논문 2005-42CI-1-1

시드 클러스터링 방법에 의한 유전자 발현 데이터 분석

(Gene Expression Data Analysis Using Seed Clustering)

신 미 영*

(Miyoung Shin)

요 약

마이크로어레이 데이터의 클러스터 분석은 생물학적으로 연관성 있는 유전자 그룹을 찾기 위해 종종 사용되는 방법이다. 기능적으로 연관된 유전자들이 대개 유사한 발현 패턴을 나타내는 특징을 이용하여 유사한 발현 프로파일을 가진 유전자 그룹을 찾아냄으로써 알려지지 않은 유전자들의 기능을 같은 그룹에 속한 다른 유전자로부터 유추할 수 있기 때문이다. 본 논문에서는 클러스터 분석을 위해 시드 클러스터링 알고리즘을 새로이 제안하고, 이 방법을 마이크로어레이 데이터 분석에 적용해본다. 시드 클러스터링 방법은 주어진 데이터를 계산적으로 분석하여 시드 패턴을 자동 추출하고, 이러한 시드 패턴을 목적 클러스터의 프로토타입 벡터로서 간주하여 클러스터를 생성하는 방법이다. 이러한 시드 클러스터링 방법은 수학적 원리에 기초하고 있기 때문에, 매우 체계적인 방법으로 안정적이며 일관성 있는 클러스터링 결과를 생성할 수 있다. 또한, 실제 마이크로어레이 데이터 분석에 적용해본 결과 데이터에 내재된 각 클러스터를 대표하는 시드 패턴을 매우 효과적으로 자동 추출할 수 있었으며, 클러스터링 결과 또한 타 방법에 비해 다소 우월한 경향을 나타내었다.

Abstract

Cluster analysis of microarray data has been often used to find biologically relevant groups of genes based on their expression levels. Since many functionally related genes tend to be co-expressed, by identifying groups of genes with similar expression profiles, the functionalities of unknown genes can be inferred from those of known genes in the same group. In this paper we address a novel clustering approach, called seed clustering, and investigate its applicability for microarray data analysis. In the seed clustering method, seed genes are first extracted by computational analysis of their expression profiles and then clusters are generated by taking the seed genes as prototype vectors for target clusters. Since it has strong mathematical foundations, the seed clustering method produces the stable and consistent results in a systematic way. Also, our empirical results indicate that the automatically extracted seed genes are well representative of potential clusters hidden in the data, and that its performance is favorable compared to current approaches.

Keywords: microarray data, gene expression data analysis, clustering algorithm, seed clustering

I. 서 론

최근 DNA 마이크로어레이 기술의 급격한 발달은 수천에서 수만 개에 이르는 유전자들의 발현 양상을 동시에 관찰할 수 있게 하였고, 이렇게 생산된 대량의 발현 데이터를 효과적으로 분석하기 위한 고급 마이닝 기술이 절실히 필요하게 되었다. 유전자 발현 프로파일에 내재된 알려지지 않은 생물학적인 지식을 찾아내기 위한 중요한 방법 중의 하나는 유전자 발현 프로파일에 대한 클러스터 분석을 수행하는 것이다. 기능적으로 연관된 유전자들의 경우 그 발현 패턴이 매우 유사한 경향이 있기 때문에[1], 많은 연구자들은 다양한 환경에서 동일한 패턴으로 발현되는 유전자 그룹을 찾기 위해 노력해왔다. 만약, 동일한 패턴으로 발현되는 유전자 그룹을 찾을 수 있다면, 현재까지 잘 알려진 유전자들의 기능적 특징으로부터 아직 알려져 있지 않은 유전자들의 기능적 특징을 어느 정도 유추할 수 있기 때문이다.

유전자 발현 프로파일은 일반적으로 행렬의 형태를 가지며, 각 열(row)은 하나의 유전자에 관한 여러 가지 발현 결과들을 나타내고 각 행(column)은 서로 다른 실

정회원, 한국전자통신연구원 바이오정보연구팀
(Bioinformatics Research Team, Electronics and Telecommunications Research Institute)
접수일자: 2004년9월10일, 수정완료일: 2004년1월3일

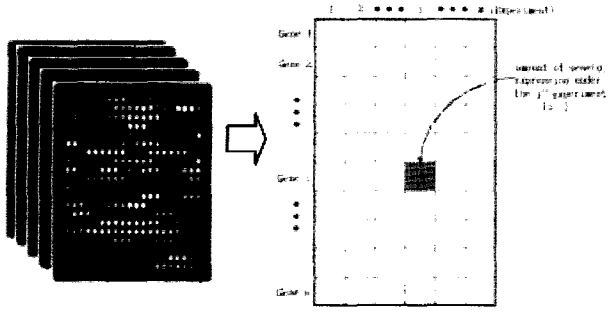


그림 1. 유전자 발현 프로파일의 형태
Fig. 1. Form of Gene Expression Profiles.

험 조건하에서 얻은 결과를 나타낸다 (그림 1). 행렬의 각 구성값은 특정한 샘플에서 특정 유전자의 발현 정도를 수치화한 결과이다.

지금까지 유전자 발현 프로파일 분석을 위해 다양한 클러스터링 알고리즘들이 연구되어 왔다 [1-3, 10-13]. 계층적 클러스터링, k-means 클러스터링, self-organizing maps 등이 대표적으로 가장 많이 사용되고 있는 방법이며, 최근에는 Expectation-Maximization과 같은 확률적 모델 기반 방법[10,13]이나 Quantum clustering[11]과 같은 다양한 방법들이 유전자 발현 프로파일 분석에 이용되고 있다.

본 논문에서는 효과적인 클러스터 분석을 위해 시드 클러스터링 알고리즘을 새로이 제안하고, 이것을 마이크로어레이 실험으로부터 생성된 유전자 발현 프로파일 분석에 적용해 본다. 먼저 제 2절에서는 시드 클러스터링 알고리즘을 상세히 기술하고, 제 3절에서는 가상 데이터 및 실제 유전자 발현 데이터를 이용한 클러스터링 실험 환경을 기술하며, 제 4절에서는 시드 클러스터링 및 다른 방법들을 이용한 클러스터 분석 실험 결과를 보여준다. 마지막으로 제 5절에서는 상기 실험 결과를 바탕으로 시드 클러스터링의 유용성에 대해 토의한다.

Ⅱ. 시드 클러스터링

시드 클러스터링 알고리즘은 두 가지 주요 단계인 (1) 시드 추출 단계와 (2) 클러스터 생성 단계로 구성된다. 첫 번째 단계인 시드 추출 단계는 주어진 데이터로부터 k개의 시드를 계산적인 방법에 의해 자동 추출하는 단계이다. 여기서 시드의 선택 기준은 전체 데이터에 내재된 특징적 패턴을 잘 대표할 수 있고, 그 패턴이 서로 충분히 상이하여 선택된 시드들 간에 나타내는 정보의 중복성이 가능한 적은 개체들로 구성된다. 두 번째 단계인 클러스터 생성 단계는 추출된 시드 패턴을 생성될 클러스터의 프로토타입 벡터로서 간주하고, 각 개체를 가장 가까운 프로토타입 벡터를 가진 클러스터에 할당한 후, 이를 기반으로 시드 패턴을 정제하고 재할당하는 과정을 반복함으로써 이루어진다. 시드 클러스터링 알고리즘에 대한 보다 상세한 내용은 다음과 같다.

가령, 분석할 데이터 D가 다음과 같이 주어졌다고 가정해보자. $\mathbf{D} = \{\mathbf{x}_i, i = 1,\mathrm{K}, n : \mathbf{x}_i = (x_{i1},\mathrm{K}, x_{id}) \in R^d\}$. 이러한 데이터 D로부터 시드 클러스터링 방법을 이용하여 k개의 클러스터를 생성하고자 할 때, 먼저 k개의 시드 추출과정이 필요하다. 이를 위해서, D에 속한 각 데이터 벡터 $\mathbf{x}_i$는 가우시안 함수 변환 $\mathbf{\Phi}$에 의해 다음과 같이 가우시안 특징 행렬 $\tilde{\mathbf{D}}$를 생성한다.

$$\tilde{\mathbf{D}} = \begin{bmatrix} \tilde{\mathbf{x}}_1 \\ \tilde{\mathbf{x}}_2 \\ \mathrm{M} \\ \tilde{\mathbf{x}}_n \end{bmatrix} = \begin{bmatrix} \Phi(\mathbf{x}_1) \\ \Phi(\mathbf{x}_2) \\ \mathrm{M} \\ \Phi(\mathbf{x}_n) \end{bmatrix} = \begin{bmatrix} \phi_1(\mathbf{x}_1) & \phi_2(\mathbf{x}_1) & \Lambda & \phi_n(\mathbf{x}_1) \\ \phi_1(\mathbf{x}_2) & \phi_2(\mathbf{x}_2) & \Lambda & \phi_n(\mathbf{x}_2) \\ \mathrm{M} & \mathrm{M} & & \mathrm{M} \\ \phi_1(\mathbf{x}_n) & \phi_2(\mathbf{x}_n) & \Lambda & \phi_n(\mathbf{x}_n) \end{bmatrix} \quad (1)$$

여기서 가우시안 함수 $\phi_j(\mathbf{x}_i)$는 j번째 데이터 벡터 $\mathbf{x}_j$를 중심축으로 하고 함수 폭이 σ에 의해 조절되는 가우시안 함수로서, 임의의 데이터 벡터 $\mathbf{x}_i$, $i = 1,\mathrm{K}, n$에 대해 $\phi_j(\mathbf{x}_i) = \exp(-\|\mathbf{x}_i - \mathbf{x}_j\|^2 / 2\sigma^2)$ 와 같이 정의된다.

이 때, 가우시안 함수 폭 조절 변수 σ는 $0 < \sigma < \sqrt{d/2}$의 범위 내에서 설정된다.

일단, 수식 (1)에서와 같이 가우시안 특징 행렬 $\tilde{\mathbf{D}}$가 생성되면, 이것을 구성하는 n개의 열벡터(column vector) 중에서 상호 충분히 독립적인 k개의 열벡터 $\mathbf{\Phi}_{s_1}, \mathbf{\Phi}_{s_2}, \mathrm{K}, \mathbf{\Phi}_{s_k}$를 특이값 분해(singular value decomposition)를 이용하여 선정한다(참조 [4]).

가우시안 특징 행렬 $\tilde{\mathbf{D}}$에서 j번째 열벡터 $[\phi_j(\mathbf{x}_1), \phi_j(\mathbf{x}_2), \mathrm{K}, \phi_j(\mathbf{x}_n)]^T$는 j번째 데이터 벡터 $\mathbf{x}_j$를 중심으로 가지는 가우시안 함수에 의해 변환된 결과이기 때문에, 앞에서 선정된 열벡터 $\mathbf{\Phi}_{s_1}, \mathbf{\Phi}_{s_2}, \mathrm{K}, \mathbf{\Phi}_{s_k}$로부터 각 열벡터를 생성하는 데에 사용되어졌던 가우시안 함수의 중심 벡터인 $\mathbf{s}_1, \mathbf{s}_2, \mathrm{K}, \mathbf{s}_k$을 추출할 수 있고 이것이 시드로서 최종 결정된다.

시드가 추출되면, 다음은 이를 기반으로 클러스터를 생성하는 단계이다. 클러스터 생성을 위해 선정된 k개의 시드 $\mathbf{s}_1,\mathbf{s}_2,\mathrm{K},\mathbf{s}_k$는 k개의 최종 클러스터 $\mathbf{C}_1,\mathbf{C}_2,\mathrm{K},\mathbf{C}_k$의 프로토타입 벡터로서 간주된다. 그리하여, 각각의 데이터 벡터 $\mathbf{x}_i$는 각 클러스터의 프로토타입 벡터와 비교하여 그 중 가장 유사한 클러스터에 할당된다. 즉, 각 데이터 벡터 $\mathbf{x}_i$, $i=1,\mathrm{K},n$은 $\tilde{k}=\arg\min_j(\|\mathbf{x}_i-\mathbf{s}_j\|)$ 을 만족하는 클러스터 $\mathbf{C}_{\tilde{k}}$에 할당된다. 이렇게 초기 할당 단계가 끝나면, 각 클러스터의 프로토타입 벡터들은 현재 할당된 클러스터 구성원들의 평균 벡터로서 재구성된다. 이렇게 재구성된 프로토타입 벡터들을 기반으로 재할당 단계가 진행되며, 이러한 반복 과정은 프로토타입 벡터가 안정될 때까지 계속된다. 즉, 프로토타입 벡터의 변화가 더 이상 없다면, 현재 클러스터 멤버쉽이 최종 클러스터 결과로서 확정된다.

III. 실 험 환 경

3.1 실험 데이터

클러스터 분석을 위해 정답 클러스터를 알고 있는 가상 데이터 집합과 실제 효모 세포 주기 데이터 집합을 각각 실험에 사용하였다. 사용된 가상 데이터 집합은 다섯 개의 클러스터를 이루는 250개의 시간열 데이터 벡터를 포함하며, 각 데이터 벡터는 10개의 서로 다른 시점에서 정의된 값으로 구성된다. 이 데이터는 미리 선정된 다섯 개의 시간열 패턴에 정규분포 $N(0,0.5^2)$를 지닌 잡음을 추가하여 각각 생성되었다.

한편, 효모 세포 주기 데이터 집합은 DNA 칩 실험을 통하여 Cho et al.[5]에 의해 얻어진 효모 세포 주기 관련 유전자 발현 데이터로서 두 번의 세포 주기에 해당하는 17개의 서로 다른 시점에서 효모 유전자 6000여개에 대해 측정한 발현값을 포함하고 있다. 이 중에서 최고 발현량을 보이는 시점이 다섯 구간으로 구분된 세포 주기 중 한 곳에서 뚜렷하게 나타나는 384개의 유전자들을 발췌하여[6] 본 실험에 사용하였다. 그리하여, 최고치를 보이는 세포 주기의 구간에 따라 이 데이터는 5개의 그룹으로 구분된다.

일반적으로 세포 주기는 G1기 → S기 → G2기 → M기의 순서로 진행된다고 알려져 있다[9]. G1기는 DNA를 합성하기 위한 준비 기간, S기는 DNA를 복제하는 기간, G2기는 세포 분열을 준비하는 기간, M기는 세포분열이 일어나는 기간을 말한다. 본 실험에서 사용된 효모 세포 주기 데이터는 G1기가 다시 전반기와 후반기로 구분되어 전체 세포 주기가 다섯 개의 구간으로 분류되어 있다. 즉, G1 전반기 → G1후반기 → S기 → G2기 → M기로 구성된다. 그리하여 본 실험 데이터에 포함된 384개의 효모 유전자는 최고의 발현량을 보이는 기간에 따라 G1 전반기에 67개, G1 후반기에 135개, S기에 75개, G2기에 52개, M기에 55개의 유전자들로 구성된다.

3.2 클러스터링 결과의 평가

본 논문에서 클러스터링의 결과는 adjusted rand index (ARI)를 이용하여 평가한다. ARI는 서로 다른 두 그룹 간의 일치하는 정도를 평가하는 통계적 측정치로서, 최근 유전자 발현 데이터 분석 연구[7,8]에서도 사용된 바 있다. 특히, ARI는 주어진 데이터에 관해 이미 알려진 그룹화 기준 (즉, 정답 클러스터)이 존재하는 경우에 사용가능하며, 클러스터링 결과가 알려진 그룹화 기준에 부합할수록 1에 가까운 값을 가진다. 가령, $\mathbf{U}=\{u_1,\mathrm{K},u_R\}$가 정답 클러스터에 나타난 분할 그룹(이하 클래스)이고, $\mathbf{V}=\{v_1,\mathrm{K},v_C\}$가 알고리즘에 의해 생성된 분할 그룹(이하 클러스터)라고 가정하면, ARI는 아래 수식 (2)과 같이 정의된다.

$$\mathrm{ARI}(\mathrm{U},\mathrm{V})=\frac{\sum_{i,j}\binom{n_{ij}}{2}-\left[\sum_i\binom{n_i}{2}\sum_j\binom{n_j}{2}\right]/\binom{n}{2}}{\frac{1}{2}\left[\sum_i\binom{n_i}{2}+\sum_j\binom{n_j}{2}\right]-\left[\sum_i\binom{n_i}{2}\sum_j\binom{n_j}{2}\right]/\binom{n}{2}} \quad (2)$$

여기서 n은 데이터 셀에 속한 전체 데이터의 크기를 나타내며, n_{ij}는 클래스 u_i와 클러스터 v_j 둘 다에 속하는 데이터의 개수를, n_i와 n_j는 클래스 u_i와 클러스터 v_j에 속하는 데이터 개수를 각각 나타낸다.

IV. 실 험 결 과

4.1 가상 데이터 분석 실험

클러스터 분석을 위해, 먼저 가상 데이터의 실제 클러스터 개수인 k=5에 대하여 시드 클러스터링 알고리즘을 수행하였다. 이 때, 가우시안 함수의 조절 변수 σ는 $0<\sigma<\sqrt{d/2}$의 범위에서 선택되어졌다. 즉, 가

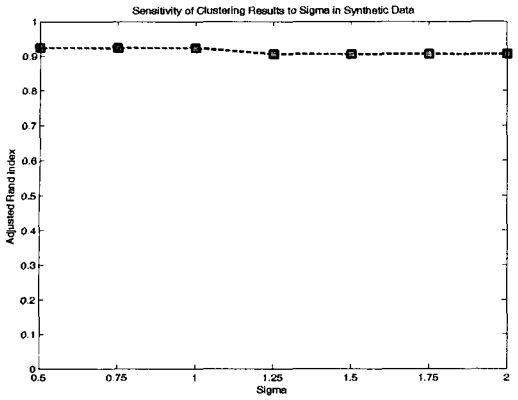


그림 2. 가상 데이터 분석 실험에서 가우시안 함수 조절 변수 σ 값의 변화에 따른 클러스터링 결과의 변화추이 (k=5인 경우)

Fig. 2. Sensitivity of clustering results to a control parameter σ in synthetic data.

상데이터의 데이터 벡터 차원 수가 d=10이므로 실제 σ 값은 σ = (0.25:0.25:2.0)과 같이 선택되어졌다.

가상 데이터에 대한 가우시안 함수의 조절 변수 σ 가 시드 클러스터링 결과에 미치는 영향을 알아보기 위하여 σ =(0.25:0.25:2.0)의 범위에서 σ 값의 변화에 따라 k=5인 클러스터를 생성하면서 클러스터링 결과의 변화 추이를 살펴보았고, 이에 대한 결과는 그림 2와 같다. 그림 2에 나타난 바와 같이, 가상 데이터의 경우, 클러스터링 결과는 σ 의 값에 그다지 변화가 크지 않았다. 그리하여 가상 데이터에 대한 다른 클러스터링 방법들과의 성능 비교 시 σ =1인 경우의 클러스터링 결과를 사용하였다.

이와 같이 가상 데이터에 대해 가우시안 함수 조절 변수 σ =1을 사용하여 다섯 개의 클러스터를 생성하고자 할 때, 본 논문에서 제안한 시드 클러스터링 방법에 의해 자동 추출된 시드는 그림 3과 같다. 그림 3에서 왼편의 그래프들은 가상 데이터 집합을 구성하고 있는 개체를 다섯 가지의 시간열 패턴별로 각각 분류하여 나타낸 것이며, 오른편의 그래프는 시드 클러스터링 알고리즘에 의해 자동 추출된 다섯 개의 시드 패턴들을 그에 대응하는 시간열 패턴들과 연계하여 보여주고 있다. 그림 3에 나타난 바와 같이, 자동 추출된 시드 패턴은 실제 데이터에 내재된 다양한 대표 패턴을 매우 잘 표현하고 있음을 알 수 있다.

한편, 시드 클러스터링의 성능을 평가하기 위하여, 현재 마이크로어레이 분석에서 대중적으로 활용되고 있는 세 가지 계층적 클러스터링 방법과 두 가지 분할 클러스터링 방법을 사용하여 가상 데이터를 분석하고 이 결

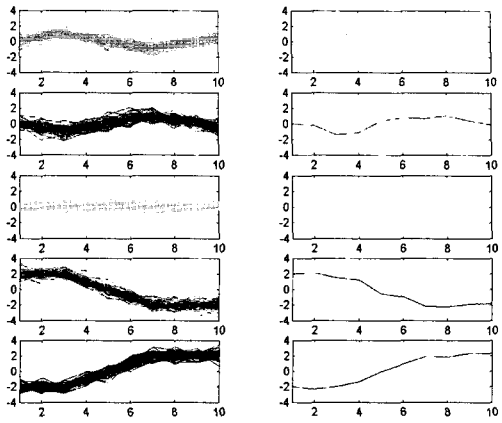


그림 3. 가상 데이터 집합에 실제 포함된 다섯 가지 시간열 패턴과 시드 클러스터링에 의해 자동으로 추출된 .시드 패턴과의 비교

Fig. 3. five different time-series patterns hidden in synthetic data and their corresponding seeds automatically extracted by seed-clustering method.

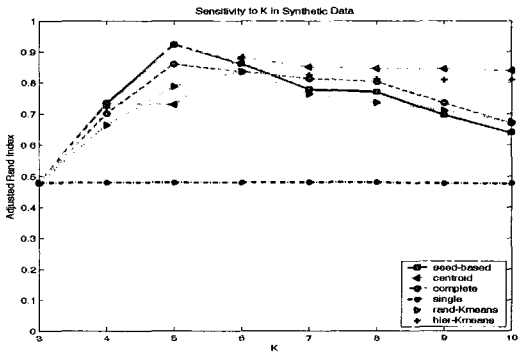


그림 4. 가상 데이터 분석 실험에서 클러스터 개수 k의 변화에 따른 클러스터링 결과 비교

Fig. 4. Sensitivity to the number k of clusters in synthetic data.

과를 시드 클러스터링 결과와 비교하여 보았다. 계층적 클러스터링을 위해 중심 결합 (centroid-linkage), 완전 결합 (complete-linkage), 그리고 단일 결합 (single-linkage) 방법을 사용하였으며, 분할 클러스터링 방법으로는 무작위로 선정된 초기값을 사용하는 k-means 방법(이하 rand-kmeans)과 중심 결합에 의한 계층적 클러스터링 결과를 k-means 방법의 초기값으로 사용하는 하이브리드 방식(이하 hier-kmeans)을 이용하여 각각 클러스터 분석이 이루어졌다. 무작위로 선정된 초기값을 사용하는 k-means 방법의 경우, 10번의 수행 결과에 대한 평균치를 비교 대상으로 고려하였다.

시드 클러스터링 방법 및 위에서 언급한 다섯 가지 다른 방법들에 의해 생성된 클러스터의 ARI 평가 결과는 그림 4과 같다. 먼저, 각 방법들에 대해 k=5일 때의

클러스터 생성 결과를 분석하였고, k값의 선택이 클러스터링 결과에 미치는 영향을 분석하기 위해 위와 동일한 실험을 k=3부터 k=10까지 반복하여 수행하였다. 여기서 ARI의 값이 클수록 클러스터링 결과가 정답 클러스터와 가깝게 일치한다는 것을 의미한다. 그림 4에 나타난 바와 같이, 현재 실험에서 시도된 모든 클러스터링 방법들 중에서 본 논문에서 제안한 시드 클러스터링 방법과 hier-kmeans 방법이 최상의 결과를 보여주었으며, 특히 여러 후보 클러스터의 개수 중에서 정답 클러스터의 실제 개수인 k=5일 때 차별화된 좋은 결과를 나타내었다.

4.2 효모 세포 주기 데이터 분석 실험

데이터 분석을 위해, 효모 세포 주기 데이터는 평균값 0와 분산 1을 가지도록 정규화하여 사용하였다. 이 데이터 집합에 포함되어 있는 각 효모 유전자는 최대 발현량이 다섯 구간으로 이루어진 세포 주기 중의 하나에서 뚜렷하게 나타나는 특성이 있기 때문에 본 실험에서는 먼저 실제 클러스터 개수인 k=5를 사용하여 클러스터를 생성하였다. 이 때, 시드 추출을 위해 가우시안 함수 조절 변수 σ = (0.25:0.25:2.75)가 사용되었으며, σ 값의 변화에 따른 k=5에 대한 클러스터링 결과의 변화 추이를 살펴보면 아래 그림 5와 같다. 이전의 가상 데이터 실험에서와는 달리 σ 값의 변화에 따라 클러스터링 결과가 다소 차이를 나타내는 경향이 있었고, 최상의 결과는 σ =2.0에서 얻어졌다.

σ =2.0일 때, 시드 클러스터링 알고리즘에 의한 시드 추출 결과는 그림 6과 같다. 그림 6에서 상단의 그래프는 효모 세포 주기를 이루는 다섯 구간별 발현 패턴을 나타낸 것이며, 하단의 그래프는 알고리즘에 의해 자동 추출된 시드 패턴의 형태를 보여주고 있다. 추출된 시드 패턴은 효모 세포 주기 데이터에 내재된 그룹별 대표 패턴을 잘 묘사하고 있음을 알 수 있다.

또한, 이러한 시드 패턴을 이용하여 클러스터를 생성한 결과는 아래 그림 7과 같다.

그림 7에서는 시드 클러스터링 방법 이외에 4.2절에서 설명하였던 다섯 가지 클러스터링 방법을 이용한 분석 결과도 함께 보여주고 있다. 이 결과에 따르면, 시드 클러스터링 방법은, 다른 방법과 달리, 실제 클러스터의 개수인 k=5에서 최적의 결과를 나타내었으며, 다른 클러스터 개수(예, k=3, 4, 6, 7.. 등인 경우)에서는 상대적으로 낮은 평가치를 보임으로써 사용자가 적절한 클러스터 개수를 선정할 수 있는 변별력을 제공해줄 수 있

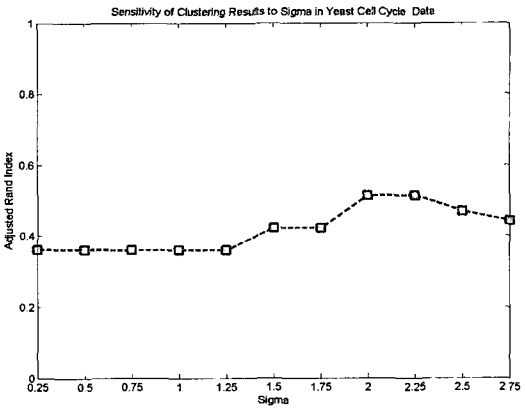


그림 5. 효모 세포주기 데이터 분석 실험에서 가우시안 함수 조절 변수 σ 값의 변화에 따른 클러스터 결과의 변화 추이 (k=5인 경우)

Fig. 5. Sensitivity of clustering results to a control parameter σ in yeast cell cycle data (when k=5).

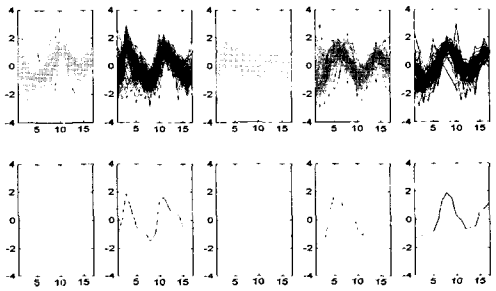


그림 6. 효모 세포주기 데이터 분석 실험에서 세포주기별 발현 패턴과 이에 대응하는 시드 클러스터링 알고리즘에 의해 자동 추출된 시드 패턴 (k=5인 경우)

Fig. 6. Gene expression patterns regarding yeast cell cycles and their corresponding seeds automati-cally extracted by seed-clustering algorithm (when k=5).

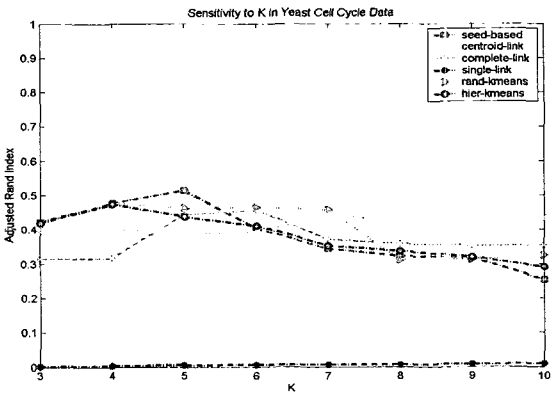


그림 7. 효모 세포주기 데이터 분석 실험에서 클러스터 개수 k의 변화에 따른 클러스터링 결과 비교

Fig. 7. Clustering results with respect to different number k of clusters in yeast cell cycle data.

표 1. 효모 세포 주기 데이터에 관한 알고리즘별 유전자 세포 주기 예측 정확도 비교 (k=5인 경우)

Table 1. Prediction accuracy of several clustering algorithms regarding yeast cell cycles (when k=5).

클러스터링 알고리즘	예측 정확도
seed-based	74.7%
centroid-linkage	57.5%
complete-linkage	65.3%
single-linkage	38.2%
rand-kmeans	57.5-73.0%
hier-kmeans	68.2%

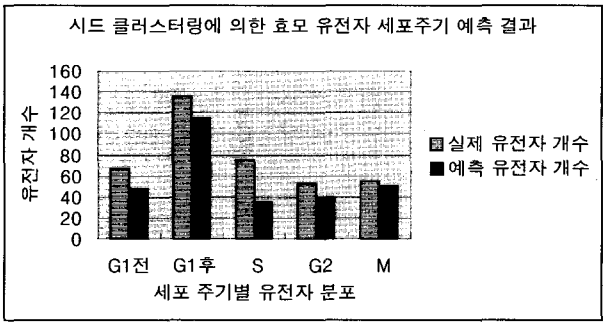


그림 8. 효모 세포주기 데이터 분석 실험에서 시드 클러스터링에 의한 유전자 세포주기 예측 결과

Fig. 8. Prediction results by seed-clustering method regarding yeast cell cycles.

organizing였다. 이에 반해, 다른 방법들은 클러스터 개수의 변화에 따른 결과 평가치의 차별성을 뚜렷하게 나타내지 못하거나, 혹은 클러스터링 결과의 평가치에 있어서 전반적으로 낮은 값을 나타내었다.

상기 그림 7에서 클러스터의 실제 개수인 k=5일 때의 클러스터링 결과들을 기반으로, 효모 유전자들의 세포 주기에 대한 예측 정확도를 계산해 보면 아래 표 1과 같다. 이 표에 의하면, 예측 정확도 측면에서도 시드 클러스터링 방법이 월등하게 우월한 결과를 나타냄을 알 수 있다.

$$\text{예측 정확도} = \frac{\text{정확하게 예측된 유전자 개수의 총합}}{\text{전체 유전자 개수(384)}} \times 100$$

표 1의 예측 정확도는 각 클러스터링 방법에 따라 생성된 다섯 개의 클러스터들을 세포 주기의 각 구간(G1 전반기, G1 후반기, S기, G2기, M기) 중 하나로 대응시키고, 이미 알려진 유전자들의 세포 주기 구간과 비교하여 정확하게 예측된 비율을 측정한 값이다. 여기서 클러스터와 세포 주기 구간의 대응 방식은 예측 정확도가 가능한 높도록 하며 대응 구간이 서로 겹쳐지지 않도록 하였다. 즉, 예측 정확도는 다음과 같이 계산된다.

한편, 이 때의 시드 클러스터링 방법에 의해 생성된 클러스터 결과를 바탕으로 각 세포 주기별로 정확하게 예측된 유전자 분포를 살펴보면 아래 그림 8과 같다.

Ⅳ. 결 론

지금까지 본 논문에서는 클러스터 분석을 위한 시드 클러스터링 알고리즘을 새로이 소개하고, 이에 대한 유용성을 가상 데이터 및 효모 세포 주기 관련 마이크로어레이 발현 데이터에 대한 분석 실험을 통해 살펴보았다. 실험 결과에서 나타난 바와 같이 본 논문에서 제안한 시드 클러스터링 방법은 주어진 데이터에 내재된 각 클러스터를 대표하는 시드 패턴을 매우 효과적으로 자동 추출할 수 있었으며, 이를 기반으로 매우 인상적인 클러스터링 결과를 생성할 수 있었다. 더욱이, 이 방법은 수학적 이론에 기초를 두고 있기 때문에, 매우 체계적인 방법으로 안정적이고 일관성 있는 클러스터링 결과를 생성한다는 특징이 있다. 이러한 매력적인 특성 때문에 시드 클러스터링 방법은 사용이 용이하고 클러스터 성능 또한 타 방법에 필적할 만하므로 실사용자들에게 더욱 어필할 수 있으리라 생각된다.

참 고 문 헌

[1] M.B. Eisen, P.T. Spellman, P.O. Brown and D. Botstein, "Cluster analysis and display of genome-wide expression patterns," Proc. Natl. Acad. Sci., Vol. 95, pp.14863-14868, 1998.

[2] S. Tavazoie, J.D. Hughes, M.J. Campbell, R.J. Cho and G.M. Church, "Systematic determination of genetic network architecture," Nature Genetics, Vol. 22, pp. 281-285, 1999.

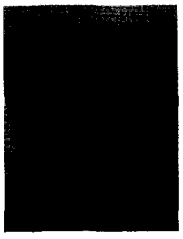
[3] P. Tamayo, D. Slonim, J. Mesirov, Q. Zhu, S. Kitareewan, E. Dmitrovsky, E.S. Lander and T. R. Golub, "Interpreting patterns of gene expression with self-organizing maps: Methods and application to hematopoietic differentiation," Proc. Natl. Acad. Sci., Vol. 96, pp. 2907-2912, 1999.

[4] Golub, G.H. and Van Loan, C.F., "Matrix Computation (3rd edition)," The Johns Hopkins University Press, pp. 590-595, 1996.

[5] R.J. Cho, M.J. Campbell, E.A. Winzeler, L. Steinmetz, A. Conway, L. Wodicka, T.G.

Wolfsberg, A.E. Gabrielian, D. Landsman, D.J. Lockhart, and R.W. Davis, "A genome-wide transcriptional analysis of the mitotic cell cycle," Molecular Cell, Vol. 2, pp. 65-73, 1998.
[6] http://staff.washington.edu/kayee/cluster/.
[7] K. Y. Yeung and W.L. Ruzzo, "Principle component analysis for clustering gene expression data," Bioinformatics, Vol. 17, no. 9, pp.763-774, 2001.
[8] K.Y. Yeung, D.R. Haynor and W. L. Ruzzo, "Validating clustering for gene expression data," Bioinformatics, Vol. 17, no. 4, pp.309-318, 2001.
[9] 오쯔까 기치비, 아비꼬 요시미쯔, "비주얼 생화학. 분자 생물학," 해돋이, pp. 94-95, 2000.
[10] Sus. Datta and Som. Datta, "Comparisons and validation of statistical clustering techniques for microarray gene expression data," Bioinformatics, Vol. 19, no. 9, pp.459-466, 2003.
[11] D. Horn and I. Axel, "Novel clustering algorithm for microarray expression data in a truncated SVD space," Bioinformatics, Vol. 19, no. 9, pp. 1110-1115, 2003.
[12] H. Toh and H. Horimoto, "Inference of a genetic network by a combined approach of cluster analysis and graphical Gaussian modeling," Bioinformatics, Vol. 18, no. 2, pp. 287-297, 2002.
[13] C. Ding, X. He, H. Zha, and H.D. Simon, "Adaptive dimension reduction for clustering high dimensional data," Proceedings of 2nd IEEE International Conference on Data Mining, 2002.

저 자 소 개



신 미 영(정회원)
1991년 연세대학교 전산과학과 학사 졸업.
1993년 연세대학교 대학원 전산과학과 석사 졸업.
1998년 미국 Syracuse Univ. 전산학 박사 졸업.
1999년~현재 한국전자통신연구원 바이오정보 연구팀 선임연구원
<주관심분야: 패턴인식, 데이터마이닝, 바이오인포매틱스>