1

Information-Theoretic Approach to Interpret Internal Representations of Self-Organizing Maps

Ryotaro Kamimura
IT Education Center, 1117 Kitakaname Hiratsuka Kanagawa 259-1292 Japan

1. Introduction

In this chapter, we propose a new method to measure the importance of input variables and to examine the effect of the input variables on other components. We applied the method to competitive learning, in particular, self-organizing maps, to demonstrate the performance of our method. Because our method is based upon our information-theoretic competitive learning, it is easy to incorporate the idea of the importance of input variables into the method. In addition, by using the SOM, we demonstrate visually how the importance of input variables affects the outputs from the other components, such as competitive units. In this section, we first state that our objective is to interpret the network configurations as clearly as possible. Then, we show why the importance of input variables should be taken into account. Finally, we will briefly survey our information-theoretic competitive learning and its relation to the importance of input variables.

The objective of the new method is to interpret network configurations, focusing upon the meaning of input variables in particular, because we think that one of the most important tasks in neural learning is that of interpreting network configurations explicitly (Rumelhart et al., 1986; Gorman & Sejnowski, 1988). In neural networks' applications, we have had much difficulty to explain how neural networks respond to input patterns and produce their outputs due to the complexity and non-linear nature of data transformation (Mak & Munakata, 2002), namely, the low degree of human comprehensibility (Thrun, 1995; Kahramanli & Allahverdi, 2009) in neural networks. One of the major approaches for interpretation is rule extraction from trained neural networks by symbolic interpretations with three types of methods, namely, *decompositional*, *pedagogical* and *eclectic* (Kahramanli & Allahverdi, 2009). In the decompositional approach (Towell & Shavlik, 1993; Andrews et al., 1993; Tsukimoto, 2000; Garcez et al., 2001), we analyze the hidden unit activations and connection weights for better understanding of network configurations. On the other hand, in the pedagogical approach (Andrews et al., 1993), the neural network is considered to be a black box, and we only focus upon the imitation of input-output relations exhibited by the neural networks. Finally, in the eclectic approach (Andrews et al., 1993; Barakat & Diederich, 2005), both pedagogical and decompositional approaches are incorporated. In the popular decompositional approach, much attention has been paid to hidden units as well as connection weights. The importance of input variables has been implicitly taken into account. For example, Tsukimoto (Tsukimoto, 2000) used the absolute values of connection weights or the squared connection weights to input variables (attributes) for measuring the importance of input variables. In addition,

(Garcez et al., 2001) pointed out that the pruning of input vectors maintained the highest possible precision.
On the other hand, in machine learning, variable selection or the interpretation of input variables has received much attention. In data processing, the number of input variables has become extremely large (Guyon & Elisseeff, 2003). Thus, it is important to estimate which input variable should be taken into account in actual data processing. Variable selection aims to improve the prediction performance, to reduce the cost in prediction and to understand the main mechanism of data processing (Guyon & Elisseeff, 2003). The third aim is more related to the present paper. To cope with this variable selection, many methods have been developed (Steppe & K. W. Bauer, 1997; Belue & K. W. Bauer, 1995; Petersen et al., 1998) so far. However, we have had few attempts made in the field of unsupervised learning, for example, competitive learning and SOM, to take into account the effect of input variables. The methods for input variables in neural networks are mainly related to supervised learning, because of the easy implementation of the measures to represent the importance of input variables (Guyon & Elisseeff, 2003). Few attempts have been made to apply variable selection to unsupervised learning. Thus, it is necessary to examine the effect of input variables through the visualization abilities of the SOM.
In unsupervised learning, explicit evaluation functions have not been established for variable selection (Guyon & Elisseeff, 2003). We have introduced variable selection in unsupervised competitive learning by introducing a method of information loss (Kamimura, 2007; 2008b;a) or information enhancement (Kamimura, 2008c; 2009). In the information loss method, a specific input unit or variable is temporarily deleted, and the change in mutual information between competitive units and input patterns is measured. If the difference between mutual information with and without the input unit is increased, the target input unit certainly plays a very important role. On the other hand, in information enhancement, a specific input unit is used to enhance competitive units or to increase the selectivity of competitive units. If the selectivity measured by mutual information between competitive units and input patterns is large, the target input unit is important to increase the selectivity.
One of the major difficulties with these information-theoretic methods is that it is extremely difficult to determine how much information should be contained in explicit ways. In those methods, there are some parameters to determine how much information should be acquired. However, there are no ways to adjust the parameters and to determine the appropriate amount of information to be acquired. We must adjust the parameters heuristically by examining final results such as competitive unit output and connection weights. In this context, we propose a new method to measure information content to be stored in input variables. The parameters in the methods are changed to increase this information content as much as possible. The basic principle to determine the parameters is how these parameters can maximize the information of the input variables. Compared with the previous methods, the criterion to determine the parameters is more explicit. With the ability to explicitly determine the information content, we can interpret network configurations with more confidence, because our method presents a network configuration with maximum possible information state.
Our method has been developed based on information-theoretic competitive learning. Thus, our method is the most suited for competitive learning. However, we applied the method to the self-organizing maps, for two reasons. First, the self-organizing map is a convenient tool to visualize the good performance of our method, better than pure competitive learning because the good performance can be intuitively understood by visualization techniques related to the SOM. Second, we think that the self-organizing map is also an attempt to

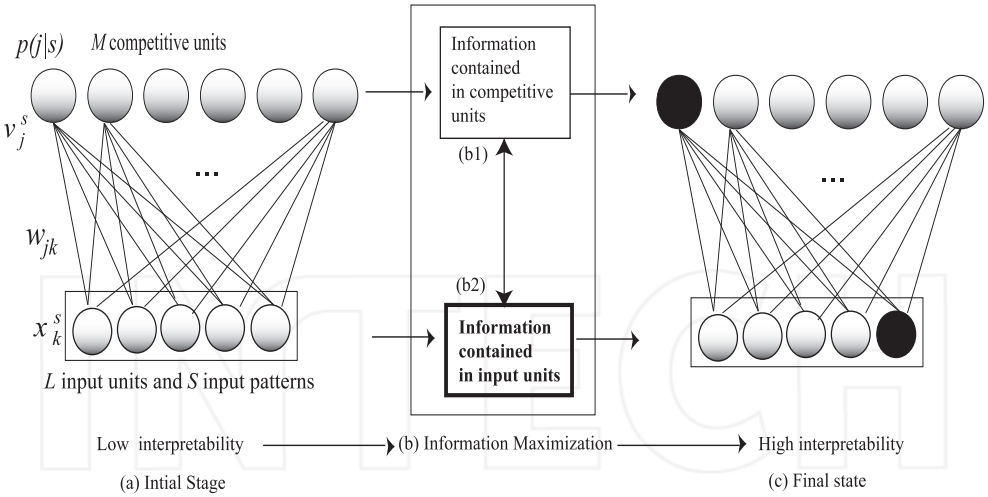


Fig. 1. A concept of the information-theoretic approach.

interpret network configurations not by symbolic but by visual representation. Though the SOM has been developed for clustering and data mining of high-dimensional data (Kohonen, 1988; 1995; Tasdemir & Merenyi, 2009), the SOM's main contribution consists in the visualization of high dimensional data in terms of the lower dimensions with various visualization techniques. In the SOM, different final configurations are made explicit by using various visualization techniques, taking into account codebooks and data distribution (Polzlbauer et al., 2006; Vesanto, 1999; Kaski et al., 1998; Mao & Jain, 1995; Ultsch & Siemon, 1990; Ultsch, 2003). From our point of view, the approach of visual representations to interpret network configurations corresponds conceptually to the decompositional approach in rule extraction, though symbolic representations are not extracted. We think that visualization is an effective tool for interpreting final configurations, corresponding to the extraction of symbolic rules in rule extraction.

2. Theory and computational methods

2.1 Information-theoretic approach

We aim to apply our information-theoretic principle to the detection of the importance of input variables. Principally, our objective is to maximize any information contained in components in a network, hoping that condensed information contained in the components is simpler and more interpretable than that before information maximization. In our sense, information maximization means strictly that information on input patterns is represented in a small number of components, such as competitive units and input units. Figure 1 shows a schematic diagram of our objective. In the figure, from the initial to the final state, the number of important units represented in black is smaller. First, information contained in competitive units must be as large as possible, as shown in Figure 1(b1). We have already shown that this information on competitive units, more exactly, mutual information between competitive units and input patterns, represents competitive processes (Kamimura & Kamimura, 2000; Kamimura et al., 2001; Kamimura, 2003a;b;c;d). Thus, this information, or more exactly, mutual information, should be as large as possible. On the other hand, we can consider

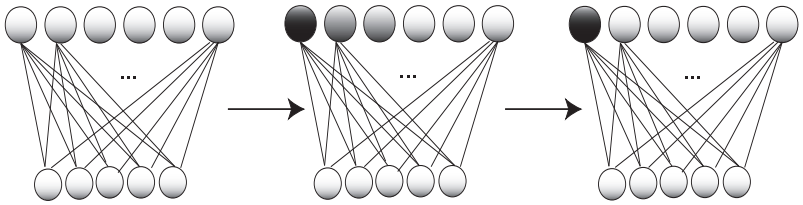


Fig. 2. Competitive unit outputs for an initial state (a), an intermediate state (b) and a state with maximum mutual information (c). The black and white competitive units represent the strong and weak firing rates, respectively.

information content in input units. As shown in Figure 1(b2), this information should be increased as much as possible. When this information is increased, the number of important input variables is decreased. We focus here on input units, or variables, and then information maximization should be biased toward information contained in input units. Thus, mutual information in competitive units should be increased under the condition that the increase in the mutual information prevents a network from increasing information in input units. In the following section, we first explain mutual information between competitive units and input patterns. Then, using the mutual information, we define the importance of input units, by which the information of input variables is defined. Finally, we explain how to compromise these two types of information.

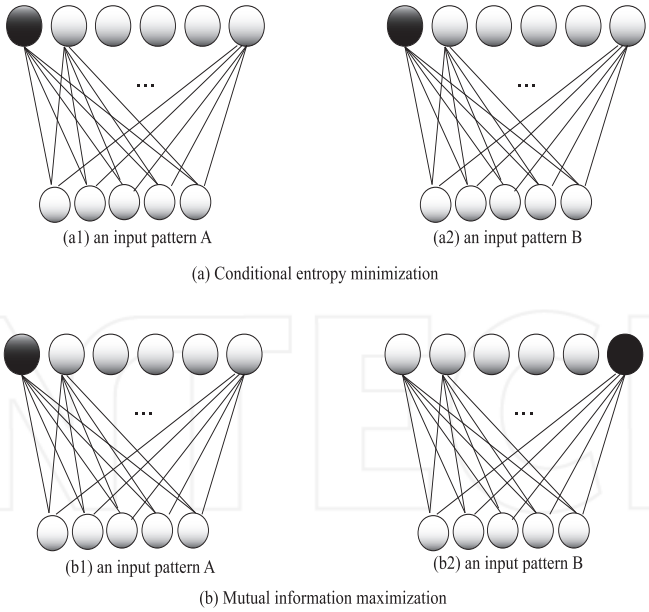


Fig. 3. Competitive unit outputs for conditional entropy minimization (a) and mutual information maximization (b). The black and white competitive units represent the strong and weak firing rates, respectively.

2.2 Information-theoretic competitive learning

We begin with information for competitive units, because information of input units is defined based upon the information for competitive units. We have so far demonstrated that competitive processes in competitive learning can be described by using the mutual information between competitive units and input patterns(Kamimura & Kamimura, 2000; Kamimura et al., 2001; Kamimura, 2003a;b;c;d). In other words, the degree of organization of competitive units can be described by using mutual information between competitive units and input patterns. Figures 2 (a), (b) and (c) show three states that depend on the amount of information stored in competitive unit outputs. Figure 2(a) shows an initial state without any information on input patterns, where competitive unit outputs respond equally to all input patterns. When some quantity of information is stored in competitive unit outputs, several neurons tend to fire at the corners, shown in Figure 2(b). When mutual information between input patterns and competitive units is maximized, shown in Figure 2(c), only one competitive unit is turned on for specific input patterns.

We explain this mutual information more exactly by using the network architecture shown in Figure 1. In the network, x_k^s, w_{jk} and v_j^s represent the kth element of the sth input pattern, connection weights from the kth input to the jth competitive unit and the jth competitive unit output for the sth input pattern. The competitive unit outputs can be normalized as $p(j \mid s)$ to represent the firing probability of the jth competitive unit. In the network, we have L input units, M competitive units and S input patterns.

First, the jth competitive unit outputs v_j^s for the sth input pattern can be computed by

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

The firing probability of the jth competitive unit for the sth input pattern can be obtained by normalizing these competitive unit outputs

$$p(j \mid s) = \frac{v_j^s}{\sum_{m=1}^{M} v_m^s}. \tag{2}$$

Then, mutual information between competitive units and input patterns can be defined by

$$\begin{aligned} MI &= \sum_{s=1}^{S}\sum_{j=1}^{M} p(s)p(j \mid s)\log\frac{p(j \mid s)}{p(j)} \\ &= -\sum_{j=1}^{M} p(j)\log p(j) + \sum_{s=1}^{S}\sum_{j=1}^{M} p(s)p(j \mid s)\log p(j \mid s). \end{aligned} \tag{3}$$

Mutual information is decomposed into the first term of entropy and the second term of conditional entropy. As shown in Figure 3(a), when only conditional entropy is minimized, we have the high possibility that only one competitive unit at the corner in the figure is always turned on. On the other hand, when mutual information is maximized, different competitive units respond to different input patterns, as shown in Figure 2(b). Thus, mutual information maximization can realize a process of competition in competitive learning.

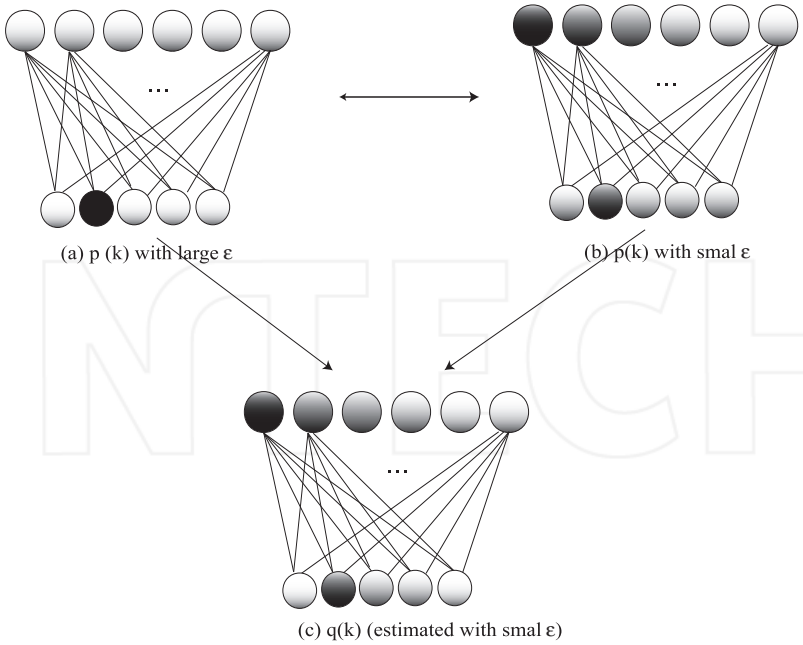


Fig. 4. Importance $p(k)$ with large ϵ (a), small ϵ and estimated importance (c).

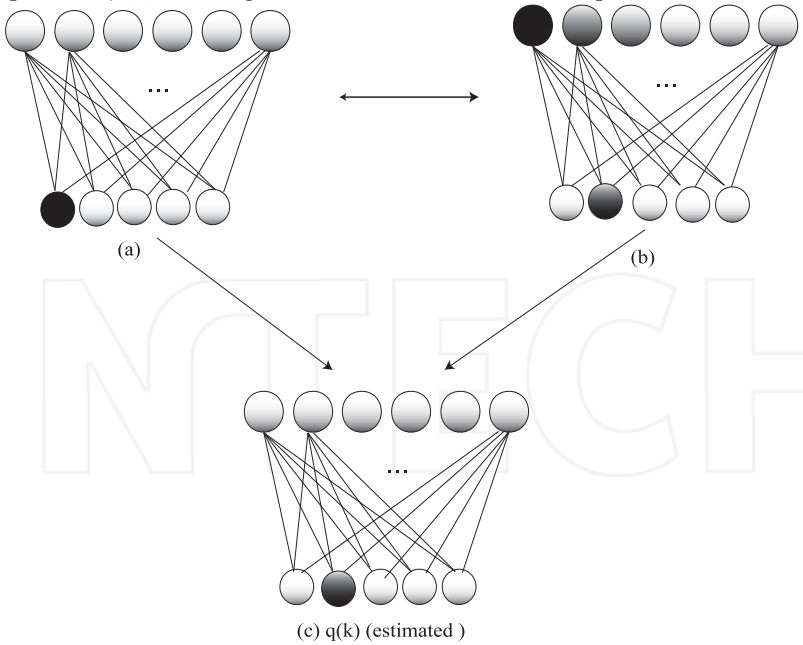


Fig. 5. Importance $p(k)$ with large ϵ (a), small ϵ and estimated importance (c).

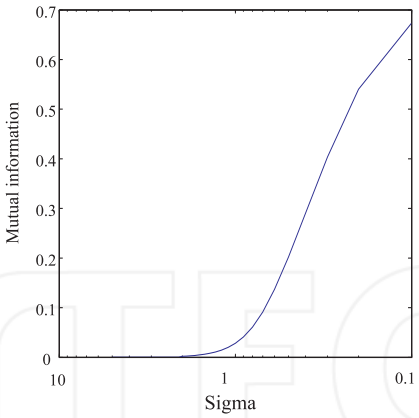


Fig. 6. Mutual information as a function of the parameter σ.

2.3 Estimated information for input variables

Using the mutual information described in the previous section, we try to estimate the importance of input variables. For this purpose, we initially suppose the importance of input units by using the parameter ϵ

$$p(k;t,\epsilon) = \begin{cases} \epsilon, & \text{if } k = t; \\ (1-\epsilon)/(L-1), & \text{otherwise,} \end{cases}$$

where ϵ is a parameter to determine the degree of *attention* paid to the kth input unit. As the parameter ϵ is increased, more attention is paid to the kth target input unit or variable. Figure 4(a) shows a case where the parameter ϵ is the largest value, one, for the second input unit, and the importance of the second input unit is the largest. However, no explicit patterns in terms of competitive unit outputs can be seen. On the other hand, in Figure 4(b), the parameter ϵ is small, the intensity of the second competitive unit is weakened and the other competitive units fires slightly. However, competitive unit outputs are slightly organized. Thus, in this case, the small parameter value of ϵ is better to organize competitive units. Then, the actual importance shown in Figure 4(c) can be modeled by using this small parameter value. Figure 5 shows a case where the first input unit produces no effect on competitive unit output patterns (a), while the second unit produces an organized competitive unit output pattern (b). Thus, the second input unit of the estimated ones is large (c).

To estimate the information, we must introduce the mutual information between competitive units and input patterns. Now, the distance between input patterns and connection weights, when focusing upon the tth input unit, is computed by

$$d_j^s(t,\epsilon) = \sum_{k=1}^{L} p(k;t,\epsilon)(x_k^s - w_{jk})^2. \tag{4}$$

By using this equation, we have competitive unit outputs for the tth input unit

$$v_j^s(t;\sigma,\epsilon) = \exp\left(-\frac{\sum_{k=1}^{L} p(k;t,\epsilon)(x_k^s - w_{jk})^2}{2\sigma^2}\right). \tag{5}$$

Normalizing these outputs, we have

$$p(j \mid s;t,\sigma,\epsilon) = \frac{v_j^s(t;\sigma,\epsilon)}{\sum_{m=1}^{M} v_m^s(t;\sigma,\epsilon)}. \tag{6}$$

The firing probability of the jth competitive unit is defined by

$$p(j;t,\sigma,\epsilon) = \sum_{s=1}^{S} p(s)p(j \mid s;t,\sigma,\epsilon). \tag{7}$$

By using these probabilities, we have mutual information MI when the tth input unit is focused on:

$$MI(t;\sigma,\epsilon) = \sum_{s=1}^{S}\sum_{j=1}^{M} p(s)p(j \mid s;t,\sigma,\epsilon)\log\frac{p(j \mid s;t,\sigma,\epsilon)}{p(j;t,\sigma,\epsilon)}. \tag{8}$$

This mutual information shows how well the tth input unit contributes to a process of competition among competitive units (Kamimura, 2003b).

2.4 Importance of input variables

Mutual information $MI(t;\sigma,\epsilon)$ represents how well the tth input variable contributes to the process of competition. As this mutual information gets larger, the tth input variable plays a more essential role in realizing competitive processes, and the variable should be considered to be important in competition. We approximate the importance of input units with this mutual information, and we have

$$q(t;\sigma,\epsilon) \approx \frac{MI(t;\sigma,\epsilon)}{\sum_{l=1}^{L} MI(l;\sigma,\epsilon)}. \tag{9}$$

Then, using the importance, $q(t;\sigma,\epsilon)$, the estimated information can be defined by

$$EI(\sigma,\epsilon) = \sum_{k=1}^{L} q(k;\sigma,\epsilon)\log\frac{q(k;\sigma,\epsilon)}{q_0(k;\sigma,\epsilon)}. \tag{10}$$

In this equation, q_0 is supposed to be equi-probable, namely, 1/L. As this estimated information gets larger, the number of important input variables gets smaller. Thus, we must increase this estimated information as much as possible, because we are trying to find a small number of important input variables.

2.5 Ratio to determine the parameters

This estimated information EI is based upon mutual information between competitive units and input patterns. Then, mutual information is dependent on the spread parameter σ and ϵ, and in particular, the mutual information is changed by the spread parameter σ. Generally, mutual information can be increased by decreasing the spread parameter σ. Thus, for the parameter σ, the parameter should be as small as possible, meaning that mutual information is as large as possible. Mutual information between competitive units and input patterns represents the degree of organization of a network; as the parameter σ gets smaller, the corresponding mutual information gets larger. This means that, when the parameter σ is small, the organization of a network is large. In addition, the importance of input variables

must be increased as much as possible. Thus, we introduce the ratio RE of the estimated information to the parameter σ

$$RE(\sigma,\epsilon) = \frac{EI(\sigma,\epsilon)}{\sigma}. \tag{11}$$

We try to increase this ratio as much as possible by changing the parameter σ and ϵ. This ratio means that we must increase the estimated information as much as possible. In addition, the mutual information between competitive units and input patterns must be as large as possible, which is realized by the property that, when the parameter σ is smaller, the mutual information is larger.

2.6 Self-organizing maps

Finally, we should note the conventional self-organizing maps (SOM) used in this chapter. Principally, the SOM is a method to increase mutual information that takes into account interaction among competitive units. The reason why we use the SOM as a basic learning method is that we have some difficulty in implementing lateral interactions in competitive output units from information-theoretic points of view[1]. In the SOM, at each training step, the data set is partitioned according to the Voronoi regions of map vectors. First, we must select the best matching unit (BMU), denoted by c:

$$c = \mathrm{argmin}_j \sum_{k=1}^{L} (x_k^s - w_{jk})^2. \tag{12}$$

This selection of the BMU corresponds to a case where mutual information between competitive units and input patterns is maximized. Then, we must compute a neighborhood kernel, h, around the winning unit c.

$$h_{jc} = \exp\left(-\frac{\| \mathbf{r}_c - \mathbf{r}_j \|^2}{2\sigma^2}\right), \tag{13}$$

where $\mathbf{r}_c$ and $\mathbf{r}_j$ denote vectors representing the position of the winner and jth competitive unit, respectively, and σ is a neighborhood radius. Connection weights w_{jk} are computed by

$$w_{jk} = \frac{\sum_{s=1}^{S} h_{jc} x_k^s}{\sum_{s=1}^{S} h_{jc}}. \tag{14}$$

We can say that the SOM is also one of the methods that increases mutual information between competitive units and input patterns.

3. Results and discussion

3.1 Experimental results

3.1.1 Symmetry data

We first applied the method to symmetric data in which input patterns are symmetric, as shown in Figure 7(a). Therefore, the method must detect this symmetric property at least. Figure 7(b) and (c) show a U-matrix and labels obtained by the conventional SOM. As can be seen in the figure, in the middle of the U-matrix, clear boundaries in warmer colors can be

[1] We will discuss this problem in the discussion section.

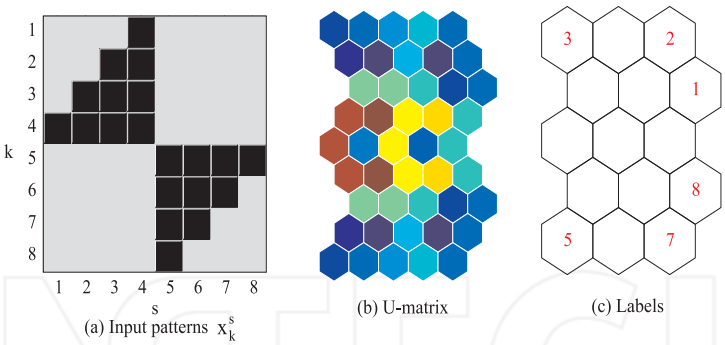


Fig. 7. Original data x_k^s (a), U-matrix (b) and labels (c) for the symmetric data obtained by the SOM.

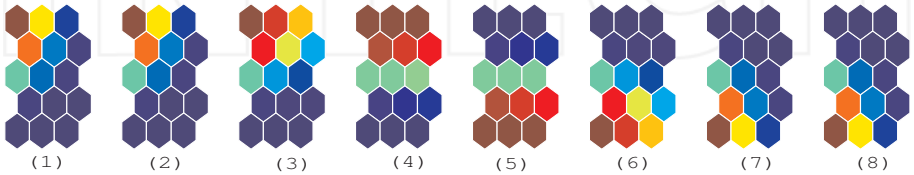


Fig. 8. Component planes along eight input variables obtained by the SOM for the symmetric data.

detected. We can see from the labels in Figure 7(c) that input patterns are naturally classified into two classes. Figure 8 shows component planes along six input units. As component planes move from the first input unit to fourth input unit, they show a gradual increase in the number of strong connection weights (in warmer colors) on the upper part of the map. On the other hand, component planes move from the fifth input unit to the eighth input unit, and then they show a gradual increase in the number of strong connection weights on the lower part of the map. This means that the importance of component planes becomes larger as the component planes move to the center. This property of component planes explains well the symmetric property of the original data.

Figure 9(a) shows estimated information $EI(\sigma,\epsilon)$ as a function of the spread parameter σ for six different values of the parameter ϵ. The computational procedure is as follows. First, the parameter ϵ is chosen; for example, ϵ is 0.2. Then, we try to increase the estimated information EI as much as possible. As shown in Figure 9(a), when the parameter ϵ is set to 0.2, then the other parameter σ is increased up to 1.1, where the estimated information reaches its steady state. Beyond this point, the estimated information cannot be increased. Learning is considered to be finished when the difference in estimated information between the present and the previous state is less than 0.001. We can see that, when the parameter ϵ is larger, the estimated information is larger. In other words, when we focus upon a specific input variable more intensely, the estimated information becomes larger. In addition, we can see that, when the estimated information is larger, the other parameter σ is also increased. To see the situation more exactly, we plot the relations between the two parameters, σ and ϵ. Figure 9(b) shows the final estimated information, with the final value of the parameter σ as a function of the parameter ϵ. The estimated information is increased and reaches its steady state as the parameter ϵ is increased. Figure 9(c) shows the values of the parameter σ as a

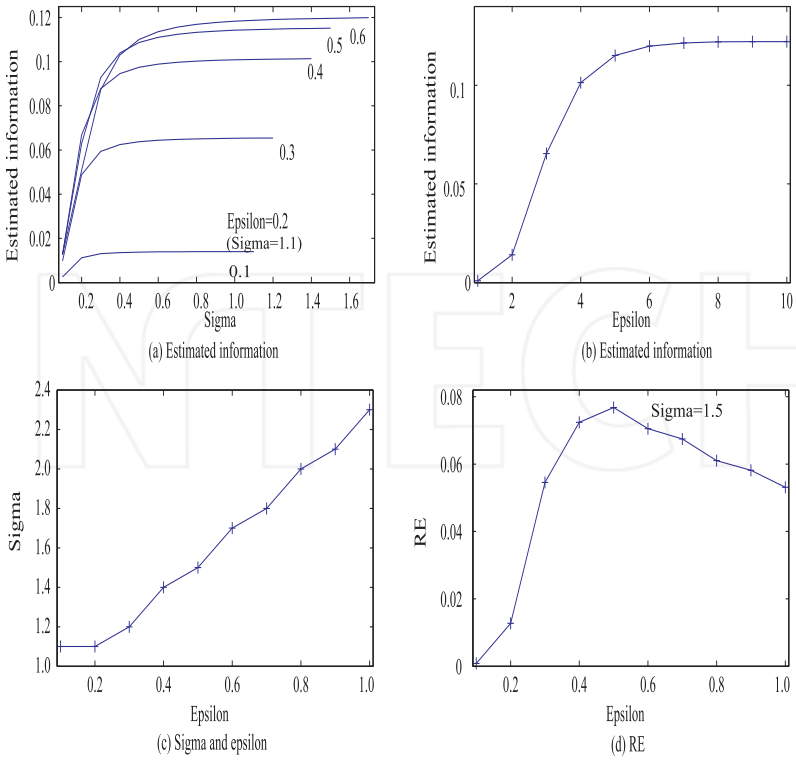


Fig. 9. Information as a function of the parameter σ (a) and the parameter ϵ (b). Optimal values of the parameter σ as a function of the parameter ϵ (c). The ratio RE as a function of the parameter ϵ.

function of the other parameter ϵ. The parameter σ is increased constantly as the parameter ϵ is increased. As mentioned above, for the mutual information between competitive units and input patterns to be increased, the parameter σ should be as small as possible. Therefore, we have introduced the ratio RE. This ratio is gradually increased, and it reaches its peak when the parameter ϵ is 0.5. Thus, this value of 0.5 produced the optimal information, where the estimated information is sufficiently high, and in addition, mutual information between competitive units and input patterns is not so small, because the parameter σ is relative small. Figure 10 shows the estimated importance $q(k)$ for four different values of the parameter ϵ. Figure 10(a) shows the importance when the parameter ϵ is 0.1. The estimated importance is flat, and little difference can be seen. As the parameter is increased from 0.2 (b) to 0.3 (c), gradually, the importance of different input units is made clearer. Finally, when the parameter ϵ is 0.5 (d, optimal), the range of the importance is the largest, and we can easily see the symmetric property of the data. As the input variable moves to the center, the importance of input variable naturally increases. These results demonstrate that the ratio of the estimated information to the parameter σ shows the most interpretable importance of input variables. In addition, we plot the estimated firing probability $p(j)$ with the optimal values of the parameters in Figure 10(e) and (f). As the probability $p(j)$ increases, the corresponding

competitive unit responds to the larger number of input patterns. As can be seen in the figure, the higher values of the probability can be seen in the middle of the map. This means that competitive units in the middle respond to many input patterns. On the other hand, competitive units on the upper and lower parts of the map respond to a fewer number of input patterns, which means that competitive units on the upper and lower parts of the map respond very selectively to input patterns. Thus, we can say that the high probabilities represent a boundary between classes. As can be seen in the figure, input patterns can be classified into two groups by the competitive units with high probabilities in the middle.
Figure 11 shows results when the network size is small (a) and large (b). When the network is small, the same results in terms of U-matrix, RE, importance and $p(j)$ can be obtained. On the other hand, when the network is large, the obtained U-matrix in Figure 11(b1) shows the detailed classification of input patterns, while the probability $p(j)$ (Figure 11(b4)) clearly shows a boundary in the middle of the map. The ratio RE and the importance show almost the same results as those obtained by the normal-sized network.

3.1.2 Student survey No. 1: an image of a university

Second, we applied the method to a student survey in which 39 students at a university were asked to answer nine questions on the good points of a university[2]. The evaluation scores ranged between five (the most favorable) and one (least favorable). Figures 12(a) and (b) show a U-matrix and labels by the conventional SOM. As can be seen in the figure, some class boundaries seem to be present on the lower part of the matrix. Figure 13 shows component planes along nine input variables. Component planes become larger for the lower part of the map. This means that the lower part of the map is a group of students with a more favorable image concerning the input variables. However, we could not estimate the characteristics of student groups separated by these boundaries. Thus, the visualization ability of the SOM is incompetent at dealing with this problem.
Figure 14(a) shows the estimated information as a function of the parameter σ for six different values of the other parameter ϵ. As the parameter ϵ is increased, the estimated information is increased, and the corresponding value of the parameter σ tends to increase also. Figure 14(b) shows the final estimated information for the parameter σ as a function of the parameter ϵ. Information is increased gradually and reaches its steady state as the parameter ϵ is increased. Figure 14(c) shows the value of the parameter σ as a function of the parameter ϵ. As the parameter ϵ is increased, the parameter σ is linearly increased. Figure 14(d) shows the ratio of the estimated information by the parameter σ. When the parameter ϵ is 0.4, the largest value of the ratio is obtained. This means that with this value of 0.4, the estimated information is the largest, with reasonable high mutual information. Figure 15 shows the importance $q(k)$ of nine input variables. The values of the importance are gradually increased from $\epsilon = 0.1$ (a) to $\epsilon = 0.4$ (optimal). Then, the largest range of the importance can be obtained when the ratio takes its optimal value of 0.4. As can be seen in the figures, gradually, input variable No. 4 (teachers' attitude toward students) becomes larger and plays important roles to make self-organizing maps more organized. Finally, Figure 15(e) shows that, in the middle of the map, there are competitive units with higher probabilities $p(j)$. This means that input patterns are classified on this boundary in the middle of the map.
Then, we examine where these measures of the importance and estimated information are independent of map size. For this purpose, we prepare a small- and large-sized network and

[2]This survey was conducted by Mr. Kenta Aoyama in 2010 for 41 students. We deleted two students whose evaluation scores toward the variables were zero or five for all questions.

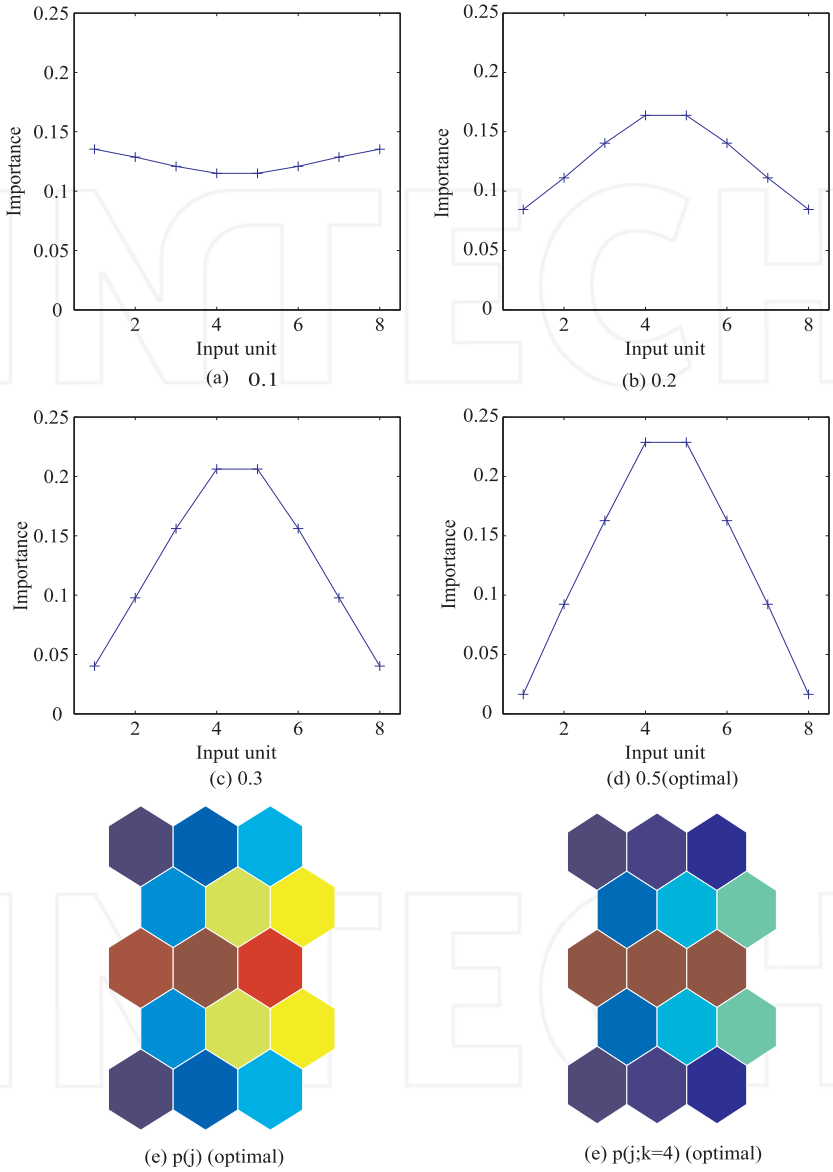


Fig. 10. Estimated importance for four different values of the parameter ϵ (a)-(d) and the estimated $p(j)$ with the optimal values of two parameters (e), (f).

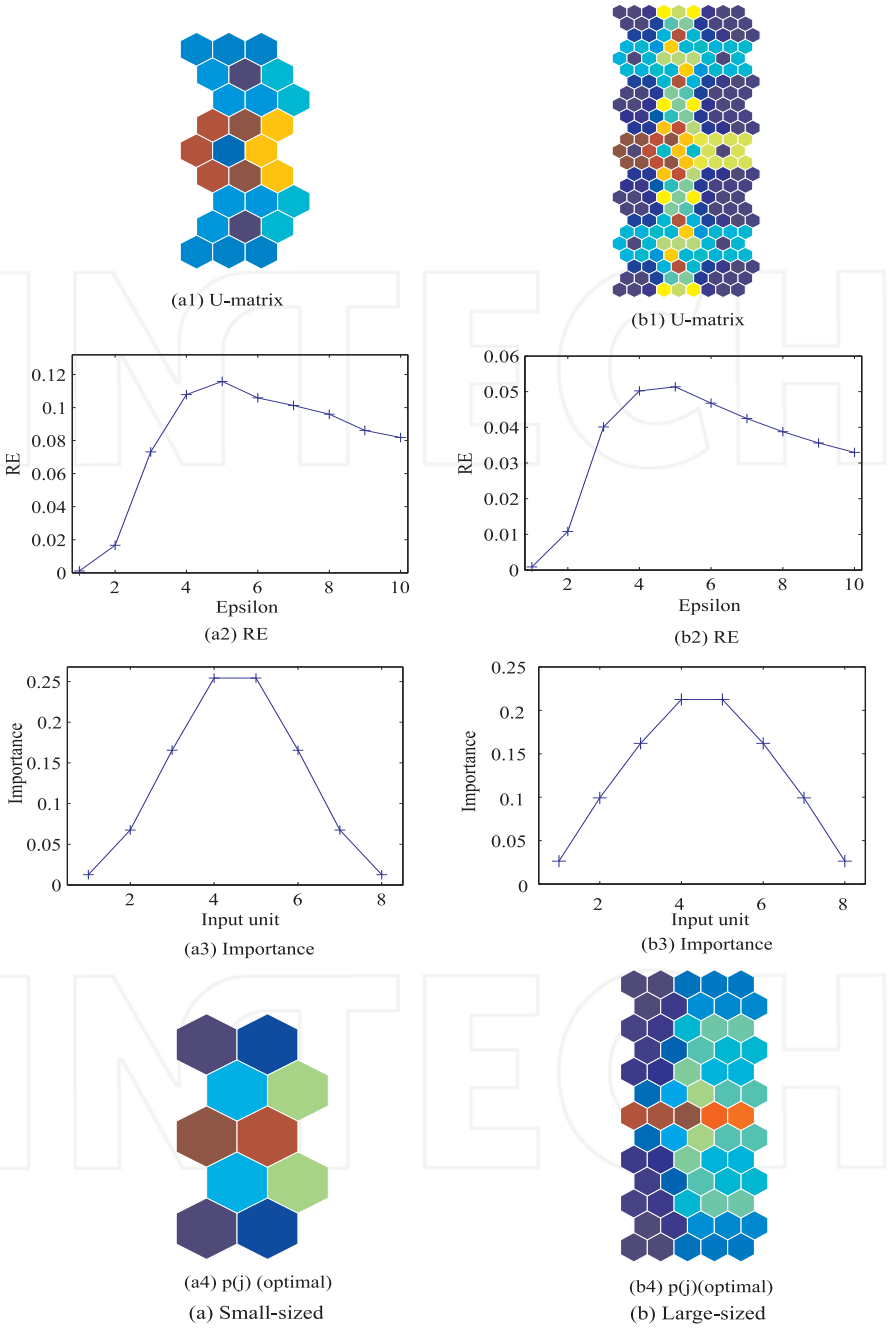


Fig. 11. U-matrices (1), the ratio RE (2), the values of importance (3) and probabilities $p(j)$ for the small-sized (a) and large-sized network (b).

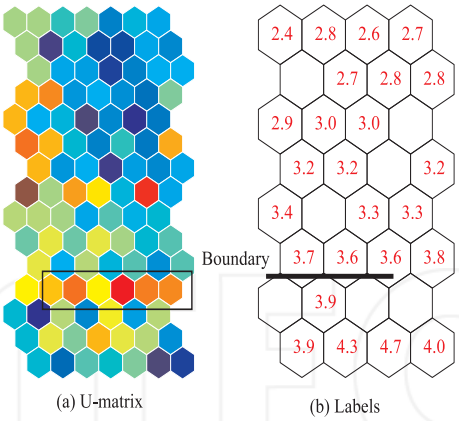


Fig. 12. Original data (a), U-matrix (b) and labels (student No.) (c) obtained by the SOM for student survey No. 1.

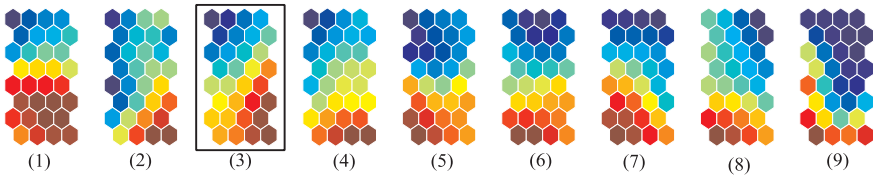


Fig. 13. Component planes along nine input variables obtained by the the SOM for student survey No.1.

then compute the estimated information. Figures 16(a1) and (b1) show final U-matrices for the small and large network, respectively. Figures 16(a2) and (b2) show the ratio RE as a function of the parameter ϵ. We can see that the best values are obtained when the parameter ϵ is 0.4, which is the same as that with the normal-sized network. Figures 16(a3) and (b3) show the values of importance for nine input variables. As can be seen in the figures, variable No. 4 (teachers' attitude) plays the important role, as is the case with the normal-sized network. In addition, we can observe that the range of importance $q(k)$ is slightly decreased when the network size is larger. Figures 16(a4) and (b4) show the probabilities $p(j)$. When the size is small, in Figure 16(a4), a clear boundary in the middle can be seen. However, when the size is large, in Figure 16(b4), the boundary becomes very wide.

3.1.3 Student survey No. 2: mass media and urbanization

We conducted a student survey on to what extent students wanted to live in an urban area or abroad and this desire's relation to mass media[3]. Figure 17 shows the U-matrix (a) and the corresponding labels (b). Some class boundaries in warmer colors seem to be located on the upper part of the map. Figure 18 shows component planes along eight input variables. Though those on the lower part of the maps tend to be stronger, in warmer colors, eight component planes are different from each other. In this problem, we cannot detect the major characteristics among groups separated by the supposed boundaries in warmer colors in the U-matrix.

[3]This survey was also conducted by Kenta Aoyama, December 2009.

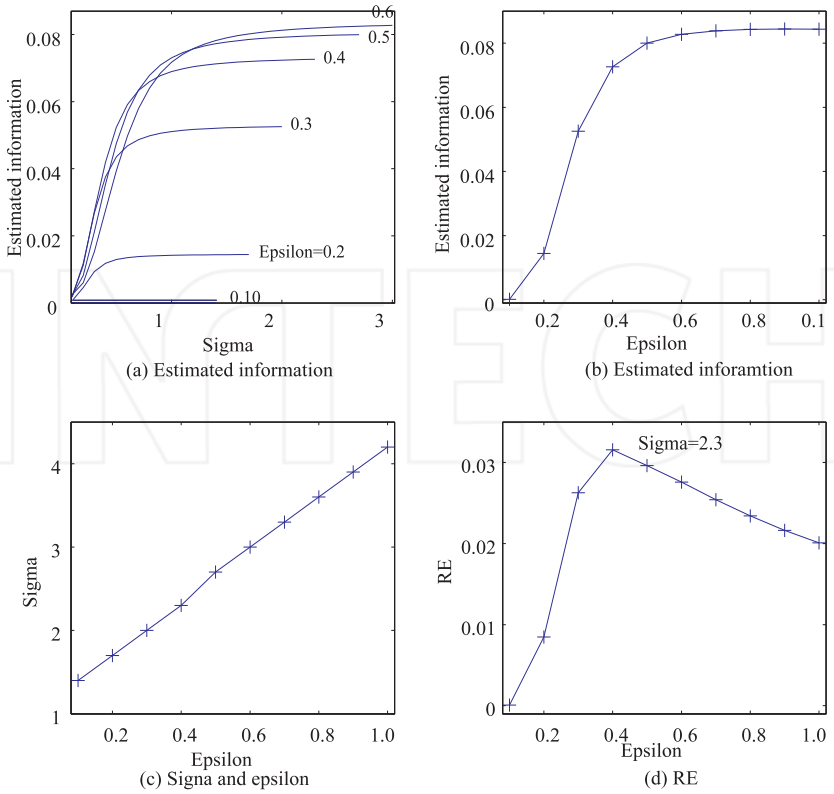


Fig. 14. Information as a function of the parameter σ (a) and the parameter ϵ (b); optimal values of the parameter σ as a function of the parameter ϵ (c); the ratio RE as a function of the parameter ϵ; and competitive unit output $p(j)$ when the information is maximized (d).

Figure 19(a) shows the estimated information as a function of the parameter σ for six different values of the parameter ϵ. As the parameter ϵ gets larger, the estimated information becomes larger and the other parameter σ is larger. Figure 19(b) shows the estimated information as a function of the parameter ϵ. As can be seen in the figure, the estimated information is gradually increased as the parameter ϵ grows larger. Figure 19(c) shows the values of the parameter σ as a function of the parameter ϵ. As the parameter ϵ is increased, the corresponding values of the parameter σ become larger. Figure 19(d) shows the ratio RE of the estimated information to the parameter σ. When the parameter ϵ is 0.4, the largest value can be obtained. Figure 20 shows the importance $p(k)$ for eight input variables. When the parameter ϵ is 0.1, the values of importance is quit small, and we cannot see any characteristics in the figure. As the parameter ϵ is increased from 0.2 (b) and 0.3 (c), the importance of the value of input variable No. 4 rises gradually. Then, when the parameter ϵ is 0.4, we have the largest value of the importance. As can be seen in the figures, input variable No. 4 shows the largest value of importance, namely, "to live abroad." This feature plays the most important role in organizing competitive unit outputs. This means that students are classified into several groups based upon this feature. Figure 20(e) shows the probability $p(j)$ with the optimal

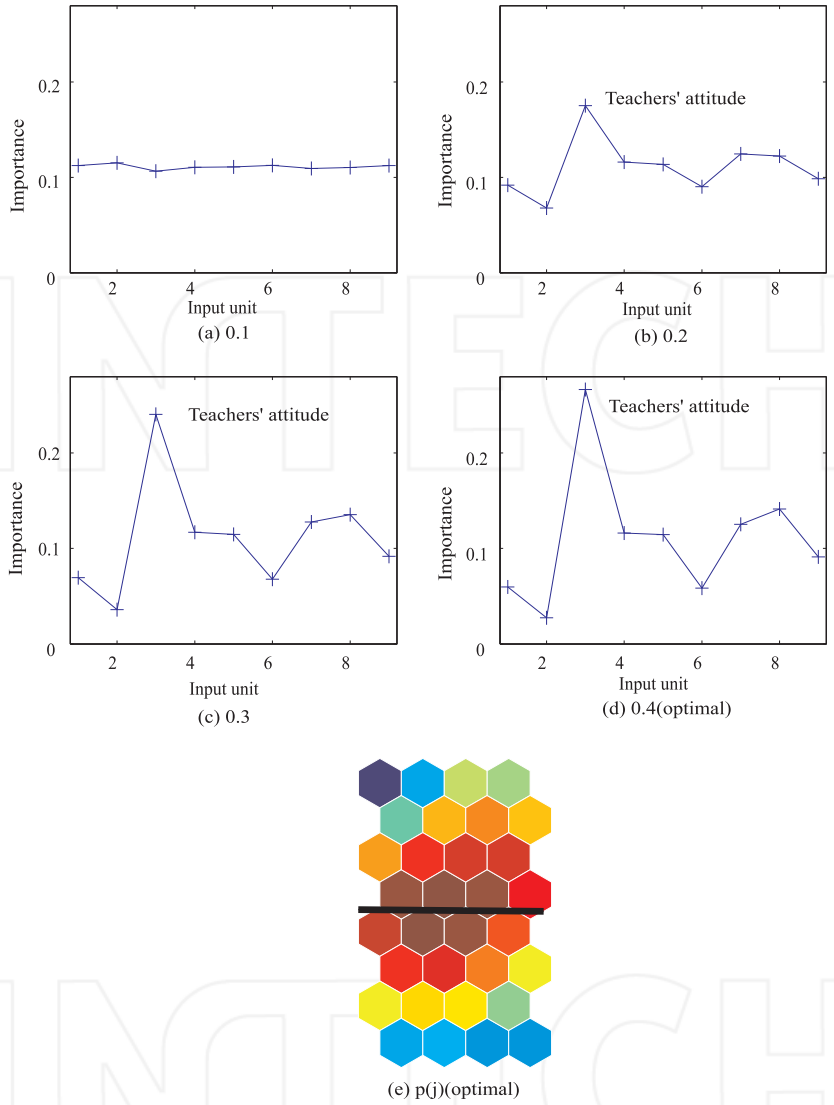


Fig. 15. Estimated importance for four different values of the parameter ϵ (a)-(d) and the probability $p(j)$ (e).

values of the parameters. As can be seen in the figure, a clear, diagonal boundary in warmer colors is located on the map. Students with higher scores and lower scores are classified by this boundary.

Then, we try to see whether these characteristics can be obtained when the network size is different. Figure 21 shows the U-matrices, the ratio RE and the importance for two different sizes of network. When a network is small or large, the optimal value of the parameter ϵ is 0.4, as shown in Figures 21(a2) and (b2), and the importance of input variable No. 4 is the

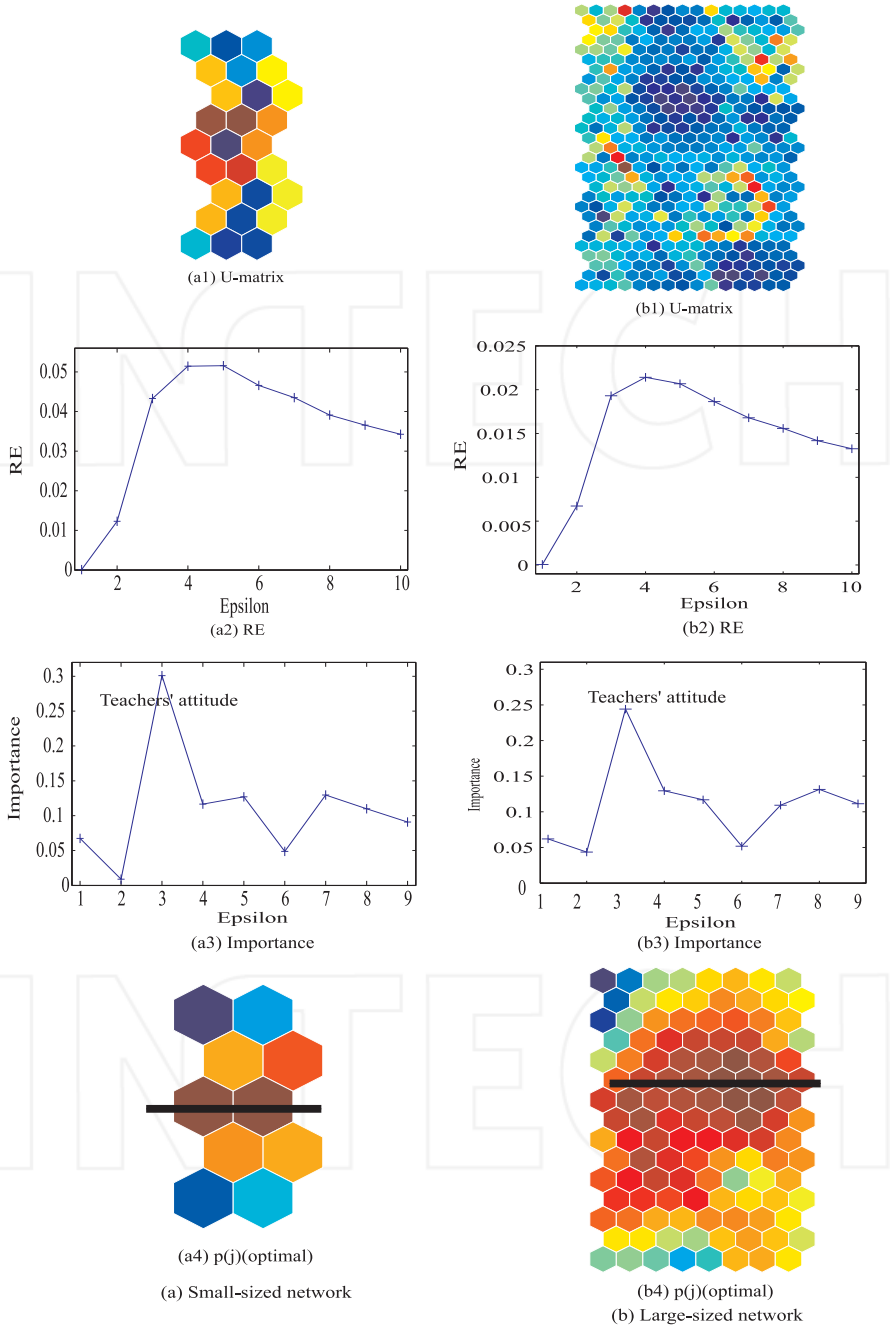


Fig. 16. U-matrices (1), the ratio RE (2), the values of importance (3) and the probability $p(j)$ (4) for the small-sized (a) and large-sized network (b).

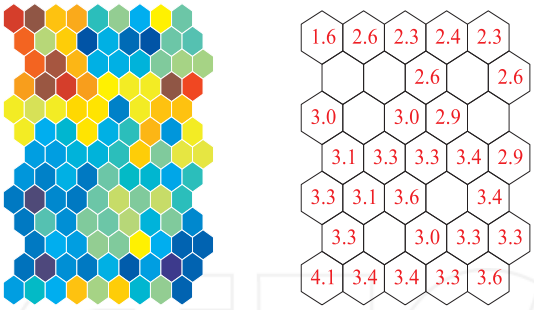


Fig. 17. Original data (a), U-matrix (b) and labels (c) obtained by the SOM for student survey No. 2.

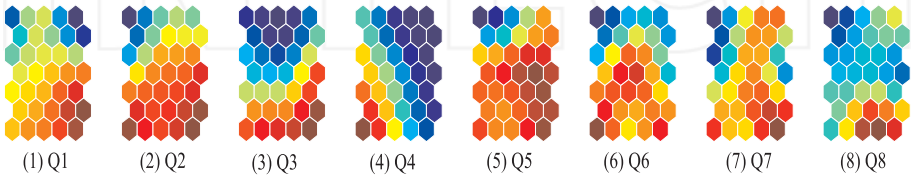


Fig. 18. Component planes obtained by the SOM for student survey No. 2.

largest for both networks in Figures 21(c1) and (c2). These values are exactly the same as those obtained by the normal-sized network, though the range of the importance values becomes smaller as the network size grows larger. In addition, when the network is large, in Figure 21(b4), a clear boundary is diagonally located. On the other hand, in the U-matrix in Figure 21(b1), this boundary cannot be seen. The experimental results suggest that the characteristics obtained by our methods are greatly independent of the size of the network.

3.2 Voting attitude problem

In this experiment, we use the voting attitude data from the machine learning data base[4]. The data set includes votes for each of the U.S. House of Representatives Congresspersons on 16 key votes, and the number of input patterns is 435. Figures 22(a) and (b) show the U-matrix and labels for the voting attitude problem obtained by the conventional SOM. As can be seen in Figure 22(a), a boundary in warmer colors can be detected in the middle of the U-matrix. The labels in Figure 22(b) show that input patterns are classified into two groups, Republicans and Democrats, in the middle of the map. Figure 23 shows component planes along the 16 input variables. As later shown in Figure 25, the input variables No.5, No.8 and No.4 have the larger values of information. They clearly represent two groups in the component planes. Figure 24(a) shows the estimated information as a function of σ for six different values of the parameter σ. We notice two important points. First, the parameter σ is decreased when the parameter ϵ is increased from 0.1 to 0.3, and then the parameter σ is increased. Second, the estimated information is not necessarily increased as the parameter ϵ is increased. Figure 24(b) shows the estimated information as a function of the parameter ϵ. As can be seen in the figure, the information is increased to a maximum point when the parameter ϵ is increased to 0.3, and

[4]http://www1.ics.uci.edu/ mlearn/MLRepository.html

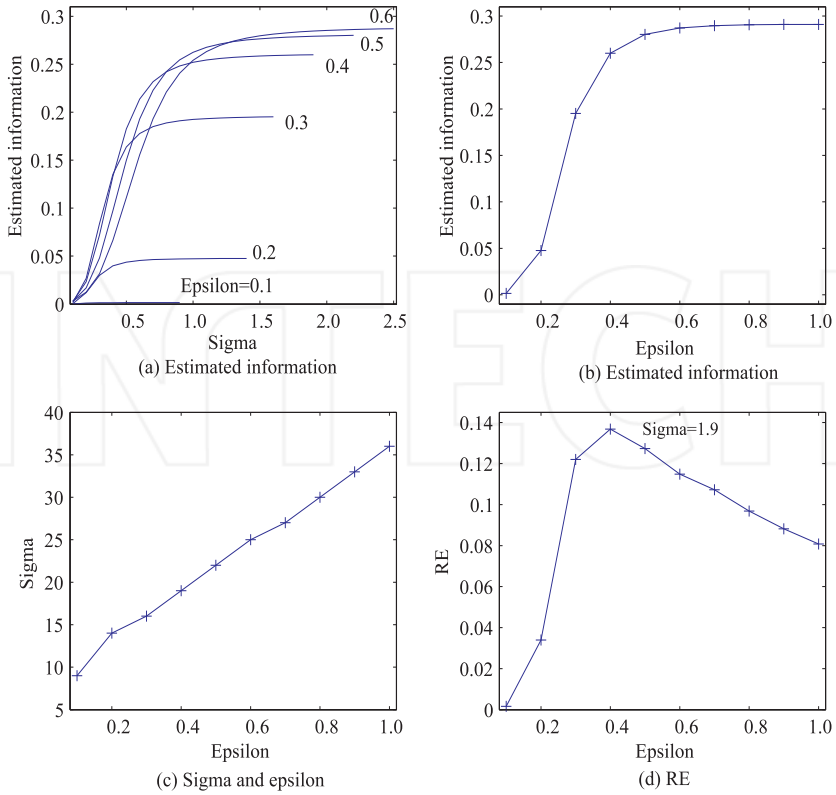


Fig. 19. Information as a function of the parameter σ (a) and the parameter ϵ (b). Optimal values of the parameter σ as a function of the parameter ϵ (c). The ratio RE as a function of the parameter ϵ (d).

then information slowly decreases. Figure 24(c) shows a relation between the parameter σ and the ϵ. When the parameter ϵ is 0.2 and 0.3, the parameter σ takes the lowest value. Figure 24(d) shows the ratio of information to the parameter σ. As can be seen in the figure, the ratio takes a maximum value when the parameter ϵ is 0.3.

Figure 25 shows the importance of 16 input variables. As the parameter ϵ gets larger, the range of values of the importance is increased. Finally, when the parameter ϵ becomes 0.3, the range of the importance becomes the largest. As can be seen in the figure, variables No. 5, No. 8 and No. 4 have larger importance. Figures 26(a) and (b) show the probability $p(j)$ and $p(j;k=5)$ with the optimal values of the parameters. In the figures, a clear boundary in warmer colors can be seen in the middle of the map.

Figure 27 shows results when the network size is small (a) and big (b). Figures 27(a1) and (b1) shows the U-matrices for the small size and big size, respectively. As can be seen in the figures, for the small network, a clear boundary in warmer colors clearly shows two groups, while for the large network, smaller boundaries are scattered on the U-matrix and clear boundaries disappear. Figures 27(a2) and (b2) show the ratio RE as a function of the parameter ϵ. Though the values for the large network become smaller, almost the same tendency can be detected.

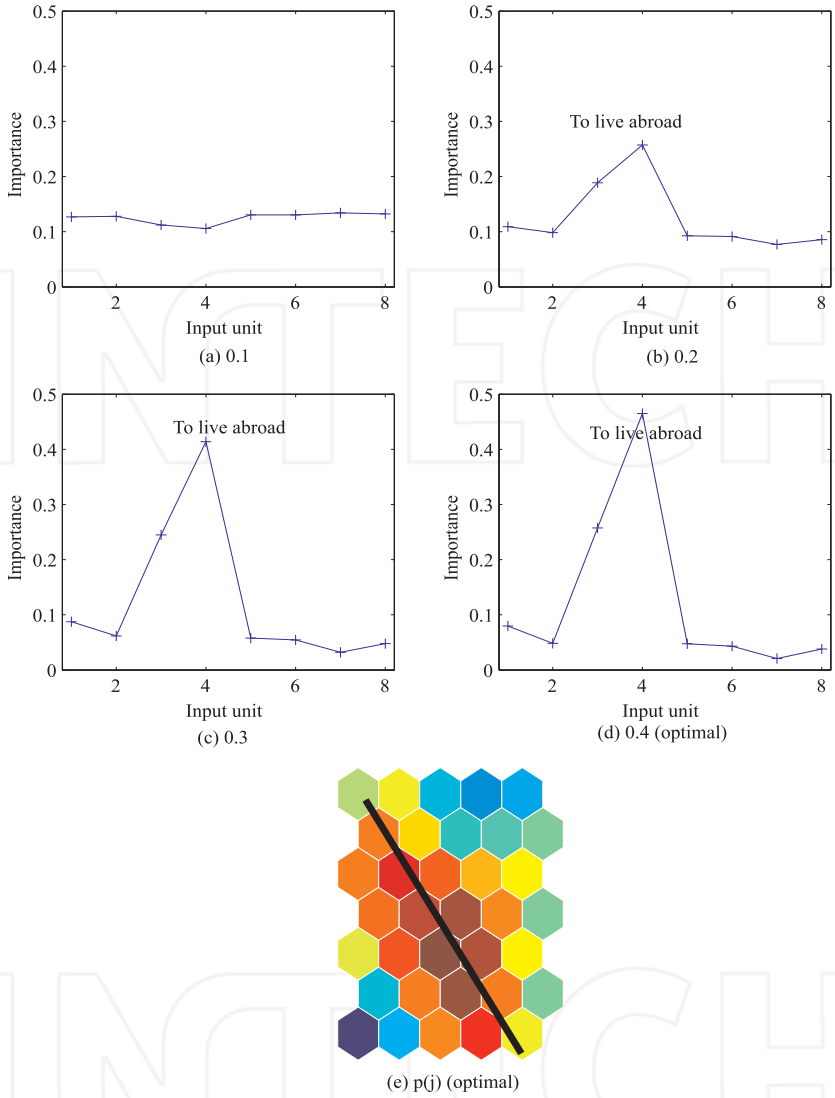


Fig. 20. Estimated importance for four different values of the parameter ϵ (a)-(d) and the probability $p(j)$.

Figures 27(a3) and (b3) show the values of the importance. As can be seen in the figure, the values for the large network become smaller, but the tendency of the importance remains the same. Figures 27(a4) and (b4) show the values of $p(j)$ for the small and large size. As can be seen in the figure, even if the network is large, a boundary in the middle of the map can be seen, while for the small size, a clear boundary can be generated.

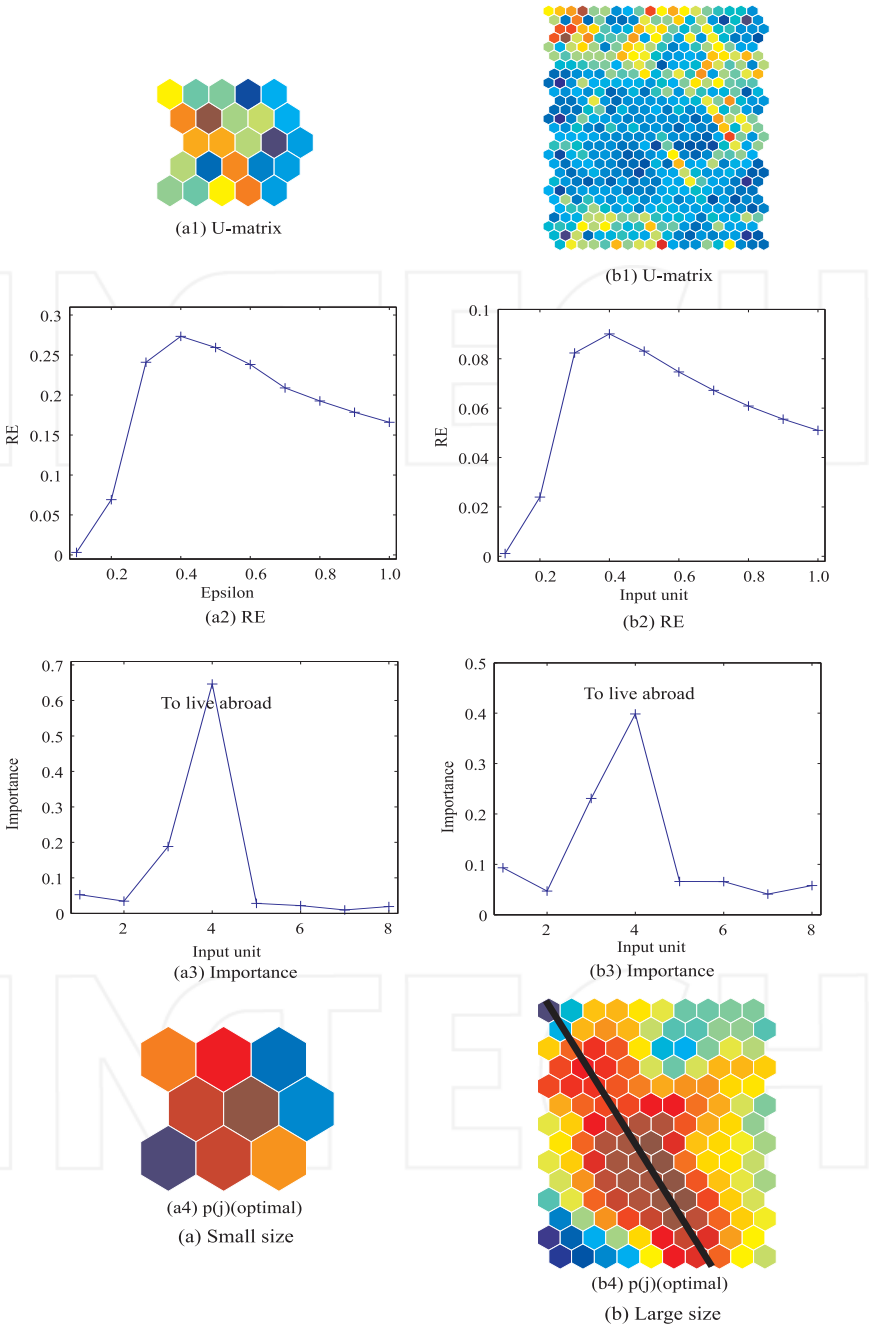


Fig. 21. U-matrices (1), the ratio RE (2), the values of importance (3) and $p(j)$ (4) for the small-sized (a) and large-sized network (b).

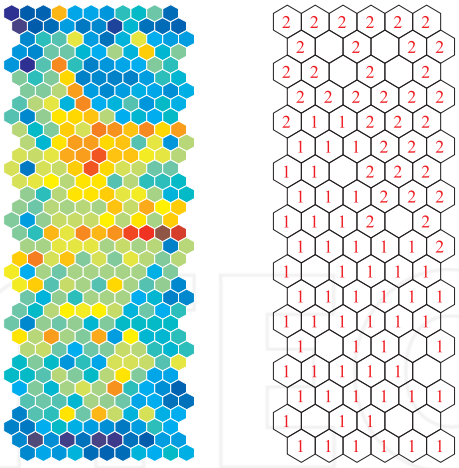


Fig. 22. Original data (a), U-matrix (b) and labels (c) obtained by the SOM for the voting attitude problem.

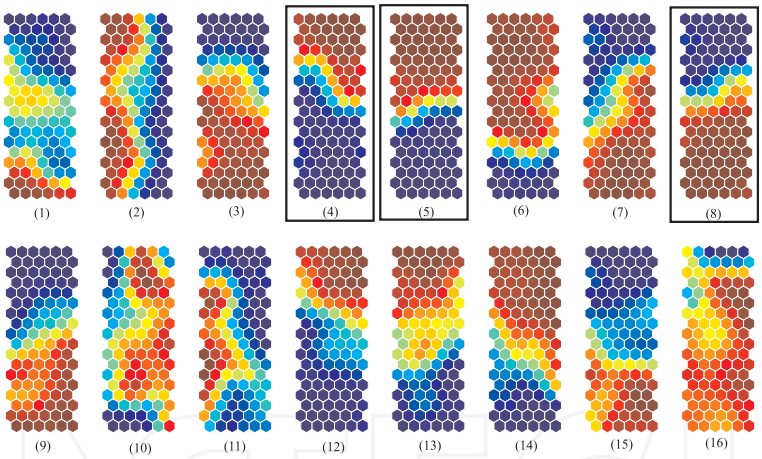


Fig. 23. Component planes obtained by the SOM for the voting attitude problem obtained.

3.3 Discussion

3.3.1 Validity of experimental results

In this paper, we have proposed a new type of information-theoretic method to measure the importance of input variables. The importance of input variables is approximated by mutual information, focusing upon a specific input variable. Then, using this importance, the information content of input variables is computed. As the information gets larger, the number of important input variables becomes smaller. Thus, we try to increase this information as much as possible. We have applied the method to four problems, namely, a symmetric data set, two actual student survey data sets and the voting attitude problem. Experimental results have clarified four points, namely, the number of important input variables, the determination of optimal information, the independency from the network size

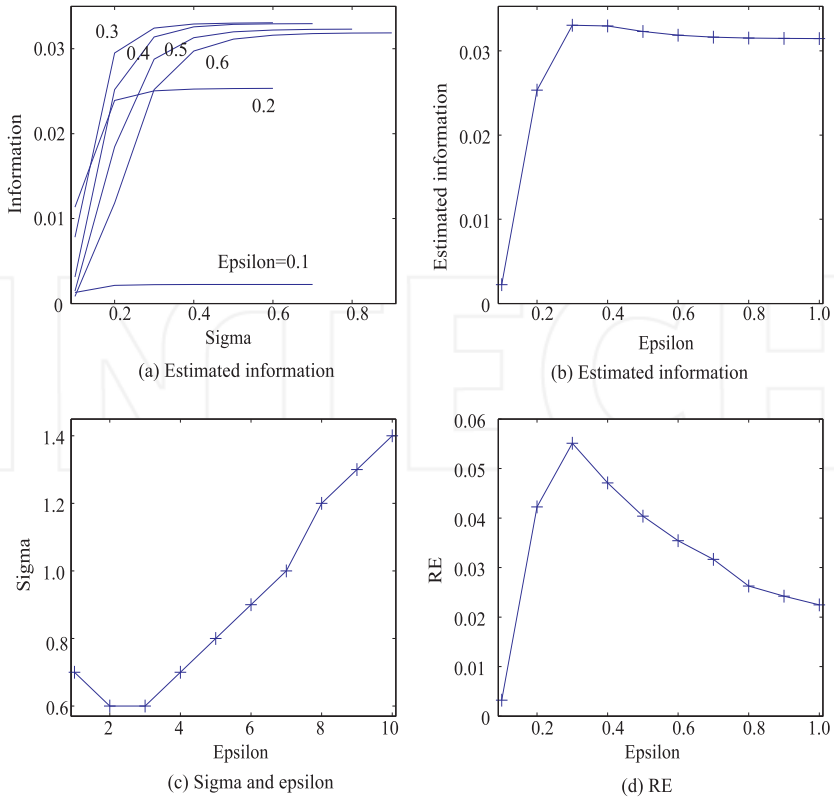


Fig. 24. Information as a function of the parameter σ (a) and the parameter ϵ (b). Optimal values of the parameter σ as a function of the parameter ϵ (c) for the voting attitude problem. The ratio RE as a function of the parameter ϵ (d).

and relations to variance. First, experimental results have confirmed that the smaller number of important input variables is detected for two data sets of an actual student survey. In student survey No. 1, the input variable representing "teachers' attitude" shows by far the largest value of importance. On the other hand, in student survey No. 2, the input variable "to live abroad" has by far the largest value of importance. These input variables are measured by the mutual information between competitive units and input patterns. The large importance also means large mutual information, meaning that input variables with large importance play more important roles to make competitive units fire in more organized ways. Second, the optimal amount of information can be estimated. To determine the optimal value of the estimated information, we have proposed the ratio of the estimated information to the spread parameter σ

$$RE = \frac{EI(\sigma,\epsilon)}{\sigma}. \tag{15}$$

When the estimated information is larger, the number of important input variables is smaller. Because the number of important variables must be as small as possible, we must increase the estimated information as much as possible. In addition, to make the ratio large, we

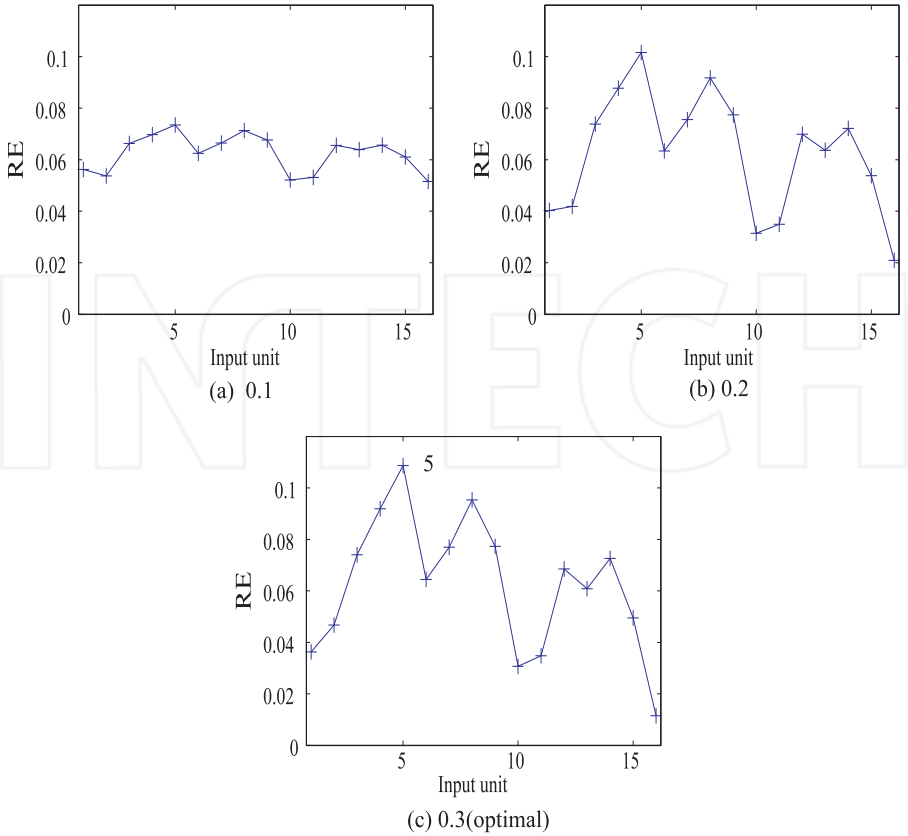


Fig. 25. Estimated importance for four different values of the parameter ϵ for the voting attitude problem.

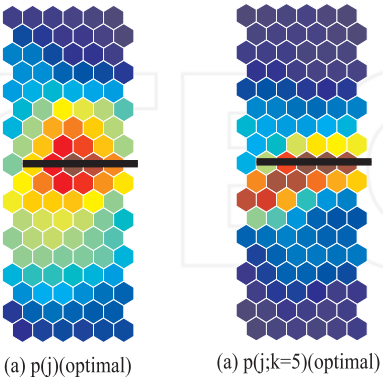


Fig. 26. The estimated probability $p(j)$ (a) and $p(j;k=5)$ for the voting attitude problem.

must decrease the parameter σ as much as possible. The parameter σ is related to the amount of mutual information between competitive units and input patterns. To increase

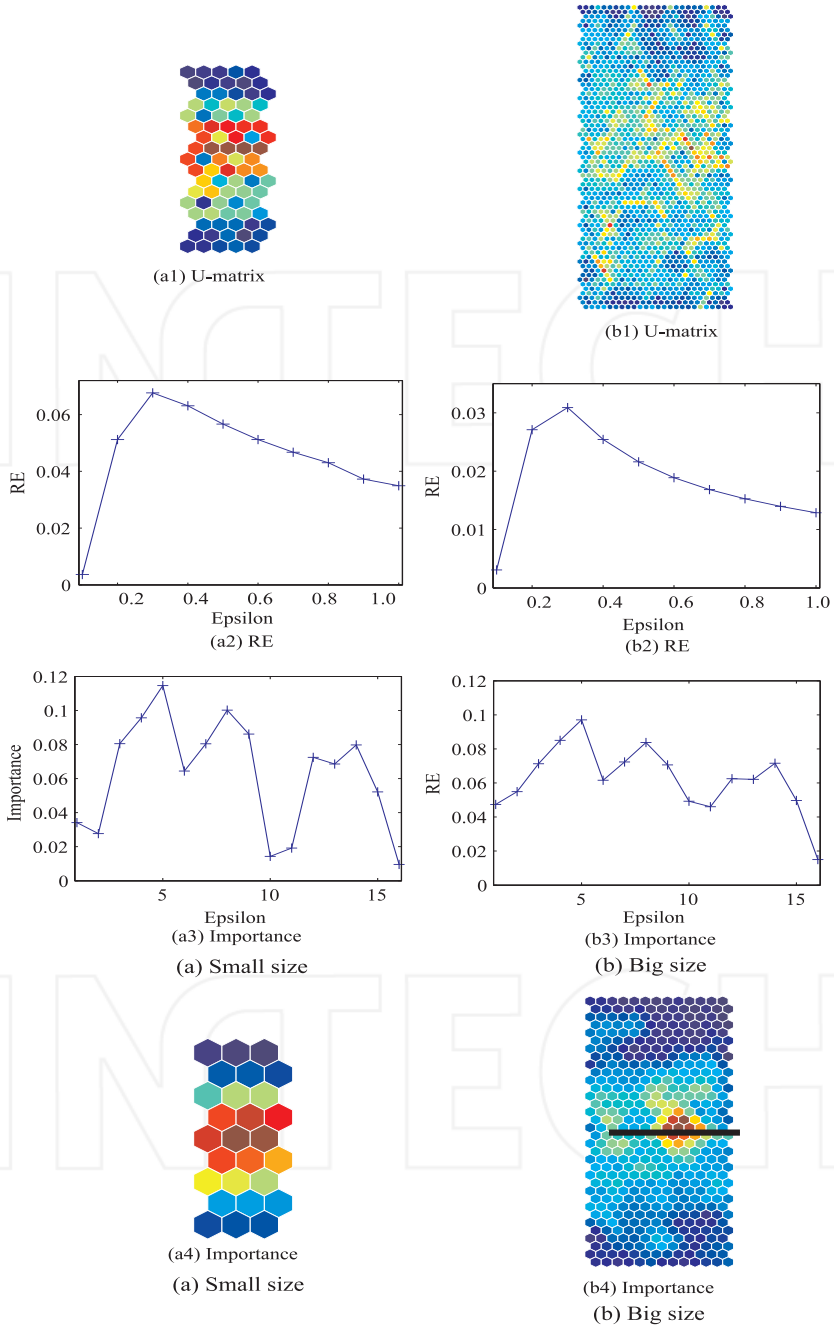


Fig. 27. U-matrices (1), the ratio RE (2), the values of importance (3) and $p(j)$ (4) for the small-sized (a) and large-sized network (b).

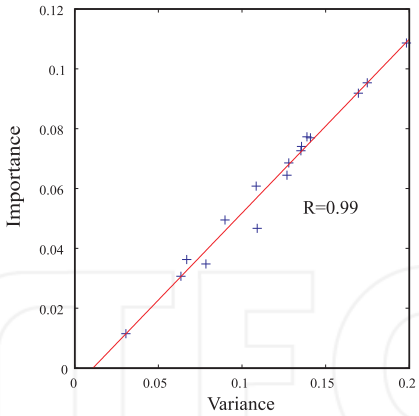


Fig. 28. Relations between importance and variance of connection weights for the voting attitude problem.

this mutual information, we must decrease the parameter σ as much as possible. Thus, the ratio is a reasonable criterion to obtain the optimal amount of estimated information. Four experimental results have shown clearly that the ratio can show peak values by which we can specify explicitly the values of two parameters. The validity of these optimal parameter values is endorsed by our obtaining the largest range of importance when the ratios have the maximum values. Third, the characteristics obtained by the optimal values of the parameters have been confirmed to be independent of the network size. To examine this property, we prepared larg- and small-sized networks for the two student survey data sets and the voting attitude problem. We observed that, even if the network size is changed, the optimal parameter values are exactly the same for the three problems. Though the values of the importance tend to be smaller when the network size is larger, the obtained values of the importance are completely the same independently of the network size, except for the range of the importance. This property of independency from the network size is important for the application to practical problems. Fourth, we have observed that the importance is closely correlated with the variance of connection weights. Figure 28 shows relations between importance and variance of connection weights for the voting attitude problem. As can be seen in the figure, the correlation coefficient between the two measures amounts to 0.99 for the problem. Generally speaking, the variance is a kind of measure to represent information in input patterns; this fact of a high correlation has supported the validity of our importance measure.

3.3.2 Limitation of the method

In our experiments, we have observed three problems or limitations of our method, namely, a problem of the parameter setting, information maximization for competitive units and the wrapper method. First, the parameters are precisely determined by examining the ratio of the estimated information to the parameter σ. However, to obtain the best parameter values, we must change the two parameters σ and ϵ extensively. Thus, this extensive search for the parameters may be a burden when we try to apply the method to practical and large-scaled problems. We think that, to reduce the computational complexity, we need to simplify the parameter setting as much as possible. Second, we have the problem of information

maximization for competitive units. As mentioned in the introduction section, because we have focused on the importance of input units, information in input units is more strongly maximized compared with information in competitive units. However, mutual information between competitive units and input patterns shows a kind of organization of competitive units. As this mutual information is more increased, more organized patterns of competitive units are generated. Because we focus upon information maximization in input units, we have paid restrained attention to the increase in this mutual information. Thus, we need to maximize mutual information in competitive units more strongly in addition to information maximization in input units. The third problem is closely related to the second one. Our method is a kind of wrapper method; we can use any learning method for learning, and then we use the information-theoretic method. In our method, we suppose two types of information, namely, mutual information between competitive units and input patterns. If it is possible to maximize two types information simultaneously, the final network is one with much information included in input units as well as competitive units. To realize this situation, we must train a network in learning, while increasing two types of information. Thus, we need an embedded system in which both learning and information maximization are simultaneously applied.

3.3.3 Possibility of the method

One of the main possibilities of our method can be summarized by two points, namely, its simplicity and the possibility of new learning. First, the importance is actually defined by focusing upon a specific input pattern. This means that the measure of information-theoretic importance can be applied to any elements or components of a network, such as connection weights, competitive units and so on. All we have to do is focus upon a specific element or component and compute mutual information between competitive units and input patterns. In particular, the applicability to the components in which several elements are combined with each other is one of the main possibilities or potentialities of our method. Second, our method opens up a new perspective for learning. In the present study, we have restricted ourselves to the detection of the importance of input variables. Now that the importance can be determined by the mutual information between competitive units and input patterns, the obtained information on the importance of input variables can be used to train networks. In that case, the learning can be done with due consideration to the importance of input variables.

4. Conclusion

In this chapter, we have proposed a new type of information-theoretic method to estimate the importance of input variables. This importance is estimated by mutual information between input patterns and competitive units, with attention paid to the specific input units. As this mutual information becomes larger, more organized competitive units are generated by the input units. Then, the information content of input variables is computed by using the importance. When this information is maximized, only one input variable plays an important role. Thus, we should increase this information as much as possible to obtain a smaller number of important input variables. To increase this information on input variables and mutual information between competitive units and input patterns, we have proposed the ratio RE of the information to the parameter ϵ to determine an optimal state. As this ratio is increased, the information on input variables is naturally increased and the corresponding mutual information between competitive units and input patterns is increased. We applied the

method to four problems, namely, a symmetric data, two data sets of actual of student surveys and the voting attitude problem. In all the problems, we have shown that, by maximizing the ratio, we can have the largest values of importance for easy interpretation. In addition, these values of the importance are independent of the network size. Finally, experimental results have confirmed that the importance of input variables is strictly correlated with the variance of connection weights. Though the parameter tuning requires an extensive search procedure to find an optimal state of information, these results certainly show that our information-theoretic method can be applied to many practical problems, because the importance can be determined based upon an explicit criterion and its meaning assured in terms of the variance of connection weights.

5. Acknowledgment

The author is very grateful to Kenta Aoyama and Mitali Das for their valuable comments.

6. References

Andrews, R., Diederich, J. & Tickle, A. B. (1993). Survey and critique of techniques for extracting rules from trained artificial neural networks, *Knowledge-based systems* 8(6): 373–389.

Barakat, N. & Diederich, J. (2005). Eclectic rule-extraction from support vector machines, *International Journal of Computational Intelligence* 2(1): 59–62.

Belue, L. M. & K. W. Bauer, J. (1995). Determining input features for multiplayer perceptrons, *Neurocomputing* 7: 111–121.

Garcez, A. S. d., Broda, K. & Gabbay, D. (2001). Symbolic knowledge extraction from trained neural networks: a sound approach, *Artificial Intelligence* 125: 155–207.

Gorman, R. P. & Sejnowski, T. J. (1988). Analysis of hidden units in a layered network trained to classify sonar targets, *Neural Networks* 1: 75–89.

Guyon, I. & Elisseeff, A. (2003). An introduction to variable and feature selection, *Journal of Machine Learning Research* 3: 1157–1182.

Kahramanli, H. & Allahverdi, N. (2009). Rule extraction from trained adaptive networks using artificial immune systems, *Expert Systems with Applications* 36: 1513–1522.

Kamimura, R. (2003a). Information theoretic competitive learning in self-adaptive multi-layered networks, *Connection Science* 13(4): 323–347.

Kamimura, R. (2003b). Information-theoretic competitive learning with inverse Euclidean distance output units, *Neural Processing Letters* 18: 163–184.

Kamimura, R. (2003c). Progressive feature extraction by greedy network-growing algorithm, *Complex Systems* 14(2): 127–153.

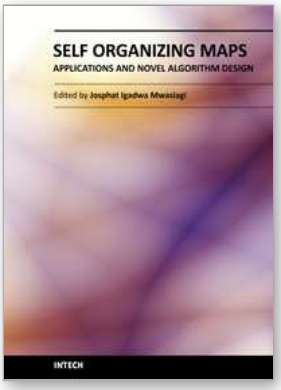
Kamimura, R. (2003d). Teacher-directed learning: information-theoretic competitive learning in supervised multi-layered networks, *Connection Science* 15: 117–140.

Kamimura, R. (2007). Information loss to extract distinctive features in competitive learning, *Proceedings of IEEE Conference on Systems, Man, and Cybernetics*, pp. 1217–1222.

Kamimura, R. (2008a). Conditional information and information loss for flexible feature extraction, *Proceedings of the international joint conference on neural networks(IJCNN2008)*, pp. 2047–2083.

Kamimura, R. (2008b). Feature detection and information loss in competitive learning, *Proceedings of the international conference on soft computing and intelligent systems and the international symposium on advanced intelligent systems(SCIS and ISIS2008)*,

pp. 1144–1148.
Kamimura, R. (2008c). Feature discovery by enhancement and relaxation of competitive units, *Intelligent data engineering and automated learning-IDEAL2008(LNCS)*, Vol. LNCS5326, Springer, pp. 148–155.
Kamimura, R. (2009). Enhancing and relaxing competitive units for feature discovery, *Neural Processing Letters* 30(1): 37–57.
Kamimura, R. & Kamimura, T. (2000). Structural information and linguistic rule extraction, *Proceedings of ICONIP-2000*, pp. 720–726.
Kamimura, R., Kamimura, T. & Uchida, O. (2001). Flexible feature discovery and structural information control, *Connection Science* 13(4): 323–347.
Kaski, S., Nikkila, J. & Kohonen, T. (1998). Methods for interpreting a self-organized map in data analysis, *Proceedings of European Symposium on Artificial Neural Networks*, Bruges, Belgium.
Kohonen, T. (1988). *Self-Organization and Associative Memory*, Springer-Verlag, New York.
Kohonen, T. (1995). *Self-Organizing Maps*, Springer-Verlag.
Mak, B. & Munakata, T. (2002). Rule extraction from expert heuristics: a comparative study of rough sets with neural network and ID3, *European journal of operational research* 136: 212–229.
Mao, I. & Jain, A. K. (1995). Artificial neural networks for feature extraction and multivariate data projection, *IEEE Transactions on Neural Networks* 6(2): 296–317.
Petersen, M., Talmoon, J. L., Hasman, A. & Ambergen, A. W. (1998). Assessing the importance of features for multi-layer perceptrons, *Neural Networks* 11: 623–635.
Polzlbauer, G., Dittenbach, M. & Rauber, A. (2006). Advanced visualization of self-organizing maps with vector fields, *Neural Networks* 19: 911–922.
Rumelhart, D. E., Hinton, G. E. & Williams, R. (1986). Learning internal representations by error progagation, *in* D. E. Rumelhart & G. E. H. et al. (eds), *Parallel Distributed Processing*, Vol. 1, MIT Press, Cambridge, pp. 318–362.
Steppe, J. M. & K. W. Bauer, J. (1997). Feature saliency measures, *Computers and Mathematics with Applications* 33(8): 109–126.
Tasdemir, K. & Merenyi, E. (2009). Exploiting data topology in visualizations and clustering of self-organizing maps, *IEEE Transactions on Neural Networks* 20(4): 549–562.
Thrun, S. (1995). Extracting rules from artificial neural networks with distributed representations, *Advances in Neural Processing Systems*.
Towell, G. G. & Shavlik, J. W. (1993). Extracting refined rules from knowledge-based neural networks, *Machine learning* 13: 71–101.
Tsukimoto, H. (2000). Extracting rules from trained neural networks, *IEEE Transactions on Neural Networks* 11(2): 377–389.
Ultsch, A. (2003). U*-matrix: a tool to visualize clusters in high dimensional data, *Technical Report 36*, Department of Computer Science, University of Marburg.
Ultsch, A. & Siemon, H. P. (1990). Kohonen self-organization feature maps for exploratory data analysis, *Proceedings of International Neural Network Conference*, Kulwer Academic Publisher, Dordrecht, pp. 305–308.
Vesanto, J. (1999). SOM-based data visualization methods, *Intelligent Data Analysis* 3: 111–126.



Self Organizing Maps - Applications and Novel Algorithm Design
Edited by Dr Josphat Igadwa Mwasiagi

ISBN 978-953-307-546-4
Hard cover, 702 pages
Publisher InTech
Published online 21, January, 2011
Published in print edition January, 2011

Kohonen Self Organizing Maps (SOM) has found application in practical all fields, especially those which tend to handle high dimensional data. SOM can be used for the clustering of genes in the medical field, the study of multi-media and web based contents and in the transportation industry, just to name a few. Apart from the aforementioned areas this book also covers the study of complex data found in meteorological and remotely sensed images acquired using satellite sensing. Data management and envelopment analysis has also been covered. The application of SOM in mechanical and manufacturing engineering forms another important area of this book. The final section of this book, addresses the design and application of novel variants of SOM algorithms.

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821