# Supervised Information-Theoretic Competitive Learning by Cost-Sensitive Information Maximization

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

***Abstract*— In this paper, we propose a new supervised learning method whereby information is controlled by the associated cost in an intermediate layer, and in an output layer, errors between targets and outputs are minimized. In the intermediate layer, competition is realized by maximizing mutual information between input patterns and competitive units with Gaussian functions. The process of information maximization is controlled by changing a cost associated with information. Thus, we can flexibly control the process of information maximization and to obtain internal representations appropriate to given problems. The new method is considered to be a hybrid model similar to the counter-propagation model, in which a competitive layer is combined with an output layer. In addition, this is considered to be a new approach to radial-basis function networks in which the center of classes can be determined by using information maximization. We applied our method to an artificial data problem, the prediction of long-term interest rates and yen rates. In all cases, experimental results showed that the cost can flexibly change internal representations, and the cost-sensitive method gave better performance than did the conventional methods.**

**Keywords: Information maximization, Gaussian, cost, error minimization, hybrid model, competitive layer, output layer**

## I. Introduction

In this paper, we propose a new supervised learning method in which information is controlled by the associated cost in the intermediate layer, and in the second layer, errors between targets and outputs are minimized. In the intermediate layer, units or neurons compete with each other by maximizing mutual information. The method is considered to be a new type of hybrid system or a new approach to radial-basis function networks. The new method can contribute to neural computing from four perspectives: (1) this is a new type of information-theoretic competitive learning; (2) the activation function is Gaussian; (3) a process of information maximization is controlled by a cost; (4) the new model is a hybrid model in which information maximization and minimization are combined; and (5) the method is considered to be a new approach to the radial-basis function networks, in which information maximization is used to determine the center of radial basis function.

First, our method is based upon a new type of information-theoretic competitive learning. We have so far proposed a new type of competitive learning based upon information-theoretic approaches [1], [2], [3], [4]. In the new approaches, competitive processes have been realized by maximizing mutual information between input patterns and competitive units. When information is maximized, just one unit is turned on, while all the others are off. Thus, we can realize competitive processes by maximizing mutual information. In addition, in maximizing mutual information, the entropy of competitive units must be maximized. This means that all competitive units must equally be used on average. This entropy maximization can realize equiprobabilistic competitive units without the special techniques that have so far been proposed in conventional competitive learning [5], [6], [7], [8], [9], [10], [11].

Second, we use in this new approach Gaussian activation functions to produce competitive unit outputs. When we first introduced information-theoretic competitive learning, we used the sigmoidal activation function $1/(1+\exp(-u))$, where $u$ is the net input to competitive units [12], [13], [14] [1], [2], [3]. When we try to increase information, the sigmoidal approaches produce strongly negative connection weights, and they can inhibit many competitive units, except some specific competitive units. Thus, it is relatively easy to increase information content. However, because strongly negative connection weights are produced almost independently of input patterns, final representations are not necessarily faithful to input patterns. Thus, we tried to use the inverse Euclidean distance between input patterns and connection weights [15]. Though this method produced faithful representations, it was sometimes very slow in learning. In particular, as problems become more complex, networks with the inverse Euclidean distance activation functions showed difficulty in increasing information to a sufficiently high level. At this point, we try to replace Euclidean distance functions by Gaussian functions, because we can easily increase information by decreasing the Gaussian width.

Third, a process of information maximization is controlled by a cost associated with information. We have observed that information maximization is achieved at the expense of similarity to input patterns. As information is increased, connection weights tend to be away from input patterns. We should say that information maximization exaggerates some parts of input patterns. This property is useful to obtaining some important features in input patterns. However, it sometime happened that obtained features did not represent faithfully input patterns. Thus, we introduce a cost that is defined as difference between

input patterns and connection weights. Then, by controlling the cost, we can control internal representations obtained by information maximization.

Fourth, our new model is a hybrid model in which information maximization and minimization are combined with each other. There have been many attempts to model supervised learning based upon competitive learning. For example, Rumelhart and Zipser [16] tried to include teacher information in competitive learning. They called this method the "correlated teacher learning" method, in which teacher information is included in input patterns. However, one of the main shortcomings of this method is that we sometimes need an overwhelmingly number of large correlated teachers to supervise learning. On the other hand, Hechet-Nielsen tried to combine competitive learning directly with error minimization procedures [17], [18] [19] in what are called "counter-propagation networks." However, error minimization procedures are realized by the gradient descent, and usually a large number of competitive units are needed. In our method, we can use the pseudo-inverse matrix operation to produce outputs, and learning is much faster than with counter-propagation.

Fifth, the new method is also considered to be a radial-basis function network approach in which the center of radial-basis function is determined by maximizing information content. The radial basis function approach has been applied to many problems, such as function approximation, speech recognition and so on, because of rapid convergence and generality [20], [21], [22]. In this paper, we use Gaussian functions, and we can consider this computational method a new approach to radial-basis function networks. One of the problems of this approach is that it is difficult to determine the center of radial-basis functions. The center has been determined by unsupervised learning methods such as K-means, competitive learning, vector quantization [23], [20], [22], [24]. Thus, our method is considered to be a new approach to the radial-basis function, in which information maximization is used to determine the center of radial-basis functions.

## II. Information Acquisition

### A. General Cost-Sensitive Information Maximization

We suppose that information on the outer environment can be obtained only at the expense of the cost associated with the acquisition process. For example, if we want to obtain some information on an object, we should illuminate it by using some energy that corresponds to the cost in information acquisition. Though information can surely be obtained with the associated cost, there have been no attempts to take into account the cost in information-theoretic approaches to neural computing. As naturally inferred, one of the most favorable situations is one in which much information is obtained with relatively small cost. Thus, our problem is to maximize information, and at the same time the associated cost should be minimized. We define this concept by the equation:

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

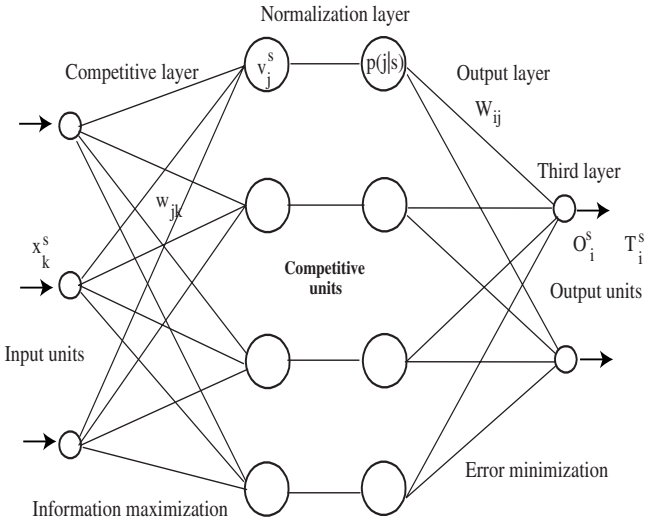


Fig. 1. A network architecture to control information content.

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. And $C$ denotes the associate cost. This equation means that information is maximized, and at the same time the associated cost must be minimized.

### B. Competition by Information Maximization

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 $v_j^s$. The $j$th competitive unit receives a net input from input units, and an output from the $j$th competitive unit can be computed by

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

where $L$ is the number of input units, $w_{jk}$ denote connections from the $k$th input unit to the $j$th competitive unit, and $\sigma$ controls the width of the Gaussian function. The output is increased as connection weights come closer to input patterns. The conditional probability $p(j \mid s)$ is computed by

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

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

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

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

$$\begin{aligned} I &= -\sum_{j=1}^{M} p(j)\log p(j) \\ &\quad + \frac{1}{S}\sum_{s=1}^{S}\sum_{j=1}^{M} p(j \mid s)\log p(j \mid s). \end{aligned} \tag{5}$$

To maximize mutual information, entropy must be maximized, and at the same time conditional entropy must be minimized. When conditional entropy is minimized, each competitive unit responds to a specific input pattern. On the other hand, when entropy is maximized, all competitive units are equally activated on average.

### C. Cost-Sensitive Information Maximization

In this paper, a cost is considered to be one representing the difference between input patterns and connection weights. Thus, a cost function is defined by

$$C = \frac{1}{2S}\sum_{s=1}^{S}\sum_{j=1}^{M} p(j \mid s)\sum_{k=1}^{L}(x_k^s - w_{jk})^2. \tag{6}$$

Thus, we must maximize the following function:

$$\begin{aligned} I &= -\sum_{j=1}^{M} p(j)\log p(j) + \frac{1}{S}\sum_{s=1}^{S}\sum_{j=1}^{M} p(j \mid s)\log p(j \mid s) \\ &\quad -\frac{1}{2S}\sum_{s=1}^{S}\sum_{j=1}^{M} p(j \mid s)\sum_{k=1}^{L}(x_k^s - w_{jk})^2. \end{aligned} \tag{7}$$

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

$$\begin{aligned} \Delta w_{jk} &= -\alpha\sum_{s=1}^{S}\left(\log p(j) - \sum_{m=1}^{M} p(m \mid s)\log p(m)\right) Q_{jk}^s \\ &\quad +\beta\sum_{s=1}^{S}\left(\log p(j \mid s) - \sum_{m=1}^{M} p(m \mid s)\log p(m \mid s)\right) \\ &\quad \times Q_{jk}^s + \frac{\gamma}{S}\sum_{s=1}^{S} p(j \mid s)(x_k^s - w_{jk}), \end{aligned} \tag{8}$$

where $\alpha$, $\beta$ and $\gamma$ are the learning parameters, and

$$Q_{jk}^s = \frac{(x_k^s - w_{jk})p(j \mid s)}{S\sigma^2}. \tag{9}$$

### D. Error Minimization

In the output layer, errors between targets and outputs are minimized. The outputs from the output layer are computed by

$$O_i^s = \sum_{j=1}^{M} W_{ij}p(j|s), \tag{10}$$

where $W_{ij}$ denote connection weights from the $j$th competitive unit to the $i$th output unit. Errors between targets and outputs can be computed by

$$E = \frac{1}{2}\sum_{s=1}^{S}\sum_{i=1}^{N}\left(T_i^s - O_i^s\right)^2, \tag{11}$$

where $T_i^s$ denote targets for output units $O_i^s$ and $N$ is the number of output units. This linear equation is directly solved by using the pseudo-inverse of the matrices of competitive unit outputs. Following the standard matrix notation, we suppose that $\mathbf{W}$ and $\mathbf{T}$ denote the matrices of connection weights and targets and $\mathbf{P}^\dagger$ shows the pseudo-inverse of the matrix of competitive unit activations. Then, we can obtain final connection weights by $\mathbf{W} = \mathbf{P}^\dagger\mathbf{T}$.

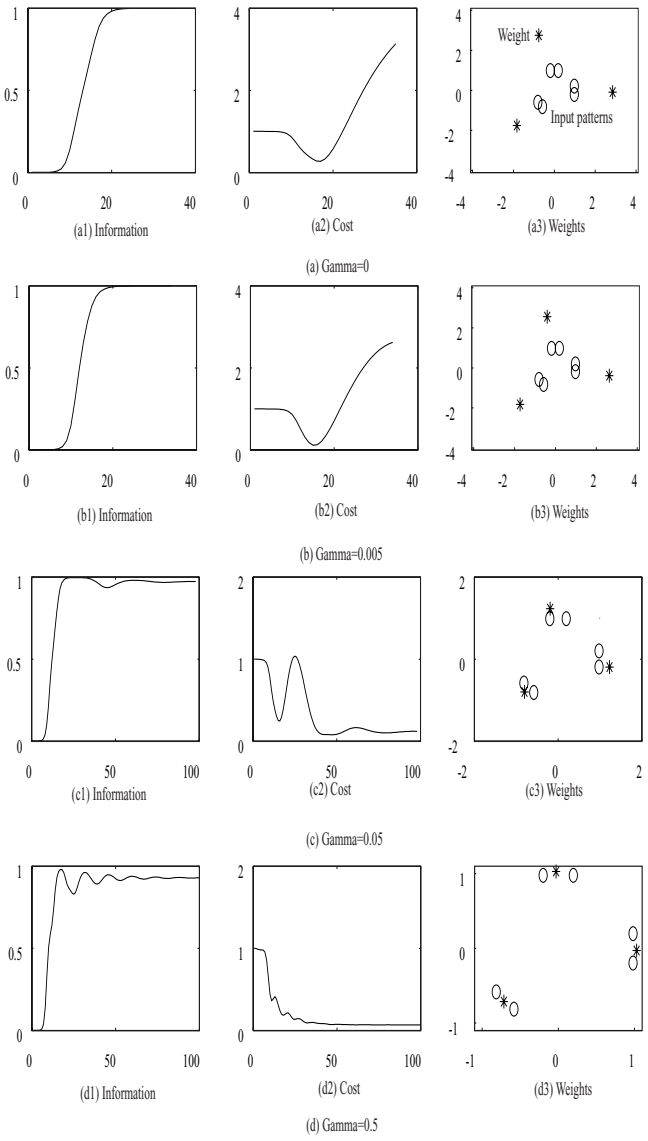


Fig. 2. Information, cost as a function of the number of epochs and connection weights: (a) $\gamma = 0$ (pure information maximization), (b) $\gamma = 0.005$, (c) $\gamma = 0.05$ and (d) $\gamma = 0.5$.

## III. Experiment No.1: Artificial Data Problem

In this section, we try to show how the cost changes final representations obtained by information maximization. The first example is a classification problem in which six patterns must be classified into three classes, as shown on the right hand side of Figure 2. Figure 2(a) shows final results by using pure information maximization. As information is increased, the associated cost is also increased. Though final connection weights classify input patterns into three classes, the weights are away from input patterns. When $\gamma$ is increased to 0.005, the cost is slightly decreased, and final connection weights are also slightly close to input patterns. When $\gamma$ is further

increased to 0.05, the cost is apparently decreased, and final connection weights are much closer to input patterns. Finally, when $\gamma$ is 0.5, information is immediately increased to almost a maximum point, and then fluctuates in the later stages of learning. The cost is decreased significantly to a smaller point, and connection weights are located in the middle of each class. These results seem to show that cost-sensitive information maximization is much better than the pure information maximization. However, the next example will explicitly show the utility of the cost as well as information maximization.

The second example is concerned with artificial data composed of common and distinctive features. We can see five input patterns given into networks in Figure 4(e). As shown in the figure, the figures are composed of distinctive features (horizontal lines) and common features (vertical lines). Figure 3 (a) shows information and cost as a function of the number of epochs. Information is rapidly increased to a maximum point with less than 10 epochs, but the associated cost decreases more slowly than the other cases. As the parameter $\gamma$ is increased from 0.0001 (Figure 3(b)) to 0.1 (Figure 3(e)), information is more slowly increased, but the cost is more rapidly decreased. Figure 4 shows connection weights obtained by changing the parameter $\gamma$. When the parameter $\gamma$ is zero, distinctive features can be obtained (Figure 4(a)). As the parameter is increased, networks tend to capture input patterns themselves. Thus, by changing the parameter $\gamma$, the properties of obtained features can flexibly be controlled.

## IV. Experiment No. 2: Long Term Interest Rate and Yen Rate Prediction

In this problem, we predict Japanese long-term interest rates and yen-to-dollar rates during 1990 and 2002. Figure 5 shows a network architecture for the prediction. The number of input units is six, representing the previous six months' rates, and the number of output unit is one, representing a rate at the current state. The number of input and competitive units were experimentally determined so as to maximize prediction performance. For example, even if we took into account more than the previous six months' rates, no improvement could been seen. We reduced the number of training patterns as much as possible. By extensive experiments, we found that 20 input patterns were the minimum number of patterns required to estimate rates by our method. Even if we increased the number of training patterns, we could not obtain better performance. On the other hand, if we decreased the number of training patterns below 20 input patterns, performance significantly degraded.

Figure 6(a) shows original long-term interest rates during 1990-2002. When the parameter $\gamma$ is 0.01[1], networks could predict the long-term interest rates quite well, as shown in Figure 6(b). When the parameter $\gamma$ is decreased to zero (Figure 6(c)), that is, pure information maximization is used, some fluctuations could be seen. However, we can say that networks still predict the long-term interest rates well. On the other hand, by using two different kinds of the radial-basis function networks (Figure 6(d) and (e)), networks failed to predict the rates. Figure 7(a) and (b) shows relations between targets and actual outputs when the parameter $\gamma$ is 0.05 and zero. The regression lines are close to lines with which outputs becomes equal to targets. However, a slightly better regression line can be obtained by the cost-sensitive method. Figure 7(c) and (d) shows regression lines by the exact and the incremental radial-basis function networks. The regression lines are far from the target lines.

Figure 8(a) shows original yen rates during 1990-2002. When the parameter $\gamma$ is 0.01, networks could predict the yen rates quite well, as shown in Figure 8(b). When the parameter $\gamma$ is decreased to zero (Figure 8), with some fluctuations, networks still predict the yen rates well. On the other hand, by using the radial-basis function networks (Figure 8(d) and (e)), networks failed to predict the rates. Figure 9(a) and (b) shows relations between targets and actual outputs when the parameter $\gamma$ is 0.05 and zero. Though two networks can predict targets well, networks with the cost-sensitive method shows

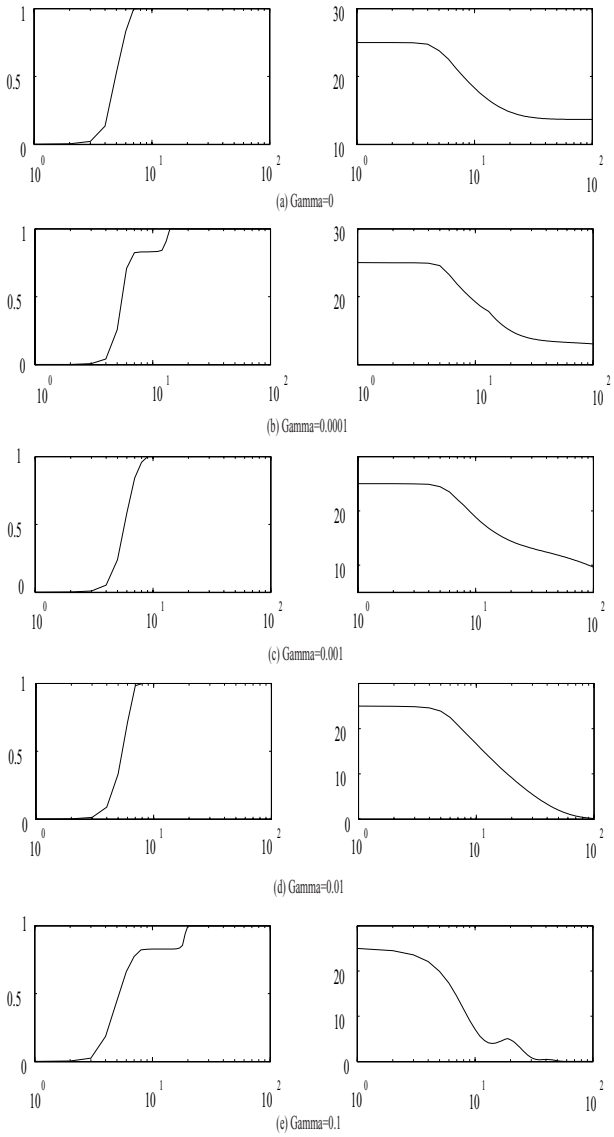


Fig. 3. Information and cost as a function of the number of epochs by four different values of the parameter $\gamma$: (a) $\gamma = 0$, (b) $\gamma = 0.0001$, (c) $\gamma = 0.001$, (d) $\gamma = 0.01$, and (e) $\gamma = 0.1$.

[1]The parameter $\alpha$ and $\beta$ were always set to 0.1.

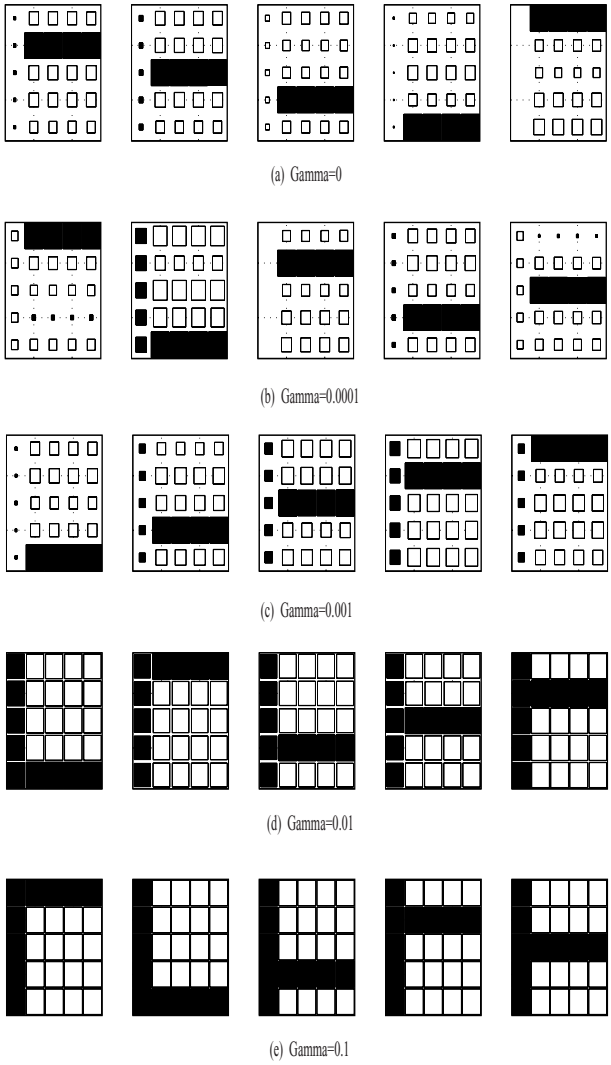


Fig. 4. Connection weights by five different parameter values: (a) $\gamma = 0$, (b) $\gamma = 0.0001$, (c) $\gamma = 0.001$, (d) $\gamma = 0.01$, and (e) $\gamma = 0.1$.

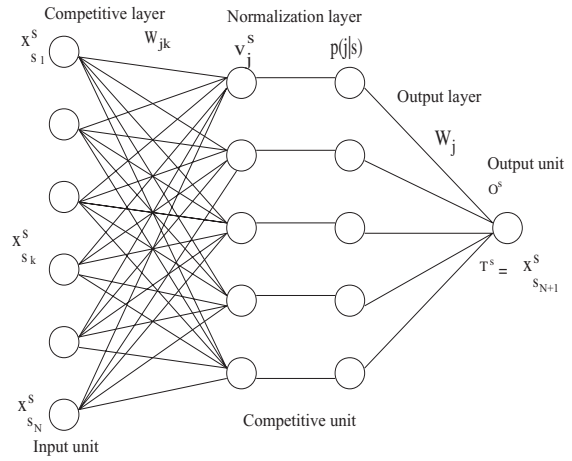


Fig. 5. A network for predicting long-term interest rates and yen rates during 1999-2002.

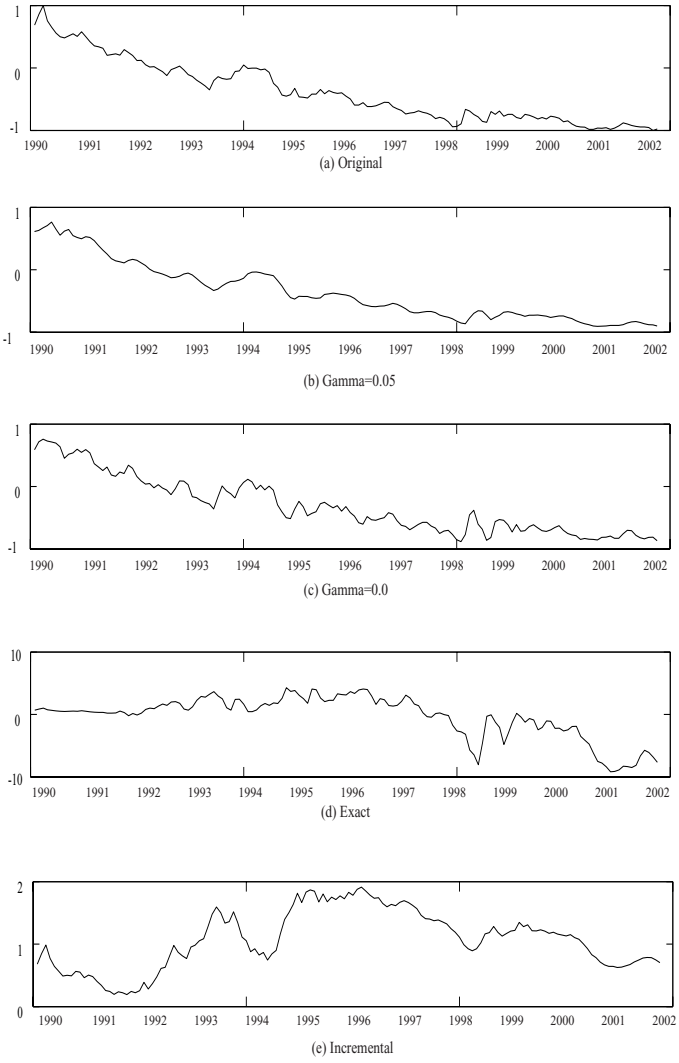


Fig. 6. Original long-term interest rates (a), estimated interest rates by the new method with $\gamma = 0.05$ (b), with $\gamma = 0$ (c), by the conventional radial basis network (exact method, $\sigma = 0.5$) (d) and the conventional radial basis method with an incremental method $\sigma = 1$ (e).

slightly better performance. Figure 9(c) shows a regression line by the exact radial-basis function networks. The regression line is far from the target line. Figure 9(d) shows a regression line by the incremental radial-basis function networks. The regression line is close to the target line, but the variation of outputs is larger.

## V. Conclusion

In this paper, we have tried to control information-theoretic competitive learning by introducing the associated cost. In our networks, we have a competitive layer, a normalization layer and an output layer. In the competitive layer, information is increased to realize competitive processes. In the normalization layer, competitive unit outputs are normalized to produce the probabilities. Then, in the output layer, errors between targets and outputs are minimized by using the least square method. In the paradigm of competitive learning, this is a hybrid model in which unsupervised and supervised learning are combined with each other, which is close to the counter-propagation networks. The difference is that in our method competitive unit

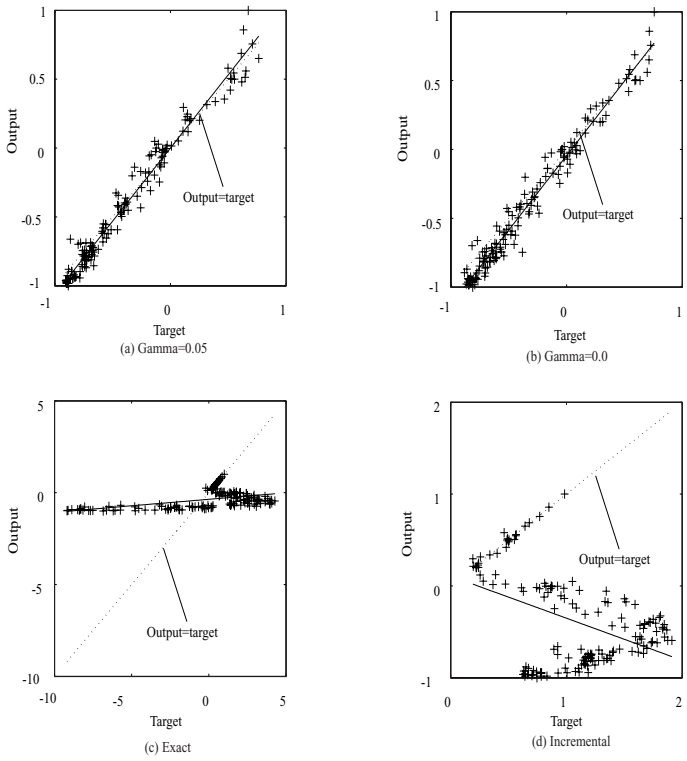


Fig. 7. Relations between targets and outputs by four methods.

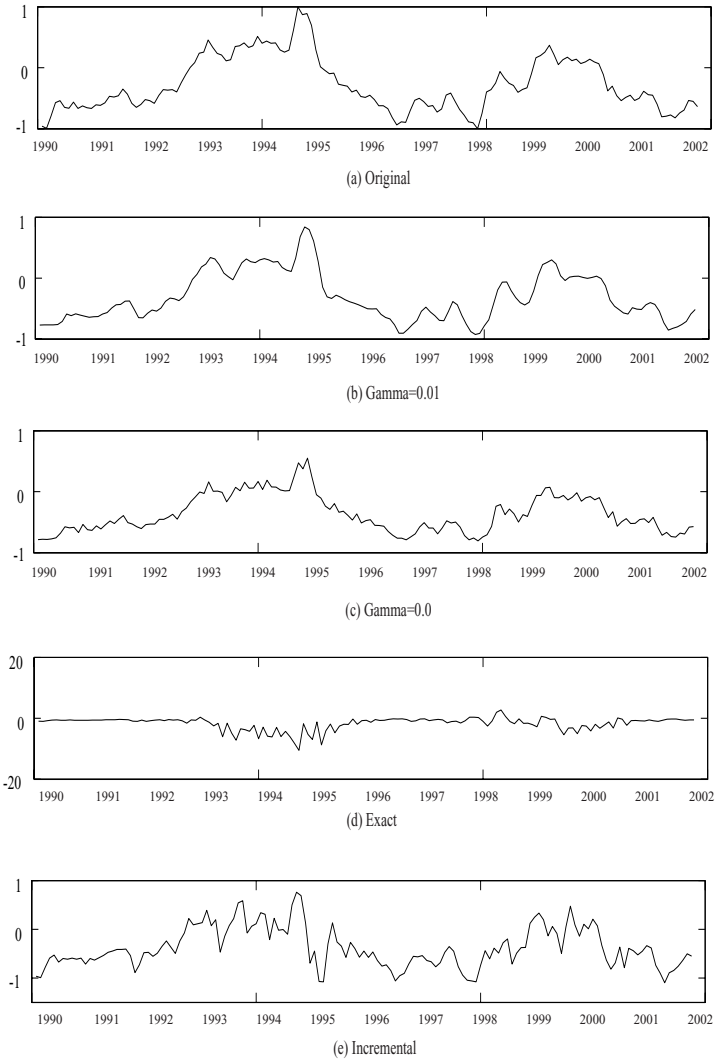


Fig. 8. Original yen rates (a), estimated rates by the new method with $\gamma = 0$ (b), $\gamma = 0.01$, $\gamma = 0.0$ (c), by the conventional radial basis network (exact method) (d) and the conventional incremental method (e).

outputs are computed by using the Gaussian functions, and in the output layer, the least square method is used. Thus, in the paradigm of the radial-basis function networks, this is a new approach to the radial-basis function to determine the center of classes. We have applied competitive learning with Gaussian functions to an artificial data problem and the prediction of long-term interest rates and yen rates. In these problems, we have shown that the new method can possibly predict future rates with a small number of past data. Finally, though some problems remain unsolved, I think that the approach outlined here is a step toward a new information-theoretic approach to neurocomputing.

## REFERENCES

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

[2] R. Kamimura, T. Kamimura, and O. Uchida, "Flexible feature discovery and structural information," *Connection Science*, vol. 13, no. 4, pp. 323–347, 2001.

[3] R. Kamimura, T. Kamimura, and H. Takeuchi, "Greedy information acquisition algorithm: A new information theoretic approach to dynamic information acquisition in neural networks," *Connection Science*, vol. 14, no. 2, pp. 137–162, 2002.

[4] R. Kamimura, "Progressive feature extraction by greedy network-growing algorithm," *Complex Systems*, vol. 14, no. 2, pp. 127–153, 2003.

[5] 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.

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

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

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] R. Kamimura and S. Nakanishi, "Improving generalization performance by information minimization," *IEICE Transactions on Information and Systems*, vol. E78-D, no. 2, pp. 163–173, 1995.

[13] R. Kamimura and S. Nakanishi, "Hidden information maximization for feature detection and rule discovery," *Network*, vol. 6, pp. 577–622, 1995.

[14] R. Kamimura, "Minimizing $\alpha$-information for generalization and interpretation," *Algorithmica*, vol. 22, pp. 173–197, 1998.

[15] R. Kamimura, "Information-theoretic competitive learning with inverse euclidean distance," *to appear in Neural Processing Letters*, 2003.

[16] D. E. Rumelhart and D. Zipser, "Feature discovery by competitive learning," *Cognitive Science*, vol. 9, pp. 75–112.

[17] R. Hecht-Nielsen, "Counterpropagation networks," *Applied Optics*, vol. 26, pp. 4979–4984, 1987.

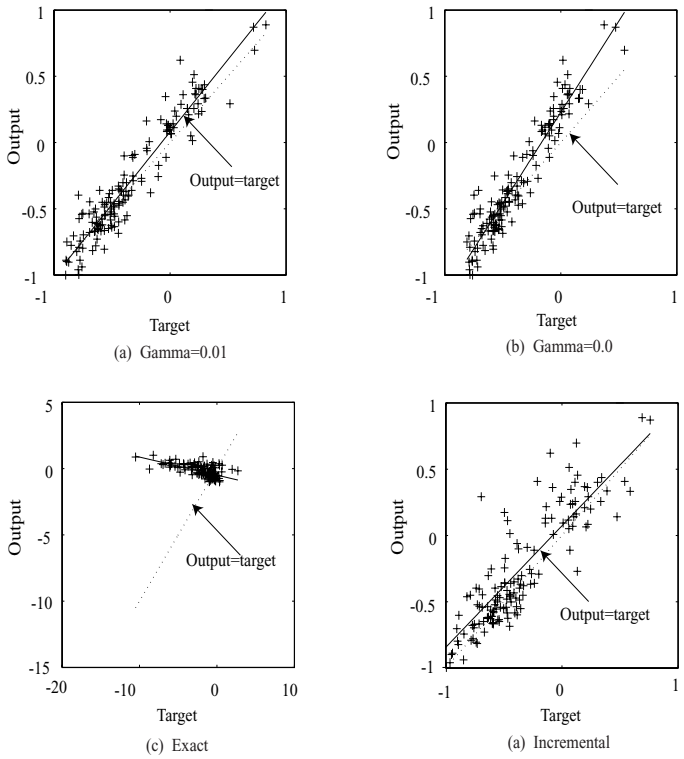


Fig. 9. Relations between targets and outputs by four methods.

[18] R. Hecht-Nielsen, "Applications of counterpropagation networks," *Neural Networks*, vol. 1, no. 2, pp. 131–139, 1988.
[19] M. Novic and J. Zupan, "Investigation of infrared spectra-structure correlation using Kohonen and counterpropagation neural network," *Journal of Chemical Information and Computer Sciences*, vol. 35, pp. 454–66, May-June 1995.
[20] J. Moody and C. J. Darken, "Fast learning in networks of locally-tuned processing units," *Neural Computation*, vol. 1, no. 2, pp. 281–294, 1989.
[21] D. F. Specht, "A general regression neural network," *IEEE Transactions on Neural Networks*, vol. 2, pp. 568–576, 1991.
[22] P. Burrascano, "Learning vector quantization for the probabilistic neural network," *IEEE Transactions on Neural Networks*, vol. 2, pp. 458–461, 1991.
[23] D. F. Specht, "A general regression neural network," *IEEE Transactions on Neural Networks*, vol. 2, pp. 568–576, Nov. 1991.
[24] D. Lowe, "Radial basis function networks," in *The Handbook of Brain Theory and Neural Networks* (M. A. Arbib, ed.), pp. 779–783, Cambridge, Massachusetts: MIT Press, 1995.