제26회 한국정보처리학회 추계학술발표대회 논문집 제13권 제2호 (2006. 11)

정규화 기반 Adaptive Simulated Annealing을 이용한 마이크로어레이 데이터 분류 시스템

박수영* · 정채영*
*조선대학교 컴퓨터통계학과
e-mail : csssy@nate.com

The Classification System of Microarray Data Using Adaptive Simulated Annealing based on Normalization.

Su-Young Park* · Chai-Yeoung Jung*
*Dept. of Computer Science & Statistics, Chosun University

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 마이크로어레이 실험에서 다양한 원인에 의해 발생하는 잡음(noise)을 줄이거나 제거하는 과정인 정규화과정을 거쳐 특징 추출방법인 SVM(Support Vector Machine) 방법을 이용하여 데이터를 2개의 클래스로 나누고, 표준화 방법들의 성능 비교를 위해 Adaptive Simulated Annealing 알고리즘으로 정확도를 평가하는 분류 시스템을 설계 구현하였다.

1. 서론

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 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[2][3].

본 논문에서는 마이크로어레이 실험에서 다양한 원인에 의해 발생하는 잡음(noise)을 줄이거나 제거하는 과정인 표준화화과정을 거쳐 특징 추출방법인 SVM(Support Vector Machine) 방법을 이용하여 데이터를 2개의 클래스로 나누고, 표준화 방법들의 성능 비교를 위해 Adaptive

Simulated Annealing 알고리즘으로 정확도를 평가하는 분류 시스템을 설계 구현하였다. 논문의 2장에서는 마이크로어레이의 개요와 특징을 소개한다. 3장에서는 표준화 방법, SVM 방법 그리고 Adaptive Simulated Annealing 알고리즘을 살펴보고, 4장에서는 마이크로어레이 데이터를 대상으로 시스템을 설계하고 실제로 실험한 결과를 분석하고 비교한다. 그리고 마지막으로 5장에서는 결론을 도출하고 향후 개선되어야 할 점을 논의한다.

2. 관련연구

2.1 마이크로어레이(Microarray)

생명체의 생명 현상을 조적하는 것은 세포 내에 존재하는 DNA(DeoxyriboNucleic acid)라는 물질이다. 유전자는 DNA의 일부분으로서, 최종 산물인 단백질 생성에 필요한 정보를 담고 있다. 유전자가 mRNA 형태로 나타나는 현상을 유전자 발현(gene expression)이라 한다[4].

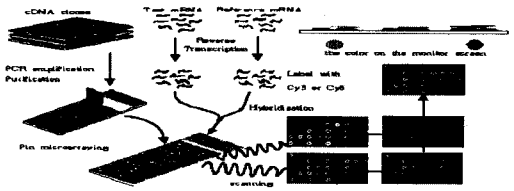


그림 1. 마이크로어레이 데이터 생성과정

3. 분류기법과 평가기법

3.1 표준화 방법

포괄적인 표준화 방법(Global normalization)은 고전적인 표준화로써 전통적인 통계적 실험에서의 표준화 방법에 근거하여 로그 변화한 값을 표준화하는 것이다.

먼저 Chen et al. (1977)에 의해 제안된 방법은 G와 R값이 한 슬라이드 내에서 일정한 비를 이루고 있다는 가정을 한 것으로 로그 비율의 분포의 중심을 상수의 가감에 의해 0에 맞추어 가는 것이다. 이는 투입되는 Cy3, Cy5 형광물질의 특성상 이러한 잡음이 첨가되리라는 것에 기초한 것이다[5].

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

포괄적인 표준화는 G나 R값 중 하나의 값을 고정한 수에 표준화하는 것으로 비율의 특성상 자료의 형태는 기본적으로 $R = G$ 에 대칭적으로 분포하게 되는데 한쪽을 고려하는 경우 대칭성이 깨질 우려가 있다. 이러한 고려 하에 Yang et al.은 M과 직교하는 척도 인텐시티 A를 제안하여 이를 기준으로 표준화하는 방법을 제안하였다. 인텐시티란 각 형광 이미지 파일에서 측정된 강도 R과 G의 로그 변환한 값의 평균으로 구한다.

가장 먼저 간단한 가정으로 선형모형에 대한 가정을 할 수 있으며 식 (4)와 같다.

$$M = \beta_0^{MA} + \beta_1^{MA} \quad (4)$$

위 식에 의해 각 유전자(j)에 대한 표준화된 값을 M_j^{MA}라 두고, 식 (5)에 의해 구한다[8].

$$M_j^{MA} = M_j - \hat{M}_j \quad (5)$$

3.2 SVM(Support Vector Machine)

SVM은 데이터를 두 개의 클래스로 분류하는 문제를 해결하기 위해 사용되는 기법이다. 이러한 문제는 이미 분류되어 있는 샘플 데이터로부터 계산된 함수를 통해서 다른 데이터를 두 개의 클래스로 분리하는 식으로 해결할 수 있다. 그림 2에서 샘플 데이터를 분리할 수 있는 Hyperplane이 매우 다양하게 존재할 수 있음을 볼 수 있다[4].

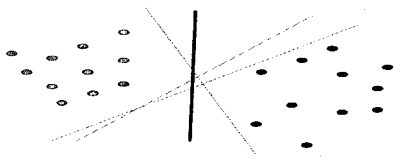


그림 2. 최적의 분리 초평면

3.3 Adaptive Simulated Annealing

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

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 $(x = \{x_i;\ i = 1, ..., D\})$
- $h(\Delta E)$: 허용 함수, 새로운 상태로의 이동 여부를 결정할 확률 분포

$h(\Delta E)$는 이전 상태의 에너지 E_k에서 다음 상태 E_{k+1}로 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.

$$h(\Delta E) = \frac{e^{-\frac{E_{k+1}}{T}}}{e^{-\frac{E_{k+1}}{T}} + e^{-\frac{E_k}{T}}} = \frac{1}{a + e^{\frac{\Delta E}{T}}} \approx e^{\frac{-\Delta E}{T}} \quad (5)$$

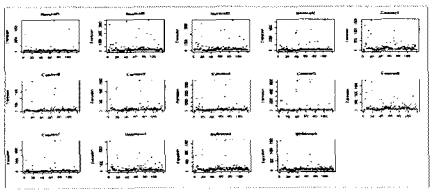
4. 성능 평가 및 결과

4.1 실험 데이터

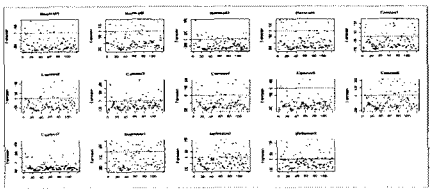
본 논문에서는 실험용 데이터로 하버드대학교의 바이오인포메틱스 코어 그룹의 샘플데이터를 사용하였다. 데이터는 12개의 조직에서의 120개의 유전자 발현 셋으로 구성되었다.

4.2 정규화(normalization)

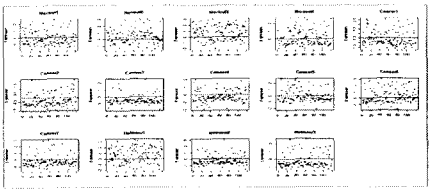
본 논문에서는 R을 이용하여 각 유전자의 발현 정도를 [0, 1] 범위로 정규화 하였고, 표준화 방법들의 실험결과를 비교 평가하기 위해 표준화 하지 않은 데이터를 사용하여 실험한 결과를 대조군으로 사용한다.



(a) 표준화 전 마이크로어레이 plot



(b) Global 표준화 후 마이크로어레이 plot



(c) Lowess 표준화 후 마이크로어레이 plot

그림 3. 표준화에 따른 유전자 발현 표준편차

4.3 Adaptive Simulated Annealing 검증 방법의 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 Adaptive Simulated Annealing 검증 방법의 분류 시스템 설계 그림이다.

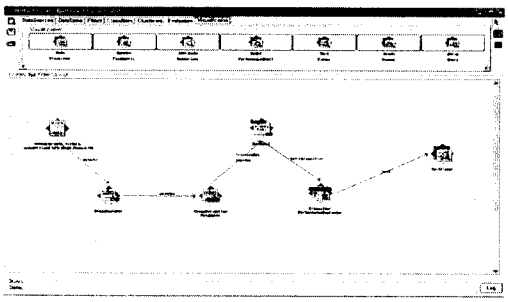


그림 4. 분류 시스템 설계

표 1은 WEKA를 이용하여 표준화 방법들의 분류 성능을 평가하기 위해 Adaptive Simulated Annealing 검증을 하였고, 10-fold cross validataion을 이용하여 정확도를 측정한 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

(%)	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

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.

5. 결론 및 향후 연구과제

본 논문에서는 바이오인포매틱스 코어 그룹의 샘플데이터를 사용하여 표준화 전과 표준화 후, SVM 알고리즘을 이용하여 클래스 분류 모델을 구축하고, Adaptive Simulatec Annealing 알고리즘을 사용하여 표준화 방법들의 성능을 비교 분석하였다. 실험결과 Lowess 표준화 후 가장 높은 정확도를 보였고, 효율적인 것으로 나타났다.

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] 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.

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