성 척도를 함께 활용하여 정보력이 있는 유전자 목록을 만든다. 그림 2은 본 논문에서 제안한 다수의 척도에서 정보력 있는 유전자로 평가받은 의미 있는 유전자들을 선택하는 알고리즘이다.

Step 1) N choice of similarity scales	조합하고자 하는 유사성 척도 N개 선택
Step 2) Compute $S_i = \sum_{j=1}^{n} E_j$ n : number of sample	유사성 척도에서 i 번째 유전자 G_i 가 평가 받은 유사도 수치의 총계
Step 3) Compute $V_i = \frac{\sum_{j=1}^{N} S_i}{N}$	합산한 결과 값의 평균
Step 4) combination value = V_i	step 3)의 결과를 유사성 척도 N개의 조합 값
Step 5) Sort of an ascending series	유사도 값이 높은 G_i 를 선택하여 유전자 목록 완성

그림 2. 정보력이 있는 유전자 선택을 위한 조합 알고리즘
Fig. 2. the combination algorithm for significant gene selection

3.3. Multi-Layer Perceptron(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].

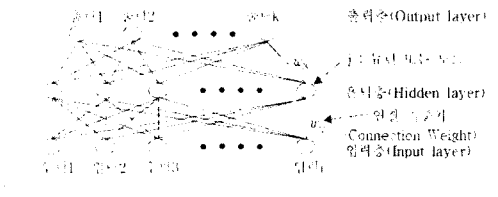

그림 3. MLP 신경망 구조
Fig. 3. Neural network structure of MLP

Ⅳ. 실험 및 결과 고찰

4.1. 시스템 구성

제안하고자 시스템의 흐름은 다음과 같다. 먼저 마이크로어레이로부터 유전자 발현 데이터를 획득한다. DNA 칩을 이용한 마이크로어레이 실험에서 얻어진 자료에는 보통 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 이상 유전자 모델이 확정되고 나면, 각각의 유전자 발현 데이터들에 대해 각 유사성 척도를 사용하여 이상 유전자 모델과의 유사한 정도를 정량적으로 평가한다. 이들 중 여러 개의 척도(최소 2개 이상의 척도)에서 정보력이 있는 유용한 유전자로 평가 받은 유전자들을 정량화된 유용성 정도에 따라 서열화 하고 이들의 상위 부분을 모아 정보력 있는 유전자 목록으로 확정한다. 기계 학습 기반 분류기로 MLP(Multi Layer Perceptron)를 사용하여 표준화의 분류 성능을 비교하는 구조이다.

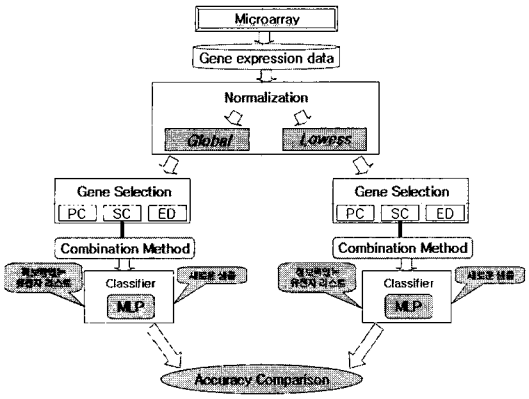

그림 4. 제안하는 분류 시스템
Fig. 4. proposing classification system

4.2. 실험 결과 및 고찰

실험용 데이터로 하버드대학교의 바이오인포메틱스 코어 그룹의 샘플데이터를 사용하였다. 통계 컴퓨터 프로그램인 R을 이용하여 Global, Lowess 표준화 하여 기존의 단일 유사성 척도 3가지를 조합한 유전자 선택 방법 4가지를 이용하여 정보력이 있는 유전자를 선택하고 목록을 만들었다. 이 유전자 목록을 이용하여 멀티퍼셉트론 신경망 분류기를 통해 학습과 테스트를 한 분류 결과를 10-fold 교차검증을 사용하여 표준화 정확도를 서로 비교 분석하였다. 그림 5은 표준화 후 각 유사성 척도에 따라 선택된 상위 200개 유전자 산점도의 일부분이다.

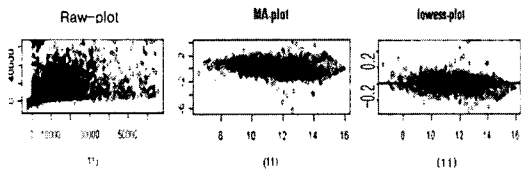

그림 5. 상위 200개 유전자 산점도
Fig. 5. plot of 200 high rank gene

그림 6은 각 유사성 척도에 따라 선택된 상위 200개 유전자산점도의 일부분이다.

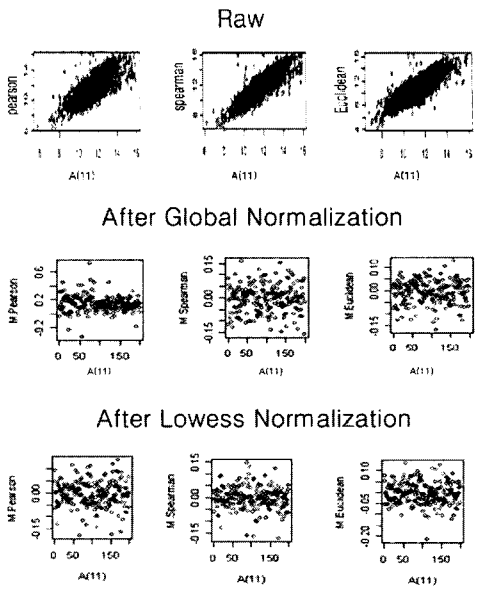

그림 6. 유사성 척도에 따른 상위 200개 유전자 산점도
Fig. 6. Gene plot of 200 high rank gene by similarity scale

4.3 분석 결과

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.

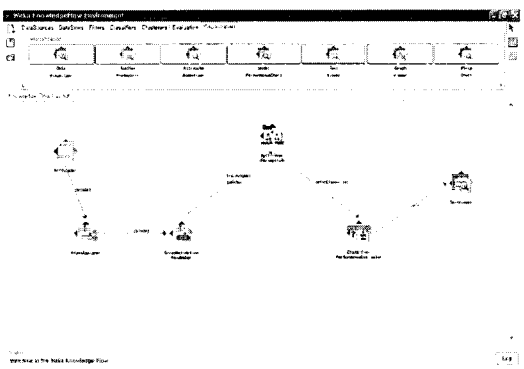

그림 7. 마이크로어레이 분류 시스템
Fig. 7. microarray classification system

'PC'는 피어슨 적률 상관 계수를 뜻하고, 'SC'는 스피어만 상관 계수를 나타내며, 'ED'는 유클리디안 거리 계수를 의미한다. 예를 들어, 'PC-ED'로 표기된 것은 기존의 유사성 척도인 피어슨 적률 상관계수와 유클리디안 거리 계수를 이용하여 선택된 유전자들을 본 논문에서 제안한 조합 방법에 의해 서로 조합하여 유전자를 새롭게 선택하고 목록을 재구성하였음을 의미한다. 그림 7은 데이터 마이닝툴 WEKA를 이용한 마이크로어레이 분류 시스템을 설계한 그림이다.

Raw 데이터와 표준화 한 데이터에 대해 기존의 단일 유사성 척도를 이용하여 상위 200개 유전자를 선택하고 목록을 만들어 MLP를 이용한 표준화 성능을 실험한 결과는 표 1와 같다. 표에 나타나 있는 분류 성능 수치는 해당 조합 조건에서 분류 성공률을 의미하며 그 단위는 퍼센트(%) 이다. 표준화에 따른 분류 성능을 비교 평가하기 위한 대조군으로 사용하였다.

표 1. 단일 척도 사용에 따른 분류 성능
table 1. classification performance according to using a single scale

(%)	PC	ED	SC
Raw	74.19	68.00	67.74
Global	83.87	80.65	79.41
Lowess	88.24	85.29	82.35

Lowess 표준화 후 피어슨 적률 상관관계 척도 척도를 사용하여 유전자 목록을 생성한 뒤 실험 한 경우 다층 퍼셉트론 기반의 분류기를 통해 분류 성능을 확인한 결과 88.24%의 분류 성공률을 보였다. 같은 방법으로 유클리디안 거리 척도를 사용하였을 경우 85.29%의 분류 성능을 보였으며, 스피어만 상관관계 척도를 사용하였을 때에는 82.35%의 분류 성능을 나타내었다.

Global 표준화 후 기존의 단일 유사성 척도에 의해 유전자 목록을 생성한 뒤 실험한 경우 분류기에서 Lowess 표준화 후 유사성 척도에 의해 유전자 목록을 생성한 뒤 실험 경우 보다 모두 낮은 분류 성능을 나타내었으며 Raw 데이터에 대해서는 모두 가장 낮은 분류 성능을 나타내었다.

그러나 표준화 후 이러한 기존의 단일 유사성 척도를 조합 방법에 의해 조합하여 보다 정보력이 있는 유전자 목록을 생성한 뒤 실험한 경우 표준화를 하지 않은 데이터 보다 분류기에서 향상된 분류 성능을 나타내었다. Raw 마이크로어레이 데이터와 표준화 한 마이크로어레이 데이터에 대해 기존의 유사성 척도를 단일하게 사용한 유전자 선택 방법 중 두 가지를 조합하여 실험한 결과, 관찰된 분류 성능은 표 2과 같다.

표 2. 2개 척도 조합에 따른 분류 성능
table 2. classification performance according to combination of 2-scale

(%)	PC-ED	ED-SC	ED-SC
Raw	94.12	91.18	85.29
Global	95.12	94.27	91.34
Lowess	98.84	95.29	97.44

Raw 데이터와 표준화 한 데이터에 단일 유사성 척도를 사용하여 실험한 결과보다 대부분 높은 분류 성능을 나타냈으며 Lowess 표준화 후 PC-ED의 경우 98.84%로 가장 높은 분류 성능을 보였다.

표 3. 3개 척도 조합에 따른 분류 성능
table 3. classification performance according to combination of 3-scale

(%)	PC-ED-SC
Raw	88.24
Global	91.33
Lowess	93.18

Raw 데이터와 표준화 한 데이터에 기존의 유사성 척도를 사용하여 유전자를 선택하는 세 가지를 모두 조합한 경우 표 3과 같은 분류 성능 향상을 보였다. 그러나 이러한 분류 성능 향상은 Lowess 표준화 후 두 가지 척도를 조합하는 경우에 가장 현저하게 나타났고 Global 표준화 후 세 가지 이상의 척도를 조합하는 경우에는 다소 소극적으로 나타났으며 Raw 데이터에 대해서는 가장 낮은 분류 성능을 나타냈다. 표준화를 적용과 상관없이 하나의 유전자 선택 방법만으로는 분류하고자 하는 해 공간을 모두 포함하지 못할 가능성이 있으나, 많은 유전자 선택 방법의 조합이 오히려 포함하지 않아야 할 해 공간까지 포함하는 경우 분류기의 분류 성능을 상대적으로 저하시킬 수도 있을 것으로 추정된다.

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.

본 논문에서는 표준화 방법을 적용한 후 정보력이 있는 유전자 목록을 조합하는 시스템을 고안하고 보다 분류 성능을 향상 시킬 수 있는 조합 방법을 제안하고, 여러 분류기들을 이용하여 실험 평가 하였다. 그 결과 Lowess 표준화 후 제안한 PC-ED조합으로 추출한 데이터를 멀티 퍼셉트론 신경망 분류기로 분류한 결과 98.84%의 정확도를 보여 Raw 데이터에 유사성 척도를 사용하여 유전자 목록을 생성하고 실험을 수행한 경우보다 표준화 후 제안한 조합 방법으로 추출한 데이터를 멀티퍼셉트론 신경망 분류기로 분류한 결과 분류 성능이 향상되었다.

참고문헌

[1] M. Brown, W. Grundy, D. Lin, N. Christianini, C. Sugnet, M. Ares Jr., 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] S. Dudoit, J. Fridlyand, 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.

[3] Dov Stekel, Microarray Bioinformatics, Cambridge University Press, 2003.

[4] Golub, T.R., Slonim, D.K, Tamayo, P., Huard, D., Gaasenbeek, M., Mesirov, J.P., Collrt, H., Loh, M.L.Downing, 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. 5439, 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, 멀티미 디어 컨텐츠, Bioinformatics
※ cyjung@chosun.ac.kr 062)230-6625