신경망 기반의 유전자 조합을 이용한 마이크로어레이 데이터 분류 시스템

박수영*, 정채영**

The System Of Microarray Data Classification Using Significant Gene Combination Method based on Neural Network.

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

이 논문은 2008년도 조선대학교 학술연구비를 지원받았음

요 약

최근 생명 정보학 기술의 발달로 마이크로 단위의 실험조작이 가능해짐에 따라 하나의 chip상에서 전체 genome의 expression pattern을 관찰할 수 있게 되었고, 동시에 수 만개의 유전자들 간의 상호작용도 연구가능하게 되었다.

본 논문에서는 암에 걸린 흰쥐 외피 기간 세포 분화 실험에서 얻어진 3840 유전자의 마이크로어레이 cDNA를 이용해 데이터의 정규화를 거쳐 본 논문에서 제안한 유사성 척도 조합 방법으로 정보력 있는 유전자들을 추출한 후, 유사성 척도 조합 방법과 결합한 멀티퍼셉트론 신경망 분류기와 기존의 DT, NB, SVM 분류기를 이용하여 클래스 분류 시스템을 구축하고, 성능을 비교분석하였다. 피어슨 적률 상관 계수와 유클리디안 거리 계수 조합을 이용하여 선택된 200 유전자들을 멀티퍼셉트론 신경망 분류기로 분류한 결과 98.84%의 정확도를 보여 다른 분류기를 이용하여 실험을 수행한 경우보다 향상된 분류 성능을 보였다.

ABSTRACT

As development in technology of bioinformatics recently makes it possible to operate micro-level experiments, we can observe the expression pattern of total genome through on chip and analyze the interactions of thousands of genes at the same time.

In this thesis, we used cDNA microarrays of 3840 genes obtained from neuronal differentiation experiment of cortical stem cells on white mouse with cancer. It analyzed and compared performance of each of the experiment result using existing DT, NB, SVM and multi-perceptron neural network classifier combined the similar scale combination method after constructing class classification model by extracting significant gene list with a similar scale combination method proposed in this paper through normalization. Result classifying in Multi-Perceptron neural network classifier for selected 200 genes using combination of PC(Pearson correlation coefficient) and ED(Euclidean distance coefficient) represented the accuracy of 98.84%, which show that it improve classification performance than case to experiment using other classifier.

키워드

microarray, significant gene list, PC-ED, MLP(multi- Layer perceptiron)

* 조선대학교 컴퓨터통계학과
** 교신저자

접수일자 2008. 05. 16

Ⅰ. 서론

DNA 마이크로어레이(microarray 또는 microchip)는 하나의 칩(chip)상에서 전체 유전체(genome)의 발현양상을 탐색할 수 있고, 동시에 수천 개의 유전자들 간의 상호작용도 관찰할 수 있다. 따라서, 수많은 유전자들로부터 실제 종양들의 세부 부류에 따라 확연하게 발현량이 변하는 표본 분류에 유용한 유전자들을 추출하기 위한 특징 추출(feature selection)방법과 이 유전자들을 이용하여 보다 정확한 종양 분류 모델(tumor classification)을 구축하는 것이 매우 중요하게 부각되고 있다[1][2].

본 논문의 구성은 다음과 같다. 2장에서 마이크로어레이에 대해 먼저 소개한다. 3장에서 본 논문이 수행한 시스템 설계 및 구현과정을 소개하고, 4장에서는 3장에서 제안한 분류 시스템을 사용한 모의실험에 대한 결과를 기술하고, 이를 분석한다. 5장에서는 결론을 도출한다.

Ⅱ. 관련 연구

2.1 마이크로어레이(Microarray)

생명체의 생명 현상을 조직하는 것은 세포 내에 존재하는 DNA(DeoxyriboNucleic acid)라는 물질이다. 유전자는 DNA의 일부분으로서, 최종산물인 단백질 생성에 필요한 정보를 담고 있다. 유전자가 mRNA 형태로 나타나는 현상을 유전자 발현(gene expression)이라 한다. [3].

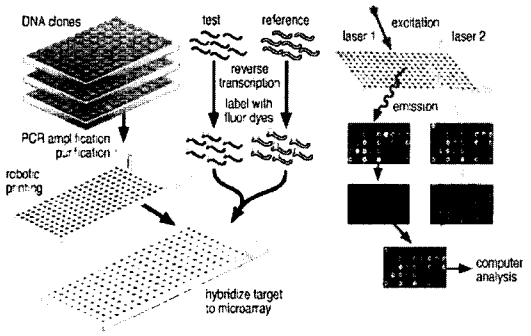


그림 1. 마이크로어레이 데이터 생성과정
Fig. 1. generation process of microarray data

2.2 마이크로어레이 기술을 이용한 암 분류 시스템 연구

마이크로어레이를 이용한 암 분류에서 최근의 전산학적 접근으로는 다중 분류기 시스템의 활용이 대표적이다. 암 분류 문제에 있어서 좀 더 높은 분류 성능을 확보하고자 기계 학습 기반 다중 분류기 시스템을 암 분류에 이용하는 사례가 많다. 그림 2는 암 분류를 위한 다중 분류기 시스템의 구조도를 나타내고 있다[4].

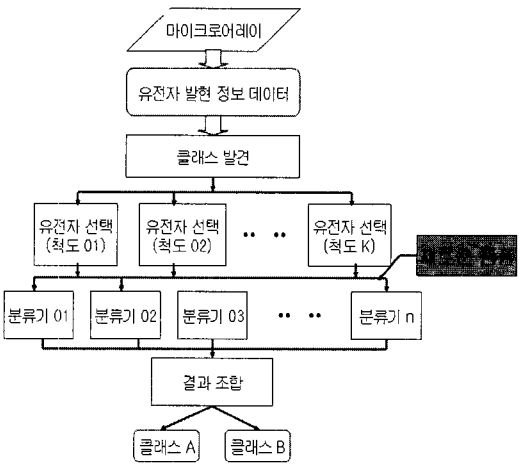


그림 2. 암 분류를 위한 다중 분류기 시스템의 구조
Fig. 2. system construction of multi-classification for cancer classification

Ⅲ. 유의한 유전자 선택 방법

1. 유의한 유전자 선택

종양 분류를 위해 마이크로어레이는 직접 종양 샘플에서부터 마이크로어레이 기술에 의해 데이터가 생성되기 때문에 종양의 특정 클래스에 연관이 큰 유전자는 그 수가 매우 적다. 따라서 분류기를 이용하여 현실적으로 효과적인 학습을 하기 위해서는 해당 클래스와의 연관성이 높은 유전자들을 시스템의 전단부인 전처리 과정에서 선택해야만 한다.

각 클래스에 대한 특징을 극단적으로 뚜렷하게 나타내면서 이상적으로 발현하는 유전자를 Gideal이라고 하면, 종양 세포의 특징을 1로 정의하고 나머지 정상세포 혹은 다른 종양 세포의 특징을 0으로 정의하여 식 (1)과

같은 벡터로 표현할 수 있다. Gideal은 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[5]. 유전자 선택을 위해 사용되는 유사성 척도는 그림 3과 같다.

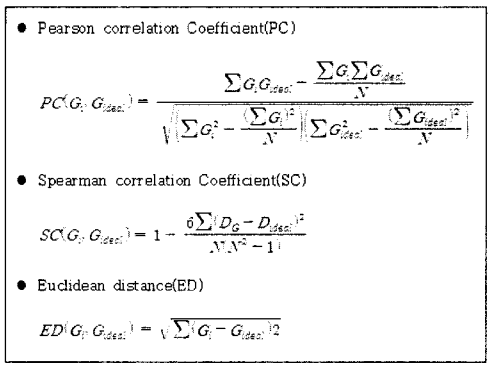


그림 3. 유전자 선택을 위한 유사성 척도
Fig. 3. the similar scale for gene selection

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 4는 본 논문에서 제안한 다수의 척도에서 정보력 있는 유전자로 평가받은 의미 있는 유전자들을 선택하는 알고리즘이다.

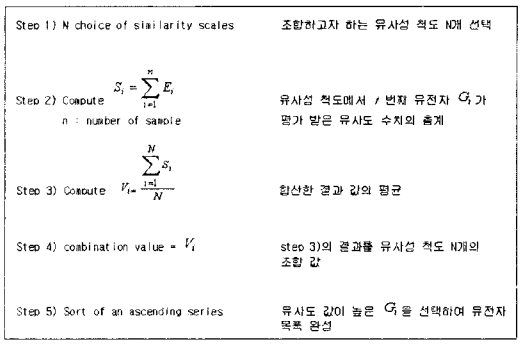


그림 4. 정보력이 있는 유전자 선택을 위한 조합 알고리즘
Fig. 4. the combination algorithm for gene being informative

3. 분류기법

3.1 Decision Tree(DT)

의사 결정 트리는 수집된 데이터의 레코드들을 분석하여 이들 사이에 존재하는 패턴, 즉 부류별 특성을 속성의 조합으로 나타내는 분류 모형을 나타내는 것이다. 그리고 이렇게 만들어진 분류 모형은 새로운 레코드를 분류하고 해당 부류의 값을 예측하는데 사용된다. 이 기법은 분류나 예측의 근거를 알려주기 때문에 이해하기가 쉽고 변수 간 상관관계와 영향을 알 수가 있어 데이터 선정이 용이하며 구조가 단순하여 모형 구축에 소요되는 시간이 짧다는 장점이 있다.

3.2 Naive Bayes(NB)

Naive Bayes는 베이지안 확률 모형에 기초한다. 이는 임의의 데이터가 특정 분류에 속할 확률을 계산하여 계산된 확률 중 가장 높은 확률을 가지는 분류를 선택하는 것을 의미한다. 가끔 우리는 어떤 실험결과에서 나온 정보를 이용하여 어떤 사건의 처음 확률을 개선시킬 수 있는데, 여기서 처음 확률은 사전확률(prior probability)이라 하고, 개선된 확률을 사후확률(posterior probability)이라고 하며, 이러한 확률의 개선을 이루는 것이 베이즈의 정리(Bayes' theorem)이다.

3.3 Support Vector Machine(SVM)

SVM은 분류(classification)와 회귀(regression)에 응용할 수 있는 지도학습(Supervised learning)이 일종으로서 기본적인 분류를 위한 SVM은 입력 공간에 maximum-

margin hyperplane을 만든다. 학습데이터와 범주 정보의 학습 진단을 대상으로 학습과정에서 얻어진 확률분포를 이용하여 의사결정함수를 추정한 후 이 함수에 따라 새로운 데이터를 이원 분류하는 것으로 VC(Vap-nik Chervonenkis) 이론이라고도 한다. 특히, SVM은 분류 문제에 있어서 일반화 기능이 높기 때문에 많은 분야에서 응용되고 있다.

3.4 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[6].

Ⅳ. 실험 및 결과 고찰

1. 제안하는 시스템 구조도

제안하고자 하는 효과적인 유전자 선택 방법의 현실적 구현을 위해서는 기존의 암 분류를 위한 유전자 발현 분석 시스템의 구조를 변경해야 한다. 그림 5는 이러한 시스템의 구조도를 나타낸 것이다.

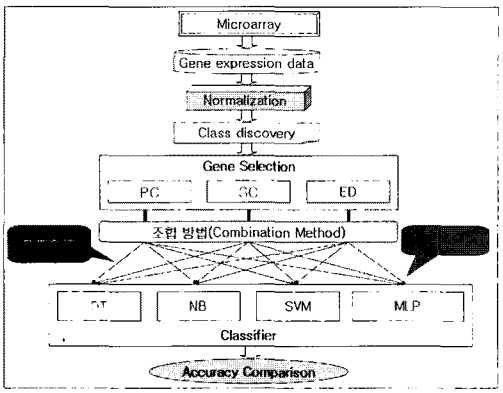
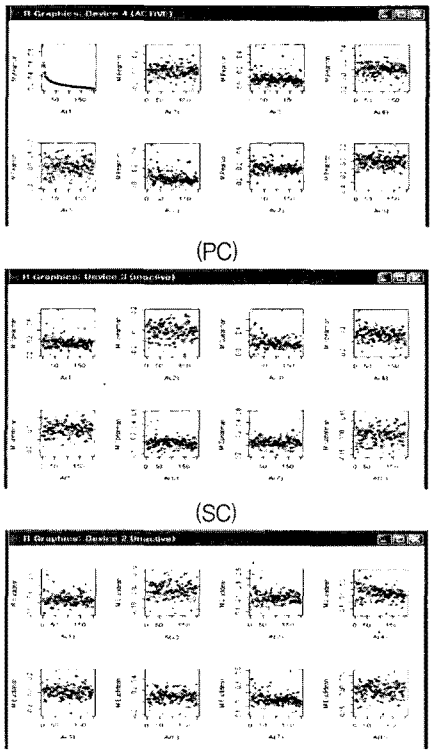


그림 5. 제안하는 분류 시스템
Fig. 5. proposing classification system

2. 실험 결과 및 고찰

본 논문에서는 암에 걸린 흰쥐와 암에 걸리지 않은 흰쥐의 각 뇌신경조직 부위에서 획득한 유전체의 조정 인자를 각각 Cy5, Cy3로 염색한 다음, 2400개 이상의 알려진 유전체와 1700여개의 새로운 유전체가 찍힌 유리칩을 이용한 cDNA 마이크로어레이 실험에서 획득한 마이크로어레이 데이터를 사용하였다. 통계 컴퓨터 프로그램인 R을 이용하여 각 유전자의 발현 정도를 [0, 1] 범위로 정규화 하였고 기존의 단일 유사성 척도 3가지를 사용한 유전자 선택 방법과 이들을 조합한 유전자 선택 방법 4가지를 이용하여 정보력이 있는 유전자를 선택하고 목록을 만들었다. 이 유전자 목록을 이용하여 멀티퍼셉트론 신경망과 비교 연구된 분류알고리즘인 DT, NB, SVM, 분류기를 통해 학습과 테스트를 한 분류 결과를 10-fold 교차검증을 사용하여 정확도를 서로 비교 분석하였다. 그림 6은 정규화 후 각 유사성 척도에 따라 선택된 상위 200개 유전자 산점도의 일부분이다.



(PC)

(SC)

(ED)

그림 6. 상위 200개 유전자 산점도
Fig. 6. plot of 200 high rank gene

'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.

MSE(Mean Square Error)는 평균제곱 오차를 나타내며, 실제 클래스와 예측한 클래스 차이를 제곱한 결과를 나타내며 이 값이 작을수록 좋은 분류를 나타낸다. 그림 7는 데이터 마이닝툴 WEKA를 이용한 마이크로어레이 분류 시스템을 설계한 그림이다.

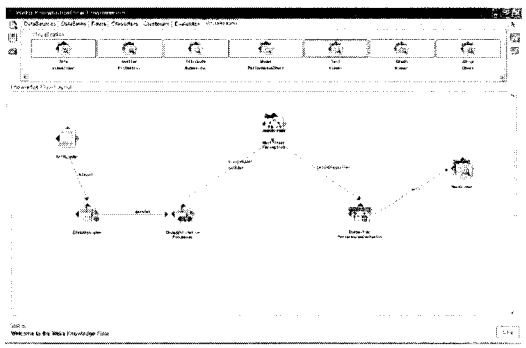


그림 7. 마이크로어레이 분류 시스템
Fig. 7. microarray classification system

3. 분석 결과

기존의 단일 유사성 척도를 이용하여 상위 200개 유전자를 선택하고 목록을 만들어 분류 성능을 실험한 결과는 표 1와 같다. 이를 두 개 이상의 기존 척도 조합에 따른 분류 성능 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
MLP	92.6	91.5	90.6
MSE	0.07	0.08	0.08
DT	89.3	89.3	88.2
MSE	0.12	0.10	0.11
NB	89.3	87.2	89.2
MSE	0.13	0.12	0.10
SVM	89.5	89.2	89.2
MSE	0.12	0.12	0.12

기존의 단일 유사성 척도를 사용하여 유전자 목록을 생성한 뒤 실험 한 경우 대부분 낮은 분류성능을 나타내었다.

그러나 이러한 기존의 단일 유사성 척도를 조합 방법에 의해 조합하여 보다 정보력이 있는 유전자 목록을 생성한 뒤 실험한 경우 분류기에서 향상된 분류 성능을 나타내었다. 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
MLP	98.84	98.29	95.14
MSE	0.02	0.04	0.04
DT	95.4	95.2	96.2
MSE	0.05	0.06	0.06
NB	96.5	93.8	94.2
MSE	0.06	0.08	0.08
SVM	97.2	94.2	95.2
MSE	0.04	0.05	0.04

단일 유사성 척도를 사용하여 실험한 결과보다 대부분 높은 분류 성능을 나타냈으며 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
MLP	97.33
MSE	0.05
DT	92.42
MSE	0.12
NB	91.12
MSE	0.23
SVM	96.14
MSE	0.12

기존의 유사성 척도를 사용하여 유전자를 선택하는 세 가지를 모두 조합한 경우 표 3과 같은 분류 성능 향상을 보였다. 그러나 이러한 분류 성능 향상은 두 가지 척도를 조합하는 경우에 가장 현저하게 나타나고 세 가지

이상의 척도를 조합하는 경우에는 다소 소극적으로 나타났다. 이는 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

본 논문에서는 정규화 후 정보력이 있는 유전자 목록을 조합하는 시스템을 고안하고 보다 분류 성능을 향상시킬 수 있는 조합 방법을 제안하고, 여러 분류기들을 이용하여 실험 평가 하였다. 그 결과 제안한 PC-ED조합으로 추출한 데이터를 멀티 퍼셉트론 신경망 분류기로 분류한 결과 98.84%의 정확도와 0.02%의 MSE를 보여 단일 유사성 척도를 사용하여 유전자 목록을 생성하고 실험을 수행한 경우보다. 본 논문에서 제안한 조합 방법으로 추출한 데이터를 멀티 퍼셉트론 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.

[5] Evertsz, E., Starink, P., Gupta, R., and Watson, D., "Technology and application of gene expression microarraysaa", Schena, M.(ed.), Microarray Biochip Technology, Eaton Publishing, MA, pp. 149-166, 2000.

[6] Martin T. Hagan, Howard B. Demuth, and Mark Beale, "Neural network design", PWS Publishing Company, 1996.

저자소개

박수영(Su-Young Park)

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

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