

Feature Discovery by Information Loss

Ryotaro Kamimura

IT Education Center, Tokai University, Kanagawa, Japan

Email: ryo@cc.u-tokai.ac.jp

Abstract—In this paper, we propose a new approach called *information loss* to feature detection in competitive learning. The information loss is defined by the difference between a full network and a network without some elements. If this deletion significantly decreases the amount of information contained in a network, the elements are considered to be important and are expected to play a very important role. The method was applied to artificial and symmetric data to show the features extracted by the information loss. Then, we applied the method to the classification of OECD countries. The experimental results confirmed that the method was efficient enough to detect main features comparable to those detected by the conventional SOM.

Index Terms—mutual information, information loss, feature detection, competitive learning, self-organizing maps

I. INTRODUCTION

There have been many attempts to apply information-theoretic methods to theoretic and practical problems in neural networks [1], [2], [3], [4], [5], [6], [7]. Though they have been successful in dealing with overall information processing in neural networks, little attention has been paid to information content in specific elements in a network, because there have been no methods to compute the information content for some elements in a network.

We have introduced mutual information as a measure of structure in neural networks [8], [9], [10]. It is important to see the change in this information, depending upon some parts in a network. For information content about specific parts of a network, we have introduced an information loss that explains the importance of given elements. The information loss is defined by the difference between information content with all elements and without some elements. If this deletion significantly decreases information content in a network, the element surely plays an important role in information processing of input patterns. Thus, the information loss indicates to what extent an element plays an important role in networks.

The information loss in this paper is an extended version of the previous one [11]. In the previous paper, we focused upon the importance of input units and tried to formulate the information loss only for input units. In this paper, the information loss is generalized to cover all elements in a network. Thus, the information loss can be defined for any element or any groups of elements. The

information loss now becomes more general and more flexible than the previous one.

In Section 2, after explaining information-theoretic competitive learning, we present how to compute the information loss. In Section 3, we present experimental results on two problems. In the first problem, we use artificial data to show intuitively the features extracted by the information loss. In the second example, the classification of OECD countries, we try to show that experimental results obtained by the information loss are comparable or superior in some cases to those obtained by the conventional computational methods, such as the U-matrix.

II. THEORY AND COMPUTATIONAL METHODS

A. Competitive Learning and Information Content

In this paper, we apply a new method, called *information loss*, to extract important features in competitive learning. First, it is necessary to explain the relations between competitive learning and mutual information.

Competitive learning has been a simple and powerful technique to extract features in input patterns. Though many methods have been developed to solve the fundamental problems of competitive learning, such as the dead neuron problem [12],[13],[14],[15],[16], much attention has been paid to classification performance. There have been few attempts to explain how competitive learning can classify input patterns, that is, to interpret internal representations by competitive learning. In actual data analyses, it is important to explain why and how classification by competitive learning is possible.

To interpret final representations obtained by competitive learning, we can consider competitive learning to be a process of information maximization in neural networks. In other words, competitive learning is only one aspect of information maximization in neural networks. We have so far proposed information-theoretic competitive learning [8], [9], [10], in which competitive processes are supposed to be equivalent to information maximization processes. When mutual information between input patterns and connection weights is maximized, only one competitive unit is active, while all the other units are inactive. We can consider the process of competition as one of information maximization. In addition, we have observed that the careful observation of information content reveals many characteristics in input patterns.

This paper is based on "Information loss to extract distinctive features in competitive learning," by R. Kamimura, which appeared in the Proceedings of IEEE Conference on Systems, Man, and Cybernetics, Montreal, Canada, 2007. © 2007 IEEE.

B. Information-Theoretic Formulation

We think that one of the major objectives of competitive learning is to store information on input patterns as much as possible. This means that the close examination of this information on input patterns reveals some important information on the features in input patterns.

Now, let us define information to be stored in a neural system. Information stored in a system is represented by a decrease in uncertainty [17]. Uncertainty decrease, that is, information I , is defined by

$$I = - \sum_{\forall j} p(j) \log p(j) + \sum_{\forall s} \sum_{\forall j} p(s)p(j | s) \log p(j | s), \quad (1)$$

where $p(j)$, $p(s)$ and $p(j|s)$ denote the probability of firing of the j th unit, the probability of the s th input pattern and the conditional probability of the j th unit, given the s th input pattern, respectively.

As shown in Figure 1, a network is composed of input units x_k^s and competitive units v_j^s . The j th competitive unit receives a net input from input units, and an output from the j th competitive unit can be computed by

$$v_j^s = \exp \left(- \frac{\sum_{k=1}^L (x_k^s - w_{jk})^2}{2\sigma^2} \right), \quad (2)$$

where σ represents the Gaussian width, L is the number of input units and w_{jk} denote connection weights from the k th input unit to the j th competitive unit. The output is increased as connection weights come closer to input patterns. To compute mutual information, we suppose that the normalized activity v_j^s represents a probability with which a neuron fires. The conditional probability $p(j | s)$ is computed by

$$p(j | s) = \frac{v_j^s}{\sum_{m=1}^M v_m^s}, \quad (3)$$

where M denotes the number of competitive units. Since input patterns are supposed to be given uniformly to networks, the probability of the j th competitive unit is computed by

$$p(j) = \frac{1}{S} \sum_{s=1}^S p(j | s), \quad (4)$$

where S is the number of input patterns. Information I is computed by

$$I = - \sum_{j=1}^M p(j) \log p(j) + \frac{1}{S} \sum_{s=1}^S \sum_{j=1}^M p(j | s) \log p(j | s). \quad (5)$$

As information becomes larger, specific pairs of input patterns and competitive units become strongly correlated. In addition, in maximizing mutual information, entropy $-\sum_{j=1}^M p(j) \log p(j)$ must be maximized. This means

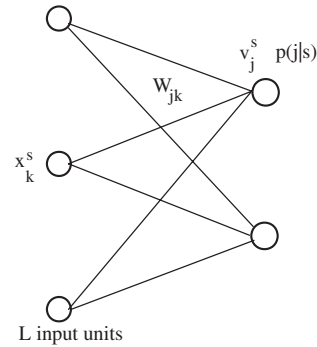


Fig. 1. A network architecture for competition.

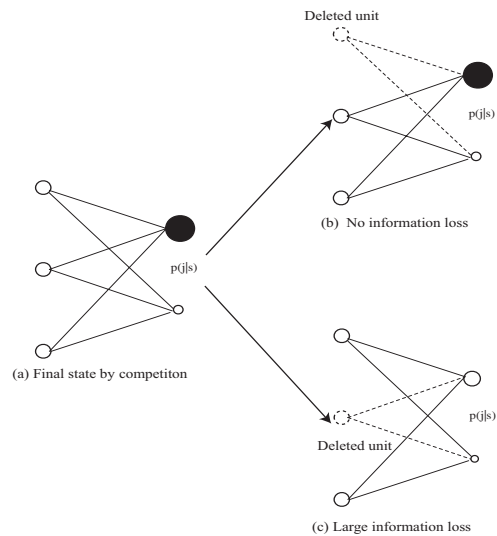


Fig. 2. A process to obtain information loss.

that all competitive units must equally be used on average. Thus, we can realize competitive processes by maximizing mutual information.

C. Information Loss

Sensitivity analysis [18], [19], [20], [21] has been well established in supervised learning, because one of the major problems of neural networks consists in the difficulty in interpreting final internal representations. However, there are few studies on unsupervised learning comparable to the sensitivity analysis in supervised learning, because it has been difficult to identify criteria comparable to those in the error terms between targets and outputs. In this context, we consider mutual information between input patterns and connection weights as one of the main criteria to identify the structure in competitive learning.

We have defined mutual information between input patterns and connection weights for competitive unit activations. Now, let us show how to compute the information loss for input units. Figure 2 presents an example to illustrate a process of information loss. In Figure 2(b), when the first input is deleted, no change in competitive unit activations can be seen, meaning that no change in information about competitive unit activation patterns can be seen. Thus, no information about input patterns is lost by deleting the first unit. Because the first input unit is

not so important, the unit can be deleted without drastic change for a network. However, when the second input unit is deleted, as seen in Figure 2(c), the first competitive unit becomes smaller and inactive, meaning that the second unit plays a very important role in information storage.

D. Information Loss for Input Units

We examine whether mutual information is changed by deleting some elements in competitive networks. Now, we focus upon the t th input unit and try to define information loss for the input unit. Distance when the t th input unit is deleted is defined by

$$d_{jt}^s = \sum_{k=1}^L \Phi_{kt}(x_k - w_{jk})^2, \quad (6)$$

where

$$\Phi_{kt} = \begin{cases} 1 - \epsilon, & \text{if } k = t; \\ \epsilon, & \text{otherwise.} \end{cases}$$

and where the parameter ϵ ranges between 0.5 and 1 for all experiments explained in the following sections. By using this equation, we have competitive unit activations for the t th input unit

$$v_{jt}^s = \exp\left(-\frac{d_{jt}^s}{2\sigma^2}\right). \quad (7)$$

We can normalize these activations, and we have

$$p^t(j | s) = \frac{v_{jt}^s}{\sum_{m=1}^M v_{mt}^s}. \quad (8)$$

The probability of the j th hidden unit is defined by

$$p^t(j) = \sum_{s=1}^S p(s)p^t(j | s). \quad (9)$$

By using these probabilities, we have mutual information when the k th input unit is deleted,

$$I_t = \sum_{s=1}^S \sum_{j=1}^M p(s)p^t(j | s) \log \frac{p^t(j | s)}{p^t(j)}. \quad (10)$$

We can compute the difference between mutual information or information loss by

$$IL_t = I - I_t. \quad (11)$$

For easy interpretation, we normalize this information loss and compute normalized information loss IL_k^{nrm} where $\sum_t IL_t^{nrm} = 1$ with all positive values, because mutual information varies greatly according to the Gaussian width and the parameter ϵ . In addition, there is a possibility of negative information loss.

E. Information Loss for Competitive Units

Then, we consider a case where a competitive unit should be deleted. Competitive unit activations when the r th unit is deleted are given by

$$V_{jr}^s = \Phi_{jr} v_j^s, \quad (12)$$

where

$$\Phi_{jr} = \begin{cases} 1 - \epsilon, & \text{if } r = j; \\ \epsilon, & \text{otherwise.} \end{cases}$$

and where $0.5 < \epsilon < 1$. By normalizing the activations, we have

$$p^r(j | s) = \frac{V_{jr}^s}{\sum_{m=1}^M V_{mr}^s}, \quad (13)$$

and

$$p^r(j) = \sum_{s=1}^S p(s)p^r(j | s). \quad (14)$$

By using these probabilities, we have mutual information for the r th competitive unit

$$I_r = \sum_{s=1}^S \sum_{j=1}^M p(s)p^r(j | s) \log \frac{p^r(j | s)}{p^r(j)}. \quad (15)$$

Information loss is defined by the difference

$$IL_r = I - I_r. \quad (16)$$

We also normalize this information, and normalized information loss IL_r^{nrm} is used in experiments for easy interpretation or comparison.

F. Information Loss for Input Patterns

Then, we consider a case where an input pattern should be deleted. Competitive unit activations when the q th input pattern is deleted are given by

$$d_j^{sq} = \Phi^{sq} \sum_{k=1}^L (x_k - w_{jk})^2, \quad (17)$$

where

$$\Phi^{sq} = \begin{cases} 1 - \epsilon, & \text{if } q = s; \\ \epsilon, & \text{otherwise.} \end{cases}$$

and where $0.5 < \epsilon < 1$. By using this equation, we have competitive unit activations for the r th competitive unit

$$v_j^{sq} = \exp\left(-\frac{d_j^{sq}}{2\sigma^2}\right). \quad (18)$$

By normalizing the activations, we have

$$p^q(j | s) = \frac{v_j^{sq}}{\sum_{m=1}^M v_m^{sq}}, \quad (19)$$

and

$$p^q(j) = \sum_{s=1}^S p(s)p^q(j | s). \quad (20)$$

By using these probabilities, we have mutual information for the r th competitive unit

$$I_q = \sum_{s=1}^S \sum_{j=1}^M p(s)p^q(j | s) \log \frac{p^q(j | s)}{p^q(j)}. \quad (21)$$

Information loss is defined by the difference

$$IL_q = I - I_q. \quad (22)$$

We also normalize this information, and normalized information loss IL_q^{norm} is used in experiments for easy interpretation or comparison.

G. Information Loss for Input and Competitive Units

We have competitive unit activations

$$V_{jrt}^s = \Phi_{jrt} v_{jt}^s. \quad (23)$$

By normalizing the activations, we have

$$p^{rt}(j | s) = \frac{V_{jrt}^s}{\sum_{m=1}^M V_{mrt}^s}, \quad (24)$$

and

$$p^{rt}(j) = \sum_{s=1}^S p(s)p^{rt}(j | s). \quad (25)$$

By using these probabilities, we have mutual information for the r th competitive unit

$$I_{rt} = \sum_{s=1}^S \sum_{j=1}^M p(s)p^{rt}(j | s) \log \frac{p^{rt}(j | s)}{p^{rt}(j)}. \quad (26)$$

Information loss is defined by the difference

$$IL_{rt} = I - I_{rt}. \quad (27)$$

We also normalize this information, and normalized information loss IL_{rt}^{norm} is used in experiments for easy interpretation or comparison.

III. RESULTS AND DISCUSSION

A. Artificial Data

We first applied the method to simple and symmetric artificial data, as shown in Figure 3(a). In the architecture, the number of input and competitive units was eight. Because the data was symmetric, we expected similar patterns of information loss for input units and competitive units.

Figure 4 shows three types of information loss when the Gaussian width is changed from 0.1 to 5, keeping another parameter, ϵ , a constant (0.6). Figure 4(a) shows three types of information loss when the Gaussian width is 0.1. Figure 4(a2) typically shows the characteristics of this state; that is, all units show almost equal information loss. When the Gaussian width σ is increased from 0.1 to 0.5, some elements show strong information loss. When the Gaussian width is increased further to 1, 3 and 5, as in Figure 4(c) to (e), the patterns of stable information loss can be seen. For example, moving from the corner to the center, the information loss is gradually decreased,

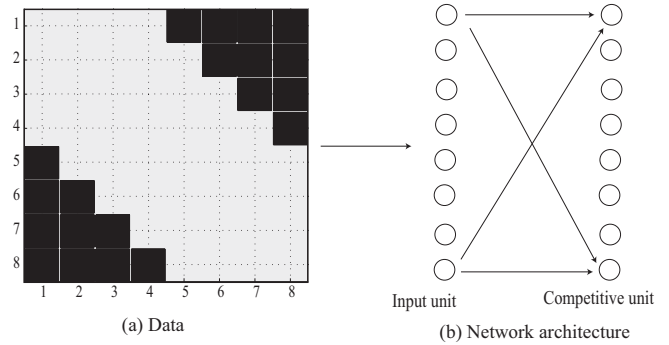


Fig. 3. Data (a) and a network architecture (b) for artificial data.

corresponding to the input pattern shown in Figure 3. Figure 5 shows three types of information loss when the parameter ϵ is increased from 0.7 to 0.99, keeping the Gaussian width at 5. We can immediately see that the final patterns of information loss remain stable independently of the parameter values. This means that final patterns are strongly influenced by the Gaussian width σ . Figure 6 shows information loss for input and competitive units. Figure 6(1) to (4) shows information loss for the first four input patterns. Though the first four competitive units show relatively stable information loss, the second four competitive units clearly show a decrease in the number of competitive units with high information loss. Figure 6(5) to (8) shows information loss for the fifth to the eighth input units. In this case, information loss on the right-hand side remains stable, while information loss on the left-hand side shows that the amount of strong information loss corresponds to input patterns.

B. OECD Data Description

We tried to classify 23 OECD countries by four variables, that is, the total fertility rate, the women's labor rate, the tertiary industry labor ratio and the gender development index [22]. Figure 7(a) shows the U-matrix obtained by the SOM¹. On the U-matrix, we can see a boundary in red or brown separating the countries into two parts.

Figure 8 shows three types of information loss when the Gaussian width σ is changed from 1 to 20, keeping the parameter ϵ at a constant (0.6). As the Gaussian width σ is gradually increased from one to 20, as shown in Figure 8(a1) to (e1), a more stable pattern of information loss appears, which is indicated in the middle of the map as a dark blue boundary. On the other hand, the other types of information loss, that is, information loss for input units and input patterns, are quite stable. For example, Figure 8(a2) to (e2) shows the information loss for input units. For any parameter values of the Gaussian width, the information loss for the third and the fourth input unit shows large values. This means that the countries are classified mainly based upon the women's labor rate and

¹We used SOM Toolbox 2.0, February 11th, 2000, by Juha Vesanto <http://www.cis.hut.fi/projects/somtoolbox/>. No special options were used for easy reproduction.

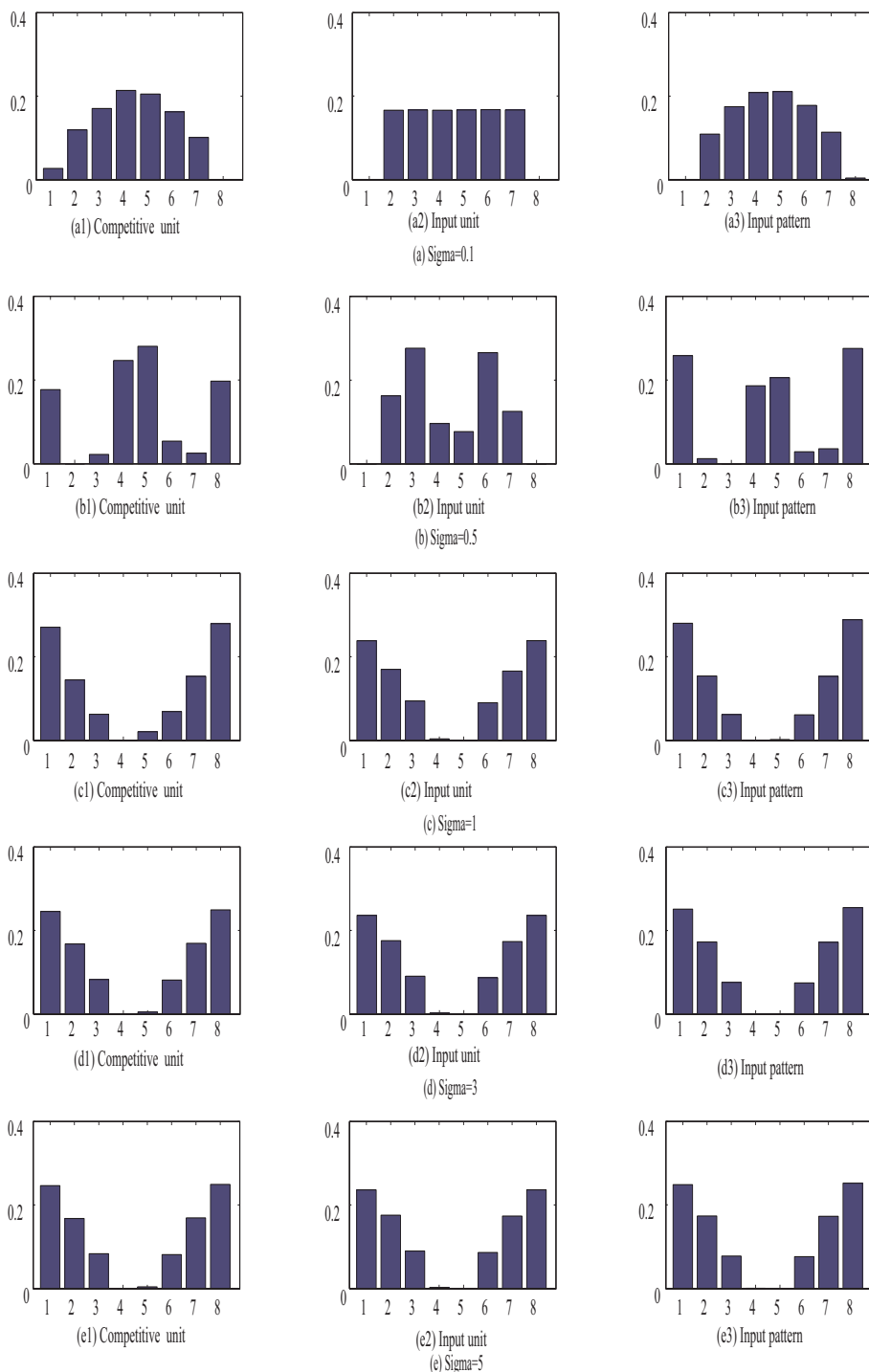


Fig. 4. Three types of information loss for an artificial data when the Gaussian width σ is changed from 0.1 to 5, keeping another parameter, ϵ , at 0.6.

the tertiary industry labor ratio of the four variables. For the information loss for input patterns, when the Gaussian width σ is below 5, rather random patterns are generated, as shown in Figure 8(a3) and (b3). When the Gaussian width is larger than 5, a stable pattern is generated. In addition, we can see a clear relation between this information loss and labels obtained by the conventional SOM. Countries No. 12 (Italy) and No. 9 (Greece) have the highest values of information loss, as shown in Figure 8(e3). As shown in Figure 7(b), Italy and Greece, with

the highest information loss, are located on the upper end of the map. When the information loss is smaller, the corresponding competitive units are located on or near a boundary. On the other hand, when the information loss is larger, the competitive units are located at the corners.

Then, we tried to change the information loss parameter ϵ from 0.6 to 0.99. As can be seen in Figure (a2) to (e2) and (c3) to (e3), we can see that information loss for input units and input patterns is very stable. On the other hand, we can see that the information loss for competitive units

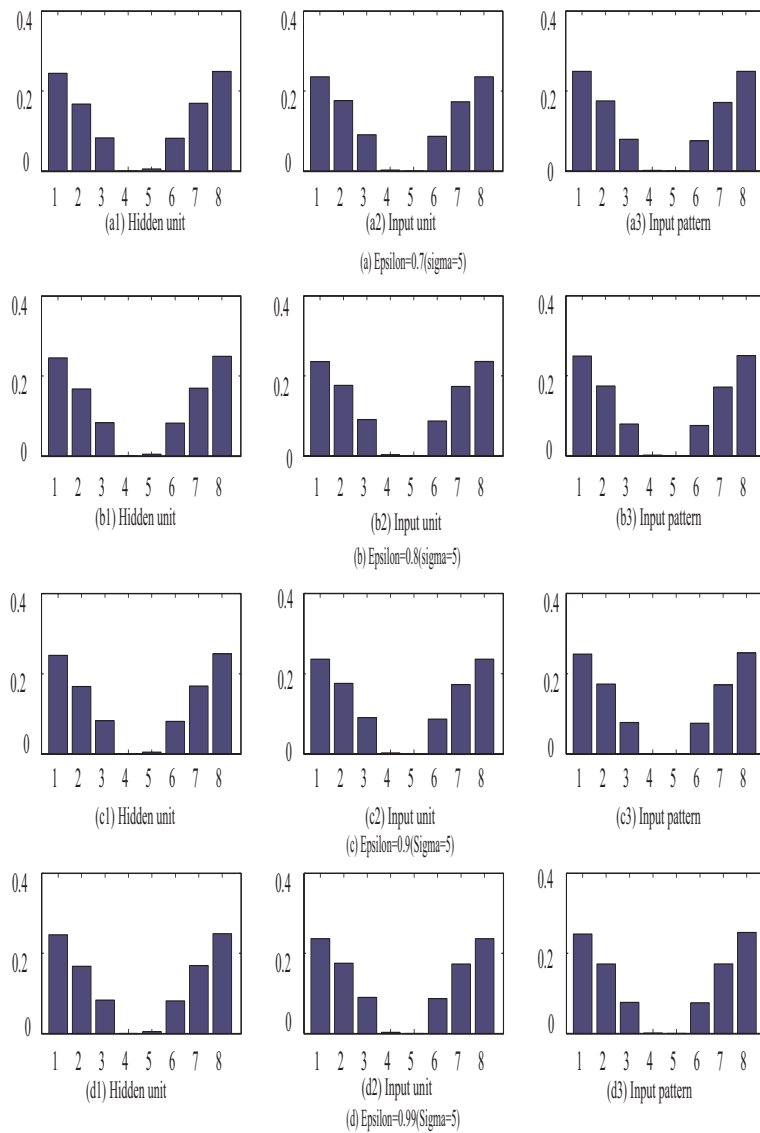


Fig. 5. Three types of information loss when the parameter σ is changed from 0.6 to 0.99.

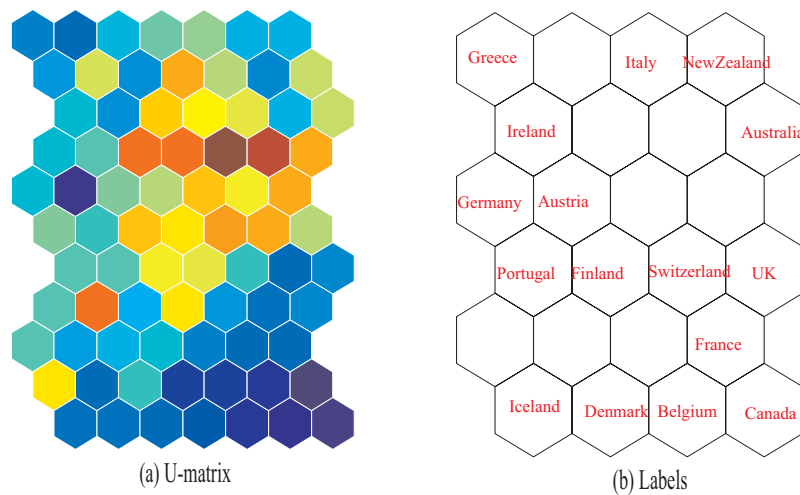


Fig. 7. U-matrix (a) and a map with labels (b). Warmer and cooler colors show larger and smaller values of the U-matrix.

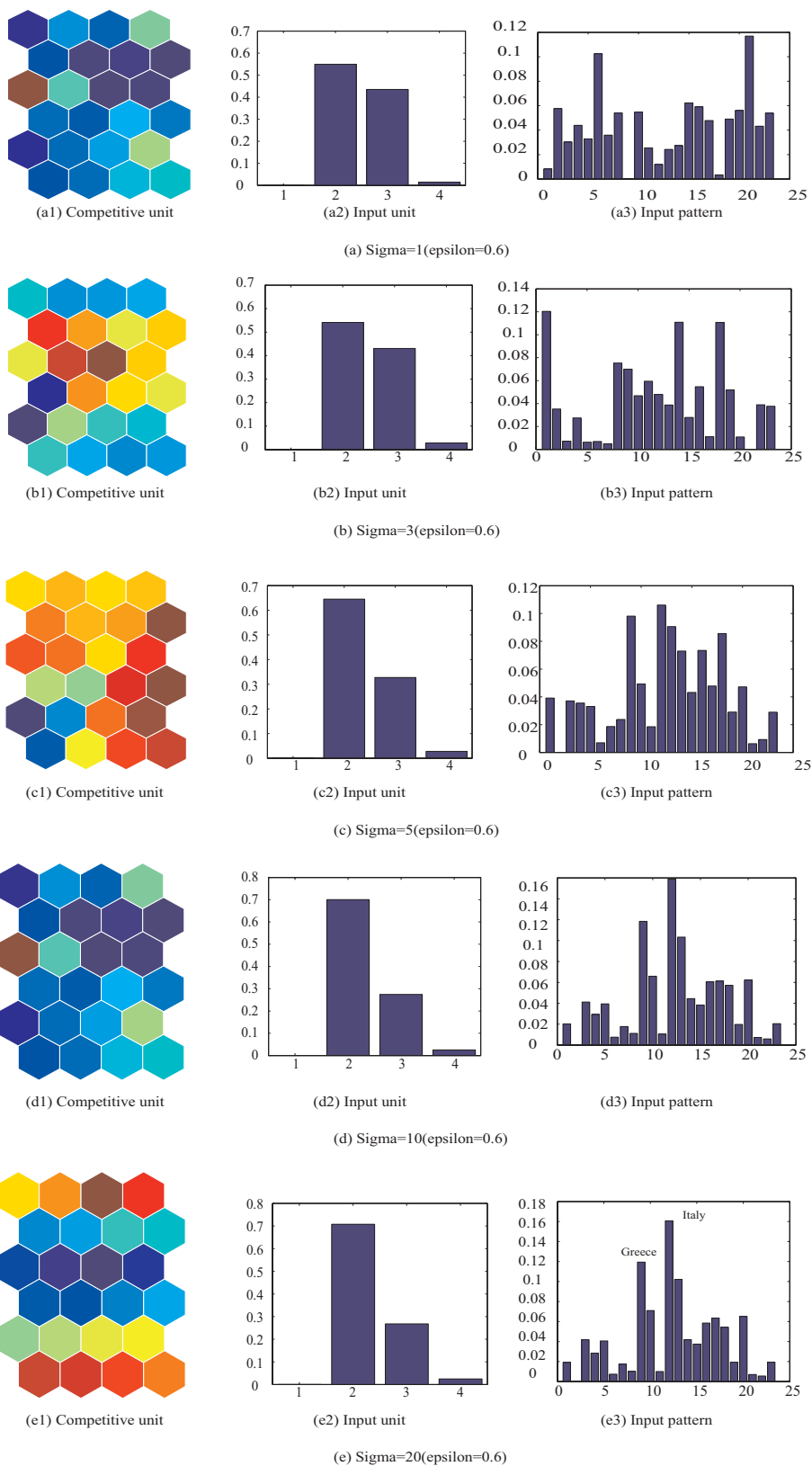


Fig. 8. Information loss for competitive units, input units and input patterns when the Gaussian width is increased from 1 to 20 ($\epsilon = 0.6$). Warmer and cooler colors show larger and smaller values of information loss.

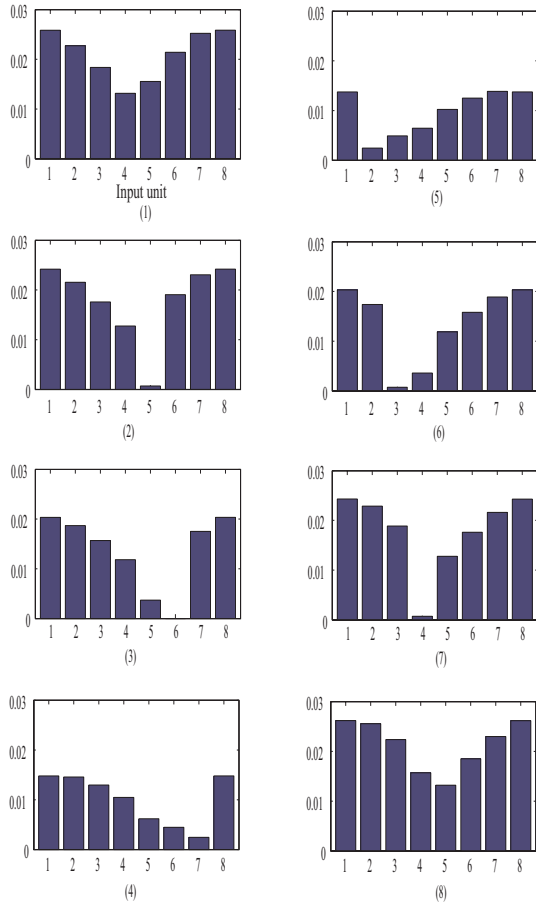


Fig. 6. Information loss for input-competitive units.

in Figure 9 (a1) to (e1) varies greatly. The information loss becomes clearer as the parameter ϵ is increased from 0.6 to 0.99 which means that the parameter ϵ should be as large as possible to obtain clearer feature maps for competitive units. Figure 10 shows information loss for competitive units when the Gaussian width σ is 10 and the parameter ϵ is 0.99. As can be seen in the figure, information loss for competitive units becomes clearer. The information loss for competitive units shows the clearest map, in which a dark blue boundary is located in the middle.

Figure 11 shows component planes or connection weights and information loss for input units and competitive units. As can be seen in the figure, connection weights and information loss naturally show the same tendency. One of the differences is that the information loss shows the importance of competitive units and becomes stronger (brown) at the corners. This means that competitive units at the corners are very important in terms of information content about input patterns.

Figure 12 (a) shows a U-matrix, labels and information loss when the map size is increased to 6 by 6. As can be seen in the U-matrix on the left (a1), a clear boundary can be seen in the middle. Corresponding to the boundary in the middle, weak information loss in dark blue can be seen in Figure 12(a3). Figure 12 (b) shows a U-matrix, labels and information loss in a 12 by 4 matrix. As can be

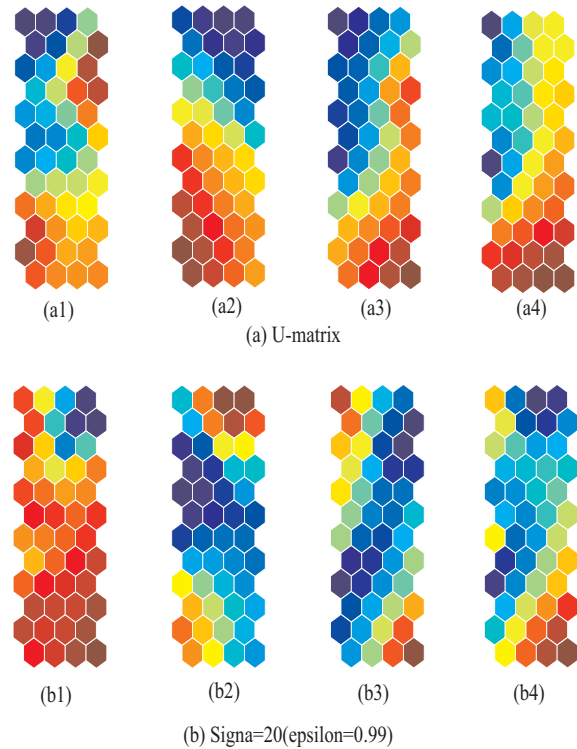


Fig. 11. Component planes or connection weights (a) and information loss for input-competitive units (b) when the Gaussian width is 20.

seen in Figure 12(b1), a boundary in the oblique line can be seen. To this boundary correspond weak competitive units in Figure 12(b3).

Experimental results presented here can be summarized by three points. First, the information loss can produce clearer boundaries by which input patterns are classified into several groups. The boundaries become more stable as the Gaussian width becomes larger. Second, final maps obtained by information loss are greatly dependent upon the Gaussian width σ and the loss parameter ϵ . The Gaussian width σ should be relatively large enough to detect the macroscopic features of input patterns. On the other hand, the loss parameter ϵ should be as large as possible to show the effect of the loss of some elements more clearly. Third, we can obtain two types of maps by information loss: a map in which a boundary is located in the middle of the map, and on the other hand, another map in which an oblique line can be seen.

IV. CONCLUSION

In this paper, we have proposed a new type of information-theoretic approach to feature detection in competitive learning. The new method is called *information loss* and defined by the difference between mutual information with all elements and without some elements in a network. When these elements are deleted, and information lost is significantly large, the elements surely play an important role in information processing. We have applied the method to artificial and symmetric data to show intuitively how well the method extracts the features of input patterns. In addition, we have applied the

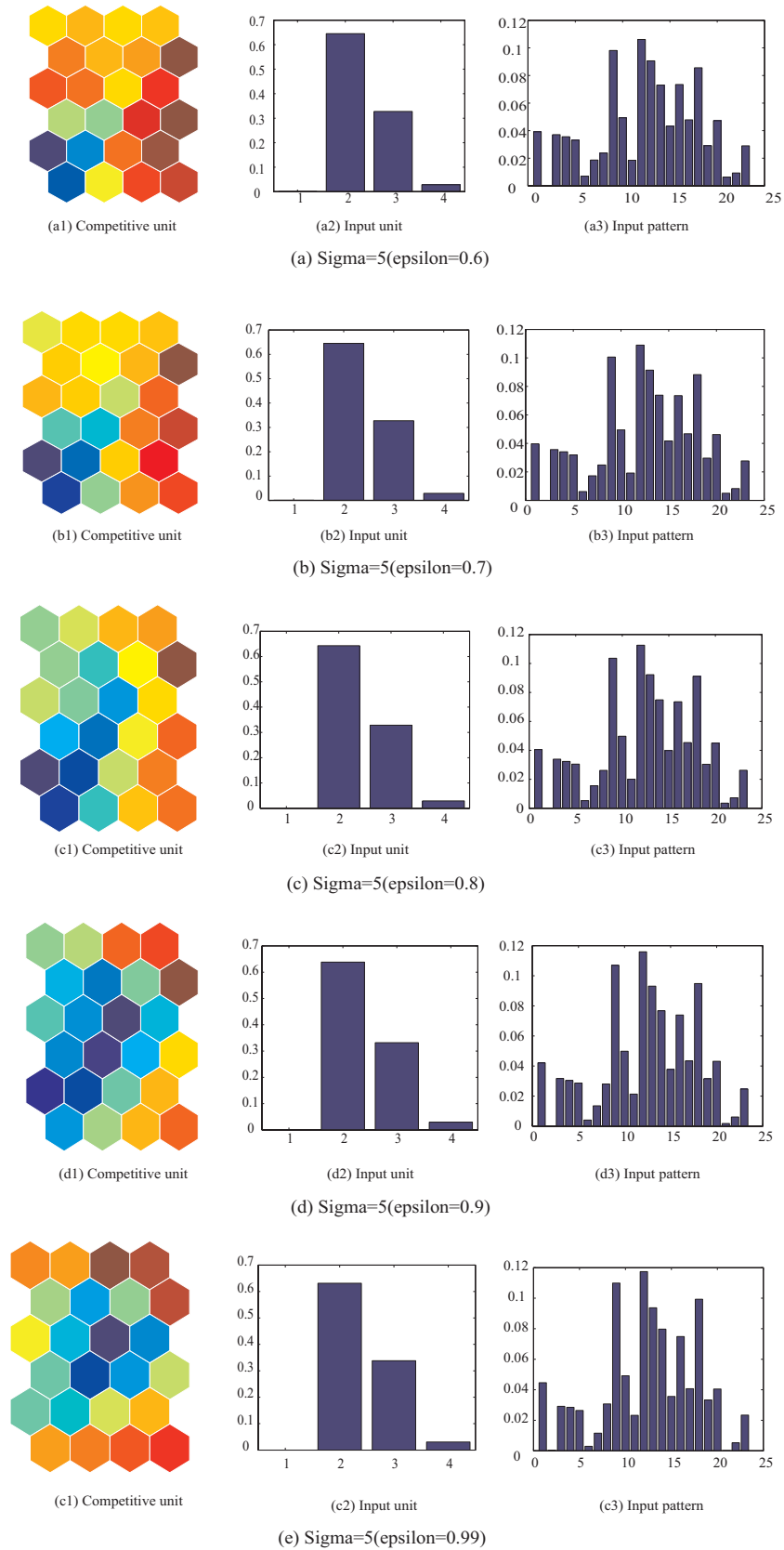


Fig. 9. Three types of information loss when the Gaussian width σ is 5 and the other information loss parameter ϵ is changed from 0.7 to 0.99. Warmer and cooler colors shows higher and lower values of information loss.

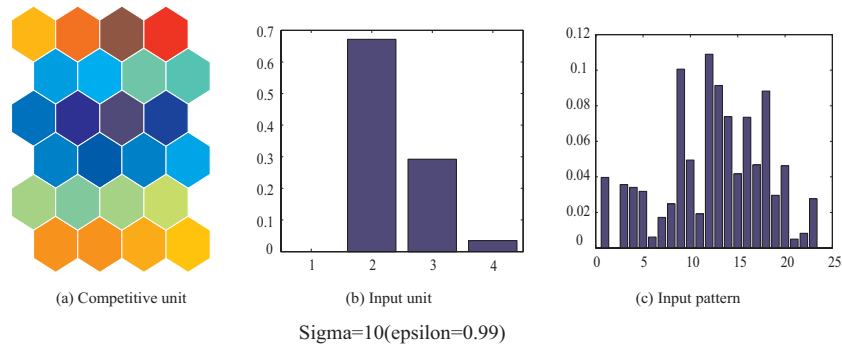


Fig. 10. Information loss for input-competitive units when the Gaussian width σ is 10. Warmer and cooler colors show higher and lower values of information loss.

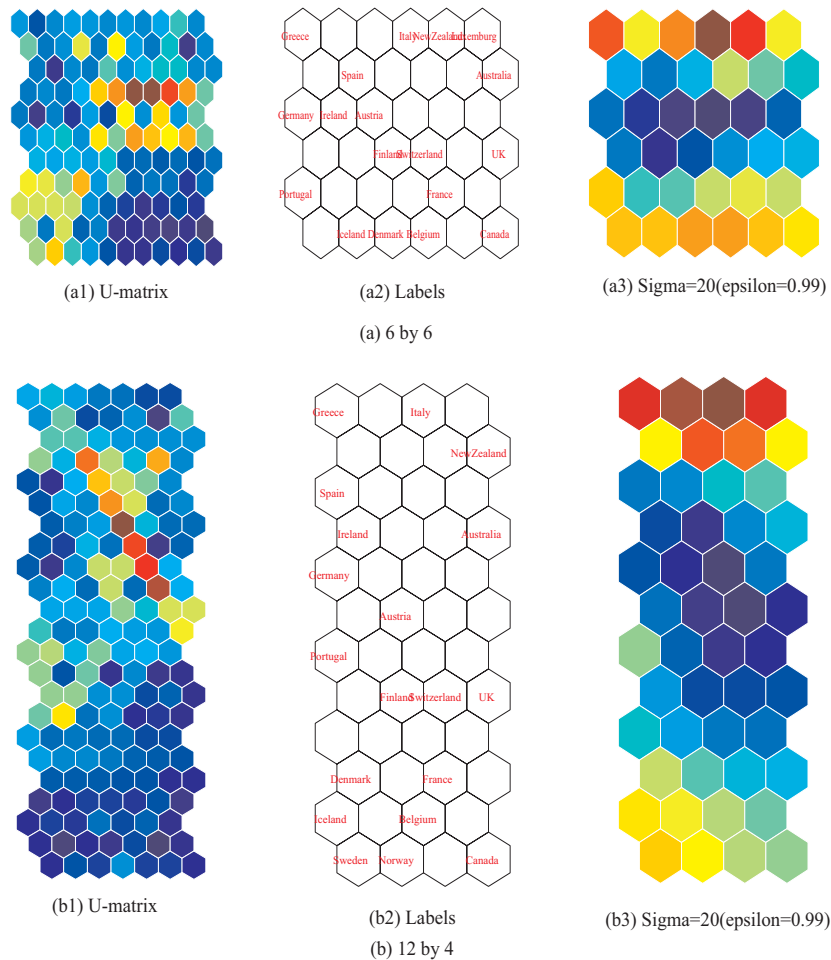


Fig. 12. U-matrices, labels and information loss for competitive units when the size of maps is increased to 6 by 6 (a) and 12 by 4 (b).

method to the classification of several OECD countries. Experimental results have confirmed that the information loss can clearly classify the countries. These results have been comparable to those obtained by the conventional SOM.

For the method to be applicable to more practical and large-scale problems, we should solve three problems. First, final representations are naturally dependent upon two parameters, σ and ϵ . We have experimentally determined two parameters. If explicit relations between the parameters and final representations are determined

theoretically, the new method can be more easily applied to complex problems. In addition, we should more extensively compare the results obtained by the new method with those by the conventional neural methods as well as statistical methods. Finally, we should apply the information loss to larger building blocks of neural networks. Though several problems should be solved, the information loss can be applied to many practical problems because of the simplicity and flexibility of the method.

REFERENCES

- [1] E. Gokcay and J. Principe, "Information theoretic clustering," *IEEE Transactions on Pattern Analysis and Machine*, vol. 24, no. 2, pp. 158–171, 2002.
- [2] D. E. T. Lehn-Schioler, Anant Hegde and J. C. Principe, "Vector-quantization using information theoretic concepts," *Natural Computation*, vol. 4, no. 1, pp. 39–51, 2004.
- [3] D. Erdogmus and J. Principe, "Lower and upper bounds for misclassification probability based on renyi's information," *Journal of VLSI signal processing systems*, vol. 37, no. 2/3, pp. 305–317, 2004.
- [4] K. Torkkola, "Feature extraction by non-parametric mutual information maximization," *Journal of Machine Learning Research*, vol. 3, pp. 1415–1438, 2003.
- [5] N. Slonim and N. Tishby, "Agglomerative information bottleneck," 1999.
- [6] R. Linsker, "Self-organization in a perceptual network," *Computer*, vol. 21, pp. 105–117, 1988.
- [7] R. Linsker, "How to generate ordered maps by maximizing the mutual information between input and output," *Neural Computation*, vol. 1, pp. 402–411, 1989.
- [8] R. Kamimura, T. Kamimura, and O. Uchida, "Flexible feature discovery and structural information," *Connection Science*, vol. 13, no. 4, pp. 323–347, 2001.
- [9] R. Kamimura, T. Kamimura, and H. Takeuchi, "Greedy information acquisition algorithm: A new information theoretic approach to dynamic information acquisition in neural networks," *Connection Science*, vol. 14, no. 2, pp. 137–162, 2002.
- [10] R. Kamimura, "Information theoretic competitive learning in self-adaptive multi-layered networks," *Connection Science*, vol. 13, no. 4, pp. 323–347, 2003.
- [11] R. Kamimura, "Information loss to extract distinctive features in competitive learning learning," in *Proceedings of IEEE Conference on Systems, Man, and Cybernetics*, pp. 1217–1222, 2007.
- [12] D. DeSieno, "Adding a conscience to competitive learning," in *Proceedings of IEEE International Conference on Neural Networks*, (San Diego), pp. 117–124, IEEE, 1988.
- [13] S. C. Ahalt, A. K. Krishnamurthy, P. Chen, and D. E. Melton, "Competitive learning algorithms for vector quantization," *Neural Networks*, vol. 3, pp. 277–290, 1990.
- [14] L. Xu, "Rival penalized competitive learning for clustering analysis, RBF net, and curve detection," *IEEE Transaction on Neural Networks*, vol. 4, no. 4, pp. 636–649, 1993.
- [15] A. Luk and S. Lien, "Properties of the generalized lotto-type competitive learning," in *Proceedings of International conference on neural information processing*, (San Mateo: CA), pp. 1180–1185, Morgan Kaufmann Publishers, 2000.
- [16] M. M. V. Hulle, "The formation of topographic maps that maximize the average mutual information of the output responses to noiseless input signals," *Neural Computation*, vol. 9, no. 3, pp. 595–606, 1997.
- [17] L. L. Gatlin, *Information Theory and Living Systems*. Columbia University Press, 1972.
- [18] M. C. Mozer and P. Smolensky, "Using relevance to reduce network size automatically," *Connection Science*, vol. 1, no. 1, pp. 3–16, 1989.
- [19] E. D. Karnin, "A simple procedure for pruning back-propagation trained neural networks," *IEEE Transactions on Neural Networks*, vol. 1, no. 2, 1990.
- [20] J. S. D. Y. Le Cun and S. A. Solla, "Optimal brain damage," in *Advanced in Neural Information Processing*, pp. 598–605, 1990.
- [21] R. Reed, "Pruning algorithms-a survey," *IEEE Transactions on Neural Networks*, vol. 4, no. 5, 1993.
- [22] E. Kumagai and N. Funao, *Data Mining by R (in Japanese)*. 9-Ten Publishing Company, 2007.

Ryotaro Kamimura is currently a Professor of IT education center of Tokai University, Japan. His current research interests include information-theoretic competitive learning.