정보병목기법에 기반한 유전자 발현 데이터의 이중 클러스터링

김병희○, 황규백, 장정호, 장병탁
서울대학교 컴퓨터공학부 바이오지능 연구실
{bhkim○, kbhwang, jhchang, btzhang}@bi.snu.ac.kr

Double Clustering of Gene Expression Data Based on the Information Bottleneck Method

Byoung-Hee Kim○, Kyu-Baek Hwang, Jeong-Ho Chang, and Byoung-Tak Zhang
Biointelligence Lab, Dept. of Computer Sci. & Eng., Seoul National University

요 약

기능 유전체학에서 클러스터링 기법은 고차원의 마이크로 어레이 데이터 분석을 위한 주된 도구 중의 하나이다. 본 논문에서는 정보병목(information bottleneck)기법 기반의 이중 클러스터링에 의한, 유전자 발현 데이터의 계층적 병합방식 클러스터링 기법을 제안한다. 정보병목기법은, 두 랜덤변수의 결합확률분포가 주어진 경우 두 변수의 상호 정보량을 최대한 보존하면서 한 변수를 압축하는 기법이며, 두 변수를 차례로 압축하는 것이 이중 클러스터링이다.

실제 마이크로 어레이 데이터인 NCI60 데이터(암세포 내 유전자 발현 데이터)에 대한 실험에서, 먼저 유전자를 그 발현패턴에 따라 클러스터링 한 후 이를 이용하여 표본들을 클러스터링하고 그 성능을 다각도로 분석하였다. 상호 정보량과 유전자 및 표본 클러스터 수와 엔트로피 척도에 의한 성능을 검토해 본 결과, 표본이 추출 조직에 따라 구분 가능할 것이라는 가정을 검증할 수 있었으며, 적절한 클러스터의 수를 결정할 수 있는 임계점의 기준을 설정할 수 있었다.

1. 서 론

DNA 마이크로어레이 데이터는 다수의 조직이나 세포의 수천~수만에 이르는 유전자의 발현도를 측정한 것으로 생물학과 의학 연구에 다양하게 이용되고 있다. 이러한 마이크로어레이 데이터의 분석에는 분석의 목적에 따라 기계학습과 통계학의 다양한 분석기법들이 적용되고 있다. 다양한 분석 기법들 중 가장 널리 이용되고 있는 것은 클러스터링이다. 특히, 조직이나 세포에서의 발현 양상에 의한 유전자의 클러스터링은 유전자의 기능 예측, co-regulation, 단백질사이의 상호작용 등을 밝히는데 매우 유용하게 사용되어 왔다. 이밖에 클러스터링은 다른 분석 방법의 적용을 위한 데이터 전처리, 유전자의 발현 양상에 기반한 조직의 분류 등에도 널리 이용되어 왔다. 대표적인 예로는 계층적 클러스터링, *k*-means 기법, SOM(self-organizing map)을 들 수 있다.

이러한 클러스터링 기법의 적용에 있어서 가장 중요한 것은 유사도 및 거리 측정 기준이다. 일반적으로 널리 이용되는 거리측정 기준은 유클리드 거리(Euclidean distance)와 피어슨 상관계수(Pearson correlation coefficient)이며, 이는 마이크로어레이 데이터의 분석에도 그대로 적용이 되어 왔다. 물론, 이러한 기준들이 유전자 발현이나 조직의 유사도를 측정하는데 가장 적절한지의 여부는 알려져 있지 않다.

본 논문에서는 정보병목기법(information bottleneck method)[5]을 마이크로어레이 데이터의 클러스터링에 적용한다. 정보병목기법은 두 랜덤변수의 결합확률분포가 주어진 경우, 두 변수 간의 상호 정보량을 최대한 보존하면서 한 변수를 압축하는 기법이다. 특히, 이 기법은 피어슨 상관계수로는 측정이 어려운 비선형 관계의 적절한 측정을 원칙적으로 가능케 한다. 실제 클러스터링 알고리즘 구현에 있어서는 [4]의 이중 클러스터링(double clustering) 기법을 이용한다. 이 기법에서는 두 변수를 차례로 정보병목기법을 이용해 압축함으로써 고차원 데이터의 노이즈를 줄이고, 보다 조밀(dense)하고 엄정하며(robust) 데이터에 내재된 구조를 더 잘 반영하는 축약된 자료를 산출한다.

실험에서는 실제 마이크로어레이 데이터인 NCI60데이터에서 표본(sample)에 대한 클러스터링을 실시하였으며, 유전자 클러스터 수의 변화에 따른 엔트로피 및 상호 정보량을 관찰하여 클러스터 결과의 의미를 파악하고, 적절한 클러스터의 수를 결정하기 위한 기준을 살펴보았다.

2. 이론적 배경

2.1 주요 개념 정리

정보병목기법과 이중클러스터링의 핵심이 되는 정보이론의 주요 개념으로, 엔트로피, 상대적 엔트로피, 상호 정보량을 간단히 정리한다. 엔트로피는 랜덤 변수의 불확실성에 대한 척도로서 이산 랜덤 변수 X의 엔트로피 $H(X)$는 다음과 같다.

$$H(X) = -\sum_{x \in X} p(x) \log p(x).$$

상대적 엔트로피는 두 확률분포 사이의 거리에 대한 척도이며, 두 확률질량함수 $p(x)$와 $q(x)$ 사이의 상대적 엔트로피 또는 *Kullback-Leibler divergence*는

$$D_{KL}[p \parallel q] = \sum_{x \subset X} p(x) \log \frac{p(x)}{q(x)}.$$

로 정의한다.

상호 정보량(mutual information)은 한 랜덤 변수가 다른 랜덤 변수에 대해 담고 있는 정보량에 대한 척도이다. 두 랜덤변수 X와 Y의 *상호 정보량* $I(X;Y)$은 결합확률분포 $p(x,y)$와 확률분포의 곱 $p(x)p(y)$ 사이의 상대적 엔트로피이다.

$$I(X;Y) = \sum_{x \in X} \sum_{y \in Y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)}$$
$$= D_{KL}[p(x,y) \parallel p(x)p(y)].$$

2.2 정보 병목기법

대부분의 클러스터링 알고리즘의 출발점이자 결과에 가장 큰 영향을 주는 요소는 데이터를 표현하는 두 점 사이의 '거리'의 쌍 또는 클래스의 중심점과 데이터 위치 사이의 이격도 측정이다.

"분포(distributional) 클러스터링"[1]의 관점에서 이격도 또는 거리 측정에서의 임의 선택에 따른 문제를 피할 수 있는 기법으로 제시된 것이 '정보 병목 기법'[5]이다. 이 접근 방법에서는 두 확률변수 간에 실험적으로 얻어진 결합 확률분포 $p(x,y)$가 주어졌을 때, 관련 변수 Y의 정보를 최대한으로 보유하고 있는 X의 축약된 표현을 찾기 위해 다음 문제에 대한 해답을 제시한다.

<문제> 집합 X 구성원의 클러스터 집합 $\tilde{X}$를 찾아라. 단, $\tilde{X}$는 X에서 추출한 정보에 대한 제한조건 $I(\tilde{X};X)$을 만족하면서, 상호 정보량 $I(\tilde{X};Y)$가 최대가 되도록 해야 한다.

[표 1] 병합식 정보병목 기법 알고리즘의 pseudo-code

입력 : 결합 확률분포 $p(x, y)$
출력 : X를 분할하여 생성한 m개의 클러스터, $\forall m \in \{1...|X|\}$

초기화 :

- 클러스터 초기화. $\tilde{X} \equiv X$
- 병합비용 행렬 초기화
  $\forall i, j = 1...|X|$, $i < j$ 에 대해 다음을 계산
  $d_{i,j} = (p(\tilde{x}_i) + p(\tilde{x}_j)) \cdot D_{JS}[p(y|\tilde{x}_i), p(y|\tilde{x}_j)]$

반복 :

- For $m = |X| - 1...1$
  - $d_{i,j}$를 최소로 하는 인덱스 쌍 $\{i, j\}$를 찾는다.
  - 병합: $(\tilde{x}_i, \tilde{x}_j) \rightarrow \tilde{x}_*$
  - 클러스터 갱신: $\tilde{X} = \{\tilde{X} - \{\tilde{x}_i, \tilde{x}_j\}\} \cup \{\tilde{x}_*\}$
  - 병합비용행렬 갱신: $\tilde{x}_*$와 관련된 $d_{i,j}$를 갱신.
- End For

문제의 해는 동시에 풀어야 하는 세 개의 분포(distribution) 방정식으로 주어지며, hard 클러스터링의 경우는 다음과 같다.

$$\begin{cases} p(\tilde{x}|x) = \begin{cases} 1 & \text{if } x \in \tilde{x} \\ 0 & otherwise \end{cases} \\ p(y|\tilde{x}) = \dfrac{1}{p(\tilde{x})} \sum_{x \in \tilde{x}} p(x)p(y|x) \\ p(\tilde{x}) = \sum_{x \in \tilde{x}} p(x). \end{cases}$$

이 해를 적용한 클러스터링 방식은 계층적 병합식 알고리즘이며 [표 1]에 정리하였다. 이 알고리즘은 X의 원소 각각을 하나의 클러스터로 나눈 상태, 즉 $|X|$ 개의 클러스터에서 시작한다. 각 단계에서 현재 파티션 상에 있는 두 성분(component)을 병합하여 하나의 새로운 성분으로 만드는데, 이 때의 제한조건은 국부적으로(locally) 상호 정보량 $I(\tilde{X};Y)$의 손실을 최소화하는 것이다. 모든 병합과정 $(\tilde{x}_i, \tilde{x}_j) \rightarrow \tilde{x}_*$ 은 다음의 방정식에 의해 명확하게 정의된다.

$$\begin{cases} p(\tilde{x}_*|x) = \begin{cases} 1 & \text{if } x \in \tilde{x}_i \, or \, x \in \tilde{x}_j \\ 0 & otherwise \end{cases} \\ p(y|\tilde{x}_*) = \dfrac{p(\tilde{x}_i)}{p(\tilde{x}_*)} p(y|\tilde{x}_i) + \dfrac{p(\tilde{x}_j)}{p(\tilde{x}_*)} p(y|\tilde{x}_j) \\ p(\tilde{x}_*) = p(\tilde{x}_i) + p(\tilde{x}_j). \end{cases}$$

이 병합 과정에서의 상호 정보량 $I(\tilde{X};Y)$의 감소 정도를 계산하면 다음과 같으며, 이 항목을 "병합에 필요한 비용"으로 해석할 수 있다.

$$\delta I(\tilde{x}_i, \tilde{x}_j) \equiv (p(\tilde{x}_i) + p(\tilde{x}_j)) \cdot D_{JS}[p(y|\tilde{x}_i), p(y|\tilde{x}_j)].$$

각 항목은 다음과 같이 구할 수 있다. 함수 D_{JS}는 Jensen-Shannon (*JS*) divergence이다.

$$D_{JS}[p_i, p_j] = \pi_i D_{KL}[p_i \| \bar{p}] + \pi_j D_{KL}[p_j \| \bar{p}].$$

$$\begin{cases} p_i, p_j \equiv p(y|\tilde{x}_i), p(y|\tilde{x}_j) \\ \pi_i, \pi_j \equiv \dfrac{p(\tilde{x}_i)}{p(\tilde{x}_*)}, \dfrac{p(\tilde{x}_j)}{p(\tilde{x}_*)} \\ \bar{p} = \pi_i p(y|\tilde{x}_i) + \pi_j p(y|\tilde{x}_j) \quad . \end{cases}$$

[표 2] 이중 클러스터링 절차

입력 : 결합확률분포 $p(x, y)$

1단계 :

- $\{p(x|y)\}$를 이용해 클러스터 집합 $\tilde{Y}$를 찾는다.

2단계 :

- 모든 $x \in \tilde{X}$에 대해, $p(y|x)$를 보다 축약된 표현 $p(\tilde{y}|x)$로 대체한다.
- $\{p(\tilde{y}|x)\}$를 이용해 클러스터 집합 $\tilde{X}$를 찾는다.

2.3 이중 클러스터링 기법

정보 병목 기법의 목적은 변환 $p(\tilde{x}|x)$에 의해 정의되는 X의 파티션을 얻어내되, 이 파티션이 상호 정보 함수의 최대값이 되도록 하는 것이다. 이 때, X와 Y는 대칭적인 관계를 가지며. 어떤 변수를 압축해야 되는지에 대한 아무런 전제조건이 없다. 두 변수를 모두 압축하는 방법이 '이중 클러스터링(double clustering)'이며, [표 2]와 같이 두 단계의 클러스터링을 실시한다.

첫 번째 단계에서 두 변수 간의 조건부 확률분포 $p(x|y)$를 구한다. 다음으로, [표 1]의 알고리즘을 이용해 Y의 클러스터 $\tilde{Y}$를 얻는다. 이 때 $|\tilde{Y}| \ll |Y|$ 가 되도록 한다. 두 번째 단계에서는 이 클러스터를 이용해 X의 표현 방식을 바꾼다. 즉, X를 $p(y|x)$를 이용해 표현하는 대신, $p(\tilde{y}|x)$를 이용해 [표 1]의 알고리즘을 다시 적용하여 원하는 문서 클러스터 $\tilde{X}$를 얻는다.

이중 클러스터링을 통해, 고차원의 데이터에서 필수불가결하게 나타나는 노이즈를 상당히 줄일 수 있다. 즉, 다음 식에서와 같이 변수간의 상호 정보량의 손실은 크지 않으면서도 기존 변수들의 차원(dimension)을 둘 모두 상당히 줄일 수 있다.

$$I(\tilde{X};\tilde{Y}) \lesssim I(X;\tilde{Y}) \lesssim I(X;Y).$$

3. 실험 및 결과

3.1 데이터 선정 및 처리

미국 국립 암센터(National Cancer Institute, NCI)에서는 암 치료제 개발을 위해 'NCI60 Cell Lines Data Set'을 구축하였다. 여기에는 9개 조직 60여종의 인간 암세포의 유전자 발현정도를 마이크로어레이를 이용해 생성한 자료가 포함되어 있으며, 이는 60여 표본(sample)에 대해 9,706개 유전자의 발현정도를 수치화한 2차원 행렬 형태(9706×60)의 데이터이다. 실험에는 [3]에서 적용한 기준에 따라 1,364개의 유전자만을 걸러낸 1,364×60 행렬 데이터를 사용하였다.

이 논문에서는 이 데이터에 정보병목기법을 이용해 표본에 대해 단일 및 이중 클러스터링을 적용하였으며, 이중 클러스터링에서는 유전자의 클러스터 수를 변경하며 결과를 살펴보았다. 정보병목기법을 적용하기 위해서는 발현수치를 확률분포로 변환해야 하며, 이를 위해 $\log_2$변환되어 있는 행렬의 각 원소값 x에 다음 식을 적용한 후(양수로 만든다),

$$f(x) = \frac{1}{1 + 2^{-x}}.$$

행렬상의 값의 총합으로 나눈 값을 $p(x)$로 두었으며, missing value의 경우 $x = 0$으로 처리하였다.

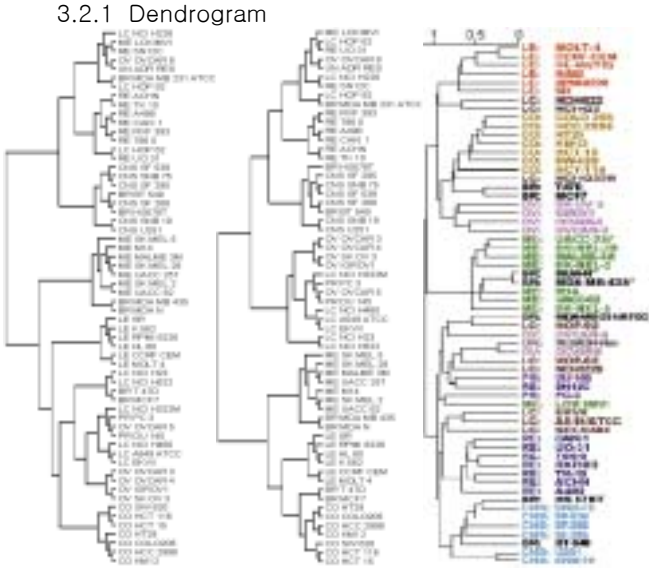
클러스터링 및 분석은 (1) dendrogram을 통한 [3]의 결과와의 비교 및 (2) 이중 클러스터링의 특성과 클러스터링 결과에 대한 품질평가에 초점을 두었다. 표본이 추출 조직에 따라 구분될 것이라는 가정 하에, 60개의 표본을 9개의 클러스터로 구성하여 [표 3]과 같은 표를 구성하였다. 표본의 이중 클러스터링 결과에 대해 엔트로피 측정을 통해 품질 평가를 한 것이 [그림 2]-(a)이며, 클러스터링 과정에서의 상호 정보량의 변화를 살펴본 것이 [그림 2]-(b)와 [그림 3]이다. 이하 유전자 클러스터의 수는 'G=100'으로, 표본 클러스터의 수는 'S=9'와 같이 표기한다. 모든 로그 계산에서 자연로그를 사용하였다.

[표 3] 표본에 대한 이중 클러스터링 결과(G=100). 행은 9가지의 암세포 추출조직이며, 열은 클러스터이다.

class	#1	#2	#3	#4	#5	#6	#7	#8	#9	합계
1:BR	2	2			2				2	8
2:CNS					4		2			6
3:CO									7	7
4:LC	3			4		2				9
5:LE			6							6
6:ME	1	7								8
7:OV	1			5						6
8:PR				2						2
9:RE	2							6		8
합계	9	9	6	11	6	2	2	6	9	60

3.2 결과 및 분석

3.2.1 Dendrogram



(a) (b) (c)

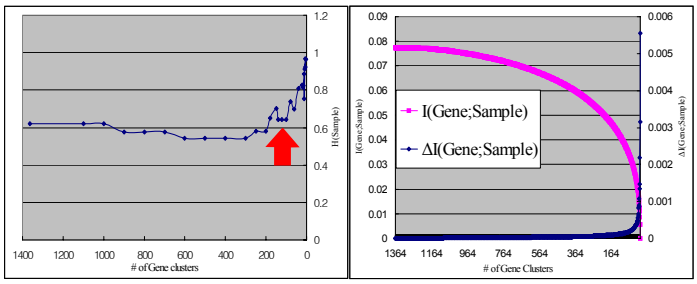
[그림 1] 클러스터링 결과를 dendrogram으로 구성. (a)는 단일 클러스터링, (b)는 G=100인 이중 클러스터링, (c)는 [3]의 결과이며 Pearson correlation을 거리 측정기준으로 사용

[표 3]과 [그림 3]에서 표본을 암세포 종류를 기준으로 살펴보면, CNS, CO, LE, ME(LOXIMVI 제외), PR의 표본은 잘 묶이며, RE는 SN12C와 UO-31이 약간 차이가 있다. OV는 OVCAR-5가 다른 표본들과 약간 거리가 있으며, OVCAR-8은 별도의 행동을 보인다. LC는 부분적으로 모인 표본들이 관찰되며, BR은 가장 특이한 경우로서, HS578T와 BT-549는 CNS와, MDA-N과, MDA-MB-435는 ME와, T-47D와 MCF7은 CO와 항상 묶이는 것을 볼 수 있다. BR:MDA-MB-231/ATCC는 언제나 LC:HOP-92와 묶인다.

이를 종합하면 같은 조직의 암세포는 유전자의 발현 정도가 유사할 것이라는 직관적인 예측에 대한 한 답변으로 볼 수 있으며, 다른 조건에서 클러스터링한 결과가 일치하는 표본간의 관계는 생물학적으로 검증해볼만한 의미 있는 자료가 되리라 기대한다.

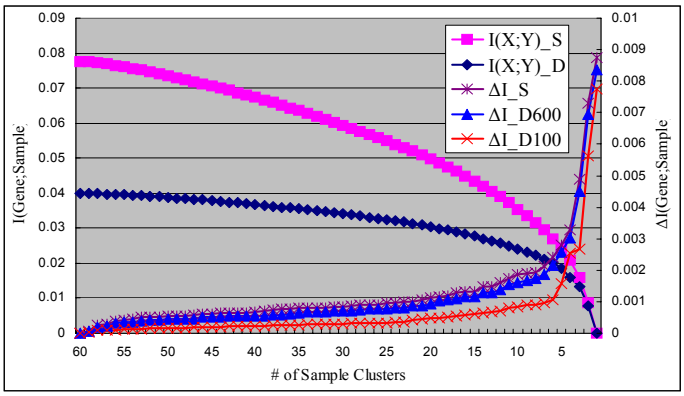
3.2.2 클러스터링 품질 평가

[그림 2]-(a)는 G=1364에서 3까지 다양한 값에서 [표 3]과 같은 표를 구성하여, 클러스터의 엔트로피를 계산한 것이다. 화살표시한 곳은 G=100~140인 지점으로서, 단일 클러스터링(G=1,364)과 비교해 엔트로피의 편차가 8% 이내로 유지되는 한계선이다. 유전자 클러스터링 진행 과정에서의 상호 정보량의 감소 정도([그림 2]-(b))는 예상과는 달리, 꾸준히 감소를 한다. 그러나, 상호 정보량이 최초의 50%정도(G=100일 때 51.5%)가 되더라도 클러스터의 품질은 유지되며, 이는 단일 클러스터링과 비교한 [표 3], [그림 1]-(b)에서도 확인할 수 있다. 종합하면, 유전자의 클러스터링 과정에서 차원을 줄임에 따라 자료의 중복성(redundancy)이 제거되었다고 해석할 수 있다.



(a) (b)

[그림 2] (a)는 표본이 9개의 클래스로 구분된다는 가정 하의 엔트로피 측정 결과. (b)는 이중 클러스터링의 첫 단계인 유전자 클러스터링에서의 상호 정보량(I) 및 병합비용(ΔI)의 변화.



[그림 3] 표본에 대한 단일 및 이중 클러스터링 과정에서의 상호 정보량 및 병합비용의 변화. I(X;Y)_S는 표본의 단일 클러스터링, I(X;Y)_D는 G=100인 이중 클러스터링에서의 상호 정보량의 변화. ΔI_S, ΔI_D600, ΔI_D100은 모두 병합비용의 변화로서, 각각 단일, 이중(G=600, G=100)의 결과이다.

표본의 경우에 상호 정보량 및 병합비용의 변화를 [그림 3]에서 살펴보면, 단일 클러스터링과 이중 클러스터링의 상호 정보량이 근접하는 S=6~7정도에서 병합비용이 급등하는 것을 볼 수 있다. 이는 S가 8 이상일 때까지의 클러스터링 과정에서는 정보의 중복성이 감소하지만, 8 미만에서는 주요 정보의 손실이 발행하는 것으로 해석된다. 즉, 적절한 S값은 8~9정도라고 볼 수 있으며, 표본이 추출 조직에 따라 구분될 것이라는 가정 하에 9개의 클러스터로 구분한 것이 의미가 있다고 해석할 수 있다.

4. 결론 및 논의

이중 클러스터링의 DNA 마이크로어레이 데이터의 중복성과 차원을 줄이고 내재된 구조를 파악하는 도구로서의 가능성을 지금까지 확인해보았다. 이 점에서 PCA(principal component analysis)와 비교될 만하며, 정보병목기법이 PCA와는 달리 비선형적인 접근법이라는 점에서 대조적이다. 향후, 다양한 데이터에 확대적용하는 연구를 수행할 예정이다.

감사의 글

본 연구는 과학기술부 국가지정연구실사업, 뇌신경정보학연구사업과 교육부 BK21-IT 프로그램에 의하여 일부 지원되었음.

참고문헌

[1] Pereira, F.C., Tishby, N., and Lee, L., "Distributional clustering of English words", In *Proceedings of the 30th Annual Meeting of the Association for Computational Linguistics,* pp.183-190, 1993.

[2] Ross, D.T. *et al.*, "Systematic variation in gene expression patterns in human cancer cell lines", *Nature Genetics*, Vol. 24, pp. 227-235, 2000.

[3] Scherf, U. *et al.*, "A gene expression database for the molecular pharmacology of cancer", *Nature Genetics*, Vol. 24, pp. 236-244, 2000.

[4] Slonim, N. and Tishby, N., "Document clustering using word clusters via the information bottleneck method", In *Proceedings of SIGIR-2000*, pp.208-215, 2000.

[5] Tishby, N., Pereira, F.C., and Bialek, W., "The Information bottleneck method", In *Proceedings of the 37th Allerton Conference on Communication and Computation*, pp.368-377, 1999.