**Proceedings of the 9th International Conference on Neural Information Processing (ICONIP'02), Vol. 1**
Lipo Wang, Jagath C. Rajapakse, Kunihiko Fukushima,
Soo-Young Lee, and Xin Yao (Editors)

# PROGRESSIVE FEATURE EXTRACTION BY EXTENDED GREEDY INFORMATION ACQUISITION

*Ryotaro Kamimura, †Haruhiko Takeuchi and †† Osamu Uchida*

Information Science Laboratory
and Future Science and Technology Joint Research Center, Tokai University,
1117 Kitakaname Hiratsuka Kanagawa 259-1292, Japan
ryo@cc.u-tokai.ac.jp
† Human-Computer Interaction Group,
Institute for Human Science and Biomedical Engineering,
National Institute of Advanced Industrial Science and Technology,
1-1-1 Higashi, Tsukuba 305-8566, Japan
takeuchi.h@aist.go.jp
‡‡Department of Information Science
and Future Science and Technology Joint Research Center, Tokai University,
1117 Kitakaname Hiratsuka Kanagawa 259-1292, Japan
ouchida@keyaki.cc.u-tokai.ac.jp

## ABSTRACT

In this paper, we propose a new network growing method to detect salient features in input patterns. The new method is based upon the previous network growing model[1] and introduced to overcome some problems in the previous model. We have so far tried to build a model that can learn input patterns as efficiently as possible. To realize this efficiency, we impose upon networks a constraint that only connections into new competitive units must be updated to absorb as much information as possible from outside. However, one of the problems is that the previous improper feature extraction prevents networks from extracting appropriate features in the later learning stages. To overcome this problem, we relax the condition of the previous model, and we permit networks to update all connections for gradual feature extraction at the expense of computational efficiency. We applied the new method to a simple problem that the previous model cannot solve, and information education data analysis. In both problems, we found that the new method can appropriately extract features from input patterns.

## 1. INTRODUCTION

In this paper, we extend our greedy information acquisition algorithm to overcome a shortcoming of the previous method that was developed as a new type of unsupervised network growing method [1]. The new model is characterized by the following two points: (1) the method is based upon information theoretic competitive learning; (2) the new model can solve a shortcoming of our previous greedy information acquisition model. Let us explain these two points in more details.

First, the method is based upon information theoretic competitive learning [2]. Competitive learning is one of the most familiar techniques in unsupervised learning. However, some problems have already been pointed out. For example, the dead neuron and the number of competitive units are serious problems in conventional competitive learning. When initial conditions are not appropriately given, some neurons can not be activated for ever, that is, dead neuron. In addition, when the number of competitive units exceeds the number of classes in input patterns, final results tend to deteriorate. In our model, dead neuron problems are solved, because entropy for competitive units must be maximized, and this prevents some neurons from beeing inactive for all input patterns. In addition, the number of competitive units is gradually increased, and this makes it possible to determine the number of classes flexibly.

Second, the present method can grow a network gradually, while detecting different features in input patterns. In our previous model on greedy information acquisition, we tried to grow networks as efficiently as possible [1]. For this purpose, in maximizing information content, only connections into a new competitive unit are updated. This means that networks always update connections into one competitive unit except the initial cycle. Thus, the method 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 growing cycles, the feature extraction of the later stages degrade, because all connections after learning are frozen in the algorithm. To overcome this shortcoming, we relax the severe restriction of the previous 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.

## 2. GROWING NETWORK

### 2.1. Growing Algorithm

The greedy information acquisition algorithm increases gradually network complexity. In the algorithm, a network tries to absorb as much information as possible from outer environment. When no more additional information can be obtained, the network recruits another unit, and then it again tries to absorb information maximally. This general idea of networks with greedy information acquisition algorithm has been realized in neural networks [1]. The method aimed to develop a growing mechanism that is as efficient as possible in computation time. However, the inappropriate feature detection in the early stage of learning prevents networks from developing explicit internal representations. For solving this problem, we relax the condition on the previous model for speeding up learning.

Let us show a fundamental difference between two methods. Figure 1 shows an actual network architecture for greedy algorithm. Figure 1(1) 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 to make 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 the previous model. 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(2). Because 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 1(3)). 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 very efficient in terms of computational time, because networks have only to update connections into one competitive unit. If it is possible to extract features gradually by this method, this is one of the most efficient methods ever developed for feature extraction. We have found that for some problems, this efficient method can extract basic features gradually as the number of growing cycles is increased. However, one of the serious problems is that if networks capture inadequate features at the beginning of learning, networks fail to extract appropriate features in the later cycles. Once learning is finished, all connections are frozen in this algorithm. These frozen connections prevent networks from extracting appropriate features in the later cycles. For overcoming this shortcoming, we propose a novel greedy information acquisition algorithm in which all connections are updated to maximize information content at the expense of computation efficiency.

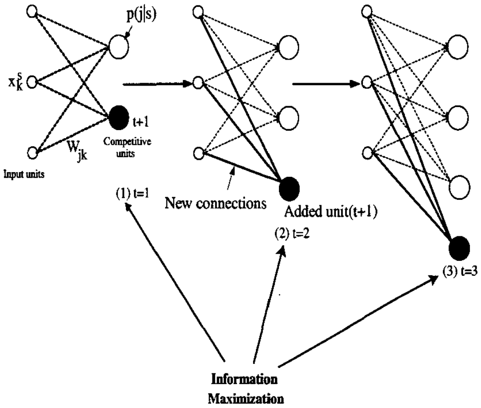


Figure 1: A process of network growing by the new method.

### 2.2. Information Maximization

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 decrease in uncertainty [3]. Uncertainty decrease, that is, information $I(t_n)$ at the $t_n$ epoch ($n$th learning cycle in the $t$th growing cycle), is defined by

$$\begin{aligned} I(t_n) &= -\sum_{\forall j} p(j;t_n)\log p(j;t_n) \\ &\quad +\sum_{\forall s}\sum_{\forall j} p(s)p(j \mid s;t_n)\log p(j \mid s;t_n), \quad (1) \end{aligned}$$

where $p(j;t_n)$, $p(s)$ and $p(j|s;t_n)$ denote the probability of the $j$th unit in a system at the $n$th learning cycle in the $t$th growing cycle, the probability of the $s$th input pattern and the conditional probability of the $j$th unit, given the $s$th input pattern at the $t_n$th epoch, respectively.

Let us present update rules to maximize information content in every stage of learning. As shown in Figure 1, a network at the $t_n$th epoch (the $n$th learning cycle and the $t$th growing cycle) is composed of input units $x_k^s$ and competitive units $v_j^s(t_n)$. The $j$th competitive unit receives a net input from input units, and an output from the $j$th competitive unit can be computed by

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

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

$$p(j \mid s; t_n) = \frac{v_j^s(t_n)}{\sum_{m=1}^{t+1} v_m^s(t_n)}. \qquad (3)$$

Since input patterns are supposed to be uniformly given to networks, the probability of the $j$th competitive unit is computed by

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

Information $I(t_n)$ is computed by

$$\begin{aligned} I(t_n) &= -\sum_{j=1}^{t+1} p(j; t_n) \log p(j; t_n) \\ &+ \frac{1}{S}\sum_{s=1}^{S}\sum_{j=1}^{t+1} p(j \mid s; t_n) \log p(j \mid s; t_n). \qquad (5) \end{aligned}$$

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

$$\begin{aligned} \Delta w_{jk}(t_n) &= -\beta \sum_{s=1}^{S} Q_{jk}^s(t_n) \log p(j; t_n) \\ &+ \beta \sum_{s=1}^{S} \sum_{m=1}^{t+1} p(m \mid s; t_n) \\ &\times Q_{jk}^s(t_n) \log p(m; t_n) \\ &+ \beta \sum_{s=1}^{S} Q_{jk}^s(t_n) \log p(j \mid s; t_n) \\ &- \beta \sum_{s=1}^{S} \sum_{m=1}^{t+1} p(m \mid s; t_n) \\ &\times Q_{jk}^s(t_n) \log p(m \mid s; t_n), \qquad (6) \end{aligned}$$

where $\beta$ is the learning parameter and

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

Finally, we should state how to stop learning and how to add a new competitive unit. Now, let $I(t_n)$ denote information content computed at the $n$th learning cycle in the $t$th growing cycle. Relative increase in information $R(t_n)$ is computed by $R(t_n) = |I(t_n) - I(t_{n-1})|/I(t_{n-1})$, where $n = 1, 2, 3, \ldots$. If $R(t_n)$ is less than a certain point $\epsilon$ for three consecutive epochs, the first information maximization process is finished.

## 3. ARTIFICIAL DATA ANALYSIS

In the first experiment, we try to show that a problem cannot be solved by the conventional greedy information acquisition algorithm. In the problem, we have 30 input patterns shown in Figure 2. The number of competitive units is increased gradually up to 6 units with 5 growing cycles (left figure in Figure 2). The number of input units is 30. The learning parameter $\beta$ is 1 for updates for the new method. For the conventional method, the learning parameter $\beta$ is the same, but only connections into the new unit is updated. In learning, the momentum is used with the parameter value: 0.8 to speed up and stabilize learning.

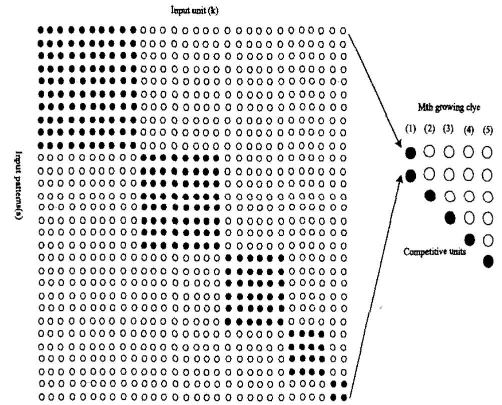


Figure 2: An artificial data (left) and an example of growing cycles (right).

Figure 3(a) shows information as a function of the number of epoch $t_n$ by the conventional greedy information acquisition method. Information is increased greatly at the first epoch, and then information is slowly increased. On the other hand, Figure 3(b) shows information as a function of the number of epochs by the new method. As shown in the figure, information is significantly increased during different learning cycles.

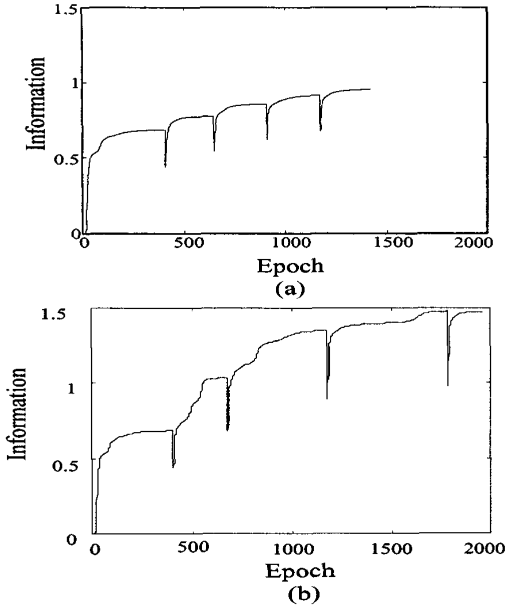


Figure 3: Information by the conventional greedy algorithm (a) and novel algorithm (b).

Figure 4 shows competitive unit activation patterns obtained at the 1st to 5th growing cycles by the conventional (a) and the new method (b). As shown in Figure 4(a1) and (b1), in the first growing cycle, input patterns are classified into two groups arbitrarily. From the 2nd growing cycle, some troubles happen with the conventional method. As expected, a network tries to classify input patterns into three classes with the second growing cycle with three competitive units. However, as shown in Figure 4(a2), classification is not possible by using the conventional method. On the other hand, by the new method, input patterns are clearly classified into three classes with some minor exceptions (Figure 4(b2)). With the third growing cycles, difference becomes much clearer. As shown in Figure 4(a3), clear classification is impossible by the conventional method, because weights frozen after the 1st growing cycle prevent the network from classifying input patterns correctly. On the other hand, by using the new method, four clear classes can be seen, as shown in Figure 4(b3). From the 4th growing cycle on, the conventional method shows great confusion in classification, as shown in Figure 4(a4) and (a5). On the other hand, the new method continues to classify input patterns correctly, as shown in Figure 4(b4) and (b5).

We have shown that the new method is superior to the convention one in terms of feature extraction. However, we should note that the conventional method has been developed to be as efficient as possible in computation. Thus, the conventional method is a fast and computationally inexpensive method, compared with the new method. For example, we computed the number of epochs to finish the five growing cycles by the new and the conventional method. As can be seen in the figure, by the conventional method, the number of epochs to finish the 5th growing cycles is 1468 in 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 updates always connections into one competitive unit, the efficiency of the conventional method is much prominent. We expect both methods to be used for different purposes.

## 4. INFORMATION SCIENCE EDUCATION ANALYSIS

The second experiment shows to what extent the greedy algorithm can extract some characteristics of students who take a introductory course in information technology. We prepared a questionnaire composed of 7 group features such as attribute, use of information, personal computers and so on. In these 7 groups, there were 26 basic features such as sex, major, concept of information and so on. The number of students participating in this questionnaire is 89 students. From 89 samples, we extracted only 71 samples by eliminating extraordinary samples such as samples with too regular response patterns.

Figure 5 shows information by the new method(a) and the previous method (b). As shown in Figure 5(a), information is gradually increased, and seems to reach a stable point with 8th growing cycle. On the other hand, by the previous method, only 5th growing cycle produce a saturated state.

Figure 6(a) shows connection weights obtained by the 1st growing cycle. We can immediately see that students are classified as two groups. One is a group that has much interest in information technology and related matter in general, because they respond positively to the majority of the features, as shown in connection weights on the upper side of Figure 6(a). Especially, they have much 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 female students whose major is science and engineering. We should note that this does not exclude the male students in this group, but the typical student represented by the first competitive units is a female student. Another group is composed of students who have little interest in information technology or related matter, as shown in the lower figure in 6(a). Students in the group respond negatively to the majority of features except Japanese word processing. The use of Japanese word processors is absolutely necessary in the university life. Thus, even students with little interest in information technology feel the necessity to study word processing in

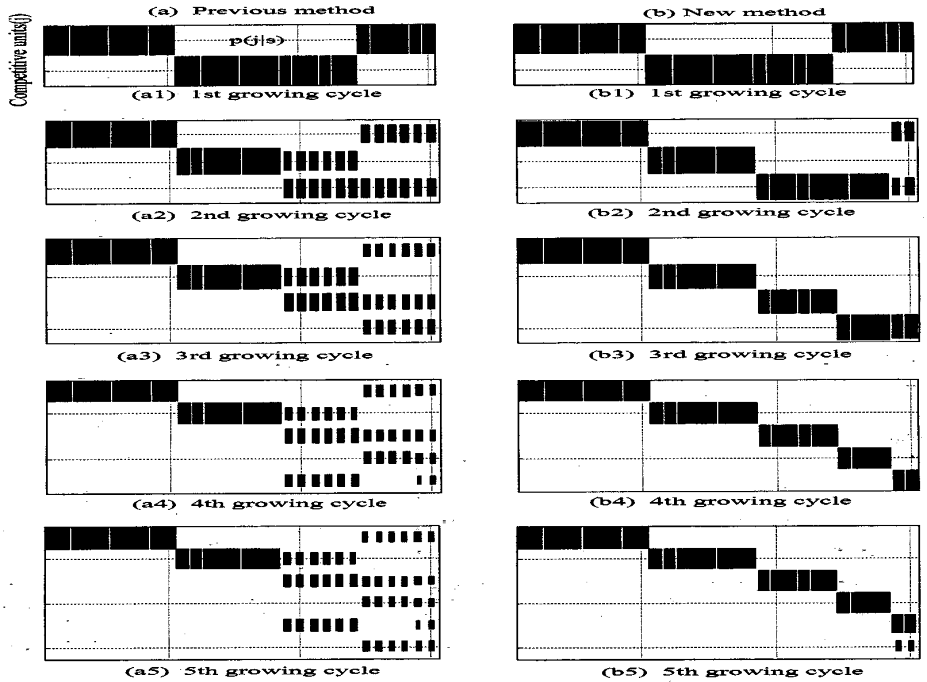


Figure 4: Competitive unit activation patterns by the conventional (a) and the novel method (b).

the university life. In addition, since the first and the second feature is on, this competitive unit represents male students whose major is the humanities, economics and so on (not students of science and engineering). We computed the response rate to examine the ratio of these students. The response rate is computed by

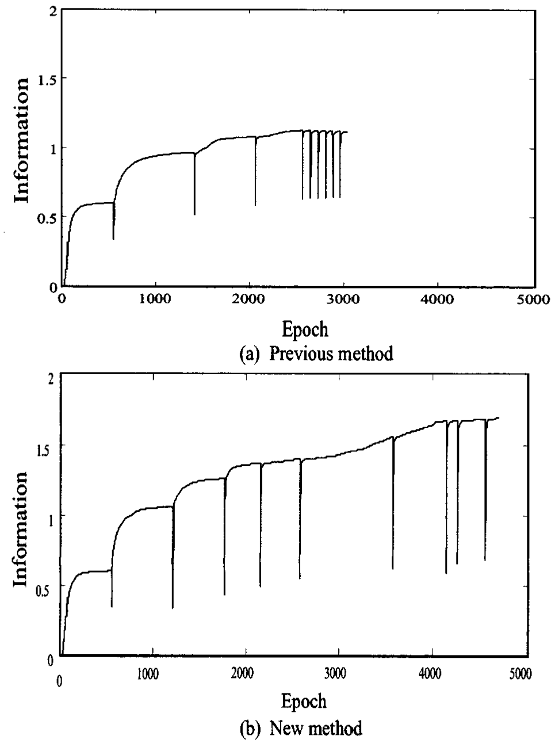
$$R_s = \frac{1}{S}\sum_{s=1}^{S} U(j \mid s), \tag{8}$$

where $U(j \mid s) = 1$ for $p(j \mid s) \geq 0.5$, and 0, otherwise. Figure 7(a) shows the response rates by the 1st growing cycle. As shown in the figure, the first group is about 70 percent, and the second group is about 30 percent of total students.

Figure 6(b) shows connection weights after the 2nd growing cycle. We can see that connection weights into the 3rd competitive unit are strongly positive especially for features related to information society (Feature No.24, 25 and 26). Type of students in this class is almost the same as the first group. However, the freshman feature(Feature No.3) plays some importance in this group, because a connection weight into the future No.3 is slightly positive. Figure 7(b) shows that the ratio of the first group is decreased from 70 to 40 percent, the difference 30 percent is for the students with special interest in information society. Figure 6(c) and (d) show connection weights after the 3rd and 4th growing cycle. These connections seem to represent students that want to have basic knowledge on computer. It is, however, difficult to interpret explicitly the meaning of connections.

These results show that students are mainly classified into three groups, as shown in Figure 8. The first group is composed of student with much interest in information technology and its related matter. This group is further divided into two subgroups. One is a group that has much interest in information technology in general. However, we should note that especially these students' major is science and engineering, and that they have special interest in information technology. Another group is composed of students who have interest in information society. Especially, the freshmen are the typical students in this class. 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 the campus life, because the students in this class respond strongly to Japanese word processing.

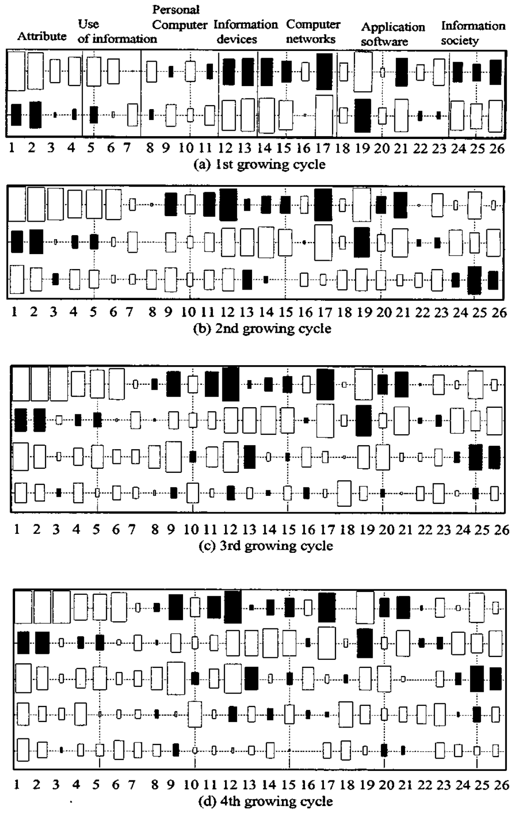
We also analyzed the education data by the principle component analysis, which is one of the conventional multivariate data analysis methods. Figure 9 shows the contribution rate of the first ten principal components. Each principal component has only a small value, and there are no dominant principal components. Figure 10 represents the



(a) Previous method

(b) New method

Figure 5: Information (a) and information gain (b) for the education problem.



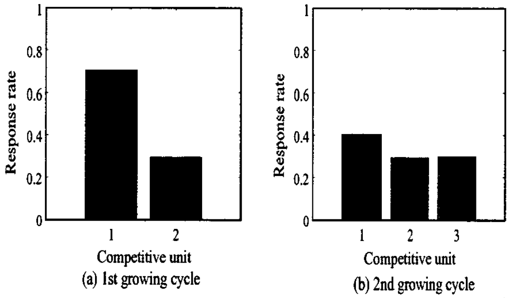
(a) 1st growing cycle

(b) 2nd growing cycle

(c) 3rd growing cycle

(d) 4th growing cycle

Figure 6: Connection weights from 1st to 5th growing cycles.



(a) 1st growing cycle

(b) 2nd growing cycle

Figure 7: Response rates from 1st to 2th growing cycles.

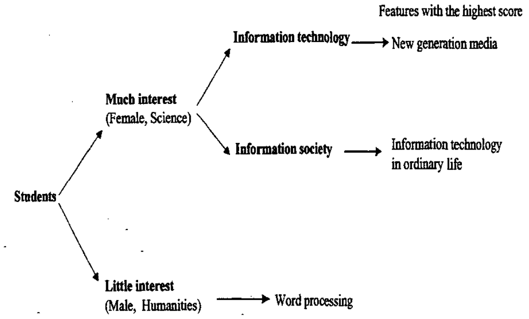


Figure 8: An interpretation of connection weights obtained by the greedy information acquisition algorithm.

scores of the first five principal components. The first principal component separates the subject's attributes such as sex, major and grade (Feature No. 1 to 4) from the answers to the questionnaire (Feature No. 5 to 26). The strongly positive values represent the answers to the questionnaire, while the other values show the subject's attributes. The 2nd principal component separates the features on the application software (Feature No. 19 to 23) from the features on the information society (Feature No. 24 and 25). Though we can interpret the meaning of the remaining principal components in the same way, it is difficult to show clearly the meaning of the principal components. These results show that the greedy information acquisition algorithm detects features in a way fundamentally different from the conventional data analysis, and can detect the substantial information which principal component analysis does not provide.

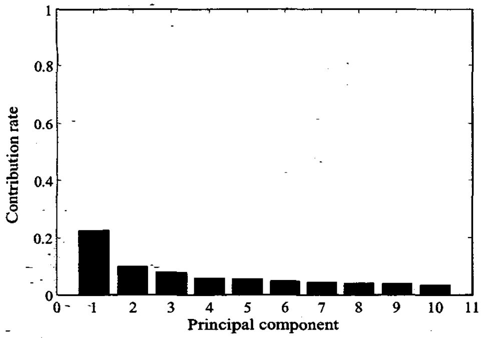


Figure 9: Contribution rate of the first ten principal components by the principal component analysis.

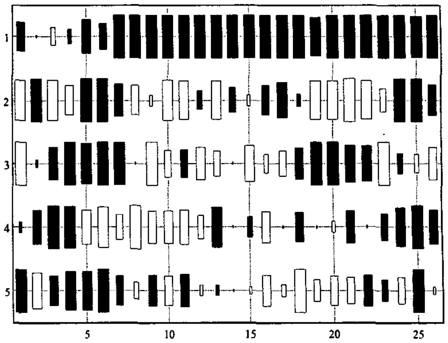


Figure 10: Category scores of the first five principal components.

## 5. REFERENCES

[1] R. Kamimura and T. Kamimura, "Greedy information algorithm," in *Proceedings of International Joint Conference on Neural Networks*, 2002.

[2] R. Kamimura, T. Kamimura, and T. R. Shultz, "Information theoretic competitive learning and linguistic rule acquistion," *Transactions on the Japanese Society for Artificial Intelligence*, vol. 16, no. 2, pp. 287–298, 2001.

[3] L. L. Gatlin, *Information Theory and Living Systems.* Columbia University Press, 1972.