Progressive Feature Extraction with a Greedy Network-growing Algorithm

Ryotaro Kamimura*
Information Science Laboratory and
Future Science and Technology Joint Research Center,
Tokai University,
1117 Kitakaname Hiratsuka Kanagawa 259-1292, Japan

In this paper, a new information theoretic method called the greedy network-growing algorithm is proposed. The method is called "greedy," because a network with this algorithm grows while absorbing as much information as possible from outside. The method is based upon information theoretic competitive learning and can solve the fundamental problems inherent in competitive learning, such as the dead neurons and inappropriate number of neurons problems. The new model can grow networks by repeatedly maximizing information content and by gradually extracting salient features from input patterns. Because the new model can cope with inappropriate feature detection in the early stage of learning, extracted features should cover most of the input patterns. The new method is applied to political data analysis, medical data analysis, and information science education analysis. Experimental results confirm that the new method can acquire significant information and that more explicit features can be extracted.

1. Introduction

In this paper, we propose a novel approach to information acquisition in neural networks. The method is called the greedy network-growing algorithm (GNGA) because a network with this algorithm grows while absorbing as much information as possible from outside.[1] This new network-growing method is based upon information theoretic competitive learning. It has many characteristics different from conventional competitive and network-growing learning methods. We discuss the characteristics of the GNGA in three perspectives: (1) this is a new competitive learning method; (2) it is a new type of network-growing algorithm that repeatedly maximizes information content; and (3) the

*Electronic mail address: ryo@cc.u-tokai.ac.jp.

[1] We have already proposed a similar network-growing algorithm [1]. However, the method in [1] is less greedy than the present one because the number of connections to absorb information was limited to speed up learning.

method can extract salient features explicitly and flexibly. Let us discuss these three points in more detail.

First, the GNGA is based upon a new competitive learning technique in which competition processes are simulated by maximizing mutual information between input patterns and competitive units. Competitive learning is one of the most popular methods in neural networks, and is used for pattern classification as well as for regularity detection [2]. Though conventional competitive learning methods have turned out to be useful and powerful in many applications, they have shown many serious problems. When conventional competitive learning methods are applied to complex problems underutilized neurons or dead neurons can cause serious problems. The determination of the appropriate number of competitive units is another serious problem, because conventional competitive learning techniques significantly degrade with an inappropriate number of classes. This means that unless the number of competitive units is appropriately given before learning performance significantly degrades. There have been many attempts to overcome these problems, such as leaky learning [2, 3], conscience learning [4], frequency-sensitive learning [5], rival penalized competitive learning [6], and lotto-type competitive learning [7]. However, these methods do not necessarily give satisfactory solutions to the problems. On the other hand, the GNGA can give clear solutions to these problems. The dead neurons can be suppressed in our method, because it includes the entropy maximization of competitive units. By maximizing the entropy of competitive units, all competitive units tend to be equally used. In addition, because neurons can be added during learning, the appropriate number of neurons can easily be determined.

Second, the GNGA is a new type of information theoretic method that always maximizes information during network growth [8]. Network constructive or growing approaches have been used extensively, because constructive approaches are computationally economical and have a good chance of finding smaller networks for given problems. For example, in supervised learning, one of the most popular constructive methods is the cascade correlation network [9] that grows a network by maximizing the covariance between the outputs of newly recruited hidden units and the errors of the network outputs [9, 10]. In the popular back propagation method, a technique called constructive back propagation was developed in which only connections to a newly recruited hidden unit are updated [11]. In the RBF networks, Fritzke tried to develop incremental RBF networks for the fast learning method [12]. In self-organizing maps, Fritzke [13–15] tried to extend Kohonen's self-organizing map into a network growing model. This model has been extended to the self-organized tree algorithm [16, 17].

Contrary to these conventional network-growing methods, we use an information theoretic method to grow networks, and thus, it is con-

siderably different from those mentioned. Information theoretic approaches have been introduced in various ways into neural computing; for example, a principle of maximum information preservation [18–20], a principle of minimum redundancy [21], and spatially coherent feature detection [22, 23]. In conventional information theoretic methods, maximum possible information is determined before learning. It is impossible to increase information beyond a predetermined point, and thus, information theoretic methods have not been used for network growing algorithms. However, when considering the self-organization processes in living systems, information capacity should be increased with growth. We can intuitively see that, as living systems develop, their complexity becomes larger; that is, their information capacity is larger. In GNGA, the number of competitive units can be increased during learning, and accordingly, the maximum possible information is increased. Because information is defined directly for the competitive units, it can be increased, as the number of competitive units becomes larger. This permits networks to adapt flexibly to new input patterns.

Third, this method aims to flexibly extract salient features. In a previously developed network-growing algorithm [1] we tried to develop a growing method that was as efficient as possible in computation. For this purpose, in maximizing information content, only connections into a new competitive unit are updated, as is the case with the constructive back propagation method [11]. This means that networks always update connections into one competitive unit, except during the initial cycle. Thus, the method from [1] is very efficient and is expected to be applied to many large-scale problems. However, one of the serious problems is that, if networks extract inadequate features at the beginning of the growth cycles, feature extraction during later stages degrades, because all connections after learning are frozen in the algorithm. To overcome this shortcoming, we relax the severe restriction of the earlier algorithm in which all previous connections must be fixed in the later learning stage. By this relaxation, we have an algorithm in which inappropriate feature extraction at the beginning of learning is gradually remedied at the expense of computational efficiency. Thus, the GNGA can be used to analyze actual complex data.

In section 2, we present the new algorithm, comparing it with our previous model. In section 3, we show that similar results can be obtained with either algorithm for a political data problem. In section 4, we apply the method to artificial data whose features cannot be extracted by using our previous model. In section 5, we apply our new method to an actual data analysis, that is, the analysis of information science education where we try to show that the GNGA can extract salient features more explicitly than can conventional competitive learning or multivariate analyses.

2. Growing network

2.1 Growing algorithm

In the GNGA, we suppose that a network attempts to absorb as much information as possible from the outer environment, because in the outer environment there are many destructive forces against artificial systems. Therefore, we assume at least that artificial systems must absorb as much information as possible as defense against the destructive forces. To absorb information from the outer environment, the systems gradually increase their complexity until no more complexity is needed. When no more additional information can be obtained, the network recruits another unit, and then it again tries to maximally absorb information (Figure 1). This general idea of networks with a GNGA has been realized for neural networks in [1]. The method in [1] aims at developing a growing mechanism that is as efficient as possible in computation. Figure 2 shows an actual network architecture of a greedy algorithm. Figure 2(a) shows a process of growing by our previous method. Figure 2(a1) represents an initial state of information maximization in which only two competitive units are used. We need at least two competitive units, because our method aims at making neurons compete with each other. First, information is increased as much as possible with these two competitive units. When it becomes difficult to increase information, the first cycle of information maximization is finished, and all connections are frozen. In this case, just one unit wins the competition, while the other loses. Then, a new competitive unit is added, as shown in

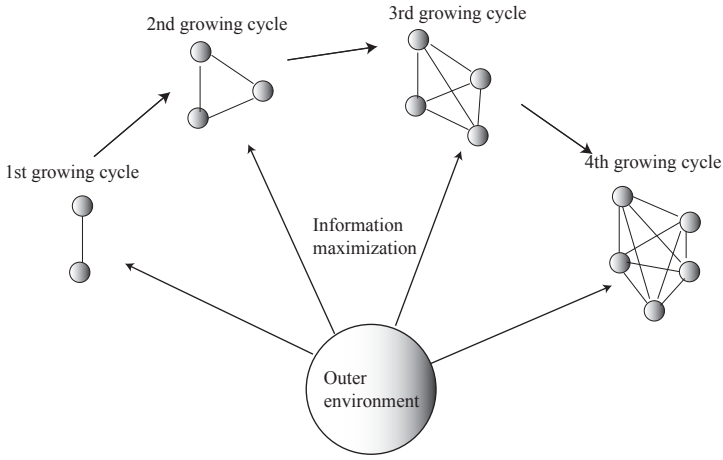

Figure 1. A process of greedy network growing in which the network grows by absorbing maximum information from the outer environment.

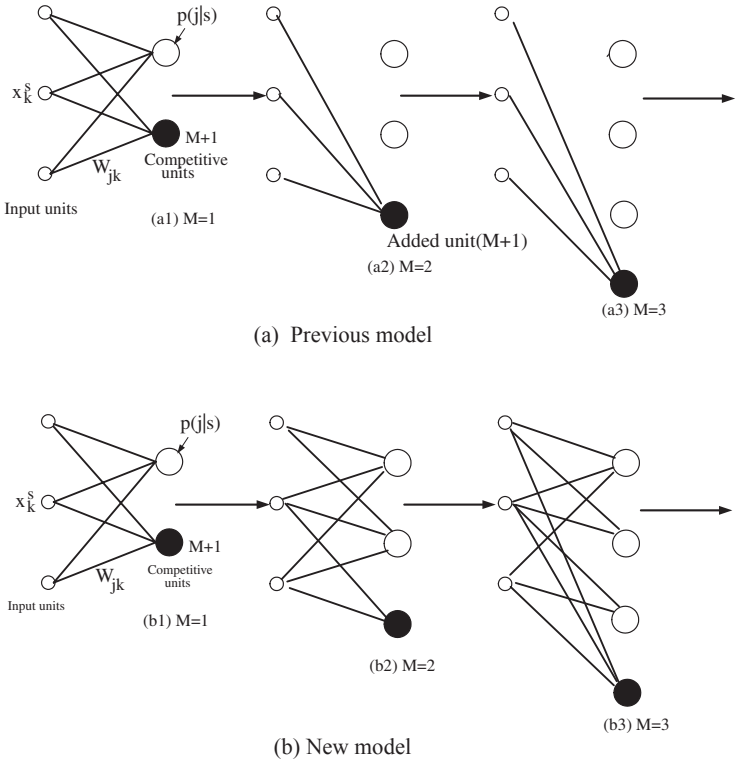

Figure 2. A process of network growing by (a) the method from [1] and (b) the GNGA. In the conventional method (a), only variable connections are shown.

Figure 2(a2). Since connections into the previous two competitive units are frozen, connections into the new competitive unit must be changed to maximize information. Then, another competitive unit is added (Figure 2(a3)). At this stage, connections into the previous three competitive units are fixed, and only connections into the new competitive unit are adjusted to increase information as much as possible. These processes continue until no more increase in information content is possible. The algorithm is computationally very efficient, because networks only have to update connections into one competitive unit. If it is possible to extract features gradually, this will be one of the most efficient methods ever developed for feature extraction. We have found that, for some problems, this method can extract basic features gradually as the number of growing cycles is increased [1].

However, once learning is finished in a particular growth cycle, all connections at that point are frozen in this algorithm. These frozen connections prevent networks from extracting appropriate features in later cycles. For example, if networks capture inadequate features at

the beginning of learning, they fail to extract appropriate features in later cycles. One of the simplest ways to overcome this shortcoming is to update all connections in every growth cycle at the expense of computational time. Thus, we propose a novel GNGA in which all connections are updated to maximize information content at the expense of computational efficiency. Figure 2(b) shows a growing process using the present model. As can be seen in Figure 2(b1)–(b3), all connections are updated at every growth cycle. This allows the network to adjust for inappropriate connections during later stages.

2.2 Learning rule

Now we can compute information content in a neural system. We consider information content stored in competitive unit activation patterns. For this purpose, let us define information to be stored in a neural system. Information stored in the system is represented by a decrease in uncertainty [24]. Uncertainty decrease, that is, information $I(t)$ at the tth epoch, is defined by

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

where $p(j;t)$, $p(s)$, and $p(j|s;t)$ denote the probability of the jth unit in a system at the nth learning cycle in the tth growth cycle, the probability of the sth input pattern and the conditional probability of the jth unit, given the sth input pattern at the tth epoch, respectively.

Let us present update rules to maximize information content during every stage of learning. For simplicity, we consider the Mth growing cycle, and t denotes the cumulative learning epochs throughout the growth cycles. As shown in Figure 2, a network at the tth epoch is composed of input units x_k^s and competitive units $v_j^s(t)$. The jth competitive unit receives a net input from the input units, and an output from the jth competitive unit can be computed by

$$v_j^s(t) = f\left(\sum_{k=1}^{L} w_{jk}(t)x_k^s\right), \quad (2)$$

where $w_{jk}(t)$ denotes connections from the kth input unit to the jth competitive unit, and the sigmoid activation function $f(u) = 1/(1 + \exp(-u))$ is used. The conditional probability $p(j \mid s;t)$ at the tth epoch is computed by

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

where M denotes the Mth growth cycle. Since input patterns are supposed to be given uniformly to networks, the probability of the jth

competitive unit is computed by

$$p(j;t) = \frac{1}{S}\sum_{s=1}^{S} p(j \mid s;t). \tag{4}$$

Information $I(t)$ is computed by

$$I(t) = -\sum_{j=1}^{M+1} p(j;t)\log p(j;t) + \frac{1}{S}\sum_{s=1}^{S}\sum_{j=1}^{M+1} p(j \mid s;t)\log p(j \mid s;t). \tag{5}$$

As information becomes larger, specific pairs of input patterns and competitive units become strongly correlated. Differentiating information with respect to input–competitive connections $w_{jk}(t)$, we have

$$\begin{aligned}\frac{\partial I(t)}{\partial w_{jk}(t)} = &-\sum_{s=1}^{S}\left(\log p(j;t) - \sum_{m=1}^{M+1} p(m \mid s;t)\log p(m;t)\right) Q_{jk}^{s}(t) \\ &+\sum_{s=1}^{S}\left(\log p(j \mid s;t) - \sum_{m=1}^{M+1} p(m \mid s;t)\log p(m \mid s;t)\right) Q_{jk}^{s}(t)\end{aligned} \tag{6}$$

where

$$Q_{jk}^{s}(t) = \frac{1}{S}p(j \mid s;t)(1 - v_j^s(t))x_k^s. \tag{7}$$

Thus, we have the update rules:

$$\begin{aligned}\Delta w_{jk}(t) = &-\sum_{s=1}^{S}\left(\log p(j;t) - \sum_{m=1}^{M+1} p(m \mid s;t)\log p(m;t)\right) Q_{jk}^{s}(t) \\ &+\sum_{s=1}^{S}\left(\log p(j \mid s;t) - \sum_{m=1}^{M+1} p(m \mid s;t)\log p(m \mid s;t)\right) Q_{jk}^{s}(t).\end{aligned} \tag{8}$$

Finally, we should state how to stop learning and add a new competitive unit. Let $I(t)$ denote information content computed at the nth learning cycle in the tth growth cycle. Relative increase in information $R(t)$ is computed by

$$R(t) = \frac{|I(t) - I(t-1)|}{I(t-1)}, \tag{9}$$

where $t = 1, 2, 3, \ldots$. If $R(t)$ is less than a certain point ϵ for three consecutive epochs, the first information maximization process is finished.

Information maximization realizes the processes of competitive units. When maximizing mutual information, the conditional entropy

$$-\sum_{s} p(s)\sum_{j} p(j \mid s)\log p(j \mid s)$$

should be as small as possible. In this case, a competitive unit is turned on for a specific input pattern. Conditional entropy minimization does not exclude a case where a competitive unit is always turned on for any input pattern, corresponding to a dead neuron. However, because entropy $-\sum_j p(j) \log p(j)$ must be maximized, and competitive units must be equally utilized on average, it is impossible for a competitive unit to respond to all input patterns. When mutual information is maximized, that is, entropy is maximized and conditional entropy is minimized, different competitive units tend to respond to different input patterns.

3. Political data analysis

We attempted to classify congressmen in terms of their voting attitude. The data were represented as a qualitative matrix of United States congressmen with their voting attitude toward 19 environmental bills [25]. The first eight congressmen are Democrats, while the latter seven are Republicans. In the data, 1, 0, and 0.5 represent *yes*, *no*, and *undecided*, respectively (Table 1). By a cluster analysis [25],[2] it was confirmed that these data can be classified into two groups except for Republican No. 6(14), who responds to the bills according to the line of Democrats. In addition, Democrat No. 7 was misclassified as a Republican, as shown in Figure 3. The number of input units corresponds to the 19 environmental bills, and the number of competitive units is gradually increased. Figure 4 shows information as a function of the number of epochs by the model from [1] and the GNGA, respectively. We can see that at the beginning of each growing step, information drops temporarily. At the beginning of each growing cycle, a new competitive unit with new connections is added. Because new connections are initialized with small random values, the activation of a new unit is close to the intermediate level of 0.5. Thus, the probability distribution of competitive units at the new stage are away, or more distributed, from that of the previous growth stage. Thus, information temporarily drops at the beginning of each growing cycle. As can be seen in Figure 4, there is no difference between the two methods. Information is increased gradually in the first to the third growth cycles and reaches a stable point in the fourth.

Figure 5 shows competitive unit activations $p(j \mid s)$ by the previous and present models. In the figure, as the magnitude of black squares is larger, the probability $p(j \mid s)$ is higher. The white part shows that the corresponding probabilities are almost zero. As is evident in Figure 5, little difference can be seen in any competitive unit activations. In the first growth cycle, all congressmen are grouped into two classes, Republicans and Democrats. Compared to the results by the cluster analysis

[2] In the cluster analysis, the Euclid distance and Ward method were used.

No.	1	2	3	4	5	6	7	8
Party	D1	D2	D3	D4	D5	D6	D7	D8
1	0	1	0	1	1	1	1	0.5
2	0.5	0	0	0	0	0	0	0
3	0	0.5	1	1	1	1	1	1
4	1	0.5	1	1	1	1	0	0
5	1	0.5	1	1	1	1	1	0
6	1	0.5	0	0	0	0	0	0
7	1	1	1	0.5	1	1	1	1
8	0	0	0.5	0	0	0	0	0
9	1	1	0.5	1	1	1	0.5	1
10	1	1	1	1	1	1	0	1
11	1	1	1	1	1	1	0	1
12	1	0.5	1	1	1	1	1	1
13	0	1	0	0	0	0	0	1
14	0	0	0	0	0	0	0	0
15	0	0	0	0	0	0	0	0
16	0	0.5	0	0	0.5	0	0.5	0
17	0	0	0	0	0	0	0	0
18	1	1	1	1	1	1	1	1
19	1	1	1	1	0	0	0	1

No.	9	10	11	12	13	14	15
Party	R1	R2	R3	R4	R5	R6	R7
1	0	0	1	1	1	1	0
2	1	0	0	0	1	0	1
3	1	0.5	1	1	0.5	1	1
4	0	0	0	0	0	1	0
5	0	0.5	0	0	0	1	1
6	1	0.5	0	1	1	0	1
7	0	0	1	1	1	1	1
8	1	1	1	0	0	0	1
9	0	0	0	0	0	1	1
10	0	0	0	0	0	1	0
11	0	0	1	0	0	1	0
12	0	0.5	0	0	0	1	0
13	1	0.5	0.5	0	1	0	0
14	0	0.5	0	0	1	0	0
15	1	0.5	1	0	1	0	1
16	0.5	0.5	1	1	1	0	1
17	0.5	0.5	1	0	1	0	0
18	1	1	0.5	0	1	1	0
19	0	0	0.5	0	0	1	0

Table 1. United States congressmen with their voting attitude on 19 environmental bills. The first eight congressmen are Democrats, while the latter seven (from 9 to 15) congressmen are Republicans [25]. In the table, 1, 0, and 0.5 represent *yes*, *no*, and *undecided*, respectively.

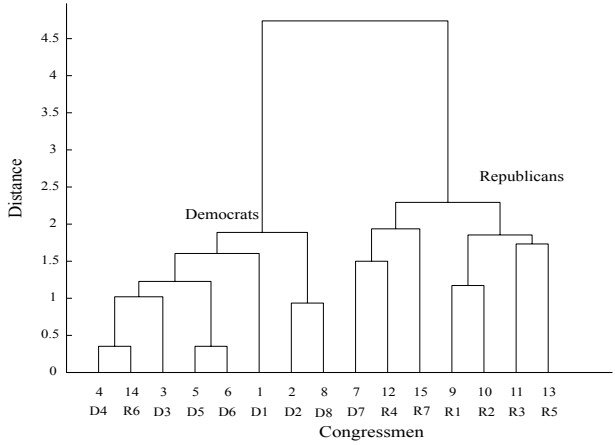

Figure 3. A dendrogram by a cluster analysis for the congressional data.

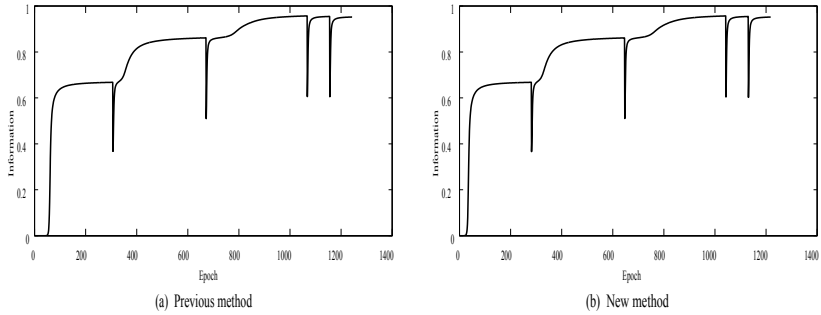

Figure 4. Information as a function of the number of epochs for the political analysis. (a) and (b) denote information by the previous and the present model, respectively.

in Figure 3, Democrat No. 7(7) is appropriately classified as a Democrat. In the second growth cycle, Democrat No. 7(7) and Republican No. 4(12) are detected by the third competitive unit as shown in Figure 5(a2) and (b2). By the cluster analysis (Figure 3), these politicians are grouped together. In the third growth cycle (Figures 5(a3) and (b3)), the fourth competitive unit detects Republicans No. 1(9) and 2(10). As expected, the cluster analysis grouped these politicians together as shown in Figure 3.

Because there is no objective criterion for correctness, this makes algorithm comparisons difficult. However, we can say at least that the greedy methods can classify the politicians into two groups with the same performance as the cluster analysis. However, we could not see any difference between the GNGA and the model in [1]. This means that the previous model is much better than the present model, because

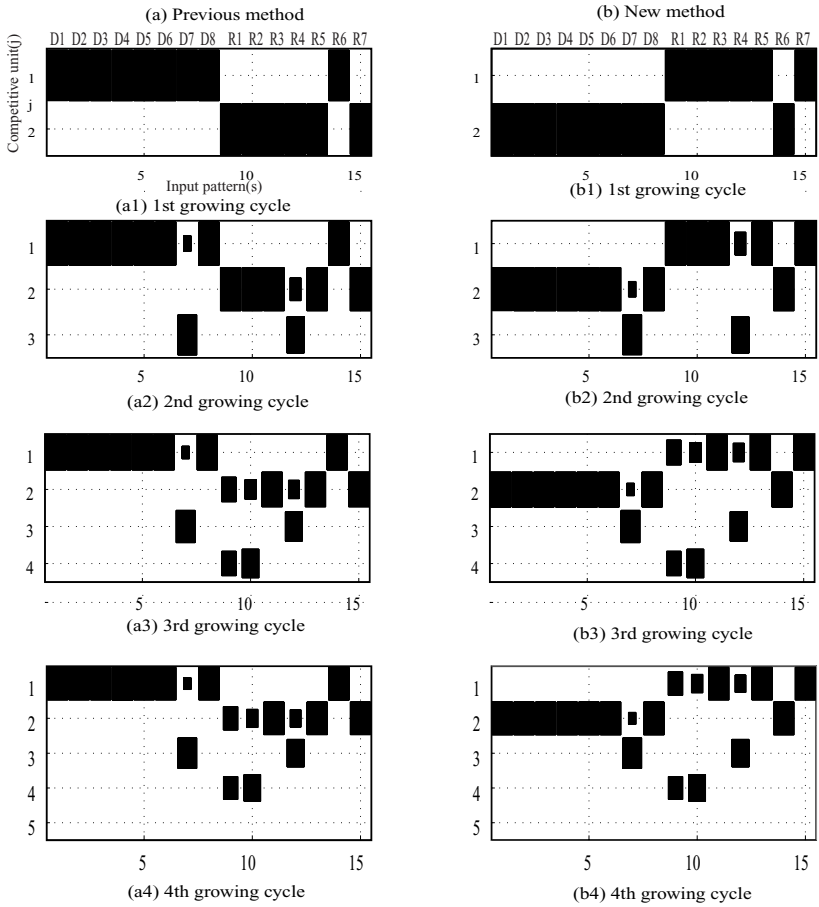

Figure 5. Competitive unit activations by (a) the previous model and by (b) the present model. Black squares represent normalized competitive unit activations (probabilities) $p(j \mid s)$, and their size denotes the activation level.

it only has to update connections into a newly recruited competitive unit. However, we have found that several problems cannot be solved by this method. Sections 4 and 5 show two examples in which the previous model performs poorly.

4. Artificial data analysis

In the second experiment, we attempt to show a problem that cannot be solved by the algorithm from [1]. In the problem, we have the 30 input patterns shown in Figure 6. The number of competitive units is increased gradually up to six, with five growth cycles (right portion of Figure 6). The number of input units is 30. The learning parameter

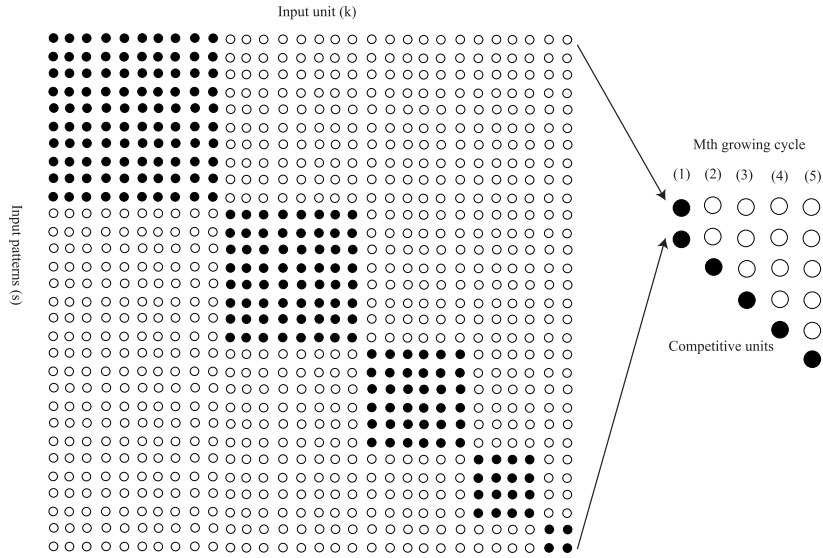

Figure 6. Artificial data (left) and an example of growing cycles (right).

β is 1 for updates using the GNGA. For the conventional method, the learning parameter β is the same, but only connections into the new unit are updated. In learning, momentum is used with the parameter value 0.8 to speed up and stabilize learning.

Figure 7(a) shows information as a function of the number of epochs t by the conventional greedy network-growing method. Information is slowly increased and reaches a stable point. On the other hand, Figure 7(b) shows information as a function of the number of epochs using the GNGA. As shown in the figure, information is significantly increased during different learning cycles and reaches its stable point in the fifth learning cycle. We can see that information is significantly larger than that achieved by the previous model.

Figure 8 shows competitive unit activation patterns obtained after the first to the fifth growth cycles by (a) the conventional and (b) the new method. As shown in Figure 8(a1) and (b1), in the first growth cycle, input patterns are arbitrarily classified into two groups. Beginning with the second growth cycle, some problems occur with the conventional method. As expected, a network tries to classify input patterns into three classes in the second growth cycle with three competitive units. However, as shown in Figure 8(a2), classification is not possible by using the conventional method. On the other hand, by the GNGA, input patterns are clearly classified into three classes, with some minor exceptions (Figure 8(b2)). With the third growth cycle, the difference becomes much clearer. As shown in Figure 8(a3), clear classification is impossible by the conventional method, because weights frozen after the

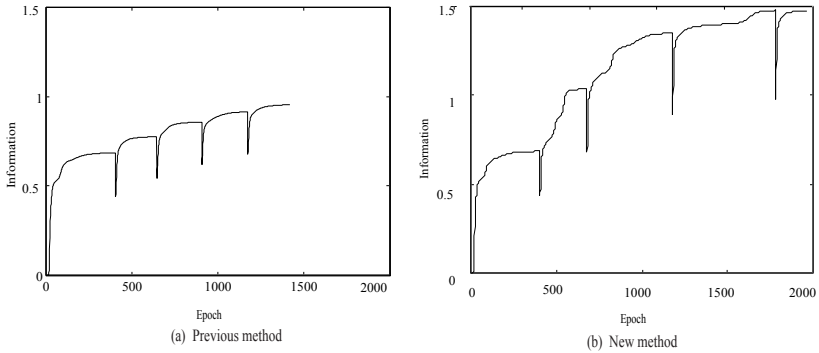

Figure 7. Information by the conventional greedy algorithm (a) and the novel algorithm (b).

first growth cycle prevent the network from classifying input patterns correctly. On the other hand, with the GNGA, four clear classes can be seen, as shown in Figure 8(b3). From the fourth growth cycle on, the conventional method shows great confusion in classification, as shown in Figure 8(a4) and (a5). On the other hand, the GNGA continues to classify input patterns correctly, as shown in Figure 8(b4) and (b5).

We have shown that the new method is superior to the previous one in terms of feature extraction. However, we should note that the method in [1] had been developed to be as efficient as possible in computation. Thus, the conventional method is a fast and computationally inexpensive method, compared with the GNGA. For example, we computed the number of epochs to finish five growing cycles by the new and conventional methods. As can be seen in Figure 7, by the conventional method, the number of epochs to finish the fifth growth cycle is 1468 on average (65 in S.D.). On the other hand, the number of epochs by the new method is increased to 2023 (116 in S.D.). Considering the fact that the conventional method always updates connections into one competitive unit, the efficiency of the conventional method is quite prominent. We expect both methods to be used according to different purposes.

5. Information science education data analysis

The third experiment shows to what extent the GNGA can extract some characteristics of students who take introductory courses in information technology. We did a survey on this subject at Tokai University in Japan in January, 2002. We prepared a questionnaire composed of seven group features, such as attribute, the use of information, personal computers, and so on. In these seven groups, there were 26 basic features, such as sex, major, the concept of information, and so on, as shown in Table 2. 89 students participated in this questionnaire. From the 89

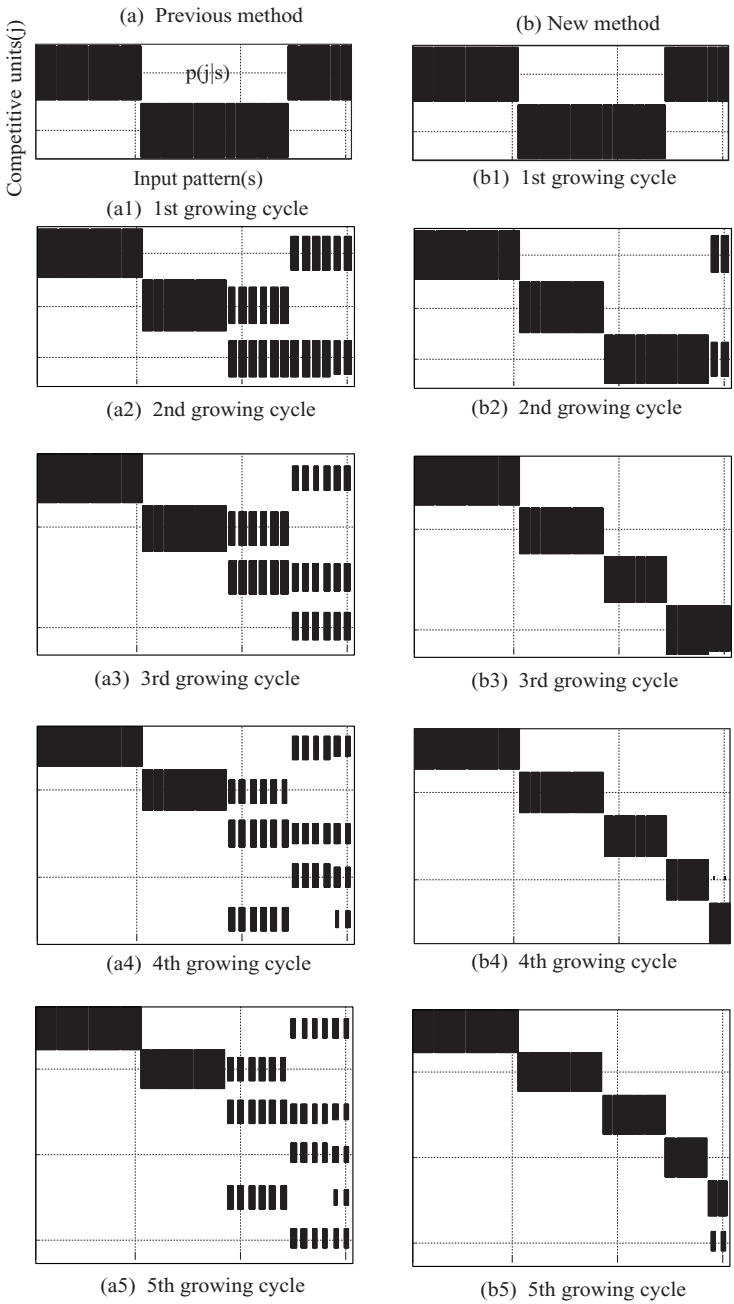

Figure 8. Competitive unit activation patterns by (a) the conventional and (b) the novel method. Black squares represent normalized competitive unit activations (probabilities) $p(j \mid s)$, and their size denotes the activation level.

No.	Macro featues	Micro features
	Attribute	
1		Sex
2		Major
3		Grade
4		Computer knowledge
	Use of information	
5		Concept of information
6		Information representation
7		Information processing procedures
	Personal computers	
8		Types of computers and their applications
9		Mechanism and performance of personal computers
10		Types of software and their characteristics
11		Roles of operating systems
	Information devices	
12		Types of information devices and their operation
13		Setting of information media and devices and their operation
	Computer networks	
14		Basic knowledge of computer networks
15		Personal computer and networks
16		Internet and intranet
17		Computer networks and the next generation media
	Use of application software	
18		Use of Windows operating systems
19		Japanese word processing
20		Spreadsheet software
21		Database software
22		DTP
23		Computer graphics
	Information society	
24		Basic knowledge of information society
25		Information technology in ordinary life
26		Problems in information society

Table 2. Feature description in our data. In the *sex* feature, 1 and 0 denote male and female students, respectively. In the *major*, 1 and 0 represent humanities, and science and engineering, respectively. In the *grade* feature, 1 and 0 represent freshmen and higher, respectively. In the *knowledge on computer* feature, 1 and 0 denote some and little computer knowledge, respectively. In all the other features, 1 and 0 denote some and little interest in the features, respectively.

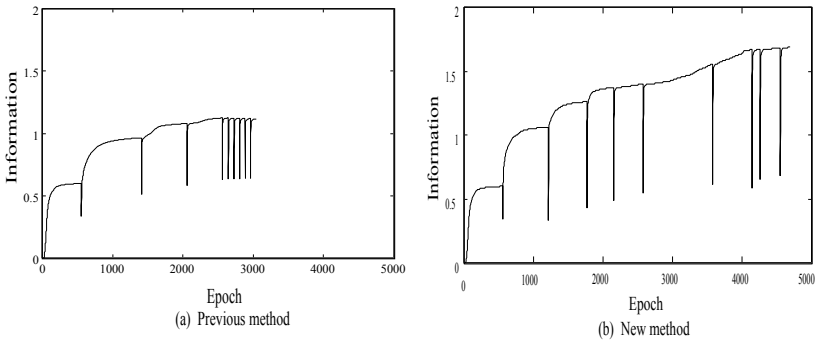

Figure 9. Information as a function of the number of epochs by (a) the previous method and by (b) the new method.

samples we extracted only 71 by eliminating exceptional samples, such as those having regular response patterns. The number of input units corresponds to the 26 basic features. The number of competitive units is gradually increased. The learning parameter β is 1, with the momentum parameter 0.8.

Figure 9 shows information as a function of the number of epochs by the method from [1] and by the GNGA. As shown in Figure 9(a), information is gradually increased and reaches its stable point in five growing cycles. On the other hand, by using the new method, information is more significantly increased and reaches its stable point in eight growing cycles, as shown in Figure 9(b). This suggests that by using the GNGA, a network can obtain much larger information.

Figures 10(a) and 11(a) show connection weights in the first growth cycle by the previous and new methods, respectively. We can immediately see that students are classified into two groups. One is a group that has a high level of interest in information technology and related matter, because the group members respond positively to the majority of the features, as shown in the upper portion in those figures. In particular, they have a high level of interest in information devices (12), (13), computer networks (14), (15), (17), and information society (24), (25), (26). In addition, the first two features, sex and major, are strongly negative, meaning that the typical students of this group are females majoring in science and engineering. We should note that this does not exclude male students from this group, but the typical student represented by the first competitive unit is female. Another group is composed of students who have little interest in information technology or related matters, as shown in the lower portion of Figures 10(a) and 11(a). Students in the group respond negatively to the majority of the features, except Japanese word processing. The use of Japanese word processors is absolutely necessary in university life. Thus, even students

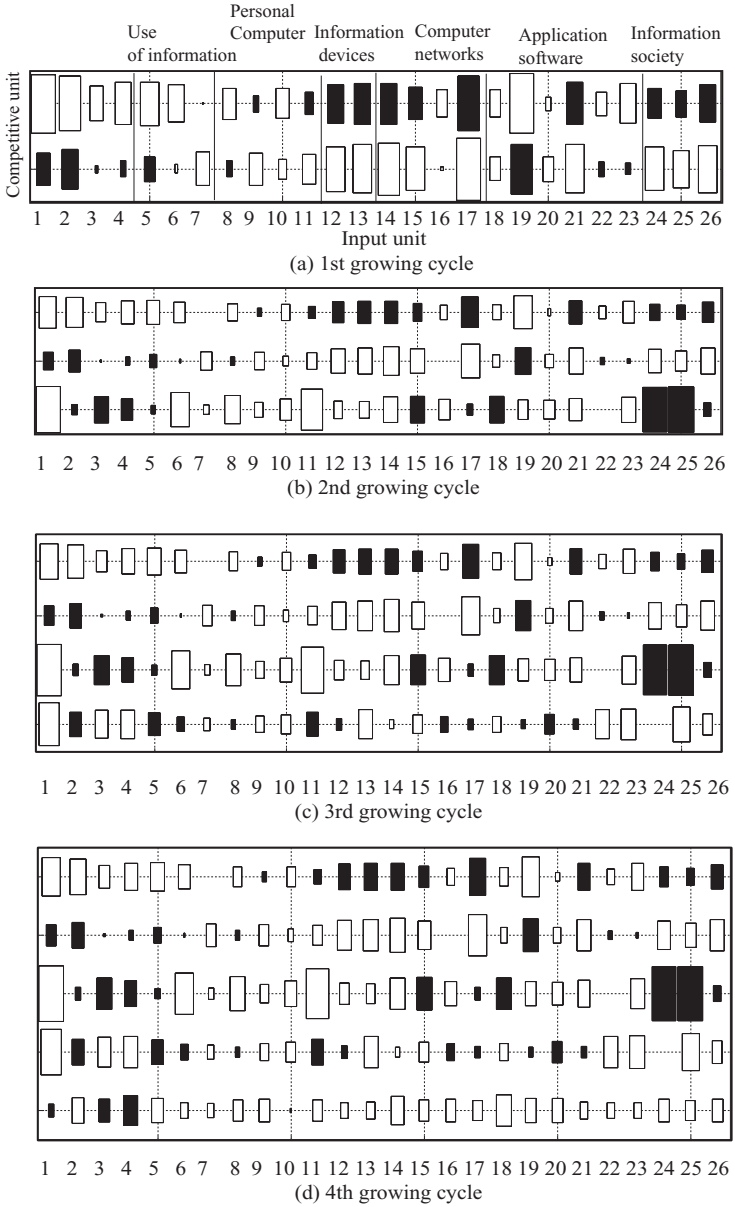

Figure 10. Connection weights obtained in the first to fifth growth cycles by the conventional greedy algorithm. Black and white squares represent positive and negative connections, respectively. Their size denotes the strength of connections.

with little interest in information technology feel the need to study word processing at the university. In addition, since connections to the first and the second feature are positive, this competitive unit represents male students that major in the humanities, economics, and so on (not students of science and engineering). Figure 12 shows response rates by the previous method (a) and the new method (b). The response rates are actually the probability of competitive units $p(j)$. These probabilities denote how many input patterns can activate a competitive unit, because conditional probabilities $p(j \mid s)$ tend to be close to 1. As shown in Figure 12, about 70 percent of all students are classified as a group that has a high level of interest in information technology. On the other hand, about 30 percent are classified as a group whose members have little interest in information technology.

As shown in Figure 11(b), the third competitive unit positively responds to features No. 24 (basic knowledge of information society), No. 25 (information technology in ordinary life), and No. 26 (problems in information society). The feature group "information society" is composed of these three features. Thus, it is certain that the third competitive unit represents students who have more interest in information society than the other two. The type of students in this class is almost the same as the first group. However, the freshman feature (No. 3) plays some importance in this group, because a connection weight into the future No. 3 is slightly positive. Figure 12(b2) shows response rates by the GNGA in the second growth cycle. The second competitive unit still responds to about 30 percent of the students, while a group that has more interest in information technology (70 percent) is distributed over two groups represented by the first and third competitive units.

Figure 10(b) shows connections obtained after the second growth cycle using the previous method. The method from [1] freezes connections after learning is finished in each growing cycle. Thus, it becomes difficult to decompose some features that were extracted in the previous growth cycle. As can be see in the figure, the third competitive unit tries to extract features on information societies. However, because connections into the previous two competitive units have already been frozen, the network shows some difficulty in extracting this feature, as shown in Figure 10(b). Figure 12(a2) shows the response rates of the previous method. Because connections are frozen, the third competitive unit seems to extract information on both the first and second competitive units. This means that the response rates for the first and second competitive units decrease to compensate for the response rate of the third competitive unit.

Figure 11(b3) shows connection weights after the third growth cycle. Because the fourth connections are relatively small, interpreting them is difficult. Relatively large positive connections correspond to features No. 9 (mechanism and performance of personal computers), No. 12

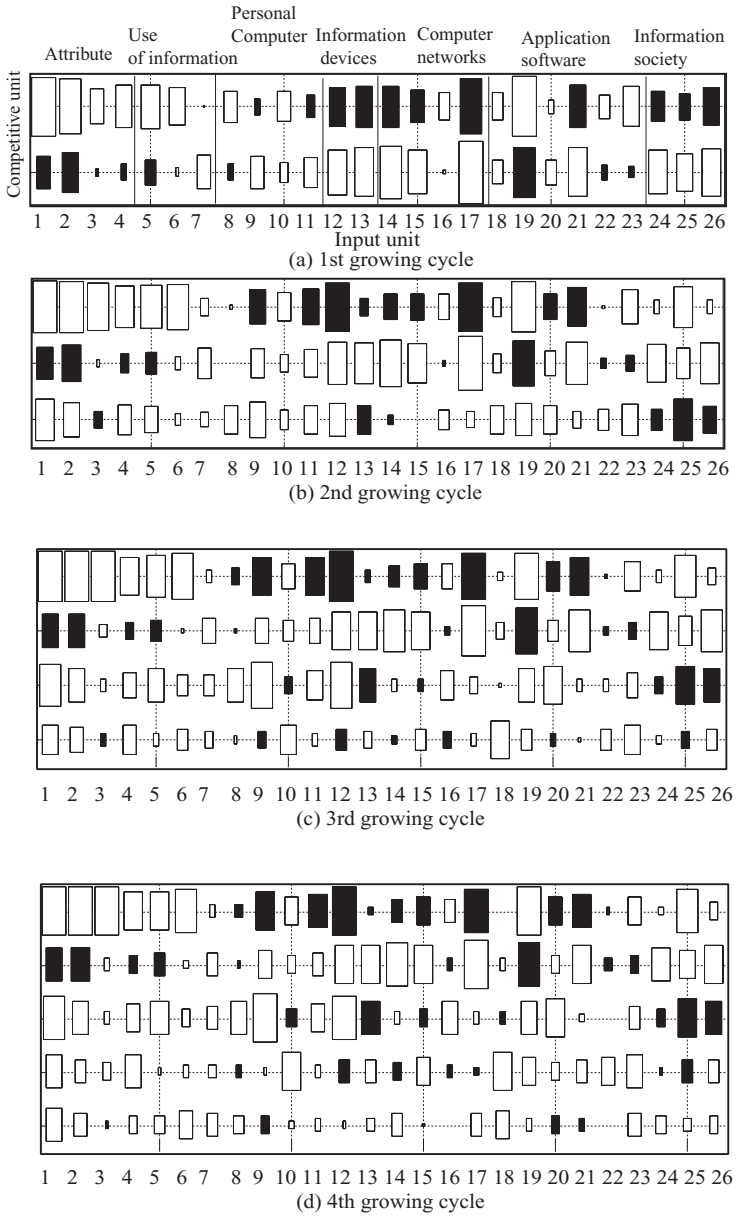

Figure 11. Connection weights obtained in the first to fifth growth cycles. Black and white squares represent positive and negative connections, respectively. Their size denotes the strength of connections.

(type of information devices and their operation), No. 16 (Internet and intranet), and No. 25 (information technology in ordinary life). We can call these connections "information technology in ordinary life." As can be seen in Figure 12(b3), the third competitive unit is divided into the new third and fourth competitive units. Finally, the fifth competitive unit responds positively to No. 9 (mechanism and performance of personal computers) and No. 20 (spreadsheet software). These connections can be called "personal computers and their applications."

Comparing these connections with those found by the GNGA, it is difficult to interpret connections and extract some features using the previous method, because positive connections are smaller. Figure 12(a3) shows the response rates by the previous method. As can be seen in Figure 12(a3), because connections are frozen, the response rates by the previous method for the previous three competitive units are the same. Thus, the competitive unit tends to respond to an extremely small number of students. Finally, in the fourth growth cycle, by the previous method Figure 12(a4), the response rate by the fifth competitive unit is further decreased, and almost all connections are strongly negative, as shown in Figure 10(d). These results show that the GNGA can more explicitly extract features from these input patterns, because the previous method cannot extract new features during later stages.

Experimental results have shown that students are mainly classified into three groups, as shown in Figure 13. The first group is composed of students with a high level of interest in information technology and its related matter. This group is further divided into two subgroups. One is a group that has a high level of interest in information technology in general. We should note especially that these students major in science and engineering and thus have a special interest in new information media. Another group is composed of students who have interest in information society. On the other hand, the second group is composed of students who have little interest in information technology and its related matter. However, even if they have no interest in information technology, they know the importance of information technology in campus life, because the students in this class respond strongly to Japanese word processing.

We compared these results with those by conventional competitive learning and the principle of component analysis. First, we present results by a conventional competitive method. The method used in the experiment was the frequency-sensitive competitive learning method by Ahalt, *et al.* in [5]. In this method, the winner is selected by the equation

$$\| \mathbf{w}_{j^*} - \mathbf{x} \| c_{j^*} \leq \| \mathbf{w}_j - \mathbf{x} \| c_j, \tag{10}$$

where j^* denotes the winning neuron and c_{j^*} denotes the total winning number. As a neuron wins frequently, the total winning number c_{j^*} increases, and then the neuron tends to win less frequently. Note that the learning parameter was 0.1. Figure 14 shows connection weights

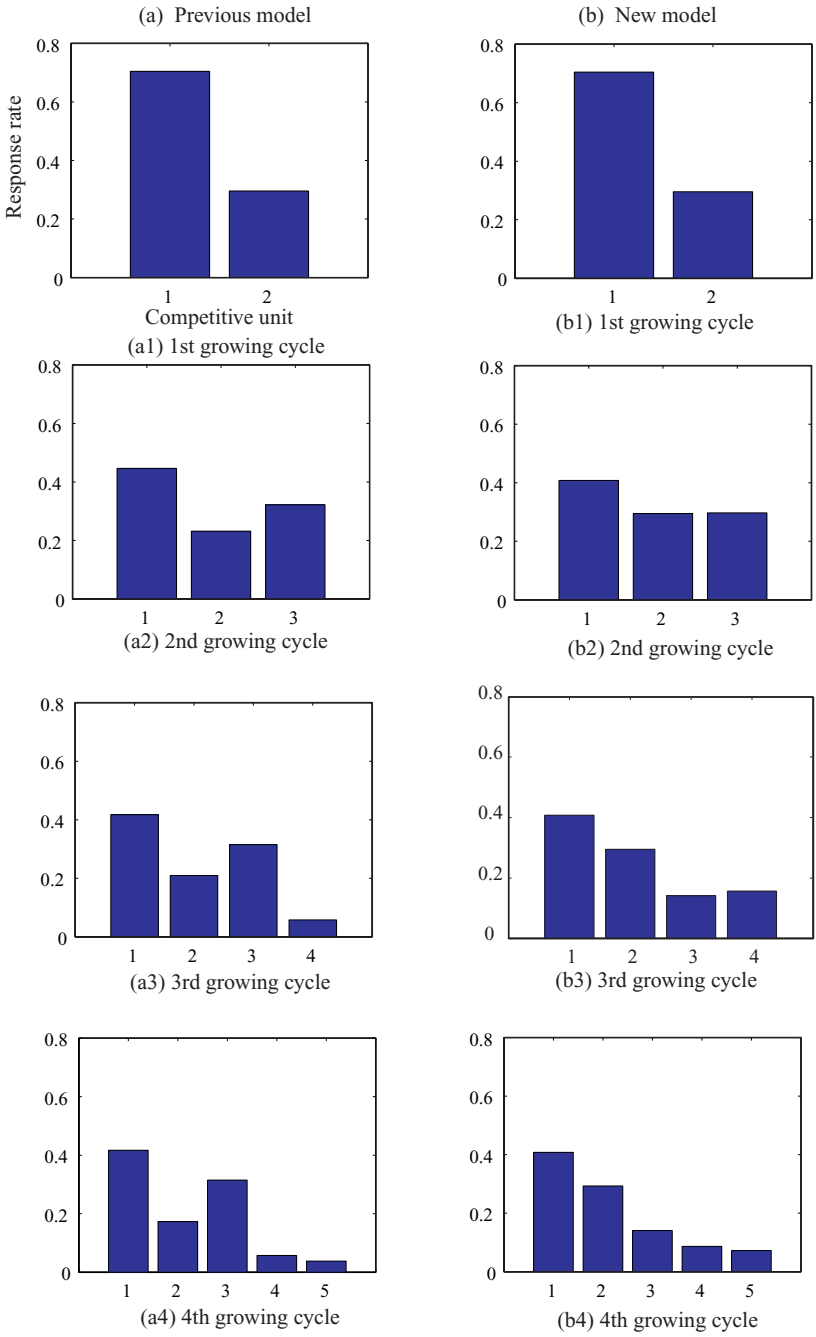

Figure 12. Response rates by the previous method (a) and the new method (b).

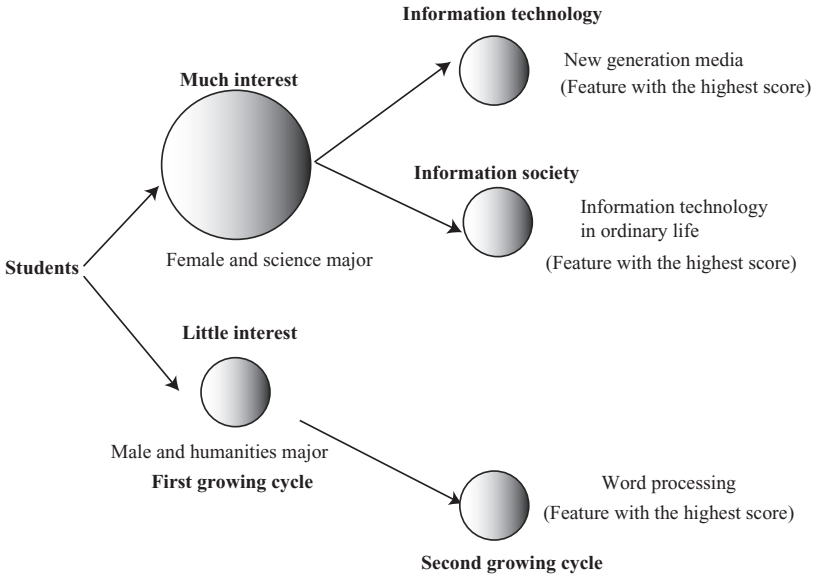

Figure 13. An interpretation of connection weights obtained by the GNGA.

using this conventional competitive learning method when the number of competitive units is increased from two to five. As can be seen in the figure, it is extremely difficult to interpret connection weights. As the number of growth cycles is larger, it is much more difficult to interpret the meaning of connections.

Then we analyzed the education data by the principal component analysis method, which is one of the most popular conventional multivariate data analysis methods. Figure 15 shows the contribution rate of the first 10 principal components. The contribution rates are very small, meaning that there are no dominant principal components. Figure 16 represents the scores of the first five principal components. The first principal component separates subject attributes such as sex, major, and grade (Features No. 1 to 4) from the answers to the questionnaire (Features No. 5 to 26). From the second principal component onwards, it is extremely difficult to interpret the meaning. These results show that the GNGA detects features in a way fundamentally different from the conventional data analysis method, and that it can detect significant information that principal component analysis does not extract.

6. Conclusion

A new network-growing algorithm called the greedy network-growing algorithm (GNGA) was proposed. This new method is based upon information theoretic competitive learning techniques that can maxi-

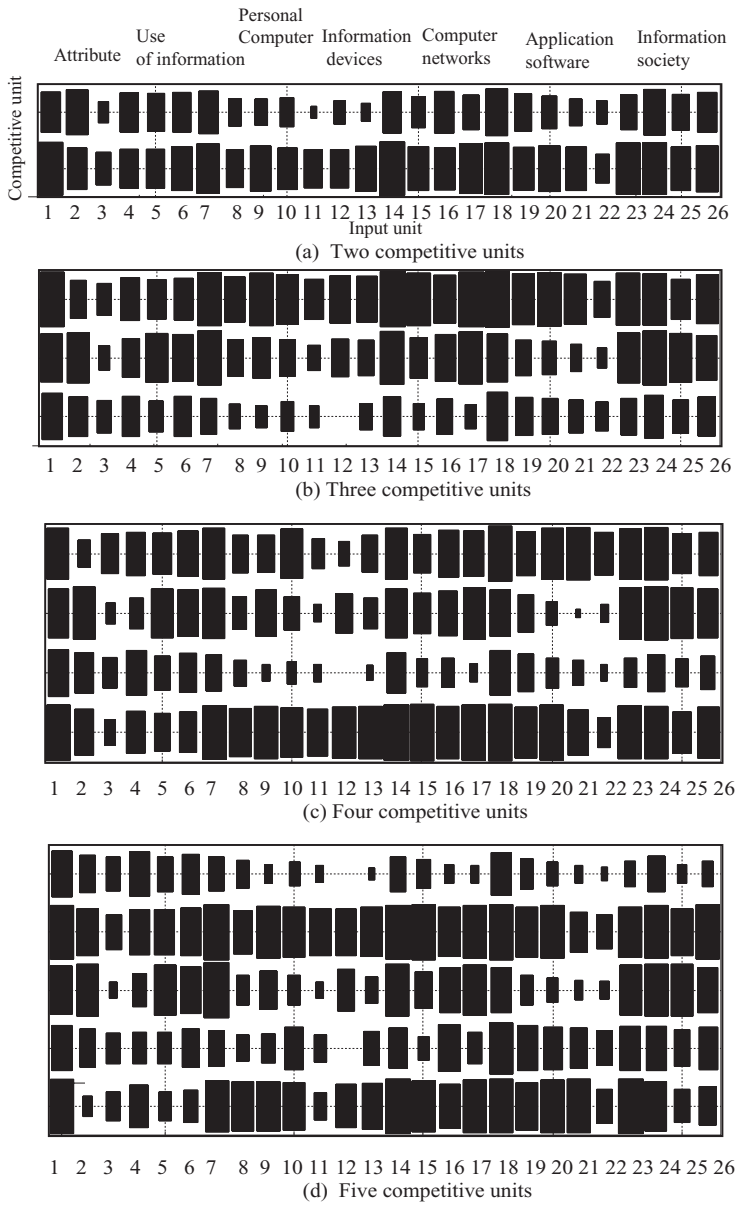

Figure 14. Connection weights with two to five competitive units by the conventional competitive method. Black squares represent positive connections, and their size denotes the strength of connections.

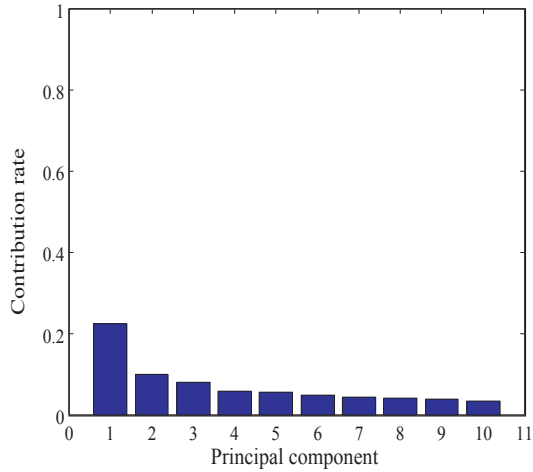

Figure 15. Contribution rate of the first 10 principal components using principal component analysis.

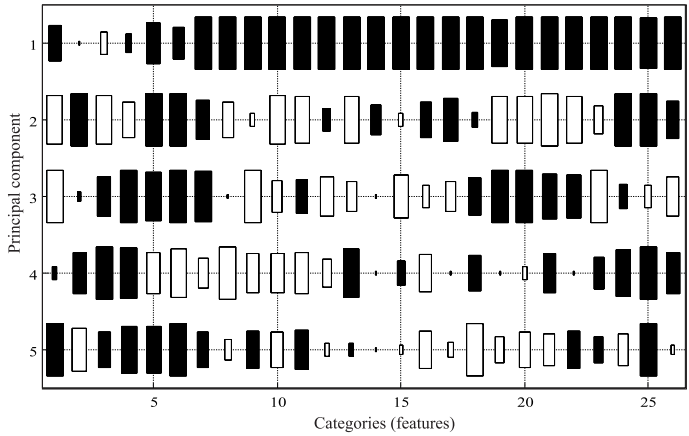

Figure 16. Category scores of the first five principal components.

mize mutual information between input patterns and competitive units. In maximizing mutual information, the entropy of competitive units is maximized. By using entropy maximization, the problem of dead neurons can be avoided, because all competitive units tend to be equally used. In addition, this model has been proposed to extend a previous model [1], whose principal objective was to learn input patterns as efficiently as possible. In the previous model, only connections into new units were updated for computational efficiency. Because connections, except those to a newly recruited competitive unit, are frozen, networks sometimes fail to extract salient features successively. To overcome this

shortcoming, we relax the condition of the previous model, and all connections are updated to capture features gradually, at the expense of efficient computing.

For further development, several problems should be mentioned. First, we did not use any acceleration learning methods, except for the momentum term. However, a more sophisticated learning method should be used to tune parameters. For example, the learning parameter should be changed according to the magnitude of information obtained. Second, we should combine the method from [1] with the GNGA for efficient and explicit feature extraction. As mentioned, our previous method is very computationally efficient because only connections into a newly recruited competitive unit are updated. However, the method in [1] cannot extract salient features in later stages of learning. To overcome this shortcoming, when inappropriate feature detection is observed, a network should take on a new method in which all connections are updated. By this technique, we can overcome the shortcoming of the previous method and extract salient features comparable to those extracted by the GNGA. Finally, though much remains to be done to find a compromise between computational efficiency and feature extraction, the present research certainly opens up a new perspective in network-growing algorithms as well as information theoretic neural computing.

Acknowledgment

The author is very grateful to an anonymous reviewer, Taeko Kamimura, and Mitali Das for valuable comments and suggestions.

References

[1] R. Kamimura and T. Kamimura, "Greedy Information Algorithm," in *Proceedings of International Joint Conference on Neural Networks* (IEEE, Honolulu, 2002).

[2] D. E. Rumelhart and J. L. McClelland, "On Learning the Past Tenses of English Verbs," in *Parallel Distributed Processing, Volume 2*, edited by D. E. Rumelhart, G. E. Hinton, and R. J. Williams (MIT Press Cambrige, 1986).

[3] S. Grossberg, "Competitive Learning: From Interactive Activation to Adaptive Resonance," *Cognitive Science*, **11** (1987) 23–63.

[4] D. DeSieno, "Adding a Conscience to Competitive Learning," in *Proceedings of IEEE International Conference on Neural Networks* (IEEE Press, San Diego, 1988).

[5] S. C. Ahalt, A. K. Krishnamurthy, P. Chen, and D. E. Melton, "Competitive Learning Algorithms for Vector Quantization," *Neural Networks*, **3** (1990) 277–290.

[6] L. Xu, "Rival Penalized Competitive Learning for Clustering Analysis, RBF Net, and Curve Detection," *IEEE Transactions on Neural Networks*, **4**(4) (1993) 636–649.

[7] A. Luk and S. Lien, "Properties of the Generalized Lotto-type Competitive Learning," in *Proceedings of International Conference on Neural Information Processing* (Morgan Kaufmann Publishers, San Mateo, CA, 2000).

[8] 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*, **14**(2) (2002) 137–162.

[9] S. E. Fahlman and C. Lebiere, "The Cascade-correlation Learning Architecture," in *Advances in Neural Information Processing, Volume 2*, (Morgan Kaufmann Publishers, San Mateo, CA, 2000).

[10] T. A. Shultz and F. Rivest, "Knowledge-based Cascade-correlation: Using Knowledge to Speed Learning," *Connection Science*, **13** (2001) 43–72.

[11] M. Lehtokangas, "Modelling with Constructive Backpropagation Architecture," *Neural Networks*, **12** (1999) 707–716.

[12] B. Fritzke, "Fast Learning with Incremental RBF Networks," *Neural Processing Letters*, **1**(1) (1994) 2–5.

[13] B. Fritzke, "Unsupervised Clustering with Growing Cell Structures," in *Proceedings of IEEE International Joint Conference on Neural Networks* (IEEE and INNS, Seattle, 1991).

[14] B. Fritzke, "Growing Cell Structures—A Self-organizing Network for Unsupervised and Supervised Learning," *Neural Networks*, **7**(9) (1994) 1441–1460.

[15] B. Fritzke, "Unsupervised Ontogenetic Networks," in *Handbook of Neural Computation*, edited by E. Fiesler and R. Beale (Institute of Physics Publishing and Oxford University Press, London, 1996).

[16] J. Dopazo and J. M. Carazo, "Phylogenetic Reconstruction Using a Growing Neural Network that Adopts the Topology of a Phylogenetic Tree," *Journal of Molecular Evolution*, **44** (1997) 226–233.

[17] J. Herrero, A. Valencia, and J. Dopazo, "A Hierarchical Unsupervised Growing Neural Network for Clustering Gene Expression Patterns," *Bioinformatics*, **17**(2) (2001) 126–136.

[18] R. Linsker, "Self-organization in a Perceptual Network," *Computer*, **21** (1988) 105–117.

[19] R. Linsker, "How to Generate Ordered Maps by Maximizing the Mutual Information between Input and Output," *Neural Computation*, **1** (1989) 402–411.

[20] R. Linsker, "Local Synaptic Rules Suffice to Maximize Mutual Information in a Linear Network," *Neural Computation*, **4** (1992) 691–702.

[21] J. J. Atick and A. N. Redlich, "Toward a Theory of Early Visual Processing," *Neural Computation*, **2** (1990) 308–320.

[22] S. Becker, "Mutual Information Maximization: Models of Cortical Self-organization," *Network: Computation in Neural Systems*, **7** (1996) 7–31.

[23] S. Becker and G. E. Hinton, "Learning Mixture Models of Spatial Coherence," *Neural Computation*, **5** (1993) 267–277.

[24] L. L. Gatlin, *Information Theory and Living Systems* (Columbia University Press, New York, 1972).

[25] H. C. Romesburg, *Cluster Analysis for Researchrers* (Krieger Publishing Company, Florida, 1984).