The Korean Journal of Applied Statistics (2012)
25*(3), 415–422*
DOI: http://dx.doi.org/10.5351/KJAS.2012.25.3.415

A Study of HME Model in Time-Course Microarray Data

Sungmin Myoung[1] · Donggeon Kim[2] · Jinnam Jo[3]

[1]Faculty of Health Science, Jungwon University
[2]Department of Information and Statistics, Dongduk Women's University
[3]Department of Information and Statistics, Dongduk Women's University

(Received January 25, 2012; Revised February 21, 2012; Accepted April 20, 2012)

Abstract

For statistical microarray data analysis, clustering analysis is a useful exploratory technique and offers the promise of simultaneously studying the variation of many genes. However, most of the proposed clustering methods are not rigorously solved for a time-course microarray data cluster and for a fitting time covariate; therefore, a statistical method is needed to form a cluster and represent a linear trend of each cluster for each gene. In this research, we developed a modified hierarchical mixture of an experts model to suggest clustering data and characterize each cluster using a linear mixed effect model. The feasibility of the proposed method is illustrated by an application to the human fibroblast data suggested by Iyer *et al.* (1999).

Keywords: Hierarchical Mixture of Experts, Mixture model, Linear Mixed Effect Model, Microarray.

1. 서론

마이크로어레이 자료는 수많은 유전자 발현자료들을 생성하기 때문에, 연구자들은 이들의 변화를 동시에 관찰 연구하는 것이 중요한 연구과제이다. 이를 위한 마이크로어레이 자료분석방법론의 개발 및 이에 대한 해석은 필수적이다 (Lander, 1999).

마이크로어레이 분석기법 중 군집분석(clustering)은 수많은 유전자들과 생물학적 네트워크의 복잡성을 가지는 유전자 발현자료의 분석에 대하여 유용하게 설명할 수 있는 통계학적 방법이다 (Yeung 등, 2001). 유전자 발현자료의 분석을 위하여 기존에 제안된 군집분석 알고리즘들은 Eisen 등 (1998)의 계층적 군집분석(hierarchical clustering), Tamayo 등 (1999)의 자기 구성 지도(Self-Organizing Map; SOM), Tavazoie 등 (1999)의 k-평균 군집분석, Hartuv 등 (1999)의 그래프-이론 접근방법(graph-theoretic approach; Hartuv 등, 1999) 및 Brown 등 (2000)이 제안한 SVM(support vector machine) 등이 있다.

생물학적 프로세스(biological process)가 이루어지는 동안에 어떤 특정한 셀(cell)의 발현이 어느 시점에서 이루어지는지에 대한 연구를 시간 경로(time course) 실험이라 한다 (Costa 등, 2004). 이는 생물

This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the ministry of Education, Science and Technology(2011-0013877).
[3]Corresponding author: Professor, Department of Information and Statistics, Dongduk Women's University, 23-1 Wolgok-Dong, Sungbuk-Gu, Seoul 136-714, Korea. E-mail: jinnam@dongduk.ac.kr

학적 프로세스를 통하여 동일한 형태로 발현되는 유전자들의 군들을 확인하는 것으로서, 생물학자들은 유전자의 기능과 유전자 규칙성(gene regulation)에 대한 매커니즘을 시간 경로 실험으로부터 추론하는 것이 목적이다 (Quackenbush, 2001; Slonim, 2002; Draghici, 2003; Quakenbush, 2001).

이러한 시간경로 마이크로어레이 자료에 대하여 현재까지 많은 군집분석방법들이 제안되었다. Yeung 등 (2001, 2003)은 모형에 기초한(model-based) 군집분석방법을 제안하였으며, Luan과 Li (2003)는 B-스플라인을 가지는 혼합모형(mixed effect model)을 이용하여 군집방법을 제안하였다. Storey 등 (2005)는 ODP(Optimal Discovery Procedure)에 기초한 시간경로 실험에 대한 새로운 유의성 분석(significance analysis)을, Wang 등 (2009)은 임의확률계수모형(random coefficients model)을 이용한 통합된 통계적 모형을 제안하였다.

Jordan과 Jacobs (1994)에 의해 제안된 계층적 혼합 엑스퍼트(Hierarchical Mixture of Experts; HME) 모형은 분리와 해결의 원칙(divide and conquer)을 이용하여 입력 공간(input space)을 각 부지역(sub-region)들로 반복하여 분할하는 내포모형(nested model)을 일반화 한 것으로, EM-알고리즘을 통하여 혼합모형(mixture model)의 모수들을 추정할 수 있다.

이러한 점에 착안하여, 본 연구에서는 시간 경로에 대한 마이크로 어레이 실험을 시간에 따라 반복측정된 유전자 발현자료의 형태로 파악하여 선형혼합모형(linear mixed model)기반의 HME 모형을 제안하고, 이 방법을 통하여 유전자들의 각 군집을 계층적 확률모형의 형태로 표현하는 동시에, 각 군집들의 선형추세를 파악하는 모형을 제시하고자 한다 (Schlattmann, 2009).

본 논문의 구성은 다음과 같다. 2절에서는 HME 모형에 대하여 설명하고, 3절에서는 시간에 따라 반복측정된 마이크로어레이 자료에 대하여 선형혼합모형을 고려한 수정된 HME 모형에 대하여 제안한다. 4절에서는 실제자료에 대한 분석으로, Iyer 등 (1999)에 의해 사용된 혈청 섬유아세포(fibroblast)에 대하여 제안된 모형을 적용하며, 5절에서는 결론 및 본 연구의 성과에 대해 설명한다.

2. HME 모형

HME 모형은 Jordan과 Jacobs (1994)에 의하여 제안된 나무모형 기반의 지도학습(supervised learning) 알고리즘이다. HME 모형과 나무모형과의 주요한 차이점은 분할점을 이산적으로 탐색하지 않고 부드러운 확률적 형태(soft probabilistic)로 탐색한다는 점이다 (Hastie 등, 2001). 부드러운 확률적 형태의 의미는 입력 변수(input variable)에 의해 추정된 확률값을 기준으로 왼쪽 또는 오른쪽 노드로 분류되는 것을 의미하는데, 이를 통하여 나무모형보다 자료에 대한 유용한 해석을 제공할 수 있다는 장점이 있다. 또한 나무 모형에서는 각 노드에서의 추정량이 상수형태로 제공되는 것과는 다르게, HME 모형에서는 선형모형을 이용하여 각 종단 노드의 추정량을 제공한다.

요약하면, HME 모형은 분리와 해결의 원칙(divide and conquer)에 의해, 입력변수를 각 부지역(sub-region)들로 반복하여 분할하는 내포모형(nested model)을 일반화 한 것으로, EM-알고리즘을 통하여 혼합모형(mixture model)의 모수들을 추정한다 (Jordan과 Jacobs, 1992).

Figure 2.1은 수준이 2개인 HME 모형을 나타낸 것이다. 이는 각각의 비종단 노드에서 부드러운 분할을 가지는 나무모형이다. 그러나 Jordan과 Jacobs (1992)은 새로운 용어를 이용하였다는데, 종단노드를 엑스퍼트 네트워크(expert network), 비종단 노드는 입력 네트워크(gating network)으로 정의하였다. 이와 같이 정의했을 때, 각 엑스퍼트 네트워크들은 반응변수에 대한 예측을 의미하며, 이러한 엑스퍼트 네트워크들은 입력 네트워크에 의해서 서로 연결되어 있다.

입력네트워크는 벡터 x를 입력받아 왼쪽 또는 오른쪽 노드로 분류될 확률을 의미하며 g_i로 정의한다.

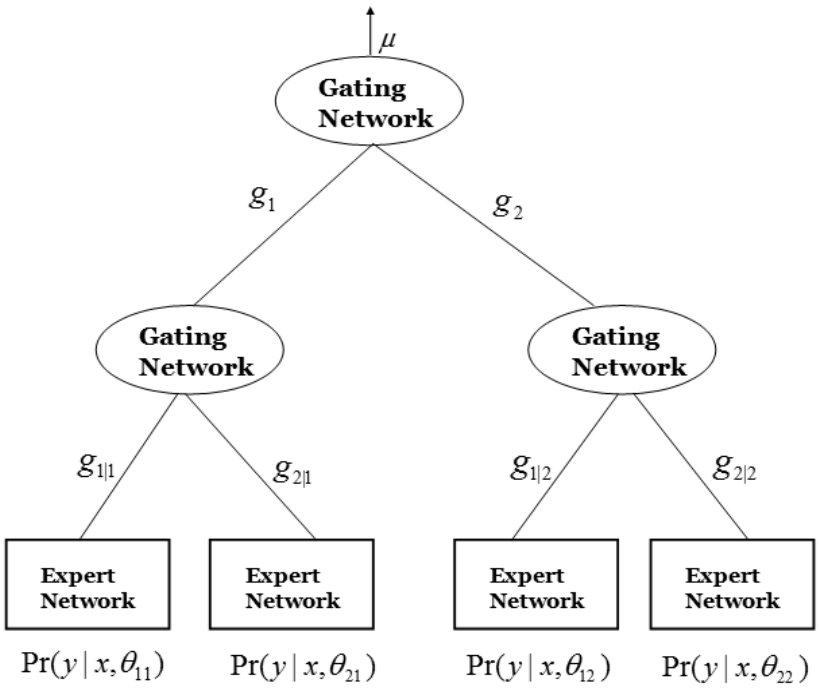

Figure 2.1. A two-level hierarchical mixture of experts(HME) model

g_i의 결과는 소프맥스(softmax)함수이며, 각 입력벡터 x에 대해서 g_i들의 합은 1이다 (Bridle, 1989; McCullagh와 Nelder, 1983).

엑스퍼트 네트워크에서 반응변수에 관한 모형은 $Y \sim \Pr(y|x,\theta_{ij})$이며, 가우시안 선형 회귀모형 혹은 선형 로지스틱 모형이 사용된다. 입력벡터 x로부터 y를 생성하는 HME의 확률모형은 다음과 같다.

$$P\left(y|x,\theta^0\right) = \sum_i g_i\left(x,v_i^0\right) \sum_j g_{j|i}\left(x,v_{ij}^0\right) P\left(y|x,\theta_{ij}^0\right),$$

여기서 기호 '0'가 의미하는 것은 모수의 실제 값이며, θ는 모수벡터이다. 위의 모형은 입력 네트워크 모형에 의해 결정되는 혼합 비율(mixing proportion)을 가지는 혼합 모형(mixture model)이다 (Jordan과 Jacobs, 1994). 모수벡터의 최대우도추정치는 Dempster 등 (1977)이 제안한 EM-알고리즘을 이용한다.

먼저 잠재변수(latent variable) z_{ij}를 정의하는데, 이는 z_i와 $z_{i|j}$의 곱으로 표시되며, 확률모형에서의 엑스퍼트를 나타낸다. E-step은 모수의 현재값이 주어진 상태에서 z_i와 $z_{i|j}$의 기대값을 계산한다. E-step에서 계산된 기대값은 M-step에서의 엑스퍼트 네트워크들의 모수들을 추정하기 위한 관찰 가중치로 사용된다. M-step에서는 E-step에서 계산된 관찰가중치를 가지고 일반화 선형모형(generalized linear model)에 대한 가중치 우도함수를 최대화 시키는 방법인 반복 재가중 최소제곱(iteratively reweighted least square; IRLS)을 이용하여 최대우도추정치를 구하고 이를 수렴할 때까지 반복한다 (Little과 Rubin, 2002; McLachlan, 2008)).

3. 시간에 따른 반복측정된 마이크로어레이 자료에 대한 HME 모형의 수정

시간 경로에 대한 마이크로 어레이 실험의 경우, 기존의 HME 모형을 적용하기 위해서는 가정된 엑스퍼트가 가우시안 회귀모형 혹은 로지스틱 회귀모형이기 때문에 분석의 한계가 있다. 그러므로 본 절에서는 시간 경로에 대한 마이크로어레이 실험을 시간에 따라 반복측정된 유전자 발현자료의 형태로 파악하여 선형혼합모형(linear mixed model)기반의 HME 모형을 제안한다.

자료 $(x_i^{(t)}, y_i^{(t)})$에서, y_i는 연속형 반응변수이고 x_i는 입력벡터이다. 즉, 마이크로어레이 자료에서 x_i는 i번째 유전자의 시간에 관한 변수라고 할 수 있고, y_i는 i번째 유전자의 각 시간에 대한 유전자의 발현수준(expression level)인 Cy3와 Cy5의 로그 비(log ratio)다. 또한 n개 유전자들의 HME 모형의 상하위 수준이 각각 k개의 유전자 군집을 이루어져 있다고 가정한다.

HME 모형의 상위수준 입력네트워크는 다음과 같다.

$$g_i(x, \gamma_j) = \frac{e^{\gamma_j^T x}}{\sum_{k=1}^{K} e^{\gamma_k^T x}}, \quad j = 1, \ldots, k.$$

위의 식에서 $g_j(x, \gamma_j)$는 입력벡터 x가 j번째 가지의 관찰치를 할당하는 확률을, γ_j는 알려지지 않은 모수벡터이며, $g_j(x, \gamma_j)$는 γ_j의 소프맥스 함수이다.

하위수준 입력네트워크는 상위 수준에서와 비슷한 형태로 제공되며, 이는 상위수준에서 j번째 가지로 할당되어 졌고, 하위수준에서의 ℓ번째 가지로 할당될 확률이며 아래와 같이 정의한다.

$$g_{\ell|j}(x, \gamma_{j\ell}) = \frac{e^{\gamma_{j\ell}^T x}}{\sum_{k=1}^{K} e^{\gamma_{jk}^T x}}, \quad \ell = 1, \ldots, k.$$

각 엑스퍼트에서 반응변수에 대한 모형은 $Y \sim \Pr(y|x, \theta_{j\ell})$로서 나타나며, Laird과 Ware (1982)이 제안한 선형 혼합모형(linear mixed effect model)을 적용한다. 여기서 고정효과(fixed effect) X_i는 시간에 대한 효과를 의미하고, 임의효과(random effect)는 각 유전자들을 의미한다.

$$y_i = X_i\beta + Z_i b_i + \epsilon_i, \quad i = 1, \ldots, n \ \text{ where } b_i \sim N(0, D),\ \epsilon \sim N(0, \Sigma_i).$$

모든 모수들의 집합을 $\psi = \{\gamma_j, \gamma_{j\ell}, \theta_{j\ell}\}$로서 정의하면, $Y = y$일 전확률(total probability)은 다음과 같으며, 이는 입력 네트워크 모형에 의해서 결정되는 혼합모형의 형태가 된다.

$$\Pr(y|x, \psi) = \sum_{j=1}^{K} g_i(x, \gamma_j) \sum_{\ell=1}^{K} g_{\ell|j}(x, \gamma_{j\ell}) \Pr(y|x, \theta_{j\ell}).$$

각 모수들을 추정하기 위해서 자료의 로그우도함수 $\sum_i \ln \Pr(y_i|x_i, \psi)$를 ψ에 관하여 최대화시키는 방법으로 EM-알고리즘을 이용한다.

즉, 선형 혼합모형의 혼합 형태를 추정하기 위해서 잠재변수 z가 알려져 있다고 가정한다면, 최대우도추정치의 계산은 각 엑스퍼트 네트워크에 관한 선형 혼합모형 문제와 입력 네트워크에 관한 다중분류(multiway classification)문제로 귀결되어, 서로 독립적으로 해를 구해야 한다.

완전 자료 로그우도함수(complete data log likelihood)는 다음과 같다.

$$\ln L_c(\psi) = \sum_t \sum_j \sum_\ell z_{j\ell}^{(t)} \left\{ \ln g_j^{(t)} + \ln g_{\ell|j}^{(t)} + \ln \Pr\left(y^{(t)}|x^{(t)}, \theta_{j\ell}\right) \right\}.$$

Table 4.1. Parameter estimation for the suggested HME model

Expert Network	$\hat{\beta}$ (β_0, β_1)	$\hat{\sigma}^2$	$\hat{g_j}$	$\hat{g_{\ell\|j}}$
$\Pr(y\|x, \theta_{1\|1})$	(0.11, −0.18)	0.077	380(70.96%)	147(27.47%)
$\Pr(y\|x, \theta_{2\|1})$	(0.02, −0.02)	0.041		233(43.49%)
$\Pr(y\|x, \theta_{1\|2})$	(−0.24, 0.21)	0.117	137(25.89%)	111(21.92%)
$\Pr(y\|x, \theta_{1\|2})$	(−0.71, 0.73)	0.122		26(3.97%)

EM-알고리즘의 E-step에서는 ψ에 대한 k번째 $\psi^{(k)}$를 이용하여 얻어진 조건부 기대값 $\tau_{j\ell}^{(t)}$에 의해 완전자료 로그우도함수에서의 잠재변수 $z_{j\ell}^{(t)}$를 대체한다. 조건부 기대값 $\tau_{j\ell}^{(t)}$는 아래와 같다.

$$\tau_{j\ell}^{(t)} = \Pr_{\psi}\left\{z_{j\ell}^{(t)} = 1 | y^{(t)}, x^{(t)}\right\} = \frac{g_j g_{\ell|j} \Pr\left(y^{(t)}|x^{(t)}, \theta_{j\ell}^{(k)}\right)}{\sum_j g_j \sum_\ell g_{\ell|j} \Pr\left(y^{(k)}|x^{(k)}, \theta_{j\ell}^{(k)}\right)}.$$

M-step은 세 가지의 최대화 문제로 구성되어지며, 업데이트된 k번째 $g_j^{(k+1)}$, $g_{\ell|j}^{(k+1)}$, $\theta_{j\ell}^{(k+1)}$는 아래와 같이 풀 수 있다.

$$\sum_t \sum_j \frac{\tau_j^{(t)} \partial \ln g_j^{(t)}}{\partial \gamma_j} = 0$$

$$\sum_t \sum_j \sum_\ell \frac{\tau_{j\ell}^{(t)} \partial \ln g_{\ell|j}^{(t)}}{\partial \gamma_{j\ell}} = 0$$

$$\sum_t \frac{\tau_{j\ell}^{(t)} \partial \ln \Pr\left(y^{(t)}|x^{(t)}\right)}{\partial \theta_{j\ell}} = 0.$$

4. 혈청 섬유아세포에 대한 제안된 HME 모형의 적용

본 절에서는 시간 경로에 대한 실제 마이크로어레이 실험자료에 대하여 제안된 HME 모형을 적용하여 그 결과를 제시한다. 자료에 대한 분석은 R-package 2.14.0을 이용하였으며, Pinheiro와 Bates (2009)가 제안한 nlme library를 이용하였다.

분석에 이용한 자료는 Iyer 등 (1999)이 분석했던 것으로서, 혈청에 관한 섬유아세포(fibroblast)에 관한 생리적 반응을 12시간 동안 관찰한 cDNA 마이크로어레이 자료이다. 원 자료분석 결과는 혈청에 관한 섬유아세포 자극 후 12시점 동안 8613개의 유전자들을 가지고 관찰하여, 혈청자극에 반응하는 유전자 517개를 발견하였다. 그러나, 본 연구에서는 위의 자료를 처음시점부터 4번째 시점까지로 제한하였고, 유의하다고 발견된 517개의 유전자를 이용하여 제안된 HME 모형을 사용하였다. 4개의 시점으로 제한한 이유는 시간에 따른 유전자 발현을 선형적인 변인으로 고려하였기 때문이다. 만약 모든 시점에 대하여 고려할 경우 선형적인 변인이 아닌 비선형모형으로 적용해야 하므로 선형혼합모형을 사용할 수 없기 때문이다.

HME 모형의 설정은 상위 입력네트워크의 수준이 2개이고, 하위 입력네트워크의 수준이 각 2개씩 가지는 엑스퍼트 네트워크를 가지도록 고려하였다. 이와 같이 고려한 이유는 본 연구에서 이용하는 자료는 기존 연구에서 이미 유의하다고 발견된 유전자 만을 이용하였기 때문이다. 제안된 HME 모형을 적용한 결과는 Table 4.1과 같다.

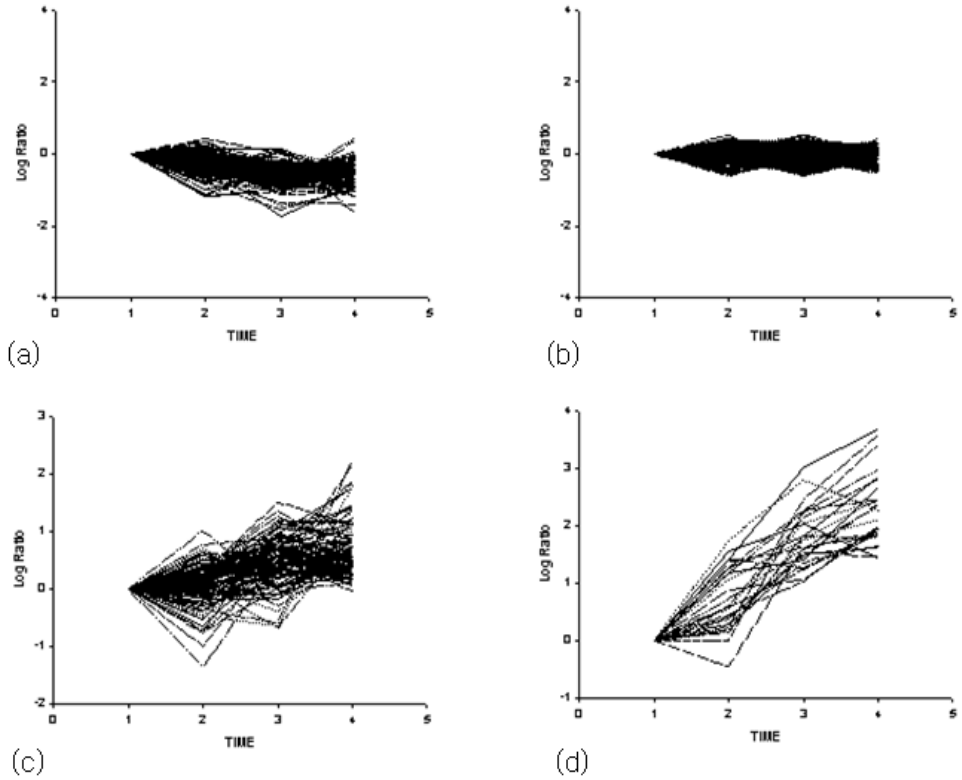

Figure 4.1. Gene expression profiles of the expert networks for human fibroblast

Table 4.2. Distribution of the 517 cell-cycle regulated genes of two clusters defined by Iyer *at al.* (1999) over the four estimated experts using the proposed method

	Expert 1|1	Expert 1|2	Expert 2|1	Expert 2|2
Cluster 1	111(75.51%)	172(73.81%)	57(51.35%)	4(15.38%)
Cluster 2	36(24.46%)	61(26.19%)	54(48.65%)	22(84.62%)
Total	147	233	111	26

Table 4.1의 결과를 살펴보면, 엑스퍼트 1|1의 경우 절편이 양이면서 감소하는 추세를 가지고, 엑스퍼트 2|1인 경우 1|1 보다는 절편이 작으면서 완만하게 감소하는 추세를 가진다. 엑스퍼트 1|2, 2|2인 경우 음의 절편을 가지면서 증가하는 추세를 가지는데, 엑스퍼트 1|2 보다는 2|2가 더 강한 양의 증가추세가 존재함을 확인 할 수 있다.

각 엑스퍼트 네트워크들에 해당되는 유전자들을 도식화 하면 Figure 4.1과 같다. Figure 4.1에서 위의 Table 4.1에 대한 엑스퍼트 네트워크들의 추정치와 이에 대한 추세를 확인 할 수 있다. Figure 4.1의 (a), (b)는 감소하는 경향성을 가지며, (c), (d)는 증가하는 경향성이 존재한다. 또한, (b)보다는 (a)가 더욱 감소하는 추세를 가지며, (c)보다는 (d)가 더욱 증가하는 추세를 가진다는 것을 알 수 있다.

Iyer 등 (1999)은 계층적 군집분석(hierarchical clustering analysis)를 통하여 본 517 유전자들에 대하여 2개의 군집을 정의하였다. 직접적인 비교는 시점이 다르기 때문에 고려할 수는 없지만, Iyer 등에 의

해 제안된 2개의 군집과 본 연구에서 제안된 HME 모형을 적용시켰을 때의 분포는 Table 4.2와 같다.

Table 4.2에서 '군집 1/2'는 Iyer 등이 제시한 계층적 군집분석의 결과로서, '군집 1'은 시간에 따라 발현이 감소하는 집단이며, '군집 2'는 반대로 발현이 시간에 따라 증가하는 경향을 나나내는 군집이었다. 제안된 HME 모형과 비교한 결과 엑스퍼트 1|1, 1|2가 '군집 1'에 집중되어 있으며, 엑스퍼트 2|1은 '군집 1', '군집 2'에 비슷하게 분포해 있으나 엑스퍼트 2|2는 '군집 2'에 거의 집중되어 있음을 알 수 있다.

5. 결론

본 연구에서는 마이크로어레이 자료가 시간에 따라 반복되는 형태를 가질 때, 이에 따른 각 유전자의 군집과 이에 따른 선형추세를 기존의 HME 모형의 엑스퍼트 네트워크 모형을 혼합모형으로 수정하여 EM-알고리즘을 이용하여 추정하는 방법을 제안하였다. 제안된 모형에 대하여 Iyer 등 (1999)이 제안한 혈청 섬유아세포에 대한 cDNA 마이크로어레이 자료를 적용함으로서 본 방법의 유용성을 확인하였다.

이를 기초로 하여 앞으로 몇 가지의 확장된 접근이 가능하다고 판단되는데, 첫째로 공변수(covariate)를 시점으로 고려하는 것이 아닌 처리(treatment), 혹은 기타 제반 변수들을 고려할 수 있다. 둘째로, 임의효과를 단순히 유전자간으로 고려하는 것이 아닌 Kerr와 Churchill (2001) 등이 제시한 어레이(array)간, 혹은 염색(dye)효과들도 고려하여 적용할 수 있으며, 셋째로, 실제 microarray 자료에서는 시간에 따른 반복측정자료는 일반적으로 세포주기(cell-cycle)자료들이 주가 되는데, 이러한 세포주기자료들은 시점이 적어도 10시점 이상이라는 것이다. 이렇게 시점이 많은 경우에 대한 자료의 선형성은 존재하지 않을 가능성이 높기 때문에, 추후 연구에서는 선형성만을 고려하는 것이 아닌 곡선의 형태를 적용할 수 있는 이차형태(quadratic form) 등의 비선형 문제로 확장시켜야 할 것이며, 계산적(computational)인 측면에서 보다 효율적이고 일반화될 수 있는 구체적인 방법을 제안하여야 할 것이다.

References

Bridle J. (1989). *Probabilistic Interpretation of Feedforward Classification Network Outputs, with Relationships to Statistical Pattern Reconition*, In Neurocomputing: Algorithms, Architectures, and Applications, Springer

Brown, M., Grundy, W., Lin, D., Cristianini, N., Sugnet, C., Furey, T., Ares, M. and Haussler, D. (2000). Knowledge-based analysis of microarray gene expression data by using support vector machines, *Proceedings of the National Academy of Sciences of the United States of America*, **97**, 262–267.

Costa, I. G., Carvalho, F. and Souto, M. (2004). Comparative analysis of clustering methods for gene expression time course data, *Genetics and Molecular Biology*, **27**, 623–631.

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm, *Journal of Royal Statistical Society Series B*, **39**, 1–38.

Draghici, S. (2003). *Data Analysis Tools for DNA Microarrays*, Chapman & Hall.

Eisen, M. B., Spellman, P. T., Brown, P. O. and Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns, *Proceedings of the National Academy of Sciences of the United States of America*, **95**, 14863–14868.

Hartuv, E., Schmitt, A., Lange, J., Meirer-Ewert, S., Lehrach, H. and Shamir, R. (1999). An algorithm for clustering cDNAs for gene expression analysis, *IN RECOMB99: Proceedings of the Third Annual International Conference on Computational Molecular Biology*, Lyon, France, 188–197.

Hastie, T., Tibshirani, R. and Friedman, J. (2001). *The Elements of Statistical Learning*, Springer.

Iyer, V. R., Eisen, M. B., Ross, D. T., Schuler, G., Moore, T., Lee, J. C., Trent, J. M., Staudt, L., Hudson, J., Boguski, M., Lashkari, D., Shalon, D., Botstein, D. and Brown, P. O. (1999). The transcriptional program in the response of human fibroblasts to serum, *Science*, **283**, 83–87.

Jordan, M. I. and Jacobs, R. A. (1992). Hierarchies of adaptive experts, *Advances in Neural Information Processing Systems*, **4**, 985–993.

Jordan, M. I. and Jacobs, R. A. (1994). Hierarchical mixtures of experts and the EM algorithm, *Neural Computation*, **6**, 181–214.

Kerr, M. K. and Churchill G. A. (2001). Experimental design for gene expression microarrays, *Biostatistics*, **2**, 183–201.

Laird, N. M. and Ware, J. H. (1982). Random effect models for longitudinal data, *Biometrics*, **38**, 963–974.

Lander, E. S. (1999). Array of hope, *Nature Genetics*, **21**, 3–4.

Little, R. J. and Rubin, D. B. (2002). *Statistical Analysis with Missing Data*, Wiley.

Luan, Y. and Li, H. (2003). Clustering of time-course gene expression data using a mixed-effects model with B-splines, *Bioinformatics*, **19**, 474–482.

McCullagh, P. and Nelder, J. A. (1983). *Generalized Linear Models*, Chapman & Hall, London.

McLachlan, G. J. (2008). *The EM Algorithm and Extensions*, Wiley.

Pinheiro, J. and Bates, D. (2009). *Mixed-Effects Models in S and S-PLUS 2nd Ed.*, Springer.

Quackenbush, J. (2001). Computational analysis of cDNA microarray data, *Nature Review Genetics*, **6**, 418–428.

Schlattmann, P. (2009). *Medical Applications of Finite Mixture Models*, Springer.

Slonim, D. (2002). From patterns to pathways: Gene Expression data analysis come of age, *Nature Genetics*, **32**, 502–508.

Storey, J. D., Xiao, W., Leek, J., Tomkins, R. G. and Davis, R. W. (2005). Significance analysis of time course microarray experiments, *Preceedings of the National Academy of Sciences*, **102**, 12837–12842.

Tamayo, P., Slonim, D., Mesirov, J., Zhu, Q., Kitareewan, S., Dmitrovsky, E., Lander, E. S. and Goulb, T. R. (1999). Interpreting patterns of gene expression with self-organizing maps: Methods and application to hematopoietic differentitation, *Proceedings of the National Academy of Sciences of the United States of America*, **96**, 2907–2912.

Tavazoie, S., Huges, J. D., Campbell, M. J., Cho, R. J. and Church, G. M. (1999). Systematic determination of genetic network architecture, *Nature Genetics*, **22**, 281–285.

Wang, L., Chen, X., Wolfinger, R. D., Franklin, J. L., Coffey, R. J. and Zhang, B. (2009). A unified mixed effects model for gene set analysis of time course microarray experiments, *Statistical Applications in Genetics and Molecular Biology*, **8**, Article 47.

Yeung, K. Y., Fraley, C., Murua, A., Raftery, A. E. and Ruzzo, W. L. (2001). Model-based clustering and data transformations for gene expression data, *Bioinformatics*, **17**, 977–987.

Yeung, K. Y., Medvedovic, M. and Bumgarner, R. E. (2003). Clustering gene-expression data with repeated measurements, *Genome Biology*, **4**, R34.