마이크로어레이 데이터를 이용한 암 분류 표지 유전자 선별 시스템

박수영* · 정채영**

An Intelligent System of Marker Gene Selection
for Classification of Cancers using Microarray Data

Su-Young Park* · Chai-Yeoung Jung**

이 논문은 2010년도 조선대학교 학술연구비의 지원을 받아 연구되었음

요 약

마이크로어레이를 기반으로 하는 암 분류 방법은 암 종류에 따라 다르게 발현되는 유전자 양상을 통계적으로 발견함으로써 정확한 암 분류에 기여할 수 있다. 따라서 현재의 마이크로어레이 기술을 이용해서 효과적으로 암을 분류하기 위해서는 특정 암과 밀접하게 관련이 있는 정보력 있는 유전자를 선택하는 과정이 필수적이다. 본 논문에서는 난소 암 마이크로어레이 데이터를 이용하여 암에 영향을 미치는 가장 다르게 발현할 가능성이 있는 표지 유전자를 추출할 수 있는 시스템을 고안하고, 다층퍼셉트론 분류기를 이용하여 기존의 마이크로어레이 시스템과 분류 성능을 비교분석하였다. 그 결과 ANOVA를 이용하여 선택된 표지 유전자를 포함하는 마이크로어레이 데이터 셋에서 98.61%의 향상된 분류 성능을 보였다.

ABSTRACT

The method of cancer classification based on microarray could contribute to being accurate cancer classification by finding differently expressing gene pattern statistically according to a cancer type. Therefore, the process to select a closely related informative gene with a particular cancer classification to classify cancer using present microarray technology with effect is essential. In this paper, the system can detect marker genes to likely express the most differentially explaining the effects of cancer using ovarian cancer microarray data. And it compare and analyze a performance of classification of the proposed system with it of established microarray system using multi-perceptron neural network layer. Microarray data set including marker gene that are selected using ANOVA method represent the highest classification accuracy of 98.61%, which show that it improve classification performance than established microarray system.

키워드

마이크로어레이, t-테스트, 아노바, 다층 퍼셉트론

Key word

Microarray, t-test, ANOVA, MLP(Multi-Layer Perceptron)

* 조선대학교 컴퓨터통계학과

** 조선대학교 (교신저자, cyjung@chosun.ac.kr)

접수일자 : 2009. 05. 20

심사완료일자 : 2010. 08. 12

I. 서 론

마이크로어레이 데이터는 실제 표본의 개수에 비해 유전자의 개수가 훨씬 많다는 특성을 가지고 있다. 따라서 현재의 마이크로어레이 기술을 이용해서 효과적으로 암을 정확하게 분류하기 위해서는 특정 암과 밀접하게 관련이 있는 정보력 있는 유용한 유전자(informative gene)를 선택하고 이 유전자들을 이용하여 보다 정확한 암 분류 모델을 구축하는 것이 매우 중요하게 부각되고 있다[1].

본 논문에서는 난소 암 마이크로어레이 데이터에서 특별한 표지로서 행동할 수 있는 표지 유전자 선택과 분류를 위한 시스템을 제안하였다. 제안된 시스템에서, t-test는 암과 밀접하게 관련된 표적 유전자를 발견하기 위해 처음으로 적용되었고, 표적 유전자로부터 가장 다르게 발현 할 가능성이 있는 표지 유전자를 발견하기 위해 ANOVA 방법은 실시되었다. MLP 방법은 선택된 표지 유전자의 암 조직을 검증하는데 적용되었고, 기존의 가공하지 않은 데이터 셋과 표적 유전자 데이터 셋을 사용하는 마이크로어레이 시스템과 분류 정확도 비교를 위해 사용되었다.

본 논문의 2장에서는 관련연구를 소개하고, 3장에서 시스템을 제안하고 시스템을 위해 필요한 중요한 방법들을 설명한다. 4장에서는 3장에서 제안한 시스템이 실제로 어떤 성능을 보이는지에 대해 실험한 과정과 그 결과를 제시하고 이를 분석한다. 5장에서는 결론을 도출한다.

II. 관련연구

2.1 마이크로어레이(Microarray)

유전자는 DNA의 일부분으로서, 최종산물인 단백질 생성에 필요한 정보를 담고 있다. 유전자가 mRNA 형태로 나타나는 현상을 유전자 발현(gene expression)이라 한다.

기존의 분자 생물학적 방법과 같이 한 번에 유전자 하나에 대한 접근을 시도하고 이로부터 획득한 데이터들을 수집해서 전체적인 의미를 도출하는 형태의 방법은 대량의 데이터를 포함하는 종합적인 분석에 적합하

지 않다. 분자 생물학과 공학 기술을 결합하여 고안된 마이크로어레이 기술은 동시에 대량의 유전자에 대해 특정 조건에 따른 발현 정보를 관찰할 수 있게 해 줌으로써 이러한 분야에서의 효과적인 분석을 가능하게 하고 있다[2].

그림 1. 마이크로어레이 데이터 생성과정
Fig. 1. a creation process of microarray data

2.1 암 분류를 위해 요구되는 최소 시스템 개요

마이크로어레이 데이터를 이용하여 암 분류를 하기 위해 요구되는 최소한의 시스템 개요는 그림 2와 같다.

그림 2. 요구되는 최소 시스템의 개요
Fig. 2. summary of the required minimum system

우선 마이크로어레이를 이용하여 해당 종양 샘플에 대해 특정 조건에 따른 유전자 발현 데이터를 확보한다. 이 마이크로어레이 데이터를 기반으로 클래스(유전기능 분류) 발견 작업을 수행한다. 클래스 발견 작업이 완료되면 각 클래스로 분류를 하는 데에 있어 극단적으로 큰 연관성을 갖는 가상의 유전자를 이상 유전자 모델(ideal set)로 결정한다. 이 이상 유전자 모델의 정보를 기준으로 분류를 하는 데에 큰 관련성을 갖는 유전자를 선택하는 작업을 한다. 이렇게 선택된 유전자 리스트를 기반으로 분류기 혹은 예측기를 구현한다.

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.

그림 3. 제안하는 분류 시스템
Fig. 3. proposing classification system

MLP 방법은 선택된 표지 유전자의 암 조직을 검증하는데 적용되었고 기존의 마이크로어레이 시스템과 분류 정확도 비교를 위해 사용되어졌다. 제안된 시스템 구성도는 그림 3과 같다.

3.2 t-test

유전자 i 의 t-score(TS)는 다음처럼 정의된다[3].

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

여기에서

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

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

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

K 개의 클래스가 있고, $\max\{y_k, k = 1, 2, \dots, K\}$ 는 모두 $y_k, k = 1, 2, \dots, K$ 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 i 의 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-score는 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.

3.3 Analysis Of Variance(ANOVA)

ANOVA는 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 건본으로 조사하는 반면, ANOVA에서 각 요소는 제한된 수의 전혀 다른 값(수준)에서 취해진다.

ANOVA 결과는 시뮬레이션 디자인이 잘 안정된다면 특히 직교한다면 해석하기가 더 쉽고 완전히 같은 반복인수를 갖는 ANOVA 모델 디자인은 훌륭한 통계 특징을 갖지만 시뮬레이션의 수는 빨리 증가한다. 왜냐하면 p 개의 수준을 갖는 n 개 요소의 완전한 디자인은 p^n 개의 시뮬레이션 실행을 요청하기 때문이다[4].

본 논문에서 ANOVA는 표지 유전자를 선택하기 위해 수행되어졌다.

3.4 Multi-level 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 백프로파게이션(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[5].

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 마이크로어레이 데이터를 사용하였다. 그림 4는 실험에서 획득한 가공하지 않은 마이크로어레이 데이터 산점도이다.

그림 4. 원자료의 산점도

Fig. 4. plot of raw data

그림 5는 통계 프로그램 R을 사용하여 t-테스트 결과 획득한 유의수준 0.05%에 속하는 3795개의 표적 유전자 산점도이다.

그림 5. 표적 유전자의 산점도

Fig. 5. plot of target gene

선택된 표적 유전자들은 표지 유전자 선택을 위해 ANOVA 방법에 사용되어졌다. ANOVA 방법을 적용한 결과 GeneNumber 'HG2538-HT2636, D87448, HG3914-HT4186, D49400, D78129, D83669, HG2247-HT2333, D14689, D38462, HG3264-HT3444, AD000092, D89859, HG3342-HT3523' 등 14개의 표지 유전자들이 추출되었다.

4.2 분석 결과

본 논문의 목적은 표지 유전자의 분류 정확도를 측정하기 위해 가공하지 않은 마이크로어레이 데이터 셋과 표적 유전자가 선택된 마이크로어레이 데이터 셋을 사

용하는 기존의 마이크로어레이 시스템과 표지 유전자가 선택된 마이크로어레이 데이터 셋을 사용하는 제안된 시스템에 각각 MLP를 적용하여 분류 정확도와 예측 에러 차이를 조사하는 것이다.

기계 학습 툴 WEKA를 이용하여 각각의 데이터 셋의 분류 정확도와 예측 에러 차이를 평가하기 위해 MLP 신경망을 구현하고 모델은 0.09로, 총 레이어 수는 3으로 고정된 후, 학습률을 0.01에서 0.05로 변화시켜가며 실험하여 분류 정확도와 예측 에러를 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 제공한 결과를 나타내는 MSE(the Mean Squared Error)와 치우침을 나타내는 bias로 수행되었고, 이 값이 작을수록 좋은 분류를 나타낸다. MSE와 bias는 다음 식 (1)와 (2)와 같다.

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

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

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

가공하지 않은 마이크로어레이 데이터 셋에는 20280 개의 유전자가 사용되었고, 표적 유전자가 선택된 마이크로어레이 데이터 셋에는 3795개의 표적 유전자가 사용되었고, 표지 유전자가 선택된 마이크로어레이 데이터 셋에는 14개의 유전자가 사용되었다.

마이크로어레이 데이터 셋에 MLP 신경망으로 측정 한 분류 정확도와 예측 에러 추정치 결과는 표 1과 같다. 가공하지 않은 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

DataSet	Accuracy	MSE	Bias
raw Microarray Data	60.70	0.47	0.43
Microarray Data Selected Target gene	91.43	0.26	0.13
Microarray Data Selected Marker gene	98.61	0.12	0.02

98.61%의 가장 높은 정확도와 0.12%, 0.02%의 가장 낮은 MSE와 bias는 표지 유전자가 선택된 마이크로어레이 데이터 셋에서 나타났다. 반면, 표지 유전자를 선택하지 않고 가공하지 않은 데이터 셋을 사용한 기존의 마이크로어레이 시스템에서는 60.70%의 가장 낮은 정확도와 0.47%, 0.43%의 가장 높은 MSE와 bias를 보였으며, 표적 유전자가 선택된 마이크로어레이 데이터 셋에서는 가공하지 않은 마이크로어레이 데이터 셋 보다 높은 정확도와 낮은 MSE와 bias를 보였다.

V. 결 론

암 연구에 있어, 민감하고 특별한 표지 유전자를 발견한다는 것은 어려운 일이다. 본 논문에서는 MLP를 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 셋의 분류 정확도와 예측 에러를 비교 평가하는 시스템을 고안하고 각 데이터 셋의 결과를 비교 분석하였다.

ANOVA를 이용하여 선택된 표지 유전자를 포함하는 마이크로어레이 데이터 셋에서 98.61%의 가장 높은 분류 정확도와 0.12% MSE와 0.02% bias의 가장 낮은 예측 에러 추정치를 보였다.

제안한 시스템은 ANOVA 방법에 의해 표지 유전자가 선택될 수 있다는 것을 증명하였고, 선택된 표지 유전

자는 난소 종양(OVT)과 난소암(OVCAI)을 가장 잘 분류하였다.

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, 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] Dov Stekel, Microarray Bioinformatics, Cambridge University Press, 2003.
- [3] J.Devore, and R. Peck, Statistic: the Exploration and Analysis of Data, 3rd ed. Pacific Grove, CA.:Duxbury Press, 1977.
- [4] Kobilinsky, A., 1997. Les plans factoriels. In: Dreesbeke, J.-J., Fine, J., Saporta, G. (Eds.), Plans d'Experiences, Applications a l'Entreprise. Editions Technip, Paris, pp. 69-209.
- [5] Martin T. Hagan, howard B. Demuth, and Mark Beale, "Neural network design", PWS Publishing Company, 1996.
- [6] Steven Huang G, Hung Y-C, Chen A, Hong M-Y. Microarray analysis of ovarian cancer. In: IEEE international conference on systems, man and cybernetics. New York: IEEE Systems, Man and Cybernetics Society; 2005. p. 1036-42

저자소개

박수영(Su-Young Park)

2001년 조선대학교
컴퓨터통계학과 이학사
2003년 조선대학교
컴퓨터통계학과 이학석사

2007년 조선대학교 컴퓨터통계학과 이학박사
※관심분야 : 신경망, 인공지능, 정보보호, 멀티미디어, 멀티미디어 콘텐츠, Bioinformatics

정채영(Chai-Yeoung Jung)

1987년 조선대학교 컴퓨터공학과
공학석사
1989년 조선대학교 컴퓨터공학과
공학박사

1986년~현재 조선대학교 컴퓨터통계학과 교수
※관심분야 : 신경망, 인공지능, 정보보호, 멀티미디어, 멀티미디어 콘텐츠, Bioinformatics