응용통계연구 제 16권 2호, 2003년, pp. 321-333

수성분 자기조직화 지도 PC-SOM*

허명회[1)]

요 약

자기조직화 지도(SOM)은 T. 코호넨의 주도하에 개발된 비지도 학습 신경망 모형이다. 그 동안 패턴인식과 문서검색 분야에 주로 응용되어 왔기 때문에 통계학 분야에서는 덜 알려졌으나, 최근 K-평균 군집화에 대한 대안적 데이터 마이닝 기법으로 활용되기 시작하였다. 본 연구에서는 SOM의 한 버전인 PC-SOM(주성분 자기조직화 지도)을 제안하고 활용 예를 제시하고자 한다. PC-SOM은 1차원적 SOM 알고리즘을 반복 수행하여 2차원, 3차원 등의 SOM을 얻는 방법이기 때문에 기존 SOM과는 달리 사전 Map의 크기를 확정할 필요가 없다. 또한, 기존 SOM에 비하여 향상된 시각화를 가능하게 한다.

주요용어: 자기조직화 지도(SOM), 코호넨(T. Kohonen), 신경망, 비지도 학습, 시각화.

1. 연구 배경 및 목적

코호넨의 자기조직화 지도(self-organizing map, SOM)는 1980년대 초반 핀란드의 전기공학자 T. 코호넨(Teuvo Kohonen)에 의해 개발된 비지도 학습(unsupervised learning) 신경망(neural network) 모형의 한 종류이다 (Kohonen, 1995). 스스로의 소개 논문에서 코호넨은 SOM의 특성을 시각화(visualization)와 축약화(abstraction)의 두 가지로 뽑았다(Kohonen, 1998). SOM이 고차원 다변량 자료의 저차원 시각화 기법으로 다차원 공간에서 비선형적 관계의 주요한 위상적·계량적 특성을 저차원 공간에서 축약적으로 보여준다는 것이다.

SOM은 그 동안 전자공학의 패턴인식 분야나 문서검색 정보분야에서 응용되어 왔기 때문에 통계학 분야에서는 덜 알려졌으나 최근 K-평균 군집화에 대한 대안적 데이터 마이닝 기법으로 활용되기 시작하였다. 그런 과정에서, 통계학자들은 코호넨 SOM이 어떤 무엇의 예측을 목표로 하지 않는 자료 탐색적 방법이라는 점에서 K-평균 군집화와 마찬가지이되 유사한 개체들을 서로 이웃하는 위치에 오도록 스스로 배치한다는 점에서 K-평균 군집화와 차이가 있다는 것을 이해하게 되었다 (Ripley, 1996; Hastie, Tibshirani and Friedman, 2001).

국내 통계학계에서도 SOM의 안정적 해 산출을 위한 알고리즘을 제안한 전성해·전홍석·황진수(2002)의 연구가 있었다. 이어서, 본 연구자는 SOM이 표준적인 통계적 방법으로 활용되기 위해서는 몇 가지 사항에서 개선이 필요할 것으로 생각한다. 본 논문에서는 이런 관점에서 주성분 자기조직화 지도(PC-SOM, principal components self-organizing map)이라는 SOM의 한 버전을 제안할 것이다.

* 본 연구는 2002년도 고려대학교 특별연구비 지원을 받았음.

1) (136-701) 서울시 성북구 안암동 5가 1, 고려대학교 통계학과, 교수
E-mail : stat420@korea.ac.kr

이에 앞서, 일반적인 SOM 알고리즘을 간단히 소개함으로써 본 논문에서 사용할 표기와 용어를 일러두고자 한다.

분석 데이터가 n개의 p차원 입력개체(input units) $x_1, \cdots, x_n$으로 구성되어 있다고 하자 ($p \geq 3$). 저차원 그리드 상에 m개의 p차원 중량(weights) $w_1, \cdots, w_m$을 산출하기 위한 코호넨 SOM (K-SOM) 알고리즘은 다음과 같다. [그리드(grid)는 일정 간격을 갖는 망(網, net)을 의미하며 그리드 상에서 각 마디(knot)는 이웃하는 마디들에 연결된다. 그리드는 신경망(neural net), 마디는 뉴런(neuron)에 대한 표상이다. 이하 '마디' 대신 노드(node)를 기본 용어로 사용하기로 한다. 각 노드는 중량으로 대표되며 SOM 산출과정에서 계속 업데이트(학습)된다. 그리드는 흔히 2차원으로 간주되지만 그렇게 제한되어야 하는 것은 아니다.]

코호넨 SOM (K-SOM) 알고리즘

0) 시점 t의 첫 값을 1로 둔다. 그리고 m개의 p차원 중량(weights) $w_1, \cdots, w_m$에 초기값을 부여한다. 초기 값으로 임의 벡터를 주어도 되지만 Kohonen (1998)의 권고에 따라 주성분분석을 활용하기로 한다.

1) 입력개체 $i(= 1, \cdots, n)$가 제시되면 m개의 중량 $w_1, \cdots, w_m$ 중 가장 가까운 것을 찾아내고 그 중량이 속한 노드를 $k(i)$라고 하자. 여기서 유클리드 거리를 사용한다. 즉,

$$||x_i - w_{k(i)}|| \leq ||x_i - w_j||, \quad \text{모든 } j = 1, \cdots, m \text{에 대하여.}$$

노드 k를 승자(winner)라고 한다.

2) 승자 노드 $k(i)$와 그 주변 노드의 중량들이 다음과 같이 업데이트된다.

$$w_j \leftarrow w_j + \alpha_t \; h_t(k(i), j) \; (x_i - w_j), \quad ||r_j - r_{k(i)}|| \leq d_t \text{인 모든 } j \text{에 대하여.}$$

여기서 r_j는 노드 j의 그리드 위치점이고 학습률(learning rate) α_t는 초기값 α_0에서 최종값 α_1까지 t에 따라 감소하도록 세팅된다. $h_t(k, j)$는 국소가중치(local weight)로서 t와 $||r_j - r_k||$에 따라 감소하도록 세팅되는데 흔히 다음 패턴의 함수가 적용된다.

$$h_t(k, j) = \exp\{-||r_j - r_k||^2/(2\sigma_t^2)\}, \quad \sigma_t^2 \text{은 } t \text{에 따라 감소.}$$

마지막으로 d_t도 t에 따라 감소하는 계단함수이다.

3) 한 개체가 단계 1과 2에 따라 처리되면 시점 t를 1만큼 증가시킨다. 마지막 개체가 처리되면 첫 개체로 되돌아간다. 단, 모든 중량 값의 변화가 거의 소멸하거나 시점 t가 미리 지정된 최대 한계에 도달하게 되면 진행을 멈추고 각 개체의 승자 노드 $k(1), \cdots, k(n)$ 및 중량 $w_1, \cdots, w_m$을 출력한다.

위 알고리즘의 적용에 앞서 입력개체 및 중량이 위치하는 p차원 공간에서 유클리드 거리가 간주되었으므로, 특별한 이유가 없는 한 분석 데이터를 사전 표준화하는 것이 좋다. 표준화의 방법은 1) 평균 0, 분산 1의 정규화, 2) 최소값 0, 최대값 1의 범위 표준화 등이다.

K-SOM을 통계적 자료분석에 적용할 때 분석자는 Map의 차원을 포함, 그리드의 크기를 정해야 한다. 분석자가 2차원 사각형 그리드를 상정하였다고 해도 그리드의 행 수 c_1과 열 수 c_2를 미리 정해야 하는 부담을 갖는다. K-SOM에서 입력은 연속형이나 이에 대한 출력은 이산형이다. 때문에 통상적인 K-SOM은 시각화(visualization)를 중시하는 데이터 마이너를 만족시키기 어려웠다.

본 연구에서는 앞의 두 문제를 해결하고자 SOM의 한 버전으로 PC-SOM(주성분 자기조직화 지도)을 제안하고 활용 예를 제시하고자 한다. PC-SOM은 1차원적 SOM 알고리즘을 반복 수행하여 2차원, 3차원 등의 SOM을 얻는 방법으로, 통상적인 K-SOM과는 달리 Map의 크기를 사전에 모두 정해야 할 필요가 없다는 데 그 특징이 있다. 또한, 연속적 출력을 생성하므로 기존 SOM에 비하여 향상된 시각적 그래프를 제공할 것이다. 연속적 K-SOM에 대한 기존 연구로는 Goppert and Rosenstiel (1997), Campos and Carpenter (2000) 등이 있으나 주변 중량들에 사영하는 방법에 의존함으로써 위상이 복잡한 지역에서는 일부 문제가 있었다. 그러나 본 연구의 PC-SOM은 이 문제를 1차원으로 축소시켜 쉽게 근사적 최적 해를 산출해낼 것이다.

2. PC형 SOM의 제안

자기조직화 지도(self-organizing map, SOM)는 p차원 입력개체 공간을 2차원, 3차원 등으로 축약하는 것이 보통이지만 최소 1차원 축약도 가능하고 이에 대한 기존 연구들이 여럿 있다 (예컨대 Koikkalainen (1999) 등). c_1개의 노드를 단선으로 연결한 SOM을 만들었다고 하자. 좌우 이웃하는 노드들은 유사한 중량을 갖게 될 것이다.

본 연구에서 제안하는 PC-SOM의 기본 아이디어는 1차원 SOM 산출이후 입력개체들을 각각 대표점과 잔차로 분리하고 여기서 발생한 잔차들로 별개의 1차원 SOM을 산출하여 기존 SOM에 교차적으로 붙이면 결과적으로 2차원 SOM이 된다는 것이다. 이런 식으로 한·두 차례 더 반복하면 3차원 이상의 SOM 산출도 가능하다. 이와 같은 순차적 방식은 각 단계에서 주성분 분석(principal components analysis, PCA)의 선형적 사영을 SOM의 비선형적 축약으로 대체하는 것이기에 새 버전의 SOM을 PC-SOM(principal components self-organizing map)으로 명명한다.

PC-SOM 알고리즘: 기본형

0) 제1축 노드 수($= c_1$)가 입력되면 선형 주성분분석(PCA)을 수행하여 제1 고유값 λ_1을 얻었다고 하자. 그러면 제1 주성분은 평균이 0, 표준편차 $\lambda_1^{0.5}$인 분포를 갖는다. 따라서 입력개체들의 제1축 사영점들의 실질적 범위는 $(-2.5\cdot\lambda_1^{0.5},\ 2.5\cdot\lambda_1^{0.5})$가 되고 제1축 상에 c_1개의 노드들을 등간격으로 배치한다면 노드간 초기 간격은 다음과 같이 된다.

$$\text{interval} \ = \ 5\cdot\lambda_1^{0.5}/(c_1-1).$$

1) c_1개의 노드를 갖는 1차원 SOM을 만들어 입력개체 i에 대한 승자 노드 $k_1(i)$와 해당하는 중량 $w_{k_1(i)}$의 리스트를 출력한다 $(i = 1,\cdots,n)$. 이로부터 입력개체 x_i는 SOM

에서 그것을 대표하는 $w_{k_1(i)}$와 잔차 $x_i - w_{k_1(i)}$로 분리된다. 입력개체 x_i를 잔차로 대체한다.

2) 새로 PCA를 수행하여 제1 고유값 λ_2를 얻는다. 여기서의 제1 주성분은 실질적으로 구간 $(-2.5 \cdot \lambda_2^{0.5},\ 2.5 \cdot \lambda_2^{0.5})$에 놓이게 된다. 따라서 제2축의 그리드 간격을 제1축 그리드 간격과 가급적 같게 만들기 위하여 요구되는 제2축 노드 수는 다음과 같다.

$$c_2 \ = \ \text{round}(5 \cdot \lambda_2^{0.5}/\text{interval}) + 1.$$

여기서 round$(\cdot)$는 반올림 함수이다.

3) c_2개의 노드를 갖는 1차원 SOM을 만들어 각 입력개체에 대한 승자 노드와 중량들의 리스트를 출력한다. 이에 따라 입력개체 x_i는 2차원 SOM에서 노드 $(k_1(i), k_2(i))$에 배속된다.

필요하다면 이 과정을 반복하여 제3축, 제4축, $\cdots$ 등에 필요한 노드 수를 구할 수 있고, 이로부터 필요한 차원 수 및 노드 수를 알아낼 수 있다.

선형적인 구조를 갖는 다변량 데이터에 적용되는 경우 기본형 알고리즘은 PCA와 유사하지만 동일하지는 않다. 한 가지 이유는 PCA가 각 입력개체를 개별적 사영점으로 대체시키는 반면 PC-SOM 기본형에서는 각 축의 결정시 입력개체를 c개의 덩어리 점 중 하나로 대체시키기 때문이다. 이와 같은 PC-SOM 기본형의 문제를 해결하기 위하여 본 연구자는 다음 PC-SOM 보간형을 제안한다.

PC-SOM 알고리즘: 보간형

알고리즘의 대부분은 기본형과 동일하다. 차이는 단 하나인데 기본형에서는 입력개체를 해당 노드의 중량으로 대표시키는 반면 새 버전에서는 인접 중량들의 가중값으로 대체한다는 점이다. 입력개체 x에 대한 승자 노드 중량을 w_k, 이것의 왼쪽 인접 노드 중량을 w_j, 오른쪽 인접 노드 중량을 w_l이라고 하자. [승자노드가 끝 노드인 경우, 일반성을 잃지 않고 승자노드가 노드 1(가장 왼쪽 노드)인 경우에는, 노드 1(승자 노드)의 중량을 w_1, 노드 2의 중량을 w_2라고 하자. 이 경우 w_1과 w_2를 연결하는 직선상에서 노드 0(= 노드 1의 왼쪽 노드)의 중량 w_0을 잡는 것이 자연스럽다. 따라서, $w_0 = 2w_1 - w_2$.] 따라서 w_j과 w_k를 연결하는 선분과 w_k와 w_l를 연결하는 선분상에서 입력개체 x에 가장 가까운 점을 찾자. 그럼으로써 이산적 출력의 문제를 해결할 수 있겠다. 실용적으로, 각 선분(노드 길이 1)을 7등분하여 승자 노드를 중심으로 왼쪽 그리드 연결선 상에 -6/7, -5/7, -4/7, -3/7, -2/7, -1/7과 0(중심 노드), 오른 쪽 연결선 상에 1/7, 2/7, 3/7, 4/7, 5/7, 6/7을 타점하여 -3/7, -2/7, -1/7, 0, 1/7, 2/7, 3/7에 해당하는 부노드들(subnodes) 중에서 x에 가장 가까운 승자를 찾는 방법을 제안한다. 구체적으로 승자 노드가 k일 때, 최종 승자 부노드(final winner subnode)를 찾기 위해 비교해보는 중량들은 다음과 같다.

$$(3w_j + 4w_k)/7,\quad (2w_j + 5w_k)/7,\quad (w_j + 6w_k)/7,$$
$$w_k,\quad (6w_k + w_l)/7,\quad (5w_k + 2w_l)/7,\quad (4w_k + 3w_l)/7.$$

여기서는 7등분 보간법을 고려하였지만 더 조밀한 보간법도 필요하다면 쓸 수 있을 것이다.

PC-SOM에서 기본형 대신 보간형을 채택함으로써 그리드 크기를 확대시키지 않고도 입력개체와 그것의 Map 상에서의 출력간 거리를 줄일 수 있다. 따라서 다음과 같이 비적합도를 정의할 수 있다.

비적합도(Measure of Unfitness)

PC-SOM의 중간/최종 단계, 또는 K-SOM에서 입력개체 x_i에 대하여 잔차 e_i가 산출되었다고 하자 $(i = 1, \cdots, n)$. 비적합도(Measure of Unfitness)를 잔차들의 평균 제곱합 $\sum_{i=1}^{n} e_i' e_i/(n-1)$로 정의한다. 비적합도는 PC-SOM의 산출과정에서 단계가 진행됨에 따라 감소하게 된다. 적합도(Measure of Fitness)는 $\sum_{i=1}^{n} x_i' x_i/(n-1) - \sum_{i=1}^{n} e_i' e_i/(n-1)$로 정의될 수 있겠다.

PC-SOM의 두드러진 장점은 SOM의 각 축이 어떤 변수적 특성을 갖는지를 볼 수 있다는 점이다. PC-SOM의 기본형을 보자. 제1축이 형성되면서 c_1개의 중량 $w_1, \cdots, w_{c_1}$이 만들어진다. $w_k (k = 1, \cdots, c_1)$가 p개의 변수 성분을 가지므로 변수 $j(= 1, \cdots, p)$가 제1축에 어떻게 부하되어 있는가를 보려면 c_1개의 중량 $w_1, \cdots, w_{c_1}$의 j번째 성분을 순서대로 훑으면 된다. 다른 축에 대하여도 마찬가지이다. PC-SOM의 보간형에서도 노드들의 중량들만 순서대로 관찰하는 것으로 충분하다. 왜냐하면 이웃하는 노드간 보간점들의 성분은 양 끝 노드들 성분 사이에 있기 때문이다. 다음 절에서 예를 보이기로 하겠다.

3. 붓꽃 자료 예

앞 절에서 제안한 PC-SOM 알고리즘을 잘 알려진 Fisher의 붓꽃 자료(Iris Data)에 적용하여 봄으로써 이 방법이 예상대로 작동한다는 사실을 확인하고자 한다. 여기서 산출할 SOM은 PC-SOM 보간형, PC-SOM 기본형, 코호넨의 SOM (K-SOM)의 세 가지이다.

붓꽃 자료는 150(= n)개 개체에 대한 4(= p)개 변수(x1: sepal length, x2: sepal width, x3: petal length, x4: petal width)와 개체별 품종 분류(1: setosa, 2: versicolor, 3: virginica)로 구성되어 있다. 그림 3.1을 보라. 여기서는 SOM의 수행성을 보기 위한 것이 주목적이므로 품종 변수를 SOM에 사용하지 않기로 하겠다. 즉 입력변수는 x1~x4이고 SOM 산출에 앞서 표준화(평균 0, 표준편차 1) 변환되었다.

그림 3.2a는 제1축의 노드를 12개로 지정하여 이 연구에서 제안하는 PC-SOM 보간형을 적용한 결과이다. PC-SOM의 산출시, 사전에 제1축 노드 수 12개, 초기 학습률 0.25, 최

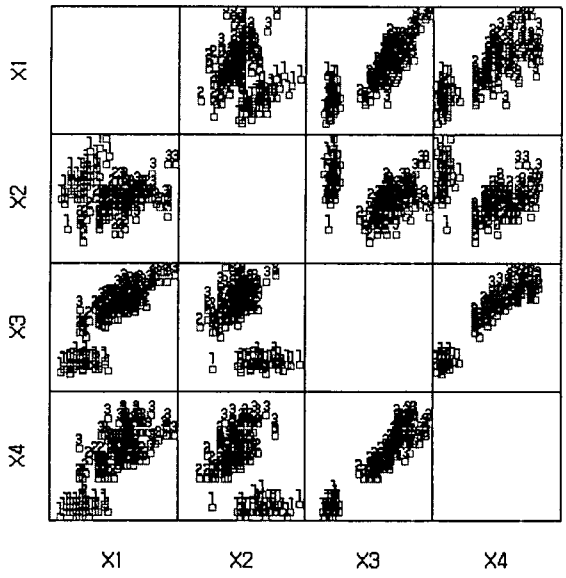

그림 3.1: 피셔의 붓꽃 자료에 대한 산점도 행렬: 숫자는 품종번호를 표시함

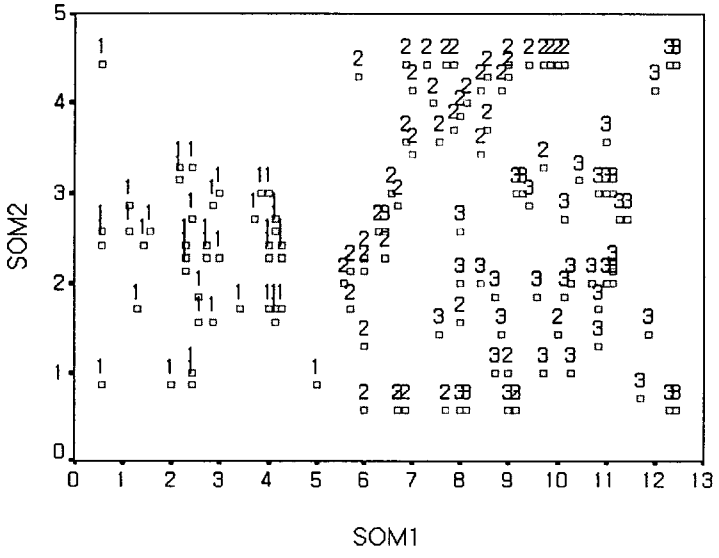

그림 3.2a: 12x4 PC-SOM: 보간형: 숫자는 품종번호를 표시함

종 학습률 0.001, 초기 주변거리=3, 최종 주변거리=1, 반복 = 50회 등으로 지정한 결과이다. 이에 따라, 최종 출력을 얻는데 150x3x50=22,500회의 업데이팅이 이루어진 셈이다. 연구자가 작성한 SAS/IML 프로그램을 사용하였다. 제2축의 노드가 4개로 나왔다. 품종 1이 품종 2, 3과 격리되어 있고 품종 2와 3은 다소 겹쳐 있다는 것이 Map의 주요 특징이다. 12x4 PC-SOM 보간형으로 적합되지 않는 잔차 평균제곱, 즉 비적합도는 0.22로 계산되었다. (사전 표준화되었으므로 적합도는 3.78이 된다. 적합도와 비적합도의 합이 4.00이므로.)

그림 3.2b는 제1축 및 제2축의 그리드 점에서 각 변수 값이 얼마인지를 보여준다. 대체로 제1축은 변수 x1, x3 및 x4에 선형적으로, 변수 x2에는 곡선형으로 나타난 것을 볼 수 있다. 반면, 제2축은 거의 변수 x2에 의하여 선형적으로 결정되는 것으로 나타났다.

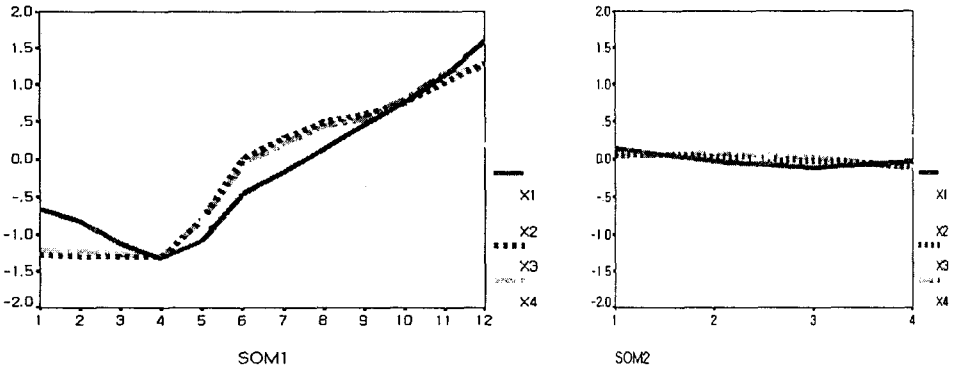

그림 3.2b: 12x4 보간형 PC-SOM에서 제1축과 제2축 변수 부하

이와 대비하기 위하여 PC-SOM의 기본형을 적용하여 보았다. 제1축의 노드를 12개로 지정하였더니 앞에서와 같이 제2축 노드가 4개로 결정되었다. 그림 3.3a의 왼쪽 그림을 보라. 그리드 점에서 다수의 개체가 겹쳐 출력되므로 150 개체 모두를 볼 수 없다. (-0.5, 0.5)의 균일난수를 덧붙이는 진동(agitation, jittering)으로 이 문제를 해결하여 본 것이 그림 3.3a의 오른쪽 그림이다. 그림 3.3b는 2개의 진동 그래프를 추가적으로 보여주는데 진동의 임의성 때문에 약간씩 다른 느낌을 준다. 그림 3.3c는 PC-SOM 기본형에서의 축별 변수 부하를 보여준다. 그림 3.2b와 거의 같다. PC-SOM 기본형과 보간형간 시각적으로는 차이가 없어 보이지만 자세히 보면 기본형에서 품종 1의 상하 산포가 보간형에 비해 크다. 기본형12x4 PC-SOM 기본형으로 표현되지 않은 비적합도는 0.27로 계산되었다. 따라서 PC-SOM 보간형이 기본형에 비해 23% (=0.27/0.22-1) 더 효율적이라고 말할 수 있다.

그림 3.4는 12x4 K-SOM의 원형과 진동형을 본 것이다. 비적합도는 0.13으로 계산되었다. 따라서 보간형 PC-SOM에 비해 적합도가 상당히 좋다고 하겠다. 그러나 과다 적합일 가능성이 있으며 PC-SOM과는 달리 축의 변수적 특성을 파악하기가 어렵다.

붓꽃 자료 사례에서 보듯이 K-SOM이 PC-SOM과 동일 크기인 경우 일반적으로 적합도가 좋을 것이라는 것을 알 수 있다. 상대적으로 PC-SOM은 적합도가 다소 떨어지겠지만 그만큼 안정도가 좋을 것이고 더욱이 SOM의 그리드 방향을 변수적 의미로 살필 수 있다는 점, 즉 추가적 자료 해석이 가능하다는 장점을 갖는다.

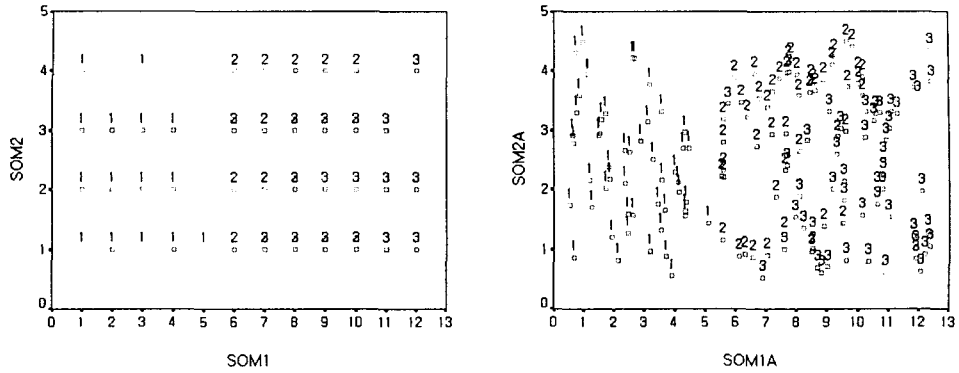

그림 3.3a: 12x4 PC-SOM의 기본형: 원형과 진동형

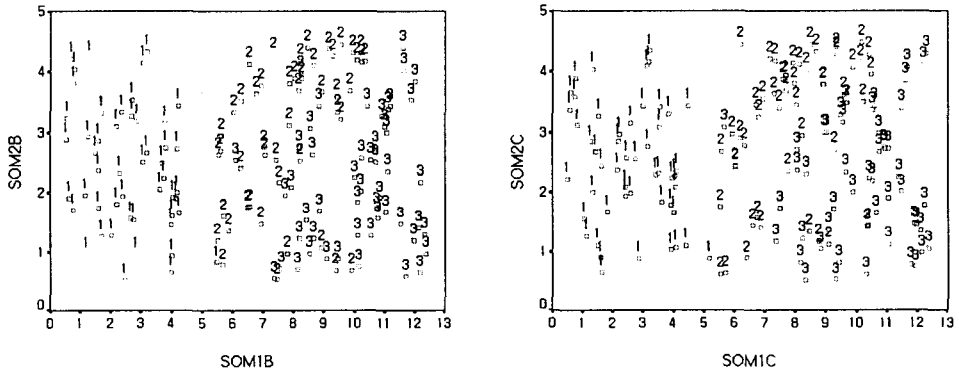

그림 3.3b: 12x4 PC-SOM의 기본형: 다른 진동 그래프

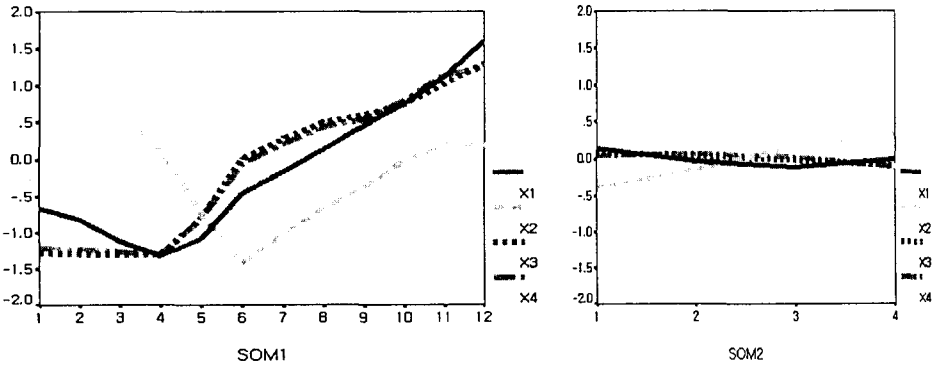

그림 3.3c: 12x4 기본형 PC-SOM에서 제1축과 제2축 변수 부하

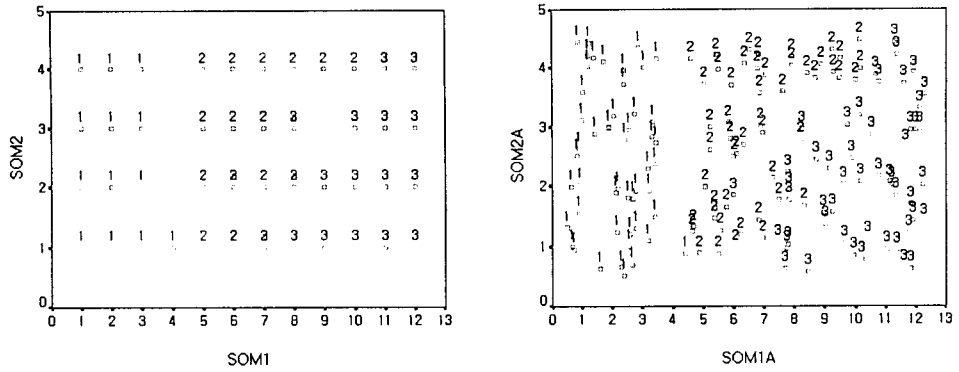

그림 3.4: 12x4 K-SOM: 원형과 진동형

앞의 K-SOM 분석에서는 Map의 크기를 PC-SOM가 제시한 12x4로 정하였다. Map 크기를 7x7로 하면 어떤 결과가 나오는지를 살펴보기로 하자 (7x7 그리드의 총 노드 수는 12x4 그리드의 총 노드 수와 거의 같다). 그림 3.5가 결과인데 품종 2와 3 그룹이 각각 두 군데로 나뉘어져 있다 (비적합도는 0.13으로 12x4 K-SOM과 동일하게 나왔다). 그림 3.1의 산점도 행렬에서 보듯이 각 품종 그룹은 각기 한 덩어리를 이루고 있으므로 그림 3.5와 같은 동일 품종군의 분리된 표출은 바람직하지 못하다.

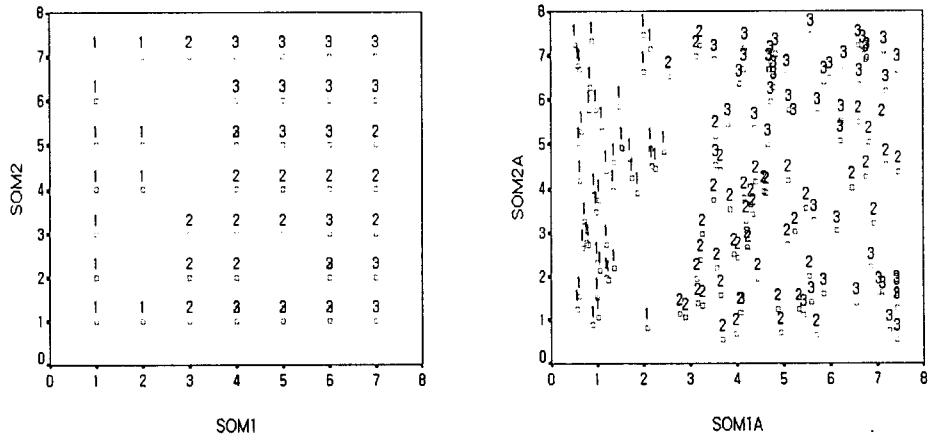

그림 3.5: 7x7 K-SOM: 원형과 진동형

이 실험은 K-SOM에서 Map의 양 축 크기가 적정하지 못한 경우 불안정한 결과가 초래될 수 있다는 교훈을 상기시킨다. PC-SOM은 한 축의 노드 수만 적절하게 정하면 되므로 이 점에서 상대적으로 자유롭다. 다음 절에서는 모의자료(simulated data) 예를 다루어 봄으로써 PC-SOM의 성질을 좀 더 탐구해 볼 것이다.

4. 모의자료 예

이 절에서는 구조가 잘 알려진 모의자료에 대하여 PC-SOM을 만들어 보기로 한다. 그 결과가 기대하는 바와 일치하는가를 확인하는 것이 목적이다.

$Z_1, \cdots, Z_5$를 표준정규분포 N(0,1)로부터의 확률변수라고 하자. 그리고 $X_1, \cdots, X_5$를 다음과 같이 정의한다.

$$\begin{aligned} X_1 &= (Z_2+Z_3+Z_4+Z_5)/2, \quad X_2 = (Z_1+Z_3+Z_4+Z_5)/2, \\ X_3 &= (Z_1+Z_2+Z_4+Z_5)/2, \quad X_4 = (Z_1+Z_2+Z_3+Z_5)/2, \\ X_5 &= (Z_1+Z_2+Z_3+Z_4)/2. \end{aligned}$$

그러면 $(X_1, \cdots, X_5)$는 평균이 $(0, \cdots, 0)$이고 공분산행렬이 $0.25I_5 + 0.75J_5$인 다변량 정규분포를 따르게 된다. 이에 따라 제1주성분 PC1은 $X_1 + \cdots + X_5$에 비례하게 되고, 이에 직교하는 선형결합이 제2주성분 PC2가 된다.

그림 4.1은 이런 등상관 구조를 따르는 $(X_1, \cdots, X_5)$ 입력개체 400개로부터 생성된 10x3 보간형 PC-SOM에서 각 축의 변수 부하를 보여준다. 제1축에서 다섯 변수의 부하가 동일하게 나옴으로써 이론적 기대와 일치함을 확인할 수 있다.

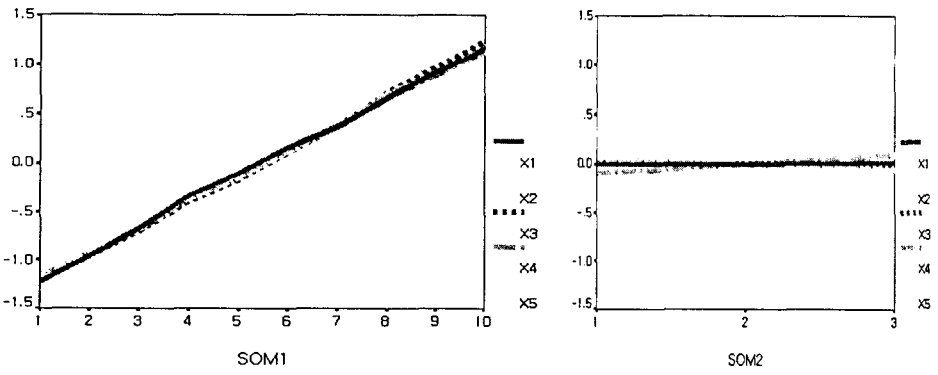

그림 4.1: 등상관구조의 5변량 자료에 대한 10x3 PC-SOM에서 각 축의 변수 구성

이제 $(X_1, \cdots, X_5)$의 각 성분에 지수변환을 적용한 뒤에 PC-SOM을 적용해 보도록 하겠다. 즉 여기서 분석할 자료는 400개의 $(\exp(X_1), \cdots, \exp(X_5))$이다. 그림 4.2가 변환자료로부터 생성된 10x5 보간형 PC-SOM의 각 축 변수 부하를 보여주는 그래프이다. 기대할 수 있는 바와 같이, 제1축 변수 부하가 5개 변수 모두에서 지수적으로 균일하게 증가하는 패턴을 보여준다. 이에 따라 모의생성 원자료나 변환자료에 대한 Map의 제1축 좌표점들은 대체적으로 동일하게 나타난다. 그림 4.3을 보라. 즉, PC-SOM은 입력변수들의 단조변환에 거의 불변적(invariant)으로 작동한다. PC-SOM의 산출시, 사전에 제1축 노드 수 10개, 초기 학습률 0.25, 최종 학습률 0.001, 초기 주변거리=5, 최종 주변거리=2, 각 단계당 주기 =

50 등으로 지정한 결과이다. 이에 따라 출력을 얻는데 총 400x4x50=80,000회의 업데이팅이 이루어진 셈이다. 모든 계산은 연구자가 작성한 SAS/IML 프로그램으로 수행되었다.

이 모의자료 예에서 PC-SOM이 필요한 변수 변환을 스스로 한다는 것을 알 수 있다. 이 점은 선형 주성분분석에서 기대하기 어려운 비선형적 기능을 PC-SOM이 갖는다는 사실을 보여준다.

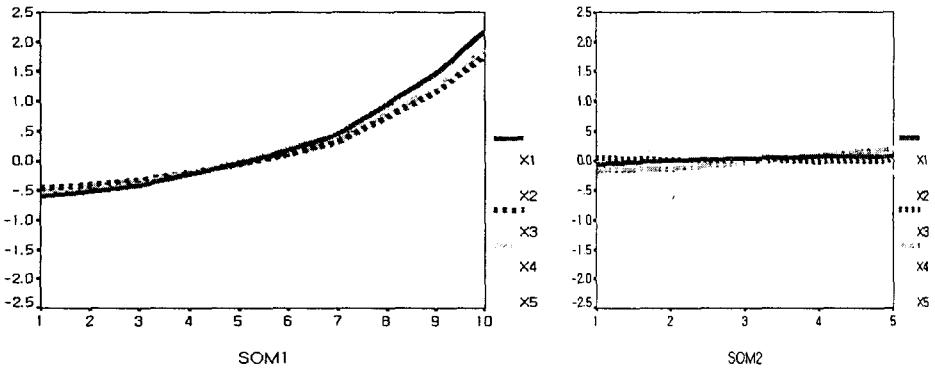

그림 4.2: 지수변환 자료에 대한 10x5 PC-SOM에서 각 축의 변수 구성

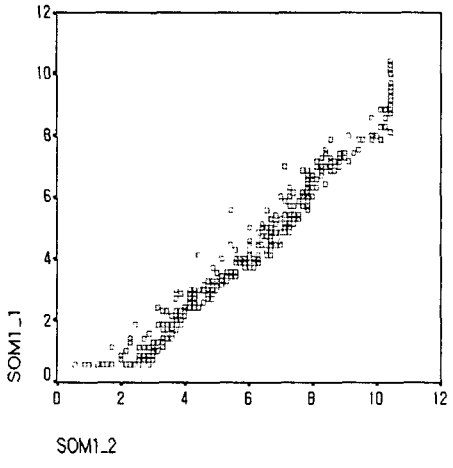

그림 4.3: 변환이전 자료(수평)와 지수변환 자료(수직)의 제1축 좌표점들간 관계

5. 맺음 말

본 연구에서 제안하는 PC-SOM은 여러 장점을 갖는다. 첫째, Map의 축이 갖는 변수적 특성을 제시하는데 이는 K-SOM이 지니지 못하였던 점이다. 둘째, K-SOM에 비하여 향상된 시각화를 가능하게 함으로써 연속적 출력을 원하는 통계적 자료 분석자들에게 호응을 얻을 수 있다는 점이다.

참고문헌

[1] 전성해·전홍석·황진수 (2002) "자기조직화 지도를 위한 베이지안 학습", 응용통계연구 15권. 252-267.

[2] Campos, M.M., and Carpenter, G.A. (2000). "Building adaptive basis functions with a continuous self-organizing map," *Neural Processing Letter*, **11**. 59-78.

[3] Goppert, J. and Rosenstiel, W. (1997). "The continuous interpolating self-organizing map," *Neural Processing Letter*, **5**. 185-192.

[4] Hastie, T., Tibshirani, R. and Friedman, J. (2001). *The Elements of Statistical Learning.* Springer, New York. 480-485 (Section 14.4).

[5] Kohonen, T. (1998). "The self-organizing map," *Neurocomputing*, **21**, 1-6.

[6] Kohonen, T. (1995). *Self-Organizing Maps.* Springer, Berlin.

[7] Koikkalainen, P. (1999). "Tree structured self-organizing maps," in *Kohonen Maps* (Edited by E. Oja and Kaski, S.). Elsevier Science, B.V.

[8] Ripley, R.D. (1996). *Pattern Recognition and Neural Network.* University Press, Cambridge. 322-326 (Section 9.4).

[2003년 1월 접수, 2003년 5월 채택]

Principal Components Self-Organizing Map PC-SOM*

Myung-Hoe Huh [1)]

ABSTRACT

Self-organizing map (SOM), a unsupervised learning neural network, has been developed by T. Kohonen since 1980's. Main application areas were pattern recognition and text retrieval. Because of that, it has not been spread to statisticians until late. Recently, SOM's are frequently drawn in data mining fields.

Kohonen's SOM, however, needs improvements to become a statistician's standard tool. First, there should be a good guideline as for the size of map. Second, an enhanced visualization mode is wanted.

In this study, principal components self-organizing map (PC-SOM), a modification of Kohonen's SOM, is proposed to meet such needs. PC-SOM performs one-dimensional SOM during the first stage to decompose input units into node weights and residuals. At the second stage, another one-dimensional SOM is applied to the residuals of the first stage. Finally, by putting together two stages, one obtains two-dimensional SOM. Such procedure can be easily expanded to construct three or more dimensional maps.

The number of grid lines along the second axis is determined automatically, once that of the first axis is given by the data analyst. Furthermore, PC-SOM provides easily interpretable map axes. Such merits of PC-SOM are demonstrated with well-known Fisher's iris data and a simulated data set.

Keywords: Kohonen's self-organizing map (SOM); unsupervised learning; neural network; visualization; PC-SOM; Fisher's iris data.

* This research was supported by a Korea University Grant during the academic year 2002.

1) Professor, Dept. of Statistics, Korea University. Anam-Dong 5-1, Seoul 136-701, Korea.
E-mail : stat420@korea.ac.kr