마이크로어레이 발현 데이터 분류를 위한 베이지안 검증 기법

박수영* · 정종필** · 정채영*

A Bayesian Validation Method for Classification of Microarray Expression Data

Su-Young Park* · Jong-Pil Jung** · Chai-Yeoung Jung*

요 약

생물정보는 사람의 능력을 넘어섰으며 데이터 마이닝과 같은 인공지능 기법이 필수적으로 요구된다. 한 번에 수천 개의 유전자 발현 정보를 획득할 수 있는 DNA 마이크로어레이 기술은 대량의 생물정보를 가진 대표적인 신기술로 질병의 진단 및 예측에 있어 새로운 분석방법들과 연계하여 많은 연구가 진행 중이다. 이러한 새로운 기술들을 이용하여 유전자의 메커니즘을 규명하는 것은 질병의 치료 및 신약의 개발에 많은 도움을 줄 것으로 기대된다.

본 논문에서는 마이크로어레이 실험에서 다양한 원인에 의해 발생하는 잡음(noise)을 줄이거나 제거하는 과정인 표준화과정을 거쳐 표준화 방법들의 성능 비교를 위해 특징 추출방법인 베이지안(Bayesian) 방법을 이용하여 마이크로어레이 데이터의 분류 정확도를 비교 평가하여 Lowess 표준화 후 95.89%로 분류성능을 향상시킬 수 있음을 보였다.

ABSTRACT

Since the bio-information now even exceeds the capability of human brain, the techniques of data mining and artificial intelligent are needed to deal with the information in this field. There are many researches about using DNA microarray technique which can obtain information from thousands of genes at once, for developing new methods of analyzing and predicting of diseases. Discovering the mechanisms of unknown genes by using these new method is expecting to develop the new drugs and new curing methods.

In this paper, We tested accuracy on classification of microarray in Bayesian method to compare normalization method's performance after dividing data in two class that is a feature abstraction method through a normalization process which reduce or remove noise generating in microarray experiment by various factors. And We represented that it improve classification performance in 95.89% after Lowess normalization.

키워드

microarray expression data, normalization, Bayesian validation method

* 조선대학교 컴퓨터통계학과

** 교신저자

접수일자 : 2006. 7. 24

Ⅰ. 서 론

복잡한 조절 기능을 갖는 생명 현상에 대한 분자 수준의 단편적인 이해는 한계가 있기 때문에 인간 게놈 프로젝트(Human Genomic profect: HGP)와 같이 전체적인 이해를 위한 연구의 필요성이 대두되었다. DNA 칩 기술의 발전은 유전자 정보의 대량 생산을 가능하게 하였고, 특정한 실험 환경[Harrington2000]과 조건에 따른 수천 개의 유전자 발현 정도를 동시에 파악할 수 있게 하였으며, 이를 대량으로 처리함으로써 수천 개의 유전자 정보를 굉장히 빠르고 정확하게 분석할 수 있게 되었다[1].

일반적으로 DNA 마이크로어레이 데이터는 방대한 양의 유전자 정보를 갖고 있으므로, 이를 효율적으로 분석하기 위하여 다양한 데이터 마이닝 기법, 기계학 알고리즘, 통계적인 방법 등을 적용하고 그 성능을 평가하려는 연구가 진행되고 있다[2].

본 논문에서는 하버드대학교의 바이오인포메틱스 코어 그룹의 샘플데이터에 대해 마이크로어레이 실험에서 다양한 원인에 의해 발생하는 잡음(noise)을 줄이거나 제거하는 과정인 표준화과정을 거쳐 특징 추출방법인 베이지안 방법을 이용하여 표준화 방법들의 정확도를 비교 평가하였다.

논문의 2장에서는 마이크로어레이의 개요와 표준화의 정의와 중요성을 소개하고, 3장에서 표준화 방법, 베이지안 검증 방법을 살펴본다. 4장에서는 마이크로어레이 데이터를 대상으로 실제로 실험한 결과를 보이고 분석한다. 그리고 마지막으로 5장에서는 결론을 도출하고 향후 개선되어야 할 점을 논의한다.

Ⅱ. 관련 연구

2.1. 마이크로어레이 개요

DNA 마이크로어레이는 유전자 발현 정보를 얻기 위해 고형 지지체(substrate) 위에 실험하고자 하는 대량의 유전자를 고정해 놓은 것이다. DNA 마이크로어레이 기술은 고형 지지체 위에 유전자를 고정하는 방법에 따라 cDNA 마이크로어레이 기술과 oligonucleotide 마이크로어레이 기술로 나뉜다.

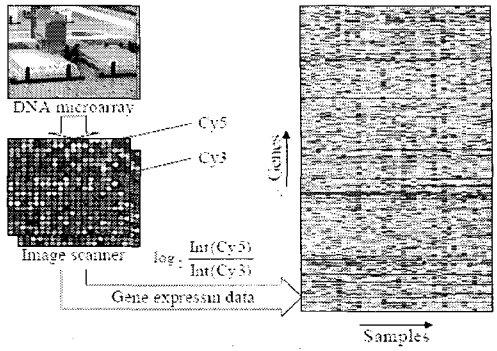


그림 1. DNA 마이크로어레이로부터 유전자 발현정보를 얻는 과정

Fig. 1. an obtained procedure of genes expression information from DAN microarray

1) cDNA 마이크로어레이

두 가지 다른 환경에서 세포들로부터 추출한 mRNA를 역전사(reverse transcription) 시킬 때 두 가지의 형광 물질을 띤 염기(IdUTP)를 집어넣어 빨간색(Cy5)이나 녹색(Cy3)을 띤 cDNA를 합성한 다음, 합성된 두 개의 cDNA를 똑같은 양으로 섞어서 하나의 cDNA 마이크로어레이 결합(hybridization)시킨다. 결합 반응이 끝나면 cDNA 마이크로어레이는 스캐너에 의하여 읽혀지고 그 결과 화상 데이터를 얻을 수 있는데, 이 화상 데이터는 각 유전자의 형광 정도를 담고 있으므로 이를 컴퓨터 소프트웨어를 통해 분석함으로써 유전자의 발현(expression) 정도를 수치화할 수 있다. 유전자 발현정보 데이터로 사용하기 위하여 Cy3와 Cy5형광정도의 상대적 강도를 식 (1)과 같이 구한다[3,4,5].

$$\text{gene\_expression} = \log_2 \frac{Cy5}{Cy3} \tag{1}$$

2.2. 표준화의 정의와 중요성

실험에 의해서 얻어진 마이크로어레이 자료는 자료의 특성상 많은 잡음을 포함하고 있다. 크게 잡음에는 두 가지가 있는데 하나는 자료자체에 내재하는 잡음과 다른 하나는 실험 수행에서 발생하는 잡음이다. 전자의 경우 해당 자료의 산포를 증대시켜 통계적 검정력을 약화시키나 개개의 자료에 대해 랜덤하게 발생하므로 제거가 어려운 반면 후자의 실험에 의한 잡음의 경우 통계적 분석의 결과에 유의한 영향을 미칠 수 있는 위험성을 가지고 있으

므로 표준화를 통하여 다양한 형태의 잡음을 찾아내어 제거해야 한다. 즉, 표준화란 마이크로어레이 실험에서 슬라이드 내 자료의 유전자 발현 수준을 정확하게 측정할 수 있도록 유전자 발현 수준에 영향을 미치는 여러 잡음을 찾아 제거함과 동시에 각 슬라이드 혹은 실험 간의 결과를 비교할 수 있도록 자료를 보정하는 것이다.

슬라이드 내에서의 표준화는 보통 로그 변환한 두 값 $\log_2 R$과 $\log_2 G$의 균형을 맞추기 위해 행해지는데, 위치 보정의 경우는 일반적으로 로그비율 M의 평균이나 중위수를 사용하여 보정하며, 스케일 보정의 경우 분산값을 보정하여 표준화시켜준다[6].

Ⅲ. 분류기법과 평가기법

3.1. 표준화 방법

포괄적인 표준화 방법(Global normalization)은 Chen et al. (1977)에 의해 제안된 방법으로 고전적인 표준화로써 전통적인 통계적 실험에서의 표준화 방법에 근거하여 로그 변화한 값을 표준화하는 것이다.

Global 표준화 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[7].

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

포괄적인 표준화는 G 나 R 값 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 M 과 직교하는 척도 인텐시티 A 를 제안하여 이를 기준으로 표준화하는 방법을 제안하였다. 인텐시티란 각 형광 이미지 파일에서 측정된 강도 R 과 G 의 로그 변환한 값의 평균으로 구한다.

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

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

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

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

3.2. 베이지안 검증 기법

베이지안 검증 방법은 주어진 데이터에 대해 각 클래스의 사후확률이 최대가 되는 것을 최적의 클래스 파티션으로 한다.

$$\max P(Class|Dataset) \tag{8}$$

이 사후확률 값을 구하기 위하여 Bayes' Theorem을 적용하면 식 (9)와 같이 사전확률을 이용하여 사후확률 값을 계산할 수 있다.

$$P(Class|Dataset) = \frac{P(Class)P(Dataset|Class)}{P(Dataset)} \tag{9}$$

각 데이터들이 서로 독립이라 가정하면 multiplication rule과 independent rule에 의해 식 (9)는 식 (10)와 같이 표현될 수 있다.

$$\begin{aligned} &P(Class|Dataset) = P(Class|d_1, d_2, \cdots, d_N) \\ &= P(Class|d_1) \text{ x } P(Class|d_2) \text{ x } \cdots \text{ x } P(Class|d_N) \end{aligned} \tag{10}$$

이러한 과정을 이용하여 식 (11)와같이 베이지안 스코어(Bayesian score)를 계산하여 이를 베이지안 검증 방법에 사용할 수 있으며, 그 값이 클수록 각 클래스의 사후확률이 커지므로 좋은 클래스 파티션임을 나타낸다.

$$\begin{aligned} BS &= \frac{\sum_{i=1}^{c} P(C_i|D_i)}{C} = \frac{\sum_{i=1}^{c} P(C_i|d_{i1}, d_{i2}, \cdots, d_{iN})}{C} \\ &= \frac{\sum_{i=1}^{c} P(C_i|d_{i1})P(C_i|d_{i2})\cdots P(C_i|d_{iN})}{C} \\ &= \frac{\sum_{i=1}^{c}\prod_{j=1}^{N_i} P(C_i)p(d_{ij}|C_i)/p(d_{ij})}{C} \end{aligned} \tag{11}$$

여기서 $D_i = \{d_{ij} | \mu_{ij} > \alpha, 1 \le j \le n\}, N_i = n(D_i)$ 이며, 각 확률들을 다음과 같이 계산할 수 있다[10].

$$P(C_i) = \frac{\sum_{j=1, \mu_{ij} > \alpha}^{n} \mu_{ij}}{\sum_{i=1}^{c}\sum_{j=1}^{n} \mu_{ij}}$$

$$P(d_{ij}) = \sum_{i=1}^{c} P(C_i)p(d_{ij}) = \sum_{i=1}^{c} (C_i)\mu_{ij} \quad (8)$$

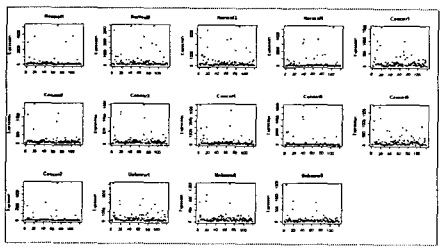
Ⅳ. 성능 평가 및 결과

4.1. 실험 데이터

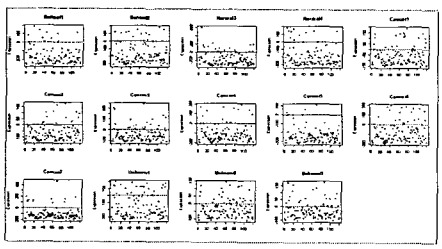
본 논문에서는 실험용 데이터로 하버드대학교의 바이오 인포메틱스 코어 그룹의 샘플데이터를 사용하였다. 데이터는 12개의 조직에서 120개의 유전자 발현 셋으로 구성되었다.

4.2. 표준화(normalization)

표준화 방법들의 실험결과를 비교 평가하기 위해 표준화 하지 않은 데이터를 사용하여 실험한 결과를 대조군으로 사용한다. R을 이용하여 마이크로어레이 데이터를 표준화 하였고, 표준화 하지 않은 마이크로어레이 데이터와 표준화 후 마이크로어레이 데이터에 대해 BayesNet을 사용하여 Lowess 표준화 후 분류성능을 향상시킬 수 있음을 보이고자 한다.



(a) 표준화 전 마이크로어레이 plot



(b) Global 표준화 후 마이크로어레이 plot

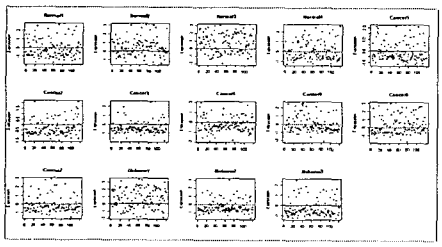


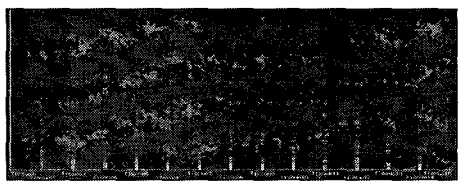
그림 2. 표준화에 따른 유전자 발현 표준편차
Fig. 2. a standard deviation of gene expression by normalization

4.3. 베이지안 검증 방법의 성능 분석

본 논문에서는 데이터마이닝 툴 WEKA를 이용하여 표준화 방법들의 분류 성능을 평가하기 위해 BayesNet 검증을 하였고, 10-fold cross validation을 이용하여 정확도를 측정하였다. 표준화 방법들의 분류 성능을 비교하기 위한 베이지안 검증 방법의 실험 결과가 그림 3과 표 1에 나와 있다. 표에서 사용된 *MSE(Mean Square Error)*는 평균제곱 오차를 나타내며, 실제 클래스와 예측한 클래스 차이를 제곱한 결과를 나타내며 이 값이 작을수록 좋은 분류를 나타낸다.

표 1. 검증 방법 결과
table 1. the result of verification method

	raw 데이터	Global 표준화	Lowess 표준화
정확도	89.46	93.52	95.89
MSE	0.3	0.21	0.18



(a) 표준화 전 마이크로어레이 베이지안 검증

(b) Global 표준화 후 마이크로어레이 베이지안 검증

(c) Lowess 표준화 후 마이크로어레이 베이지안 검증

청색 : normal tissue □ : negative classification
빨간색 : cancer tissue × : positive classification

그림 3. 검증 방법 결과
Fig. 3. the result of verification method

실험 결과 표준화 전 데이터는 89.46%의 정확도를 보였고 Global 표준화 후 93.52%의 정확도를 보였으며 Lowess 표준화 후에는 95.89%의 정확도를 보였다. Global 표준화 후 데이터 검증을 했을 때 보다 Lowess 표준화 후 데이터를 검증 했을 때 분류성능이 더 높다는 것을 알 수 있다.

□는 정상조직 유전자를 암 조직 유전자로 암 조직 유전자를 정상 조직 유전자로 잘못 분류한 경우를 나타내며, × 는 올바른 분류를 나타낸다.

V. 결론 및 향후 연구과제

본 논문에서는 바이오인포매틱스 코어 그룹의 샘플데이터를 사용하여 표준화 전과 표준화후, 베이지안 알고리즘을 이용하여 클래스 분류 모델을 구축하고, 표준화 방법들의 정확도를 비교 분석하였다. 그 결과 Lowess 표준화 방법으로 데이터 검증을 하였을 때 95.89%로 Global 표준화 후 데이터 검증을 하였을 때 93.52%보다 더 높은 정확도를 보였고, 0.18%의 더 낮은 MSE를 보였다. 클래스와 연관성이 높고 분별력이 있는 유전자를 사용하고 여러 요인에 의한 잡음을 고려한 경우 마이크로어레이 데이터 분류에 있어 정확도의 차이가 크게 날 것으로 생각된다.

향후 연구과제로는 다양하고 보다 체계적인 많은 데이터의 획득과 분석을 통해 좀 더 효율적인 조합을 찾는 연구가 계속 되어야 하고, 이와 더불어 이렇게 제안된 최적의 조합이 다른 종류의 데이터를 대상으로 한 검증이 이루어져야 할 것이다.

참고문헌

[1] M.B. Eisen and P.O.Brown,"DNA arrays for analysis of gene expression," Methods Enzambol, vol.303, pp.179-205, 1999

[2] U.Alon, N.Barkai, D.A.Notterman, K.Gish, S.Ybarra, D.Mack, A.J. Levine, "Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays," Proc. of the Natl. Acad. of Sci. USA, vol.96, no.12, pp.6745-6750, 1999.

[3] J.Derisi, V.Iyer and P.Brosn, "Exploring the metabolic and genetic control of gene expression on a genomic scale," Science, vol.278, pp.680-686, 1997.

[4] D.Lashkari, J.Derisi, J.McCusker, A.Namath, C.Gentile, S.Hwang, P.Brown, and R.Davis, "Yeast microarrays for genome wide parallel genetic and gene expression analysis," Proc. of the Natl. Acad. of Sci. USA, vol.94, pp.13057-13062, 1997.

[5] R.J. Lipshutz, S.P.A.Fodor, T.R.Gingeras and D.J.Lockhart, "High density synthetic oligonucleotide arrays.", Nature Genetics, vol.21, pp.20-24. 1999.

[6] Ben-Dor, A., Bruhn, L., Friedman, N., Nachman, I., Schummer, M., and Yakhini, Z. "Tissue classification with gene expression profiles," Journal of Computational Biology, vol.7, no.3/4, pp.559-584, 2000.

[7] Y. Chen, E.R. Dougherty and M.L. Bittner, "Ratio-Based Decision and the Quantitative Analysis of cDNA Microarray Images," Journal of Biomedical Optics, vol. 2, no. 3, pp. 364-374, 1997.

[8] Y.H. Yang, S. Dudoit, P. Luu and T.P. Speed, "Normalization for cDNA Microarray data," SPIE BiOs, 2001.

[9] Mangasarian, O.L., Musicant D.R., "Lagrangian support vector machines," Journal of machine learning Research, vol.1, pp.161-177, 2001.

[10] Cooper, G.F. and Herskovits, E., "A Bayesian Method for the induction of probabilistic networks form data,"Machine Learning, 9, pages 309-347, 1992.

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 종 필(Chai-Yeoung Jung)

1995년 조선대학교 전산통계학과 이학석사
2002년 조선대학교 전산통계학과 이학박사
1997년~현재 순천청암대학 컴퓨터정보과 교수
※관심분야 : 영상처리, 컴퓨터비젼, 멀티미디어컨텐츠, 모바일이터넷, 디지털미디어, Bioinformatics

정채영(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
※ cyjung@chosun.ac.kr 062)230-7964