베이지안 기법을 적용한 마이크로어레이 데이터 분류 알고리즘 설계와 구현

박수영* · 정채영**

The Algorithm Design and Implement of Microarry Data Classification using the Byesian Method

Su-Young Park* · Chai-Yeoung Jung**

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. 이처럼 DNA 마이크로어레이 기술은 복잡한 생물체를 이해하는 새로운 방향을 제시해주게 되었다. 따라서 이러한 기술을 통해 얻어진 대량의 유전자 정보들을 효과적으로 분석하는 방법이 시급하다.

본 논문에서는 실험용 데이터로 하버드대학교의 바이오인포메틱스 코어 그룹의 샘플데이터 이용하여 마이크로어레이 실험에서 다양한 원인에 의해 발생하는 잡음(noise)을 줄이거나 제거하는 과정인 표준화 과정을 거쳐 특징 추출방법인 베이지안 알고리즘 ASA(Adaptive Simulated Annealing) 방법을 이용하여 데이터를 2개의 클래스로 나누고, 정확도를 평가하는 시스템을 설계하고 구현하였다. Lowess 표준화 후 98.23%의 정확도를 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. Thus, DNA microarray technology presents the new directions of understandings for complex organisms. Therefore, it is required how to analyze the enormous gene information obtained through this technology effectively.

In this thesis, We used sample data of bioinformatics core group in harvard university. It designed and implemented system that evaluate accuracy after dividing in class of two using Bayesian algorithm ,ASA, of feature extraction method through normalization process, reducing or removing of noise that occupy by various factor in microarray experiment. It was represented accuracy of 98.23% after Lowess normalization.

키워드

microarray expression data, normalization, ASA(Adaptive Simulated Annealing)

Ⅰ. 서 론

최근 생물정보학(Bioinformatics)의 발전은 생명체 정보들을 대량으로 얻어내는데 큰 역할을 하고 있다. 특히 DNA에 있는 유전자(gene) 정보들을 분석해내기 위한 고도의 생명과학 실험 기술인 DNA 마이크로어레이 기술은

* 조선대학교 컴퓨터통계학과
** 교신저자

접수일자 : 2006. 8. 9

대량의 유전자 발현 정보를 만들어 내게 된다. DNA hybridization은 세포 내의 수천 개의 유전자들의 발현 정도를 동시에 측정하는 기술로, 이렇게 측정된 유전자 발현 데이터를 마이크로어레이(microarray) 데이터라 한다[1].

수천 개에서 수만 여개의 유전자들이 들어있는 마이크로어레이 데이터는 종양 샘플을 구하기가 쉽지 않을 뿐만 아니라 실험 비용도 매우 비싸 실제 표본의 개수에 비해 유전자의 개수가 훨씬 많다는 특성을 가지고 있다. 따라서 수많은 유전자들로부터 실제 종양들의 세부 분류에 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 model)을 구축하는 것이 매우 중요하게 부각되고 있다[2][3].

본 논문에서는 마이크로어레이 실험에서 다양한 원인에 의해 발생하는 잡음(noise)을 줄이거나 제거하는 과정인 표준화과정을 거쳐 특징 추출방법인 ASA 방법을 이용하여 데이터를 2개의 클래스로 나누고, 정확도를 평가하는 시스템을 설계하고 구현하였다. 논문의 2장에서는 마이크로어레이의 개요와 특징을 소개한다. 3장에서는 표준화 방법, ASA 검증 방법을 살펴보고, 4장에서는 마이크로어레이 데이터를 대상으로 알고리즘을 설계하고 실제로 실험한 결과를 분석하고 비교한다. 그리고 마지막으로 5장에서는 결론을 도출하고 향후 개선되어야 할 점을 논의한다.

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. 아무리 많은 양의 DNA 정보를 획득하였어도 그것만으로는 유전자가 무슨 일을 하는지, 세포가 어떤 역할을 하고 어떻게 유기체를 형성하며 어떻게 노화되는지 등에 대한 해답을 얻을 수는 없다. 따라서 이와 같이 방대한 양의 DNA 서열 정보를 의미 있게 이용하기 위한 기술이 필요한데, 이를 위해 분자 생물학 지식과 기계 및 전자 공학 기술이 결합된 것이 DNA 마이크로어레이이다[4].

마이크로어레이라 함은 샘플이 고밀도 순서대로 잘 정리된 것을 말하나 실제적으로 DNA-Chip 자체를 일컫기도 한다. DNA 마이크로어레이 실험의 기본적인 원리는 염기결합이다 (DNA의 경우 A와 T base 간의 이중 수소결합, G-C간의 3중 수소결합, RNA인 경우 A-U, G-C). 이는 염기결합 법칙에 근거하여 마이크로어레이 상의 probe에 알려 지거나 알려지지 않은 유전자 targe들을 결합시켜 유전자 발현 정도를 알아내는 매개체로 사용된다. 그림 1은 cDNA 마이크로어레이 실험과정이다[5].

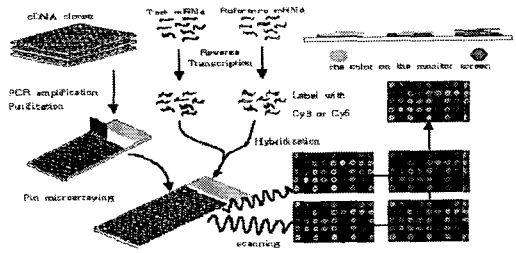


그림 1. 마이크로어레이 데이터 생성과정
Fig. 1. a creation process of microarry data

2.2. 마이크로어레이 활용분야

마이크로어레이는 짧은 시간에 적은 양의 시료로부터 많은 양의 정보를 얻을 수 있고, 자동화할 수 있다. 마이크로어레이의 응용분야는 유전자 발현 모니터링이다. 유전자 발현 모니터링은 여러 genome project를 통해 밝혀진 DNA Sequence를 이용하여 chip을 만들고, 이것을 이용하여 세포내 신진대사와 생 · 물리적 현상, 그리고 각 유전자들 간의 상호연관성을 밝혀내는데 활용될 수 있다. 이는 환경에 따른 발현 정도를 연구하는데 유용하게 쓰일 수 있으며, 기능을 알지 못하는 유전자에 대해 해석할 수 있는 실마리를 제공할 수 있을 것이다. 마이크로어레이가 가장 기대가 되는 응용분야는 유전병 진단에 활용인데, 마이크로어레이로 수많은 유전자를 빨리 시간 내에 정확하게 분석함으로써 유전자 변형에 의해 생기는 유전병을 진단하는데 쓰일 수 있기 때문이다. 이와 관련 Affymetrix에서는 제한적이지만 BRCA1과 같은 암 관련 유전자 진단 chip을 개발하여 선보였다. 마이크로어레이 응용분야는 생물공학, 환경, 농업, 화학공학분야 등 산업 전반에 걸쳐 확대되어가고 있다[5].

Ⅲ. 분류기법과 평가기법

3.1. 표준화(Normalization)

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 발생 할 수 있다. 표준화는 유전자 발현 값에 영향을 미치는 다양한 형태의 잡음을 찾아내어 제거하는 과정이라고 할 수 있다.

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.

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

표준화방법은 global(G) 표준화방법과 A를 고려하는 intensity dependent(ID) 표준화방법으로 구분한다. G 표준화 방법은 Chen et al. (1997)에 의해 제안된 방법으로 G와 R값이 한 슬라이드 내에서 일정한 비를 이루고 있다는 가정을 한 것으로 로그 비율의 분포의 중심을 상수(c)의 가감에 의해 0에 맞추어 가는 것으로 고전적인 표준화로써 전통적인 실험에서의 표준화 방법에 근거하여 로그 변화한 값을 표준화하는 것이다. 각 유전자별로 M을 다음과 같이 표준화한다[6].

$$M = \log_2\frac{R}{G} \Rightarrow \log_2\frac{R}{G} - c = \log_2\frac{R}{(k \bullet G)} \tag{2}$$

따라서 각 유전자(j)마다 표준화된 값을 M_j^{Global} 이라 하면 그 값은 식 (3)과 같이 정리할 수 있다.

$$M_j^{Global} = M_j\,\hat{c} \tag{3}$$

A를 고려하는 intensity dependent(ID) 표준화방법은 Yang et al. (2001)에 제안된 방법으로 M과 직교하는 척도 인텐시티 A를 제안하여 이를 기준으로 표준화하는 방법이다. 인텐시티란 각 형광 이미지 파일에서 측정된 강도 R과 G의 로그 변환한 값의 평균으로 구한다. 가장 먼저 간단한 가정으로 선형 모형에 대한 가정을 할 수 있으며 식 (4)와 같다[6].

$$M = \beta_0^{MA} + \beta_1^{MA} A \tag{4}$$

위 식에 의해 각 유전자(j)에 대한 표준화된 값을 M_j^{MA} 라 두고, 식 (5)에 이해 구한다.

$$M_j^{MA} = M_j - \widehat{M}_j \tag{5}$$

여기서 $\widehat{M}_j$은 회구분석 추정치를 이용하고, M_j^{MA}는 잔차값을 이용하게 된다. 이 방법을 A를 고려한 선형 표준화 방법이라고 한다.

선형모형에서 더 나아가 비선형적인 잡음이 첨가되는 경우에 대해 로버스트 smoother lowess 산점도를 사용하여 A에 의존한 비선형모형으로 확장할 수 있다. 즉, $M = k(A)$와 같은 A에 의존하는 일반적인 함수형태를 가정하고, 이상점에 대해 상대적으로 덜 민감한 LOWESS 함수 추정법으로 추정하는 방법이다. 각 유전자(j)별로 LOWESS 함수를 추정하며 각 유전자(j)에 대한 표준화는 M_j^{LOWESS}는 식 (6)과 같이 로그비율에서 함수에 의해 추정된 값을 제거함으로써 구해진다.

$$M_j^{LOWESS} = M_j - k(\widehat{A_j}) \tag{6}$$

3.2. 베이지안 알고리즘 ASA 검증기법

ASA는 파라미터 공간을 무작위 탐색하여 전영 최소점을 찾는 방법이다. 1989년 Ingber가 Very Fast Simulated Annealing(VFSR)을 고안한 이후로, quenching의 개념이 추가되어 Adaptive Simulated Annealing이라는 이름으로 알려져있다. ASA는 D 차원의 파라미터 공간을 가진 문제를 서로 다른 파라미터 범위에 대하여, 서로 다른 annealing-time-dependent한 감도에 따라 스케줄링하여 수렴 속도를 높인 시뮬레이티드 어닐링 방법이다.

Simulated Annealing의 성능에 영향을 미치는 것은 온도 파라미터의 초기값, 감소비율 그리고 정지조건의 세

가지이다. 이 세 가지 파라미터를 잘 조절하는 것을 annealing schedule이라 한다. 이상적으로 여겨지는 annealing schedule은 아래 그림 2와 같다[7].

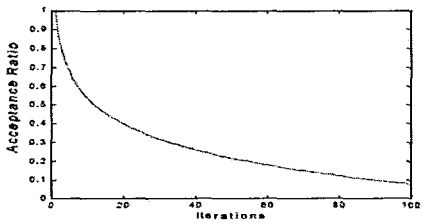


그림 2. 이상적인 annealing schedule
Fig. 2. a ideal annealing schedule

ASA를 비롯한 모든 시뮬레이티드 어닐링 방법은 다음과 같은 3가지의 중요한 요소로 구성된다.

- $T_{(k)}$: annealing 시간 t에서의 온도 T를 얻기 위한 스케줄
- $g_{T(\Delta E)}$: 생성 함수, D차원 공간에서 파라미터 (x = {xi; i = 1, ..., D})
- $h(\Delta E)$: 허용 함수, 새로운 상태로의 이동 여부를 결정할 확률 분포

$h(\Delta E)$는 이전 상태의 에너지 Ek에서 다음 상태 Ek+1로 전이될 확률이다. ΔE는 이전 상태와 다음 상태의 에너지의 차이이다($\Delta E = E_{k+1} - E_k$). $h(\Delta E)$가 수렴 속도에 매우 중요함을 알 수 있다. Boltzmann Annealing을 비롯한 많은 어닐링 알고리즘에서 다음의 허용함수를 사용한다[7].

$$h(\Delta E) = \frac{e^{-\frac{E_{k+1}}{T}}}{e^{-\frac{E_{k+1}}{T}} + e^{-\frac{E_k}{T}}} = \frac{1}{a + e^{\frac{\Delta E}{T}}}$$
$$\approx e^{\frac{-\Delta E}{T}} \quad (7)$$

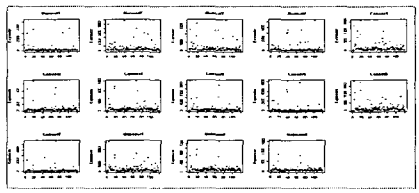
Ⅳ. 성능 평가 및 결과

4.1. 실험 데이터

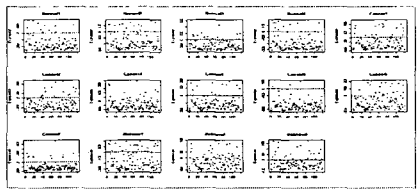
본 논문에서는 실험용 데이터로 하버드대학교의 바이오인포메틱스 코어 그룹의 샘플데이터를 사용하였다. 데이터는 12개 조직에서의 120개의 유전자 발현 셋으로 구성되었다.

4.2. 표준화(normalization)

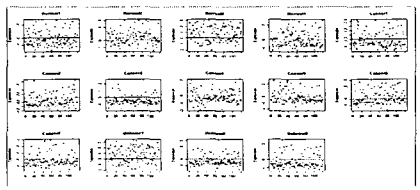
본 논문에서는 R을 이용하여 각 유전자의 발현 정도를 [0, 1] 범위로 표준화 하였고, 표준화 방법들의 실험결과를 비교 평가하기 위해 표준화 하지 않은 데이터를 사용하여 실험한 결과를 대조군으로 사용한다.



(a) 표준화 전 마이크로어레이 plot



(b) Global 표준화 후 마이크로어레이 plot



(c) Lowess 표준화 후 마이크로어레이 lot

그림 3. 표준화에 따른 유전자 발현 표준편차
Fig. 3. a standard deviation of gene expression by normalization

4.3. 베이지안 알고리즘 ASA의 성능 분석

본 논문에서 WEKA를 이용하여 표준화 방법들의 분류 성능을 평가하기 위해 Adaptive Simulated Annealing 검증을 하였고, 10-fold cross validation을 이용하여 정확도를 측정하였다. 그림 4는 표준화 방법들의 분류 성능을 비교하기 위한 베이지안 방법의 Adaptive Simulated Annealing 검증 방법의 분류 시스템 설계 그림이다.

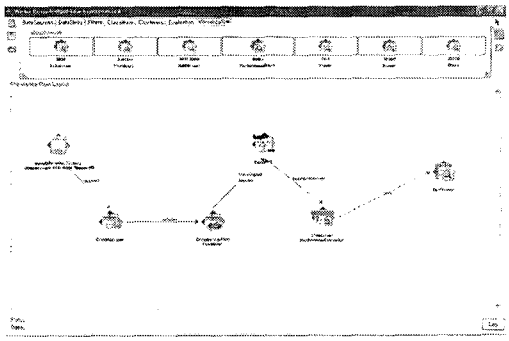


그림 4. 분류 시스템 설계
Fig. 4. a design of classification system

표 1은 WEKA를 이용하여 표준화 방법들의 분류 성능을 평가하기 위해 10-fold cross validataion을 이용한 Adaptive Simulated Annealing 검증 실험 결과이다. 표에 사용된 MSE(Mean Square Error)는 평균 제곱 오차를 나타내며, 실제 클래스와 예측한 클래스 차이를 제곱한 결과를 나타내며 이 값이 작을수록 좋은 분류를 나타낸다.

표 1. 검증 방법 결과
Table 1. a result of verification method

(%)	raw 데이터	Global 표준화	Lowess 표준화
정확도	84.23	95.33	98.23
MSE	0.35	0.18	0.04

실험 결과 표준화 전 데이터는 84.23%의 정확도를 보였고, Global 표준화 후 93.52%의 정확도를, Lowess 표준화 후에는 98.23%의 정확도를 보였다. 두 가지 표준화 후 정확도의 차이가 크지 않고 알고리즘의 결과가 좋게 나왔기 때문에 성능 대비 시간을 고려하여 좀 더 효율적인 표준화 방법을 구별하였다. 성능 대비 시간을 계산하기 위해 평균 실행 시간을 계산하였다.

표 2. 평균 실행 시간
Table 2. a mean of run time

초	raw 데이터	Global 표준화	Lowess 표준화
평균실행시간	10	7	2

두 표준화 모두 표준화를 하지 않은 데이터에 비해 대부분 시간이 짧게 걸렸다. 또한 Lowess 표준화 후에는 성능 대비 시간이 가장 효율적인 것으로 나타났다.

V. 결론 및 향후 연구과제

본 논문에서는 바이오인포매틱스 코어 그룹의 샘플데이터를 사용하여 표준화 후, Adaptive Simulatec Annealing 알고리즘을 사용하여 표준화 방법들의 성능을 비교 분석하였다. 실험결과 Lowess 표준화 후 98.23% 가장 높은 정확도를 보였고, 실행 시간이 적게 걸려 효율적인 것으로 나타났다. 데이터의 수가 증가 할 경우 Lowess 표준화 후 데이터의 수에 총 계산량의 영향을 받지 않는 Simulated Annealing 알고리즘의 성능이 더 좋아질 것으로 생각된다.

향후 연구과제로는 다양하고 보다 체계적인 많은 데이터의 획득과 분석을 통해 좀 더 효율적인 조합을 찾는 연구가 계속 되어야 할 것이다.

이에 아직 사용해보지 못한 또 다른 특징 추출방법과 Simulated Annealing 알고리즘의 파라미터를 달리하여 더 많은 연구를 진행하고자 한다.

참고문헌

[1] D.J.Duggan, M.Bittner, Y.Chen, P.Meltzer, J.M.Trent, "Expression profiling using cDNA microarray", Nature genetics supplement, Vol.21, pp.10-14, 1999.

[2] Jane Jijun Liu, Gene Cutler, Wuxiong Li, Zheng Pan, Sihua Peng, Tim Hoey, Liangbiao Chen and Xuefeng BruceLing, "Multiclass cancer classification and biomaker discovery using GA-based algorithms", Bioinformatics,vol.21, no.11, pp.2691-2697, 2005.

[3] 원홍희, 조성배, "암 분류를 위한 기계학습 분류기의 성능평가", 한국정보처리학회 추계 학술대회, vol.09, no.02. 2002.

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

[5] DNA chip 분석, http://www.bio.davidson,edu/courses/genomics/chip/chip.html

[6] Yang, Y.h., Dudoit, S., Luu, D.M., Peng, V., Ngai, J., and Speed, T.P., "Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variatation, Nucleic Acides Research, vol.30, no.c15, 2002.

[7] L. Ingber, "Very Fast Simulated Re-Annealing", Math1, Comput. Modeling, Vol. 12, pp, 967-973, 1989

저자소개

박 수 영(Su-Young Park)

2003년 조선대학교 컴퓨터통계학과 이학석사
2005년 조선대학교 컴퓨터 통계학과 박사과정수료

※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