퍼지 및 지능시스템학회 논문지 2007, Vol. 17, No. 2, pp. 167-172

마이크로 어레이 데이터에 적용된 2단계 K-means 클러스터링의 소개

An Introduction of Two-Step K-means Clustering Applied to Microarray Data

박대훈* · 김연태* · 김성신* · 이춘환**
Daehoon Park, Yountae Kim, Sungshin Kim, and Choon-Hwan Lee

* 부산대학교 전기공학과
** 부산대학교 분자생물학과

요 약

많은 유전자 정보와 그 부산물은 많은 방법을 통해 연구되어 왔다. DNA 마이크로어레이 기술의 사용은 많은 데이터를 가져왔으며, 이렇게 얻은 데이터는 기존의 연구 방법으로는 분석하기 힘들다. 본 논문에서는 많은 양의 데이터를 처리할 수 있게 하기 위하여 K-means 클러스터링 알고리즘을 이용한 분할 클러스터링을 제안하였다. 제안한 방법을 쌀 유전자로부터 나온 마이크로어레이 데이터에 적용함으로써 제안된 클러스터링 방법의 유용성을 검증하였으며, 기존의 K-means 클러스터링 알고리즘을 적용한 결과와 비교함으로써 제안된 알고리즘의 우수성을 확인할 수 있었다.

키워드 : K-means 클러스터링, 마이크로어레이, 쌀

Abstract

Long gene sequences and their products have been studied by many methods. The use of DNA(Deoxyribonucleic acid) microarray technology has resulted in an enormous amount of data, which has been difficult to analyze using typical research methods. This paper proposes that mass data be analyzed using division clustering with the K-means clustering algorithm. To demonstrate the superiority of the proposed method, it was used to analyze the microarray data from rice DNA. The results were compared to those of the existing K-meansmethod establishing that the proposed method is more useful in spite of the effective reduction of performance time.

키워드 : K-means Clustering, Microarray, Rice

1. 서 론

생명 공학의 발달로 현재 우리는 박테리아로부터 시작하여 인간에 이르기까지 엄청난 수의 새로운 유전자 정보들을 얻을 수 있게 되었다. 이러한 정보들은 생명 현상에 대한 많은 실마리를 제공한다. 하지만 지금까지의 대부분의 유전공학 방법들은 한계가 있기 때문에 새로운 기술의 개발이 절실히 요구되고 있다. 기존의 방법에 있어서의 문제점들을 극복하기 위해 개발된 방법 중의 하나가 바로 DNA chip을 이용한 유전자 검색 방법이다. DNA chip은 붙이는 유전 물질의 크기에 따라 cDNA chip oligonucleotide chip으로 나뉘는데 우리는 이러한 DNA chip 분석 기술을 이용하여 엄청난 양의 유전자 정보를 얻을 수 있게 되었다[1].

DNA 마이크로어레이 실험 데이터에 대한 효율적인 클러스터링(clustering) 알고리즘 개발은 유전자의 기능 분석(functional genomics), 유전자의 상호관련성 분석 (genetic networks)등의 중요한 분야의 연구에 크게 기여할 수 있다는 의의를 가진다. 더욱이 DNA 마이크로어레이 데이터에는 굉장히 많은 양의 유전자 정보로 이루어져 있으므로, 다양한 데이터마이닝 기법으로 분석하여야 하며, 또한 이러한 분석된 결과를 평가할 수 있는 다양한 방법이 연구되고 있다.

기존에 연구된 데이터마이닝 기법을 살펴보면 우선 Hartuv와 Ben-Dor 등에 개발된 그래프 이론과 알고리즘을 바탕으로 한 데이터 클러스터링 알고리즘[2][3]이 있으며, Tamayo 등은 SOM(Self-Organizing Maps)라는 알고리즘을 개발하고 구현하였다[4]. 또한 Eisen 등은 Hierachical 클러스터링을 이용한 방법을 제안하고 개발하였다[5].

본 논문에서는 기존에 연구된 많은 클러스터링 방법 들이 비교적 작은 양의 데이터에서는 우수한 성능을 보이나 많은 양의 데이터를 처리하기에는 처리 속도가 느리고 처리 능력이 부족하다는 점에서 K-mean 클러스터링 알고리즘을 이용한 분할 클러스터링을 제안한다. 또한 본 논문에서 제안한 K-means 클러스터링 알고리즘을 이용한 분할 클러스터링의 우수성을 검증하기 위하여 기존의 K-means 클러스터링을 시행한 결과와 비교하였으며, 수행 결과 본 논문에서 제안한 알고리즘을 이용한 결과가 기존의 결과에 비해 우수한 결과를 얻을 수 있었다.

접수일자 : 2006년 10월 21일
완료일자 : 2007년 2월 26일
감사의 글 : 이 논문은 부산대학교 자유과제 학술연구비(2년)에 의하여 연구되었음.

2. 마이크로어레이

마이크로어레이 또는 DNA chip은 한 연구자가 동시에 많은 수의 유전자를 이용해서 실험하는데 한계가 있는 기존의 대부분의 노동집약적인 유전공학 방법들의 한계를 넘은 방법으로 기존의 분자 생물학적 지식에 전자 공학 및 기계 공학의 기술들을 접목하여 만들어지게 되었다. 현재에는 전자 집적 기술과 기계 자동화로 인하여 수백 개에서 수십만 개의 DNA 클론을 아주 작은 공간에 집적시킬 수 있게 된 것이다[7]. DNA 마이크로어레이 기술은 고형 지지체 위에 유전자를 고정하는 방법에 따라 cDNA 마이크로어레이 기술과 oligonucleotide 마이크로어레이 기술로 나뉜다.

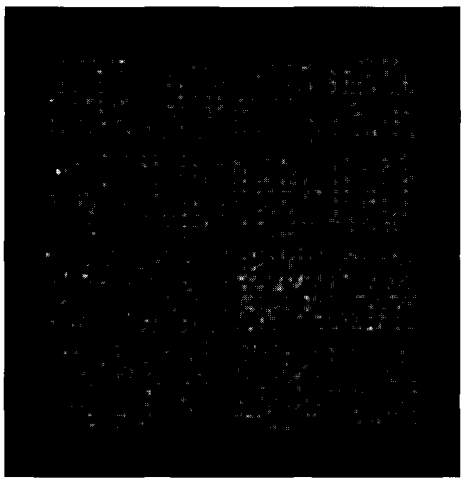

그림 1. 마이크로어레이 발현 분석을 위한 이미지 처리 결과

마이크로어레이를 이용한 유전자 발현 실험으로 얻은 유전자의 발현 정보를 데이터 클러스터링을 통해 비슷한 기능의 유전자끼리 분류하는데 이렇게 분류된 유전자 정보를 통해 우리는 질병 또는 특정 유전자에 대한 정보 등을 얻을 수 있어 여러 분야에 활용할 수 있다.

본 논문에서는 17,000여개로 이루어진 벼 유전자로부터의 마이크로어레이 데이터를 사용하였다. 마이크로어레이 데이터의 이미지는 그림 1과 같으며, 이를 수치화한 데이터는 그림 2와 같다.

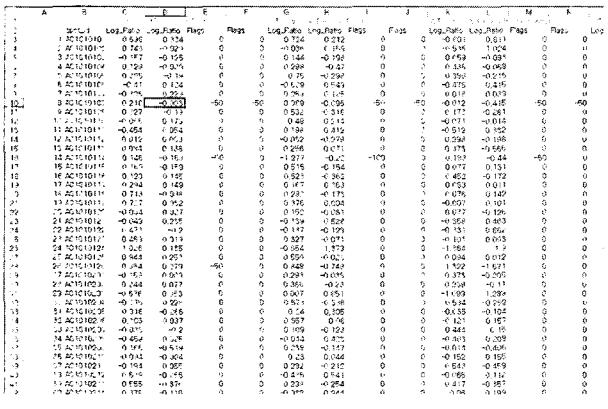

그림 2. 수치화된 쌀 마이크로어레이 데이터

3. 클러스터링 알고리즘

3.1 클러스터링 알고리즘

클러스터링 알고리즘은 주어진 전체 데이터 집합을 유사한 성질을 갖는 몇 개의 클러스터로 분할하는 것이며, 대량의 데이터를 분석하는데 용이하다. 이러한 클러스터링 알고리즘은 패턴 분석 및 분류(pattern analysis and classification), 그룹화(grouping), 의사결정(decision making), 학습 시스템(machine-learning situations), 데이터 마이닝(data mining) 등 여러 가지 분야에서 많이 사용되고 있는 알고리즘이다[8]. 이러한 클러스터링 알고리즘은 수많은 통계학자와 전산학자, 그리고 생물학자 등 많은 분야에 걸친 전문가들에 의해 개발되고 있으며 현재에는 클러스터링 알고리즘 자체에 대한 연구보다는 클러스터링 알고리즘을 이용한 여러 가지 응용에 대한 연구가 활발히 진행되고 있다.

클러스터링 알고리즘은 크게 집적적인 방법인 Hierarchical 클러스터링 알고리즘과 분할적인 방법인 Patitional 클러스터링 알고리즘으로 구분할 수 있다[7].

클러스터링 방법에 있어서 두 클러스터간의 유사도를 정의하는 것은 매우 중요한데 본 논문에서는 식 (1)에서 본 바와 같이 유사도의 측정을 일반적으로 많이 사용하는 방법인 유클리드 거리를 사용하여 유사도를 나타내었다.

$$d(x_i,x_j)=\left(\sum_{k=1}^{d}(x_{i,k}-x_{j,k})^2\right)^{\frac{1}{2}} \tag{1}$$

3.2 Hierarchical 클러스터링 알고리즘

Hierarchical 클러스터링 알고리즘은 통계학에서 사용되어 온 고전적이며 일반적인 알고리즘으로 우선 각각의 데이터를 하나의 클러스터로 생각하고, 각 클러스터들의 유사도가 가장 높은 한 쌍을 선택하여 병합하고, 그 클러스터에 대한 다른 모든 클러스터들 간의 유사도 값을 재계산한다. 주어진 데이터 전체가 하나의 클러스터로 모두 병합될 때까지 이러한 과정을 반복해서 수행하며, 알고리즘이 종료되면 그 결과를 수형도(dendrogram)으로 표현한다. Hierachical 클러스터링 알고리즘에서는 클러스터간의 유사도를 측정하는 방법에 따라 single link, complete link, average link 등의 방법으로 나눌 수 있다[6].

3.3 Partitional 클러스터링 알고리즘

Partitional 클러스터링 방법은 Hierachical 클러스터링의 결과로 얻어 지는 수형도와 같은 클러스터링 구조 대신 데이터의 최종적 분할을 클러스터링 결과로 구하는 방식이다. Partitional 클러스터링은 수형도를 만들기엔 계산적 한계가 있는 많은 용량의 데이터 클러스터링에 적합한 방식이다. 분할 클러스터링의 예로는 Hard c-means 클러스터링 알고리즘과 K-means 클러스터링 알고리즘, Fuzzy c-means 알고리즘 등이 있다.

본 논문에서는 많은 양의 정보를 가진 마이크로어레이 데이터를 사용하기 때문에 Hierarchical 클러스터링 알고리즘이 아닌 Partitional 클러스터링 알고리즘을 사용하였고, 이 중 가장 많이 쓰이는 K-means 클러스터링 알고리즘을 사용하였으며, 성능 개선을 위하여 데이터를 두 번에 걸쳐 클러스터링을 하는 분할 클러스터링 알고리즘을 적용하였다.

3.4 K-means 클러스터링 알고리즘

본 논문에서 사용한 K-means 클러스터링 알고리즘은 partitional 클러스터링 알고리즘 중에서 가장 많이 쓰이는 알고리즘으로 클러스터의 개수 k를 지정하면 지정된 클러스터의 수에 분류된 데이터가 클러스터 내의 원소들과 그 중심과의 거리가 최소가 되도록 분류하는 방법이다. K-means 알고리즘은 클러스터의 중심을 어떻게 잡느냐에 따라 K-means와 K-medoid 방법으로 나뉘는데 여기서 K-means 방법은 클러스터의 중심을 평균값으로 잡는 방법이며, K-medoid 방법은 클러스터의 중심을 무게 중심이 되는 원소로 잡는 방법이다. 본 논문에서는 클러스터의 중심을 평균값으로 잡는 K-means 클러스터링을 사용하였다.

4. 2단 구조 K-means 알고리즘을 이용한 분할 클러스터링 시스템 구현

본 논문에서는 Matlab 7.1을 이용하여 프로그래밍하였으며 구성된 GUI는 그림 3과 같다. GUI는 크게 세 부분으로 분류되며, 그림의 왼쪽에는 결과를 그래프로 보여 주는 결과 그래프 창과 결과를 문자로 나타내 주는 결과 문자 창이 있으며, 그림의 오른쪽에는 각 수행 과정에 대응되는 실행 버튼이 있어 알고리즘의 수행 순서에 맞게 하나씩 실행해 나갈 수 있도록 프로그램을 구성하였다.

프로그램을 실행하는 순서는 우선 'Read Start' 실행 버튼을 누르면 클러스터링을 수행할 엑셀 데이터를 읽어 내고, 클러스터링 패널 안의 'Start' 버튼을 누르면 첫 번째 클러스터링이 수행된다. 그리고 그 아래의 'Result' 버튼을 누르면 두 번째 클러스터링이 수행되어 클러스터링 최종 결과가 출력된다. 그리고 그 아래의 'Analysis' 패널에서는 본 논문에서 제안한 알고리즘에 의한 클러스터링 결과를 이용하여 전체 데이터를 원하는 패턴에 대해 분류하는 과정을 수행할 수 있다.

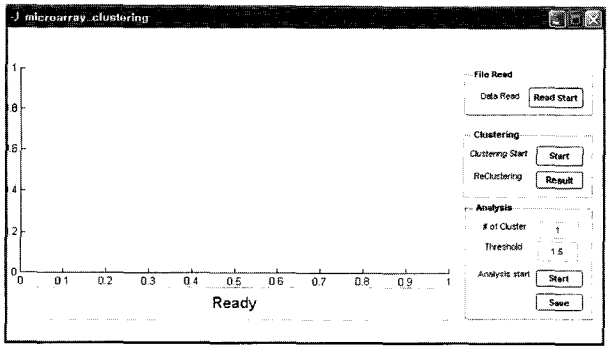

그림 3. 분할 클러스터링 시스템의 GUI

본 논문에서 구현한 2단 구조 K-means 클러스터링 알고리즘을 이용한 분할 클러스터링은 그림 4와 같다. 그림 4에서 본 바와 같이 본 논문에서 사용한 쌀 유전자로부터의 마이크로어레이 데이터를 우선 정규화 과정을 거친 후 두 번의 클러스터링 과정을 거친다. 그렇게 두 번의 클러스터링 과정을 거치면 클러스터링 과정이 완료되고, 이렇게 분류된 클러스터링의 중심값을 이용하여 관심 있는 클러스터링 모양에 대한 데이터를 추출하며, 또한 데이터로 저장할 수 있다.

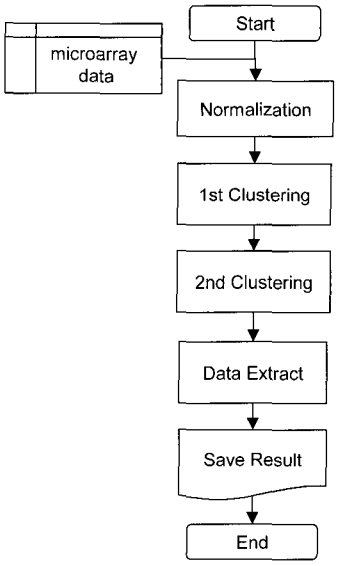

그림 4. 알고리즘의 전체 구성도

4.1 정규화(Normalization)

본 논문에서 사용된 마이크로어레이 데이터는 일반적으로 -1에서 1 사이의 값을 가지고 있으나, 범위를 벗어나는 값 또한 소수 존재하고 있다. 여기서 쓰일 클러스터링 알고리즘은 데이터 간의 유사도를 기반으로 클러스터링을 하기 때문에 편차가 큰 데이터가 있을 경우 그 데이터가 전체 유사도를 결정하게 된다. 따라서 이런 경우에는 유전자의 표준편차를 조정하여 -1과 1 사이의 값으로 조정해 줄 필요가 있다. 따라서 본 논문에서는 -1과 1 값의 범위를 넘어 서는 값에 대해서는 소수의 값들이 유사도에 많은 영향을 끼치므로 최대값 혹은 최소값의 절대값으로 나누어 줌으로써 정규화하는 과정을 거치게 하였다. 이러한 과정을 그림 5에 나타내었다.

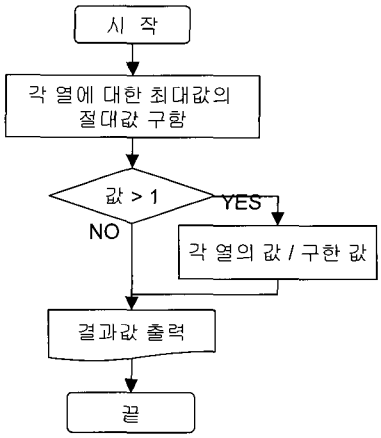

그림 5. 정규화 과정

4.2 분할 클러스터링 알고리즘

본 논문에서 쓰인 데이터는 17,000여개의 쌀 유전자에 대한 마이크로어레이 데이터이므로 한 번에 클러스터링을 하기엔 연산양이 많아 시간이 오래 걸린다는 단점이 있으므로 이것을 두 번에 걸쳐 분할하여 클러스터링을 수행하였으며, 그 과정은 그림 6과 같다. 우선 17,000개의 데이터를 1,000개씩 17개로 분할하였으며, 그 각각의 1000개의 데이터에 대해서 클러스터링을 수행한다. 그리고 클러스터링 수행 시 k=36으

로 하며 36개로 분할된 클러스터링에 대한 대푯값을 이용하여 다시 한 번 클러스터링을 수행한다. 이러한 두 번에 걸친 클러스터링을 통하여 17,000개에 대한 클러스터링 과정을 수행할 수 있다.

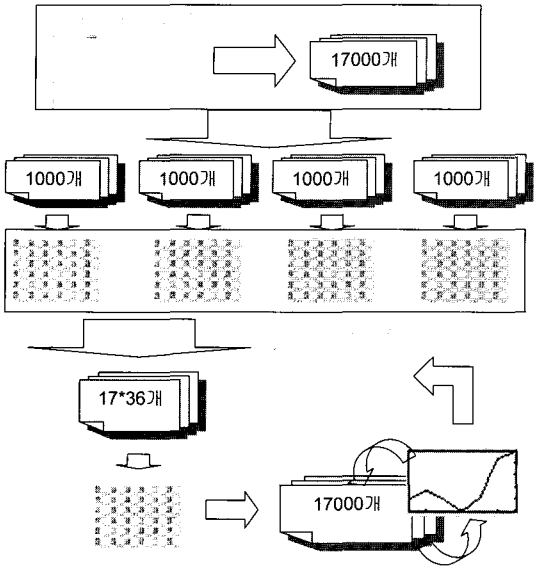

그림 6. 분할 클러스터링 과정

4.3 데이터 분석 및 저장

본 논문에서는 클러스터링이 된 결과를 이용하여 원하는 데이터에 대한 값들을 저장할 수 있게 만들었다. 저장되는 데이터는 대푯값에 대한 그래프와 원하는 클러스터링 값에 가까운 유전자 번호들을 나열한 텍스트 파일의 두 가지 파일을 저장하였다.

5. 시뮬레이션 및 결과 고찰

본 논문에서는 쌀에 대한 마이크로어레이 데이터와 클러스터링 방법에 대해 살펴보고, 쌀 유전자로부터의 마이크로어레이와 같은 많은 데이터에 대한 클러스터링 방법에 대해 제안하였다. 제안된 알고리즘을 이용하여 마이크로어레이 데이터에 대해 클러스터링을 수행하였으며 수행된 결과는 그림 7과 같다.

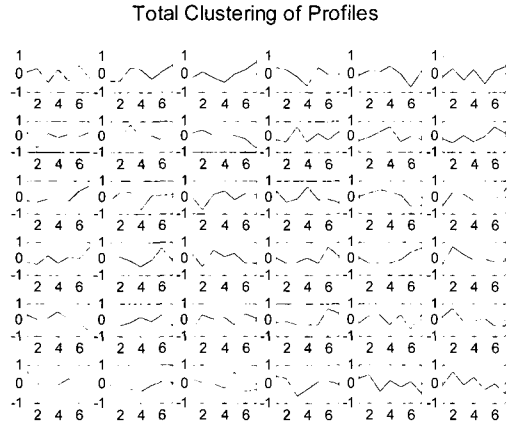

그림 7. 제안된 방법에서의 클러스터링 수행 결과

그림 8은 기존의 방법인 17,000개의 모든 데이터를 한꺼번에 K-means 클러스터링을 수행한 결과이다. 상세한 비교를 위해서 아래의 그림 9, 10에서와 같이 분류된 여러 패턴 중 비슷한 패턴끼리 겹쳐 보이게 하였다.

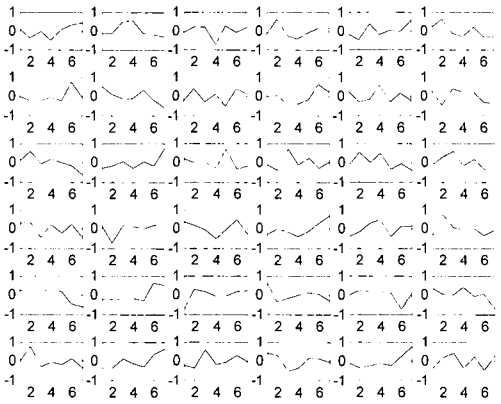

그림 8. 기존의 방법에서의 클러스터링 수행 결과

그림 9는 이전의 방법과 제안된 방법에서의 클러스터링에 대한 대푯값을 비교한 것이다. 그림 9에서 본 바와 같이 두 방법에 대해서 클러스터링은 거의 비슷한 모습으로 나타남을 알 수 있다. 또한 그림 10에서 본 바와 같이 'k=16'일 경우에도 마찬가지로 비슷한 패턴의 클러스터링이 이루어짐을 알 수 있다.

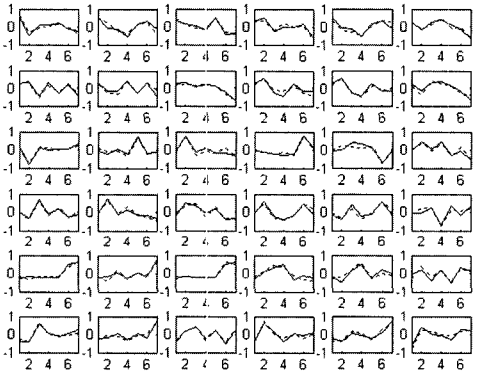

그림 9. 두 방법 간의 클러스터링 대푯값 비교 k=36

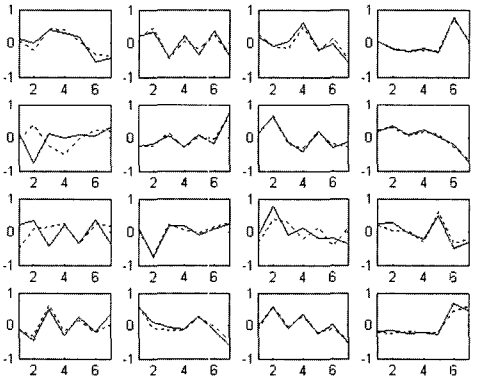

그림 10. 두 방법의 클러스터링 대푯값 비교 k=16

2단 구조 클러스터링 과정에 있어서 첫 번째 클러스터링을 할 경우에 분할하는 그룹의 개수를 17개로 수행하였으나, 이러한 그룹의 개수가 클러스터링의 수행 결과에 영향을 미치는지 알 기 위하여 본 논문에서는 그 그룹의 개수를 다르게 하여 클러스터링을 수행하였다. 그 수행 결과는 그림 11과 그림 12에서 본 바와 같다. 그림 11과 12에서 점선은 기존의 방법으로 수행하였을 때의 클러스터링 결과를 나타낸 것이며, 실선은 그림 11에서는 첫 번째 클러스터링에서 그룹을 25개로 하여 한 그룹 당 680개의 데이터를 가지고 클러스터링을 수행한 결과를 나타냈으며, 그림 12에서는 첫 번째 클러스터링에서 그룹을 20개로 하여 한 그룹 당 850개의 데이터를 가지고 클러스터링한 결과를 나타낸 것이다. 그림에서 본 바와 같이 그룹의 수를 달리하여도 클러스터링 결과는 크게 변하지 않아 그룹의 수에 대해서 클러스터링 결과가 크게 다르지 않음을 알 수 있다.

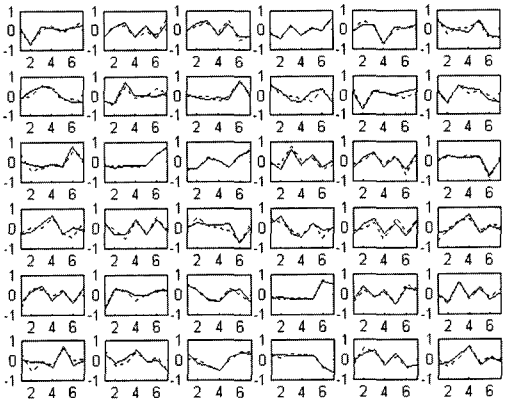

그림 11. 25 그룹으로 나누었을 때의 두 방법 간의 클러스터링 대푯값 비교, k=36

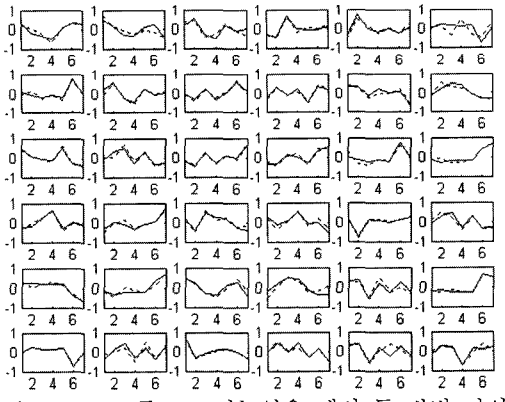

그림 12. 20 그룹으로 나누었을 때의 두 방법 간의 클러스터링 대푯값 비교, k=36

클러스터링의 결과에 대한 비교는 표 1에서도 볼 수 있듯이 두 가지 방식에 의해 이루어졌다. 첫째는 알고리즘을 수행한 전체 시간에 대해서 비교를 하였으며 수행 시간이 짧을수록 우수한 알고리즘이라 할 수 있다. 또한 알고리즘 수행 이후에 생성된 클러스터링 대푯값들 상호간의 표준편차를 구해 비교하였으며, 이 표준편차가 클수록 더 명확하게 분류되었음을 의미한다. 우선 알고리즘 수행 시간을 측정한 결과 기존의 방법에서는 수행 시간이 251초가 나왔으며, 제안된 방법에서는 46.45초가 나와 수행시간이 대폭 줄었음을 알 수 있다. 또한 이전의 클러스터링 방법과 제안된 클러스터링 방법에 대한 표준편차는 기존의 클러스터링 방법에 대해서는 0.7702가 나왔고, 제안된 방법에서는 0.7987이 나와 제안된 방식이 더 우수했음을 확인할 수 있다.

표 1. 이전의 방법과 제안된 방법의 성능 비교.

클러스터링 방법 / 평가 기준	기존의 방법	제안된 방법
수행 시간	251초	46.45초
표준 편차	0.7702	0.7987

기존에 연구된 방법에 대해서는 우선 Hierachical 클러스터링 방법에서는 그러한 클러스터링 알고리즘의 단점에서도 지적되듯이 쌀 마이크로어레이 데이터와 같은 많은 데이터에 대해서는 수행하지 못하는 한계를 가지고 있었으며, partitional 클러스터링 방법에서도 분류는 가능하나, 제안한 방법에 비해 상당히 많은 시간이 걸렸음을 확인할 수 있었다. 본 논문에서 제안한 방법인 K-means 클러스터링을 이용한 분산 클러스터링 방법은 기존의 전체 데이터를 사용한 K-means 클러스터링에 비해 훨씬 빠른 속도로 동작함에도 좀 더 나은 성능을 냄을 확인할 수 있었으며, 더 많은 데이터에 대해서도 충분히 효과적인 방법이라 할 수 있다.

참 고 문 헌

[1] 황승용, "DNA chip 기술," 한국정보과학회지, 제18권 8호, pp .23- 28, 2000.8

[2] E. Hartuv, A. Schumitt, J. Lange, S. Meier-Ewert, H. Legrach and R. Shamir , "An Algorithm for Clustering cDNAs for Gene Expression Analysis," Proceed ings of the Third International Conference on Computational Molecular Biology (RECOMB 99),pp.188- 197, 1999

[3] A. Ben-Dor , R.. Shamir, Z. Yakhini, "Clustering Gene Expression Patterns," Journal of Computational Biology , 6:281- 297, July 14, 1999

[4] P. Tamayo, D. Slonim, J. Mesirov, Q. Zhu, S. Kitareewan, E. Dmitrov sky, E. S. Lander and T. R. Golub, "Interpreting patterns of gene expression with self- organizing maps : Methods and application to Hematopoeitic differentiation," Proceedings of National Academy of Sciences of the USA , v ol.96, pp.2907- 2912, March 1999

[5] M. B. Eisen, P. T. Spellman, P. O. Brown and D. Botstein, "Cluster analysis and display of genome-wide expression patterns," Proceedings of National Academy of Sciences of the USA , vol.95, pp.14863- 14868, December 1998

[6] http://gene-chips.com

[7] A.K.Jain, M.N.Murty, and P.J.Flynn. "Data clustering: A review," ACM Computing Surveys, 31(3):264-323, September 1999

저 자 소 개

박대훈(Daehoon Park)
2003년 : 부산대학교 전자전기정보컴퓨터 공학부 졸업
2006년~현재 : 동대학원 전기공학과 석사 과정

관심분야 : 클러스터링 알고리즘, 고장진단, 지능제어
E-mail : dhsmile@pusan.ac.kr

김연태(Yountae Kim)
2003년 : 부산대학교 전자전기통신공학부 졸업
2005년 : 동대학원 전기공학과 졸업 (공학석사)
2005년~현재 : 동대학원 전기공학과 박사 과정

관심분야 : 신호처리, 영상처리, 고장진단, 지능제어
E-mail : dream0561@pusan.ac.kr

김성신(Sungshin Kim)
1986년 : 연세대학교 전기공학과 졸업 (공학석사)
1996년 : Georgia Institute of Technology, 전기공학과 졸업(공학박사)
1998년~현재 : 부산대학교 전기공학과 부교수

관심분야 : 지능 시스템, 데이터 마이닝
Phone : +82-51-510-2374
Fax : +82-51-513-0212
E-mail : sskim@pusan.ac.kr

이춘환(Choon-Hwan Lee)
1977년 : 서울대학교 식물학과 졸업
1979년 : 동 대학원 식물학과 석사졸업
1988년 : Ph.D., Biophysics, Biophysics Program, Ohio State University, Columbus, Ohio, USA

관심분야 : 광합성 기구의 구조 및 기능, 광저해 방어 및 표현 기구
Phone : +82-51-510-2230
Fax : +82-51-513-9258
E-mail : chlee@pusan.ac.kr