The Korean Communications
in Statistics Vol. 10 No. 2, 2003
pp. 507-517

Validity Study of Kohonen Self-Organizing Maps

Myung-Hoe Huh [1)]

Abstract

Self-organizing map (SOM) has been developed mainly by T. Kohonen and his colleagues as a unsupervised learning neural network. Because of its topological ordering property, SOM is known to be very useful in pattern recognition and text information retrieval areas. Recently, data miners use Kohonen's mapping method frequently in exploratory analyses of large data sets.

One problem facing SOM builder is that there exists no sensible criterion for evaluating goodness-of-fit of the map at hand. In this short communication, we propose valid evaluation procedures for the Kohonen SOM of any size. The methods can be used in selecting the best map among several candidates.

KeyWords: Kohonen Self-Organizing Map (SOM), Data Mining, Partitioned Data Sets, Valid Measure of Lack-of-Fit, Re-sampling.

1. Background and Aim

Self-organizing map (SOM) is a unsupervised learning neural network method developed by Teuvo Kohonen of Finland and his colleagues since 1980's (Kohonen, 1995). SOM is known to be very useful in pattern recognition and text information retrieval areas, as demonstrated by numerous studies (http://www.cis.hut.fi/research/som-bibl/, http://www.soe.ucsc.edu/NCS). The main virtue of SOM is the topological ordering property, which enables visualization and abstraction of data sets at the same time (Kohonen, 1998).

SOM has been neglected in statistical community, because of its engineering orientation. Only a few years ago, applied statisticians began to use Kohonen's mapping method in exploratory analyses of large data sets or data mining. Readable writings are available now in several statistical text books such as Ripley (1996) and Hastie, Tibshirani and Friedman (2001). In Korean statistical circle, Jun, Jorn and Hwang (2002) seems to be the first article on SOM. Recently, the author finished writing an article on the topic (Huh, 2003).

1) Professor, Dept. of Statistics, Korea University. Anam-Dong 5-1, Seoul 136-701, Korea.
E-mail: stat420@korea.ac.kr
* The author would like to thank reviewers who pointed out that an important article is not referred in original manuscript and helped to clarify several arguments. This work is done during author's research year at Korea University.

We will briefly sketch SOM algorithm to introduce the terms and notations adopted in this article. Suppose that the input data set consists of n units of p-dimensional measurements $x_1, \cdots, x_n$. The aim of SOM is to arrange $m (\ll n)$ neurons that respond to input units in a self-organized manner. The following algorithm is for typical SOM built on two-dimensional grid of rectangular shape.

0) $m (= c_1 c_2)$ neurons are allocated to one of the nodes on $c_1 \times c_2$ grid. Inside of nodes, p-dimensional latent vectors called the node *weights* $w_{11}, \cdots, w_{c_1 c_2}$ reside. Initially, node weight vectors are set randomly or spaced systematically on major principal components (Kohonen, 1998). Set the time t=1. Node weights are becoming refined as time passes.

1) Once the input unit x_i arrives at the map or the net ($i = 1, \cdots, n$), find the node that has the closest weight vector. Denote the *hit* or *winner* node by $j(i)$ and its weight by $w_{j(i)}$. That is,

$$\| x_i - w_{j(i)} \| \leq \| x_i - w_j \|, \quad \text{for all } j = (1,1), \cdots, (c_1, c_2).$$

Here, $\| \cdot \|$ is Euclidean distance. Because of this, the input data set is normally standardized before being processed by the algorithm.

2) The node weights are updated as follows:

$$w_j \leftarrow w_j + \alpha_t h_t(j, j(i))(x_i - w_j), \quad \text{for all } j \text{ such that } \| r_j - r_{j(i)} \| \leq d_t$$

where r_j is coordinate point of grid node j, α_t is learning rate that decreases from starting value α_0 to ending value α_1 as t increases. $h_t(j, j')$ is local weighting constant that decreases as t or $\| r_j - r_{j'} \|$ increases. For convenience, assume

$$h_t(j, j') = \exp \{ - \| r_j - r_{j'} \|^2 / (2 \sigma_t^2) \},$$

where σ_t^2 is a decreasing function of t. Finally, node neighborhood radius d_t is a decreasing step function of t. Each step value of d_t defines the phase containing a fixed number of cycles.

3) As one input unit is processed in Step 1 and Step 2, increase t by one, and get the next unit ready. As the last unit is processed, start either new cycle or phase with the first unit. In all cases, return to Step 1 unless the following conditions are met: i) The changes in node weights become negligible or ii) the time t reaches pre- determined limit. Finally, output the list of hit nodes for each unit and the array of weight vectors for all grid nodes.

When one draws several SOM's with a given data set, he/she naturally want to compare the maps. For instance, which is a better map between 12x4 SOM and 7x7 SOM? [These two have nearly same number of nodes.] The aim of this short communication is to develop a

valid method for measuring goodness-of-fit's.

In Section 2 and Section 3, data set partitioning and re-sampling procedures are proposed for valid evaluation of the lack-of-fit of given SOM with numerical examples. Concluding remarks follow in Section 4.

2. Validity Evaluation from Partitioned Data Sets

Suppose we have the SOM portraying n input units $x_1, \cdots, x_n$ on the $c_1 \times c_2$ grid of nodes. If one considers

$$G_0 = \frac{1}{n} \sum_{i=1}^{n} \| x_i - w_{j(i)} \|^2$$

as a "lack-of-fit" measure of the map, then he/she will certainly suffer from under-evaluation since the node weights are derived from input units. Hence it is necessary to separate the input units from the node weights of the map statistically to get a valid measure of lack-of-fit. For this purpose, we propose following data set partitioning procedure.

Step 0. Split the n $(= n_1 + n_2)$ input units randomly into two non-overlapping subsets, called Half-sample 1 and Half-sample 2, of nearly equal sizes n_1 and n_2, respectively. Write

Half-sample 1: $x_1^{(1)}, \cdots, x_{n_1}^{(1)}$, and Half-sample 2: $x_1^{(2)}, \cdots, x_{n_2}^{(2)}$.

Step 1-1. Construct a $c_1 \times c_2$ map with Half-sample 1 units. As result, one obtains the flow connections from the unit i_1 to the hitting node $j_1(i_1)$, $i_1 = 1, \cdots, n_1$. Also, node weights $w_{11}^{(1)}, \cdots, w_{c_1c_2}^{(1)}$ become available.

Step 1-2. Parallel works are done with Half-sample 2 units. Denote the hit node of unit i_2 as $j_2(i_2)$ and write $w_{11}^{(2)}, \cdots, w_{c_1c_2}^{(2)}$ for node weights produced.

Step 2. Compute

$$G_2^{(1)} = \frac{1}{n_1} \sum_{i_1=1}^{n_1} \| x_{i_1}^{(1)} - w_{j_1(i_1)}^{(2)} \|^2$$

as the lack-of-fit of Half-sample 1 units. Similarly, compute

$$G_2^{(2)} = \frac{1}{n_2} \sum_{i_2=1}^{n_2} \| x_{i_2}^{(2)} - w_{j_2(i_2)}^{(1)} \|^2$$

as the lack-of-fit of Half-sample 2 units.

It is quite clear that node weights $w_{11}^{(2)}, \cdots, w_{c_1c_2}^{(2)}$ of Half-sample 2 SOM are statistically independent of Half-sample 1 input units $x_1^{(1)}, \cdots, x_{n_1}^{(1)}$. Hence, $G_2^{(1)}$ is an honest estimate of the "lack-of-fit". By symmetry, $G_2^{(2)}$ provides another estimate of the same thing.

The above method is in agreement with a general cluster validation procedure, in which the data set is randomly divided into two subsets and one evaluates the difference between analysis results of subsets. Refer to Gordon (1999).

Iris Data Example

Fisher's iris data set consists of 150 units of 4-dimensional measurements (sepal length, sepal width, petal length, petal width). These four variables are standardized before being processed. Also, observation (input) units are classified into one of three species (1: setosa, 2: versicolor, 3: verginica), but species codes will not be used in the construction of SOM's.

Imposing 12x4 grid over all units, we obtained the apparent lack-of-fit G_0 = 0.130. [In constructing this SOM, we used following specifications: Initial learning rate = 0.25, terminal learning rate = 0.001, starting radius of node neighborhood = 3, ending radius of node neighborhood = 1, number of cycles per phase = 50. The author wrote a SOM program in SAS/IML.]

To evaluate the lack-of-fit not in a overlapping way, the input data was split into two equal halves of 75 units each, Half-sample 1 and Half-sample 2. It turned out that the lack-of-fit statistics are

$$G_2^{(1)} = 0.995,\ G_2^{(2)} = 1.088 \quad \text{[Average 1.042]}.$$

Based on half samples, the apparent lack-of-fit G_0 were 0.094 and 0.089, seriously under-valued compared to cross-validated estimates, 0.995 and 1.088 (= $G_2^{(1)}, G_2^{(2)}$).

As another trial, we constructed a 7x7 SOM, of which G_0 = 0.132 [In constructing the SOM, same specifications are used as before]. We obtained the lack-of-fit statistics:

$$G_2^{(1)} = 1.225,\ G_2^{(2)} = 1.163 \quad \text{[Average 1.194]}.$$

Hence, we may tentatively conclude that 7x7 SOM is inferior to 12x4 SOM, since the former's lack-of-fit's are about 15% larger than those of the latter.

Figure 1 shows the 12x4 and 7x7 SOM's split by species. Of course, one cannot discern different degrees of the lack-of-fit in SOM's with figures only. However, there is one point worthy to note: On 12x4 SOM, input units of the same species are clustered in square shapes, contrasting long rectangular-shape clusters on 7x7 SOM.

(a)

Species1	Species2	Species3
3 3 4 7	. . . .	. . . .
4 1 4 2	. . . .	. . . .
6 6 6 3	. . . .	. . . .
1 . . .	. . . .	. . . .
. . . .	4 2 1 4	. . . .
. . . .	4 2 4 2	. 1 . .
. . . .	2 1 1 5	1 . . .
. . . .	1 1 1 3	2 . 1 .
. . . .	. . . 7	5 2 3 .
. . . .	. . 1 4	4 1 3 .
. . . .	. . . .	. 4 2 4
. . . .	. . . .	6 3 5 3

(b)

Species1	Species2	Species3
3 8 4 6 6 6 6	. . . . .	. . . . .
1 . 1 3 5 1 .	. . . . .	. . . . .
. . . . .	4 1 1 . . 1 3	. . . . .
. . . . .	5 1 6 4 7 . .	. . . . . 1 4
. . . . .	2 6 3 1 1 . .	1 . . 1 . . 5
. . . . .	. . 1	. 1 2 3 3 2 3
. . . . .	3	2 2 4 4 3 3 6

Figure 1. (a) 12x4 SOM and (b) 7x7 SOM for Fisher's Iris Data, Split by Species. The cell numbers represent the number of input units hit on the nodes.

3. Re-Sampling Procedure

In the lack-of-fit evaluation procedure of Section 2, one computed the measures with half samples. So, strictly speaking, they are not for the sample of size n, even though we reasonably expect the measures derived from half samples do proper roles in relative comparison of different SOM's. To overcome such detail problem, we propose a re-sampling validity evaluation procedure for full sample SOM as follows.

Step 0. Re-sample n units independently from $x_1, \cdots, x_n$, and repeat (B times). Then one obtains

$$\text{Subsample } b\text{: } x_1^{*(b)}, \cdots, x_n^{*(b)} \text{ for } b = 1, \cdots, B.$$

Step 1. Construct $c_1 \times c_2$ maps with Subsample b $(=1, \cdots, B)$ units. Output the list of units and corresponding nodes for each subsample SOM. Denote $j_b(i_b)$ for the hitting node for the unit i_b $(=1, \cdots, n)$ on Subsample b SOM. Write node weights of Subsample b SOM as $w_{11}^{*(b)}, \cdots, w_{c_1c_2}^{*(b)}$.

Step 2. For the lack-of-fit of Subsample b units, compute

$$G_B^{(b)} = \frac{1}{n} \sum_{i_b=1}^{n} \| x_{i_b}^{*(b)} - W_{j_b(i_b)}^{*(-b)} \|^2, \quad \text{for } b = 1, \cdots, B$$

where

$$W_{j_b(i_b)}^{*(-b)} = \sum_{b' \neq b} w_{j_b(i_b)}^{*(b')} / (B-1)$$

is the average of all subsample SOM weights on the node $j_b(i_b)$ except that of Subsample b SOM. Note that $W_{11}^{*(-b)}, \cdots, W_{c_1c_2}^{*(-b)}$ are conditionally independent of Subsample b input units $x_1^{*(b)}, \cdots, x_n^{*(b)}$, given the observed sample $x_1, \cdots, x_n$.

Note: de Bote, Cottrell and Verleysen (2002) proposed using bootstrap method for assessment of SOM's reliability. They measured the degree how well neighboring pairs of input units are represented in close nodes on bootstrapped SOM's. In contrast, my use of bootstrap method is for fair assessment of the lack-of-fit sum of squares.

Iris Data Example (Continued)

For the 12x4 SOM on Fisher's iris data, we generated five($=B$) subsamples, of which

$G_B^{(b)}$ = 0.42, 0.31, 0.39, 0.44, 0.40 [Median 0.40, Range 0.13] for $b=1, \cdots, 5$.

In contrast, 7x7 SOM yielded

$G_B^{(b)}$ = 0.98, 0.50, 0.61, 0.39, 0.58 [Median 0.58, Range 0.59] for $b=1, \cdots, 5$.

So, we may conclude clearly that 12x4 SOM is better than 7x7 SOM in terms of validity (honest goodness-of-fit) perspective.

Since it is quite possible that a better SOM compared to the current winner exists, we tried out several SOM's including 12x4. See Table 1 for the result of three repetitions in each case with B = 5. As judged by lack-of-fit's average of medians, the best size for Fisher's iris data seems to be 15x5. See Figure 2 for 15x5 SOM.

Simulated Data Example

We are going to generate simulated data from known structures to assess whether the methods work. Suppose that $Z_1, \cdots, Z_5$ are iid N(0,1) random variables. Define

$$X_1 = (Z_2 + Z_3 + Z_4 + Z_5)/2, \qquad X_2 = (Z_1 + Z_3 + Z_4 + Z_5)/2,$$
$$X_3 = (Z_1 + Z_2 + Z_4 + Z_5)/2, \qquad X_4 = (Z_1 + Z_2 + Z_3 + Z_5)/2,$$
$$X_5 = (Z_1 + Z_2 + Z_3 + Z_4)/2.$$

Then $(X_1, \cdots, X_5)$ follows the mulivariate normal distribution with mean vector 0 and covariance matrix $0.25\,I_5 + 0.75\,J_5$. So, the first principal component PC1 is proportional to the sum of $X_1, \cdots, X_5$. PC1 has variance 4, which is 80% of the total variance. Remaining 20% of total variance is shared equally by four other principal components. Hence, for the one-dimensional reduction of the data, the lack-of-fit measure is equal to 1, if one knows the true model.

Table 1. Subsample Lack-of-Fit's of Fisher's Iris Data in Several SOM's

	Subsample Lack-of-Fit's								
SOM	1	2	3	4	5	Median	Average Median	Range	Average Range
6x2 [2]	0.71	0.53	0.64	0.79	0.79	0.71		0.26	
	0.76	0.59	0.62	0.59	0.85	0.62		0.26	
	0.68	0.80	0.66	0.82	0.89	0.80	0.71	0.23	0.25
9x3 [2]	0.41	0.47	0.44	0.49	0.54	0.47		0.13	
	0.43	0.42	0.45	0.66	0.33	0.43		0.33	
	0.37	0.48	0.86	0.35	0.54	0.48	0.46	0.51	0.32
12x4 [3]	0.42	0.31	0.39	0.44	0.40	0.40		0.13	
	0.43	0.25	0.50	0.60	0.28	0.43		0.35	
	0.32	0.54	0.94	0.36	0.41	0.41	0.41	0.62	0.37
15x5 [3]	0.44	0.31	0.24	0.41	0.36	0.36		0.20	
	0.27	0.21	0.32	0.35	0.27	0.27		0.13	
	0.26	0.66	0.56	0.24	0.34	0.34	0.32	0.32	0.22
18x6 [4]	0.52	0.46	0.24	0.44	0.35	0.44		0.28	
	0.39	0.24	0.38	0.40	0.26	0.38		0.26	
	0.29	0.68	0.62	0.29	0.32	0.32	0.38	0.39	0.31
21x7 [4]	0.57	0.48	0.21	0.21	0.41	0.41		0.36	
	0.24	0.22	0.43	0.45	0.35	0.35		0.23	
	0.26	0.58	0.66	0.29	0.44	0.44	0.40	0.40	0.33

* The number in brackets is the starting radius of node neighborhood. Ending radius is set to 1 and initial/terminal learning rates are 0.25/0.001, number of cycles per phase is 50.

Table 2 shows subsample lack-of-fit's of Simulated Data Set consisting of 400 units on several SOM's such as 10x4, 20x2 and 40x1 [Computer program was written in SAS/IML]. Each case was repeated three times and the number B of subsamples was set to five. By the lack-of-fit criterion of this section, the best map size among several candidates is 40x1, which is one-dimensional array of nodes. It is remarkable that the map's lack-of-fit is equal to 1.18. close to the ideal value 1.0 for one-dimensional reduction.

Species1	Species2	Species3
3 2 1 3 4	. . . . .	. . . . .
. . 3 3 3	. . . . .	. . . . .
5 1 4 2 .	. . . . .	. . . . .
3 7 2 2 1	. . . . .	. . . . .
1	. . . . .	. . . . .
. . . . .	3 1 2 1 3	. . . . .
. . . . .	3 2 4 2 2	. 1 . . .
. . . . .	3 . 1 1 3	1 . . . .
. . . . .	1 . . 2 2	1 1 . . .
. . . . .	. 1 . . 4	4 2 1 . .
. . . . .	. . . 1 4	1 . 3 . .
. . . . .	 4	3 1 2 1 .
. . . . .	. . . . .	1 1 1 2 .
. . . . .	. . . . .	3 1 4 1 3
. . . . .	. . . . .	4 1 2 1 3

Figure 2. 15x5 SOM for Fisher's Iris Data, Split by Species.
The cell numbers represent the number of input units hit on the nodes.

To secure a related data set, we transformed $(X_1, \cdots, X_5)$ into $(Y_1, \cdots, Y_5)$, where

$$Y_1 = \exp(X_1), \cdots, Y_5 = \exp(X_5).$$

Therefore, each of $Y_k = \exp(X_k)$, $k = 1, \cdots, 5$, follows a log-normal distribution.

Table 3 shows subsample lack-of-fit's of Transformed Simulated Data Set of size 400 on several SOM's of same sizes as before. Each case was repeated three times and the number B of subsamples was set to five. Although the 40x1 map showed sightly poor performance compared to the 20x2 map for the transformed data, we may say that the 40x1 grid is still one of the bests.

Figures 3 and 4 show the weight vector's coordinates and frequencies in the nodes of 40x1 SOM, for raw/transformed data sets. In both figures, for each of five variables, node values are running smoothly in an increasing mode. There is a notable difference in the figures, however: The pattern in Figure 3 is fairly linear, while the pattern in Figure 4 is exponential. This meets our expectation.

Table 2. Subsample Lack-of-Fit's of Raw Simulated Data Set in Several SOM's

	Subsample Lack-of-Fit's								
SOM	1	2	3	4	5	Median	Average Median	Range	Average Range
10x4 [3,1]	1.98	1.66	1.40	1.69	1.80	1.69		0.58	
	1.61	1.62	2.21	1.61	1.69	1.62		0.60	
	1.69	1.35	1.33	1.81	1.90	1.69	1.67	0.57	0.58
20x2 [5,3]	1.32	1.30	1.15	1.12	1.29	1.29		0.20	
	1.17	1.30	1.26	1.08	1.25	1.25		0.22	
	1.12	1.21	1.23	1.24	1.18	1.21	1.25	0.12	0.18
40x1 [10,8]	1.16	1.27	1.21	1.07	1.28	1.21		0.21	
	1.15	1.10	1.06	1.20	1.24	1.15		0.18	
	1.12	1.18	1.22	1.21	1.13	1.18	1.18	0.09	0.16

* The numbers in brackets are the starting/ending radii of node neighborhood. Initial/ terminal learning rates are set to 0.25/0.001, number of cycles per phase is 50.

Table 3. Subsample Lack-of-Fit's of Transformed Simulated Data Set in Several SOM's

	Subsample Lack-of-Fit's								
SOM	1	2	3	4	5	Median	Average Median	Range	Average Range
10x4 [3,1]	2.02	2.28	2.25	2.34	1.87	2.25		0.47	
	1.66	2.28	2.54	2.23	2.65	2.28		0.99	
	1.71	2.04	1.72	2.20	2.40	2.04	2.19	0.69	0.72
20x2 [5,3]	1.67	1.63	1.85	1.77	1.68	1.68		0.22	
	1.60	2.10	1.86	1.58	1.70	1.70		0.52	
	1.56	1.54	1.84	1.73	1.55	1.56	1.65	0.30	0.35
40x1 [10,8]	1.63	2.06	1.88	1.61	1.73	1.73		0.45	
	1.71	1.80	1.71	1.68	1.66	1.71		0.14	
	1.58	1.55	1.78	1.68	1.63	1.63	1.69	0.23	0.27

* The numbers in brackets are the starting/ending radii of node neighborhood. Initial/ terminal learning rates are set to 0.25/0.001, number of cycles per phase is 50.

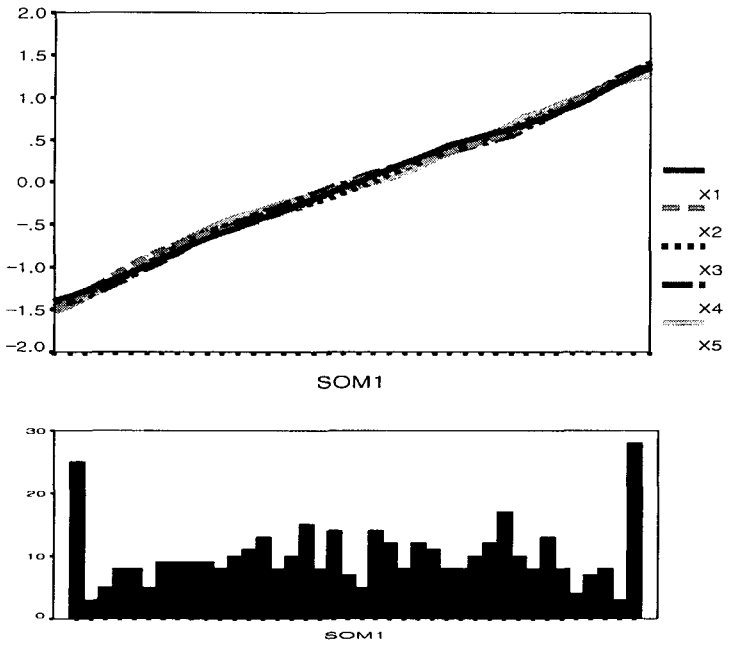

Figure 3. Node Values (Above) and Counts (Below) for Simulated Data Set

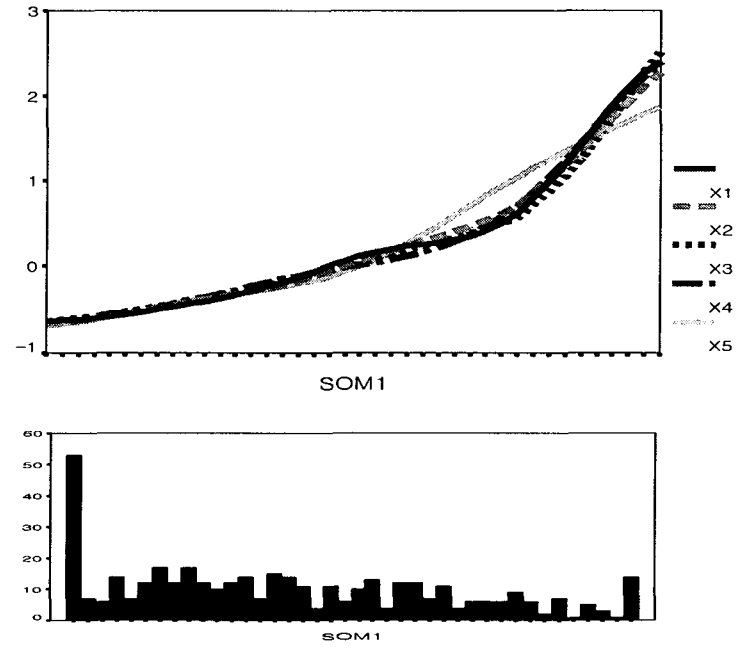

Figure 4. Nodes Values (Above) and Counts (Below) for Transformed Simulated Data Set

4. Concluding Remarks

Self-organizing map is computationally intensive compared to K-means clustering. Ten times or more is quite common. Hence, comprehensive searches for the choice of the best map size are not recommendable in practice. Instead, one may compare two or three shapes of SOM with approximately same number of nodes, and, subsequently, try two or three similar shape maps with different number of nodes. In doing so, one should keep in mind that SOM is somewhat sensitive on the starting/ending parameters of the node neighborhood.

In this short communication, we intended to provide a method for determining the size of SOM, which is one of fundamental issues in this research area (Koikkalainen, 1999). It is quite certain, however, that the lack-of-fit is not all things in the choice of the best map. For instance, interpretability and visualization should be considered simultaneously, which are not covered in this article.

References

[1] de Bodt, E., Cottrell, M. and Verleysen, M. (2002). "Statistical tools to assess the reliability of self-organizing maps," *Neural Networks*, 15, 967-978.

[2] Gordon, A.D. (1999). *Classification* (Second Edition). Chapman & Hall, London. (Chapter 7).

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

[4] Huh, M.H. (2003). "Principal Components Self-Organizing Map PC-SOM", To appear in *Korean Journal of Applied Statistics*. 16. (Written in Korean).

[5] Jun, S.H., Jorn, H., and Hwang, J. (2002) "Bayesian Learning for Self Organizing Maps", *Korean Journal of Applied Statistics*, 15. 252-267 (Written in Korean).

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

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

[8] Koikkalainen, P. (1999). "Tree structured self-organizing maps," in *Kohonen Maps* (Edited by E. Oja and Kaski, S.). Elsevier Science, B.V. 121-129.

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

[Received January 2003, Accepted May 2003]