정보력 있는 유전자 선택 방법 조합을 이용한 마이크로어레이 분류 시스템 구현

박수영* · 정채영**

The Implement of System on Microarry Classification Using Combination of Signigicant Gene Selection Method

Su-Young Park* · Chai-Yeoung Jung**

요 약

오늘날 인간 genome 프로젝트와 같은 종합적인 연구의 궁극적 목적을 달성하기 위해서는 이들 연구로부터 획득한 대량의 관련 데이터에 대해 새로운 현실적 의미를 부여할 수 있어야 한다. 이러한 맥락에서 유전자 발현 분석 시스템과 염기 서열 분석 시스템의 구축이 포스트 genome 시대를 맞이하여 새롭게 주목을 받고 있다. 최근에는 종양의 특정 부 클래스가 특정 염색체와 관련되어 있다는 사실이 밝혀지면서, 마이크로어레이는 유전자 발현 정보를 기반으로 암의 분류와 예측을 통한 진단 분야에도 활용되기 시작했다.

본 논문에서는 암에 걸린 흰쥐 외피 기간 세포 분화 실험에서 얻어진 3840 유전자의 마이크로어레이 cDNA를 이용하여 데이터의 정규화를 거쳐 정보력 있는 유전자 목록을 별도로 추출할 수 있는 시스템을 고안하고 보다 정보력 있는 유전자를 선택하기 위해 조합 방법을 제안하였다. 그리고 제안한 시스템과 방법론의 가능성을 실험을 통해 검증하였다. 그 결과 PC-ED 조합이 98.74%의 정확도와 0.04%의 MSE를 보여 단일 유사성 척도를 사용하여 유전자 목록을 생성하고 실험을 수행한 경우보다 분류 성능이 향상되었다.

ABSTRACT

Nowadays, a lot of related data obtained from these research could be given a new present meaning to accomplish the original purpose of the whole research as a human genome project. In such a thread, construction of gene expression analysis system and a basis rank analysis system is being watched newly. Recently, being identified fact that partircular sub-class of tumor be related with particular chromosome, microarray started to be used in diagnosis field by doing cancer classification and predication based on gene expression information.

In this thesis, we used cDNA microarrays of 3840 genes obtained from neuronal differentiation experiment of cortical stem cells on white mouse with cancer, created system that can extract informative gene list through normalization separately and proposed combination method for selecting more significant genes. And possibility of proposed system and method is verified through experiment. That result is that PC-ED combination represent 98.74% accurate and 0.04% MSE, which show that it improve classification performance than case to experiment after generating gene list using single similarity scale.

키워드

microarray, PC-ED combination method, MLP(Multi-Perceptron)

* 조선대학교 컴퓨터통계학과

** 교신저자

접수일자 : 2007. 9. 7

I. 서 론

2000년 6월 인간 유전체지도의 완성으로, 인간유전체로부터 각 유전자들의 생체기능을 밝히고 개인, 생물간 유전체(Post-Genome) 시대가 열리게 되었다.

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 소개하고, 기존의 마이크로어레이 기술을 이용한 암 분류 시스템 대해 소개한다. 3장에서 유전자 선택 방법과 멀티퍼셉트론 대하여 설명하고, 분류 시스템을 제안한다. 4장에서는 3장에서 제안한 분류 시스템을 사용한 모의실험에 대한 결과를 기술하고, 이를 분석한다. 5장에서는 결론을 도출하고 향후 연구 과제에 대핸 논하도록 한다.

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. 마이크로어레이란 형태상으로 현미경 슬라이드 정도 크기의 유리판과 같이 투명하고 딱딱한 판 위에 수천 혹은 수만 개의 DNA 조각을 격자 모양으로 가지런히 배열해 놓은 분자 생물학의 도구로서 특정 조건에서 대량의 염기서열이 보이는 경향을 관찰하고 대량의 유전자 발현 데이터를 획득하고 처리한다[3].

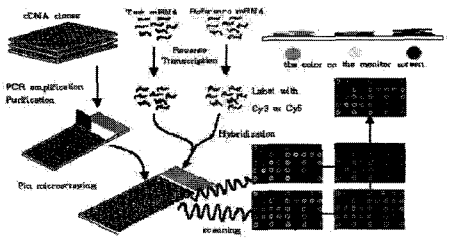


그림 1. 마이크로어레이 데이터 생성과정
Fig. 1. a creation process of microarry data

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].

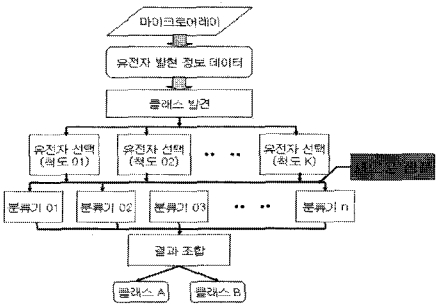


그림 2. 암 분류를 위한 다중 분류기 시스템의 구조
Fig. 2. system construction of multi-classification for cancer classification

위와 같은 형태의 다중 분류기 시스템을 이용하여 암 분류를 시도하게 되면 상대적으로 높은 분류 성능을 얻을 수 있다. 그러나 여러 분류기를 운용하는 데에 따르는 고비용과 느린 속도가 단점으로 지적된다. 또한 정보력 있는 유전자 리스트를 별도로 추출한다는 것 자체가 거의 불가능하다.

Ⅲ. 유의한 유전자 선택 방법

3.1. 유의한 유전자 선택

종양 분류를 위해 마이크로어레이는 분류기를 이용하여 현실적으로 효과적인 학습을 하기 위해 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 백터로 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.

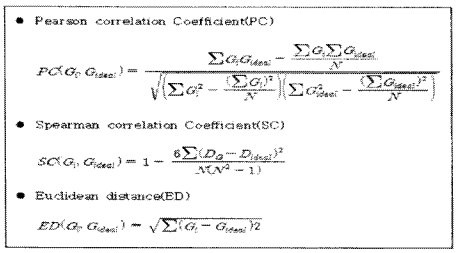


그림 3. 유전자 선택을 위한 유사성 척도
Fig. 3. the similarity measure for gene choice

3.2. 조합 방법

기존 방법과 같이 각 유사성 척도를 개별적으로 사용하여 유용한 유전자 목록을 만들게 되면, 정확한 암 분류에 있어서 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.

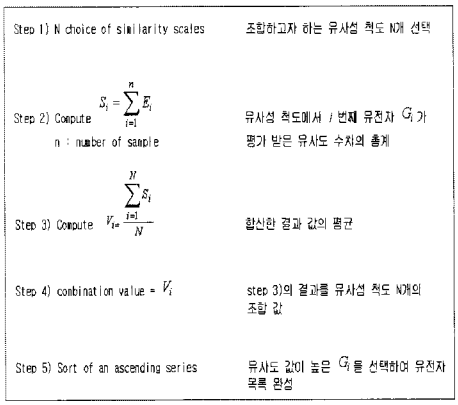


그림 4. 정보력이 있는 유전자 선택을 위한 조합 알고리즘
Fig. 4. the combination algorithm for significant gene selection

3.3. 다층퍼셉트론

본 논문에서는 제안하는 시스템의 분류기로 기계 학습 기반 분류기를 사용하여 이를 구현하였다.

인공 신경망의 대표적인 기계 학습 알고리즘인 다층 퍼셉트론은 대두분의 패턴 인식 문제에 대해 안정적인 성능을 보이며, 일단 학습이 끝나면 응용 단계에서는 매우 빠르게 결과를 출력한다. 다층 퍼셉트론은 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].

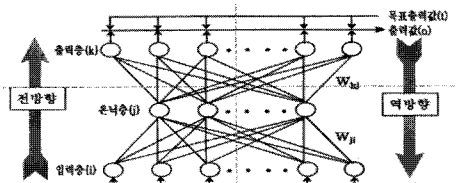


그림 5. 다층퍼셉트론의 구조
Fig. 5. structure of multi-perceptron

3.4. 제안하는 시스템 구조도

제안하고자 하는 효과적인 유전자 선택 방법의 현실적 구현을 위해서는 기존의 암 분류를 위한 유전자 발현 분석 시스템의 구조를 변경해야 한다. 그림 6은 이러한 시스템의 구조도를 나타낸 것이다.

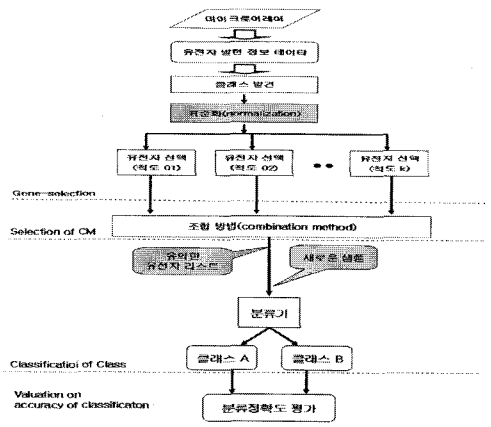


그림 6. 제안하는 분류 시스템 구조도
Fig. 6. the construction of proposing classification system

변경된 시스템에서 데이터의 흐름은 다음과 같다. 먼저 마이크로어레이로부터 유전자 발현 데이터를 획득한다. 정규화 과정을 거쳐 잡음을 제거한 후 클래스 발견 단계에서 이상 유전자 모델을 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)유용한 유전자로 평가 받은 유전자들을 정량화된 유용성 정도에 따라 서열화 하고 이들의 상위 부분을 모아 정보력 있는 유전자 목록으로 확정한다.

Ⅳ. 실험 및 결과 고찰

4.1. 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 만들어 다층퍼셉트론을 기반으로 하는 분류기를 통해 학습과 테스트를 한 분류 결과를 10-fold 교차검증을 사용하여 정확도를 서로 비교 분석하였다.

'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. 표에 나타나 있는 분류 성능 수치는 해당 조합 조건에서 분류기의 분류 성공률을 의미하며 그 단위는 퍼센트(%)이다.

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

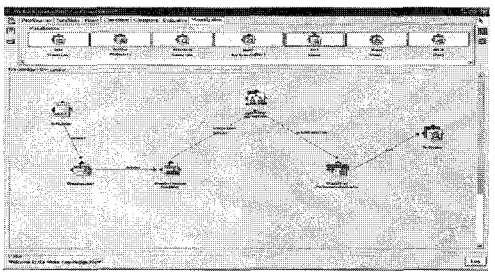
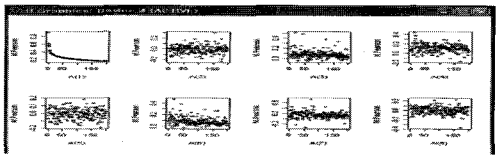
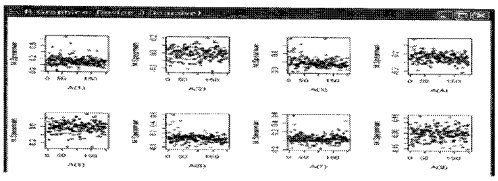


그림 7. 마이크로어레이 분류 시스템
Fig. 7. microarray classification system

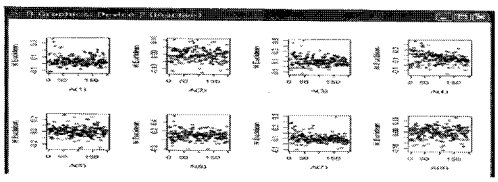
그림 8은 정규화 후 각 유사성 척도에 따라 선택된 상위 200개 유전자산점도의 일부분이다.



(PC)



(SC)



(ED)

그림 8. 상위 200개 유전자 산점도
Fig. 8. the plot of top 200 gene

4.2. 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 by single scale use

조 합	PC	ED	SC
분류 성능	92.14	91.35	91.24
MSE	0.16	0.18	0.2

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 by 2-scale combination

조 합	PC-ED	PC-SC	ED-SC
분류 성능	98.74	96.29	95.14
MSE	0.04	0.08	0.12

단일 유사성 척도를 사용하여 실험한 결과보다 대부분 높은 분류 성능을 나타냈으며 PC-ED의 경우 98.74%로 가장 높은 분류 성능을 보였다.

표 3. 3개 척도 조합에 따른 분류 성능
table 3. classification performance by 3-scale combination

조 합	PC-ED-SC
분류 성능	95.25
MSE	0.16

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. 결론 및 향후 연구과제

오늘날 인간 genome 프로젝트와 같은 종합적인 연구의 궁극적 목적을 달성하기 위해서는 이들 연구로부터 획득한 대량의 관련 데이터에 대해 새로운 현실적 의미를 부여할 수 있어야 한다.

최근 마이크로어레이는 유전자 발현 정보를 기반으로 암의 분류와 예측을 통한 진단 분야에도 활용되고 있다. 현재의 마이크로어레이 기술을 이용해서 효과적으로 암을 정확하게 분류하기 위해서는 특정 암의 분류와 밀접하게 관련이 있는 유용한 유전자를 선택하는 과정이 필수적이다.

이에 본 논문에서는 정규화 후 정보력이 있는 유전자 목록을 조합하는 시스템을 고안하고 보다 분류 성능을 향상 시킬 수 있는 조합 방법을 제안하였다. 그 결과 제안한 시스템과 방법론으로 PC-ED 조합이 98.74%의 정확도와 0.04%의 MSE를 보여 단일 유사성 척도를 사용하여 유전자 목록을 생성하고 실험을 수행한 경우보다 분류 성능이 향상되었다.

따라서 마이크로어레이 데이터를 이용한 암 분류에 있어서 암의 분류와 예측을 통한 진단 분야뿐만 아니라 정확한 분류 이후의 치료 분야에도 본 논문에서 제안한 시스템과 조합 방법이 효과적으로 적용될 수 있을 것으로 생각된다.

이에 아직 사용해보지 못한 또 다른 유사성 척도 방법과 기계 학습 알고리즘에 더 많은 연구를 진행하고자 한다.

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 Uiversity 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.

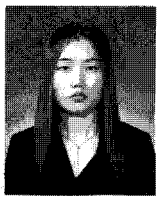
[6] Martin T. Hagan, Howard B. Demuth, and Mark Beale, "Neural network design", PWS Publishing Company, 1996.

[7] J. Hertx, A. Krogh, and R. G. palmer, "Introduction to The Theory of Nerualcomputation, Vol. 1, Addison-Wesley Publishing Co., 1991.

[8] Marco Gori and Alberto Tesi, "On the Problem of Local Minima in Backpropagation," IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 14, no. 1, pp. 318-362, 1986.

[9] Robert Gentleman, Vincent J. carey, Wolfgang Huber, Rafael, A Irizarry, Sandrin Dudoit, "Bioinformatics and Computational Biology Solutions Using R and Bioconductor," Sptringer.

저자소개



박 수 영(Su-Young Park)

2007년 조선대학교 컴퓨터 통계학과 박사

※관심분야 : 신경망, 인공지능, 정보보호, 멀티미디어, 멀티미디어 컨텐츠, Bioinformatics

정 채 영(Chai-Yeoung Jung)

1987년 조선대학교 컴퓨터공학과 공학석사

1989년 조선대학교 컴퓨터공학과 공학박사

1986년~현재 조선대학교 컴퓨터통계학과 교수

※관심분야 : 신경망, 인공지능, 정보보호, 멀티미디어, 멀티미디어 컨텐츠, Bioinformatics