Hierarchical Clustering of Gene Expression Data Based on Self Organizing Map
자기 조직화 지도에 기반한 유전자 발현 데이터의 계층적 군집화

Changbeom Park[1*], Donghwan Lee[2], Seongwhan Lee[2,3]

[1] **WatchVision, Inc., Seoul, Korea**
[2] **Department of Computer Science and Engineering, Korea University, Seoul, Korea**
[3] **Interdisciplinary Graduate Program in Bioinformatics, Korea University, Seoul, Korea**

[*]**To whom correspondence should be addressed. E-mail: cbpark@watchvision.com**

Abstract

Gene expression data are the quantitative measurements of expression levels and ratios of numberous genes in different situations based on microarray image analysis results. The process to draw meaningful information related to genomic diseases and various biological activities from gene expression data is known as gene expression data analysis. In this paper, we present a hierarchical clustering method of gene expression data based on self organizing map which can analyze the clustering result of gene expression data more efficiently. Using our proposed method, we could eliminate the uncertainty of cluster boundary which is the inherited disadvantage of self organizing map and use the visualization function of hierarchical clustering. And, we could process massive data using fast processing speed of self organizing map and interpret the clustering result of self organizing map more efficiently and user-friendly. To verify the efficiency of our proposed algorithm, we performed tests with following 3 data sets, animal feature data set, yeast gene expression data and leukemia gene expression data set. The result demonstrated the feasibility and utility of the proposed clustering algorithm.

Introduction

DNA 칩 혹은 DNA 마이크로어레이는 자동화된 기계와 전자 제어 기술 등을 이용하여 적게는 수백 개부터 많게는 수십만 개의 cDNA를 유리 표면에 매트릭스 형태로 부착한 것으로 유전자 수준에서 변이를 측정하는데 매우 유용한 도구로 최근 각광받고 있다[1]. DNA 마이크로어레이를 이용하면 최소 수백 개 이상의 유전자를 동시에 비교 분석할 수 있으므로 관심 있는 몇 개의 유전자만을 분석하던 기존 방법과 비교하여

실험에 소요되는 비용과 시간을 획기적으로 절약할 수 있다. DNA 마이크로어레이 실험 영상을 분석하여 정량화한 자료와 유전자 정보를 결합하면 특정 상황에서 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 생물체의 특성을 파악할 수 있는 매우 중요한 자료이며 이를 분석하여 질병이나 다양한 생명현상과 관련된 유용한 정보를 얻어내는 과정을 유전자 발현 데이터 분석이라고 부른다. 한 개체 내에서 비슷한 발현 값을 가지는 유전자는 비슷한 기능을 가진다는 사실에 기초하여 한 개체 내의 유전자를 기능별로 분류할 수도 있고[3], 또한 같은 유전자라도 서로 다른 환경에서는 서로 다른 발현 값을 가진다는 사실을 이용하여 환자의 병을 진단하기도 한다[2]. 아직까지 유전자나 단백질에 대한 특성이나 기능의 파악이 거의 이루어지지 않은 상태이므로 유전자 발현 데이터를 분석하는 경우에는 일반적으로 군집화 방법을 사용한다.

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 명확하게 해주는 동시에 군집화된 데이터의 상호 관계를 트리 구조로 표현하여 사용자가 직관적으로 군집화 결과를 이해할 수 있도록 하는 기법을 제안하고자 한다. 본 논문에서는 먼저 유전자 발현 데이터를 군집화하는 관련된 연구에 대해 살펴보고 이러한 관련 연구를 바탕으로 제안한 SOM에 기반한 유전자 발현 데이터의 계층적 군집화 기법에 대해 소개한다. 다음으로 제안된 기법의 성능을 실험을 통해 기존 연구 결과와 비교 분석하고 결론 및 향후 연구 방향에 대해 언급한다.

Related Works

현재까지 다양한 유전자 발현 데이터 군집화 기법들이 개발되어 사용되고 있는데 그중에서 가장 널리 사용되고 있는 계층적 군집화 방법과 K-평균 방법 그리고 SOM(Self Organizing Map)에 대해서 알아보고 혼합형 군집화 방법에 대해 알아보기로 하겠다.

계층적 군집화 방법

군집화 방법의 기본이 되는 방법으로 크게 가장 가까운 관측 값들을 묶는 병합 방법과 가장 먼 관측 값들을 나누어가는 분할 방법으로 나눌 수 있다. 계층적 군집화를 수행하면 한 군집이 다른 군집의 내부에 포함되는 형태로 매 단계 마다 계층적인 구조를 이루게 된다. 이러한 계층적 구조를 군집들간의 유사도 혹은 거리와 함께 표현하여 이진 트리 형태로 도식화한 것을 덴드로그램(dendrogram)이라고 부른다. 덴드로그램을 분석하면 군집들의 계층적 구조뿐만 아니라 군집들 간의 구조적 관계까지도 쉽게 살펴 볼 수 있는 장점이 있다.

계층적 군집화 방법은 개체수가 적을 경우에는 간단하고 가시화가 좋다는 장점이 있지만 개체수가 많은 경우에는 군집화 과정에서 잘못된 병합이 일어날 수 있는 경우가 많아지게 되고 이러한 경우에 오차 수정이 불가능하기 때문에 오류 발생 가능성이 높아지게 되는 단점이 있다.

K-평균 방법

사전에 결정된 군집수 K에 기초하여 전체 데이터를 상대적으로 유사한 K개의 군집으로 구분하는 방법이다. 이 방법은 계층적인 군집화 방법에 비해 계산량이 적어 유전자 발현 데이터와 같이 데이터의 양이 많은 집합에 보다 적합한 방법이라고 할 수 있다. 그리고 계층적 군집화 방법에서 단점으로 지적되었던 부적절한 병합이 발생한

경우에도 나중에 수정이 가능하다는 장점이 있다. 그러나 K-평균 방법은 초기에 결정된 군집 수에 따라 그 결과가 민감하게 반응하기 때문에 K값을 설정하는 일이 쉽지 않다는 단점이 있다. 이러한 이유로 일반적인 경우 K-평균 방법은 자료의 시각화를 통해 최적의 군집 수를 미리 알 수 있는 경우나 다양한 K값을 선택하여 군집 분석을 수행한 후에 가장 좋다고 생각되는 K값을 사용하는 방식으로 군집화한다[3].

SOM을 이용한 군집화 방법

SOM(Self Organizing Map)은 1981년 헬싱키대학의 Tuevo Kohonen이 소개한 무감독 경쟁 학습 방법이다[5]. SOM은 복잡한 다차원 실험 데이터의 처리에 알맞고, 가시화가 쉬우며, 빠른 처리 속도와 데이터의 크기에 많은 영향을 받지 않는다는 장점 때문에 널리 사용되고 있다.

SOM은 입력 데이터의 차원을 표현하는 n개의 입력 노드들과 2차원의 출력 노드 집합으로 구성되어 있다. 모든 입력 노드들은 모든 출력 노드들과 연결되어 있으며 연결 가중치를 가진다. 훈련 단계에서 사용된 데이터와 비슷한 데이터가 입력으로 들어오면 2차원 맵상에서 가장 유사한 노드가 승자가 되고 해당 노드로 분류된다.

SOM의 알고리즘을 간단히 요약하면, 특정한 입력에 대해 가장 잘 반응하는 출력 노드와 그 주위의 노드를 학습시킴으로써 자연스럽게 그 입력에 대한 대표자로써의 역할이 가능하게 하는 것이다. 이러한 특징 때문에 유전자 발현 데이터 군집화에 SOM을 많이 사용되고 있다.

혼합형 군집화 방법

최근 들어 유전자 발현 데이터를 군집화하는 경우에 혼합형 군집화 방법을 사용하는 경우가 늘어나고 있는데 이것은 어떤 특정한 군집화 방법을 사용하여 원하는 결과를 얻기가 어려울 뿐만 아니라 결과가 나왔다고 하더라도 해석하는데 어려운 경우가 많이 발생하기 때문이다. 이러한 문제점을 해결하고자 다양한 연구가 진행되고 있는데 그 중에서 본 논문에서 제안한 기법과 관련 있는 SOM을 이용한 혼합형 군집화 방법에 관해서 살펴보도록 하겠다.

먼저 SOM을 군집화 도구로 사용하지 않고 단지 유전자 정보를 표현하기 위한 가시화 도구로 사용하는 연구가 있다[2]. 먼저 군집화 도구로 K-평균 방법을 사용하여 K값을 2로 설정하여 계속해서 나누어 나가는 방식으로 트리 구조를 형성한다. 여기에서 SOM은 한 개체의 유전자 정보를 표현하기 위한 수단으로만 사용된다.

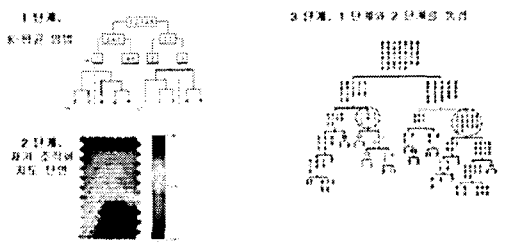

그림 1. K-평균을 이용한 군집화 및 SOM을 이용한 가시화 방법

다음의 방법들은 SOM의 2차원 결과 맵의 군집 경계가 불확실하기 때문에 이를 보완하기 위해서 방법들이다.

먼저 SOM의 결과를 U-matrix 형태로 바꿔준 후 watershed 방법과 형태처리 모폴로지 기법을 사용하여 영상을 분할한 다음 다시 각각의 군집에 SOM을 적용하여 계층화된 결과를 도출하는 연구가 있었다[4].

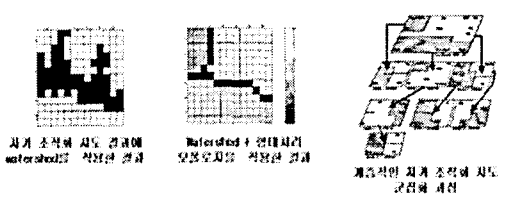

그림 2. Watershed와 형태처리 모폴로지를 사용하여 계층적으로 SOM을 적용

SOM의 결과 맵을 K-평균 방법을 이용하여 군집화하는 방법도 연구되었다[3]. 이 방법은 SOM으로 충분히 군집을 확보해 놓은 다음 K-평균 방법으로 이들을 다시 군집화하는 방법을 취한다.

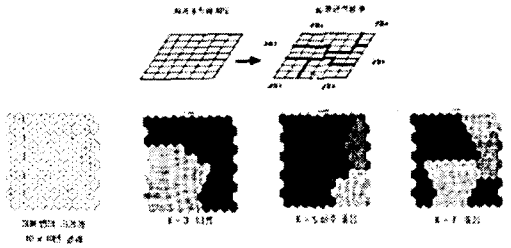

그림 3. SOM의 결과 맵에 K-평균을 적용한 군집화 방법

이 방법은 K-평균 방법의 K값에 따라 결과가 달라진다는 단점이 있다. 이 문제는 K-평균의 방법을 사용하는 경우에 발생하는 근본적인 문제점으로 미리 군집화 대상의 군집 특성을 알 수 있는 경우는 좋은 결과를 얻을 수 있지만 예측이 불가능한 경우는 좋은 성능을 보여주지 못한다.

Methods

본 논문에서 제안하는 혼합형 군집화 기법은 SOM과 계층적 군집화 방법의 장점을 최대한 살린 방법으로 SOM의 결과인 2차원 맵을 계층 구조로 다시 변환한다. SOM 결과 맵의 각 셀들은 훈련 과정을 통해 입력 데이터인 유전자 발현 데이터의 대표자 역할을 하는 값을 가지게 된다. 그러나 2차원 맵을 가지고는 각각의 셀들이 어떤 관계를 가지고 있는지 해석하기가 어렵다. 따라서, SOM의 결과에 대해서 보다 가시적이고 직관적인 해석을 가능하게 하기 위해 계층적 군집화 방법을 적용하였다. 그림 4에 본 논문에서 제안한 방법을 이용한 유전자 발현 데이터의 군집화 과정을 나타내었다.

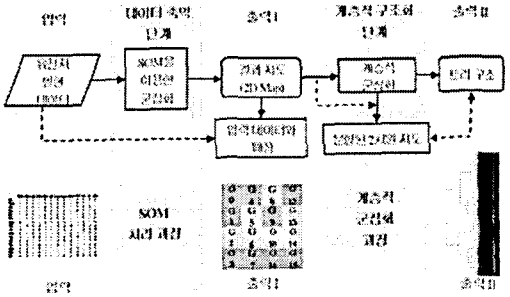

그림 4. 제안된 군집화 과정 개요도

유전자 발현 데이터 처리

SOM을 통한 데이터 추상화

마이크로어레이 실험을 통하여 얻는 유전자 발현 데이터는 그 양이 상당하므로 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 매핑하는 것이 가능한데 실제로 데이터가 이동하는 것이 아니고 2차원 맵 공간에 설정해 놓은 데이터들을 입력 데이터와 비교해가면서 지정된 횟수만큼 훈련하게 되면 2차원 맵 공간에 존재하는 모든 셀들이 입력 데이터를 대변하는 성격을 가지게 되는 것이다. 아래 그림 5에 SOM의 개요도를 나타내었다.

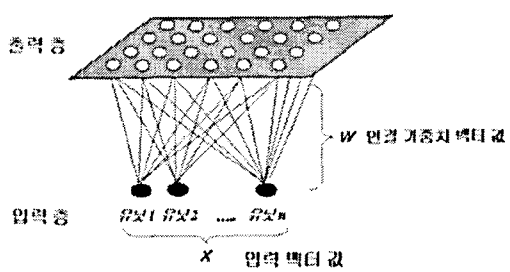

그림 5. SOM 개요도

2차원 맵 생성

훈련이 끝난 후 SOM의 결과인 2차원 맵의 각각의 셀은 입력 데이터의 군집을 대표하는 성격을 가지게 되지만 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. SOM의 결과 맵에 대한 예를 표 1에 나타내었다.

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

표 1은 [4]에서 사용한 동물의 특성을 이용하여 구별하는 간단한 데이터 집합에 대한 SOM의 결과 맵이다. 4 x 4 크기의 결과 맵 각 셀에 훈련된 데이터가 들어있다. SOM의 결과 맵을 분석하면 같은 셀안에 있는 개체들은 서로 비슷하다는 것을 알 수 있지만 각 셀간의 관계는 해석하기 어렵다. 예를 들면 [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.

SOM의 결과인 2차원 맵 처리

2차원 맵에 대한 계층적 군집화

위에서 제기한 문제점처럼 SOM을 이용하여 2차원 맵으로 매핑한 결과를 각각의 입력 데이터와 연결하고 나면 2차원 맵을 구성하는 각 셀끼리는 무언가 관계가 있음을 결과 맵만으로 짐작할 수는 있지만 전체적인 셀들의 관계는 알 수가 없다. 이러한 2차원 맵의 특성을 이용하여 각각의 셀들이 서로 어떤 관계에 있는지 직관적으로 알 수 있도록 2차원 맵을 구성하는 셀을 계층적인 군집화를 통해서 트리구조로 표현하였다.

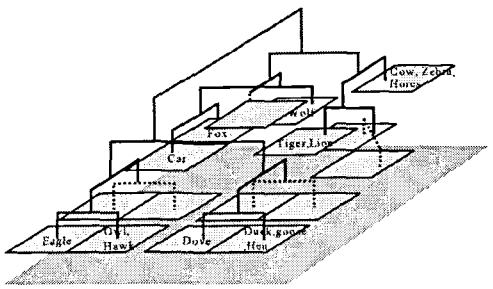

그림 6. SOM 결과 맵의 계층적 군집화

덴드로그램(Dendrogram) 생성

본 논문에서 제안한 방법을 사용하여 SOM의 결과 맵의 각 셀을 계층적 군집화를 통해 덴드로그램으로 표현하면 각 셀들간의 거리와 위치, 관련도 등을 정량화하여 쉽게 파악할 수 있다. 그림 7에 계층적 군집화를 수행하여 최종적으로 얻은 덴드로그램을 나타내었다.

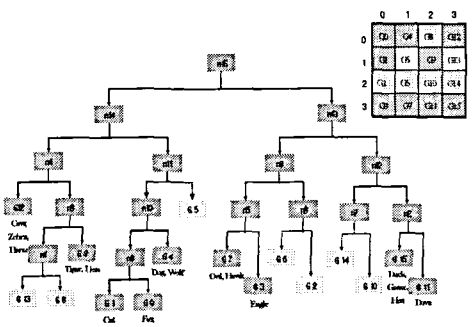

그림 7. SOM의 결과 맵에 대해 계층적 군집화를 수행하여 구한 덴드로그램

그림 7의 계층적 군집화 결과를 보면 일반적으로 볼 때 비슷한 습성을 갖는 동물인 [eagle], [owl, hawk]가 같은 군집에 속하고 [dove], [duck, goose, hen]이 계층적 군집화 과정을 통해 같은 군집에 속하는 것을 볼 수 있어 우수한 군집화 성능을 보여주고 있음을 알 수 있다.

Experiment and Results

제안된 알고리즘의 효용성을 검증하기 위해 동물의 특성 데이터, 효모 유전자 발현 데이터, 백혈병 유전자 발현 데이터를 사용하여 실험하였다.

동물의 특성 데이터

동물의 특성 데이터는 친숙한 동물을 데이터로 사용하여 일반적인 상식으로 군집화의 결과를 쉽게 판별할 수 있기 때문에 군집화의 개념을 이해하기 위하여 인용한 데이터이다. [4]에서 사용한 데이터를 인용하였으며 표 2에 실험에 사용한 동물의 특성 데이터를 표시하였다. 16개 동물을 군집화하기 위해 13가지 특성을 적용하여 실험하였다.

표 2. 동물 특성 데이터

	Small	Medium	Big	2 legs	4 legs	Hair	hooves	mane	feathers	hunt	run	fly	swim
Hen	1	0	0	1	0	0	0	0	1	0	0	0	0
Goose	1	0	0	1	0	0	0	0	1	0	0	1	1
Duck	1	0	0	1	0	0	0	0	1	0	0	1	1
Dove	1	0	0	1	0	0	0	0	1	0	0	1	0
Owl	1	0	0	1	0	0	0	0	1	1	0	1	0
Hawk	1	0	0	1	0	0	0	0	1	1	0	1	0
Eagle	0	1	0	1	0	0	0	0	1	1	0	1	0
Cat	1	0	0	0	1	1	0	0	0	1	0	0	0
Fox	0	1	0	0	1	1	0	0	0	1	0	0	0
Wolf	0	1	0	0	1	1	0	0	1	1	1	0	0
Dog	0	1	0	0	1	1	0	0	0	1	1	0	0
Tiger	0	0	1	0	1	1	0	0	0	1	1	0	0
Lion	0	0	1	0	1	1	0	0	1	1	1	0	0
Horse	0	0	1	0	1	1	1	1	1	0	1	0	0
Zebra	0	0	1	0	1	1	1	1	0	0	1	0	0
Cow	0	0	1	0	1	1	1	0	0	0	0	0	0

동물의 특성 데이터를 이용하여 실험한 결과들은 본문에서 내용을 소개하면서 예제로 보여주었기 때문에 여기에서는 생략한다. 실험 결과인 그림 7을 간단히 살펴보면 일반적인 상식으로 분류할 수 있는 기준과 유사한 결과가 나왔음을 알 수 있으며 데이터를 제공한 [4]에서의 결과와도 일치함을 확인할 수 있다.

효모의 발현 데이터

이 데이터는 스탠포드 대학에서 제공하는 효모의 유전자 발현 데이터이다(http://genome-www.stanford.edu/cellcycle). 전체 6178개의 유전자로 구성이 되어 있으며 실험 조건은 77개로 구성이 되어 있는 데이터이다[6].

제안한 방법론이 타당성이 있는지 확인하기 위해 HHF1(YBR009c)라는 특정 유전자와 가장 비슷한 발현 데이터를 갖는 50개의 유전자를 스탠포드 대학에서 웹으로 제공하는 데이터 베이스를 이용하여 추출하였다.

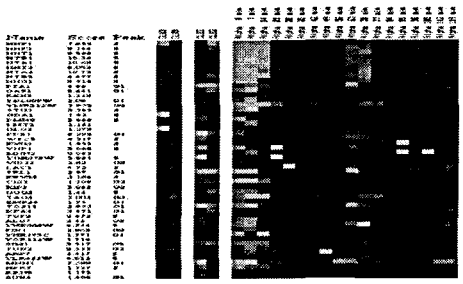

그림 8. HHF1에 해당하는 유전자 집합

6178개 효모 유전자 데이터 전체에 대해 제안된 방법을 이용하여 군집화한 결과 HHF1와 비슷한 발현 데이터를 갖는 유전자 50개중 98%에 해당하는 유전자가 전체를 계층화한 군집화 결과 중 HHF1에 해당하는 영역에 속해 있어 정확하게 군집화가 이루어진 것을 실험을 통해 알 수 있었다.

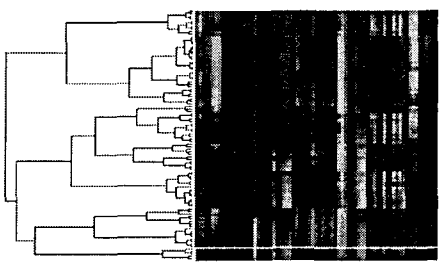

그림 9. 효모 유전자 전체를 군집화한 결과와 HHF1이 속한 영역

다음 실험은 같은 효모 유전자 발현 데이터를 가지고 수행한 실험인데 위의 실험과 약간 다른 의미를 가진다. 전체 효모 유전자를 군집화하여 8개의 클래스를 갖는 유전자가 어느 정도의 정확도로 군집화 되어있는지를 비교하는 것이다. 실험을 적용할 때 SOM의 결과 맵의 크기는 5 x 5, 7 x 7, 10

x 10, 14 x 14의 4가지로 해서 각각 25개, 49개, 100개, 196개의 군집를 가질 수 있도록 하였다. 기존의 방법과 비교하기 위해 K-평균 방법과 SOM + K-평균 방법을 사용하여 실험하였다. 다음 그림 10은 제안된 방법으로 군집화한 결과이다.

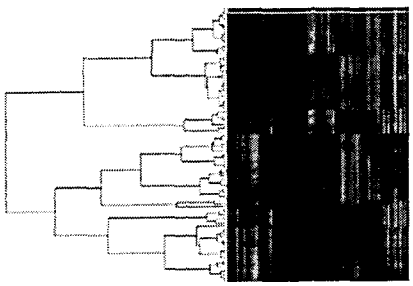

그림 10. 14 x 14 덴드로그램

군집화 방법의 정확도를 측정하기 위해 제안된 방법과 다른 군집화 방법으로 군집화된 집합에서 실제 데이터의 군집과 같은 군집에 있는 개체들의 백분율을 비교하여 표 3에 나타내었다.

표 3. 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%

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 군집화 하므로 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.

백혈병 유전자 발현 데이터 실험

이 데이터는 MIT에서 제공하는 백혈병 유전자 발현 데이터이다(http://www-genome.wi.mit.edu/). 72명의 환자의 데이터로 구성되어 있으며 한 환자의 데이터는 7129개의 유전자로 이루어져 있다. 72명의 데이터는 38명의 훈련 데이터와 34명의 시험 데이터로 구성되어 있는데 본 실험에서는 38명의 훈련 데이터만을 사용하여 실험하였다. 38개의 훈련 데이터 집합 중 27개의 데이터 집합은 ALL 형의 백혈병에 해당하는 것이고 나머지 11개의 데이터 집합은 AML 형의 백혈병에 해당한다. 이 실험은 입력 데이터를 앞에서처럼 행을 군집화 시켜주는 것이 아니라 열을 군집화시켜 준다는 점이 다르다. 제안한 방법과 기존의 방법의 성능 비교를 위해서 K-평균 방법과 SOM + K-평균 방법을 사용하였다.

표 4. 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%

실험 결과를 보면 K-평균 방법과 SOM + K-평균 방법은 AML형을 판별할 경우 제안한 방법보다 우수한 결과를 보였지만 상대적으로 빈도가 높은 ALL형에서 훨씬 많은 오류를 보여 전체적으로는 제안된 방법이 가장 좋은 성능을 보였다.

Discussion

유전자 발현 데이터는 마이크로어레이 실험을 통해 유전자 수준에서 일어나는 변이를 측정한 자료로 생명 현상과 관련된 유용한 정보를 얻을 수 있는 중요한 자료이다.

현재 유전자 발현 데이터 분석에 널리 사용되고 있는 널리 사용되고 있는 계층적 군집화 도구의 경우 가시화 및 각 군집간의 직관적인 이해에 좋은 장점은 있으나 데이터가 많아질수록 오류가 누적되고 처리 시간이 길어지는 단점이 있어 대용량의 데이터 보다는 적은량의 데이터를 군집화 하는데 좋은 장점을 가진다. 자기 조직화 지도의 경우는 대용량의 데이터 처리에 유리하고 계산이 선형적이라 처리 속도가 빠르지만 군집들간의 경계가 명확하지 않아 결과 맵의 해석이 어렵다는 단점이 있다. 이러한 자기 조직화 지도와 계층적 군집화의 방법의 장점을 취하여 유전자 발현 데이터와 같이 많은 양의 데이터와 군집간의 관계 해석이 필요한 경우 유용하게 사용할 수 있도록 두 기법을 결합한 새로운 군집화 방법을 본 논문에서 제안하였다.

본 논문에서 제안된 유전자 발현 데이터 군집화 방법의 효용성을 검증하기 위해 동물의 특성 데이터, 효모의 유전자 발현 정보와 백혈병 유전자 발현 정보를 사용하여 실험하였다. 이 세 종류의 실험 데이터를 사용하여 실험을 수행한 결과 본 논문에서 제안한 방법이 기존 방법에 비해 더 우수한 성능을 보이는 것을 확인할 수 있었다.

제안된 방법은 SOM을 이용하여 데이터를 축약하기 때문에 SOM의 구조적 특성상 제안된 방법을 사용하여 군집화하기 위해서는 초기에 트레이닝 맵을 설정하여야 한다. 트레이닝 맵의 크기에 따라 입력 데이터의 추상화 정도가 달라져서 계층적 군집화 방법을 사용하여 트리구조로 표현했을 때 잘못된 군집에 입력 데이터가 놓여질 가능성이 있다.

향후 연구 과제로는 유전자 발현 데이터의 보정 부분과 군집화 결과의 신뢰도를 측정하는 방법 등도 고려해야할 부분이다.

Acknowledgements

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

References

[1] **J. Setubal and J. Meidanis, Introduction to Computational Molecular Biology, 1997.**

[2] **M. Sultan et al., Binary Tree-structured Vector Quantization Approach to Clustering and Visualizing Microarray Data,** *Bioinformatics*, **18(1), 2002, pp. S111-S119.**

[3] **A. Sugiyama and M. Kotani, Analysis of Gene Expression Data by Using Self-organizing Maps and K-means Clustering, Proc. of the 2002 Int. Joint Conf. on Neural Networks, Hawaii, USA, May 2002.**

[4] **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.**

[5] **T. Kohonen, Self-Organizing Maps, 2001.**

[6] **P. T. Spellman et al., Comprehensive Identification of Cell Cycle-regulated Genes of the Yeast Saccharomyces Cerevisiae by Microarray Hybridization,** *Molecular Biology of the Cell*, **9(12), 1998, pp. 3273-3297.**