Proceedings of the 2007 IEEE Symposium on
Foundations of Computational Intelligence (FOCI 2007)

# Forced Information and Information Loss for a Student Survey Analysis

Ryotaro Kamimura
Information Science Laboratory,Information Technology Center
1117 Kitakaname Hiratsuka Kanagawa 259-1292, Japan
ryo@cc.u-tokai.ac.jp

## Abstract

In this paper, we propose a new computational method called *forced information* to accelerate learning and a new method called *information loss* to extract important features for information-theoretic competitive learning. Information-theoretic learning has been proposed to solve the fundamental problems of competitive learning with many applications. However, one of the main problems is that it is slower as a problem becomes more complex. To solve this problem, we introduce forced information in which information is supposed to be maximized before learning. In addition, we introduce information loss that measures the importance of input variables. The information loss is defined by difference between information content with a unit and without the unit. We apply the method to a student survey analysis. Experimental results show that learning is accelerated significantly by the forced information. Clear features are extracted over connection weights. In addition, distinctive features are extracted by the information loss. Thus, information-theoretic learning, so far confined in relatively small problems, can be applied to large and practical problems.

**Keywords:** mutual information maximization, competitive learning, forced information, information loss, winner-take-all

## 1 Introduction

In this paper, we propose a new acceleration method called *forced information* and a new method called *information loss* to extract important variables for information-theoretic competitive learning. Information-theoretic competitive learning has been developed to solve the fundamental problems of conventional competitive learning [1], [2]. For example, in conventional competitive learning, dead neurons become a serious problem, and initial conditions surely affect final competitive unit activations [3], [4], [5], [6], [7], [8], [9]. These problems have been solved by introducing mutual information in competitive learning, because in maximizing information, entropy of competitive units must be increase as much as possible. In a maximum entropy state, all competitive units must be equally activated, meaning that no dead neurons can be produced. However, one of the main problems is that it is slow in learning when the problem becomes complex by using information-theoretic learning. To overcome this problem, we propose a forced-information method in which information is supposed to be maximized before learning. By this supposition, we can significantly accelerate a process of information maximization.

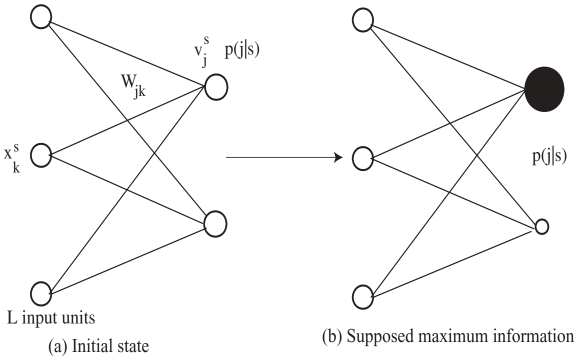


Figure 1. Supposed maximum information to accelerate a process of information maximization.

In competitive learning, little attempts have been made to extract important variables. However, to interpret final representations created by competitive learning, it is necessary to develop a method to extract important input variables. For this, we propose an information loss, that is, difference in information content for a network with an input unit and without the unit. If difference is large, the unit plays a very important role in learning. By information change, we can infer the importance of input variables.

By the forced information and information loss, we can accelerate significantly competition processes by information maximization and we can extract important input variables. Thus, information-theoretic methods are expected to be applied to practical and large-scale problems.

## 2 Theory and Computational Methods

### 2.1 Information Maximization

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 a system is represented by decrease in uncertainty [11]. Uncertainty decrease, that is, information $I$, is defined by

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

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

Let us present update rules to maximize information content. As shown in Figure 1, a network is composed of input units $x_k^s$ and competitive units $v_j^s$. We used as the output function the inverse of the square of the Euclidean distance between connection weights and outputs for facilitating the derivation. Thus, distance is defined by

$$d_j^s = \sum_{k=1}^{L}(x_k^s - w_{jk})^2, \tag{2}$$

where $L$ is the number of input units, and $w_{jk}$ denote connections from the $k$th input unit to the $j$th competitive unit. An output from the $j$th competitive unit can be computed by

$$v_j^s = \frac{1}{d_j^s}. \tag{3}$$

The output is increased as connection weights are closer to input patterns.

The conditional probability $p(j \mid s)$ is computed by

$$p(j \mid s) = \frac{v_j^s}{\sum_{m=1}^{M} v_m^s}, \tag{4}$$

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

$$p(j) = \frac{1}{S}\sum_{s=1}^{S} p(j \mid s). \tag{5}$$

Information $I$ is computed by

$$\begin{aligned} I &= -\sum_{j=1}^{M} p(j)\log p(j) \\ &\quad + \frac{1}{S}\sum_{s=1}^{S}\sum_{j=1}^{M} p(j \mid s)\log p(j \mid s). \end{aligned} \tag{6}$$

Differentiating information with respect to input-competitive connections, we have final update rules to maximize information

### 2.2 Maximum Information-Forced Learning

One of the major shortcomings of information-theoretic competitive learning is that it is sometimes very slow in increasing information content to a sufficiently large level. We here present how to accelerate learning by supposing that information is already maximized before learning. Figure 1 shows two initial states for learning. Figure 1(a) shows an initial state for conventional methods. In this case, two competitive neurons are equally activated. On the other hand, in Figure 1(b) shows an state for forced information learning. As can be seen in the figure, at the initial stage, one competitive unit is strongly activated, meaning that maximum information is already achieved. Thus, we have a conditional probability $p(j|s)$ such that the probability is set to $\epsilon$ for a winner, and $(1-\epsilon)/(M-1)$ for all the other units. We here suppose that $\epsilon$ is supposed to be close to unity. Weights are updated so as to maximize usual information content. The conditional probability $p(j \mid s)$ is computed by

$$p(j \mid s) = \frac{v_j^s}{\sum_{m=1}^{M} v_m^s}, \tag{7}$$

where $M$ denotes the number of competitive units.

$$p^\epsilon(j \mid s) = \begin{cases} \epsilon & \text{for a winner} \\ \frac{1-\epsilon}{M-1} & \text{otherwise} \end{cases} \tag{8}$$

Using these forced probabilities, we have forced information

$$\begin{aligned} I^\epsilon &= -\sum_{j=1}^{M} p^\epsilon(j)\log p^\epsilon(j) \\ &\quad + \frac{1}{S}\sum_{s=1}^{S}\sum_{j=1}^{M} p^\epsilon(j \mid s)\log p^\epsilon(j \mid s). \end{aligned} \tag{9}$$

This approach seems to be one in which the main merit of standard competitive learning is incorporated in a comprehensive way.

### 2.3 Information Loss

To interpret final representations of competitive learning, we must examine which features play more important roles in learning. For this, we now define information when a neuron is damaged by some reasons. In this case, distance without the $m$th unit is defined by

$$d_{jm}^s = \sum_{k\neq m}(x_k^s - w_{jk})^2, \tag{10}$$

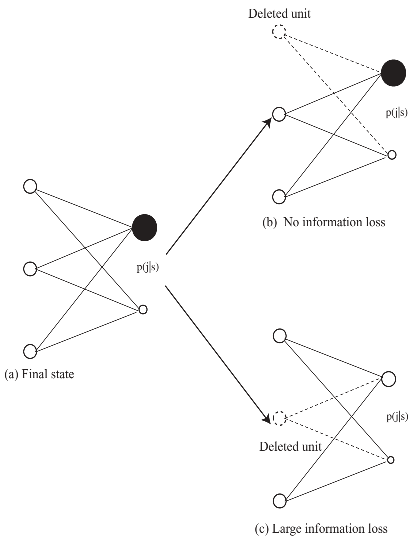


Figure 2. Maximum information state and how to measure information loss.

where summation is over all input units except the $m$th unit. The output without the $m$th unit is defined by

$$v_{jm}^s = \frac{1}{d_{jm}^s}. \tag{11}$$

The normalized output is computed by

$$p^m(j \mid s) = \frac{v_{jm}^s}{\sum_{l=1}^{M} v_{lm}^s}. \tag{12}$$

Now, let us define mutual information without the $m$th input unit by

$$\begin{aligned} I^m &= -\sum_{j=1}^{M} p^m(j) \log p^m(j) \\ &\quad + \sum_{s=1}^{S} \sum_{j=1}^{M} p(s) p^m(j \mid s) \log p^m(j \mid s), \end{aligned} \tag{13}$$

where $p^m$ and $p^m(j \mid s)$ denote a probability and a conditional probability without the $m$th input unit, given the $s$th input pattern. Information loss is defined by difference between original mutual information with full units and mutual information without a unit. Thus, we have the information loss

$$IL^m = I - I^m. \tag{14}$$

Figure 2 shows a concept of information loss. In the first place, information is supposed to be maximized in Figure 2(b). Then, the first input unit is deleted as shown in Figure 2(c). Without this input unit, no change in firing probabilities of competitive units occurs. This means that the first input unit does not play an important role in information acquisition. On the other hand, when the second input unit is deleted in Figure 2(d), we can see a drastic change in firing probabilities. This means that this input unit plays a very important role in learning.

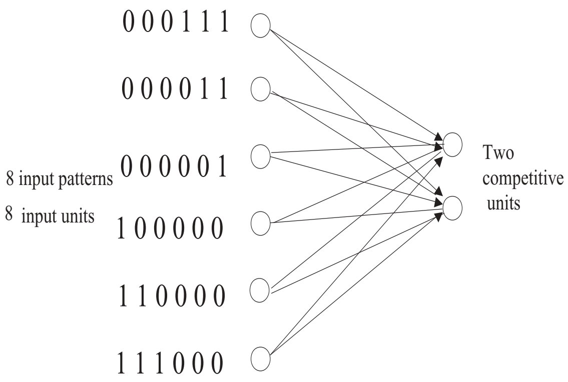


Figure 3. A network architecture for the artificial data.

## 3 Results and Discussion

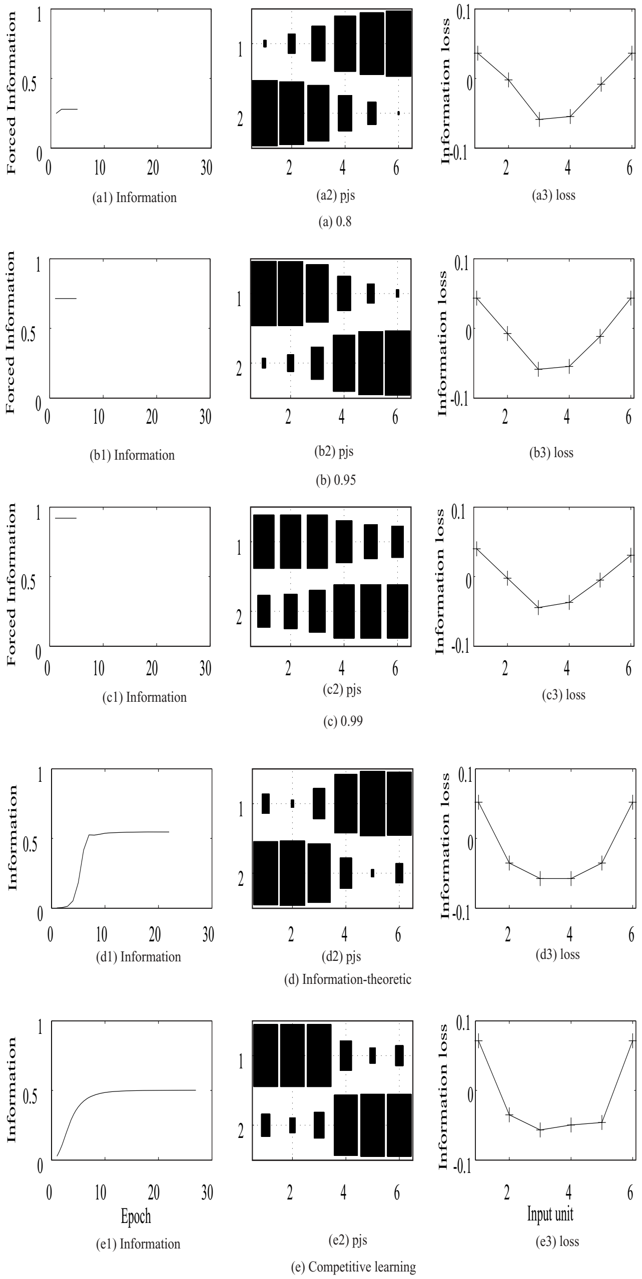
### 3.1 Symmetric Data

In this experiment, we try to show that symmetric data can easily be classified by forced information. Figure 3 shows a network architecture where six input patterns are given into input units. These input patterns can naturally be classified into two classes.

Figure 4 shows information, forced information, probabilities and information losses for the symmetric data. When the constant $\epsilon$ is set to 0.8, information reaches a stable point with eight epochs. When the constant is increased to 0.95, just one epoch is enough to reach that point. However, when information is further increased to 0.99, information reaches easily a stable point, but obtained probabilities show rather ambiguous patterns. Compared with forced information, information-theoretic learning needs more than 20 epochs and as many as 30 epochs are needed by competitive learning. We could obtain almost same probabilities $p(j \mid s)$ except $\epsilon = 0.99$. For the information loss, the first and the sixth input patterns show large information loss, that is, important. This represents quite well symmetric input patterns.

### 3.2 Student Survey Analysis

In this experiment, we report a preliminary experimental result on a student survey. The student survey was conducted in 1996 for students who attended several courses concerning information technologies. The main objective is to know students' interest in many things such as *computer*, *sport*, *music* and so on. The number of students was 580, and the number of variables (questionnaires) was 58.



(a1) Information (a2) pjs (a3) loss

(a) 0.8

(b1) Information (b2) pjs (b3) loss

(b) 0.95

(c1) Information (c2) pjs (c3) loss

(c) 0.99

(d1) Information (d2) pjs (d3) loss

(d) Information-theoretic

(e1) Information (e2) pjs (e3) loss

(e) Competitive learning

Figure 4. Information, forced information, probabilities and information losses for the artificial data.

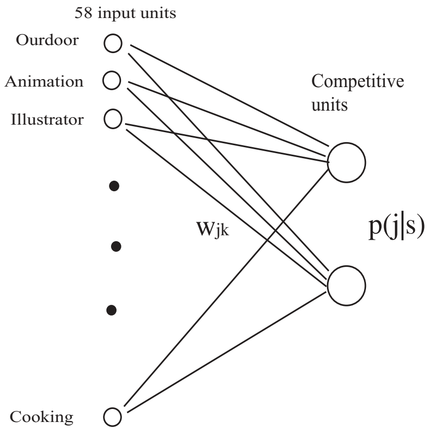


Figure 5. Network architecture for a student analysis.

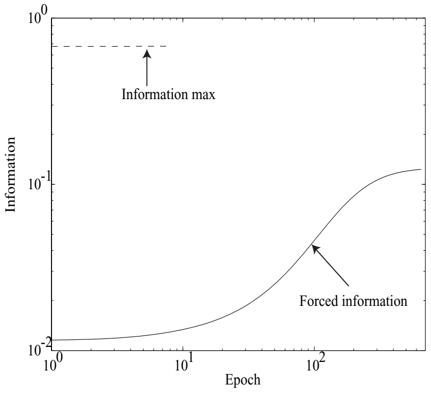


Figure 6. Information and forced information as a function of the number of epochs by the information-theoretic and forced-information method.

Figure 5 shows a network architecture with two competitive units. The number of input units is 58 units, corresponding to 58 items. The students must responds to these items with four scales. For example, 4 and 1 represent *strong interest* and *no interest*, respectively.

In the previous information-theoretic model, when the number of competitive units is large, it is sometimes impossible to attain the appropriate level of information. Figure 6 shows information as a function of the number of epochs. By using simple information maximization, we need as many as 500 epochs to be stabilized. On the other hand, by forced information, we need just eight epochs to finish learning. Thus, we can say that forced information maximization can accelerate learning almost seven times faster than the ordinary information maximization.

We examine connection weights reflecting actual responses of students. Figure 7 shows connection weights by forced information. The main difference between two groups is difference in the strength of connection weights. This means that students are divided into two groups: students with much interest in the subjects and those with

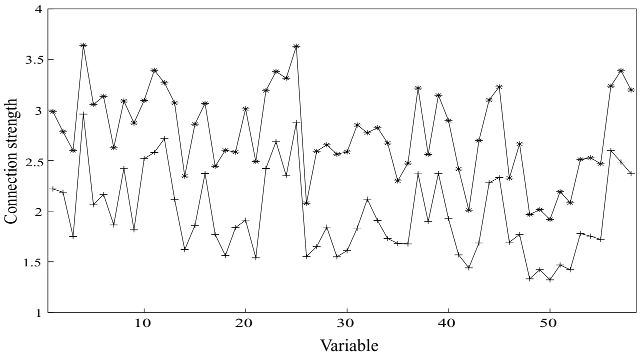


Figure 7. Connection weights for two groups by forced-information.

less interest. However, if we examine connection weights closely, we can find that two groups seem to respond to input patterns similarly. Actually, the correlation coefficient between two groups is quite high, and over 0.9. Though a network can classify students into two groups, they are similar to each other.

Table 1 shows the items in which students have some interest for a group of students with strong interest (large connection weights). As can be seen in the table, students have strong interest in something concerning pleasure or entertainment such as music, travel, movie, visual media, sport and so on. On the other hand, conventional academic subjects such as physics, statistics, and chemistry gives lower scores in the table. Table 2 shows items for students with less interest in the subjects. We can immediately see the same tendency that pleasure or entertainment are ranked top, and conventional subjects such as physics are ranked low. One of the main differences between two groups is that students with less interest have little interest in business such as marketing, management sciences, and exchange. They have strong tendency toward entertainment and pleasure.

Then, we examine the information loss. Figure 3 shows the ranking of items by information loss. The first item with the highest loss is *multimedia*. This means that two groups can be classified by this item. In addition, items concerning business such as *business*, *marketing* can be seen. This shows that two groups have been classified based upon multimedia and business. Thus, students with high interest in the subjects have strong interest in multimedia and business.

We try to interpret the information in a more concrete way. Figure 4 shows information loss and difference in magnitude of weights between two groups. Almost same patterns can be seen in two graphs except the magnitude of information loss for multimedia (No.20). Table 4 shows ranking by difference. As can be seen in the table, the ranking is quite similar to that by the information loss. Thus, the information loss represents difference between two groups in terms of the strength of connection weights.

Table 1. Ranking of items for a group of students who responded to items with a high level of interest.

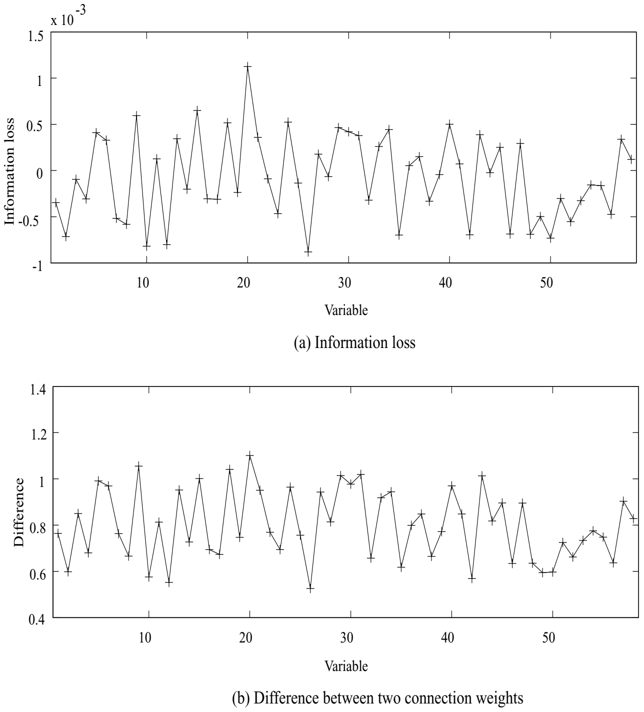
| No. | No.(Figure) | Strength | Item |
|---:|---:|---:|---|
| 1 | 4 | 3.639 | Internet |
| 2 | 25 | 3.630 | Music |
| 3 | 11 | 3.393 | Computer |
| 4 | 57 | 3.389 | Travel |
| 5 | 23 | 3.381 | Movie |
| 6 | 24 | 3.314 | Visual media |
| 7 | 12 | 3.269 | Sport |
| 8 | 56 | 3.236 | Comic |
| 9 | 45 | 3.229 | Human relations |
| 10 | 37 | 3.217 | Qualification |
| 49 | 14 | 2.347 | Trading |
| 50 | 46 | 2.327 | Mathematics |
| 51 | 35 | 2.299 | Archeology |
| 52 | 51 | 2.193 | Statistics |
| 53 | 52 | 2.083 | Physics |
| 54 | 26 | 2.078 | Chemistry |
| 55 | 49 | 2.015 | Earth science |
| 56 | 42 | 2.009 | Craftwork |
| 57 | 48 | 1.966 | Shipping |
| 58 | 50 | 1.918 | Railroad |

Table 2. Ranking of items for a group of students who responded to items with a low level of interest.

| No. | No.(Figure) | Strength | Item |
|---:|---:|---:|---|
| 1 | 4 | 2.959 | Internet |
| 2 | 25 | 2.874 | Music |
| 3 | 12 | 2.717 | Sport |
| 4 | 23 | 2.687 | Movie |
| 5 | 56 | 2.599 | Comic |
| 6 | 11 | 2.580 | Computer |
| 7 | 10 | 2.519 | Game |
| 8 | 57 | 2.486 | Travel |
| 9 | 8 | 2.423 | Entertainment |
| 10 | 22 | 2.422 | Eating and drinking |
| 49 | 18 | 1.561 | Marketing |
| 50 | 26 | 1.552 | Chemistry |
| 51 | 29 | 1.549 | Management sciences |
| 52 | 21 | 1.539 | Exchange |
| 53 | 51 | 1.468 | Statistics |
| 54 | 42 | 1.440 | Craftwork |
| 55 | 52 | 1.421 | Physics |
| 56 | 49 | 1.421 | Earth science |
| 57 | 48 | 1.331 | Shipping |
| 58 | 50 | 1.321 | Railroad |

Table 3. Ranking of information loss for two groups analysis.

| No. | No.(Figure) | Strength | Item |
|---|---|---|---|
| 1 | 20 | 0.001125 | Multimedia |
| 2 | 15 | 0.000649 | Business |
| 3 | 9 | 0.000594 | Creator |
| 4 | 24 | 0.000524 | Visual Media |
| 5 | 18 | 0.000516 | Marketing |
| 6 | 40 | 0.000501 | Photograph |
| 7 | 29 | 0.000464 | Business management |
| 8 | 34 | 0.000444 | Publicity |
| 9 | 30 | 0.000420 | Economics |
| 10 | 5 | 0.000410 | Internet business |
| 49 | 8 | -0.000581 | Entertainment |
| 50 | 46 | -0.000687 | Mathematics |
| 51 | 48 | -0.000690 | Shipping |
| 52 | 42 | -0.000695 | Craftwork |
| 53 | 35 | -0.000698 | Archeology |
| 54 | 2 | -0.000714 | Animation |
| 55 | 50 | -0.000732 | Railroad |
| 56 | 12 | -0.000802 | Sport |
| 57 | 10 | -0.000818 | Game |
| 58 | 26 | -0.000880 | Chemistry |



(a) Information loss

(b) Difference between two connection weights

Figure 8. Information loss (a) and difference between two connection weights ($w_{2k} - w_{1k}$) (b).

Table 4. Difference between two groups of students.

| No. | No.(Figure) | Strength | Item |
|---|---|---|---|
| 1 | 20 | 1.100 | Multimedia |
| 2 | 9 | 1.055 | Creator |
| 3 | 18 | 1.040 | Marketing |
| 4 | 31 | 1.020 | Arts |
| 5 | 29 | 1.014 | Business management |
| 6 | 43 | 1.014 | Information sciences |
| 7 | 15 | 1.001 | Business |
| 8 | 5 | 0.991 | Internet business |
| 9 | 30 | 0.977 | Economics |
| 10 | 40 | 0.971 | Photograph |
| 49 | 48 | 0.636 | Shipping |
| 50 | 46 | 0.634 | Mathematics |
| 51 | 35 | 0.618 | Archeology |
| 52 | 2 | 0.598 | Animation |
| 53 | 50 | 0.597 | Railroad |
| 54 | 49 | 0.595 | Earth science |
| 55 | 10 | 0.575 | Game |
| 56 | 42 | 0.569 | Craftwork |
| 57 | 12 | 0.552 | Sport |
| 58 | 26 | 0.526 | Chemistry |

Figure 9 shows a conceptual figure of final interpretation of the results. As can be seen in the figure, two groups are separated by the strength of connection weights. Because the strength of weights represent the strength of interest of students toward the items in the questionnaire, the first group represents a group of students with higher interest in the items in the questionnaire. On the other hand, the second group represents a group of students with lower interest in the items. The distinctive features to separate two groups are *multimedia* and *business*. This means that a group of students with higher interest have also some interest in multimedia and business.

## 4 Conclusion

In this paper, we have proposed forced information to accelerate learning and information loss to extract important features. We have applied mainly the methods to the student survey. We have found that students have little interest in academic subjects, and they have a strong preference toward pleasure and entertainment. One of the main point to be noted is that student with strong interest in the subjects have also some interest in multimedia and business. These results show that our method of forced information with information loss can accelerate learning and extract the main features in input patterns.

## References

[1] R. Kamimura, T. Kamimura, and O. Uchida, “Flexible feature discovery and structural information,”

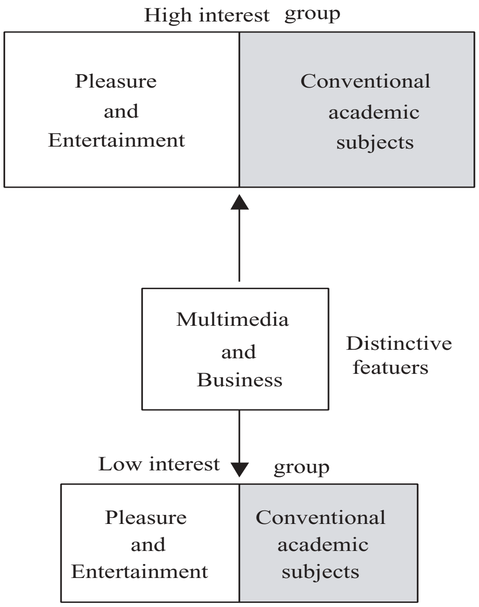

High interest group
Pleasure and Entertainment
Conventional academic subjects
Multimedia and Business
Distinctive featuers
Low interest group
Pleasure and Entertainment
Conventional academic subjects

Figure 9. Final interpretation of students'interest in the subjects.

*Connection Science*, vol. 13, no. 4, pp. 323–347, 2001.

[2] R. Kamimura, "Information-theoretic competitive learning with inverse euclidean distance," *Neural Processing Letters*, vol. 18, pp. 163–184, 2003.

[3] D. E. Rumelhart and J. L. McClelland, "On learning the past tenses of English verbs," in *Parallel Distributed Processing* (D. E. Rumelhart, G. E. Hinton, and R. J. Williams, eds.), vol. 2, pp. 216–271, Cambrige: MIT Press, 1986.

[4] S. Grossberg, "Competitive learning: from interactive activation to adaptive resonance," *Cognitive Science*, vol. 11, pp. 23–63, 1987.

[5] D. DeSieno, "Adding a conscience to competitive learning," in *Proceedings of IEEE International Conference on Neural Networks*, (San Diego), pp. 117–124, IEEE, 1988.

[6] S. C. Ahalt, A. K. Krishnamurthy, P. Chen, and D. E. Melton, "Competitive learning algorithms for vector quantization," *Neural Networks*, vol. 3, pp. 277–290, 1990.

[7] L. Xu, "Rival penalized competitive learning for clustering analysis, RBF net, and curve detection," *IEEE Transaction on Neural Networks*, vol. 4, no. 4, pp. 636–649, 1993.

[8] A. Luk and S. Lien, "Properties of the generalized lotto-type competitive learning," in *Proceedings of International conference on neural information processing*, (San Mateo: CA), pp. 1180–1185, Morgan Kaufmann Publishers, 2000.

[9] M. M. V. Hulle, "The formation of topographic maps that maximize the average mutual information of the output responses to noiseless input signals," *Neural Computation*, vol. 9, no. 3, pp. 595–606, 1997.

[10] R. Kamimura, "Improving information-theoretic competitive learning by accentuated information maximization," *International Journal of General Systems*, vol. 34, no. 3, pp. 219–233, 2006.

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