자기 조직화 지도와 계층적 군집화를 이용한 유전자 발현 데이터 군집화 기법

박창범[1] 이동환[2] 이성환[2]
[1](주)워치비전 기술연구소, [2]고려대학교 정보통신대학 컴퓨터학과
cbpark@watchvision.com, {dhlee, swlee}@image.korea.ac.kr

Clustering of Gene Expression Data by using SOM and Hierarchical Clustering

Changbeom Park[1] Donghwan Lee[2] Seongwhan Lee[2]
[1]WatchVision, Inc., [2]Dept. of Computer Science and Engineering, Korea University

요 약

본 논문에서는 유전자 발현 데이터를 분석하는데 있어서 자기 조직화 지도와 계층적 군집화 기법을 상호 보완적으로 사용하여 사용자가 보다 직관적으로 군집화 결과를 해석할 수 있는 방법을 제안한다. 제안된 방법을 사용하면 빠른 처리 속도로 대용량 데이터 처리에 적합한 자기 조직화 지도의 장점을 살릴 수 있으며 계층적 군집화의 장점인 가시화 기능을 이용하여 자기 조직화 지도의 단점인 군집 경계에 대한 불명확성을 해소하여 군집화 결과를 사용자가 쉽게 이해하고 직관적으로 해석할 수 있도록 도와준다. 본 논문에서 제안된 방법의 효용성을 검증하기 위해 세 종류의 데이터를 사용하여 실험을 수행한 결과 제안된 방법이 기존 방법에 비해 더 나은 성능을 보이는 것을 확인할 수 있었다.

1. 서 론

마이크로어레이는 수많은 DNA를 유리 표면에 매트릭스 형태로 심어 놓은 것으로 유전자의 발현 측정을 목적으로 널리 사용되는 방법이다[1]. 마이크로어레이를 이용하면 최소한 수백 개 이상의 유전자를 동시에 비교 분석할 수 있으므로 몇 개의 관련된 유전자만을 분석하던 기존 방법과 비교하여 매우 획기적인 방법으로 각광받고 있다. 마이크로어레이를 이용하여 실험한 후 촬영한 영상을 분석하면 각 유전자의 발현되는 양과 비율을 측정할 수 있는데 이것을 유전자 발현 데이터라고 부른다. 유전자 발현 데이터는 개체별, 세포별, 시기별, 건강 상태별로 각각 다른 상태를 보여 질병이나 다양한 생명현상과 관련된 생물체의 특성을 파악할 수 있는 매우 중요한 자료로 최근 활발히 연구되고 있다.

유전자 발현 데이터 분석을 통해 한 개체 내에서 비슷한 발현 값을 가지는 유전자는 비슷한 기능을 가진다는 사실에 기초하여 한 개체 내의 유전자를 기능별로 분류할 수도 있고[2], 같은 유전자라도 서로 다른 환경에서는 서로 다른 발현 값을 가진다는 사실을 이용하여 환자의 병을 진단할 수도 있다[3]. 현재까지 기능이 밝혀진 유전자의 수가 많지 않으므로 유전자 발현 데이터를 분석하는 경우에 일반적으로 군집화 방법을 사용한다. 초기에는 계층적인 군집화 방법으로 분석을 많이 시도하였지만 계층적 군집화 방법은 강건성이 부족하고 결과가 일정하지 않아서 신뢰할 수 있는 결과를 얻을 수 없었다[2]. 특히, 입력 데이터의 크기가 커질수록 이러한 문제점이 더 심해지는 경향이 있기 때문에 사용 가능한 유전자 발현 데이터가 점점 더 쌓여갈수록 이러한 단점은 더욱 커지게 되었다. 이러한 문제점을 해결하기 위해 SOM(Self Organizing Map)과 같은 신경망을 기초로 한 방법을 사용하게 되었다. SOM은 입력 데이터의 크기에 제한을 받지 않을 뿐만 아니라 빠른 수행 속도를 보여 대규모 유전자 발현 데이터 분석에 널리 사용되고 있다. SOM은 이차원 맵을 이용하기 때문에 군집화 결과의 가시화가 좋다는 장점이 있지만 군집들의 경계를 찾기가 어렵다는 점과 직관적으로 해석이 불가능하다는 단점이 있다[2,4]. 따라서, 본 논문에서는 SOM을 이용한 군집화 방법과 계층적인 군집화 방법의 장점을 살려 먼저 SOM을 사용하여 유전자 발현 데이터를 군집화하여 데이터를 축약한 후 계층적 군집화 방법을 적용하여 군집들의 경계를 명확히 해주는 동시에 군집화된 데이터의 상호 관계를 트리 구조로 표현하여 사용자가 직관적으로 군집화 결과를 이해할 수 있도록 하는 방법을 제안하고자 한다.

2. 자기 조직화 지도와 계층적 군집화를 이용한 유전자 발현 데이터 군집화 기법

2.1 제안된 기법 소개

본 논문에서 제안하는 방법은 SOM과 계층적 군집화 방법의 장점을 최대한 살린 방법으로 SOM의 결과인 2차원 맵을 계층 구조로 다시 변환해주는 방법이다. SOM의 결과 맵은 입력 데이터인 유전자 발현 데이터의 대표자 역할을 하는 셀들로 이루어져 있다. 각각의 셀들은 SOM의 연산과정에서 주어진 반복 횟수만큼 트레이닝이 되어 입력 데이터와 유사하게 바뀐 것이다. 그러나 2차원 맵을

이용해 각각의 셀들이 어떤 관계를 가지고 있는지 해석하기는 어렵다. 그래서 SOM의 결과에 대해서 보다 가시적이고 직관적인 해석을 가능하게 하기 위해 계층적 군집화 방법을 적용하였다. 계층적 군집화 방법은 데이터 양이 적은 경우에 군집화 결과를 덴드로그램을 통해 빠르고 쉽게 가시적으로 나타낼 수 있는 장점이 있다.

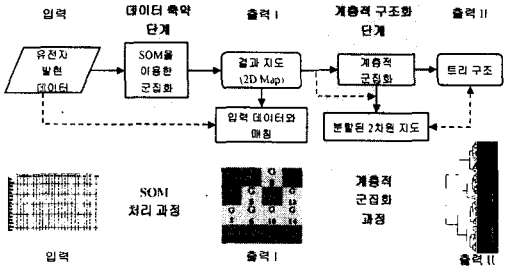

그림 1. 제안된 방법의 개요도

2.2 유전자 발현 데이터 분석

2.2.1 SOM을 통한 데이터 추상화

마이크로어레이 실험을 통하여 얻는 유전자 발현 데이터는 그 양이 상당하다. 그러므로 유전자 발현 데이터를 계층적으로 군집화를 하여 사용자가 직관적으로 이해하기 위해서는 데이터 추상화 과정을 통해 축약하는 과정이 필요하다. 본 논문에서 제안하는 유전자 발현 데이터 군집화 방법은 데이터 추상화에 좋은 특성을 보이는 SOM을 사용하여 데이터를 축약하였다. SOM은 n차원의 입력 데이터를 사용자가 설정한 2차원 맵 공간으로 매핑하는데 실제로 데이터가 이동하는 것이 아니라 2차원 맵 공간에 설정해 놓은 데이터들을 입력 데이터와 비교해가면서 지정된 횟수만큼 훈련하게 되면 2차원 맵 공간에 존재하는 모든 셀들이 입력 데이터를 대변하는 역할을 하게 되는 것이다. 아래 그림 2는 SOM을 이용하여 입력 데이터를 훈련하는 과정을 보여준다.

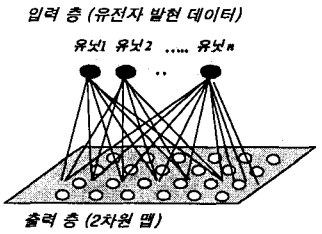

그림 2. SOM 훈련 과정

2.2.2 2차원 결과맵 생성

훈련이 끝난 후 가중치 벡터는 SOM의 결과인 2차원 맵의 각각의 셀에 해당하는 데이터가 되며 각 셀은 입력 데이터들의 군집을 대표하는 성격을 가지게 된다. 이러한 특성을 이용하여 SOM을 이용한 군집화가 가능하지만 결과 맵을 해석하는 부분이 문제로 남게 된다. 군집화는 되어있는데 각각의 군집들이 어떠한 관계를 갖는지 정확히 파악하는 것이 SOM을 군집화 도구로 사용하는데 있어 가장 어려운 부분이다. 아래 표 1은 [5]에서 사용한 동물의 특성을 이용하여 구별하는 간단한 데이터 집합에 대한 SOM의 결과 맵이다. 결과 맵을 분석하면 같은 셀 안에 있는 개체들은 서로 비슷하다는 것은 알 수 있지만 각 셀간의 관계는 해석하기 어렵다.

표 1. SOM의 결과 맵

	0	1	2	3
0	Fox	Dog Wolf		Horse Zebra Cow
1	Cat		Tiger Lion	
2				
3	Eagle	Owl Hawk	Dove	Duck Goose Hen

2.3 SOM의 2차원 결과 맵 처리

2.3.1 2차원 맵에 대한 계층적 군집화

위에서 제기한 문제점처럼 SOM을 이용하여 2차원 맵으로 매핑한 결과를 각각의 입력 데이터와 연결하면 2차원 맵을 구성하는 각 셀끼리 관계가 있음을 결과 맵으로 짐작할 수는 있지만 전체적인 상관 관계는 알 수 없다. 예를 들어 SOM만을 이용할 경우 [Owl, Hawk]가 같은 군집에 속하므로 서로 관계가 있다는 것은 알 수 있지만 [Eagle]과 [Owl, Hawk]간의 관계는 정확히 판단할 수 없으며 [Eagle]과 [Owl, Hawk]의 관계가 [Dove], [Owl, Hawk]와의 관계와 비교하여 어떤 차이가 있는지 알기 힘들다. 이러한 문제를 해결하기 위해 본 논문에서는 SOM의 결과 맵에 대해 각각의 셀들이 서로 어떤 관계에 있는지 직관적으로 알 수 있도록 결과 맵을 구성하는 각 셀에 대해 계층적인 군집화를 수행한다. 아래 그림 3은 SOM의 2차원 결과 맵을 계층적 군집화를 통해 트리구조로 표현한 것이다.

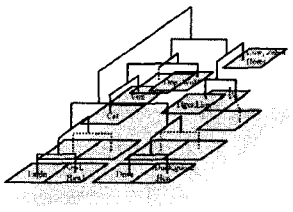

그림 3. SOM 결과 맵의 군집화

2.3.2 덴드로그램(Dendrogram) 생성

제안된 방법을 사용하여 그림 4처럼 SOM의 결과를 다시 계층적으로 군집화하여 덴드로그램으로 나타내면 두 셀간의 거리와 위치, 관련도 등을 정량화하여 쉽게 파악할 수 있어 [Eagle]과 [Owl, Hawk]가 매우 가까운 관계임을 쉽게 알 수 있다.

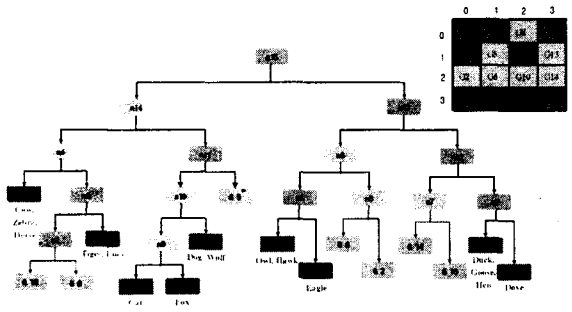

그림 4. SOM의 결과 맵을 계층적으로 표현한 덴드로그램

3. 실험 및 결과 분석

본 논문에서 제안된 방법의 효용성을 검증하기 위해 동물의 특성데이터, 효모 유전자 발현 데이터, 백혈병 유전자 발현 데이터 이렇게 세 종류의 데이터를 사용하여 실험하였다. 동물의 특성 데이터로 분석한 결과는 본 문에서 이미 소개하였기 때문에 여기에서는 생략한다.

전체 6178개의 유전자로 구성된 효모 유전자 발현 데이터를 군집화하여 8개의 클래스를 가지는 유전자가 어느 정도의 정확도로 군집화가 되는지 실험하였다. SOM을 사용하여 실험할 경우 결과 맵의 크기를 결정하는 문제가 발생하는데 이는 알고리즘의 근본적인 문제로 본 실험에서는 결과 맵의 크기를 아래 표 2와 같이 4가지로 정하여 실행하였다. 기존의 방법과 비교하기 위해서 K-평균 방법과 SOM + K-평균 방법을 사용하였다. 8개의 클래스로 구성된 것을 미리 알고 있었으므로 K-평균 방법에서 K를 8로 놓고 실험하였다. 결과를 살펴보면 7 x 7 이상의 맵에서 다른 방법에 비해 좋은 결과가 나왔으며 맵의 크기가 커질수록 보다 더 좋은 결과를 나타냄을 알 수 있다. 실제 클래스 수보다 결과 맵의 크기를 훨씬 크게 설정하여 SOM의 판별력을 이용하여 각 클래스를 몇 개의 대표적인 패턴으로 만들어 각 유전자의 미세한 차이들을 제거한 후 결과 맵 상의 셀들의 유사도를 이용하여 다시 계층적 군집화하므로 좋은 결과를 얻을 수 있었다. 실제 클래스의 수를 미리 알고 있을 경우 좋은 결과를 보인다고 알려진 K-평균 방법, SOM+K-평균 방법보다도 더 나은 결과를 나타내어 성공적인 실험 결과를 보여주였다.

표 2. 효모의 8 클래스 군집화 결과

실험 결과 비교		K-평균	SOM+ K-평균	제안한 방법			
SOM 맵 크기		K=8	K=8	5x5	7x7	10x10	14x14
Yeast Gene Class	CLN2	92%	94%	84%	88%	94%	98%
	Y'	90%	89%	82%	92%	95%	97%
	Histone	97%	90%	88%	92%	98%	100%
	MET	90%	92%	75%	86%	92%	95%
	CLB	89%	89%	84%	90%	94%	97%
	MCM	92%	94%	88%	92%	96%	98%
	SIC1	82%	92%	79%	87%	93%	96%
	MAT	98%	92%	88%	92%	97%	100%

백혈병 유전자 발현 데이터는 38명의 환자의 유전자 데이터로 구성되어 있으며 각각 7129개의 유전자로 이루어져 있다. 38개의 데이터 집합 중 27개는 ALL형이고, 11개는 AML형에 해당한다. 이 실험은 입력 데이터의 행을 군집화하는 것이 아니라 열을 군집화하는 것이 앞 실험과 다르다. ALL형과 AML형을 군집화하는 실험을 수행한 결과 K-평균 방법과 SOM+K-평균 방법은 AML형을 판별할 경우 제안한 방법보다 나은 결과를 보였지만 상대적으로 빈도가 높은 ALL형에서 훨씬 많은 오류를 보여 전체적으로는 제안된 방법이 가장 좋은 성능을 보였다.

표 3. 백혈병 데이터 실험 결과

실험 결과 비교		K-평균	SOM + K-평균	제안한 방법
백혈병 유전자 발현 데이터	ALL	60%	74.07%	96.3%
	AML	100%	100%	90.9%
	전체	71.58%	79.02%	94.74%

세 종류의 실험 데이터를 사용하여 실험을 수행한 결과 본 논문에서 제안한 군집화 방법이 기존 방법에 비해 더 나은 성능을 보이는 것을 확인할 수 있었다. 제안된 방법은 SOM을 기본으로 하기 때문에 결과 맵의 크기에 따라 입력 데이터의 추상화 정도가 달라져서 계층적 트리 구조로 표현했을 때 잘못된 군집화가 진행될 가능성이 있다. 향후 과제로 유전자 발현 데이터를 보정하는 부분과 군집화 결과의 신뢰도를 측정하는 방법 등도 고려해야할 부분이다.

감사의 글

"본 연구는 보건복지부 보건의료기술진흥사업의 지원에 의하여 이루어진 것임.(02-PJ1-PG11-VN01-SV06-0029)"

참고 문헌

[1] M. Johnston, "Gene chips: Array of hope for understanding gene regulation," Current Biology, Vol. 8, pp. R171-R174, 1998.

[2] A. Sugiyama and M. Kotani, "Analysis of gene expression data by using self-organizing maps and k-means clustering," Proc. of the 2002 Int. Conf. on Neural Networks, Hawaii, USA, May 2002, pp. 1342-1345.

[3] M. Sultan et al., "Binary tree-structured vector quantization approach to clustering and visualizing microarray data," Bioinformatics, Vol. 18, Suppl. 1, pp. S111-S119, 2002.

[4] K. Horimoto and H. Toh, "Statistical estimation of cluster boundaries in gene expression profile data," Bioinformatics, Vol. 17, No. 12, pp. 1143-1151, 2001.

[5] J. A. F. Costa and M. L. A. Netto, "Automatic data classification by a hierarchy of self-organizing maps," Proc. of IEEE Int. Conf. on Systems, Man and Cybernetics, Tokyo, Japan, October 1999, pp. 419-424.