1998 International Computer Symposium
Workshop on Artificial Intelligence
December 17-19, 1998, N.C.K.U., Tainan, Taiwan, R.O.C.

Simple Structural Information to Optimize Neural Network Architectures

Ryotaro Kamimura
Information Science Laboratory, Tokai University,
1117 Kitakaname Hiratsuka Kanagawa 259-12, Japan

Abstract

In this paper, new information called *structural* information is proposed. Structural information is composed of two types of information: the first order and the second order information. The first and second order information represent information on deviation from the equiprobable distribution and the independence respectively. We have so far dealt with total information to be stored in neural networks. However, by introducing structural information, we can control appropriate types of information, depending on methods and problems. In other words, we can control information, taking into account the quality as well as the quantity of information. We applied the structural information to information content in input-hidden connections and hidden units. Then it was applied to XOR problem to show how the structural information control affects the simplification of network architectures. In addition, we applied the methods to language acquisition problems complex enough to test the performance. Experimental results confirmed that generalization is not concerned with total information but with the second order information.

1 Introduction

Many attempts have been made to describe neural learning from information theoretic points of view [1], [2]. Information, appropriately defined, has been maximized or minimized, depending on problems [2], [3]. However, we can definitely say that information is not simply controlled in living systems. To cope with extremely uncertain living conditions, information is controlled and stored in very complicated ways [4]. We think that simple information maximization and minimization so far developed are inadequate to modeling actual information control and storage in living systems. At the present stage, we should ask the quality of information, that is, what kinds of information should be stored as well as the quantity of information to be stored.

In this context, we propose structural information composed of two types of information. The structural information is applied to information content in connections. We attempt to maximize information to be stored in connections and hidden units, considering which type of information is necessary in learning.

2 Structural Information

2.1 Utility of Structural Information

Many methods to optimize the network architecutre have been proposed, for example, connection and unit pruning and weight decay. The optimization is necessary for improving generalization performance and for interpreting clearly internal representations. For example, Figure 1 shows two typical examples of weight pruning and weight decay. (a) shows that unnecessary connections are eliminated, producing an architecture in which each connection is connected with different hidden units. (b) shows that unnecessary hidden units are eliminated, giving an architecture in which just one hidden unit is obtained. Finally, (c) shows a combined method in which hidden units and connections are both eliminated.

If it is possible to move from (a) to (b) and to (c), we can generate optimal network architectures, depending upon the problems. Structural information is introduced to control freely network architectures by changing the structural parameter α.

2.2 Simple Structural Information

We are here concerned not with information to be transmitted but with stored information [4]. Thus, information is considered to be the decrease of uncertainty. As shown in Figure 2, the first order information D_1 is the decrease from the maximum uncertainty H_0. The second order information D_2 is the decrease from the first order uncertainty. Total information is obtained by summing two types of information.

Let J and K denote discrete random variables taking the values j and k with probabilities $p(j)$ and $p(k)$, and $p(j \mid k)$ represent the conditional probability of j for k, then maximum uncertainty (H_0), the first-order uncertainty (H_1) and the second-order uncertainty (H_2) are defined by

$$\begin{aligned} H_0 &= \log M \\ H_1 &= -\sum_{j}^{M} p(j) \log p(j) \\ H_2 &= -\sum_{k}^{L} \sum_{j}^{M} p(k) p(j \mid k) \log p(j \mid k). \end{aligned} \quad (1)$$

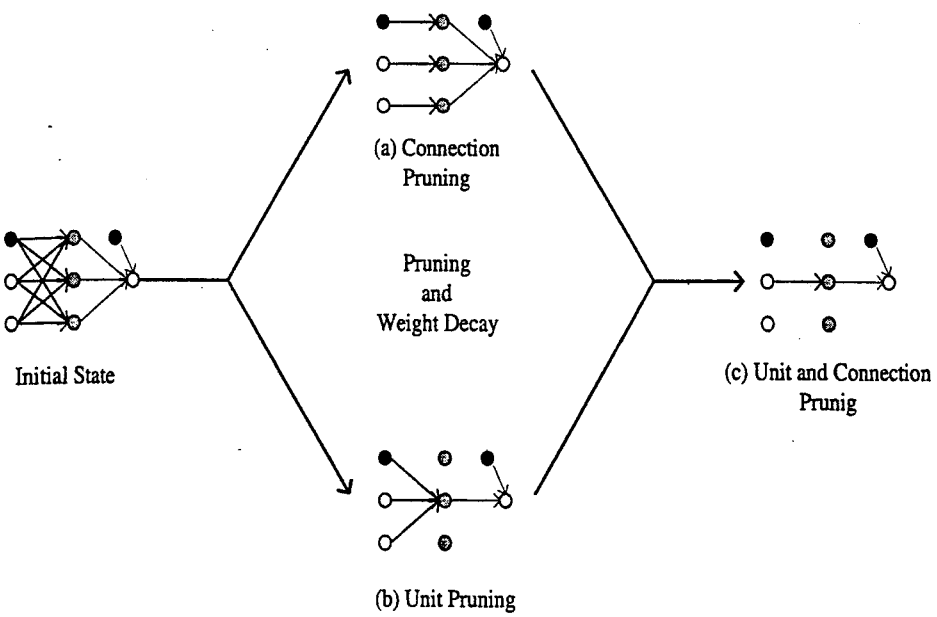


Figure 1: Three typical examples obtained by the pruning and weight decay methods. (a) shows that each connection is connected with different hidden units. (b) shows that only one hidden unit is connected with all the input-hidden connections. In (c), unit and connection elimination are combined to simplify network architectures.

The first-order information is defined by the decrease of uncertainty from maximum uncertainty

$$\begin{aligned} D_1 &= H_0 - H_1 \\ &= \log M + \sum_j p(j) \log p(j). \end{aligned} \quad (2)$$

The first order information D_1 means how much the distribution of j is deviated from the equi-probable distribution. The second-order information is defined by the decrease from the first order uncertainty

$$\begin{aligned} D_2 &= H_1 - H_2 \\ &= -\sum_j p(j) \log p(j) \\ &\quad + \sum_k \sum_j p(k) p(j \mid k) \log p(j \mid k). \end{aligned} \quad (3)$$

The second order information D_2 means deviation from the independence, that is, how much k-j pairs are related with each other. Total information is defined by

$$\begin{aligned} D &= D_1 + D_2 \\ &= H_0 - H_2 \\ &= \log M + \sum_k \sum_j p(k) p(j \mid k) \log p(j \mid k). \end{aligned} \quad (4)$$

Using the structural parameter α, structural information is defined by

$$\begin{aligned} SI &= \alpha D_1 + (1-\alpha) D_2 \\ &= \alpha \log M + (2\alpha - 1) \sum_j p(j) \log p(j) \\ &\quad + (1-\alpha) \sum_k p(k) \sum_j p(j \mid k) \log p(j \mid k) \end{aligned} \quad (5)$$

where the parameter α range between zero and one ($0 \leq \alpha \leq 1$). In this structural information, the parameter α is used to determine the ratio of each information to total information. We maximize this structural information, changing the parameter α. To interpret easily the ratio of the first and the second order information, we introduce normalized measures

$$\begin{aligned} RD_1 &= \frac{D_1}{D} \\ RD_2 &= \frac{D_2}{D}. \end{aligned} \quad (6)$$

The normalized total information is defined by the total information divided by the maximum information:

$$RD = \frac{D}{H_0}. \quad (7)$$

2.3 Generalized Structural Information

As shown in Figure 3, this second order structural information is easily extended to the nth order structural information content as follows:

$$SI^{(n)} = \sum_n \alpha_n D_n \quad (8)$$

where

$$D_n = H_{n-1} - H_n, \quad (9)$$

and

$$\sum_n \alpha_n = 1. \quad (10)$$

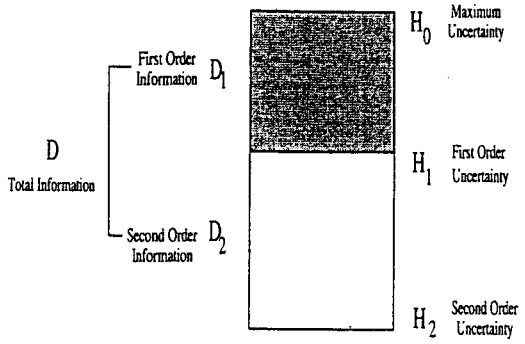


Figure 2: Structural information composed of the first order information content (D_1) and the second order information content (D_2).

The normalized nth order information is defined by

$$RD_n = \frac{D_n}{D}. \tag{11}$$

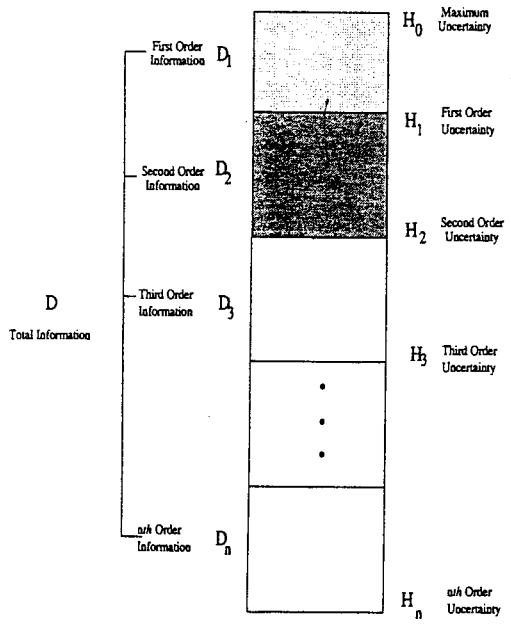


Figure 3: Generalized structural information composed of nth order information.

3 Application to Neural Networks

Weight decay or weight elimination methods have so far been used to simplify internal representations to improve generalization. This means that connections are forced to be distributed as unevenly as possible. In terms of information, information in connections is increased as much as possible.

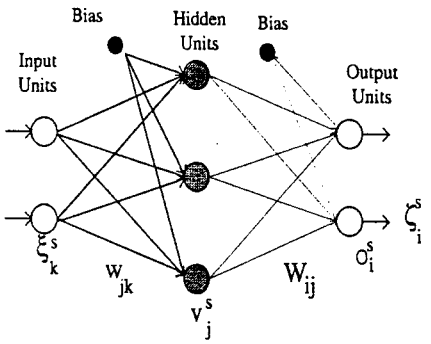


Figure 4: Network architecture for defining structural information.

We consider a network architecture shown in Figure 4. Let us define information for input-hidden connections and attempt to maximize it, that is, to distribute connections as unevenly as possible. The strength of connections w_{jk} naturally denotes the degree of relation between the jth hidden unit and the kth input unit. The strength of connections can be computed by squared connections w^2_{jk}. For probabilistic interpretation, we must normalize input-hidden connections

$$p(j \mid k) = \frac{w^2_{jk}}{\sum_m w^2_{mk}}. \tag{12}$$

Then, a probability of the hidden unit $p(j)$ can be computed by

$$p(j) = \sum_k^L p(k)p(j \mid k). \tag{13}$$

To simplify computation, let us suppose that

$$p(k) \approx \frac{1}{L}. \tag{14}$$

Thus this supposition is purely for simplification in computation. However, we can interpret it as a statement that no information on input units should be incorporated as the initial stage. Information on input units can be inferred from hidden units. Then, we have

$$p(j) \approx \frac{1}{L}\sum_k^L p(j \mid k). \tag{15}$$

Thus, the first order information is approximated by

$$D_1 \approx \log M + \sum_j^M p(j)\log p(j). \tag{16}$$

By definition, the first order information measures deviation from the equi-probable distribution. We can also say by definition that the information represents to what extent specific hidden units are affected by input units

on average. The second order information D_2 is approximated by

$$D_2 \approx -\sum_{j=1}^{M} p(j)\log p(j) + \sum_k \frac{1}{L}\sum_j p(j \mid k)\log p(j \mid k). \quad (17)$$

The second order information represents deviation from the independence. As the second order information is larger, specific pairs of input and hidden units are strongly connected, while all other connections are close to zero. Structural information is approximated by

$$\begin{aligned} SI &= \alpha D_1 + (1-\alpha)D_2 \\ &\approx \alpha \log M + (2\alpha - 1)\sum_j p(j)\log p(j) \\ &\quad +(1-\alpha)\sum_k \frac{1}{L}\sum_j p(j \mid k)\log p(j \mid k). \end{aligned} \quad (18)$$

Differentiating structural information with respect to input-hidden connections w_{jk}, we have

$$\begin{aligned} \frac{L}{2}\frac{\partial SI}{\partial w_{jk}} &= (2\alpha-1)\log p(j)Q_{jk} \\ &\quad -(2\alpha-1)\sum_m p(m \mid k)\log p(m)Q_{jk} \\ &\quad +(1-\alpha)\log p(j \mid k)Q_{jk} \\ &\quad -(1-\alpha)\sum_m p(m \mid k)\log p(m \mid k) \\ &\quad \times Q_{jk} \end{aligned} \quad (19)$$

where

$$Q_{jk} = \frac{w_{jk}}{\sum_m w_{mk}^2} \quad (20)$$

Input-hidden connections are updated so as to maximize information.

In addition to controlling information, errors (represented in a cross entropy in this paper) between targets and outputs should also be minimized.

4 Application to XOR Problem

The simple XOR problem was used to demonstrate the performance of the structural information control. First, we examine whether the parameter α can efficiently be used to control structural information. Figure 5 shows the normalized second order information RD_2 as a function of the parameter α. The second order information is decreased as the parameter α is increased. When the parameter α is zero, only the second order information D_2 can be increased. On the other hand, when the parameter α is one, the first order information D_1 can be increased. Thus, as the parameter α is increased from zero to one, the second order information D_2 is expected to be decreased as demonstrated by Figure 5.

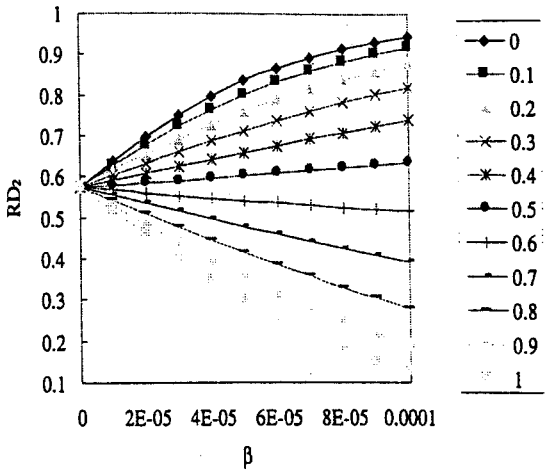


Figure 5: The second order information for ten different values of the parameter α represented on the right hand side of the figure. In the figure, β is the learning parameter.

Then, keeping the parameter α a constant, the learning parameter β is increased as much as possible at the expense of the increase in training errors. As shown in Figure 6, an original network architecture in (a) can be transformed to (b) and (c) by controlling structural information. Figure 6 (b) shows an internal representation when the parameter α is set to zero, that is, only the second order information is used to control information. As can be seen in the figure, the bias, the first, and the second input unit are connected with the first, the second, and the third hidden unit. Completely specialized connections are generated.

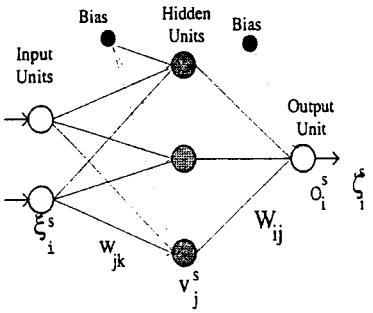
On the other hand, Figure 6(c) shows an internal representation when the parameter α is set to one, that is, only the first order information is used to control information. As can be seen in the figure, just one hidden unit is connected with all input units and bias, while all other hidden units are not used.

In both cases, the number of input-hidden connections is the same. However, completely different internal representations are obtained. Structural information can freely control internal representations: concentrated to specialized internal representations.

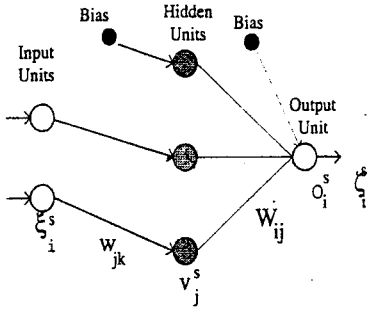
5 Application to Language Acquistion

5.1 Consonant Order Detection

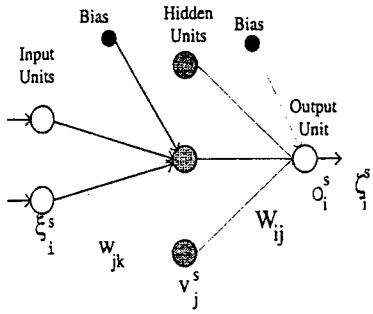
We evaluated relation between information and generalization performance by using complicated problems concerning language acquisition. The problems were complex enough to test the performance, because the actual existence of consonant clusters as well as the simple theoretical inference must be inferred by neural networks.



(a) Original Architecture

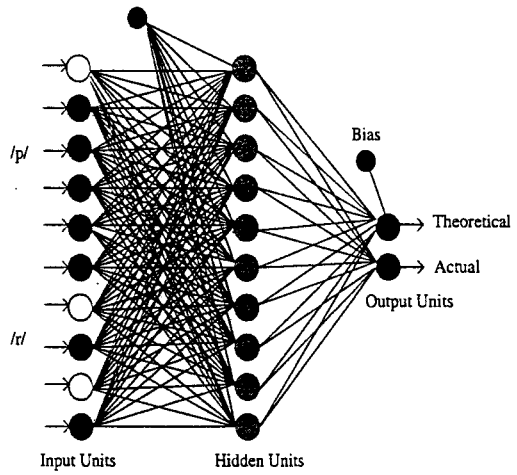


(b) $\alpha = 0$

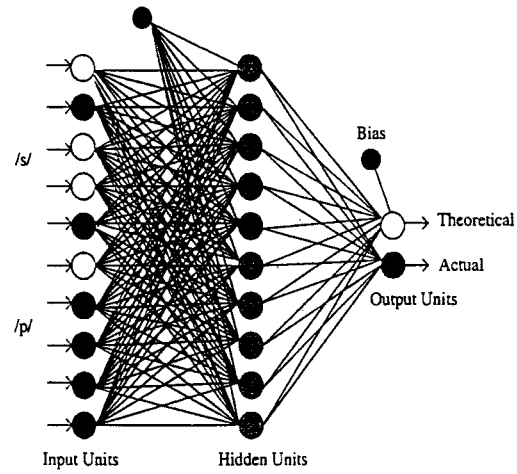


(c) $\alpha = 1$

Figure 6: Internal representations generated by structural information control. (a) depicts an original network architecture. (b) and (c) show representations for the structural parameter $\alpha = 0$, and 1.0.



(a)/*pr*/



(b) /sp/

Figure 7: Networks to infer the theoretical and actual existence of consonant clusters: (a) */pr/* and (b) */sp/*.

In natural languages, the combination of consonants is regulated by some rules. For example, the combination has been said to be regulated by a sonority principle. In actual natural languages, for example, in English, many exceptional cases to the sonority principle, have been observed. For example, a consonant cluster */sp/* (*speech*) is well-formed in English. However, by the sonority principle, this is not well-formed. In experiments, in addition to the theoretical inference by the sonority principle the inference of the actual existence of consonant clusters are incorporated. Figure 7 explains the inference for the experiments. A consonant cluster */pr/* is theoretically possible and actually exists. Thus, a network must be trained to produce an output string 11, as shown in Figure 7(a). The first bit and the second bit in 11 represent the theoretical and the actual possibility. The consonant cluster */sp/* is well-formed in spite of the theoretical impossibility. Thus, the network should produce 01, as shown in Figure 7(b).

The number of input, hidden and output unit were 10, 10 and 2 units respectively. The number of training, validation and testing patterns were 50. The fifty training patterns were so small that by using the standard BP method, at the extremely earlier stage of learning over-training occurred.

We increased total information D as much as possible. For almost all values of the structural parameter α, total information D can be close to a maximum value. However, we could see that generalization errors are approximately independent of total information D. Figure 8 shows generalization errors as a function of normalized total information RD. Values in the figure show the structural parameter α. As shown in the figure, generalization errors are independent of total information. However, we can see that as the structural parameter is smaller, generalization errors are smaller.

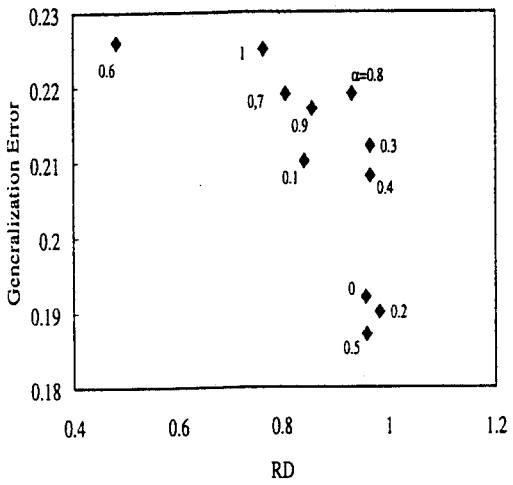


Figure 8: Generalization and total information (RD).

Figure 9 shows generalization errors as a function of the normalized second order information RD_2. We can detect two distinct groups. In one group, the second order information RD_2 is small, and generalization errors are relatively high. On the other hand, in another group, the second order information RD_2 is relatively high, and generalization errors are small. These results suggest that generalization errors are smaller, as the second order information is larger.

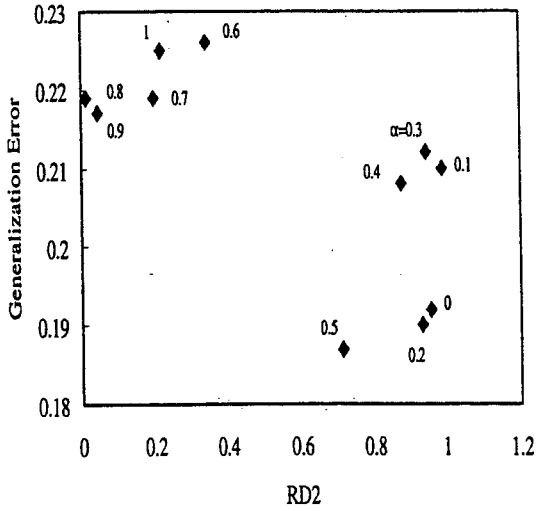


Figure 9: Generalization and the normalized second order information RD_2.

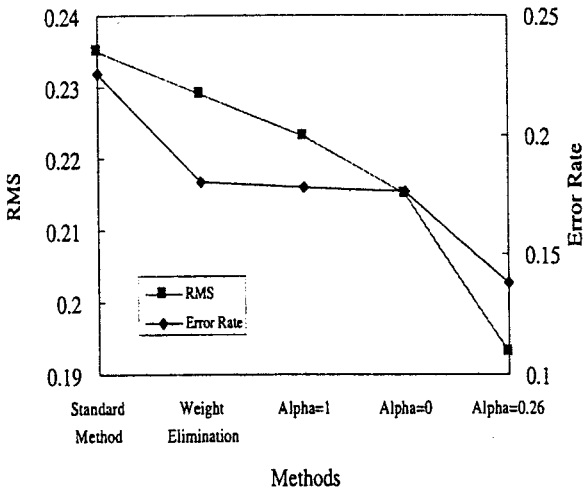


Figure 10: Generalization comparison with five methods: standard BP methods, weight elimination methods [5] and structural information methods ($\alpha = 0, 1, 0.26$).

We compared generalization performance by controlling structural information with the performance by other traditional methods. Figure 10 shows generalization errors by standard BP methods, weight elimination methods [5], methods to control structural information. As shown in the figure, generalization errors are largest (*RMS*: 0.235; error rate: 0.226) by standard BP methods. All values in the figure were averages over ten different runs. By using the weight elimination by Weigned

et al. [5] with good reputation for improved generalization, generalization errors are slightly decreased to 0.229 in *RMS* of *RMS* (error rate: 0.18). When the structural parameter α is set to one, that is, only the first order information D_1 is maximized, generalization errors are slightly decreased to 0.223 in *RMS* (error rate: 0.178). When the structural parameter α is zero, that is, the second order structural information is exclusively maximized, generalization errors are further decreased to 0.215 in *RMS* (error rate: 0.176). Finally, the lowest level of generalization errors (*RMS*: 0.193; error rate: 0.138) is obtained when the average parameter value is 0.26.

6 Conclusion

We have introduced the structural information in order to control freely the simplification process of network architectures. The structural information is composed of two kinds of information: the first and the second order information. The first and the second order information represent the distribution of hidden units and input hidden connections. By changing the structural parameter α, we can control a simplification process.

As mentioned, the simple structural information can be extended to the more general structural information, composed of nth order information. With this generalized version of the structural information, network architectures can more freely be controlled.

References

[1] H. B. Barlow, "Unsupervised learning," *Neural Computation*, vol. 1, pp. 295–311, 1989.

[2] G. Deco, W. Finnof, and H. G. Zimmermann, "Unsupervised mutual information criterion for elimination in supervised multilayer networks," *Neural Computation*, vol. 7, pp. 86–107, 1995.

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

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

[5] A. S. Weigend, D. E. Rumelhart, and B. A. Huberman, "Generalization by weight-elimination with application to forecasting," in *Neural Information Processing Systems*, vol. 3, (San Mateo: CA), pp. 875–882, Morgan Kaufmann Publishers, 1991.