유전자 발현 데이터에 대한 클러스터링과 리프오더링 연구[1)]

여상수[0] 이정원 김성권
중앙대학교 컴퓨터공학과
{ssyeo[0], biomania}@alg.cse.cau.ac.kr
skkim@cau.ac.kr

Clustering and Leaf Ordering for Gene Expression Profiles

Sang-Soo Yeo[0] Jung-Won Rhee Sung-Kwon Kim
Dept. of Computer Science & Engineering, Chung-Ang University

요 약

계층적 클러스터링(hierarchical clustering)은 유전자 발현 데이터를 분석할 때 일반적으로 사용하는 방법이다. 계층적 클러스터링의 결과물은 유전자 발현 데이터의 덴드로그램이다. 이 덴드로그램에서 인접한 리프 노드들간의 유사도는 높아지게 하고 멀리 떨어진 노드들간의 유사도는 낮아지게 하기 위해서, 리프 노드들을 재배열하는 과정을 리프오더링이라고 한다. 본 논문에서는 전체 리프 노드들을 대상으로 하는 리프오더링 알고리즘들을 변형하여 각 클러스터별로 리프오더링을 하는 접근방식을 제안하고, 기존의 리프오더링 알고리즘을 사용했을 때의 결과와 제안하는 접근방식을 사용했을 때의 결과를 비교 분석하였다.

1. 서 론

유전자 발현 데이터의 클러스터링에 대한 연구는 다양하게 이루어지고 있다[1,2,3,5,9]. 유전자를 클러스터링하는 문제는 쉽지 않다. 일반적으로 유전자 발현 데이터에서 유전자의 개수에 비해 실험 횟수가 상대적으로 매우 적은 수이기 때문이다. 여기에 대한 현재까지의 연구 결과는 [1,2]의 논문에 잘 나와 있다.

현재 연구 개발된 클러스터링 방법 중에서 가장 일반적으로 생물학자들에 의해서 사용되고 있는 것은 계층적 클러스터링(hierarchical clustering)이다[3,4]. 계층적 클러스터링의 결과물은 흔히 덴드로그램(dendrogram)이라고 불리는 것으로서 노드들간의 유사도(또는 거리)가 함께 표현된 이진 트리를 말한다. 최근에는 이 결과 트리의 리프 노드(유전자를 의미함)의 순서를 전산학적으로 생물학적으로 더 의미 있게 재배열하는 방법에 대해서 여러 연구가 있었다[5,6,10,11]. 이 문제를 일반적으로 "리프오더링(leaf-ordering) 문제"라고 한다. 리프오더링 문제에 대한 자세한 설명은 2.2절에 있다.

본 논문에서는 리프오더링 문제에 대한 기존의 해결 알고리즘들의 공통된 접근 방식이 무엇이었는지를 설명하고, 새로운 접근 방식을 제안한다. 그리고, 기존의 접근 방식과 제안하는 접근 방식을 비교 분석하였다.

2. 리프오더링 문제의 정의

먼저, 리프오더링 문제를 정형화하기 위해 필요한 표기들이다.

$D = \{d_{ij}, i=1,\ldots,m, j=1,\ldots,n\}$: DNA 마이크로어레이 실험 데이터로 구성된 $m \times n$행렬.

d_{ij} : i번째 유전자의 j번째 실험(어레이)에서의 발현도(expression level).

d_i : i번째 유전자의 모든 실험(어레이)에 대한 발현도를 담고 있는 행 벡터.

T : 계층적 클러스터링의 결과로 나온 이진 트리. m개의 리프 노드를 가진다

덴드로그램으로 T를 그리면 근노드가 맨 왼쪽에 위치하고, 자식노드들을 부모노드의 오른쪽에 그린다. 리프노드들은 맨 오른쪽에 수직으로 일렬로 그려진다. 이때 리프 노드들의 순서를 리프오더링이라 부른다. T의 각 내부 노드의 두 자식노드 중에서 어느 것을 위쪽 자식노드로 하고 어느 것을 아래쪽 자식노드로 할 것인가를 덴드로그램을 그릴 때 정해야 한다. T가 $m-1$개의 내부 노드를 가지므로 모두 2^{m-1}개의 리프오더링이 가능하다. 이 중에서 목적 함수의 값을 최적으로 하는 리프오더링을 최적 리프오더링이라 부른다. 흔히 사용하는 목적 함수는 리프오더링에서 인접한 리프노드들 간의 거리(유사도의 역수)의 총합을 최소화하는 것이다.

3. 기존의 알고리즘들과 그 접근 방식

3.1 기존의 알고리즘들

리프오더링 문제는 m의 크기(유전자의 수)가 매우 크기 때문에 전수 검색(exhaustive search)으로 해결하는 것은 어려운 일이다. 따라서, 여러 가지 휴리스틱한 알고리즘들이 제안되었고[3,4,5,10,11], 최근에는 동적 프로그래밍을 이용한 최적해 리프오더링 알고리즘이 개발되었다[6,7]. 아래는 간략하게 현재까지의 리프오더링 알

1) "이 논문은 2001년도 한국학술진흥재단의 지원에 의하여 연구되었음" (KRF-2001-041-E00265)

고리즘들을 설명한 것이다.

Eisen1 알고리즘[3] : 각각의 행 벡터 d_i에 대한 모든 어레이들의 평균 발현도를 이용하는 알고리즘

Eisen2 알고리즘[4] : 행렬 D에 대해서, 먼저 1차원 SOM (Self-Organizing Map)을 적용한 결과를 이용해서 리프오더링하는 알고리즘

Ssyeo 알고리즘[11] : 본 연구자들이 제안했던 것으로서 각각의 행 벡터 d_i 내에서 가장 발현도가 높은 실험의 번호를 이용하는 알고리즘.

Alon 알고리즘[5] : Alon의 분할 방식의 계층적 클러스터링 알고리즘에서만 적용될 수 있는 리프오더링 알고리즘. 분할 확률 함수의 이용

Random 알고리즘 : 무작위 리프오더링 알고리즘.

최적해 알고리즘[6,7] : 동적 프로그래밍 기법을 이용하는 $O(m^3)$의 알고리즘.

3.2 기존 알고리즘들의 접근 방식

3.1절에서 언급된 리프오더링 알고리즘들은 알고리즘을 적용하는데 있어서, 공통적인 접근 방식을 가진다. 그것은 결과 트리 T의 전체 노드들에 대해서 리프오더링을 한다는 것이다. 이 접근 방식을 본 논문에서는 편의상 Overall 리프오더링 방식이라고 부르기로 한다.

Overall 리프오더링 방식의 문제점은 다음과 같다. 자신의 클러스터가 아닌 다른 클러스터에 속해 있는 인접 노드들의 거리도 최소한으로 만들려고 하기 때문에 클러스터 내의 정확한 리프오더링 결과를 보장할 수 없다는 것이다.

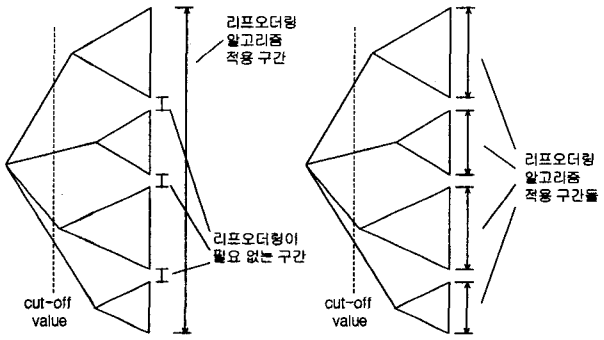

그림 1 기존 방식과 제안하는 방식의 모식도

4. 제안하는 접근 방식

제안하는 리프오더링 접근 방식은 각 클러스터별로 리프오더링 알고리즘을 따로 적용하자는 것이다. 이 접근 방식을 본 논문에서는 편의상 클러스터별 리프오더링 방식이라고 부르기로 한다.

클러스터별로 리프오더링을 하기 위해서는 먼저 결과 트리 T를 클러스터로 나누어야 한다. 트리 T의 에지들은 유사도 값을 표현하고 있으므로, 에지들을 자를 값을 정하면 자연적으로 클러스터로 나누어지게 된다. 에지들을 자를 기준 값을 절단값(cut-off value)이라 하고, 이는 유사도 범위 [-1, 1] 내에서 정해 주면 된다. 클러스터별로 나누어진 후에는 각 클러스터별로 3.1절에서 언급된 리프오더링 알고리즘들을 적용하면 된다.

5. Overall 방식과 클러스터별 방식의 비교 실험 분석

5.1 비교 실험

비교 실험에 사용되는 리프오더링 방법들은 Random 알고리즘, Eisen1 알고리즘, Ssyeo 알고리즘, 최적해 알고리즘 등이다.

표 1 실험에 사용된 데이터들

	데이터 이름	유전자수	어레이수	비고
1	r300_20	300	20	랜덤 데이터1
2	r500_20	500	20	랜덤 데이터2
3	r800_20	800	20	랜덤 데이터3
4	Spellman800	800	82	효모 데이터1
5	Spellman523	523	82	효모 데이터2
6	SpellmanFun1033	1033	82	효모 데이터3
7	Alon2000	2000	62	결장암 데이터

실험 방법 : 실험에서는 본 연구자들이 개발 중인 소프트웨어의 클러스터링 기능과 리프오더링 기능을 사용하였다. 실험 데이터에 대해서, Overall 리프오더링 방식으로 알고리즘들을 적용하였고, 클러스터별 방식으로 알고리즘들을 적용하였다.

실험 데이터 : 실험 데이터는 표 1과 같은 총 7가지의 데이터를 사용하였다. 1),2),3)은 난수를 발생시켜 만든 후, Cluster 소프트웨어[4]를 이용해서 정규화와 로그 변환 등의 필터링을 거친 데이터들이다. 4)~6)은 Spellman[8]의 효모 데이터들 중에서 선택된 데이터들이다. 4)는 세포 주기(Cell cycle)와 관련된 유전자 800개를 선택해서 만든 데이터이다. Bar-Joseph은 이 데이터들을 이용해서 생물학적인 분석을 하였다[6]. 5)는 cdc15에 대한 실험만을 선택한 데이터이다. 6)은 전체 효모 데이터 중에서 20%이상의 데이터가 없는(missing) 필드를 가지고 있는 유전자들과 적어도 로그 비율로 2를 넘는 발현도가 한 번도 나오지 않은 유전자들에 대해서는 필터링하고 난 뒤에 남은 523개의 유전자 데이터이다. 7)은 Alon[5]이 사용한 결장암(colon cancer)환자들에 대한 데이터이다.

5.2 결과 분석

Overall 리프오더링 방식과 클러스터별 리프오더링 방식을 비교한 실험의 결과가 표 2에 나와 있다. 실험 결과에서 통해서 Overall 리프오더링 방식을 사용한 결과보다 클러스터별 리프오더링 방식을 이용해서 알고리즘 적용한 것이 더 좋은 결과를 보임을 알 수 있다.

표 2 Overall 리프오더링 접근 방식과 클러스터별 접근 방식의 결과 비교
(각 항목의 수치는 결과 트리의 인접 리프 노드들 간의 거리 총합. 낮을 수록 좋음)

	Overall 리프오더링 방식				클러스터별 리프오더링 접근 방식				
방법 / 데이터	Random	Eisen1	ssyeo	Optimal	cut-off value	Random	Eisen1	ssyeo	Optimal
r300_20	155.18	157.42	148.87	122.17	0.0	151.60	153.46	145.47	118.55
r500_20	237.56	233.95	222.63	184.74	0.0	232.91	229.66	219.10	180.62
r800_20	356.61	356.18	353.33	281.21	0.0	353.56	352.84	349.40	277.51
Spellman800	336.16	334.21	332.89	289.75	0.1	332.95	331.16	330.32	286.90
Spellman523	236.20	238.55	231.91	210.31	0.1	229.15	232.03	224.88	203.43
SpellmanFun1033	567.03	564.35	561.92	498.92	0.2	554.16	552.54	548.86	486.13
Alon2000	781.32	785.56	777.39	680.12	-0.1	779.43	783.35	775.02	677.92

6. 결론 및 향후 연구 방향

본 논문에서는 계층적 클러스터링 방법에 대한 여러 가지 리프오더링 방법들의 접근 방식을 Overall 리프오더링 방식으로 정의하고, 새로운 리프오더링 알고리즘 적용 방식인 클러스터별 리프오더링 방식을 제안하였다. 그리고, 제안한 방법을 몇 가지 리프오더링 알고리즘을 직접 적용하여 비교 실험하여서 결과를 분석하였다.

앞으로의 연구에서는 본 연구에서 제시한 클러스터별 리프오더링 방식에 대한 전산학적인 분석이외에 생물학적인 가치의 분석이 필요하리라고 여겨진다. 또한, 또 다른 리프오더링 방법이 더 연구될 수 있으리라고 본다.

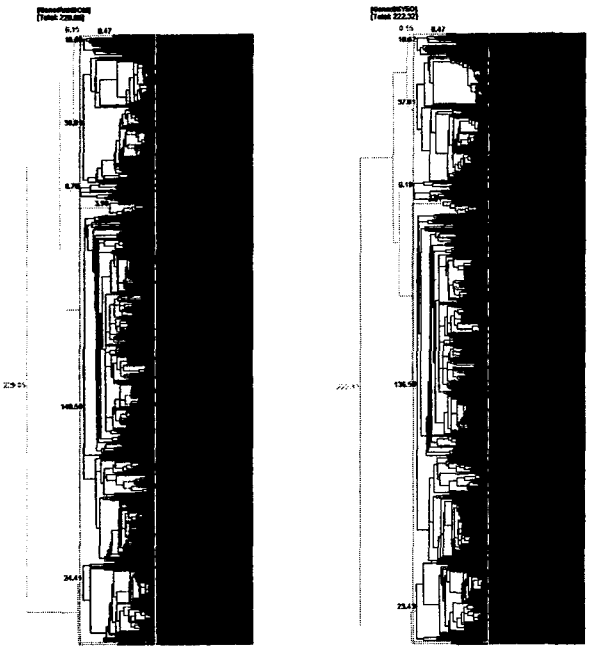

그림 2 실행 결과 화면의 예

참고 문헌

[1]여상수, 김성권, "DNA 마이크로어레이 데이터 클러스터링 알고리즘의 연구 동향", 한국정보과학회 컴퓨터이론 연구회지, 제12권 1호, pp.2-11, 2001년10월.

[2]R. Shamir and R. Sharan, "Algorithmic approaches to clustering gene expression data", *Current Topics in Computational Biology*. MIT Press, submitted.

[3]M. Eisen et al., "Cluster analysis and display of genome-wide expression patterns", *Proc. of Natl. Acad. Sci.*, 95:14863-14867, 1998.

[4]M. Eisen, "Cluster and TreeView Manual", *Eisen Lab. Homepage* (http://rana.lbl.gov/manuals/ClusterTreeView.pdf)

[5]U. Alon et al., "Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays", *Proc. Natl. Acad. Sci.*, 96:6745-6750, 1999.

[6]Z. Bar-Joseph et al., "Fast optimal leaf ordering for hierarchical clustering", *Proceedings of ISMB 2001*. pp.s22-s29.

[7]Z. Bar-Joseph, Therese Biedl, et al., "Optimal Arrangement of Leaves in the Tree Representing Hierarchical Clustering of Gene Expression Data", *Bioinformatics Research Group Homepage of University of Waterloo* (http://monod.uwaterloo.ca/supplements/01expr/art.pdf)

[8]P.T. Spellman et al. "Comprehensive identi- fication of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybri- dization", *Molecular Biology of the Cell* 9:3273-97.1998.

[9]T.R. Golub et al., "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring", *Science*, 286:531-537, 1999.

[10]C. Ding, "Analysis of gene expression profiles: class discovery and leaf ordering", *Proc. RECOMB 2002*, pp.127-136. April 2002.

[11]여상수, 이정원, 김성권, "DNA 마이크로어레이 데이터의 계층적 클러스터링에 대한 리프오더링 알고리즘 개발", *한국정보과학회 2002년 봄 학술발표논문집(A)*, 제29권 제1호, pp.706-708, 2002년 4월.