응용통계연구 (2009)
22(2), 415–424

자기조직도에서 최소생성나무의 활용

장유진[1] · 허명회[2] · 박미라[3]

[1]고려대학교 통계학과, [2]고려대학교 통계학과, [3]을지대학교 예방의학교실

(2009년 1월 접수, 2009년 1월 채택)

요 약

비지도 학습 신경망모형의 한 종류인 자기조직도(self-organizing map: SOM)는 고차원 자료를 차원축소하고 저차원지도를 통해 유사한 개체를 군집화하는 방법이며 다양한 분야의 데이터에 적용되고 있다. 한편 최소생성나무(minimal spanning tree: MST)는 개체점들을 닫힌 루프 없이 가장 짧게 선분으로 연결하는 그래프 방법이다. 본 연구에서는 부노드 자기조직도에 최소생성나무를 적용하여 부노드 간 거리를 근사적으로 나타내는 자료 시각화 방법과 자기조직도의 최적 형태와 크기를 결정하기 위한 거리 측도를 제안하였다. 또한 피셔의 붓꽃자료와 실제 유전자발현자료 및 모의생성 자료에 적용하여 이 방법의 유용성을 살펴보았다.

주요용어: 자기조직도, 최소생성나무, 자료 시각화, 거리측도.

1. 서론

자기조직도(self-organizing map: SOM)는 고차원 개체공간에 있는 개체벡터를 저차원 그리드 공간에 표현하는 방법으로, 고차원 공간에서 주요한 위상적·계량적 특성을 저차원 공간에서 축약적으로 보여준다 (Kohonen, 1995). 자기조직도는 K-평균 군집화의 대안적 방법으로 유전체 연구를 비롯한 다양한 분야의 자료분석에 활용되고 있다 (Tamayo 등, 1999; Park 등, 2005; Yan, 2006). 최근 Haese와 Goodhill (2001)은 귀납적 모수추정법을 이용하여 자동적으로 학습 모수들을 추정하는 Auto-SOM을 제안하였으며, Berglund와 Sitte (2006)는 학습률(learning rate)과 이웃의 크기(neighborhood size)를 구하지 않아도 되는 PL-SOM을 제안한 바 있다. 이 밖에도 자기조직도를 기초로 하여 개발된 여러 형태의 변종이 활발히 연구되고 있다 (허명회, 2003; 엄익현과 허명회, 2005; Hsu와 Halgamuge, 2003; Hsu, 2006; Samsonova 등, 2006). 한편 최소생성나무(minimal spanning tree: MST)는 그래프상의 점들을 연결하는 최단선분을 구하는 방법으로서 그 자체로 군집화 방법으로 사용될 수 있다 (Xu 등, 2001). 김성수 (1999)와 Kim 등 (2000)은 다차원 자료의 시각화와 탐색을 위하여 다차원척도 그래프 상에 최소생성나무를 활용한 바 있다.

본 연구에서는 자기조직도에 최소생성나무기법을 접목하는 방안을 제시할 것이다. 자기조직도상에 최소생성나무를 구현한 이 방법은 데이터의 군집 및 차원축약이라는 자기조직도의 장점을 살리면서 군집간의 실제거리에 대한 근사적 정보를 함께 제공한다. 또한 이로부터 적절한 지도의 형태 및 크기를 선정하기 위한 통계적 기준 측도를 얻을 수 있다. 2절에서 자기조직도와 최소생성나무의 기본알고리즘을 소개하고, 이들을 결합하는 방법을 제안한다. 3절에서 모의자료와 실제자료에 적용한 사례를 제시하고, 4절에서 결론을 맺는다. 분석에 사용된 프로그램은 SAS/IML로 구현되었다.

이 논문은 정부(교육인적자원부)의 재원으로 한국학술진흥재단의 지원을 받아 수행된 연구임(R14-2003-002-01001-0).

[3]교신저자: (301-832) 대전시 중구 용두동, 을지대학교 예방의학교실, 부교수. E-mail: mira@eulji.ac.kr

2. 자기조직도와 최소생성나무의 결합

2.1. 자기조직도와 최소생성나무의 기본 알고리즘

자기조직도는 분할군집화(partitioning clustering)의 일종으로서 군집에 대한 정보를 시각화된 그래프로 제공하면서 유전자발현의 패턴인식에도 유용한 것으로 알려져 있다. 분석의 목적은 고차원의 입력개체를 저차원 지도(map)상의 몇 개의 노드(node)로 표현하는데 있으며 보통 직사각형태의 2차원 그리드(grid)로 생성된다. 비지도학습 데이터가 n개의 관측과 p개의 변수로 구성되어 있다고 하자. 즉, 입력개체가 p차원의 벡터 $x_1, \ldots, x_n$으로 구성되어 있다고 할 때, 저차원 그리드상에 $r \times c$ 자기조직도를 산출하기 위한 알고리즘은 다음과 같이 요약된다 (Kohonen, 1995; 허명회, 2003).

(i) $m(= r \times c)$개의 p차원 중량벡터 $w_1, \ldots, w_m$에 초기값 $w_{1(0)}, \ldots, w_{m(0)}$을 부여한다.

(ii) 각 입력개체 $x_i (i = 1, \ldots, n)$에 대해 가장 가까운 중량벡터를 찾아 그 중량벡터가 속한 노드 $K(i)$ (승자노드; winner node)에 할당한다. 이 때 거리의 측도로 유클리드 거리를 사용할 수 있다.

(iii) 중량을 업데이트 한다. 즉, $||r_j - r_{K(i)}|| \leq d_t$인 모든 j에 대하여,

$$w_{j(t+1)} = w_{j(t)} + \alpha_t h_{(t)}(K(i), j)(x_i - w_{j(t)})$$

로 업데이트 한다. 여기서, r_j는 노드 j에 대한 그리드 위치점이고, t는 업데이트 시점 $0, 1, 2, \cdots$을 나타내는 정수이다. 학습률(learning rate) α_t는 t에 따라 감소하도록 설정되며, d_t도 t에 따라 감소하는 계단함수이다. $h_t(k, j)$는 국소가중치(local weight)로서 흔히 가우시안 함수가 적용된다.

(iv) 업데이트된 중량벡터가 속한 노드를 찾아 승자노드도 업데이트 한다.

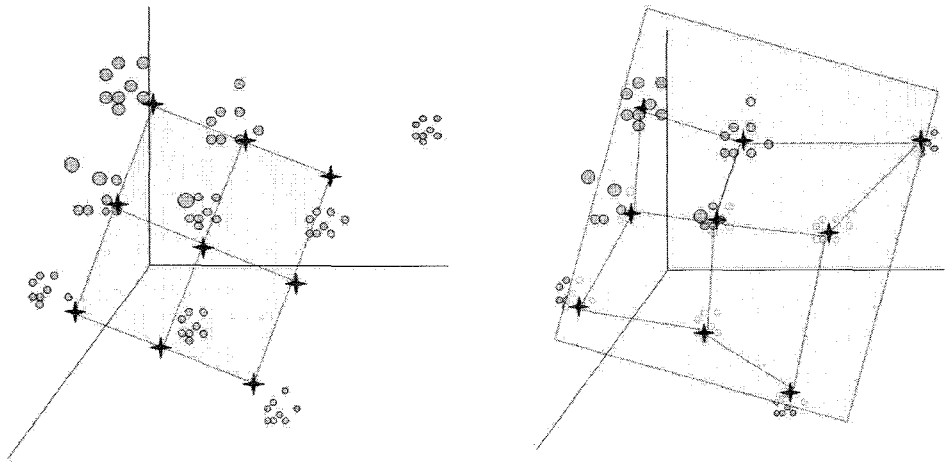
(v) 모든 중량벡터 값의 변화가 거의 없을 때까지 위의 단계를 반복하여 각 개체의 최종 승자노드 및 최종 중량벡터를 구한다. 각 개체들을 각 승자 중량벡터에 연결된 그리드 공간상의 노드로 표현한다.

그림 2.1은 3차원 데이터에서 2차원 그리드를 이용하여 자기조직도를 적용한 예를 보여준다. 그림 2.1 (a)는 3×3의 직사각노드를 설정한 초기노드의 위치를 보여주며, (b)는 반복적인 학습과정을 거쳐 안정된 최종노드와 이에 속하는 개체점들을 보여준다. 2차원 그리드를 설정했을 때에는 이와 같이 초기노드와 최종노드가 2차원 곡면 위에 위치하게 된다. 자기조직도를 적용할 때 장점 중 하나는 가까운 노드에 속하는 개체들은 먼 노드에 속하는 개체에 비해 비슷한 패턴을 갖게 되어 군집의 위치에 따라 순서의 의미를 둘 수 있다는 것이다. 반면, 문제점은 입력요건이 되는 초기학습률, 이웃의 크기, 지도의 크기, 형태 등의 세부 사항들을 모두 분석자가 결정해야 한다는 것이다.

한편, 다차원 공간상에 n개의 점이 있을 때 닫힌 루프(loop)가 없이 점들을 연결한 선분들의 집합을 생성나무(spanning tree)라 하며, $n(n-1)/2$개 선분들의 길이 합이 최소가 되는 생성나무를 최소생성나무(minimal spanning tree: MST)라고 한다 (Gower와 Ross, 1969). 본 고에서는 다음과 같은 Prim의 알고리즘을 이용하였다 (Prim, 1957).

(i) A를 n개의 점인 $a_1, \ldots, a_n$을 원소로 하는 집합이라 하자. $A = \{a_1, \ldots, a_n\}$. 처음 n개의 점 중 한 점 a_i를 선택하여 집합 B의 원소로 넣는다. 즉, $B = \{a_i\}$이 된다.

(ii) 집합 $A - B$의 원소 중 B와 가장 짧은 거리를 갖는 점 a_j를 a_i와 잇고 B에 포함시킨다. 즉, $B = \{a_i, a_j\}$이다.



(a) 초기노드 (b) 최종노드

그림 **2.1.** 자기조직도의 원리 - 3차원자료에서 3 × 3 자기조직도 적용 예 (⋆ :노드, ∘:입력개체, 점크기는 원근을 표시함)

(iii) 집합 $A - B$의 원소 중 B와 가장 거리가 짧은 점을 B에 포함시키고, 짧은 거리를 갖게 한 B의 원소와 그 점을 잇는다.

(iv) 위의 과정을 계속하여 $a_1, \ldots, a_n$이 어느 하나의 점에라도 연결이 되면 멈춘다.

이 때 거리의 측도로서 유클리드 거리를 사용할 수 있다. 최소생성나무는 최단연결법(single linkage)과 밀접한 관련성을 가지고 있어서, 이를 구하여 긴 가지(edge)를 잘라나가면 최단연결법을 이용한 군집화도 직접적으로 구할 수 있게 된다.

2.2. 군집간 거리정보의 시각화 방안

자기조직도를 적용할 때 노드간 거리가 일정한 직사각형 그리드를 상정했다고 하자. 반복절차를 거쳐 최종결정된 노드들은 위상적 순서를 저차원 공간에서 표현하지만 상호적 거리는 유지되지 않는다. 예컨대 $r \times c$ 자기조직도에서 i행 j열에 해당하는 노드를 (i, j)로 표현할 때 노드 $(2,1)$와 $(2,2)$의 실제거리는 매우 가깝지만, 노드 $(2,2)$와 $(2,3)$의 실제거리는 멀리 떨어져 있을 수도 있다. 단순한 자기조직도의 적용은 각 노드에 해당하는 개체정보를 알려줄 뿐이어서 노드 간 거리에 대한 정보를 알 수 없다.

이를 보완하기 위하여 자기조직도 상에서 최종노드를 잇는 최소생성나무를 그리는 방법을 생각하자. 최소생성나무결과에서 짧은 선으로 연결되는 노드들은 먼 거리로 연결되는 노드에 비해 보다 유사한 군집들이 될 것이다. Kohonen의 기본적인 자기조직도상에도 이를 적용할 수 있으나, 보다 유용한 시각화를 위해서는 허명회 (2003)나 엄익현과 허명회 (2005)의 방법과 같이 동일한 승자노드라도 보다 세분화하여 자료를 흩뿌려 놓는 방법을 사용하는 것이 좋다. 여기서는 엄익현과 허명회 (2005)의 부노드 (k) SOM 방법을 적용하기로 한다. 이 방법은 그리드 공간에 승자노드를 중심으로 개체를 적절히 배분함으로써 전체적으로 자료를 퍼뜨려 놓는 효과가 있다. 따라서 다음과 같은 SOM과 MST의 결합 방법을 제안한다.

(i) SOM 단계: 그리드의 형태와 초기 옵션 등을 정하고, 자기조직도를 적용하여 개체별로 승자노드를

구한다.

(ii) 부노드(k) 단계: 승자노드의 상/하/좌/우의 인접노드를 보다 세밀하게 등분하여 이들의 교차점 중 개체벡터와 가장 가까운 점을 부노드로 사용한다.

(iii) MST 단계: 부노드간을 잇는 최소생성나무를 구한다.

이러한 결과로서 자기조직도를 통해 분류된 군집간의 거리를 시각화할 수 있으며, 이 정보를 이용하여 작은 군집들끼리 다시 군집화 할 수도 있다. 또한 최소생성나무의 특성을 살려 긴 가지(edge)를 차례로 제거함으로써 외따로 떨어진 군집을 살펴볼 수도 있게 된다.

2.3. 지도의 크기와 형태 선택을 위한 측도

자기조직도를 적용하기 위해서 분석자는 사전에 지도의 차원과 그리드의 형태와 크기를 정해야 한다. 예컨대 분석자가 2차원 사각형 그리드를 상정하였다 해도, 그리드의 행 수 r과 열 수 c를 미리 정해야 하는데, r과 c를 어떻게 정하느냐에 따라 결과가 달라지므로 적절한 그리드를 찾는 것은 어려운 일이다. 자기조직도의 그리드 크기 및 형태의 선택기준으로서 다음과 같이 최소생성나무를 활용한 측도를 생각해보자.

2.2절에서 구한 최소생성나무에서 2개 개체 x_i, x_j가 있다고 하자. 두 개체 간 거리를 구하고 이를 $d(x_i, x_j)$로 표기한다. 자기조직도를 구할 때 자료를 표준화하였다면, $d(x_i, x_j)$를 다음과 같은 유클리드 거리로 정의할 수 있다.

$$d(x_i, x_j) = \sqrt{(x_i - x_j)'(x_i - x_j)}.$$

다음에 모든 연결된 개체 간 거리의 합 D_{MST}를 구한다. 즉,

$$D_{MST} = \sum_{(i,j)connected} d(x_i, x_j)$$

이다. D_{MST}가 작다는 것은 승자노드의 위치가 가까운 개체가 실제로 개체공간에서도 거리가 가깝다는 것이며, 따라서 적용된 자기조직도의 그리드가 분석 데이터에 적절하였다는 것을 의미하므로, 동일한 크기에서는 여러 형태의 그리드 중 이를 최소화하는 그리드를 선택하면 된다. 지도의 크기가 다를 때에는 크기가 커져서 군집이 세분화될 때 이의 영향으로 D_{MST}가 자연스레 작아질 수가 있으므로 절대적인 값만을 비교하기는 곤란하다. 이보다는 크기에 따른 D_{MST}의 추세를 보고 변동 폭이 줄어든 변화 점을 선택하는 것이 좋을 것이다.

3. 분석사례

본 연구에서 제안한 방법을 실제자료와 모의생성자료에 적용하여 유용성을 살펴보기로 한다. 각 데이터에서 변수들은 평균 0, 표준편차 1이 되도록 표준화되었다. 자기조직도를 구하기 위해 초기학습률 0.25, 최종학습률 0.01, 반복 = 100회로 지정하였고, 부노드(k) SOM의 경우 초기주변거리 = (그리드의 행 수)/2, 최종주변거리 = 1로 지정하였다. 중량벡터의 초기치는 주성분분석결과를 이용하여 설정하였다. 초기 주변거리는 중량벡터를 업데이트하는 범위로 노드 수가 많을수록 업데이트할 범위도 정비례하도록 지정한 것이다. 부노드(k)에서 k는 7로 정하였다.

3.1. 실제자료 사례

제안된 방법을 피셔(R.A. Fisher)의 붓꽃자료와 Alizadeh 등 (2000)의 림프구자료에 적용하여 보았다. 붓꽃자료는 세 가지 품종(Class 1: setosa, Class 2: versicolor, Class 3: virginica)에서 각 50개씩 총

표 **3.1.** 붓꽃 자료에서 그리드에 따른 D_{MST} 비교

비교1 (유사크기)	SOM	7×7	8×6	12×4	16×3	24×2
	D_{MST}	66.58	65.68	73.49	82.19	88.57
비교2 (유사형태)	SOM	4×3	5×4	7×5	8×6	9×7
	D_{MST}	82.01	75.18	67.72	65.68	69.10

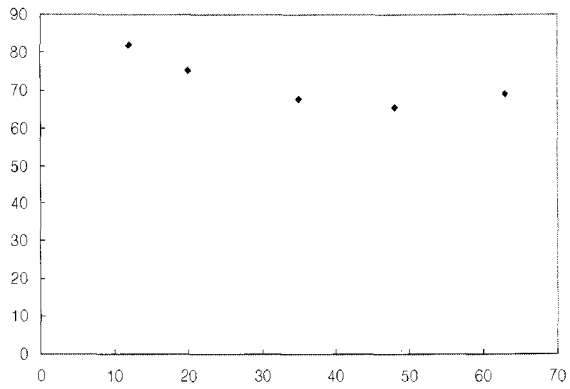
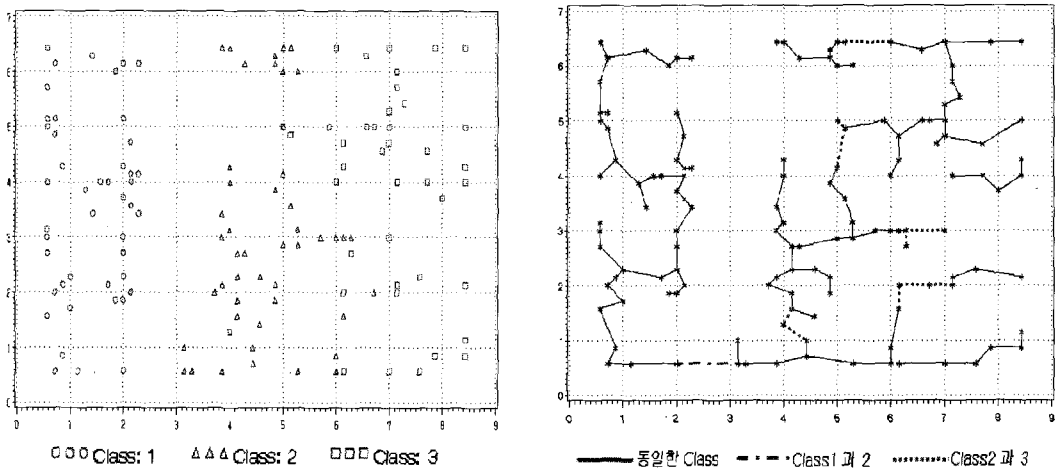


그림 3.1. 붓꽃자료에서 그리드 크기에 따른 D_{MST}

150개 개체에 대해 꽃받침 길이, 꽃받침 너비, 꽃잎 길이, 꽃잎 너비를 측정한 자료이다. 여기서는 품종변수를 제외한 4개 변수를 입력변수로 사용하여 자기조직도를 구하였다. 림프구자료는 성인 림프성 질병의 유전자발현 연구에서 나온 것으로, 림프성 질병으로 B-cell chronic lymphocytic leukemia(B-CLL), follicular lymphoma(FL), diffuse large B-cell lymphoma(DLBCL)의 세 종류를 취하여 62개 샘플(11개의 B-CLL, 9개의 FL, 42개의 DLBCL)에 대한 4,026개 유전자 발현값을 입력변수로 사용하고 분석을 실시하였다.

표 3.1은 붓꽃자료에 대해 다양한 그리드에서 구한 D_{MST}값이다. 비교1은 유사한 크기(48개 또는 49개)의 노드를 가진 자기조직도에 대한 D_{MST}인데, 8×6 그리드가 가장 작은 D_{MST}값을 가지므로 8×6 그리드가 선호된다. 비교2는 8×6 그리드와 유사한 형태(약 4 대 3의 비율)의 그리드에 대한 D_{MST}이다. 그림 3.1은 비교2의 크기에 따른 추세를 나타낸다. $48(= 8 \times 6)$에서 최소값을 가지며, $63(= 9 \times 7)$에서는 오히려 커진다. 따라서 8×6 대신 D_{MST}의 감소추세가 느려지는 지점인 $35(= 7 \times 5)$를 선택하는 것도 방안이겠지만 여기서는 최소의 D_{MST}를 갖는 8×6 SOM으로 정하기로 한다.

그림 3.2의 (a)는 붓꽃자료에 부노드(k) 자기조직도를 학습시킨 것이다. Kohonen의 자기조직도는 이러한 그림으로 나타내면 노드 $(1,1)$에서 노드 $(8,6)$까지 $48(= 8 \times 6)$개의 점으로 표현될 뿐이지만, 부노드 자기조직도의 경우에는 각 노드를 다시 7×7의 부노드로 분화하여 같은 승자노드를 갖고 있더라도 그림과 같이 개체점들이 흩뿌려지게 된다. 예컨대 노드 $(2,6)$는 하나의 값 대신 세 개의 점으로 분산되어 있음을 알 수 있다. Class 1과 Class 3은 확연히 양쪽으로 구분되며, Class 2는 Class 3과 겹치는 부분이 있는 것으로 보인다. 그림 3.2(b)는 자기조직도 위에 최소생성나무를 표현한 그림이다. 각각의 부노드 간의 관계를 알 수 있다. 예컨대 노드 $(2,5)$의 경우에 노드 $(2,4)$의 자료점들과 연결되어 있지만 노드 $(1,5)$의 점들과는 연결되어 있지 않다. 자기조직도상에서는 바로 이웃노드로서 동등하게 표현되지만, 이러한 그림을 통해서 노드간의 실제 거리는 $(2,5)$와 $(2,4)$가 더 가깝다는 것을 알 수 있다. 그림에



(a) 자기조직도 (b) 자기조직도상의 최소생성나무

그림 **3.2.** 붓꽃 자료의 8 × 6 자기조직도와 최소생성나무

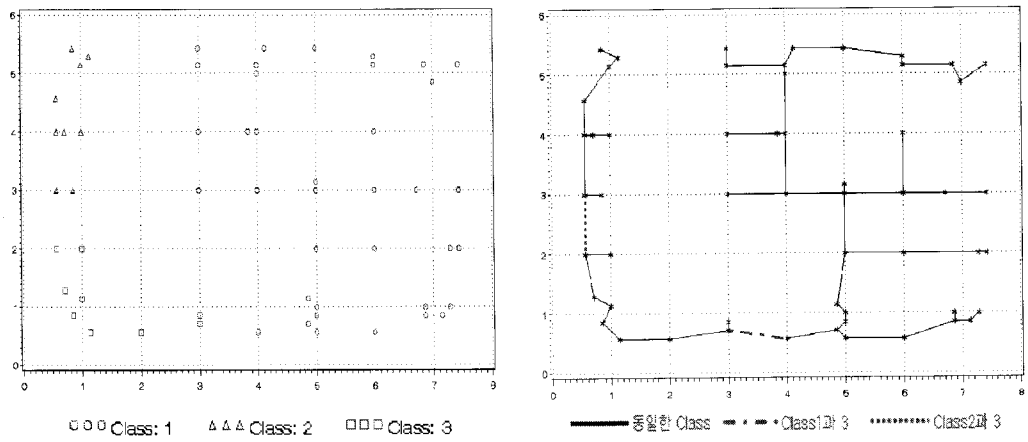
표 **3.2.** 림프구자료에서 그리드에 따른 D_{MST} 비교

비교1 (유사크기)	SOM	7 × 7	8 × 6	12 × 4	16 × 3	24 × 2
	D_{MST}	4204.90	4170.25	4275.77	4404.44	4190.53
비교2 (유사형태)	SOM	4 × 3	5 × 4	7 × 5	8 × 6	9 × 7
	D_{MST}	4187.86	4231.05	4155.19	4170.25	4263.74

서 실선은 같은 종류 개체간 연결된 선이며, 다른 종류 개체간 연결된 선은 점선으로 구분하였다. 어떤 종류끼리 연결되느냐에 따라 점선의 종류도 달리 하였는데, Class 1과 Class 3은 이어지는 부분이 없어 표현되지 않았다. 세부 군집으로 다시 분류하고 싶다면 가장 긴 가지(edge)를 자르면 되는데, 여기서는 $(2,1)$과 $(3,1)$을 잇는 선이 가장 길다. 이 지점을 자르면 Class 1과 나머지 Class가 분리됨을 알 수 있다. 또한 두 번째로 긴 가지는 $(6,1)$과 $(7,1)$ 또는 $(5,6)$과 $(6,6)$사이에서 같은 길이로 발생됨을 알 수 있다.

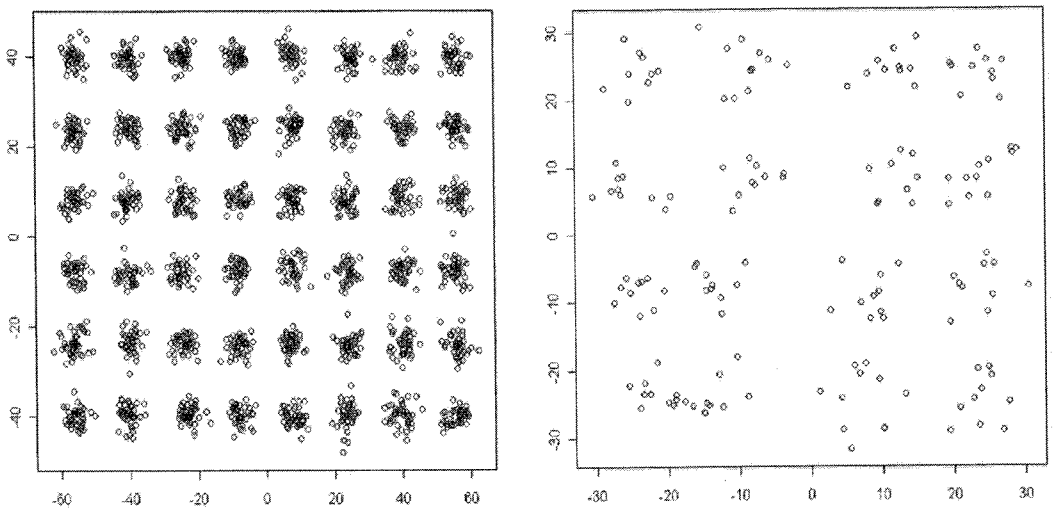
표 3.2는 림프구자료에 대한 D_{MST}이다. 여기서는 붓꽃자료와 같은 그리드에 대해 D_{MST}를 구했다. 유사한 크기의 여러 그리드를 비교한 비교1의 경우에는 8 × 6 그리드의 D_{MST}가 4170.25로 가장 작았다. 유사한 형태의 여러 그리드를 비교한 비교2의 경우에는 7 × 5 그리드가 4155.19로 최소값을 가졌다. 7 × 5 그리드는 노드 수가 35개로, 이와 같은 노드 수를 가진 6 × 6 그리드와 9 × 4 그리드에 대한 D_{MST}를 추가로 계산한 결과, 6 × 6 그리드에서는 D_{MST}가 4188.93, 9 × 4 그리드에서 4183.04으로 계산되었다. 따라서 D_{MST}를 기준으로 봤을 때 비교대상 중에서는 7 × 5 그리드가 원자료에서 가까운 거리에 있는 개체들을 자기조직도상에서도 인접하도록 하는 가장 적절한 그리드임을 알 수 있다.

그림 3.3의 (a)는 림프구 자료에 7 × 5 부노드(k) 자기조직도를 학습시킨 것이고, (b)는 이 위에 최소생성나무를 그린 것이다. Class 1(DLBCL)은 오른쪽에 분포되어 있으며 Class 2(FL)는 왼쪽 하단에 분포되어 있고 Class 3(B-CLL)은 왼쪽 상단에 분포되어 있다. 여기서도 실선은 같은 종류의 개체간 연결된 선이며 다른 종류의 개체간 연결된 선은 두 가지 점선으로 구분하였다. Class 1과 Class 2, Class 2와 Class 3이 연결된 곳은 각각 한 군데로, 군집이 비교적 잘 구분되는 것을 볼 수 있다.



(a) 자기조직도 (b) 자기조직도상의 최소생성나무

그림 3.3. 림프구 자료의 7 × 5 자기조직도와 최소생성나무



(a) 8 × 6 군집형 사례 (b) 4 × 4 군집형 사례

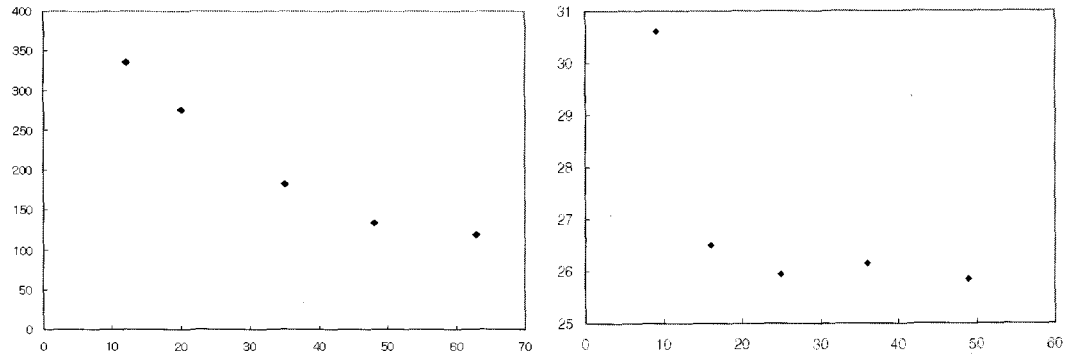
그림 3.4. 모의생성자료 사례

3.2. 모의생성자료 사례

제안된 측도 D_{MST}의 유용성을 알아보기 위해 두 개의 모의 자료를 생성하였다. 모의생성자료를 각각 8 × 6, 4 × 4의 군집 형태로 설계하였다. 즉, 두 변수 X, Y의 분포를 각각 $X \sim N(\mu_1, \sigma^2)$, $Y \sim N(\mu_2, \sigma^2)$로부터 생성하되, 8 × 6 군집형 사례의 경우 $\sigma^2 = 4$로 설정하고 μ_1과 μ_2를 각각 $(\pm 56, \pm 40, \pm 24, \pm 8)$과 $(\pm 40, \pm 24, \pm 8)$ 중에서 하나씩 선택하여, 모두 48개의 조합에 대해 각 40개씩 총 1920개의 난수를 생성하였다. 4 × 4 군집형 사례의 경우 $\sigma^2 = 9$로 설정하였고, μ_1, μ_2는 각각 $(\pm 24, \pm 8)$의 값을 갖는 16개의 조합 중 4개의 조합에서 μ_1값을 ± 8대신 ± 12를 사용하여 각 조합별로 10개씩 모두 160개의 데이터를 생성하였다. 데이터를 산점도로 나타내면 그림 3.4와 같다. 8 × 6 군집

표 3.3. 모의생성자료에서 그리드에 따른 비교

8×6 군집형	비교1	SOM	7×7	8×6	12×4	16×3	24×2
		D_{MST}	144.29	133.01	136.31	164.04	225.56
	비교2	SOM	4×3	5×4	7×5	8×6	9×7
		D_{MST}	334.41	275.37	181.70	133.01	118.39
4×4 군집형	비교1	SOM	2×8	3×5	4×4		
		D_{MST}	39.01	29.58	26.50		
	비교2	SOM	3×3	4×4	5×5	6×6	7×7
		D_{MST}	30.61	26.50	25.94	26.15	25.86



(a) 8×6 군집형 D_{MST} (b) 4×4 군집형 D_{MST}

그림 3.5. 모의생성자료에서 그리드 크기에 따른 D_{MST}

형 사례의 경우는 동일간격으로 생성되었으며, 4×4 군집형 사례는 군집간 간격이 다르게 생성되었음을 알 수 있다.

여러 그리드에 따른 D_{MST}값은 표 3.3과 같다. 8×6 군집형 사례의 경우 유사한 크기(48 또는 49개 노드)의 그리드를 비교한 결과 8×6에서 최소값을 보임을 알 수 있었다. 유사한 형태를 비교한 결과로는 8×6 이후에 감소세가 둔화되어 역시 8×6 그리드가 적절함을 나타낸다. 4×4 군집형 사례의 경우도 유사한 크기(15 또는 16개 노드)의 그리드 중에서 4×4 그리드가 최소의 D_{MST}값을 갖는 것으로 나타났다. 4×4 그리드와 형태가 유사한 정방형의 그리드에서 D_{MST}의 추세를 살펴보면, 5×5에서 국소 최소값을 보이며 4×4 이후로 감소세가 둔화되어, 4×4 그리드 또는 이웃의 5×5 그리드를 선택할 수 있다 (그림 3.5 참조). 따라서 제안된 측도 D_{MST}를 활용하여 그리드의 크기나 형태를 정하는 것이 대체로 타당하다고 볼 수 있다.

4. 결론 및 고찰

본 연구에서는 기존의 자기조직도에 최소생성나무의 원리를 추가로 활용하는 방안을 제시하였다. 즉, (i) 적절한 크기와 형태의 자기조직도를 생성하고, (ii) 이를 세분화한 부노드(k)를 구하고, (iii) 자기조직도상에 부노드간의 거리를 이용한 최소생성나무를 그리는 방법을 제안하였다. 자기조직도상에서 이웃하는 노드라 해도 공간상에서의 실제거리는 서로 상당히 다를 수 있는데, 이러한 방법을 통해 각 군집간의 관계를 추가로 시각화할 수 있었으며, 추후 긴 가지를 순서대로 잘라내는 과정을 통하여 군집들을 재군집화할 수도 있었다. 또한 자기조직도의 작성시에 적절한 지도의 크기 및 형태를 평가할 수 있는

D_{MST}측도를 제안하였고, 실제 자료와 모의 자료를 통하여 제안된 방법의 유용성을 확인하였다.

이 방법은 자기조직도와 최소생성나무의 장점을 가질 수 있는데, 자기조직도라는 분할군집방법을 우선 적용하므로 최단연결법이나 최장연결법을 사용한 계층적 군집분석과 달리 모든 자료에 대해 할당된 군집을 구할 수 있으며, 또한 군집이 확연히 분류가 되지 않는 데이터에 대해서도 할당된 군집외에 시각화된 거리를 통하여 정보를 얻을 수 있다.

제안된 방법은 약간의 수정을 가하여 다른 목적으로 활용될 수 있다. 예컨대 여기서는 자기조직도상에 최소생성나무를 작성할 때 각 노드간의 거리를 활용함으로써 노드간 관계가 보이도록 하였는데, 노드간 거리대신 각 개체간의 거리를 이용하여 최소생성나무를 형성하고 자기조직도 상에 그리면, 자기조직도의 결과와 최소생성나무의 결과를 동시에 표현할 수 있게 된다. 이러한 이 경우에는 자기조직도상에서 가까운 개체끼리 연결되지 않고 먼 개체와 연결되는 수도 있는데, 이를 통해 자기조직도가 차원축소를 통해 얼마나 변형되었는지 파악할 수 있을 것이다.

참고문헌

김성수 (1999). 통계그래픽스를 이용한 K-평균 및 계층적 군집분석, <한국분류학회지>, **3**, 13–27.

엄익현, 허명회 (2005). SOM에서 개체의 시각화, <응용통계연구>, **18**, 83–98.

허명회 (2003). 주성분 자기조직화 지도 PC-SOM, <응용통계연구>, **16**, 321–333.

Alizadeh, A. A., Eisen, M. B., Davis, R. E., Ma, C., Lossos, I. S., Rosenwald, A., Boldrick, J. C., Sabet, H., Tran, T., Yu, X., Powell, J. I., Yang, L., Marti, G. E., Moore, T., Hudson, J., Jr, Lu, L., Lewis, D. B., Tibshirani, R., Sherlock, G., Chan, W. C., Greiner, T. C., Weisenburger, D. D., Armitage, J. O., Warnke, R., Levy, R., Wilson, W., Grever, M. R., Byrd, J. C., Botstein, D., Brown, P. O. and Staudt, L. M. (2000). Different type of diffuse large B-cell lymphoma identified by gene expression profiling, *Nature*, **403**, 503–511.

Berglund, E. and Sitte, J. (2006). The parameterless self-organizing map algorithm, *IEEE Transactions on Neural Networks*, **17**, 305–316.

Gower, J. C. and Ross, G. J. S. (1969). Minimum spanning trees and single linkage cluster analysis, *Applied Statistics*, **18**, 54–64.

Haese, K., Goodhill, G. J. (2001). Auto-SOM: Recursive parameter estimation for guidance of self-organizing feature maps, *Neural Computing*, **13**, 595–619.

Hsu, A. L. and Halgamuge, S. K. (2003). Enhancement of topology preservation and hierarchical dynamic self-organising maps for data visualisation, *International Journal of Approximate Reasoning*, **32**, 259–279.

Hsu, C. C. (2006). Generalizing self-organizing map for categorical data, *IEEE Transactions on Neural Networks*, **17**, 294–304.

Kim, S. S., Kwon, S. and Cook, D. (2000). Interactive visualization of hierarchical clusters using MDS and MST, *Metrika*, **51**, 39–51.

Kohonen, T. (1995). *Self-Organizing Maps*, Springer-Verlag, Berlin.

Park, M., Jang, Y. J. and Huh, M. H. (2005). Analysis of gene expression data using PC-SOM, In *Proceedings of the 55th session of International Statistical Institute.*

Prim, R. C. (1957). Shortest connection networks and some generalizations, *Bell System Technical Journal*, **36**, 1389–1401.

Samsonova, E. V., Kok, J. N. and Ijzerman, A. P. (2006). TreeSOM: Cluster analysis in the self-organizing map, *Neural networks*, **19**, 935–949.

Tamayo, P., Slonim, D., Mesirov, J., Zhu, Q., Kitareewan, S., Dmitrovsky, E., Lander, E. S. and Golub, T. R. (1999). Interpreting patterns of gene expression with self-organizing maps: Methods and application to hematopoietic differentiation, In *Proceedings of the National Academy of Sciences*, **96**, 2907–2912.

Xu, Y., Olman, V. and Xu, D. (2001). Minimal spanning trees for gene expression data clustering, *Genome Informatics*, **12**, 24–33.

Yan, A. (2006). Application of self-organizing maps in compounds pattern recognition and combinatorial library design, *Combinatorial Chemistry & High Throughput Screening*, **9**, 473–480.

Use of Minimal Spanning Trees on Self-Organizing Maps

Yoo-Jin Jang[1] · Myung-Hoe Huh[2] · Mira Park[3]

[1]Department of Statistics, Korea University; [2]Department of Statistics, Korea University;
[3]Department of Preventive Medicine, Eulji University

(Received January 2009; accepted January 2009)

Abstract

As one of the unsupervised learning neural network methods, self-organizing maps(SOM) are applied to various fields. It reduces the dimension of multidimensional data by representing observations on the low dimensional manifold. On the other hand, the minimal spanning tree(MST) of a graph that achieves the most economic subset of edges connecting all components by a single open loop. In this study, we apply the MST technique to SOM with subnodes. We propose SOM's with embedded MST and a distance measure for optimum choice of the size and shape of the map. We demonstrate the method with Fisher's Iris data and a real gene expression data. Simulated data sets are also analyzed to check the validity of the proposed method.

Keywords: Self-organizing map(SOM), minimal spanning tree(MST), data visualization, distance measure.

This work was supported by the Korea Research Foundation Grant founded by Korean Goverment(MOEHRD, R14-2003-002-01001-0).

[3]Corresponding author: Associate Professor, Department of Preventive Medicine, Eulji University, Daejeon 301-832, Korea. E-mail: mira@eulji.ac.kr