2008 International Symposium on Nonlinear Theory and its Applications
NOLTA'08, Budapest, Hungary, September 7-10, 2008

# GHSOM with Ranking Mapping Scheme

Mitsushi Yoshida*, Masatoshi Sato*, Daisuke Shima*, Hisashi Aomori* and Mamoru Tanaka*

*Department of Electrical and Electronics Engineering, Sophia University,
7-1, Kioi-cho, Chiyoda-ku, Tokyo, 102-8554, Japan
Phone:+81-3238-3878, Fax:+81-3-3238-3321
Email: mitsus-y@hoffman.cc.sophia.ac.jp

**Abstract**—A self-organizing indicates the system producing an own structure. Especially, the map system is called the self-organizing map (SOM). The SOM can map to the low dimension by which the adjacency relation of the multidimensional data is maintained in nonlinearly. This method has been focused on because of the effectiveness for clustering, information compression, and visualization and so on. And, the growing hierarchical self-organizing map (GHSOM) is efficient way to project input data onto output map using hierarchical structure in learning stage . However, most of the SOM and GHSOM projection methods are computationally expensive when the size of the data set becomes large. In this paper we present an intuitive and effective GHSOM projection method with comparatively low computational complexity for the purpose of cluster visualization. This method is called ranking mapping scheme (RMS). This method maps data vectors on the output space based on their responses to different prototype vectors. High-resolution maps can be obtained with a relatively small network size. The effectiveness of proposed method will be demonstrated using iris data set.

## 1. Introduction

Recently, there are huge amount of information of electronic data due to the development of information processing technology. However, there is a limitation in the amount of manually treatable information, and it is difficult to get information and knowledge from such a large amount of data. A data mining is technique for getting profitable information from among data. Also, it is usually used for businesses, intelligence organizations, and financial analysts, but it is increasingly used in the sciences to extract information from the enormous data sets generated by modern experimental and observational methods.

The self-organizing map (SOM)[1] was first introduced by the Teuvo Kohonen. It creates prototype vectors which have high dimensional value and make them represent the same dimensional input data by learning process considering Euclidean distances between input data and prototype vectors. Since each prototype vectors have a low dimensional output space grid, the SOM can visualize high-dimensional data into a low-dimensional spatial grid. This dimensionality reducing mapping of the SOM makes the inter relation among the data points and clustering tendency perceptible. Growing hierarchical self-organizing map (GHSOM)[2] [3] is improvement type of SOM to aim at the solution of the problem of SOM by dynamically enhancing the map size and the layered structure according to input data. However, the original projection by the GHSOM alone represent the training results is very crude. Data vectors are mapped to the locations of corrsponding to best-matching neurons. It is usually difficult to provide much information about the global distribution of the data by observing the resulting row map. For visualizing the data structure, the SOM usually requires assistance from a separate vector projection of the prototype vectors.

In this paper we present an intuitive and effective GHSOM projection method with comparatively low computational complexity for the purpose of cluster visualization. Although RMS was taken to SOM before[4], we introduced to GHSOM this method in this paper. The purpose is to carry data mining with good accuracy.

## 2. Visualizing Algorithms based on SOM technology

### 2.1. SOM

The SOM is usually consisted of two dimentional array of neurons as shown in Fig. 1. A prototype vector associated with each neuron is described by

$$\omega_i = [\omega_1, \omega_2, \cdots \omega_n]^T, \tag{1}$$

where $n$ is the dimension of the input vectors. At each step, input vector $x$ is drawn randomly and is presented to the network. This input vector is compared with all the prototype vectors. The nearest prototype vector is called a best matching unit (BMU).

A grid number of BMU $c$ obtained by the Euclidean distance between the input vector $x$ and weight of prototype vector $\omega_i$, is expressed by

$$c = \arg\min \parallel x - \omega_i \parallel . \tag{2}$$

The neighborhood size function with time decay property is defined to decide the range of learning units. One example of neighborhood size function $\sigma(t)$ is given by

$$\sigma(t) = d_o(1 - t/T), \tag{3}$$

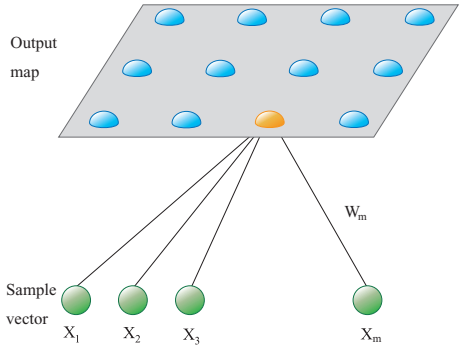

Figure 1: Concept of the SOM

where $d_o$ is a starting width of neighborhoods, $t$ is current time step, and $T$ is total learning times, respectively. Then the SOM updates the prototype vectors within the neighborhoods. The prototype vector $\omega_i$ is updated by

$$\omega_i(t+1) = \omega_i \cdot MQE_{i-1}, \tag{4}$$

where $MQE$ is mean quantization error and $MQE_{i-1}$ is the mean deviation of the parents unit. A typical smooth neighborhood function is the Gaussian function described by

$$h_{ci}(t) = \alpha(t) \exp\left(\frac{- \parallel r_c - r_i \parallel^2}{2\sigma(t)^2}\right), \tag{5}$$

where $h_{ci}$ is the time decreasing learning function, $\alpha(t)$ is the learning rate function, $\parallel r_c - r_i \parallel$ is the distance between the winner neuron $c$ and the neuron $i$. The learning processes consist of winner selection by Eq. (1) and adaptation of the prototype vectors by Eq. (3).

### 2.2. GHSOM

The growth judgment of the map is done by comparing the mean deviation for all units that exist in the map that grows up with the mean deviation of the parents unit. The end condition is shown by following equation.

$$MQE_i < \tau_1 \cdot MQE_{i-1}. \tag{6}$$

where $MQE_i$ is the mean deviation of one unit of pertinent map and $\tau_1$ is the threshold to decide degree of growth.

When all units do not satisfy Eq. (6), a unit is inserted. The new unit is inserted between Error unit $e$ and Dissimilar unit $d$. The largest mean deviation unit $e$ is given by

$$e = \arg\max_i \left( \sum_{x_j \in C_i} \|m_i - x_j\| \right), \qquad n_C = |C_i|, C_i \neq \emptyset. \tag{7}$$

where $n_C$ is the number of input data, and $C_i$ is the number of parents unit of each map. The longest distance with $d$ neighborhoods of $e$ is given by

$$d = \arg\max_i \left( \|m_e - m_i\| \right), \qquad m_i \in N_e. \tag{8}$$

After the map grows up, the hierarchization judgment is done by

$$MQE_i < \tau_2 \cdot MQE_0, \tag{9}$$

where $\tau_2$ is the threshold to adjust hierarchy level.

After the training has been completed, the weight of map should be reculculated, so that similar data items are mapped onto nearby map units. This process is very important for making the map into easily understandable data structure.

## 3. Generation of multi-dimensional lattice data and neighborhood uniting in input space

The problem of conventional GHSOM is over compressed blank space. Therefore, an actual distance relationship in the input space is not expressed in the output map. To solve the above problem, we introduced the multi-dimensional lattice data addition learning model by which the concept of the neighborhood uniting is introduced to the study of the conventional GHSOM.

The introduced method is a model to add not only to input data but also to lattice data of the same dimension as input space and to study them. This method is classified into three operations for the learning of SOM, the generation of multi-dimensional lattice data and the calculation of neighborhood uniting in input space.

For $m$ dimension input data, the multi-dimensional lattice points of $n-1$ capitation side is considered. The number of the points is given by

$$n^m \, (= K(n, m)) \, . \tag{10}$$

Although, solutions (10) increases in exponential when the dimension is increased. Therefore, when number of dimension increases, multi-dimensional lattice point increases remarkably in input data, and it becomes as a map of a lot of blank space in the output space.

Hence, so as not to consider the needless multi-dimensional lattice data, the neighborhood uniting with input data is considered to the generated multi-dimensional lattice data by the input space. Discriminant is applied to each multi-dimensional lattice data. Discriminant $D$ is defined as

$$D = \begin{cases} 1 & if\ d < d_0(1 - \frac{t}{t_{max}}) \\ 0 & otherwise \end{cases}, \tag{11}$$

where the $d$ is the distance between a multi-dimensional point data and the input point, $t_{max}$ is the total leaning number, $t$ is the present leaning number and $d_0$ is the constant. Equation (11) is applied to each multi- dimensional lattice data. When the value of $D$ is one, the SOM learning is completed.

## 4. Additional method ranking mapping scheme

A common disadvantage of the previous SOM visualization is that the map resolution depends on the size of the map. In the previous SOM, as explained above, input vectors are projected only on the neurons which have nearest value. To obtain good visualization results, the SOM should have higher number of neurons than that of data vectors. Therefore, to get a high-quality result, large number of prototype vectors is required. That is, learning process becomes impractically time-consuming.

In this paper, in order to reduce the computation cost, the proposed projection method is based on the standard SOM structure and learning procedure. In visualizing process, we consider not only closest grid, that is BMU, but also the other grids. The response of a data sample to a prototype infers its closeness, which is in turn its membership degree, to a specific unit. Thus a data sample has the highest membership degree to the prototype vector associated with the BMU and it should be mapped to a position closer to the BMU than to other units. There are usually several units with almost as good mach as the BMU. Consequently, projecting sample data only on the BMU does not provide accurate information of cluster membership. Intuitively the data item should be projected to somewhere in between the map units with a good match. Analogously, each map unit exerts an attractive force on the data item proportional to its response to that data item. The greater force, the closer data item attracted to the map unit. The data item will end up in a position where these forces reach equilibrium.

In primary method, the GHSOM projection procedure continues with directly finding the centroid of this spatial response, where the data sample is then mapped. In order to enhance the visual representation, a ranking scheme is used to visualize different degree of cluster membership. First, it is required to decide the number of units taken into the account. We set this parameter as $R$. After that, put order label on each units considering with a distance to sample data which is given by:

0 for the closest unit.

1 for the second closest unit.

$R$ for $R$-th closest unit.

Then, the coordinate $\mathbf{P} = (x_1, x_2)^T$ of output map is obtained by

$$\mathbf{P} = \frac{\sum_{i=0}^{R-1}(\sum_{j=0}^{R-1} d_j - d_i)\mathbf{W_i}}{\sum_{i=0}^{R-1} \frac{\prod_{i=0}^{R-1} d_j}{d_i}}, \tag{12}$$

where $d_i$ is Euclidean distance between input vector and weight $\mathbf{W}_i = (y_1, y_2)^T$ is coordinate of the $i$-th ranked unit, respectively. Continue to calculate the equation above for all the sample data. Then sample data is mapped on coordinate $\mathbf{P}$ of the output map. Then SOM processes are summarized as follows:

1. Initializing prototype vectors.
2. Calculate Euclidean distances between prototype vectors and input data.
3. Modify prototype vectors.
4. Set the parameter $R$.
5. Decide the coordinate $\mathbf{P}$ of output map by (12).
6. Project the sample data on $\mathbf{P}$.

In above flow, 2 and 3 repeat for certain times that is defined by user.

## 5. Simulation

In order to demonstrate the efficiency of proposed method, we present the following experiments using Iris Plants data set. The Iris data set is a widely used benchmark for pattern recognition. It contains three classes; Iris-setosa, Iris-versicolor and Iris-virginica. In Fig. 6, 7, 8, 9, Iris-setosa, Iris-versicolor and Iris-virginica are respectively represented by "red", "blue" and "green".

The error rate is required by comparing the distance within the cluster center of gravity in input space and output space.

Results simulation results by proposed method are shown in the Fig. 6, 7, 8, 9 and Table 1.

Different $R$ values result in different maps. Notice for $R$=1, where only the BMU is concerned in the projection. Because it can only project input items to map units on a rigid grid, this map does not provide much information about the global shape of the data. With $R$ getting larger, the structure and shape of the data become more prominent. When $R$=3, the resulting map become to show some clusters. Also when $R$ comes to 4, a boudary of the clusters become more visualizable. Even though four simulations are performed by using same number of prototype vector and weight, the results are getting better according to the increase of parameter $R$. In other words, to get same quality output map, processing time become shorter with increase of $R$. Moreover, the error rate has decreased as shown in Table 1.

## 6. Conclusion

In this paper, we have presented a new approach to visualization technique for the GHSOM. The proposed method is simple but effective as shown in result map. Unlike the conventional GHSOM projection method, which restrict the projection to the junction of the map grid, the proposed method maps the data samples to arbitrary positions across

Table 1: Distance ratio between cluster center of gravity

| | 1 and 2 | 2 and 3 | 3 and 1 | Error(%) |
|---|---|---|---|---|
| Theoretical | 1.000 | 0.580 | 1.545 | — |
| Conventional | 1.000 | 1.344 | 2.344 | 67.000 |
| Proposed | 1.000 | 0.367 | 1.367 | 16.800 |

the SOM grid. This enables a high-resolution out put map with a comparatively small number of map units. Thus, the computational complexity is greatly reduced. Moreover, the input data relationship in the obtained map is undersatandable.

The implementation of the proposed method is illustrated using real world high-dimensional data set. The results show the visualization technique has good potential as a tool for structure analysis encountered in high-dimensional input data.

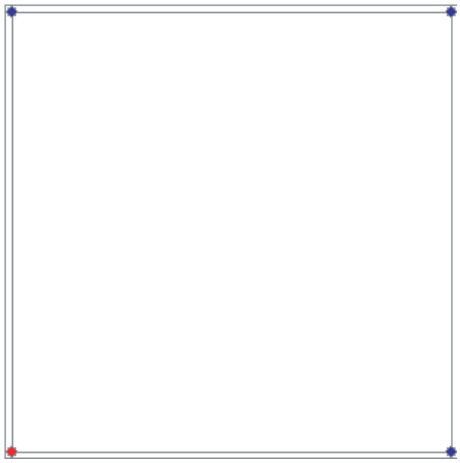

Figure 6: Output by R=1

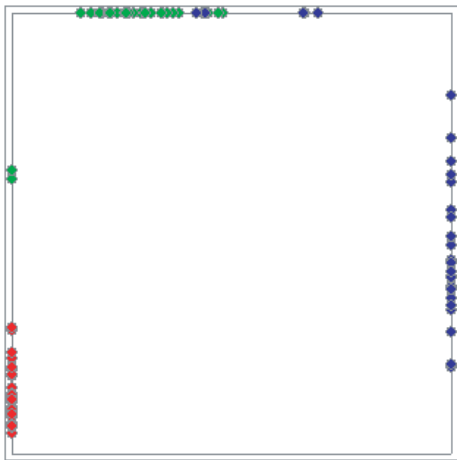

Figure 7: Output by R=2

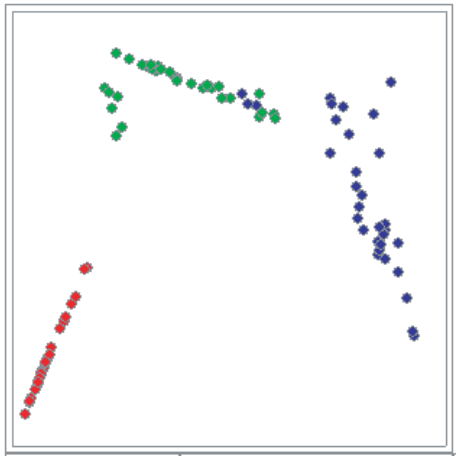

Figure 8: Output by R=3

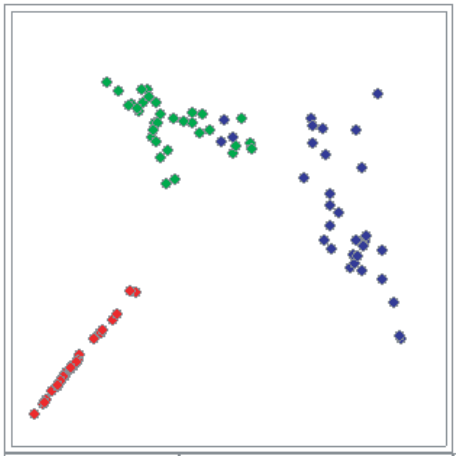

Figure 9: Output by R=4


**Acknowledgment**

This research is supported by a fund of the Open Research Center Project from MEXT of the Japanese Government (2007-2011).


**References**


[1] T. Kohonen, "*Self-organizing Maps*. Berlin, Germany." Springer, 1995, vol 30.

[2] A. Rauber, D.Merkl and M.Dittenbach, "The Growing Hierarchical Self-Organizing Map:Exploratory Analysis of High-Dimensional Data" IEEE Trans. Neural Networks, vol.13, No.6, pp.1331-1341, Nov. 2002.

[3] Z. Wu and G. Yen, "A SOM Projection Teqhnique with the Growing Structure for Visualizing High-dimensional Data." *International Journal of Neural Systems*, Vol.13, No. 5, 2003.

[4] D. Shima, K. Kurokawa, M. Yamauchi and M.Tanaka, "Efficient Implementation of the Self-Organizing Maps." , .

[5] S. Morris, C. Deyong, Z. Wu, S.Salman, and D, Yemenu, "DIVA: a Visualization System for Exploring Document Databases for Technology Forecasting," *Computer and Industrial Engineering*, vol.43, pp.841-862, 2002.