마이크로 어레이 데이터에 적용된 2단계 K-means 클러스터링의 소개

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

박대훈[1], 김연태[2], 김성신[3], 이춘환[4]

[1] 부산시 금정구 부산대학교 전기공학과
E-mail: dhsmile@pusan.ac.kr
[2] 부산시 금정구 부산대학교 전기공학과
E-mail: dream0561@pusan.ac.kr
[3] 부산시 금정구 부산대학교 전기공학과
E-mail: sskim@pusan.ac.kr
[4] 부산시 금정구 부산대학교 분자생물학과
E-mail: chlee@pusan.ac.kr

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.

Key Words : 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-mean을 이용한 분할 클러

스터링을 제안한다.

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[6].
본 논문에서는 17,000여개로 이루어진 벼 유전자로부터의 마이크로어레이 데이터를 사용하였다.

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.

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

3.2 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 1과 같다. GUI는 크게 세 부분으로 분류되며, 그림의 왼쪽에는 결과를 그래프로 보여 주는 결과 그래프 창과 결과를 문자로 나타내 주는 결과 문자창이 있으며, 그림의 오른쪽에는 각 수행 과정에 대응되는 실행 버튼이 있어 알고리즘의 수행 순서에 맞게 하나씩 실행해 나갈 수 있도록 프로그래밍하였다.

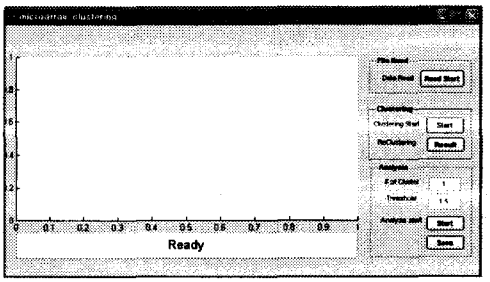

그림 1. 분할 클러스터링 시스템의 GUI.

그리고 본 논문에서 구현한 2단 구조 K-means 클러스터링 알고리즘을 이용한 분할 클러스터링은 그림 2와 같다. 그림 2에서 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.

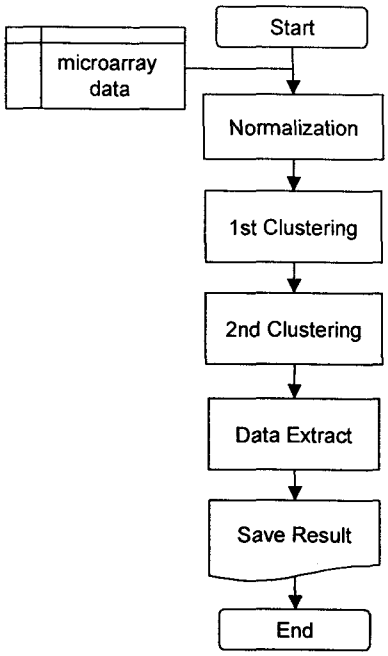

그림 2. 알고리즘의 전체 구성도.

4.1 정규화(Normalization)

본 논문에서 쓰일 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 거쳤다. 정규화하는 과정은 그림3과 같다.

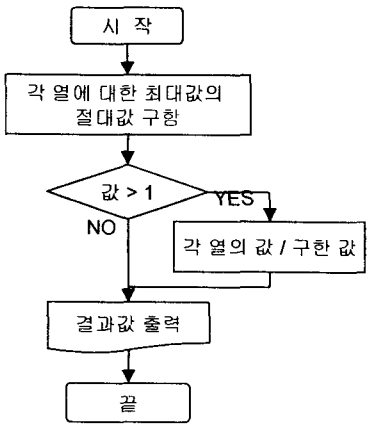

그림 3. 정규화 과정.

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 4와 같다. 우선 17000개의 데이터를 1000개씩 17개로 분할하였으며, 그 각각의 1000개의 데이터에 대해서 클러스터링을 수행한다. 그리고 클러스터링 수행 시 k=36으로 하며 36개로 분할된 클러스터링에 대한 대푯값을 이용하여 다시 한 번 클러스터링을 수행한다. 이러한 두 번에 걸친 클러스터링을 통하여 17000개에 대한 클러스터링 과정을 수행할 수 있다.

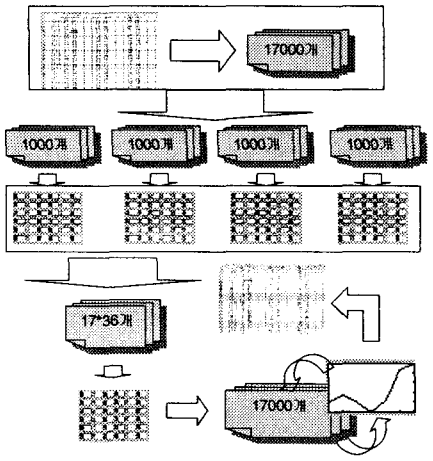

그림 4. 분할 클러스터링 과정.

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.

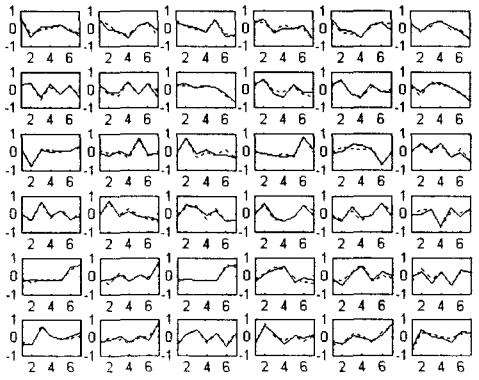

그림 5. 두 방법 간의 클러스터링 대푯값 비교 k=36.

그림 5는 이전의 방법과 제안된 방법에서의 클러스터링에 대한 대푯값을 비교한 것이다. 그림 5에서 본 바와 같이 두 방법에 대해서 클러스터링은 거의 비슷한 모습으로 나타남을 알 수 있다. 또한 그림 6에서 본 바와 같이 k=16일 경우에도 마찬가지로 비슷한 패턴의 클러스터링이 이루어짐을 알 수 있다.

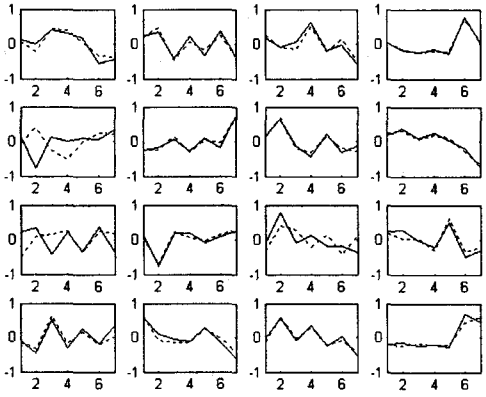

그림 6. 두 방법 간의 클러스터링 대푯값 비교 k=16.

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 47.76초가 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.

감사의 글

이 논문은 2006년 BK21 2단계 사업의 지원에 의하여 연구되었음.

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