점진적 주성분 분석을 이용한 시계열 유전자 발현 데이터의 효율적인 차원 축소

김선희*, 김만선, 양형정
*전남대학교 전산학과
e-mail : wkdal749@hanmail.net

Dimension Reduction in Time-series Gene Expression Data using incremental PCA

Sun-Hee Kim*, Man-Sun Kim*, Hyung-Jeong Yang*
*Dept. of Computer Science, Chonnam University

요 약

최근 생명 공학 기술의 발달로 마이크로 단위의 실험이 가능해지고 하나의 칩상에 수 만개의 유전자들의 발현 양상을 보다 쉽게 관찰할 수 있게 되었다. DNA 칩 기술에 의해 얻어지는 마이크로어레이(microarray) 데이터는 세포나 조직 내의 유전자 발현도(expression level)를 측정한 것으로 질병 진단이나 유전자 기능 예측 등에 이용되고 있다. 본 논문에서는 대량의 시계열 마이크로어레이 데이터 분석을 위해 효율적으로 데이터의 차원을 판단하는 점진적 주성분 분석을 이용하여 데이터의 차원을 축소 한다. 제안된 방법은 실제 시계열 마이크로어레이 데이터인 yeast cell cycle 데이터에 적용되었고, 데이터 차원 축소에 대한 효율성을 검증하기 위해 클러스터링을 수행하였다. 그 결과 데이터를 축소하여 클러스터링을 수행한 경우 학습 성능이 향상 된 결과를 보였다.

1. 서론

유전자는 DNA 분자 중 생명 활동에 의미를 갖는 부분으로 복잡한 생명 현상을 조절하는 역할을 하며 발현 양상은 세포의 형질(phenotype)과 관계가 있다. 한 생명체의 모든 세포는 같은 DNA 를 갖고 있지만, 세포 종류에 따라 발현 되는 유전자가 다르기 때문에 서로 다른 형태와 기능을 갖는다. DNA 마이크로어레이 기술은 유전자 정보들을 분석하기 위한 기술로 대량의 유전자 발현 정보를 만들어 내게 되고, 여기에서 측정된 유전자 발현 데이터를 마이크로어레이 데이터라 한다. 마이크로어레이 데이터는 시간에 관계없이 특정 시점의 발현 정도가 측정되어 분석 된다. 지금까지 마이크로어레이 데이터는 유전자 발현에 관련된 연구를 위해 여러 분야에서 분석 목적에 따라 다양한 기법들이 적용 되어왔다. 분석을 위한 대표적인 알고리즘으로는 계층적 클러스터링(hierarchical clustering)[11], SOM(Self-Organizing Map)[14][15], 베이지안망(Bayesian network)[8], PCA(Principal Component Analysis) [9] 등이 있다.

이에 비해 시계열 마이크로어레이 데이터는 여러 시점 동안 시간의 경과에 따라 유전자 발현 정도를 측정한 것이다. 시계열 마이크로어레이 데이터는 시간의 흐름에 따라 얻어진 데이터 이지만 연속적인 흐름의 데이터가 아닌 특정 시간의 발현 정도를 측정한 데이터다.

따라서 시계열 마이크로어레이 데이터는 관찰된 데이터가 이전 데이터와 서로 연관되어 있다는 특성을 가지므로 지금까지의 일반적인 분석 방법으로는 시계열 데이터의 특성을 고려하지 못한다. 또한 시계열 마이크로어레이 데이터의 속성에 해당하는 유전자는 그 수가 수만 개 혹은 수십만 개에 달하는데 비해, 실험의 비싼 비용 및 샘플의 한정된 용량으로 인해 표본(instance)에 해당하는 슬라이드 개수는 매우 작다는 특징이자 한계점을 가지고 있다. 따라서 표본 개수에 비해 너무 많은 속성, 즉 유전자들을 분석 목적에 맞는 개수로 줄여 주는 작업이 수행되어야 하며, 시계열 데이터의 특성을 고려할 수 있는 전처리(preprocessing) 작업이 필요하다.

본 논문에서는 데이터 분석의 전처리 과정으로 데이터의 효율적인 차원을 판단하기 위해 PCA 를 시간적 특성을 고려하여 확장한 점진적인 PCA 를 이용하여 데이터의 차원을 축소한다. 또한 제안된 방법의 효율성을 검증하기 위해 데이터 클러스터링을 수행하고 그 결과를 비교 분석한다.

논문의 구성은 다음과 같다. 2 장에서는 기존의 유전자 발현 분석에 대한 연구를 살펴보고, 3 장에서는

본 논문에서 사용하는 점진적 주성분 분석에 대해 기술한다. 4 장에서는 실험에 사용된 데이터에 대한 설명과 차원 축소에 대한 군집화 실험 및 결과를 보인다. 5 장에서는 결론 및 향후 연구에 대하여 기술한다.

2. 관련 연구

본 논문에서 관련연구는 마이크로어레이 데이터의 특성을 살리면서 분석하기 쉬운 데이터로 변환하는 과정으로 데이터 전 처리 부분과 유전자 발현 분석 부분으로 나눌 수 있다.

Bar-Joseph[3], Storey[10]는 전처리 방법으로 이산적인 데이터를 연속적인 데이터로 변환할 수 있는 spline 를 이용하여 결측치를 예상하였다. Ramoni[7]는 전처리 과정으로 AR(q)모델을 사용하였으며, Aach[1]에서는 Warping 알고리즘을 이용하여 유전자 발현 정도를 정렬하였다. schlep[2]의 경우 Hidden Markov Model(HMMs)을 이용하여 시간 축을 정립하고 결측치를 예측하였다. 또한 Amato[19]는 전처리 과정으로 먼저 t-test 를 이용하여 유의하지 않은 유전자를 걸러낸 후 multi nominal mixture 모델을 응용한 비선형 PCA 방법을 이용하여 결측치를 추정하였다.

군집 분석은 보편적으로 마이크로어레이 데이터 분석에 많이 사용되고 있는 분석 방법이다.

Wen[17]는 유클리드 거리행렬을 이용하여 유전자를 분석하였으며, Tamoya[14], Toronena[15]는 자기조직도(SOM)을 이용하여 유전자를 분석하였다. 그러나 자기 조직도는 유전자 발현 데이터에서 위상적 의미 해석에 대한 문제가 있었다. Brown[16]는 support vector machine(SVM)을 이용한 클러스터링 방법을 제시하기도 하였으며, Spell[11]는 Saccharomyces cerevisiae 의 마이크로어레이 데이터에 계층적 클러스터링을 이용하였다.

Tai[13]는 반복이 있는 시계열 데이터에서 유의한 유전자를 선별하는데 이용되는 통계적 방법인 다변량 empirical Bayesian 통계기법과 Hotelling T2 소개하였다. Moller-Leveta[5]는 fuzzy short time series(FSTS) 클러스터링 분석을 제안하고 시계열 분석을 위해 short time-series(STS) 거리 척도를 제안하였다.

일반적으로 클러스터링 분석 방법에서는 서로 다른 조건에서 얻어진 유전자는 각각 독립적이라고 가정하고 분석을 진행한다. 그러나 시계열 데이터에서는 한 시점에서의 관찰 값이 이전 시점에서의 관찰 값과 연관되어 있기 때문에 시계열 데이터를 고려하지 않고 일반적인 분석 과정을 그대로 사용하는 것은 올바른 결과를 얻을 수 없게 된다[4].

따라서 시계열 데이터를 분석하기 위해서는 데이터의 시간적 특성을 고려한 클러스터링 분석이 진행 되어야 한다.

본 논문에서는 시계열 마이크로어레이 데이터에서 데이터의 성질을 고려하고 유의한 유전자를 판별하여 분석하기 쉬운 형태의 데이터로 변환하기 위한 전처리 과정으로 점진적 주성분 분석을 적용한다.

3. 점진적 주성분 분석을 이용한 데이터 차원 축소

시계열 마이크로어레이 데이터의 효율적인 분석을 위해 주어진 데이터의 정보를 잘 표현하면서 정보량을 줄일 수 있는 주성분 분석을 점진적인 방법으로 확장한 점진적 주성분 분석[6][12]을 이용한다. 시계열 유전자 발현 데이터에서 행의 값은 발현 유전자를 나타내며, 열의 경우는 각각의 측정 시점을 나타낸다. 따라서 특징 벡터는 수천 개의 발현 유전자를 의미한다. 본 논문에서는 점진적 주성분 분석을 이용하여 시계열 데이터 차원 축소를 위해 마이크로어레이 데이터를 n-차원의 특징 벡터 $\boldsymbol{G} = (\boldsymbol{g}_1, \boldsymbol{g}_2, ..., \boldsymbol{g}_n)$ 로 표현한다.

모든 특징 벡터는 n-차원 특징 벡터 공간 내에 점으로 표현되며 벡터 공간에서 모든 특징 벡터들은 n-차원 벡터들의 집합에 의해 채워 지며 벡터 공간에서 각각의 축 방향을 가리키는 단위 벡터들을 기저 벡터 $\nu = \{\vec{\nu_1}, \vec{\nu_2}, ..., \vec{\nu_n}\}$ 로 나타낸다. 즉, 특징 벡터 공간에 모든 특징 벡터들이 위치하는 것을 식 1 과 같이 기저벡터의 선형 조합 형태로 표현할 수 있다.

$$\boldsymbol{G} = \boldsymbol{c}_1 \vec{\nu_1} + \boldsymbol{c}_2 \vec{\nu_2} + ,..... + \boldsymbol{c}_n \vec{\nu_n} \quad (1)$$

식 1 에서 $\boldsymbol{c}_i$ 는 기저벡터와 관계가 있는 $\boldsymbol{G}$ 의 좌표이다. 일반적으로 기저벡터는 표준 기저를 사용한다. 그러나 시계열 마이크로어레이 데이터의 경우 일정한 측정 시점에 따라 측정된 데이터이므로 데이터를 표현하는데 표준기저는 적절하지 못하다. 따라서 시계열 데이터의 측정 시점에 따른 특성을 잘 표현 하기 위해 새로운 기저 값이 요구 된다.

주성분 분석은 주어진 데이터의 정보를 잘 표현하면서 정보량을 줄일 수 있는 효과적인 방법 중 하나이다. 주성분 분석[9]은 n 차원을 갖는 데이터 $\boldsymbol{G}$ 의 고유값(eigenvalue)과 고유벡터(eigenvector) 행렬을 얻어 이를 기반으로 주성분을 찾고 원래 데이터를 더 잘 나타내는 새로운 축을 구성한다.

본 논문에서는 새로운 기저 값을 찾기 위해 원래의 고차원 데이터를 투영하는 기저 벡터 ν 를 점진적으로 갱신한다. 즉 주성분 구성요소 분석을 기반으로 주성분 구성요소를 찾고, 평균 재현 에러율(average reconstruction error)을 최소화하는 주성분 구성요소의 기저벡터 $\nu_{i,j}$ (i = 1…p, j = 1…d)를 점진적으로 갱신하면서 새로운 기저 벡터를 찾는 점진적 주성분 분석을 적용한다.

점진적 주성분 분석은 먼저 $\boldsymbol{p}$ 개의 기저 벡터를 필요로 하며, 주성분 구성요소 $\vec{\nu_i} = [\nu_{i,1}, ..., \nu_{i,d}]^T$ (i=1...p)를 가지고 원소 $\nu_{i,j}$ 의 값을 갱신 한다. 본 논문에서는 주성분 구성요소의 벡터는 $\vec{p}$ 로 표기하며, $\boldsymbol{p}$ 는 주성분 구성요소이다. 이 후 새로운 공분산 행렬은 $\boldsymbol{d} \times \boldsymbol{p}$ 크기를 갖는다.

$$\sum_k \| (\tilde{\boldsymbol{G}}_k - \boldsymbol{G}_k) \| \quad (2)$$

위의 식(2)은 평균 재현 에러율을 의미하며, $\tilde{G}_k$ 는 p 차원 공간에 G_k 을 투영한 후에 재구성된 점을 의미한다. 또한 m 개의 샘플을 가지고 p 차원 공간에 G_k 를 투영한 것을 a_k 라고 가정할 때, a_k 는 아래와 같이 나타낸다.

$$a_{k,i} = \sum_{k=1}^{m} v_{i,d} \times g_{k,d} \quad (i=1\cdots p,\ k=1\cdots .m) \qquad (3)$$

$$\tilde{G}_{k,i} = \sum_{k=1}^{m} v_{p,j} \times a_{k,p} \quad (j=1\cdots d) \qquad (4)$$

즉, $a_{k,i}$ 는 m 개의 샘플들에 각각의 주성분 구성요소를 곱한 전체의 합을 의미하며, G_k 의 재구성은 식 4 와 같이 표현한다.

시간 변화에 따른 새로운 샘플 G_{k+1} 이 주어졌을 때 그림 1 과 같이 3 단계를 수행하여 점진적으로 주성분의 구성요소를 변경한다.

(1) 현재 주성분 구성 요소 $\vec{v}_i$ (i=1…p)에 G_{k+1} 을 투영함으로써 a'_{k+1} 을 계산한다.
(2) 재구성 에러를 평가한다.
(3) $\vec{v}_i$ 를 평가하여 갱신하고, 샘플 G_{k+1} 을 위해 실제 투영된 점으로 새로운 투영 a_{k+1} 를 계산한다.

(그림 1) 점진적 주성분 분석

주성분 구성요소 벡터를 갱신할 때, 각 주성분 구성요소 $\vec{v}_i$ 를 위한 갱신의 중요도는 샘플의 개수가 k 인 경우 식 (5)과 같이 나타낸다.

$$M_{k,i} = \frac{1}{k}\sum_{t=1}^{k} a_{t,i}^2 \qquad (5)$$

4. 실험 및 결과

데이터는 Spellman[11]에서 공개하고 있는 Saccharomyces cerevisiae 의 세포주기 관련 시계열 유전자 발현 데이터이다. 이 데이터는 각각의 측정 시점에 따른 유전자의 발현 정도를 수치 데이터로 표현하고 있다. 표 1 에서 Alpha 발현 데이터는 페로몬을 이용하여 세포들을 G1 기에 정지(arrest) 시켰다가 해제(release) 시킨 후 매 7 분마다 유전자들의 발현 양상을 측정한 것으로 6000 개 이상의 효모 유전자들에 대해 총 18 개의 시점(time point)에서의 측정치를 담고 있다. Cdc15 발현 데이터는 성장 온도를 변화시켜 측정한 것이며, Elu 데이터는 elutriation 에 의해 발현된 양상을 측정한 데이터 이다. 본 논문에서는 주어진 데이터 중 70%를 학습데이터로 나머지 30%를 테스트 데이터로 나누어 실험하였다.

<표 1> spellman 의 데이터

실험의 명칭	측정간격	측정횟수
Alpha	0분부터 119분까지 매 7분	18번
Elu	0분부터 390분까지 매 30분	14번
Cdc15	10분부터 290분가지 매 10분(중간에 측정하지 않는 시각이 있음)	24번

본 논문에서는 점진적 주성분 분석을 이용하여 데이터의 차원을 축소하였다. 그 결과 데이터 평균 재현 에러률을 최소화 하는 주성분 구성요소의 수를 표 2 와 같이 생성하였다.

<표 2> 주성분 구성요소 개수에 따른 재현 에러율

	PC 개수						
데이터	1000	900	800	700	500	100	50
Alpha	0.57	0.07	**0.04**	0.15	0.11	0.14	1.25
Cdc15	0.15	**0.11**	0.26	0.26	0.39	0.50	2.53
Elu	0.17	**0.09**	0.18	0.20	0.22	0.20	2.44

Alpha 데이터의 경우 주성분 구성요소의 수가 800 개 일 때 재현 에러율이 가장 낮은 값을 나타냈으며, Cdc15 와 Elu 데이터의 경우에는 주성분 구성요소의 수가 900 개 일 경우 가장 낮은 평균 재현 에러율을 나타냈다. 따라서 재현 에러율을 최소로 하는 주성분 구성요소를 이용하여 데이터의 차원을 축소하였고, 데이터의 차원 축소에 관한 효율성을 검증하기 위하여 계층적 클러스터링과 k-means 클러스터링을 수행하였다.

계층적 클러스터링[11]은 처음에 각각의 데이터 점을 하나의 클러스터로 설정한 후 이들 쌍 간의 거리를 기반으로 하여 분할(top down), 합병해 나가는 상향식(bottom-up) 방식이다. 군집간의 거리를 측정하는 방법에는 6 가지 측도가 있으나 본 실험에는 각 클러스터 내에 포함된 객체들 사이의 거리 중 최대 거리를 두 클러스터의 거리로 보는 계층적 클러스터링을 수행하였으며, 계층적 클러스터링은 로그-우도(log-likelihood) 값이 가장 높은 값을 가질 때 클러스터의 성능이 높다고 판단한다.

K-means(K 평균) 클러스터링[18]은 거리에 기반을 둔 클러스터링 방법으로 가까운 곳에 있는 데이터들끼리 클러스터링을 하는 방법이다. K-means 알고리즘은 제곱오차의 합을 최소화 하도록 k 개 분할한다.

실험 결과 표 3 에서와 같이 계층적 클러스터링을 적용한 Alpha 데이터의 경우 클러스터의 수가 500 이면서 투영된 데이터의 경우 로그 우도 값이 원래 데이터의 2 배정도 높게 나타난 것을 확인할 수 있었다. Cdc15 데이터는 원래 데이터의 52 배 정도 높았으며, Elu 의 경우에는 2 배 정도 높은 로그 우드 값을 나타냈다. 표 3 에서 OD 는 원래 데이터를 의미하며, PD 는 점진적 PCA 에 의해 축소된 데이터, RD 는 주성분 구성요소에 의해 재구성된 데이터를 의미한다.

<표 3> 계층적 클러스터링의 로그-우도(Log likelihood) 값

계층적		클러스터 개수 10	클러스터 개수 100	클러스터 개수 500
OD	Alpha	1.99188	3.96382	6.24472
	Cdc15	-5.46022	-2.0411	0.32204
	Elu	-0.2637	-1.63202	3.31896
PD	Alpha	-3.04051	3.48401	**15.33852**

	Cdc15	-16.46023	-1.7538	**15.67119**
	Elu	-4.0831	0.63401	**7.1368**
RD	Alpha	-12.05245	-10.08526	-8.09562
	Cdc15	-25.36644	-21.75704	-18.84018
	Elu	-11.13955	-9.66488	-8.08191

<표 4> k-means 클러스터링의 제곱 오차의 합(squared errors)

K-means		클러스터 개수 10		클러스터 개수 100		클러스터 개수 500	
		반복수	제곱 오차의 합	반복수	제곱 오차의 합	반복수	제곱 오차의 합
OD	alpha	91	337.22	44	198.42	16	134.60
	cdc	58	522.72	39	341.26	21	239.19
	elu	39	285.69	34	171.21	23	113.63
PD	alpha	17	11.83	12	3.421	5	**1.99**
	cdc	15	17.60	13	3.265	7	**3.149**
	elu	45	4.217	14	2.437	5	**0.184**
RD	alpha	135	38.66	51	18.82	25	12.69
	cdc	48	53.54	47	25.38	21	17.77
	elu	56	17.090	57	9.061	22	5.770

표 4 에서는 k-means 클러스터링을 데이터에 적용한 경우 클러스터의 수가 500 일 때 원래 데이터의 제곱 오차의 합이 투영된 데이터 보다 높은 값을 나타냈으며, 재구성된 데이터도 원래 데이터 보다 낮은 제곱 오차 합의 값을 나타냈다. 위의 결과와 같이 데이터의 차원을 축소하여 클러스터링을 수행한 결과 원래의 데이터를 가지고 클러스터링을 수행하였을 때 보다 더 좋은 군집 결과를 얻을 수 있었다.

5. 결론

마이크로어레이를 통해서 얻은 데이터는 적게는 수천에서 많게는 수만 개의 유전자 데이터를 가지고 있다. 이러한 유전자들의 기능이나 특징 등 정확한 정보를 밝혀내기 위해서는 먼저 유의한 유전자를 골라내야 한다. 따라서 본 논문에서는 시계열 마이크로어레이 데이터의 기능 분석과 유전자 상호 관련성 등의 효율적인 분석을 위한 전처리 과정으로 시간적 특성을 고려하는 점진적인 방법으로 데이터의 차원을 축소하였다. 점진적 주성분 분석을 통해 데이터의 차원을 축소함으로써 저장공간의 필요량을 줄일 수 있었고, 데이터 차원 축소에 대한 효율성을 증명하기 위해 클러스터링을 수행 한 결과 군집학습 성능을 향상시킬 수 있었다. 차후 본 논문의 연구 내용을 기반으로 시계열 유전자 발현 데이터의 전처리에 관한 유사 연구 비교와 시계열 데이터의 속성을 고려한 복합적인 성능 평가 기준에 대한 연구가 필요하다.

참고문헌

[1] Aach, J. and Church, G. M, “Aligning gene expression time series with time warping algorithms,” Bioinformatics, vol.17, pp. 495-508, 2001.

[2] Alexander S, Alexander S, Christine S, “Using hidden Markov models to analyze gene expression time course data,” Bioinformatics,vol. 9, pp.255-263, 2003.

[3] Bar-Joseph, Z, et al, “Comparing the continuous representation of time-series expression profiles to identify differentially expressed genes,” Proceedings of the National Academy of Sciences, vol.100, pp.10146-10151, 2003.

[4] Chen, G. and Dai, Y., “A New Distance Measurement for Clustering Time-Course Gene Expression Data,” Proceedings of the 26th Annual International Conference of the IEEE EMBS, 2004.

[5] Moller-Leveta, et al, “Clustering of unevenly sampled gene expression time-series data,” Fuzzy Sets and Systems, pp. 49-66, 2005.

[6] Minh-Tuan. T.H, Yonggwan. W, Hyung-Jeong. Y, “Cognitive States Detection in fMRI using incremental PCA,” ICCSA , pp. 5-11, 2007.

[7] Marco F. R, Paola S, Isaac S. K, “Cluster analysis of gene expression dynamics,” Proceedings of the National Academy of Sciences, vol.99, pp.9121-9126, 2002.

[8] Nir F, et al, “Using Bayesian Networks to Analyze Expression Data,” A technical report describing this work. Submitted to RECOMB, 2000.

[9] Raychaudhuri. S., Stuart. J.M, Altman. R.B, “Principal components analysis to summarize microarray experiments: application to sporulation time series,” Pacific Symposium on Biocomuting 5(Proceeding of PSB’00), pp. 452-463, 1999.

[10] Storey, J. D, et al, “Significance analysis of time course microarray experiments,” Proceedings of the National Academy of Sciences, vol.102, pp.12837-12842, 2005.

[11] Spellman. PT, et al, “Comprehensive identification of cell cycle-regulated genes of the yesast Saccharomyces cerevisiae by microarray hybridization,” Mloecular Biology of the Cell, vol.9, pp.3273-3298, 1998.

[12] Spiros. P, Jimeng. S, Christos. F, “Streaming pattern discovery in multiple time-series,” In Proceeding of the 31st VLDB Conference, pp.697–708, 2005.

[13] Tai, Y. C, Speed, T. P, “Multivariate Empirical Bayes Statistic for Replicated Microarray Time Course Data,” Annals of Statistics, vol.34, pp.2387-2412, 2006.

[14] Tamayo. P, et al, “Interpreting patterns of gene expression with self-organizing maps: Methods and application to hematopoietic differentiation,” Proceedings of the National Academy of Sciences,vol.96, pp.2907-2912, 1999.

[15] Törönen P, et al, “Analysis of gene expression data using self-organizing maps,” FEBS Letters, vol.451, pp.142-146, 1999.

[16] Vrown MP, et al, “Knowledge based analysis of microarray gene expression data by using support vector machines,” Proceedings of the National Academy of Sciences, pp.262-267, 2000.

[17] Xiling. W, et al, “Large-scale temporal gene expression mapping of central nervous system development,” Proc. Natl Acad. Sci. USA, vol. 95, pp.34–339, 1998.

[18] Yano.N, Kotani. M, “Clustering Gene Expression Data Using Self-organizing Maps and k-means Clustering,” SICE 2003 Annual Conference, Vol.3, pp.3211- 3215, 2003.

[19] Amato. R, et al, “A multi-step approach to time series analysis and gene expression clustering,” Bioinformatics, vol.22, pp.589-596, 2006.