응용통계연구 제 18권 1호, 2005년, pp. 83-98

SOM에서 개체의 시각화

엄익현[1] 허명회[2]

요 약

다변량 자료를 분석하는 데 있어서 관측 개체들의 분포적 양태를 파악하는 것은 자료 특성의 이해에 도움이 될 뿐만 아니라 이후 모형화 과정에도 큰 도움을 준다. 이를 위하여 다변량자료의 저차원 시각화에 대한 많은 연구가 진행되어 왔다. 그 중 하나가 코호넨(T. Kohonen)의 자기조직화지도(Self-Organizing Map; SOM)이다. SOM은 저차원 그리드 공간에 고차원 다변량 자료를 축약하여 시각적으로 나타내는 비지도 학습법의 일종으로 최근 들어 통계 분석자들이 많은 관심을 가지고 있는 분야이다. 그러나 SOM은 개체공간의 연속형으로 표현되는 개체를 저차원 그리드 공간에 승자노드에 의해 비연속적으로 표현한다는 단점을 지니고 있다. 본 논문에서는 SOM을 통계적 목적으로 사용하기 위해 요구되는 그리드 공간에 개체를 연속적으로 표현하는 방법들을 제안하고 활용 예를 제시하고자 한다.

주요용어: 자기조직화 지도(SOM), 코호넨(T. Kohonen), 비지도 학습, 시각화, IL-SOM, 부노드(k) SOM.

1. 연구 배경 및 목적

SOM은 1980년대 초반 핀란드의 전기공학자 코호넨(T. Kohonen)에 의해 개발된 비지도 학습(unsupervised learning) 신경망(neural network) 모형의 한 종류이다. 코호넨은 SOM의 특성을 시각화(visualization)과 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. 또한 SOM은 자기조직화(self organization)라는 과정을 통해 다차원 공간의 유사한 개체들을 서로 이웃하는 위치에 오도록 저차원 공간에 배치한다.

전통적인 SOM 알고리즘은 개체공간에 연속형으로 놓이는 개체를 저차원 그리드 공간에 비연속적으로 표현한다. 이는 연속형 입력개체를 이산형으로 출력하게 되므로 개체의 저차원 시각화에 목적을 두는 분석자에게는 그다지 매력적이지 못하다. 본 연구에서 우리는 이 문제를 해결하기 위해 가능도(likelihood)를 이용하는 'IL-SOM' (Interpolating using Likelihood for SOM)과 부노드(subnode)를 이용하는 '부노드(k) SOM'을 제안하고 활용 예를 제시하고자 한다.

1) (135-818) 서울시 강남구 논현동 81-10, (주) 지디에스코리아, 이사/통계학박사
E-mail: abodata@gdskorea.co.kr
2) (136-701) 서울시 성북구 안암동 5가 1, 고려대학교 통계학과, 교수
E-mail: stat420@korea.ac.kr

개체의 시각화에 대한 기존 연구로는 Goppert and Rosenstiel (1997), Campos and Carpenter (2000) 등이 있다. 이 연구들은 승자 노드 중량과 인접 노드 중량의 연결선에 개체를 사영하는 방법에 의존한다. 따라서 개체공간의 차원 p가 커지면 그리드 공간의 승자노드 또는 승자노드와 각 인접노드를 연결하는 선상에 몰리는 형태로 자료점이 집중되는 경향 등 일부 문제가 있다 (엄익현, 2003).

또 다른 연구로 허명회 (2003)가 제안한 PC-SOM이 있다. 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. PC-SOM은 그리드 맵의 축이 지니는 변수적 특성의 제시와 연속적 출력을 통해 시각적으로 향상된 그래프를 제공한다.

개체의 시각화에 대한 다른 방법으로 개체를 승자노드의 주변에 랜덤하게 분포 시키는 방법이 있다. 본 논문에서는 이 방법을 '임의 시각화 SOM'이라고 부르도록 한다. 실제로 통계소프트웨어인 SAS Enterprise Miner와 SPSS의 Clementine 등은 이 방법에 의한 개체표현 방법을 채택하고 있다. 그러나 이 방법은 승자노드의 주변에 개체의 밀도를 시각적으로 보기 좋게 해주기는 하지만 자료 정보전달의 정확성을 저해한다는 단점을 갖는다 (엄익현, 2003).

2. 개체 시각화 방법의 제안

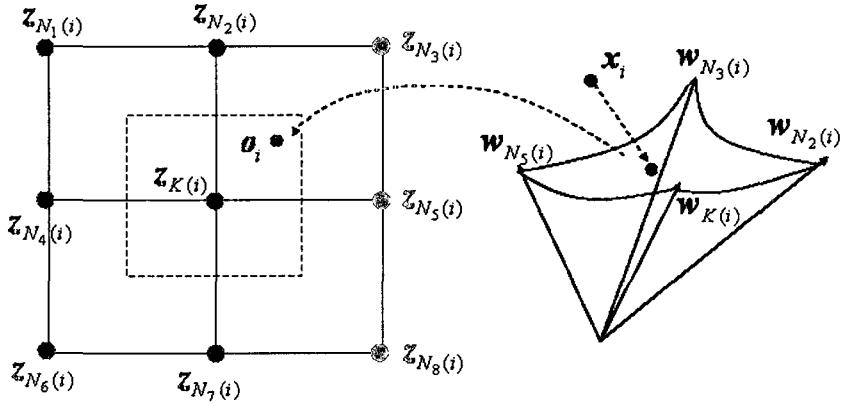
자기조직화지도(self-organizing map, SOM)는 p차원 입력개체공간을 1차원, 2차원, 3차원 등의 저차원 그리드 공간으로 축약하는 방법이다. 본 논문에서는 저차원 그리드 공간상에 개체의 시각화에 대한 방법을 2차원으로 제한하여 제안하고자 한다. 그러나 1차원 또는 3차원 이상의 경우도 자연스럽게 구현 가능하다. 또한 그리드 구성이 사각형인 경우이든 육각형인 경우이든 동일한 방식으로 처리할 수 있으므로 사각형인 경우로 제한하여 다루고자 한다.

본 논문에서 제안하는 개체 시각화 방법의 기본적인 아이디어는 개체벡터의 승자노드와 그 인접노드의 중량벡터에 대한 가능도(likelihood)를 이용하여 그리드 공간상에 개체벡터의 상대적인 위치를 표현한다는 것이다. 이에 이 방법을 IL-SOM (Interpolating using Likelihood for SOM)으로 명명한다. 또 다른 개체 시각화 방법으로 승자노드 주변에 k^2개의 부노드를 배치하여 활용하는 방법을 제안한다. 이를 부노드(k) SOM이라 명명한다.

2.1. IL-SOM

코호넨의 SOM을 수행하면 p차원 개체공간상의 m개의 중량벡터 $\boldsymbol{w}_j$와 개체벡터 $\boldsymbol{x}_i$의 승자노드 $K(i)$의 중량벡터 $\boldsymbol{w}_{K(i)}$를 얻게된다 ($i = 1, \cdots, n,\ j = 1, \cdots, m$). $\boldsymbol{w}_j$의 그리드 공간상의 이미지 좌표를 $\boldsymbol{z}_j$라고 하고, 앞으로 구하게 될 개체벡터 $\boldsymbol{x}_i$의 그리드 공간상의 이미지 좌표를 $\boldsymbol{o}_i$라고 하자. 그리고 개체 $\boldsymbol{x}_i$의 승자노드 $K(i)$를 둘러싸고 있는 8개의 인접노드를 $N_q(i)$로 표기하자 ($q = 1, \cdots, 8$). 이를 그림으로 표현하면 그림 2.1과 같다.

$L(\boldsymbol{w}_j|\boldsymbol{x}_i)$을 개체벡터 $\boldsymbol{x}_i$의 중량벡터 $\boldsymbol{w}_j$에 대한 가능도로 다음과 같이 정의하자.



(a) 2차원 그리드 공간 (b) p차원 개체공간

그림 2.1: 저차원 그리드 공간에서의 개체 시각화

$$L(w_j|x_i) = \text{상수} \cdot \exp[-1/2 \cdot (x_i - w_j)^T \Sigma^{-1} (x_i - w_j)] \tag{2.1}$$

여기서, $j = \{K(i), N_1(i), \cdots, N_8(i)\}$이고, Σ는 알려지지 않은 상수 β에 대해서 $\Sigma = \beta I$라 가정하자. 따라서 개체벡터 x_i의 승자노드 $K(i)$와 인접노드 $N_1(i), \cdots, N_8(i)$의 중량벡터, 즉 $w_{K(i)}$와 $w_{N_1(i)}, \cdots, w_{N_8(i)}$에 대한 가능도 $L(w_{K(i)}|x_i)$와 $L(w_{N_1(i)}|x_i), \cdots, L(w_{N_8(i)}|x_i)$을 각각 구할 수 있다. 여기서, 가능도에 대한 Gauss 분포를 가정하였지만 특별한 이유가 있는 것은 아니다. 그러나 Gauss 분포로써 평균과 산포를 동시에 고려할 수 있기 때문에 쉽게 생각할 수 있는 한 방편이다.

개체 x_i가 그리드 공간상에 표현될 위치 o_i를 구하기에 앞서 그림 2.2와 같이 승자노드를 중심으로 인접노드에 의해서 구성되는 네 개의 사각형 중 개체가 표현될 사각형, 즉 개체를 표현할 방향을 결정한다. 그 후에 방향을 구성하는 네 개의 노드에 대한 x_i의 가능도의 상대적 비율을 이용하여 o_i의 좌표를 산출한다.

방향을 결정하는 이유는 개체의 위치를 잡아주는데 있어서 승자노드와 8개의 인접노드를 모두 고려할 경우, 승자노드를 중심으로 반대방향에 있는 인접노드들의 잡아당김의 정도가 서로 상쇄되어 개체의 위치가 승자노드 쪽에 너무 몰려서 표현되는 경향이 있게 되기 때문이다. 이러한 문제를 해결하기 위하여 그림 2.2와 같이 승자노드를 중심으로 4개의 방향을 나누어 그 중 개체를 표현할 하나의 방향을 선택한 후 그 방향을 구성하고 있는 노드만을 이용하여 개체를 표현하는 것을 고려한다.

각 방향을 구성하고 있는 4개의 노드, 즉 승자노드와 3개의 인접노드의 중량벡터의 가능도들의 합 pd_d을 다음과 같이 구한다.

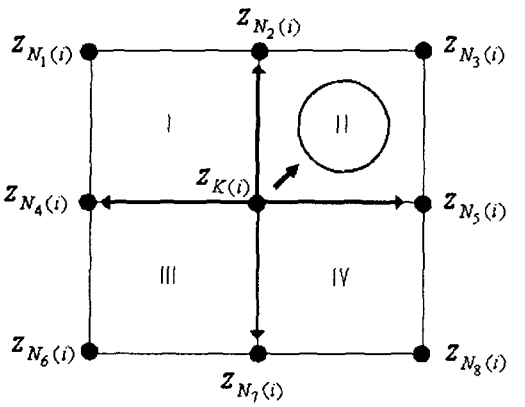


그림 2.2: 개체가 표현될 방향의 결정

$$
\begin{aligned}
pd_I &= \sum_{\{j:\, K(i),N_1(i),N_2(i),N_4(i)\}} L(\boldsymbol{w}_j|\boldsymbol{x}_i) \\
pd_{II} &= \sum_{\{j:\, K(i),N_2(i),N_3(i),N_5(i)\}} L(\boldsymbol{w}_j|\boldsymbol{x}_i) \\
pd_{III} &= \sum_{\{j:\, K(i),N_4(i),N_6(i),N_7(i)\}} L(\boldsymbol{w}_j|\boldsymbol{x}_i) \\
pd_{IV} &= \sum_{\{j:\, K(i),N_5(i),N_7(i),N_8(i)\}} L(\boldsymbol{w}_j|\boldsymbol{x}_i)
\end{aligned}
\tag{2.2}
$$

여기서 첨자 $d = \{I, II, III, IV\}$는 4개의 방향을 가리키는 인덱스이다. 다음과 같이 가장 큰 pd_d값을 가지는 방향을 개체가 표현될 방향으로 결정하기로 한다.

$$\max(pd_I, pd_{II}, pd_{III}, pd_{IV}) \tag{2.3}$$

개체가 표현될 방향을 결정한 후, 개체 $\boldsymbol{x}_i$가 그리드 공간상에 위치할 좌표 $\boldsymbol{o}_i$를 구해 보자. 설명의 편의를 위하여 제2방향이 결정된 경우를 예를 들어 구해 보자. 개체벡터 $\boldsymbol{x}_i$의 승자노드와 인접노드의 중량벡터, $\boldsymbol{w}_{K(i)}$와 $\boldsymbol{w}_{N_2(i)}$, $\boldsymbol{w}_{N_3(i)}$, $\boldsymbol{w}_{N_5(i)}$에 대한 네 개 가능도들의 전체 합 $\sum_{\{k:\, K(i),N_2(i),N_3(i),N_5(i)\}} L(\boldsymbol{w}_k|\boldsymbol{x}_i)$을 구하고, 이로부터 각각의 중량벡터가 차지하는 비율 p_j을 다음과 같이 구한다.

$$
\begin{aligned}
p_j &= \frac{L(\boldsymbol{w}_j|\boldsymbol{x}_i)}{\sum_{\{k:\, K(i),N_2(i),N_3(i),N_5(i)\}} L(\boldsymbol{w}_k|\boldsymbol{x}_i)} \\
&= \frac{\exp[-1/(2\beta)(\boldsymbol{x}_i - \boldsymbol{w}_j)^T(\boldsymbol{x}_i - \boldsymbol{w}_j)]}{\sum_{\{k:\, K(i),N_2(i),N_3(i),N_5(i)\}} \exp[-1/(2\beta)(\boldsymbol{x}_i - \boldsymbol{w}_k)^T(\boldsymbol{x}_i - \boldsymbol{w}_k)]}.
\end{aligned}
\tag{2.4}
$$

네 개의 p_j값을 이용하여 다음과 같이 개체 x_i의 그리드 공간상의 이미지 o_i를 구함으로서 개체를 그리드 공간상에 표현한다. 즉,

$$o_i = \sum_{\{j:\, K(i),N_2(i),N_3(i),N_5(i)\}} p_j \cdot z_j, \tag{2.5}$$

물론 $\sum_j p_j = 1$이다.

2.2. 가장자리 노드가 승자노드인 경우의 개체의 시각화

개체의 시각화에서 그리드 공간의 가장자리 노드가 승자노드일 때를 생각해 보자. 앞에서와는 달리, 승자노드가 그리드의 꼭지점에 위치할 경우에는 승자노드의 인접노드 수는 3개이고 꼭지점이 아닌 가장자리에 위치할 경우에는 승자노드의 인접노드 수는 5개이다. 이것들만 가지고 식 (2.2)와 (2.3)에서와 같이 방향을 결정하고 개체를 표현하는 방법을 생각하는 것은 바람직하지 않다. 왜냐하면 개체공간에서 모든 개체의 위치가 가장자리 중량벡터 안쪽에만 있지는 않을 것이며 가장자리 중량벡터 바깥쪽에도 어느 정도의 개체가 위치하는 것이 타당하기 때문이다. 따라서 그리드 공간상에 개체를 표현할 때 그리드 바깥쪽에 가상의 노드를 설정하여 가장자리 노드의 바깥쪽에도 개체가 표현될 수 있게 하는 것이 바람직하다.

가상의 노드의 개체공간상의 중량벡터를 w_{e1}라고 하고 가장자리 노드의 중량벡터를 w_1, w_1을 중심으로 w_{e1}의 정반대 방향 인접노드의 중량벡터를 w_2라 하자. 선형보간법에 의하여 w_{e1}를 설정하면 다음과 같다.

$$w_{e1} = 2w_1 - w_2 \tag{2.6}$$

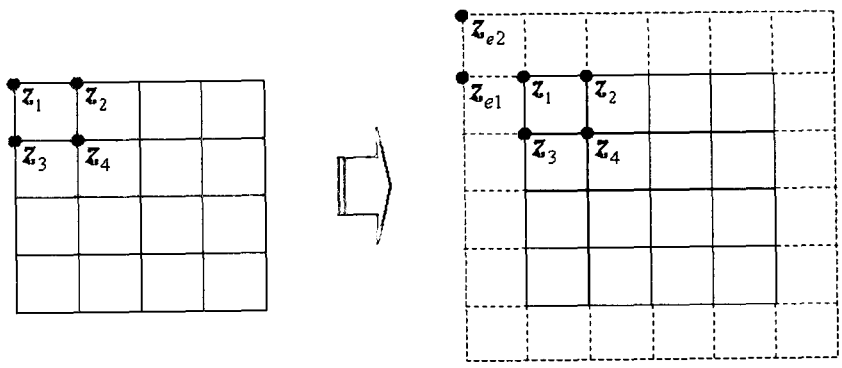


그림 2.3: 그리드의 바깥쪽에 가상의 노드 설정

중량벡터 w_{e1}에 대응하는 그리드 공간상의 노드 z_{e1}를 그림으로 표현하면 그림 2.3과 같다. 마찬가지 방식으로, 바깥쪽 가상 노드 중 4개의 모퉁이에 위치한 가상 노드는 다음과 같이 설정한다.

$$w_{e2} = 4w_1 - 2w_2 - 2w_3 + w_4 \tag{2.7}$$

그림 2.3은 중량벡터 w_{e2}에 대응하는 그리드 공간상의 노드 z_{e2}를 보여준다. 식 (2.6)과 식 (2.7)과 같은 방법으로 바깥쪽 가상노드들을 모두 설정할 수 있다. 이제 모든 가장자리 노드에 대한 인접노드의 수는 동일하게 되었으므로 가장자리 노드가 승자인 경우에서도 개체의 시각화 작업을 2.1절에서와 동일한 방식으로 처리한다.

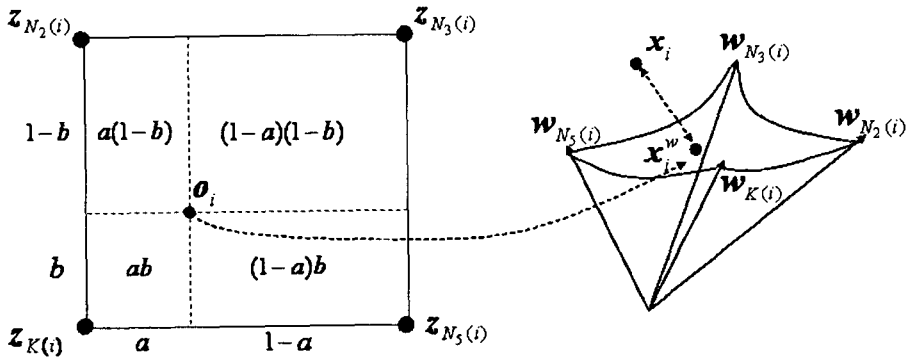
2.3. 상수 β의 결정

이제 개체를 그리드 공간상에 표현하기 위하여 식 (2.1)에서의 β값을 결정하는 문제만이 남았다. β값에 따라 그리드 공간에 표현되는 o_i들의 위치는 다르므로 적정한 β값을 선택해야 한다.

먼저 그리드 공간상에 표현된 개체 o_i를 개체공간 상의 네 개의 중량벡터에 의해서 구성되는 영역에 옮겨 놓는 문제를 생각해보자. 그림 2.4(a)와 같은 상황을 설정하자. 2차원 그리드 공간에서 o_i은 가로방향으로 승자노드 $z_{K(i)}$와 가로방향의 인접노드 $z_{N_5(i)}$를 $a:(1-a)$로 내분하는 위치에 있다. 또한 세로방향으로 승자노드 $z_{K(i)}$와 세로방향의 인접노드 $z_{N_2(i)}$를 $b:(1-b)$로 내분하는 위치에 있다. 따라서 다음과 같은 관계가 성립한다.

$$\begin{aligned} o_i = {} & (1-a)(1-b)z_{K(i)} + (1-a)bz_{N_2(i)} \\ & + a(1-b)z_{N_5(i)} + abz_{N_3(i)} \end{aligned} \tag{2.8}$$

식 (2.8)의 관계를 개체공간으로 옮겨가면 다음과 같이 표현할 수 있는데, 이를 통해서 개체공간상의 네 개의 중량벡터로 구성되는 영역에 o_i의 이미지 x_i^w를 표현할 수 있다. 그림 2.4(b)을 보라.



(a) 2차원 그리드 공간 (b) p차원 개체공간

그림 2.4: 개체공간에 o_i의 이미지 x_i^w의 표현

$$\begin{aligned} \boldsymbol{x}_i^w &= (1-a)(1-b)\boldsymbol{w}_{K(i)} + (1-a)b\boldsymbol{w}_{N_2(i)} \\ &\quad + a(1-b)\boldsymbol{w}_{N_5(i)} + ab\boldsymbol{w}_{N_3(i)} \end{aligned} \tag{2.9}$$

다음으로 개체벡터 $\boldsymbol{x}_i$와 $\boldsymbol{x}_i^w$간의 제곱거리의 합을 다음과 같이 정의하고, 제곱거리의 합을 가장 작게 하는 β을 선택하기로 한다.

$$Q_\beta = \min_\beta \sum_{i=1}^{n} ||\boldsymbol{x}_i - \boldsymbol{x}_i^w(\beta)||^2 \tag{2.10}$$

여기서, $\boldsymbol{x}_i^w$가 β의 함수라는 의미에서 $\boldsymbol{x}_i^w(\beta)$로 표현했다. 이 때 구한 최소제곱거리를 Q_β로 표기하고 이를 $r \times c$ 그리드에서의 '개체표현지수'라 명명한다.

2.4. 변수의 표현

SOM 학습을 통해서 얻어진 저차원 그리드 공간에 변수들을 표현하는 방법을 살펴보자. j번째 변수의 방향을 크기 $p \times 1$ 벡터 $\boldsymbol{v}_j$로 나타내자 ($j = 1, \cdots, p$). 표준화 자료를 사용하는 경우 j번째 변수에 대한 좋은 표현은 단위벡터이고, 비표준화 자료를 사용하는 경우는 표준편차 벡터이다. SOM의 입력자료는 사전에 표준화되는 것이 바람직하므로 j번째 변수에 대한 표현으로 단위벡터를 사용하도록 한다 (허명회, 2003).

주성분사영과 같은 선형사영에서는 변수축이 저차원공간에 하나의 직선으로 사영되기 때문에 다음과 같이 1 표준편차에 해당하는 단위벡터만을 저차원공간에 사영하여 변수를 시각적으로 관찰할 수 있다.

$$\boldsymbol{v}_j = (0, \cdots, 1, \cdots, 0) \tag{2.11}$$

그러나 SOM은 비선형사영이므로 변수축이 저차원공간에 하나의 직선이 아닌 곡선형태로 표현된다. 따라서 -3 표준편차에서 +3 표준편차까지의 범위를 0.5 표준편차 간격으로 다음과 같이 13개의 벡터를 표현하는 것을 제안한다.

$$\boldsymbol{v}_j = (0, \cdots, s, \cdots, 0) \tag{2.12}$$

여기서, $s = -3, -2.5, \cdots, 0, \cdots, +2.5, +3$.

식 (2.12)의 변수벡터 $\boldsymbol{v}_j$를 그리드 공간에 표현하는 방법은 개체벡터를 표현할 때와 동일하며 다음과 같이 요약된다.

step 0. $s = -3$으로 설정한다.

step 1. $\boldsymbol{v}_j$의 승자노드를 구한다.

step 2. 식 (2.2)와 (2.3)에서와 같은 방법으로 $\boldsymbol{v}_j$의 그리드 공간상의 좌표 $\boldsymbol{o}_{v_j}$가 놓일 방향을 선택한다.

step 3. 식 (2.4)와 (2.5)에서와 같은 방법으로 $\boldsymbol{o}_{v_j}$를 구한다.

step 4. s를 0.5 단위로 증가하여 3이 될 때까지 step 1~3을 반복한다.

이렇게 p개의 변수축을 동일한 저차원공간에 표현함으로서 개체와 변수를 시각적으로 관찰해 각 개체의 특징을 비교해 보거나 변수들 사이의 연관성을 판단해 볼 수 있겠다.

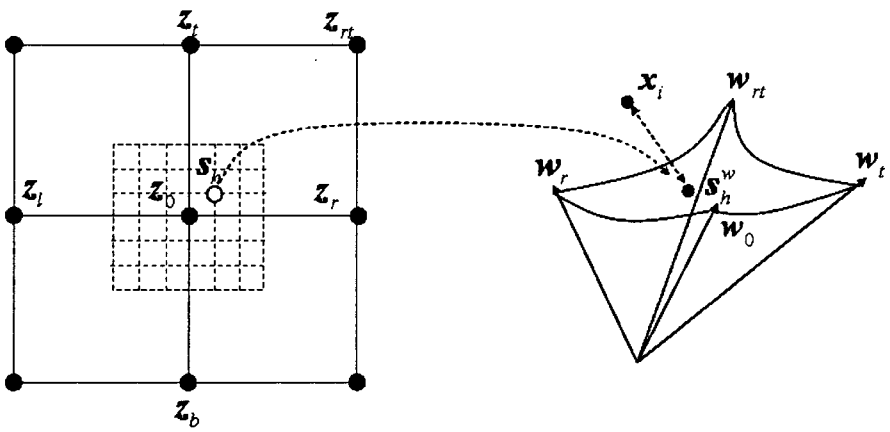
2.5. 부노드(k) SOM

이 소절에서는 개체표현 방법으로 부노드(k) SOM이라 명명한 k^2개의 부노드(subnode)를 이용하는 방법을 제안하기로 한다. 그리드 공간에서 승자노드를 z_0, 이것의 왼쪽 인접 노드를 z_l, 오른쪽 인접 노드를 z_r, 위쪽 인접 노드를 z_t, 아래쪽 인접 노드를 z_b라고 하자. 그리고 노드 z_0, z_l, z_r, z_t, z_b에 대한 개체공간에서 중량벡터를 각각 w_0, w_l, w_r, w_t, w_b라고 하자. 그림 2.5를 보라.

그리드 공간의 노드 z_0와 z_l을 연결하는 선분과 z_0와 z_r을 연결하는 선분, z_0와 z_t를 연결하는 선분, z_0와 z_b를 연결하는 선분을 k등분하는 선분을 각각의 선분에 수직으로 긋는다 (여기서 k는 홀수). 따라서 이들 선분에 의해서 만나는 점을 모두 $(2k-1)^2$개 얻게 된다. 이들 점 중에서 주변노드 z_l, z_r, z_t, z_b 보다 승자노드 z_0에 더 가까운 점을 k^2개 얻을 수 있는데 이를 부노드(subnode)라 부르고 s_h라 표기하자 ($h = 1, \cdots, k^2$).

부노드(k) SOM은 이들 k^2개의 부노드들을 2.3절에서 제안한 IL-SOM의 식 (2.8)과 같은 방법으로 개체공간에 표현한다. 개체공간에 표현된 부노드를 s_h^w로 표기하자.

부노드 s_h를 개체공간에 표현한 후 개체벡터 x_i와 s_h^w와의 제곱거리를 구한다 ($h = 1, \cdots, k^2$). 이를 k^2개 부노드에 대해서 모두 실시한 후 최소제곱거리를 갖는 s_h^w에 대응하는 부노드 s_h를 개체 x_i의 그리드 공간상의 좌표로 정하는 것을 제안한다. 즉,



(a) 2차원 그리드 공간 (b) p차원 개체공간

그림 2.5: 부노드(k) SOM을 이용한 개체의 표현

$$\min_h ||\boldsymbol{x}_i - \boldsymbol{s}_h^w||^2, \quad h = 1, \cdots, k^2 \tag{2.13}$$

3. 붓꽃 자료 예

본 논문에서 제안한 IL-SOM과 부노드(k) SOM을 코호넨의 SOM, Goppert and Rosenstiel (1997)가 제안한 ISOM과 임의 시각화 SOM 등과 함께 Fisher의 붓꽃 자료에 적용해본 후 다섯 가지 방법의 개체표현의 수행능력(performance)을 비교하고자 한다.

Fisher의 붓꽃 자료(Iris Data)는 세 가지 품종(1: setosa, 2:versicolor, 3:verginica)의 붓꽃으로부터 각각 50개, 총 150개의 개체를 추출하여 측정한 네 개의 변수 X_1: 꽃받침 조각의 길이 (sepal length), X_2: 꽃받침 조각의 폭 (sepal width), X_3: 꽃잎의 길이 (petal length), X_4: 꽃잎의 폭 (petal width)으로 구성되어 있다. 그림 3.1은 자료에 대한 산점도 행렬이다. 여기서는 각 방법들의 수행능력을 보기 위한 것이 주목적이므로 품종 변수를 SOM 생성에 사용하지 않기로 하겠다. 즉 입력변수는 $X_1 \sim X_4$이고 사전에 표준화 변환되었다 (평균 0, 표준편차 1).

그림 3.2는 그리드의 크기를 5×5로 지정하여 본 논문에서 제안한 IL-SOM을 적용한 결과이다. SOM 수행시, 사전에 그리드 크기를 5×5, 초기 학습률 0.25, 최종 학습률 0.001, 초기 주변거리 2, 최종 주변거리 1, 주기 당 반복 50회 등으로 지정한 결과이다. 이에 따라, 최종 출력을 얻는데 150×2×50=15,000회의 업데이팅이 이루어진 셈이다. 연구자가 작성한 SAS/IML 프로그램을 SOM 산출을 위하여 사용하였다. 각 그림의 노드 위의 큰 숫자는 해당노드를 승자노드로 가지는 개체의 수이며, 그리드상에 넓게 퍼져 있는 작은 숫자는 품종번호를 표시하며 해당 위치가 개체의 표현 위치이다.

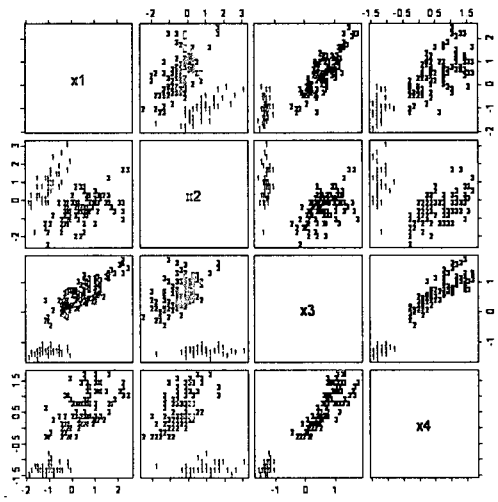


그림 3.1: 붓꽃 자료에 대한 산점도 행렬: 숫자는 품종번호를 표시함

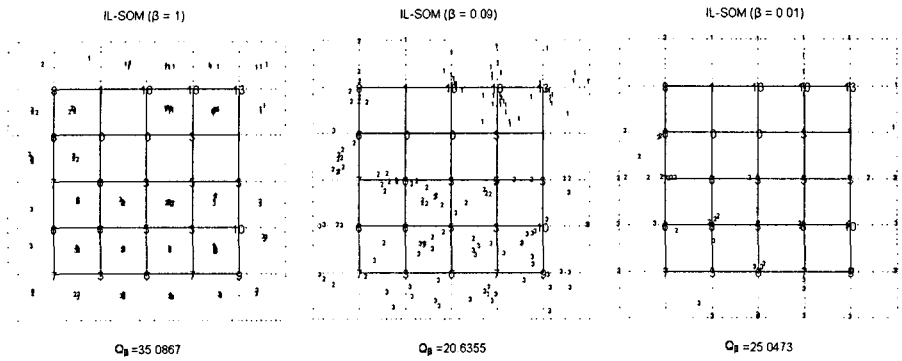


그림 3.2: 붓꽃 자료에 대한 5×5 그리드 공간에서 IL-SOM을 이용한 개체표현

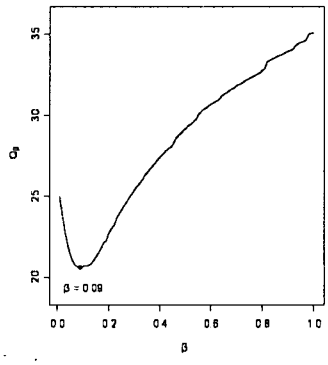


그림 3.3: 붓꽃 자료에 대한 5×5 그리드 공간에서 IL-SOM의 β선택

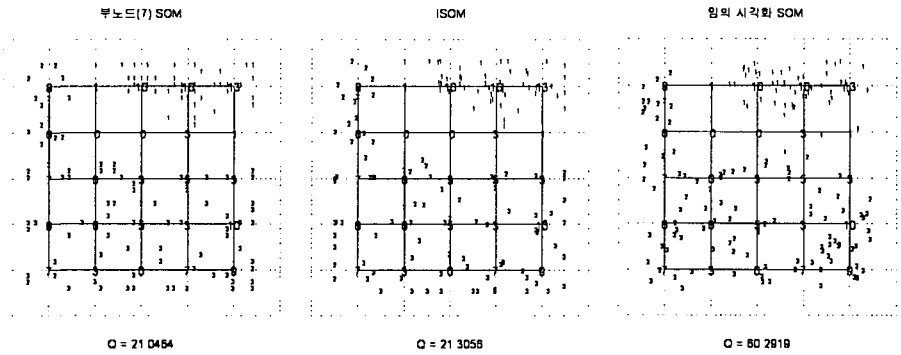


그림 3.4: 붓꽃 자료에 대한 5×5 그리드 공간에서 부노드(7) SOM과 ISOM, 임의시각화 SOM을 이용한 개체표현

표 3.1: 붓꽃 자료에 대한 5×5 그리드 공간에서 각 방법의 개체표현 결과 비교

방법	최적의 β	Q_β(또는 Q)	비교
IL-SOM	0.09	20.6355	가장 작음
부노드(7) SOM	-	21.0464	
코호넨 SOM	-	38.7422	
ISOM ($\lambda = 0.001$)	-	21.3056	
임의 시각화 SOM	-	60.2919	

그림 3.2에는 β값 중에서 1, 0.09, 0.01일 경우의 개체표현을 각각 보여주고 있다. 2.3절에서도 언급했듯이 β값에 따라 그리드 공간에 표현되는 o_i들의 위치는 다르게 된다. β값이 커질수록 o_i들은 네 개의 노드들로 구성되는 사각형의 중심 쪽으로 몰리는 경향이 있으며, β값이 작아지면 o_i들은 승자노드 쪽으로 몰리는 경향을 보인다. 그림 3.3을 보라. $\beta = 0.09$일 때 최소인 개체표현지수 Q_β를 가지며 그 때 $Q_{0.09}$는 20.6355이다.

그림 3.4는 부노드(k) SOM, ISOM과 임의시각화 SOM에 대한 결과를 보여준다. 여기서는 부노드(k) SOM으로 $k = 7$을 고려하였지만 더 조밀한 보간법도 필요하다면 쓸 수 있을 것이다. ISOM의 경우 $\lambda = 0.001$로 놓았다.

Fisher의 붓꽃 자료에 대해서 본 논문에서 제안한 두 가지 방법과 코호넨의 SOM, ISOM, 그리고 임의 시각화 SOM의 개체표현지수 Q_β(또는 Q), 즉 제곱거리의 합을 비교한 결과를 표 3.1에 정리하였다. Q_β(또는 Q)는 실제 개체 $\boldsymbol{x}_i$와 저차원에 표현된 o_i의 개체공간에서의 상대적 위치 $\boldsymbol{x}_i^w$와의 제곱거리의 합을 계산한 것으로서, Q_β(또는 Q)가 작을 수록 저차원 그리드 공간상에 개체를 더 잘 표현하는 것이라고 볼 수 있다.

각 방법들의 개체표현 수행능력을 비교해 보면, IL-SOM이 가장 좋고 ($Q_{0.09} = 20.6355$), 다음으로 부노드(7) SOM ($Q = 21.0464$), ISOM ($Q = 21.3056$)로 나타났다. 코호넨의 SOM을 그대로 적용하였을 때의 개체표현지수 Q는 38.7422이고 임의 시각화 SOM에서 Q는 60.2919로 앞의 세 가지 방법에 비해서 상당히 큰 값을 가진다.

이제 변수의 시각화에 대해서 생각해보자. 표 3.1에서 가장 작은 개체표현지수를 가지는 IL-SOM($\beta = 0.09$)을 선택하고 이를 이용하여 5×5 그리드 공간에 변수를 표현해보자.

그림 3.5는 그리드 공간에 변수를 표현한 것으로 (a)~(d)는 각각 $X_1 \sim X_4$의 변수를 표현한 그림이다. 그리드 공간에서 변수축은 비선형으로 표현된다. 그림에서 알파벳은 원점(o)를 중심으로 -3(a)~-0.5(f)와 0.5(g)~3(l)이고, 점선으로된 화살표는 변수축의 비선형적 모양을 대략적으로 그려 넣은 것이다.

위와 같은 변수의 표현은 개체의 그룹과 변수를 시각적으로 관찰해 각 그룹의 특징을 비교해 보거나 또는 그룹에 있어 큰 영향력을 행사하는 변수를 판단해 볼 수 있게 해준다. 그림 3.5를 그림 3.1의 산점도 행렬과 함께 살펴보는 것도 변수의 특성을 파악하는데 좋은 방법이다. 변수 X_1은 좌측 상단에서 우측 하단의 대각선 방향으로, X_2는 좌측에서 우측으로, 변수 X_3과 X_4는 그리드의 상단에서 하단으로 움직이는 것을 볼 수 있다.

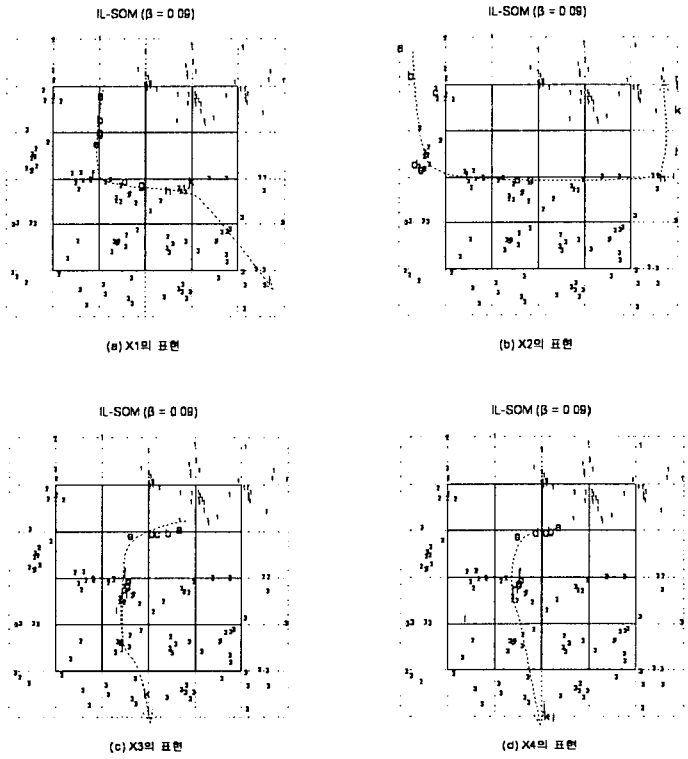


그림 3.5: 붓꽃 자료에 대한 5×5 그리드공간에서 IL-SOM을 이용한 변수표현
(a:-3, b:-2.5, c:-2, d:-1.5, e:-1, f:-0.5, o:0, g:0.5, h:1, i:1.5, j:2, k:2.5, l:3)

4. 모의자료 예

이 절에서는 구조가 잘 알려진 모의자료에 대해서 그리드공간상에 개체를 표현해 보도록 하자. 개체 및 변수의 표현이 기대하는 바와 일치하는 가를 확인하는 것이 목적이다.

모의자료는 네 개의 그룹으로 구분되며, 각 그룹별 50개의 개체, 총 200개의 개체를 생성하였다. 각 그룹은 다음과 같은 평균과 공분산행렬을 가지는 다변량정규분포를 따른다.

1 그룹: 평균이 $\left(-2.5, -\frac{2.5\cdot\sqrt{3}}{3}, -\frac{2.5\cdot 2\cdot\sqrt{2}}{4\cdot\sqrt{3}}, 0\right)$이고 공분산행렬이 $\boldsymbol{I}_4$,

2 그룹: 평균이 $\left(2.5, -\frac{2.5\cdot\sqrt{3}}{3}, -\frac{2.5\cdot 2\cdot\sqrt{2}}{4\cdot\sqrt{3}}, 0\right)$이고 공분산행렬이 $\boldsymbol{I}_4$,

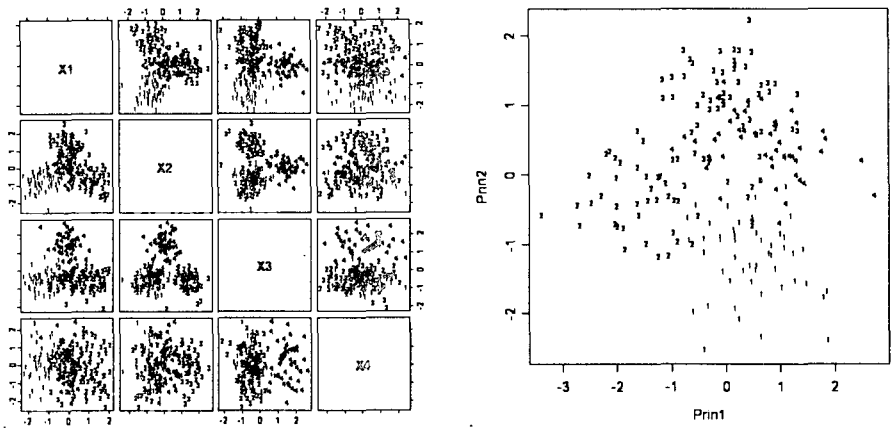
3 그룹: 평균이 $\left(0, \frac{2.5\cdot 2\cdot\sqrt{3}}{3}, -\frac{2.5\cdot 2\cdot\sqrt{2}}{4\cdot\sqrt{3}}, 0\right)$이고 공분산행렬이 $\boldsymbol{I}_4$,

4 그룹: 평균이 $\left(0, 0, \frac{2.5 \cdot 6 \cdot \sqrt{2}}{4 \cdot \sqrt{3}}, 0\right)$이고 공분산행렬이 I_4.

이 자료의 네 그룹의 평균을 연결하면 한 변의 길이(유클리드거리)가 5이고, 무게중심이 원점인 정사면체를 형성하게 된다. 여기서 변수 X_4는 그룹의 구분과는 상관이 없게 생성된 의미없는 변수이다. 이러한 정사면체 구조를 가지는 자료의 경우 2차원 평면에 선형사형을 하였을 때 어느 평면에 사영을 하던 4개의 그룹이 매끄럽게 잘 구분이 되지 않고 일부 그룹은 겹쳐서 표현된다. 그림 4.1(a)는 모의자료의 4개 변수에 대한 산점도행렬을 그림 4.1(b)은 주성분 사영을 시행한 결과이다. 작은 숫자는 그룹번호를 표시하며 해당 위치가 개체의 표현 위치이다. 1, 2, 3 그룹의 경우 잘 구분되어 사영된 것으로 보이나 4그룹은 다른 그룹에 겹쳐서 사영된 것으로 보인다. 이러한 문제가 발생하는 이유는 정사면체가 가지는 위상적인 구조가 선형사영에는 알맞지 않다는 데 기인한다. 따라서 이 경우에는 비선형사영을 고려하는 것이 더 타당해 보이며 SOM을 통해 비선형사영을 시도하고자 한다.

그림 4.2는 그리드의 크기를 5×5로 지정하여 본 논문에서 제안한 IL-SOM과 부노드(7) SOM, ISOM를 이용한 개체표현 결과를 보여준다. 그림 4.2를 통해 볼 수 있듯이 SOM의 비선형사영은 비선형의 위상을 가지는 모의자료에 대해서 그룹을 잘 구분하고 있음을 볼 수 있다. 이 결과는 선형사영을 하는 주성분 사영의 결과인 그림 4.1(b)와 대조적이다.

그림 4.3은 모의 자료에 대한 그리드공간에 변수를 표현한 것으로 (a)~(d)는 각각 $X_1 \sim X_4$의 변수를 표현한 그림이다. IL-SOM($\beta = 0.43$)을 이용하여 5×5 그리드공간에 변수를 표현하였다. 변수의 표현이 기대했던 바와 일치하는지를 알아보기 위하여 그림 4.3에서 각 변수를 살펴보자. 변수 X_4은 군집화에 별다른 영향을 미치지 못하고 있으며, 변수 X_1은 1군



(a) 산점도 행렬 (b) 주성분 사영 결과

그림 4.1: 모의자료에 대한 산점도 행렬과 주성분 사영 결과 : 숫자는 그룹번호를 표시함

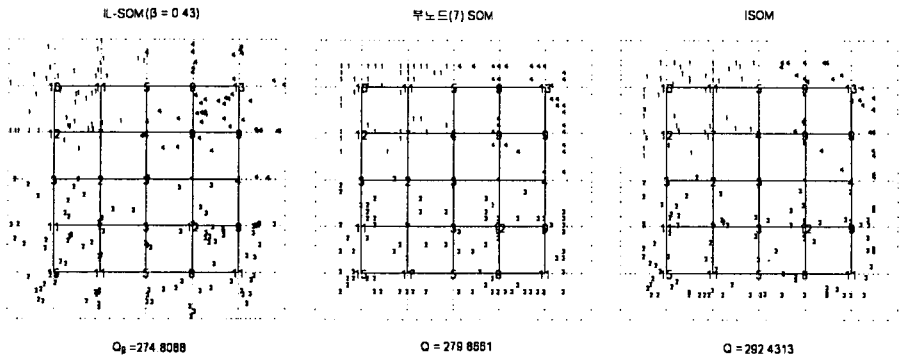


그림 4.2: 모의자료에 대한 5×5 그리드공간에서 IL-SOM과 부노드(7) SOM, ISOM을 이용한 개체표현

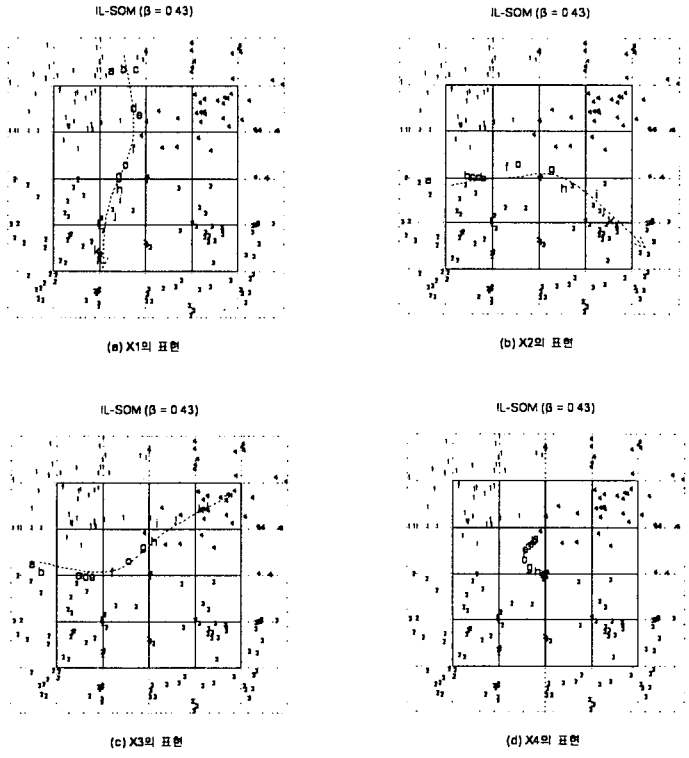


그림 4.3: 모의자료에 대한 5×5 그리드공간에서 IL-SOM을 이용한 변수표현
(a:-3, b:-2.5, c:-2, d:-1.5, e:-1, f:-0.5, o:0, g:0.5, h:1, i:1.5, j:2, k:2.5, l:3)

집과 2 군집의 대조, X_2는 1,2 군집과 3 군집의 대조, X_3는 1,2 군집과 4 군집의 대조를 의미는 것으로 나타나고 있다. 즉, 그림 4.3에서 기대했던 바와 같이 모의자료의 변수를 표현해 주고 있다.

5. 맺음 말

본 연구에서는 SOM에서 개체를 그리드공간상에 시각적으로 표현할 수 있는 두 가지 방법을 제안하였다. 이를 통해 전통적인 SOM이 가지는 이산형 출력에 대한 모순을 개선하고, 개체를 저차원 그리드공간에 표현하는 것이 가능하여 결과의 시각적 해석 및 변수의 표현을 통해 각 군집의 특성 파악이 가능하게 하였다. 따라서, 기존에 연구되어 온 SOM에 비하여 향상된 시각화를 가능하게 함으로써 연속적 출력을 원하는 통계적 자료 분석자들에게 호응을 얻을 수 있다는 장점을 가진다.

참고문헌

엄익현 (2003). 코호넨 자기조직화지도(SOM)의 통계적 활용, 고려대학교 대학원 통계학과 박사학위논문.

허명회 (2003). 주성분 자기조직화 지도 PC-SOM, <응용통계연구> **16**, 321-334.

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

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

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

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

[2004년 7월 접수, 2004년 10월 채택]

Enhancing Visualization in Self-Organizing Maps

Ick-Hyun Um [1] Myung-Hoe Huh [2]

ABSTRACT

Exploring distributional patterns of multivariate data is very essential in understanding the characteristics of given data set, as well as in building plausible models for the data. For that purpose, low-dimensional visualization methods have been developed by many researchers along various directions. As one of methods, Kohonen's SOM (Self-Organizing Map) is prominent. SOM compresses the volume of the data, yields abstraction from the data and offers visual display on low-dimensional grids. Although it is proven quite effective, it has one undesirable property: SOM's display is discrete. In this study, we propose two techniques for enhancing quality of SOM's display, so that SOM's display becomes continuous. The proposed methods are demonstrated in two numerical examples.

Keywords: Kohonen's self-organizing map (SOM), Unsupervised learning, Visualization, IL-SOM, Subnode (k) SOM.

1) Director/Ph.D., GDS Korea Inc., SinChang B/D 1F, 81-10, Nonhyun-dong, Kangnam-gu, Seoul 135-818, Korea
E-mail: abodata@gdskorea.co.kr
2) Professor, Dept. of Statistics, Korea University, 1, 5-ga, Anam-dong, Sungbuk-gu, Seoul 136-701, Korea
E-mail: stat420@korea.ac.kr