논문 2008-45CI-4-4

MCL 알고리즘을 사용한 유전자 발현 데이터 클러스터링

(Clustering Gene Expression Data by MCL Algorithm)

손 호 선*, 류 근 호**

(Ho Sun Shon and Keun Ho Ryu)

요 약

유전자 발현 데이터의 분석 기법 중 무감독 학습 기반의 클러스터링 기법은 생물학적 변화와 진의 발현 정도를 이해하는데 자주 사용되는 방법이다. 생명공학 연구에 있어서 그래프 기반의 MCL 알고리즘은 그래프 내의 노드들을 클러스터링 하는 알고리즘으로 빠르고 효과적이다. 우리는 기존의 MCL 알고리즘을 개선하여 마이크로어레이 데이터에 적용시켰다. MCL 알고리즘 수행 시 inflation과 대각선 항의 두 요인을 조정하는 시뮬레이션을 실행 하였으며, 마코브 행렬을 이용하여 변환 하였다. 또한 개선된 MCL 알고리즘에서는 더 명확한 클래스를 구분하기 위하여 각 열의 평균을 구한 후 그 값을 임계치로 사용하였다. 따라서 수정된 알고리즘은 기존의 알고리즘들보다 정확도를 높일 수 있었다. 즉, 실제 실험 결과 기존에 알려진 클래스와 비교 했을 때 평균 70%의 정확도를 보였다. 또한, 다른 클러스터링 기법, K-means 알고리즘, 계층적 클러스터링 그리고 SOM 알고리즘을 비교 분석하였으며, 그 결과 MCL 알고리즘이 다른 클러스터링 기법보다 더 좋은 결과를 보임을 알 수 있다.

Abstract

The clustering of gene expression data is used to analyze the results of microarray studies. This clustering is one of the frequently used methods in understanding degrees of biological change and gene expression. In biological research, MCL algorithm is an algorithm that clusters nodes within a graph, and is quick and efficient. We have modified the existing MCL algorithm and applied it to microarray data. In applying the MCL algorithm, we put forth a simulation that adjusts two factors, namely inflation and diagonal term, and converted them by making use of Markov matrix. Furthermore, in order to distinguish class more clearly in the modified MCL algorithm, we took the average of each row and used it as a threshold. Therefore, the improved algorithm can increase accuracy better than the existing ones. In other words, in the actual experiment, it showed an average of 70% accuracy when compared with an existing class. We also compared the MCL algorithm with the self-organizing map(SOM) clustering, K-means clustering and hierarchical clustering (HC) algorithms. And the result showed that it showed better results than ones derived from hierarchical clustering and K-means method.

Keywords : Gene Expression, MCL algorithm, Clustering, Hierarchical, K-means

I. 서 론

생명공학의 급속한 발전으로 인해 대규모 바이오 데이터가 생성됨에 따라 이를 분석하는 여러 방법들이 연구되고 있다. 이러한 바이오 데이터 분석 중 무감독 학습 기반의 클러스터링 알고리즘은 마이크로어레이 데이터의 발현 패턴을 연구하는데 많이 이용된다. 마이크로어레이 자료에서의 클러스터링은 기본적으로 비슷한 정보나 발현 형태를 갖는 유전자들이나 표본을 함께 묶는 과정이다. 같은 군집 내에 속한 표본들끼리는 유사성이 높고, 서로 다른 군집 간에 속하는 표본들끼리는 유사성을 작게 한다. 마이크로어레이 실험 데이터에 대한 클러스터링 알고리즘 개발은 유전자의 기능분석을 통해 유전자의 상호 관련성 분석 등의 중요 연구 분야에 크게 기여할 수 있다[1]. 즉, 기능적으로 연관된 유전자들

* 학생회원, ** 정회원-교신저자, 충북대학교 전기전자컴퓨터공학부
(School of Electrical and Computer Engineering, Chungbuk Nation University)

※ 이 논문은 2008년 정부(교육인적자원부)의 재원으로 한국학술진흥재단의 지원을 받아 수행된 연구임 (지방연구중심대학육성사업/충북BIT연구중심대학육성사업단)

접수일자: 2008년3월1일, 수정완료일: 2008년7월9일

은 발현 패턴도 유사한 경향이 있으므로 유사한 발현 패턴을 가진 유전자를 찾을 수 있다면, 기능이 알려진 유전자로부터 새로운 유전자의 기능을 예측할 수 있다. 기존에 연구된 유전자 발현 분석은 k-means 클러스터링, 계층적 클러스터링, SOM 알고리즘, EM 알고리즘 등 다양한 방법들이 사용되어 왔다. 그러나 기존의 방법은 데이터나 알고리즘에 따라 그 결과가 다르므로 기존의 방법 보다 좀 더 체계적이고 성능이 우수한 알고리즘의 연구가 요구된다. 따라서 본 논문에서는 그래프 내의 확률적 흐름을 시뮬레이션한 MCL 알고리즘을 유전자 발현 데이터에 적용하여 클러스터링 하므로써 더 효율적인 결과를 얻을 수 있다. 본 논문에서는 마이크로어레이 데이터의 클러스터링 방법으로 MCL 알고리즘을 적용하여 기존의 분석 방법과 비교하였다. MCL 알고리즘은 그래프 내의 확률적 흐름을 시뮬레이션을 통해 그래프의 노드들을 클러스터링 하는 알고리즘으로 빠르고 효과적이다. 지금까지 MCL 알고리즘은 여러 생물학 데이터에 적용되어 좋은 결과를 만들어 왔다. 또한 본 논문에서는 기존 MCL 알고리즘의 성능을 향상하기 위해 임계치를 사용하여 분석함으로써 성능을 높였다.

본 논문에서는 유전자 발현 정보를 이용하여 효과적인 클러스터링 분석을 위해 MCL 알고리즘을 적용하였다. Ⅱ장에서는 마이크로어레이 데이터 클러스터링에 대한 관련 연구를 기술하고, Ⅲ장에서는 유전자 발현 데이터의 클러스터링 방법에 대해 설명하고, Ⅳ장에서는 MCL 알고리즘을 기술 하였으며, Ⅴ장에서는 유전자 발현 데이터를 MCL 알고리즘에 적용하여 실험 하였으며, K-means 클러스터링, 계층적 클러스터링과 신경망으로 구분되고 있는 자기 조직화 지도 (Self-organizing map; SOM) 클러스터링 알고리즘과 MCL 알고리즘을 비교 분석 하였다. 마지막으로, 위의 실험 결과에 대한 유용성을 설명하였다.

Ⅱ. 관련 연구

유전자 발현 데이터를 이용한 클러스터링 알고리즘의 기존 연구를 살펴보면 다음과 같다. Hartuv는 그래프 이론에 바탕을 둔 알고리즘으로 가중치가 없는 유사도 그래프를 입력 받아 그 그래프에서 정점을 유전자로 나타내고, 두 정점 사이의 연결선은 두 유전자의 유사성 나타내었다. 이 유사성이 일정 임계치를 넘는 경우에만 그래프가 형성되도록 구성한 알고리즘을 제안하였다[2]. 또한, CLICK 알고리즘은 통계학적인 모델을 기반으로 설계되었다. 즉 유전자간의 유사도를 전체적으로 정규분포를 따른다고 가정하고, 유사도 그래프의 연결선 가중치에 대해서도 조건부 확률을 이용 하였다[3]. MCL 알고리즘은 그래프 내의 확률적 흐름을 simulation을 통해 그래프의 node들을 클러스터링 하는 알고리즘으로 생물학 데이터의 여러 분야에 많이 사용되고 있다. SOM(Self-Organizing Maps) 알고리즘은 Kohonen에 의해 주도적으로 개발되었으며, 비지지도 신경망 (unsupervised neural network), 위상적 순서화 성질(Topological ordering property)등을 가지며, 생명체 시스템의 외부 개입 없이 자체적으로 조직화 되는 과정이다. 즉 p 차원의 객체를 입력 받아 n 개의 노드로 구성된 저차원의 지도에 질서를 잡아 나타내는 것이다. 그러나 어떻게 SOM의 크기와 형태를 정할지는 더 연구되어야 할 것이다[4~5]. 계층적 클러스터링은 유전자 전체를 하나의 클러스터로 보고, 각 클러스터들의 유사도를 계산하여 가장 높은 한 쌍씩 병합해서 선택하고, 그 클러스터에 대한 다른 모든 클러스터간의 유사도를 재계산 하여 계층적 수형도 형태로 조직화하여 결과적으로는 하나의 클러스터가 될 때까지 반복해서 계산 한다[6]. Spellman 등은 계층적 클러스터 분석을 사용하여 유전자 발현의 주기성을 분석하여 효모에서 세포 분열 주기에 관여하는 유전자 800여 개를 새로 찾아내었다[7]. Eisen 등은 통계학에 많이 사용되는 계층적 클러스터링 알고리즘을 DNA 마이크로어레이 데이터 클러스터링에 적용하여 분석 하였으며, 이 연구를 이용하여 만든 Cluster & Treeview 소프트웨어는 많은 사람들에 의해 사용 되어 지고 있다[8].

앞에서 언급된 다양한 클러스터링 방법이 있다. 기존의 그래프 기반 클러스터링 방법은 가중치 없이 유사도 그래프를 입력 받아 그래프의 정점을 유전자로 사용하였다. 이러한 점을 개선하여 좀 더 체계적인 유전자 클러스터링 방법을 위해 가중치와 임계치를 이용하여 더 효율적인 방법을 필요로 한다. 또한, 데이터 마다 적용되는 방법에 따라 결과가 다를 수 있고, 모수에 따라서도 큰 차이가 나므로, 어떤 데이터나 모수에서도 적용 가능한 일반화된 클러스터링 분석이 이루어 져야 할 것이다.

Ⅲ. 유전자 발현 데이터 클러스터링

유전자 발현 정보의 클러스터링은 새로운 생물학적

하위 그룹이나 클래스를 발견하기 위해 사용된다. 즉 비슷한 발현 형태를 갖는 유전자 들이나 표본들을 함께 묶는 과정이다. 같은 군집 내에 속한 군집들끼리는 유사성이 높고, 서로 다른 군집에 속하는 유전자끼리는 유사성을 작게 한다.

실제 논문에서 실험된 알고리즘을 살펴보면, 먼저 계층적 클러스터링은 발현 패턴이 유사한 유전자들을 이웃하는 트리 형태로 구성하는 방법이다. 즉, 각각 한 개씩의 개체를 갖는 클러스터로부터 시작하여 적절한 개수의 클러스터가 만들어질 때까지 각 계층에서 가장 유사한 두 개의 클러스터를 병합하는 과정을 반복적으로 실행한다. 이 방법은 클러스터링 결과를 트리 모양인 덴드로그램(dendrogram)으로 시각화하여 전체적인 발현패턴을 파악할 수 있다. 본 논문에서는 평균 연결 방법을 이용하여 두 군집 사이의 거리를 각 군집에 속하는 모든 개체들의 평균거리로 정의하여 유사성이 큰 군집을 묶어 나가는 방법을 사용하였다[6].

K-means 클러스터링은 입력값으로 k를 취하고, 군집내 유사성은 높고 군집끼리의 유사성은 낮게 되도록 n개의 객체 집합을 k개의 군집으로 분해한다. 군집 유사성은 군집의 중심 혹은 무게 중심으로 객체들의 평균을 고려하여 측정하였다[9].

Self-Organizing Maps(SOM) 알고리즘은 Kohonen에 의해 개발된 신경망 학습 방법 중의 하나이다[4]. 벡터 형식의 입력 값이 주어지면, 이미 정해진 참조벡터들의 입력에 따라 학습을 통해 최종적인 벡터 값을 찾는 알고리즘이다. SOM의 목적은 점들로 구성된 고차원 공간의 모든 점들간의 거리와 인접관계는 최대한 유지하면서 저차원의 목표공간으로 나타내는 것이다[5]. SOM을 가지고 클러스터링을 할 때 참조벡터의 개수를 정해주어야 하기 때문에 클러스터의 개수를 알고 있다고 가정하고 $k+1(=n)$이라는 2차원 참조 벡터를 전달인자로 주어야 한다.

Ⅳ. MCL 알고리즘

MCL(Markov CLustering) 알고리즘은 그래프의 노드들 사이 전이 확률을 결정하기 위해 마코브 행렬을 이용하여 random work 들을 시뮬레이션 하였다[10~11]. 그래프 흐름 이론과 확률에 기반한 클러스터링 알고리즘으로 빠르고 효과적이다. 그래프를 통해 random work의 확률을 계산하는 절차는, 확률 집합들로 변환하는 두 연산자를 사용한다. 두 연산자는 inflation 요인과 expansion 연산자로 시뮬레이션을 통해 최적의 값을 찾을 수 있다. MCL 알고리즘을 설명하기 위해 그래프 내의 random work 의 확률 값으로 마코브 행렬을 사용하는데, 이 행렬은 각 열의 확률 합이 1을 넘지 않는다.

```
Procedure Markov Clustering
Input : Markov Matrix
Output : Idempotent matrix
        for i = 1 to ∞ do
            calculate Expansion
            calculate inflation
            if matrix is idempotent then
              break
            end if
        end for
```

알고리즘 1. MCL 알고리즘
Algorithm 1. MCL Algorithm.

Expansion 연산자는 다음 식 (1)과 같이 행렬의 곱으로 표현 할 수 있다.

$$Ep(M) = M^2 \tag{1}$$

Inflation 요인은 전이율(transition rate)이 높으면 확률을 더 높게 만들고, 낮으면 더 낮게 만드는 효과를 얻는데 사용된다. 즉 다음 식 (2)를 사용한다.

$$INFlation(M_{ij}) = (M_{ij})^r / \sum_{i=1}^{k} (M_{ij})^r \tag{2}$$

Expansion 연산자를 사용한 행렬을 더 이상 변하지 않게 될 때까지 즉, 각 클러스터의 수를 유지하는 역할을 할 수 있게 멱등 행렬 (idempotent matrix)을 사용한다. 멱등 행렬은 그 행렬의 곱이 같은 행렬로 식 (3)과 같이 나타낸다.

$$M = M^2 \tag{3}$$

MCL 알고리즘은 Markov 행렬을 기반으로 시작되며, 두 연산자를 이용하여 멱등 행렬이 될 때까지 두 연산자들을 수행한다. MCL 알고리즘을 실행하는 알고리즘을 살펴보면 다음과 같다.

실제 실험 데이터를 적용했을 때의 구체적인 방법을 살펴보면, 먼저 각 표본 사이의 유전자들을 전치행렬을 이용하여 바꾸고, 유클리디언 거리 즉, 식 (4)을 이용하여 정방 행렬로 변환하였다.

$$Euclidean\ distance\,(S_{ij}) = \sqrt{\sum_{k=1}^{n} (M_{ki} - M_{kj})^2} \tag{4}$$

$$IF\ S_{ij} > Threshold => NS_{ij} = S_{ij}$$
$$IF\ S_{ij} \le Threshold => NS_{ij} = 1/S_{ij}$$
$$IF (i = j),\ Then\ Diagonal\ Term = Optimal\ value$$
$$where,\ S_{ij} = Euclidence\ distance$$

조건 1. 임계치를 사용한 MCL Algorithm
Condition 1. MCL Algorithm using Threshold.

다음은 유클리디언 거리에 의해 구해진 행렬을 MCL 알고리즘에 적용한다. 변환된 행렬에 inflation과 expansion 의 두 요인을 고려하고 최적의 모수를 찾기 위하여 시뮬레이션 한다. 그리고 본 논문에서는 MCL 알고리즘을 수정하여 기존의 MCL 알고리즘에 더 명확한 클래스를 구분하기 위하여 각 열의 평균을 구한 후, 그 값을 임계치로 사용한다. 유클리디언 거리의 값이 임계치보다 크면 그대로 두고, 임계치보다 작거나 같으면 유클리디언 거리의 값에 역수를 취한다. 이때, 반복된 계산을 통해, 멱등 행렬이 될 때까지 실행하여 마코브 행렬을 만들 수 있다. 임계치를 사용하는 것은 실제 분산을 줄이는 효과가 있으므로 실제 실험에서 에러율을 감소시킨다. 따라서 분류의 성능을 향상 시킬 수 있으므로, 클러스터링의 정확도를 높일 수 있다.

대각선 항(diagonal term)은 시뮬레이션을 통해 가장 클러스터링이 잘되었을 때의 값을 찾는다. 임계치를 적용한 MCL 알고리즘을 위한 조건식 1은 위와 같다. 여기서 대각선 항은 무수히 많은 실험을 통해 얻어진 값이므로 좀 더 체계적이고, 의미 있는 모수를 만들기 위해서는 많은 유전자 데이터에 적용하여 최적의 모수를 찾는 것이 바람직하다.

V. 실험 및 평가

실험 데이터는 공개된 백혈병 데이터로 유전자의 개수는 7129개이고 표본의 수는 72개이다[12]. 이 데이터는 ALL (Acute Lymphocyte Leukemia) 과 AML (Acute myeloid Leukemia)의 두 클래스로 이루어져 있으므로 전체 표본들이 이 클래스들을 얼마나 잘 구분하는지 R-language를 이용하여 실험하였다[13]. 먼저 유클리디언 거리를 이용하여 행렬로 만들어 변환하였다. 그림 (1)은 실험 데이터이며, 행은 유전자를 나타내고 열은 표본을 나타낸다. 이 행렬은 먼저 행과 열을 바꾸어 각 행에 대한 즉 표본에 대한 정방 행렬로 만든 후, 식 (4)의 유클리디언 거리를 이용하여 그림 2의 형태로 변환하였다.

-214	-139	-76	-135	-166		15	-316	-92	-124	-135
-153	-78	-49	-114	-125		-114	-192	-48	-78	-186
-58	-1	-307	265	-76		2	-95	48	-37	-70
88	200	309	12	168		198	312	200	338	357
-295	-264	-275	-419	-230		-91	-199	-357	-198	-407
-558	-400	-559	-585	-284	...	-155	-344	-508	-425	-566
199	-330	35	156	4		29	324	-348	-31	-141
-176	-169	-367	-259	-122		-105	-287	-194	-229	-315
252	101	295	49	79		42	165	94	-92	295
195	169		248	156		179	225	96	946	209
511	697	1199	935	649		492	757	932	938	654
-125	-36	22	216	57		54	68	57	-15	-56
309	442	168	174	504		277	472	215	493	975
-37	-17	52	-110	-25	...	-19	35	-22	-2	-23
795	702	1198	627	259		279	797	588	1179	2315
825	295	777	170	814		91	227	961	284	256
86	11	41	-59	14		6	-9	-26	88	-12
191	76	228	128	56		2484	971	139	298	799
-87	-14	-41	-91	-25		-2	-81	-32	-9	-18

그림 1. M_{ij} = 마이크로어레이 원시 데이터 (72 x 72)
Fig. 1. M_{ij} = Raw data of microarray (7129 × 72).

0	84238.56	88486.18	59252.05	73384.9		94746.69	101833.4	79537.14	76635.59	96466.56
84238.56	0	84387.97	80073.6	78625.07		98521.84	93605.26	87139.42	75879.26	94001.11
88486.18	84387.97	0	82472.39	94809.8	...	111021	114629.3	92532.81	94613.08	113824.6
59252.05	80073.6	82472.39	0	60754.52		83123.32	95270.14	70046.57	71920.86	94915.14
73384.9	78625.07	94608.8	60754.52	0		79995.79	90409.3	75443.54	82968.78	96019.62
94746.69	98521.84	111021	83123.32	79995.79		0	98261.65	80665.26	90737.26	97952.79
101833.4	93605.26	114629.3	95270.14	90409.3		98261.65	0	85468.21	85221.71	89968.41
79537.14	87139.42	92532.81	70046.57	75443.54	...	80665.26	85468.21	0	67921.84	89268.49
76635.59	75879.26	94613.08	71920.86	82968.78		90737.26	85221.71	67921.84	0	78707.62
96466.56	94001.11	113824.6	94915.14	96019.62		97952.79	89968.41	89268.49	78707.62	0

그림 2. 유클리디언 거리를 이용한 행렬(72 x 72)
Fig. 2. Matrix (72 x 72) using Euclidean distance.

5.00E-04	1.19E-05	1.18E-05	1.69E-05	1.30E-05		0e-05	0.000 0e+00	1.257 4e-05	1.904 7e-05	1.030	9.00E-05
1.19E-05	5.00E-04	1.19E-05	1.25E-05	1.27E-05		9e-05	1.068 6e-05	1.147 6e-05	1.917 3e-05	1.069	7.00E-05
1.18E-05	1.19E-05	5.00E-04	1.21E-05	1.06E-05	...	0e+00	0.000 0e+00	1.000 0e-05	1.050 6e-05	0.000	0.00E+00
1.69E-05	1.25E-05	1.21E-05	5.00E-04	1.65E-05		2e-05	1.048 7e-05	1.427 2e-05	1.990 7e-05	1.050	9.00E-05
1.30E-05	1.27E-05	1.06E-05	1.65E-05	5.00E-04		6e-05	1.100 1e-05	1.925 5e-05	1.205 9e-05	1.041	4.00E-05
1.00E-05	1.02E-05	0.00E+00	1.20E-05	1.25E-05		5.00E-04	1.02E-05	1.24E-05	1.10E-05	1.02E-05
0.00E+00	1.07E-05	0.00E+00	1.05E-05	1.11E-05		1.02E-05	5.00E-04	1.17E-05	1.17E-05	1.11E-05
1.23E-05	1.15E-05	1.00E-05	1.40E-05	1.99E-05	...	1.24E-05	1.17E-05	5.00E-04	1.47E-05	1.12E-05
1.90E-05	1.92E-05	1.00E-05	1.99E-05	1.21E-05		1.10E-05	1.17E-05	1.47E-05	5.00E-04	1.27E-05
1.04E-05	1.00E-05	0.00E+00	1.05E-05	1.04E-05		1.02E-05	1.11E-05	1.12E-05	1.27E-05	5.00E-04

그림 3. NS_{ij} = 임계치를 이용한 행렬
Fig. 3. NS_{ij} = Matrix Using Threshold.

다음은 유클리디언 거리를 이용하여 변환된 72 x 72의 정방행렬이다.

여기서 우리는 그림 2 의 행렬에서 inflation과 대각선 항의 두 요인을 고려하고 최적의 요인을 찾기 위해 시뮬레이션 하였다. inflation 요인은 식 (2)를 이용하였고, 대각선 항은 시뮬레이션을 통해 구한 최적의 값이다. 또한 임계치를 이용하여 다음 조건 (5)을 수행하였다. 따라서 이 조건식이 수행된 후 마코브 행렬이 계산되고, 그림 3의 행렬을 구할 수 있다.

$$IF\ S_{ij} > Threshold => NS_{ij} = S_{ij}$$
$$IF\ S_{ij} \le Threshold => NS_{ij} = 1/S_{ij} \quad (5)$$
$$(IF\ i = j),\ then\ Diagonal\ term = 0.0005$$

즉, 기존 MCL 알고리즘 적용 시와 다른 점은 각 열의 평균을 구한 후 그 값을 임계치로 사용하였다. 그 결과 더 명확한 클래스를 구분하므로 정확도를 높일 수 있었다. 또한 MCL 알고리즘을 적용하여 inflation 요인과 대각선 항을 조정하는 시뮬레이션을 반복적으로 실

행하였다. Inflation 요인이 1.215, 대각선 항이 0.0005 일 때 가장 정확도가 높았다. 다음 그림 4를 통해 알 수 있다.

그리고 유전자 발현 데이터를 사용할 수 있는 Cluster and Treeview 도구를 사용하여 SOM 알고리즘과 계층적 클러스터링 알고리즘을 실험하였다. 그 결과 SOM 알고리즘은 8개의 노드가 만들어지고, 표본에 대한 클러스터도 구할 수 있었다. 계층적 클러스터링은 7개의 클러스터로 나누어지나 덴드로그램에 의해 최종 2개의 클러스터로 구분 할 수 있다. 그러므로 MCL 알고리즘과 비교 가능하다.

그림 5는 계층적 클러스터링을 이용한 결과이다. 표 1은 SOM을 이용한 결과로서 표본에 대해 9개의 노드가 만들어 지는데 노드 1에는 해당 표본이 없기 때문에 결과적으로 8개의 노드가 만들어 진다. SOM 클러스터

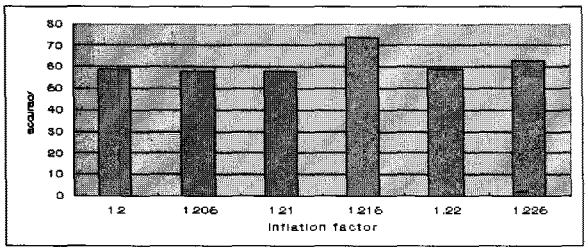

그림 4. 시뮬레이션을 통한 Inflation 실험 결과
Fig. 4. Results of Inflation Experiment Performed Using Simulation.

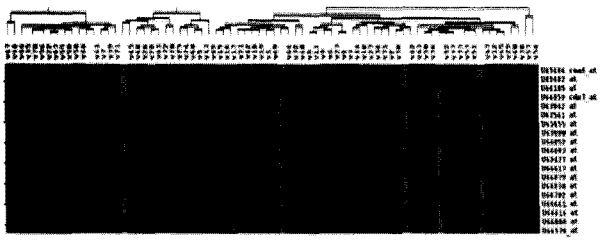

그림 5. 계층적 클러스터링의 실험 결과
Fig. 5. Results of hierarchical clustering.

표 1. SOM의 결과 노드
Table 1. Result node of SOM.

Node No	Sample
Node0	5,13,15,20,21,24,31,32,34,35,36,37,38
Node2	16,19,44,52,68
Node3	9,11,14,17,18,26,30,40,41,47
Node4	2,10,25,39,45,55
Node5	1,3,23,28,46,67,66
Node6	6,22,49,56,71
Node7	33
Node8	4,7,8,12,27,29,50,51,53,54,57,59,60,61,62,63,64,65,69,70,72

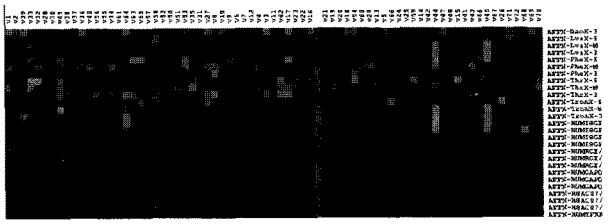

그림 6. K-mean 클러스터링의 실험 결과
Fig. 6. Result of K-means clustering.

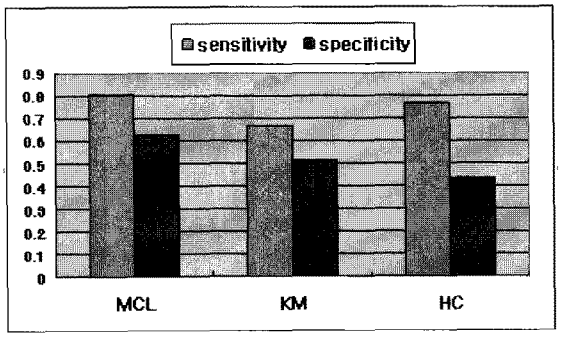

그림 7. MCL 알고리즘과 계층적 클러스터링의 민감도 특이도
Fig. 7. Sensitivity and Specificity of MCL, K-means and HC Algorithms.

링은 고차원 자료를 2차원으로 사영시켜 위상적인 성질을 이용하여 군집분석을 한 결과로 표 1에서와 같이 8개의 군집이 형성된다.

그림 6은 실험 데이터를 K-means 클러스터링을 적용하여 얻은 결과이다. K=2 일 때 두 개의 그룹으로 나누어짐을 알 수 있다.

민감도(Sensitivity)와 특이도(Specificity)는 생물학 데이터의 분류 성능을 평가하는 측도로써 사용된다. 즉, 실제 클래스와 분류기에 따른 결과를 비교하여 그 성능을 측정 할 수 있다. 그림 7은 MCL 알고리즘, K-means 클러스터링 그리고 계층적 클러스터링 알고리즘에 대한 민감도와 특이도를 비교한 것이다. 그 결과 MCL 알고리즘의 군집화가 K-means 알고리즘과 계층적 클러스터링 보다, 더 잘 분류 된 것을 알 수 있다. 특히 MCL 알고리즘의 특이도는 계층적 알고리즘 보다 월등이 높음을 알 수 있다. 따라서 MCL 알고리즘은 마이크로어레이 데이터 적용에서 다른 클러스터링 방법보다 더 효과적임을 알 수 있다.

VI. 결 론

마이크로어레이 실험의 일반화와 유전자를 이용한 연구의 빠른 발전으로 인하여 마이크로어레이 데이터들

은 계속해서 산출 되고 있다. 이러한 방대한 양의 정보들을 바탕으로 의미 있는 정보를 획득하는데 있어서 클러스터링 알고리즘이 중요한 위치를 차지하고 있다. 본 논문에서는 확률적 흐름과 그래프 이론에 기반한 MCL 알고리즘을 유전자 발현 데이터에 적용해 봄으로써 그 효과를 측정 하였다. 실험에서는 72개의 표본과 7129개의 유전자로 된 마이크로어레이 데이터를 MCL 알고리즘에 적용하여 R-language로 실험하였다. 클래스를 잘 분류하기 위해 inflation 요인과 대각선 항을 시뮬레이션 함으로써 가장 정확도가 높은 요인을 찾을 수 있었다. 또한, 기존의 MCL 알고리즘과 차별화하여 각 열의 평균을 구한 후 임계치로 사용함으로써 정확도를 높일 수 있었다. 그 결과 실제 클래스와 비교 했을 때 약 70%의 정확도를 보였다. 또한 Cluster & Treeview 도구를 이용하여 K-means 알고리즘과 계층적 클러스터링에 대한 결과를 얻어 MCL 알고리즘 결과와 비교 분석 하였다. MCL 알고리즘과 K-means 알고리즘 그리고 계층적 클러스터링 방법을 비교한 결과 MCL 알고리즘의 특이도와 민감도가 더 높음을 알 수 있었다. 향후 연구로는 시뮬레이션을 통해 얻어진 inflation 요인과 대각선 항의 모수를 다른 진 발현 데이터에 적용했을 때도 유사한 결과를 보이는지 좀 더 체계적으로 연구되어야 할 것이다.

참 고 문 헌

[1] Ho Sun Shon, Sunshin Kim, Chung Sei Rhee, Keun ho Ryu, "Clustering DNA Microarray Data by MCL Algorithm, *ISMB*, 2007.

[2] E. Hartuv et al., An Algorithm for Clustering cDNAs for Gene Expression Analysis, *RECOM* B 99, pp.188-197, 1999.

[3] R. Sharan and R. Shamir, "CLICK : A Clustering algorithm with applications to gene expression analysis", *In Proceedings ISMB*, 2000.

[4] T. Kohonen, Self-Organizing Maps, *Springer Verlag*, NewYork, 1997.

[5] T. Kohonen, "The self-organizing map," *Neurocomputing*, vol 21, pp.1-6. 1998.

[6] Q. Zhang, Y. Zhang, "Hierarchical Clustering of gene expression profiles with graphics hardware acceleration", pp.676-681, Pattern Recognition Letters, vol 27, 2006.

[7] P. T. Spellman, G. Sherlock, M. Q. Zhang, et al., "Comprehensive identification of cell cycle -regulated genes of the yeast Saccharomyces cerevisiae by Microarray hybridization", *Molecular Biology of the Cell*, vol9, no. 12, 3273 - 3297, 1998.

[8] EisenLab http://ranalbl.gov./EisenSoftware, 2008.

[9] J. Han, M. Kamber, *Data Mining: Concepts & Techniques* 2nded, March 2006.

[10] Stijn Marinus van Dongen, GRAPH Clustering by FLOW SIMULATION, 1969.

[11] Sunshin Kim, Clustering Methods for Finding Orthologs among Multiple Species, http://dblab.chungbuk.ac.kr/~sskim04/,2007.

[12] T. R Golub, D. K Slonim, P. Tamayo et al., "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring". *Science 286*, pp.531-537, 1999.

[13] The R Project for Statistical Computing, 2008 http://www.r-project.org/

[14] Sunshin Kim, Kwang Su Jung, Keun Ho Ryu, "Automatic Orthologous-Protein-Clustering from Multiple Complete-Genomes by the Best Reciprocal BLAST Hits", LNBI, vol.3916, pp.60-70, 2006.

[15] 정광수, 유기진, 정용제, 류근호, "MCL 알고리즘을 이용한 단백질 표면의 바인딩 영역 분석 기법" 정보처리학회논문지 D, 제14-D권 제7호, pp743-752, 2007.12

[16] Ho Sun Shon, Sunshin Kim, Keun Ho Ryu, "Clustering approach using MCL Algorithm for analysing Microarray Data", no1, vol 9, 2007.

저 자 소 개

손 호 선(학생회원)

1986년 성신여자대학교 통계학과 졸업.

1992년 성신여자대학교 대학원 통계학과 석사 졸업.

2006년 충북대학교 전자계산학과 박사 수료.

<주관심분야 : 시공간 데이터베이스, 데이터 마이닝, 바이오인포메틱스, 마이크로어레이 데이터 분석>

류 근 호(정회원)

1976년 숭실대학교 전산학과 졸업

1980년 연세대학교 공학대학원 전산학과 졸업

1988년 연세대학교 대학원 전산학과 박사 졸업

1976년~1986년 육군군수 지원사 전산실(ROTC 장교), 한국전자통신 연구원(연구원), 한국방송통신대 전산학과(조교수) 근무

1989년~1991년 Univ. of Arizona Research Staff (TempIS 연구원, Temporal DB)

1986년~현재 충북대학교 전기전자 및 컴퓨터공학부 교수

<주관심분야 : 시간 데이터베이스, 시공간 데이터베이스, Temporal GIS 지식기반 정보검색 시스템, 유비쿼터스컴퓨팅 및 스트림데이터처리, 데이터마이닝, 데이터베이스 보안, 바이오인포메틱스>