2007 International Symposium on Nonlinear Theory and its Applications
NOLTA'07, Vancouver, Canada, September 16-19, 2007

Self-Organizing Mapping that considers neighborhood uniting

Mitsushi Yoshida*, Daisuke Shima*, Kaname Kurokawa*, 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—Self-organizing indicates the system producing an own structure. Especially, the map system is called the self-organizing map (SOM). 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. On the other hand, since the SOM tends to compress the distance between data, the mapped data does not guarantee the actual distance relationship of the input space. Therefore, the problem is that an actual distance relationship in the input space is not expressed in the output space. In this paper, to solve the above problem, we propose the multidimensional lattice data addition learning model by which the concept of the neighborhood uniting is introduced to the study of the conventional self-organizing map.

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. Data mining is technique for getting profitable information from among data. Data mining is defined as ”the nontrivial extraction of implicit, previously unknown, and potentially useful information from data”. It is usually used by businesses, intelligence organizations, and financial analysts, but is increasingly used in the sciences to extract information from the enormous data sets generated by modern experimental and observational methods.

The SOM was first introduced by the Teuvo Kohonen[1]. The SOM 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. But, the problem of conventional SOM is compressing the blank space remarkably. Therefore, an actual distance relationship in the input space is not expressed in the mapped map.

In this paper, to solve the above problem, we propose the multi-dimensional lattice data addition learning model by which the concept of the neighborhood uniting is introduced to the study of the conventional SOM. The purpose is to carry data mining with good accuracy.

2. Visualizing Algorithm

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

$$\omega_i = [\omega_{1i}, \omega_{i2}, \cdots \omega_{in}]^T, \tag{1}$$

where m 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).

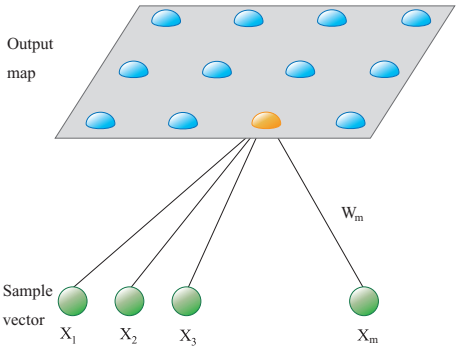

Figure 1: Concept of the SOM

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

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

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

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

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 ω_i is updated by

$$\omega_i(t+1) = \omega_i(t) + h_{ci}(t)[x(t) - \omega_i(t)], \tag{4}$$

where h_{ci} is the time decreasing learning function. A typical smooth neighborhood function is the Gaussian function described by,

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

where $\alpha(t)$ 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 equation (1) and adaptation of the prototype vectors by equation (3). After the training has been completed, the map should be topologically ordered, so that similar data items are mapped onto nearby map units. Then, the visualizing process must be carried out in order for the underlying structure of data to be perceived.

3. Mapping

The mapping is worked to decide the unit of the output space to each input data after the learning algorithm ends. Although the more the number of units increases, the more feature of input data can be mapped in high probability, while computation time is increased. Hence, to solve this problem, we used the method of ranking scheme.

In primary method, the SOM 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 Rth closest unit.

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

$$\mathbf{P} = \frac{\sum_{i=0}^{R-1}(\sum_{j=0}^{R-1} d_j - d_i)\mathbf{W_i}}{\sum_{i=0}^{R-1} \frac{\prod_{i=0}^{R-1} d_j}{d_i}}, \tag{6}$$

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.

4. Additional learning

4.1. Introduction of additional learning

The problem of conventional SOM is compressing the blank space remarkably. Therefore, an actual distance relationship in the input space is not expressed in the mapped map. In this paper, to solve the above problem, we propose the multi-dimensional lattice data addition learning model by which the concept of the neighborhood uniting is introduced to the study of the conventional SOM.

Image of problem of conventional method is shown in Fig. 2. The left figure shows the data distribution in the input space and the right shows the data distribution after maps. 1 and 2 represent the sample of the cluster and there are two clusters in those figures.

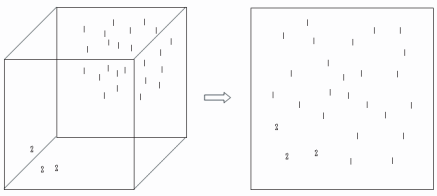

Figure 2: Image of output of conventional method

As shown in Fig. 2, it is understood that the blank space is compressed, and the distance between clusters of the input space is not reflected. Our purpose is to improve Fig. 2 to that of Fig. 3.

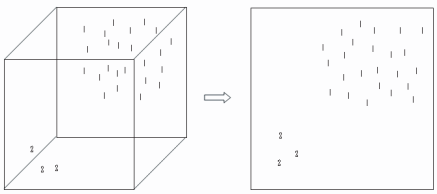

Figure 3: Image of output of proposed method

The proposed 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 3 operations for the learning of SOM, the generation of multi-dimensional lattice data and the calculation of neighborhood uniting in input space.

4.2. Generation of multi-dimensional lattice data and 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

$$n^m \,(= K(n,m)) \,. \tag{7}$$

Fig. 4 is an example of $K(3.3)$.

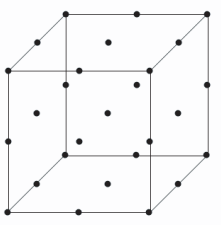

Figure 4: K(3,3)

Although, the answer of equation (7) 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

$$D = \begin{cases} 1 & if\ d < d_0(1 - \frac{t}{t_{max}}) \\ 0 & otherwise \end{cases} \,, \tag{8}$$

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 (8) is applied to each multi- dimensional lattice data. When the value of D is one, the SOM learning is completed. The image of the time variation of the neighborhood uniting is shown in Fig. 5.

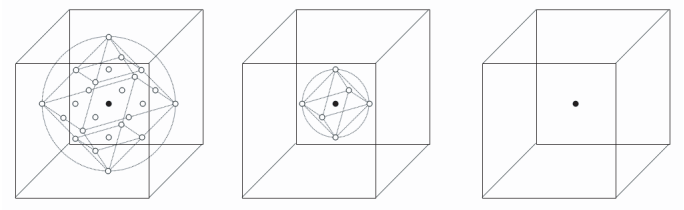

Figure 5: Image of neighborhood uniting

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 and fig. 7, Iris-setosa, Iris-versicolor and Iris-virginica are respectively represented by ”1”, ”2” and ”3”.

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

In the first experiment, the number of input data is 100 and the number of lattice data is 81. The effectiveness of the multi-dimensional lattice data was confirmed under these conditions. These results are shown in the Fig. 6, Fig. 7 and Table 1. We confirmed that the cluster has had clearly divided by comparing Fig. 6 to Fig. 7. Moreover, the error rate has decreased as shown in Table 1.

Next, we confirmed that the neighborhood uniting in the input space was effective when the number of parti-

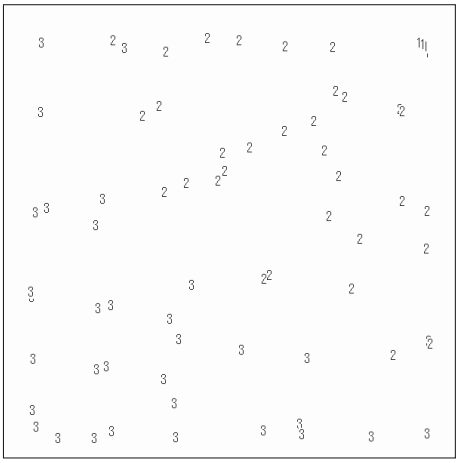

Figure 6: Output by conventional method

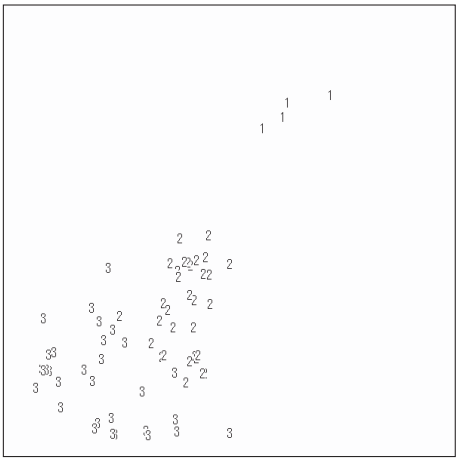

Figure 7: Output by proposed method

tions of the multi-dimensional lattice data was increased. n was changed between 2 to 6 for Iris data set with four-dimensional variable, and error rate were compared by the presence of the neighborhood uniting. These results are shown in the Fig. 8, Fig. 9 and Table 2. These graphs are result of each output at $K(5,4)$. It can be confirmed that the extra blank space has disappeared, and the error rate has decreased by comparing Fig. 8 to Fig. 9. The error rate has decreased as the number of lattice was increased as shown in Table 2. On the whole, the error rate is few when neighborhood uniting is done. Moreover, computing time can be shortened.

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

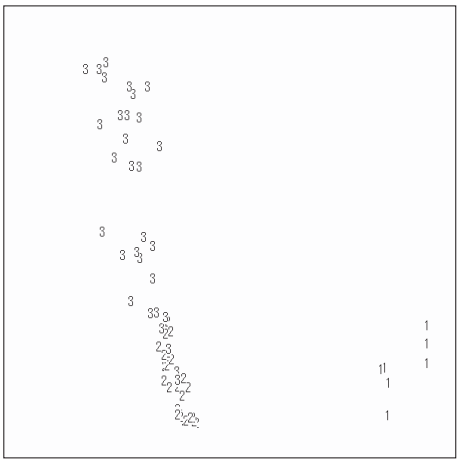

Figure 8: Output using un-neighborhood uniting

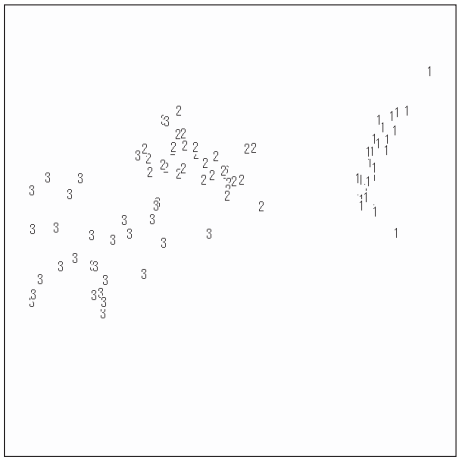

Figure 9: Output using neighborhood uniting

Table 2: Distance ratio between cluster center of gravity

	Number of lattice points	1 and 2	2 and 3	3 and 1	Error(%)
Theoretical figure	—	1.000	0.584	1.538	—
un-neighborhood uniting	K(2,4)	1.000	0.979	1.829	29.80
	K(3,4)	1.000	0.535	1.451	6.10
	K(4,4)	1.000	0.595	2.344	31.40
	K(5,4)	1.000	0.789	2.344	18.00
	K(6,4)	1.000	1.369	2.344	48.30
neighborhood uniting	K(2,4)	1.000	1.026	1.731	29.30
	K(3,4)	1.000	0.813	1.545	13.90
	K(4,4)	1.000	0.817	1.411	16.10
	K(5,4)	1.000	0.699	1.623	8.70
	K(6,4)	1.000	0.658	1.553	4.60

6. Conclusion

In this paper, we proposed multi-dimensional lattice data addition learning model. This proposed model is introduction of the concept of the neighborhood uniting with input data for the study of the conventional SOM. The problem is that the actual distance relationship in the input space is not expressed in the output space was solved by this method. In addtion, mapping the distance relationship of accuracy good input data even if remarkably the amount of multi-dimensional lattice data existed for input data became possible by introduced the neighborhood uniting.

References

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

[2] J. Vesanto, "SOM-Based Data Visualization Methods." *Intelligent Data Analysis*, Vol.3, No. 2,pp.111-126, 1999.

[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] 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.

[5] J. W. Sommon, "A Nonlinear Mapping for Data Structure Analysis," *IEEE Traqnsactions on Computers*, vol.18, pp.401-409, 1969.

[6] J. Vesanto,and E, Alhoniemo "Clustering of Data Structure Analysis," *IEEE Traqnsactions on Neural Networks*, Vol.11, pp.586-600, 2000.