종양 분류를 위한 마이크로어레이 데이터 분류 모델 설계와 구현

박수영* · 정채영**

The Design and Implement of Microarry Data Classification Model for Tumor Classification

Su-Young Park* · Chai-Yeoung Jung**

이 논문은 2007년도 조선대학교 학술연구비를 지원받아 연구되었음

요 약

오늘날 인간 프로젝트와 같은 종합적인 연구의 궁극적 목적을 달성하기 위해서는 이들 연구로부터 획득한 대량의 관련 데이터에 대해 새로운 현실적 의미를 부여할 수 있어야 한다. 마이크로어레이를 기반으로 하는 종양 분류 방법은 종양 종류에 따라 다르게 발현되는 유전자 양상을 통계적으로 발견함으로써 정확한 종양 분류에 기여할 수 있다. 따라서 현재의 마이크로어레이 기술을 이용해서 효과적으로 종양을 분류하기 위해서는 특정 종양 분류와 밀접하게 관련이 있는 정보력 있는 유전자를 선택하는 과정이 필수적이다.

본 논문에서는 암에 걸린 흰쥐 외피 기간 세포 분화 실험에서 얻어진 3840 유전자의 마이크로어레이 cDNA를 이용해 데이터의 정규화를 거쳐 정보력 있는 유전자 목록을 별도로 추출하여 보다 정확한 종양 분류 모델을 구축하고 각각의 실험 결과들을 비교 분석함으로써 성능평가를 하였다. 피어슨 적률 상관 계수를 이용하여 선택된 유전자들을 멀티퍼셉트론 분류기로 분류한 결과 98.6%의 정확도를 보였다.

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 project. The method of tumor classification based on microarray could contribute to being accurate tumor classification by finding differently expressing gene pattern statistically according to a tumor type. Therefore, the process to select a closely related informative gene with a particular tumor classification to classify tumor using present microarray technology with effect is essential.

In this thesis, we used cDNA microarrays of 3840 genes obtained from neuronal differentiation experiment of cortical stem cells on white mouse with cancer, constructed accurate tumor classification model by extracting informative gene list through normalization separately and then did performance estimation by analyzing and comparing each of the experiment results. Result classifying Multi-Perceptron classifier for selected genes using Pearson correlation coefficient represented the accuracy of 95.6%.

키워드

microarray, PC(Pearson correlation coefficient), MLP(Multi-Perceptron)

* 조선대학교 컴퓨터통계학과

** 조선대학교 컴퓨터통계학과(교신저자)

접수일자 : 2007. 5. 14

Ⅰ. 서 론

생물정보학(bioinformatic)은 컴퓨터과학과 생물학의 경계에 있는 학제적인 연구 분야로서, 특히 최근에는 이러한 종합적인 연구를 통해 분자 생물학 데이터들이 대량으로 산출됨에 따라 그 중요성이 더욱 커지고 있다. 인간 지놈 프로젝트와 같은 종합적인 연구의 궁극적 목적을 달성하기 위해서는 이들 연구로부터 획득한 대량의 데이터에 대해 새로운 현실적 의미를 부여할 수 있어야 한다. 이러한 맥락에서 유전자 발현 분석 시스템과 염기서열 분석 시스템의 구축은 포스트 지놈(post-genome)시대를 맞이하여 새롭게 주목을 받고 있다. 이러한 응용의 현실적 시도를 가능하게 했던 것은 마이크로어레이(microarray) 기술의 발전이다.

마이크로어레이 데이터는 실제 표본의 개수에 비해 유전자의 개수가 훨씬 많다는 특성을 가지고 있다. 따라서 현재의 마이크로어레이 기술을 이용해서 효과적으로 종양을 정확하게 분류하기 위해서는 특정 종양의 분류와 밀접하게 관련이 있는 정보력 있는 유용한 유전자(informative gene)를 선택하고 이 유전자들을 이용하여 보다 정확한 종양 분류 모델을 구축하는 것이 매우 중요하게 부각되고 있다[1][2].

본 논문의 구성은 다음과 같다. 2장은 관련연구를 소개하고 3장에서는 본 논문이 수행한 시스템 설계 및 구현과정을 소개한다. 그리고 4장은 실험과정을 설명하고 결과를 비교분석 할 것이다. 마지막으로 5장에서는 결론을 내리고 향후 연구방향을 제시한다.

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. 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 적합하지 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

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 도구이다[3].

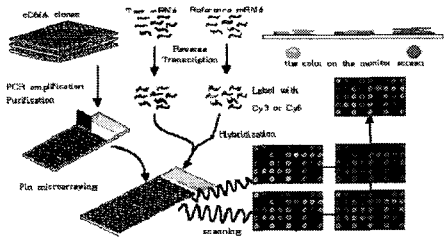


그림 1. 마이크로어레이 데이터 생성과정
Fig. 1. a creation process of microarry data

2.2 마이크로어레이 기반 암 분류 연구 사례

Golub은 지난 30년간 진행되어 온 기존의 암 분류화 작업을 전체적이고 시스템적이며 일반적인 접근으로 시도하려 하였다. 당시에는 새로운 클래스를 정의하고 알려진 클래스에 종양(tumor)을 할당하기 위한 일반적인 접근이 없었다.

Golub은 DNA 마이크로어레이에 의한 유전자 발현 정보를 관찰하는 것을 기반으로 암 분류에 대한 일반적인 접근을 시도하는 방법을 급성 백혈병에 적용하여 그 가능성을 타진하였다[4].

Ⅲ. 제안하는 시스템

3.1 제안하는 시스템 구조도

본 논문에서 구성한 시스템은 마이크로어레이 데이터를 사용해 효과적인 유전자 선택 방법과 데이터마이닝 분류기법들을 이용하여 보다 정확한 종양 분류 모델(tumor classification model)을 구축하기 위해 기존의 종양 분류를 위한 유전자 발현 분석 시스템의 구조를 변경해야 한다. 변경된 시스템에서 데이터 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. 유용한 유전자로 평가 받은 유전자들을 정량화된 유용성 정도에 따라 서열화 하고 이들의 상위 부분을 모아 정보력 있는 유전자 목록으로 확정한다. 이 정보력 있는 유전자 목록을 여러 분류기로 분류하고 성능을 비교 분석하였다.

그림 2는 이러한 시스템의 구조도를 나타낸 것이다.

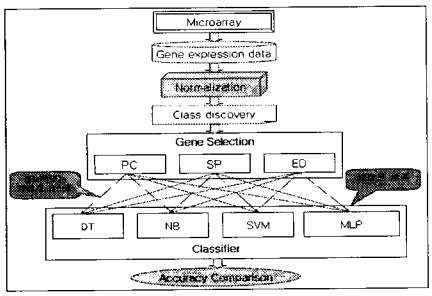


그림 2. 시스템 구성도
Fig. 2. a construction of system

3.2 유전자 선택을 위한 유사성 척도

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 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.

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

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.

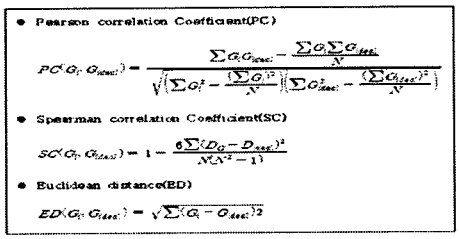


그림 3. 유전자 선택을 위한 유사성 척도
Fig. 3. the similarity measure for gene choice

3.3 분류 기법

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.

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) 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(Vapnik-Chervonenkis) 이론이라고도 한다. 특히, SVM은 분류 문제에 있어서 일반화 기능이 높기 때문에 많은 분야에서 응용되고 있다.

4) Multi-Layer Perceptron(다층퍼셉트론)

인공 신경망의 대표적인 기계 학습 알고리즘인 다층퍼셉트론은 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[5].

3.4 성능지표

성능 평가 기준(performance measures)은 얼마나 정확한 예측을 했는가를 평가하는 것이다. 분류 결과가 얼마나 정확했는지에 대한 평가 기준은 표 1에 정리되어 있는 실제 항목과 결과로써 나온 예측 항목을 비교한 비율에 기반 해서 계산된다.

표 1. 평가표
Table 1. Evaluation Table

		Actual Class	
		Positive	Negative
Predicted class	Positive	True Positive (TP)	False Positive (FP)
	Negative	False Negative (FN)	True Negative (TN)

Ⅳ. 실험 및 결과 고찰

암에 걸린 흰쥐 외피 기간 세포 분화 실험에서 얻어진 3840 유전자의 마이크로어레이 cDNA 데이터 분석을 위해 데이터의 정규화를 거쳐 PC, SP, ED 3가지의 유사성 척도를 이용하여 200개의 유전자를 선택하였다. 이 유전자를 DT, NB, SVM, MLP의 학습데이터로 하여 4가지 알고리즘의 성능을 비교분석하고 모델 구축 시간을 계산하여 최적의 조합을 찾는다. 또한 정확도에 대한 성능 지표를 계산한다.

4.1 실험 결과 및 고찰

본 논문에서는 R을 이용하여 각 유전자의 발현 정도를 [0, 1] 범위로 정규화 하였고 PC, SP, ED를 기반 해서 순위대로 나열한 후에, 유전자를 1위부터 50위, 90위, 200위까지 샘플링한 3개의 서브 데이터 셋을 만들었다. 이렇게 선택된 각 유사성 척도 별로 분별력 있는 유전자 개수에 따라 WEKA를 이용해 앞서 설명한 4가지의 기계학습 알고리즘으로 종양 클래스 분류 모델일 만들고 10-fold cross-validation을 사용하여 정확도를 측정하고 서로 비교분석 하였다.

여기서 말하는 cross-validation은 교차검증을 뜻하는 것이다. k-fold cross-validation은 k개의 부분샘플로 데이터집합을 나눈 후 k-1개의 부분샘플은 train(훈련)데이터로 하나의 부분샘플은 test(테스트)데이터로 사용하여 하나의 서브 데이터 셋을 돌아가며 테스트 데이터 셋으로 사용하여 여러 모델을 만들어 평균을 내는 방법이다. 그림 4는 raw data의 산점도와 정규화 후 data의 산점도이다.

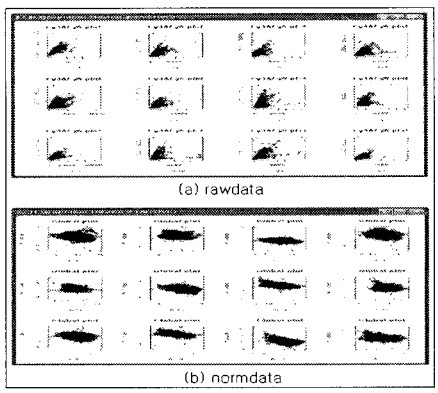
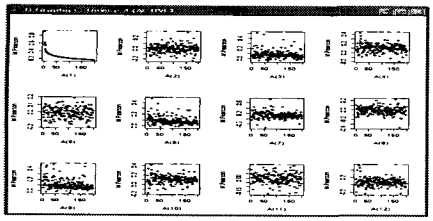
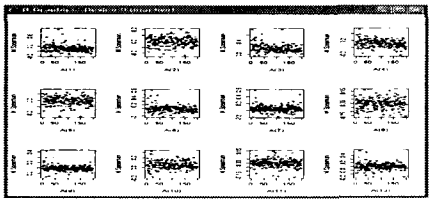


그림 4. 실험에 사용된 데이터 산점도
Fig. 4. the simulated data plot

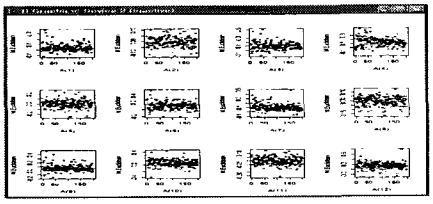
그림 5는 정규화 후 각 유사성 척도에 따른 유전자산점도일부분이다.



(a) PC



(a) SP



(a) ED

그림 5. 유사성 척도에 따른 유전자 산점도
Fig. 5. the gene plot according to similarity measure

그림 6은 WEKA를 이용하여 만든 종양 클래스 분류 모델의 일부분이다.

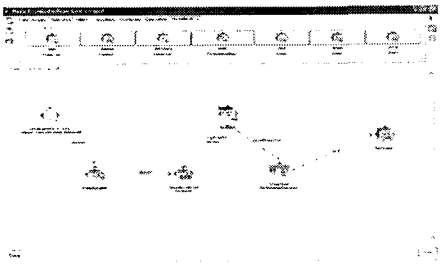


그림 4. 종양 분류 모델
Fig. 4. the model of tumor classification

4.2 분석 결과

전체 데이터 셋도 알고리즘에 적용하여 분석한 결과를 유사성 척도를 이용하여 유전자를 선택하였을 때의 성능 비교를 위한 대조군으로 사용하였다.

표 2. 검증 방법 결과 1
table 2. a result of verification method 1

	DT	NB	SVM	MLP
Total Data	93.6%	95.6%	93.1%	93.1%

표 2는 유전자를 선택하지 않고 전체 데이터 셋을 알고리즘에 적용한 결과이다. 대부분 성능이 높다. 또한 Naive Bayes와 Decision Tree이 다른 알고리즘 보다 성능이 좋다.

표 3. 검증 방법 결과 2
table 3. a result of verification method 2

gene=50				
	DT	NB	SVM	MLP
PC	93.1%	96.3%	97.6%	98.6%
SP	91.7%	95.2%	96.8%	97.2%
ED	93.1%	95.8%	97.2%	98.6%
gene=90				
	DT	NB	SVM	MLP
PC	91.2%	96.3%	97.6%	98.6%
SP	90.3%	95.2%	97.5%	98.6%
ED	90.3%	96.3%	97.5%	98.6%
gene=200				
	DT	NB	SVM	MLP
PC	90.3%	95.2%	97.5%	98.6%
SP	90.3%	95.2%	96.8%	98.6%
ED	90.3%	95.2%	96.8%	98.6%

표 2처럼 각각의 특징추출.방법에 대해 적용한 알고리즘의 성능이 대부분 좋게 나왔다. PC를 사용해 유전자를 추출한 방법이 다른 유사성 척도보다 높은 성능을 보였으며 PC-MLP 조합이 모든 경우에서 대부분 98.6%로 가장 높은 성능을 나타냈다.

또한 유사성 척도를 사용하여 정보력 있는 유전자를 추출하여 분석을 하는 경우가 전체 데이터 셋을 적용한 결과 보다 성능이 향상되었다.

V. 결론 및 향후 연구과제

본 논문에서는 암에 걸린 흰쥐 외피 기간 세포 분화 실험에서 얻어진 3840 유전자의 마이크로어레이 cDNA 데이터를 사용하여 유사성 척도 방법으로 정보력 있는 유전자들을 추출한 후, DT, NB, SVM, MLP 알고리즘을 이

용하여 클래스 분류 모델을 구축하고, 성능을 비교분석 하였다. 모든 경우의 성능을 비교 분석했을 때 PC-MLP 조합이 98.6%의 정확도를 보여 가장 최적의 조합을 보였다. 향후 연구 과제로는 다양하고 체계적인 많은 데이터의 획득과 분석을 통해 좀 더 효율적인 조합을 찾는 연구가 계속되어야 하고, 이와 더불어 이렇게 제안된 최적의 조합이 다른 종류의 종양을 대상으로 한 검증이 이루어져야 할 것이다.

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.

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 컴퓨터 통계학과 교수
※관심분야 : 신경망, 인공지능, 정보보호, 멀티미디어, 멀티미디어 컨텐츠, Bioinformatics