Hindawi Publishing Corporation
Mathematical Problems in Engineering
Volume 2016, Article ID 3015087, 17 pages
http://dx.doi.org/10.1155/2016/3015087

Research Article

Simplified Information Maximization for Improving Generalization Performance in Multilayered Neural Networks

Ryotaro Kamimura

IT Education Center and School of Science and Technology, Tokai University, 1117 Kitakaname, Hiratsuka, Kanagawa 259-1292, Japan

Correspondence should be addressed to Ryotaro Kamimura; ryo@keyaki.cc.u-tokai.ac.jp

Received 30 July 2015; Accepted 21 February 2016

Academic Editor: Antonino Laudani

A new type of information-theoretic method is proposed to improve prediction performance in supervised learning. The method has two main technical features. First, the complicated procedures used to increase information content are replaced by the direct use of hidden neuron outputs. Information is controlled by directly changing the outputs of the hidden neurons. In addition, to simultaneously increase information content and decrease errors between targets and outputs, the information acquisition and use phases are separated. In the information acquisition phase, the autoencoder tries to acquire as much information content on input patterns as possible. In the information use phase, information obtained in the acquisition phase is used to train supervised learning. The method is a simplified version of actual information maximization and directly deals with the outputs from neurons. The method was applied to the three data sets, namely, Iris, bankruptcy, and rebel participation data sets. Experimental results showed that the proposed simplified information acquisition method was effective in increasing the real information content. In addition, by using the information content, generalization performance was greatly improved.

1. Introduction

(1) Information-Theoretic Methods. Information-theoretic methods in neural networks have received due attention ever since Linsker stated the so-called "InforMax" principle in living systems [1–4]. The Informax principle holds that living systems try to maximize information content at every stage of information processing. In other words, living systems should acquire as much information as possible in order to maintain their existence. Following this principle, there have been many attempts to use information-theoretic methods in neural networks [5–9]. Following this was the development of information-theoretic methods to control hidden neuron activation, with the aim of interpreting internal representations as much as possible and examining relations between information and generalization [10–15]. The method was successful in increasing information content, keeping training errors between targets and outputs relatively small. However, there were several limitations of these information-theoretic methods. Among them, the inability to increase information, computational complexity, and compromise between information maximization and error minimization were the most serious ones.

First, several cases were observed where the information-theoretic methods did not necessarily succeed in increasing information content. For example, when the number of neurons increases, the adjustment among neurons becomes difficult, preventing the neural networks from increasing information content. Second, there is the problem of computational complexity. As expected, information or entropy functions require complex learning formulas. This suggests that information-theoretic methods can be effective only for the relatively small sized neural networks. Third is the problem of compromise between information maximization and error minimization. From an information-theoretic point of view, information on input patterns should be increased as much as possible. On the other hand, neural networks should minimize errors between targets and outputs. Because information maximization and error minimization are sometimes contradictory, it can be difficult to compromise between the two in one framework.

(2) Simplified Methods. To solve the above-mentioned problems, a new information-theoretic method is here proposed to simplify the procedure of information maximization. The proposed procedure is composed of two steps, namely, realization of information maximization by directly controlling hidden neuron outputs and the separation of the information acquisition and use phases.

First, the process of information maximization can be realized by simulating the actual process of information maximization. In the information-theoretic method, when information increases, a small number of hidden neurons are activated. This number should be decreased in the course of learning as much as possible in order to increase the information. For this purpose, hidden neurons are arranged according to the magnitude of their variance. Specifically, hidden neurons with larger variance are more strongly activated. Much importance is placed on neurons with larger variances. This direct use of outputs can also facilitate the process of information maximization and reduce computational complexity.

Second, the information acquisition and use phases are separated. This is because it has been difficult to increase information maximization and achieve error minimization at the same time. First, information content in input patterns is acquired. Information content is then used to train supervised neural networks. This eliminates the contradiction between information maximization and error minimization in the same learning processes. The effectiveness of separation has been demonstrated in the field of deep learning [16–19].

Finally, relations between the present method and sparse coding should be noted as well. In deep learning, sparse representations play an important role when only a small number of components are nonzero and the majority is forced to be zero. Sparse coding is said to be related to improved separability and interpretation and is biologically motivated [20–26]. One of the main differences between the present method and sparse coding methods is that sparse coding usually aims to suppress the majority of components and eventually realize a small number of nonzero components. On the other hand, the present method aims only to find a small number of important components and, eventually, make the majority of components zero. In terms of detecting important components, the present method is an active one, while the others are passive ones.

(3) Outline. In Section 2, the information content in hidden neurons is introduced. Then, the procedure of information maximization is simplified by directly controlling hidden neuron outputs. In Section 3, three experimental results of the Iris, bankruptcy, and rebel participation data sets are discussed. In all experimental results, it is shown that information could be increased using the present simplified method. This information increase is shown to be in direct proportion to generalization performance for higher layers in particular. Though abrupt decreases and increases in information can be observed, the simplified method can increase information for higher-layered neural networks.

2. Theory and Computational Methods

2.1. Simplified Information Maximization. Information-theoretic methods were originally developed to increase information content in hidden neurons on input patterns. Various methods have been successfully applied for increasing information content to a certain quantity [27–29]. However, these methods have been typically limited to networks with a relatively small number of hidden neurons because of the computational complexity involved. In addition, it has been observed that the obtained information content did not necessarily contribute to improved prediction performance. The present paper proposes a method to directly control the outputs from the neurons to weaken the computational complexity of the information-theoretic methods. The procedure of information maximization can be approximated by producing a smaller number of activated hidden neurons in a concrete way.

2.1.1. Information in Hidden Neurons. Though multilayered neural networks are supposed, the learning procedures are explained by using the simple layered network, because the same procedures are repeated in the multilayered networks. Let x_k^s and w_{jk} denote the kth element of the sth input pattern and connection weights from the kth input neuron to the jth hidden neuron in Figure 1; then the net input is computed by

$$u_j^s = \sum_{k=1}^{L} w_{jk} x_k^s, \tag{1}$$

where L is the number of input neurons. The output from jth hidden neuron for sth input pattern is computed by

$$v_j^s = f\left(u_j^s\right), \tag{2}$$

where the sigmoid activation function is here used. The averaged output for jth hidden neuron is defined by

$$v_j = \frac{1}{S}\sum_{s=1}^{S} v_j^s, \tag{3}$$

where S is the number of input patterns. In addition, the variance is computed by

$$V_j = \frac{1}{S}\sum_{s=1}^{S}\left(v_j^s - v_j\right)^2. \tag{4}$$

The firing probability of jth hidden neuron is obtained by

$$p\left(j\right) = \frac{v_j}{\sum_{m=1}^{M} v_m^s}. \tag{5}$$

The entropy is defined by

$$H = -\sum_{j=1}^{M} p\left(j\right)\log p\left(j\right), \tag{6}$$

where M is the number of hidden neurons. The information is defined as the decrease of entropy from its maximum value:

$$I = H^{\max} - H. \tag{7}$$

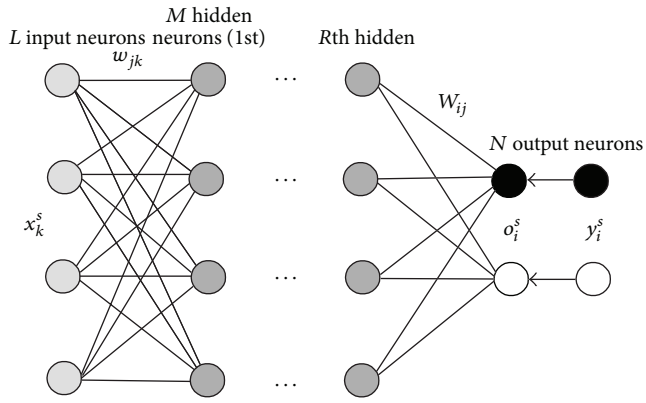


Figure 1: Network architecture for supervised learning.

2.1.2. Simplified Information Maximization. The information or entropy function in (6) can directly be differentiated. However, in actual situations, some difficulty exists in increasing the information or decreasing the entropy [27–29]. In particular, when the number of hidden neurons is large, much difficulty has been observed in increasing the information content.

Thus, this information increase is realized by using the actual outputs from hidden neurons. When the information becomes larger and entropy becomes smaller, a small number of hidden neurons tend to fire, while all the others become inactive. To realize this situation, the winners of hidden neurons are considered. The first winner is defined as a hidden neuron with the highest variance and the second winner has the second highest variance and so on. Let c_j denote the index of jth winner; then the rank order of the winners is

$$c_j = j, \quad j = 1, 2, \ldots, M, \tag{8}$$

where M is the number of hidden neurons. The winning neurons are supposed to keep the following relations:

$$V_{c_1} > V_{c_2} > V_{c_3} > \cdots > V_{c_M}. \tag{9}$$

Thus, when the variance of neurons becomes larger, the degree of winning becomes higher. For higher information, only a small number of hidden neurons fire, while all the others cease to fire. Thus, the winning neurons should have the following outputs (Figure 3):

$$\rho_j = \exp\left(-\frac{c_j - 1}{\sigma}\right), \tag{10}$$

where σ is a parameter to control the degree of winning and is larger than zero. To decrease the entropy, the following KL divergence must be decreased:

$$\mathrm{KL} = \sum_{j=1}^{M}\left[\rho_j \log\frac{\rho_j}{v_j} + \left(1-\rho_j\right)\log\frac{1-\rho_j}{1-v_j}\right]. \tag{11}$$

When the KL divergence becomes smaller, a smaller number of winning neurons tend to fire, while all other neurons become inactive.

2.2. Separation of Information Acquisition and Use Phase. Information maximization is sometimes contradictory to error minimization. This means that when maximizing information, the errors between targets and outputs cannot easily be decreased. Recently, it has been shown that unsupervised learning is effective in training multilayered networks [16–19]. Thus, the information acquisition procedure is separated from the information use. Figure 2 shows this separation. In the information acquisition phase in Figure 2(a), the autoencoder is used and information content in hidden neurons is increased as much as possible. Then, using connection weights obtained in the information acquisition phase, learning is performed in supervised ways, as in Figure 2(b).

2.2.1. Information Acquisition Phase. The computational procedures for the information acquisition phase are here explained. The output from the output neuron in the autoencoder in Figure 2(a) is computed by

$$o_k^s = f\left(\sum_{j=1}^{M} W_{kj} v_j^s\right), \tag{12}$$

where W_{kj} denote connection weights to output neurons. Thus, the error is computed by

$$E = \frac{1}{2S}\sum_{s=1}^{S}\sum_{k=1}^{L}\left(x_k^s - o_k^s\right)^2. \tag{13}$$

To increase information, the entropy should be decreased. In the information acquisition phase, the autoencoder is used. Thus, the following equation should be decreased:

$$J = \frac{1}{2S}\sum_{s=1}^{S}\sum_{k=1}^{L}\left(x_k^s - o_k^s\right)^2 - \gamma\sum_{j=1}^{M} p\left(j\right)\log p\left(j\right), \tag{14}$$

where γ is a parameter to control the effect of the entropy term.

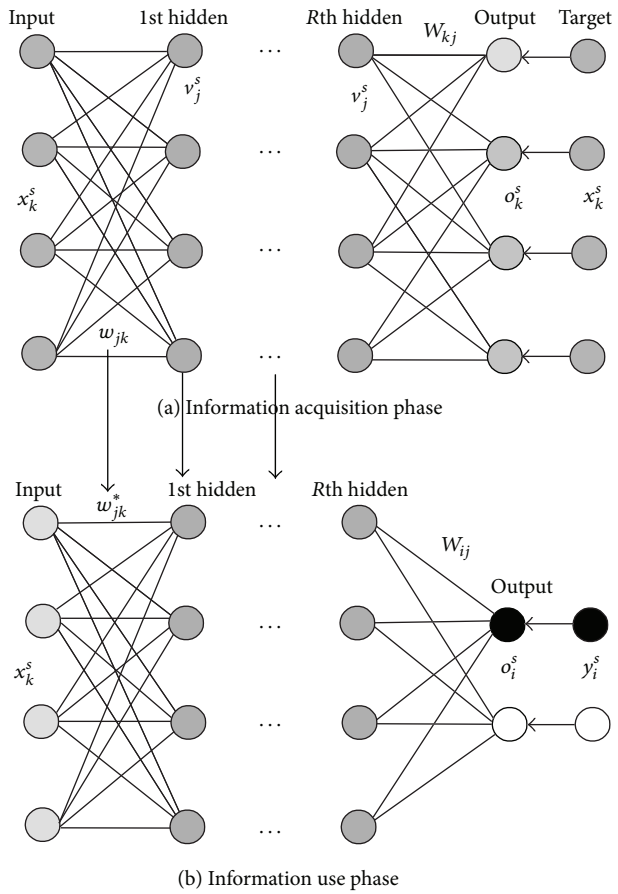


Figure 2: Network architecture for supervised learning with an information acquisition (a) and use phase (b).

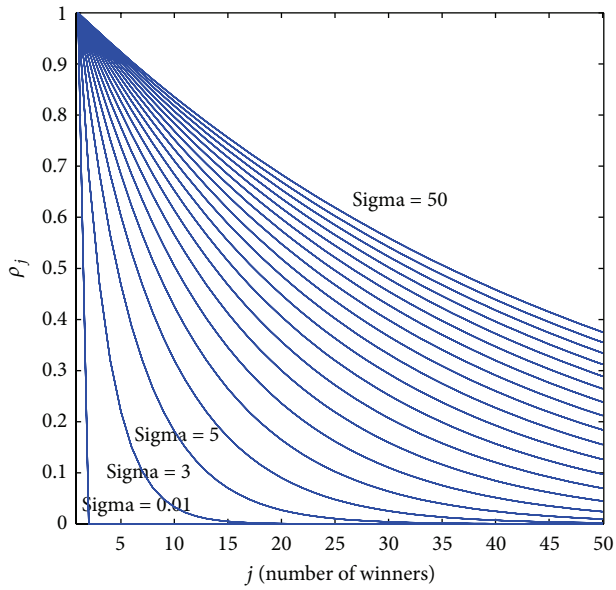
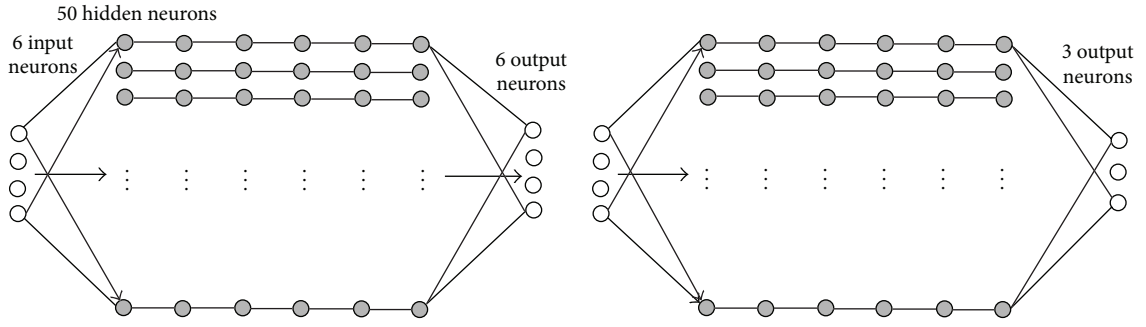


Figure 3: Targets ρ_j when the spread parameter σ increases from 0.01 to 50.



(a) Information acquisition phase (b) Information use phase

Figure 4: Network architecture with information acquisition phase (a) and use phase (b) for the Iris problem.

2.2.2. Simplified Information Acquisition Phase. The equation to be minimized is

$$J = \frac{1}{2S}\sum_{s=1}^{S}\sum_{k=1}^{L}\left(x_k^s - o_k^s\right)^2 + \gamma\sum_{j=1}^{M}\left[\rho_j \log\frac{\rho_j}{v_j} + \left(1-\rho_j\right)\log\frac{1-\rho_j}{1-v_j}\right], \tag{15}$$

where γ is a parameter to control the effect of the KL divergence. By differentiating the equation, we have

$$\frac{\partial J}{\partial w_{jk}} = \frac{1}{S}\sum_{s=1}^{S}\delta_j^s x_k^s, \tag{16}$$

where

$$\delta_j^s = \left[\sum_{k=1}^{L} W_{kj}\delta_k^s + \gamma\left(-\frac{\rho_j}{v_j} + \frac{1-\rho_j}{1-v_j}\right)\right] f'\left(u_j\right), \tag{17}$$

where δ_k^s denote the error signals from the output layer.

2.2.3. Information Use Phase. In the information use phase, the connection weights obtained in the information acquisition phase are used initially. Let w_{jk}^* denote initial connection weights provided by the information acquisition phase. Then, the output from the hidden neuron is computed by

$$v_j^s = f\left(\sum_{k=1}^{L} w_{jk}^* x_k^s\right). \tag{18}$$

In the output layer, the sofmax output is used and computed by

$$o_i^s = \frac{\exp\left(\sum_{j=1}^{M} W_{ji} v_j^s\right)}{\sum_{m=1}^{N}\exp\left(\sum_{j=1}^{M} W_{jm} v_j^s\right)}, \tag{19}$$

where W_{ji} are connection weights from the hidden neurons to the output ones. The error is computed by

$$E = -\sum_{s=1}^{S}\sum_{i=1}^{N} y_i^s \log o_i^s, \tag{20}$$

where y is the target and N is the number of output neurons. This error function can be differentiated with respect to connection weights in the competitive and output layer. The update formula for the first competitive layer is shown here:

$$\frac{\partial J}{\partial w_{jk}} = \frac{1}{S}\sum_{s=1}^{S}\delta_j^s x_k^s, \tag{21}$$

where

$$\delta_j^s = \sum_{i=1}^{N} W_{ij}\delta_i^s, \tag{22}$$

where δ is the error signal sent from the output layers and η is a learning parameter.

3. Results and Discussion

3.1. Iris Data

3.1.1. Experiment Outline. First, the method was applied to the well-known Iris problem data set [30]. Because this data set is well-known, it is easy to evaluate intuitively the good performance of the present method. In the experiment, due attention was given to the differences between our method and conventional multilayered networks. Then, the number of hidden layers was fixed to 50 for any hidden layers. In this way, it was possible to obtain better generalization errors by changing the number of hidden neurons in each layer. However, our objective in the present experiment was to show that clear differences could be seen between the conventional method and our method given the same situation. Figure 4 shows a network architecture where the number of input, hidden, and output neurons was 4, 50, and 4, respectively. In the unsupervised phase, the autoencoder was used. In the supervised phase, the ordinary training method was used.

3.1.2. Information Acquisition. First, an experiment was conducted to see whether our simplified method was effective in increasing information content:

$$I = \log M + \sum_{j=1}^{M} p\left(j\right)\log p\left(j\right). \tag{23}$$

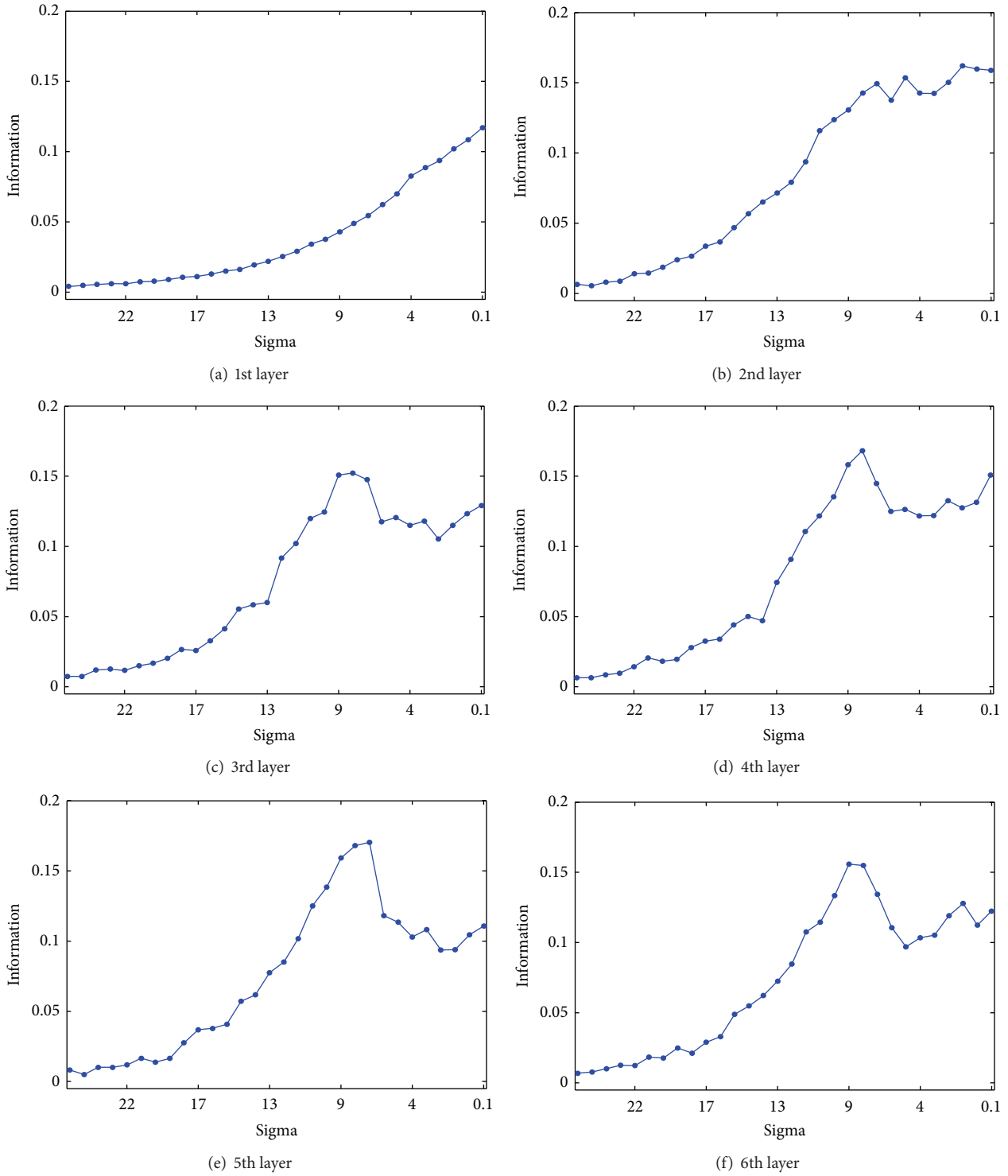


Figure 5: Information for six hidden layers as a function of the parameter σ for the Iris data.

Figure 5 shows information as a function of the parameter σ. Information should increase when the parameter σ decreases because a smaller number of neurons tend to fire. On the other hand, when the parameter σ increases, the firing rates of all hidden neurons become larger. For the first layer in Figure 5(a), the information increased gradually when the parameter σ increased. For the second layer in Figure 5(b), the information increased gradually, but in the end some fluctuations could be seen when the parameter σ became larger. For the third to the sixth layers in Figures 5(c)–5(f),

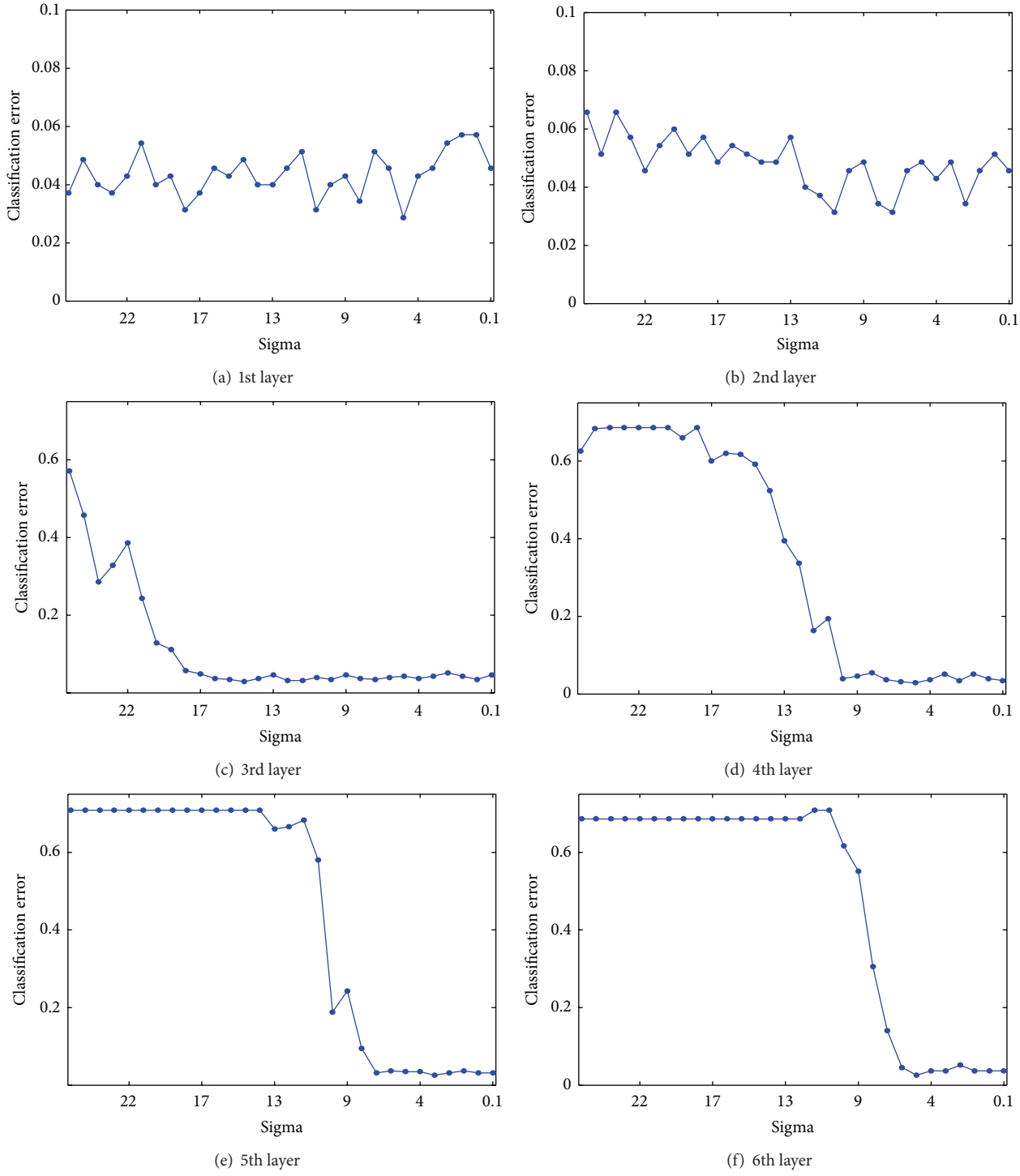


Figure 6: Generalization errors for six hidden layers as a function of the parameter σ for the Iris data.

this tendency became more apparent. In sum, information could be increased, though there were some fluctuations for the higher layers.

3.1.3. Information Use. As mentioned above, the information could be increased when the parameter σ decreased. The experimental results presented here show whether generalization errors are related to the information content. Figure 6 shows generalization errors as a function of the parameter σ. For the first layer in Figure 6(a), the generalization errors fluctuated when the parameter σ increased. For the second layer in Figure 6(b), the generalization errors decreased slightly when the parameter σ became larger. For the third layer in Figure 6(c), generalization errors gradually decreased when the parameter σ increased. For the fourth, fifth, and sixth layers in Figures 6(d)–6(f), generalization errors were

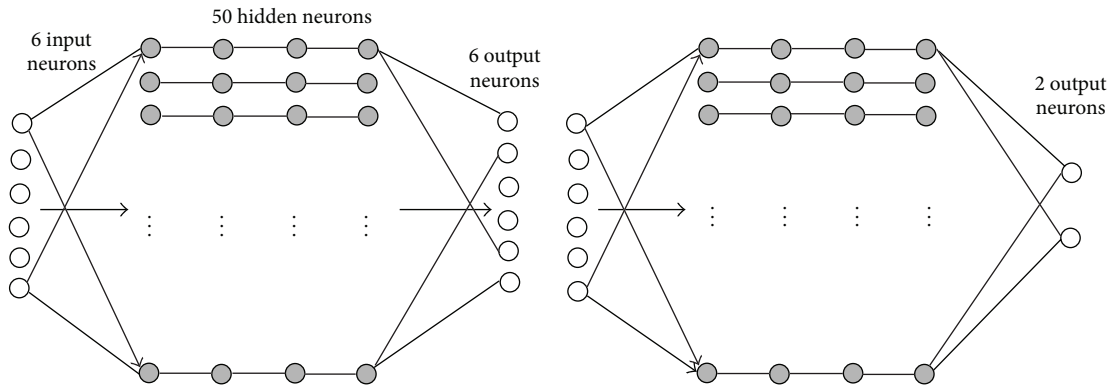


Figure 7: Network architecture with information acquisition phase (a) and use phase (b) for the bankruptcy problem.

unchanged for the smaller values of the parameter; when the parameter further decreased, the generalization decreased rapidly. Thus, for the higher layers, generalization errors decreased in proportion to information increase.

3.1.4. Summary of Generalization Errors. Table 1 shows the summary of experimental results on generalization performance. The generalization errors by the information-theoretic method were lower than those by the standard method for all layers. In particular, the lowest errors (0.026) were obtained when the number of layers was five and six. In addition, maximum generalization errors by the information-theoretic method were smaller than those by the conventional method except in the second and the third layer, which had the same maximum values. Thus, generalization errors could be decreased even for the higher layers by increasing information.

3.2. Bankruptcy Data. The second type of data was the bankruptcy data (http://pub.nikkan.co.jp/tahenryo/tahenryou.html), where companies are labeled as bankrupt or sound. The number of input, hidden, and output neurons was 6, 50, and 2, respectively. The number of input patterns was 130. 70 of them were used as training patterns, and the remaining ones were equally divided into validation and testing data. Figure 7 shows a network architecture where 6 input, 50 hidden, and 2 output neurons were used. The number of hidden layers was increased up to four because no improvement could be seen beyond the fourth layer.

3.2.1. Information Acquisition. Figure 8 shows the information as a function of the parameter σ. For the first layer in Figure 8(a), information continued to increase when the parameter σ decreased from 25 to 0.1. For the second layer, information also increased constantly, though with some small fluctuations for smaller values of the parameter σ. For the third layer, information increased constantly to a certain point, decreased, and again increased in Figure 8(c). For the fourth layer, information showed the same tendency in that it increased to a certain point and fluctuated, as in Figure 8(d). In sum, though our simplified method could increase information, there were some fluctuations.

3.2.2. Information Use. Figure 9 shows the generalization errors as a function of the parameter σ. For the first layer, the generalization errors remained to be a constant almost independently of the parameter σ. For the second layer, the generalization errors decreased gradually when the parameter σ increased. For the third layer, generalization errors were unchanged when the parameter σ was large and then decreased rapidly. For the fourth layer, the generalization errors were also unchanged for a wide range of the parameter and then decreased rapidly. Thus, the generalization errors tended to decrease in proportion to information increase except in the first layer.

3.2.3. Summary of Generalization Errors. Table 2 shows the summary of the experimental results on generalization performance. All generalization errors were well smaller than those by the conventional method. The lowest generalization error of 0.207 (with two hidden layers) was much smaller than the 0.233 (with two hidden layers) obtained by the conventional method. In addition, by the conventional method, the generalization errors increased from 0.270 in the first layer to 0.320 in the fourth layer. However, by the information-theoretic method, the generalization errors did not increase as much. For example, from the first to the fourth layer, the generalization errors decreased from 0.247 to 0.233. Thus, as shown, the present simplified method was able to decrease generalization errors by increasing information in multilayered neural networks.

3.3. Rebel Participation Data

3.3.1. Experiment Outline. Finally, the method was applied to the rebel participation data set [31]. The total number of patterns was 1340, split as follows: the number of input patterns was 500, the number of validation patterns was 400, and the remaining 440 patterns were for testing data. Figure 10 shows a network architecture of 19 input, 100

Table 1: Summary of experimental results by the simplified information maximization for the first to the sixth competitive layer for the Iris data. The simplified information maximization and standard multilayered neural networks are represented by "SIM" and "STD," respectively.

Methods	Layers	Average	Std. dev.	Max	Min
SIM	1	0.029	0.019	0.057	0.000
	2	0.031	0.025	0.086	0.000
	3	0.029	0.030	0.086	0.000
	4	0.029	0.019	0.057	0.000
	5	0.026	0.025	0.057	0.000
	6	0.026	0.021	0.057	0.000
STD	1	0.049	0.036	0.114	0.000
	2	0.029	0.027	0.086	0.000
	3	0.043	0.028	0.086	0.000
	4	0.046	0.039	0.114	0.000
	5	0.034	0.030	0.086	0.000
	6	0.049	0.027	0.114	0.029

Table 2: Summary of experimental results by the simplified information maximization for the first to the fourth hidden layers and the method without unsupervised information acquisition phase for the bankruptcy data. The simplified information maximization and standard multilayered neural networks are represented by "SIM" and "STD," respectively.

Methods	Layers	Average	Std. dev.	Max	Min
SIM	1	0.247	0.107	0.400	0.100
	2	0.207	0.068	0.300	0.100
	3	0.217	0.065	0.300	0.100
	4	0.233	0.093	0.333	0.100
STD	1	0.270	0.090	0.400	0.167
	2	0.233	0.065	0.333	0.167
	3	0.260	0.080	0.367	0.133
	4	0.320	0.115	0.533	0.133

hidden, and 2 output neurons. The number of hidden layers increased from one to eight, because no further improvement could be seen.

3.3.2. Information Acquisition. Then, the experimental results showed to what extent information content could be increased for the higher hidden layers. Figure 11 shows information as a function of the parameter σ. For the first layer in Figure 11(a), information gradually increased and stabilized. For the second layer, information gradually decreased and then increased rapidly in Figure 11(b). For the third layer in Figure 11(c) to the eighth layer, information first increased then deceased rapidly. Finally, information began to increase in the end. Around the lower points of information, it seems that some kind of phase transition occurred. This means that the connection weights changed drastically around these points. From these results, it can be said that information could be increased by the present simplified method and that information increase may go through several phase transitions in the process of learning.

3.3.3. Information Use. In the information use phase, connection weights obtained in the information acquisition phase were used to train multilayered neural networks. Figure 12(a) shows generalization errors as a function of the parameter σ. Though information increased in Figure 11(a), generalization errors were not correlated with this information increase. Figure 12(b) shows generalization errors as a function of the parameter σ for the second layer. As shown in Figure 11(b), information decreased in the beginning and generalization errors decreased correspondingly in Figure 12(b). Then, when information increased in Figure 11(b), generalization errors increased in Figure 12(b). Figure 12(c) shows generalization errors as a function of the parameter σ for the third layer. Generalization errors deceased gradually and reached their stable points with some fluctuations. Thus, generalization errors were not affected by the abrupt changes in information in Figure 11(c). For the fourth layer in Figure 12(d) to the eighth layer, close relations could be seen between generalization errors and information. As can be seen in Figures 11(d)–11(h) and 12(d)–12(h), generalization errors only decreased to the points where drastic changes in information occurred. From these points on, generalization errors increased, though only slightly.

3.3.4. Summary of Generalization Errors. Here we present the summary of the results on generalization performance. As shown in Table 3, all types of generalization errors including the average, minimum, and maximum values by the information-theoretic method were well lower than those

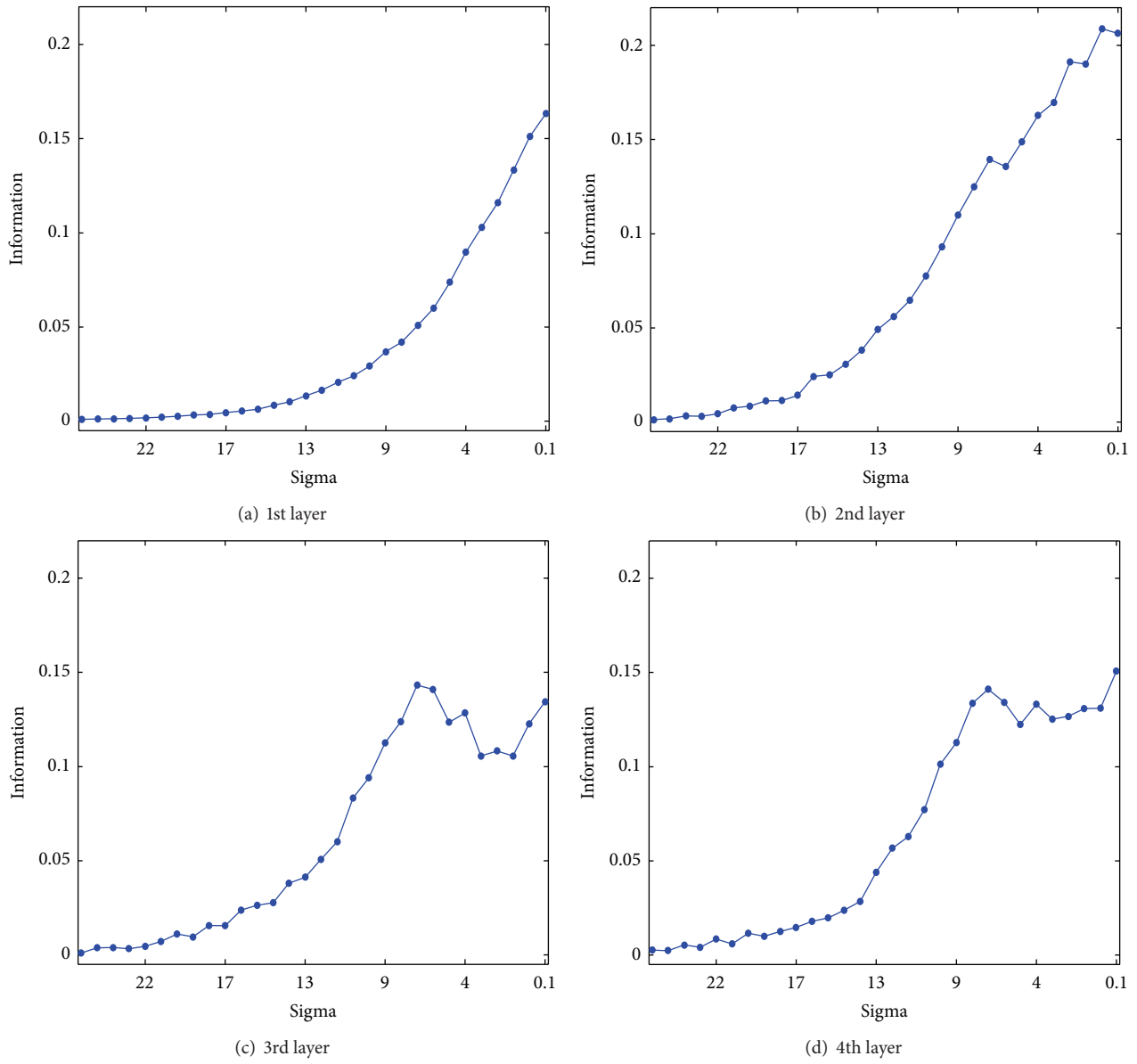


FIGURE 8: Information for four competitive layers as a function of the parameter σ for the bankruptcy data.

by the conventional method. The averaged generalization errors took the lowest value of 0.172 with six layers. However, even with eight layers, generalization errors had the second best value of 0.173. In addition, the standard deviation of the generalization errors became the smallest when the number of layers was seven and eight. The minimum generalization error of 0.134 was obtained with six layers, and the maximum error was at its lowest 0.193 with eight layers by the conventional method. On the other hand, generalization errors increased gradually from 0.189 (one layer) to 0.225 by the conventional method (eight layers). The corresponding maximum and minimum values increased gradually. This means that, by the conventional method, the effectiveness of multiple layers was not observed.

3.4. Discussion

3.4.1. Validity of Methods and Results. The results presented in this study demonstrate that the simplified information acquisition method was effective in increasing information content, accompanied by improved generalization performance. The method is simpler than those which directly differentiate the information or entropy function [27–29]. Thus, it can be applied to many problems, in particular, to large-sized data.

One of the main findings of the analysis is that generalization performance was not degraded when the number of competitive layers was larger. In all three experimental results in Tables 1, 2, and 3, generalization errors tended to

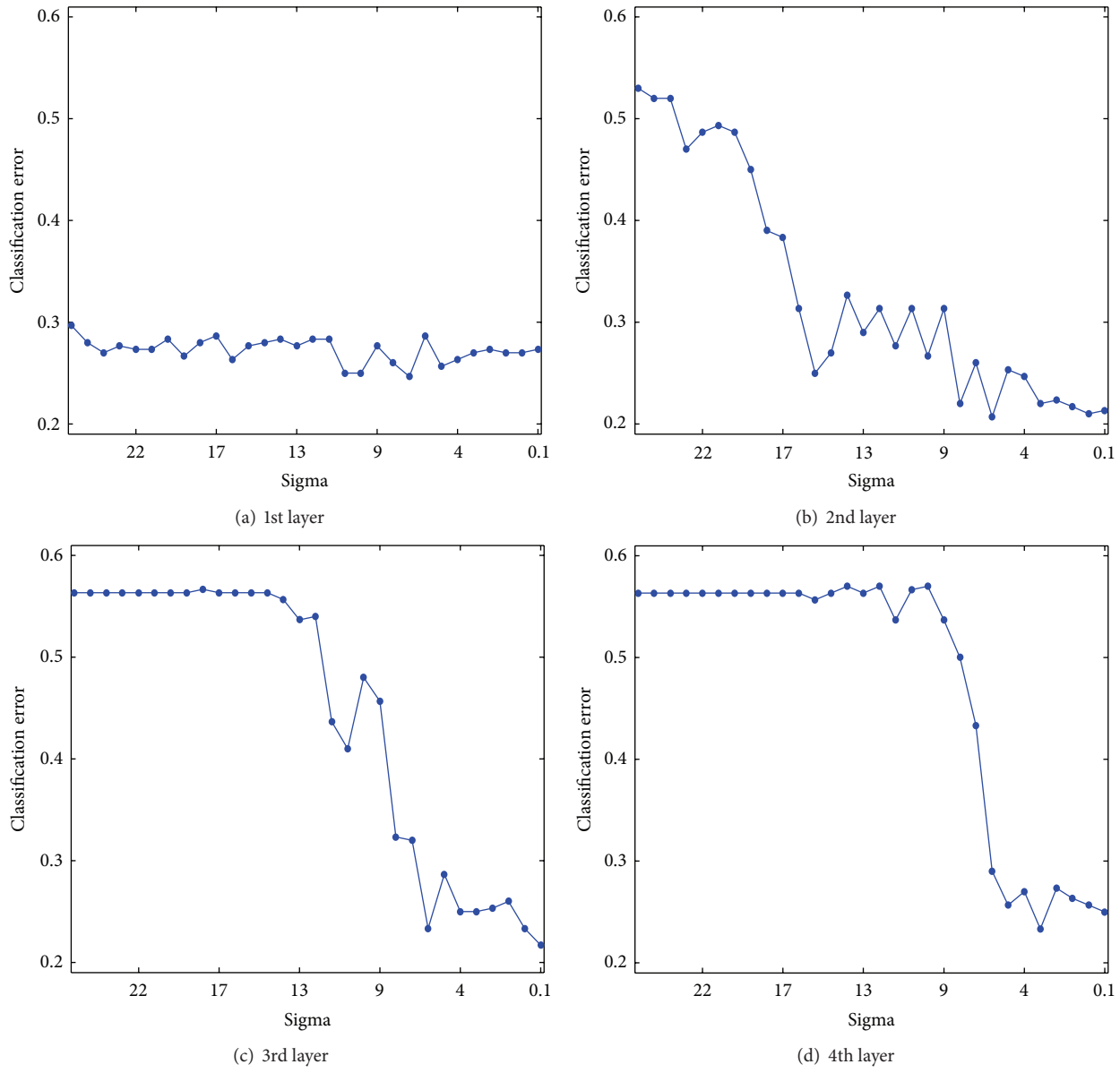


Figure 9: Generalization errors for four hidden layers as a function of the parameter σ for the bankruptcy data.

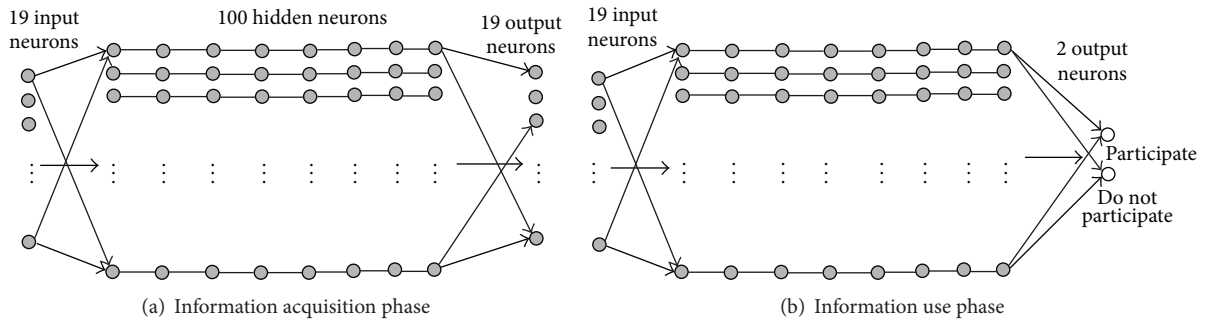


Figure 10: Network architecture with information acquisition phase (a) and use phase (b) with 19 input, 100 hidden, and 2 output neurons for the rebel participation problem.

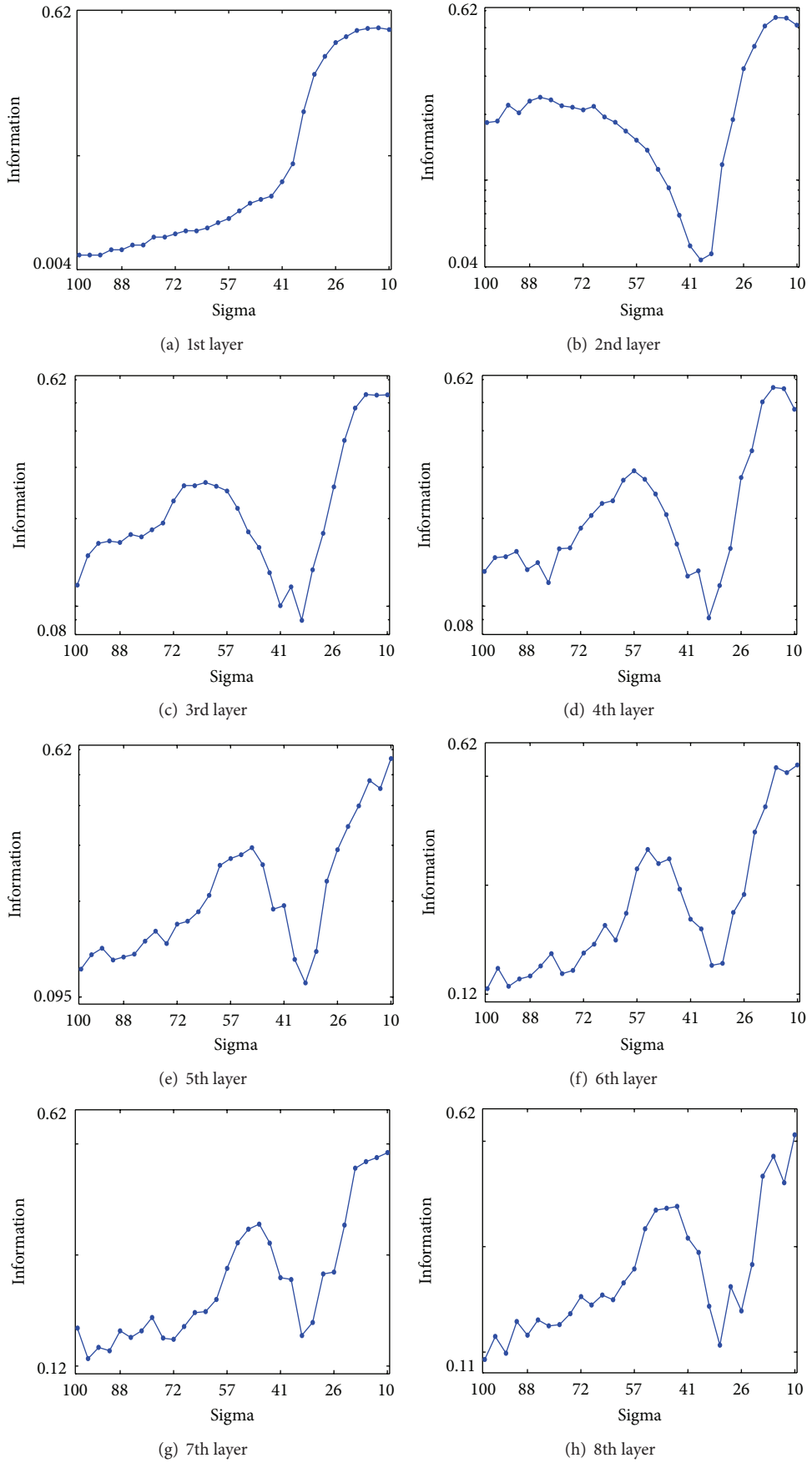


Figure 11: Information for eight hidden layers as a function of the parameter σ for the rebel participation data.

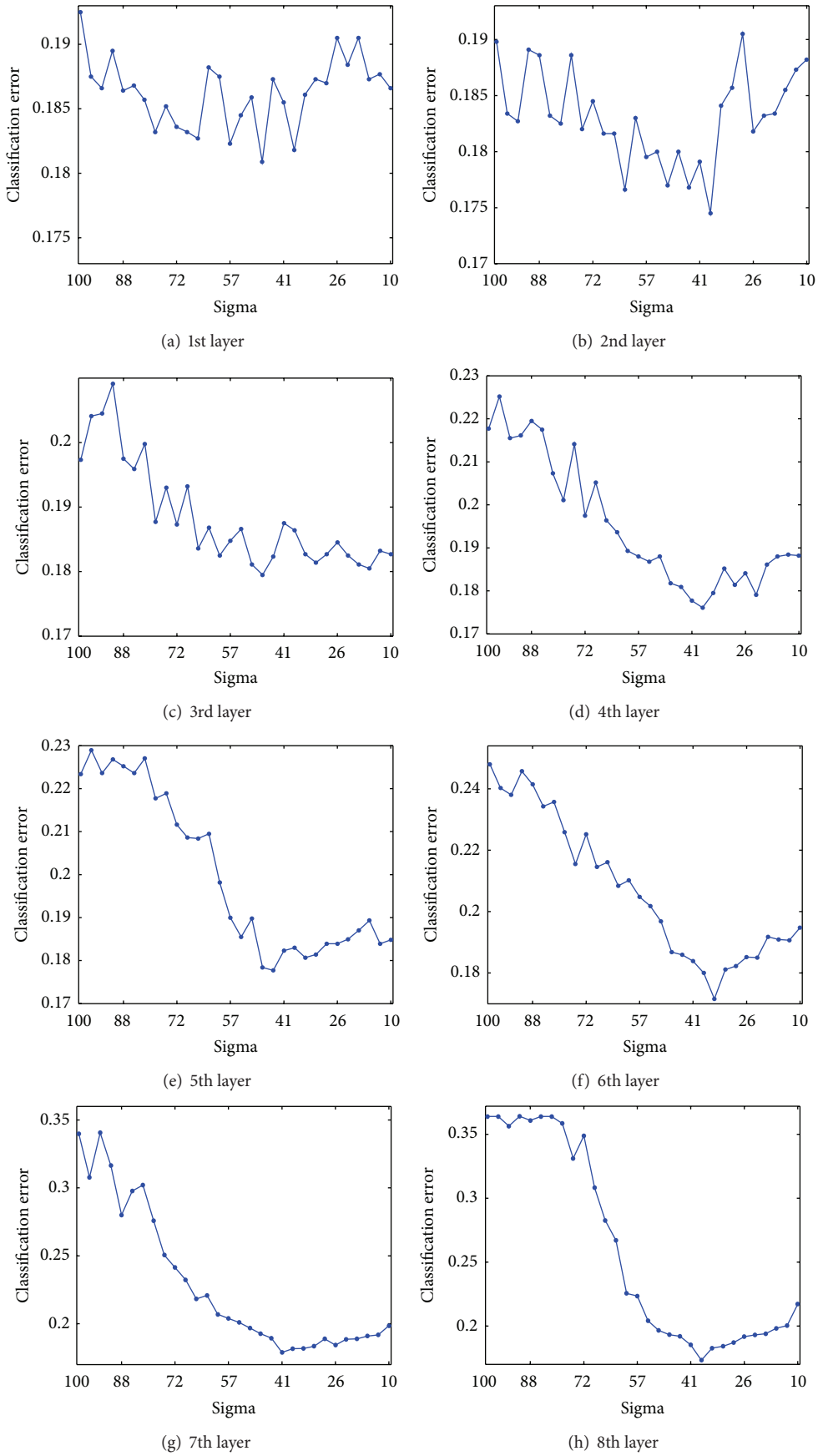


Figure 12: Generalization errors for eight competitive layers as a function of the parameter σ for the rebel participation data.

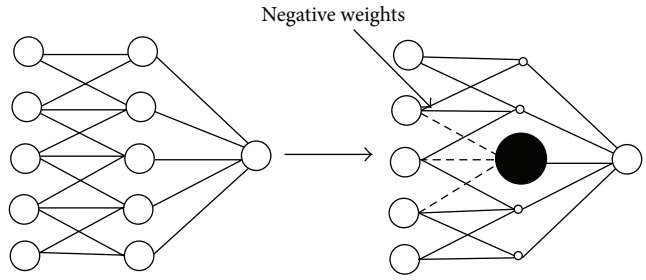


Figure 13: From an initial state to a maximum information state.

Table 3: Summary of the experimental results for the rebel participation data by information maximization and the method without the unsupervised information acquisition phase for the first to eighth competitive layers. The simplified information maximization and standard multilayered neural networks are represented by "SIM" and "STD," respectively.

Method	Layer	Average	Std. dev.	Max	Min
SIM	1	0.181	0.017	0.207	0.155
	2	0.175	0.016	0.200	0.155
	3	0.180	0.016	0.205	0.150
	4	0.176	0.016	0.196	0.141
	5	0.178	0.019	0.209	0.148
	6	0.172	0.020	0.200	0.134
	7	0.179	0.012	0.200	0.155
	8	0.173	0.012	0.193	0.152
STD	1	0.189	0.016	0.218	0.166
	2	0.184	0.020	0.209	0.157
	3	0.187	0.017	0.211	0.164
	4	0.190	0.017	0.216	0.166
	5	0.214	0.025	0.264	0.180
	6	0.216	0.023	0.252	0.184
	7	0.221	0.016	0.255	0.202
	8	0.225	0.021	0.271	0.198

increase for the higher layers by the conventional method. On the other hand, by the present method, generalization errors decreased almost constantly for the higher layers. This improved generalization performance can be explained in two ways, namely, the number of hidden neurons and the separation of the information acquisition and use phases.

First, close relations should be pointed out between information increase and the number of hidden neurons. Figures 5, 6, 8, and 9 show that information increase was accompanied by improved generalization except the first layer. In the information maximization method, when the information of hidden neurons increased, the number of firing neurons decreased. Finally, when the information was completely maximized, just one neuron fired, while all others ceased to do so, as shown in Figure 13. Information increases corresponded to decreases in the number of activated neurons. The results show that information increase is related to improved generalization. This means that when the number of activated hidden neurons decreases, improved generalization can be obtained. The present method can be used to change the number of neurons flexibly according to the quantity of information.

Second, the separation of the information acquisition and use phases was shown to be effective in improving generalization. As information maximization methods were originally developed for neural networks [27–29], they have been easily interpretable; however, these methods have not necessarily led to improved prediction performance. In the information maximization methods, hidden neurons are inhibited by generating strongly negative connection weights, as in Figure 13. The magnitude of negative connection weights becomes larger and, thus, strongly negative connection weights have unfavorable effects on generalization performance. The separation of those two phases weakens the bad effects of strongly negative connection weights. Due to this separation, the process of information maximization can concentrate on its own process of information maximization.

From this consideration, it can be concluded that improved generalization, in particular, for higher layers, is due to the flexible control of the number of hidden neurons

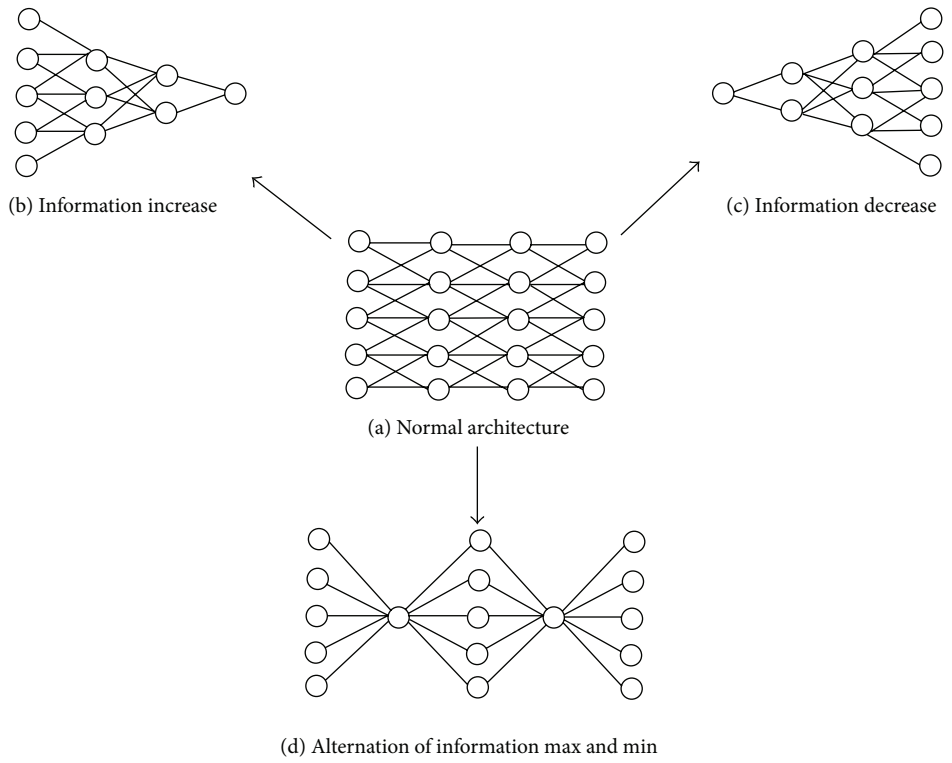


Figure 14: A variety of architectures by the information-theoretic method.

and the separation of information increasing processes and error minimization.

3.4.2. Problems of the Method. There are at least three problems with the present method, namely, winner determination, how to determine the quantity of information, and drastic changes in information. First, there is the problem of determining winners. As explained, the winners can be determined by using the variance of outputs from hidden neurons. When the variance becomes larger, the degree of winning becomes larger. The variance was adopted heuristically; thus, there is the possibility of some other options. For example, the simplest ways to choose winners is based on the magnitude of hidden outputs. When the magnitude of hidden neuron outputs becomes larger, the degree of winning becomes larger as well. This hypothesis seems to be natural for determining winners. However, even if the magnitude of hidden neuron output is large, the neuron may be not so important. For example, if all the output values of a neuron are the same, it is of no use. Thus, a method should be developed to determine the degree of winning neurons more naturally. Further comparison of those possible candidates for determining winners is needed to explain the necessity of the winners.

In addition, no explicit criteria exist to define the quantity of information content or the number of hidden neurons. As is explained in the experimental results, the information is forced to increase as much as possible and examine relations between information and generalization. It would be very convenient to have some criteria to stop this information increase. This is related to the determination of the appropriate number of hidden neurons.

Third, drastic changes should be explained in information increases, in particular, for the rebel data set in Figure 11. The information increased gradually as the parameter σ increased, suddenly decreased, and finally increased again. As mentioned in the experimental results, it is possible that a kind of phase transition occurred at this point. It is thus necessary to carefully examine the mechanisms of this phase transition and provide an explanation for why it occurred. In particular, the effect of this drastic change on generalization performance should be carefully examined.

3.4.3. Possibility of the Method. One of the main possibilities of the present method is that a number of different types of network architectures can be created simply by changing the information content. As mentioned, information maximization corresponds to the firing of a single neuron in the end. This means that the number of activated neurons can be changed by modifying the information content. Figure 14 shows this situation of flexible control of network architecture. Figure 14 shows a normal architecture in which all layers are equally treated in terms of information content. Figure 14(b) shows the case where information increases gradually. In this case, the number of activated neurons decreases gradually. On the other hand, Figure 14(c) shows the inverse case, where information decreases gradually and the number of activated neurons increases gradually. A number of different cases can be imagined with different quantities of information. Figure 14(d) shows only an example

where maximum and minimum information are alternatively applied. The present method has the possibility of producing different levels of network architecture flexibility.

The property is related to the production of appropriate network architectures for different problems. As is well-known, the learning of multilayered networks can be fasciculated by unsupervised learning [16–19]. However, no methods exist which can determine the number of neurons in intermediate layers; the number of neurons must be chosen by trial and error. The present method can flexibly control the information content or actively control information content in the hidden layers. The flexible control of information can be used to train multilayered networks more easily, because information content in each hidden layer can be controlled.

4. Conclusion

In this paper, a new type of information-theoretic method to improve generalization performance was proposed. In the method, the complex procedures of information maximization were replaced by simpler ones. The method directly deals with outputs from hidden neurons. In the process of information maximization, a small number of neurons actually become activated. This process is realized by activating neurons in accordance with the magnitude of the neurons' variance. When the neurons become more important (larger variance), they become more activated and larger.

In addition, the information acquisition and use phases are separated. In the information acquisition phase, information content in hidden neurons is increased by producing a small number of active hidden outputs. On the other hand, in the information use phase, the information obtained in the information acquisition phase is used to train supervised learning. As is well-known, information maximization has been sometimes contradictory with error minimization for supervised learning. The separation between the two phases showed that it was possible to compromise between error minimization and information maximization. This is because, in each phase, neural networks can focus solely on either the processing of information maximization or error minimization.

The method was applied to the Iris data set, bankruptcy data set, and rebel data set. Experimental results showed that the information increase and improved prediction performance were possible through the present method. In particular, for higher layers, information increase was directly related to generalization performance, though some abrupt changes in information occurred in learning.

Though information-theoretic methods have been used extensively in neural networks, in particular to examine how neural networks acquire information content on input patterns, their learning rules have been complicated for actual applications. The proposed method is simple enough that it can be applied to many problems, in particular, to large-sized data. This opens up new possibilities for information-theoretic methods.

Competing Interests

The author declares that there are no competing interests regarding the publication of this paper.

References

[1] R. Linsker, "Self-organization in a perceptual network," *Computer*, vol. 21, no. 3, pp. 105–117, 1988.

[2] R. Linsker, "How to generate ordered maps by maximizing the mutual information between input and output signals," *Neural Computation*, vol. 1, no. 3, pp. 402–411, 1989.

[3] R. Linsker, "Local synaptic learning rules suffice to maximize mutual information in a linear network," *Neural Computation*, vol. 4, no. 5, pp. 691–702, 1992.

[4] R. Linsker, "Improved local learning rule for information maximization and related applications," *Neural Networks*, vol. 18, no. 3, pp. 261–265, 2005.

[5] H. B. Barlow, "Unsupervised learning," *Neural Computation*, vol. 1, no. 3, pp. 295–311, 1989.

[6] H. B. Barlow, T. P. Kaushal, and G. J. Mitchison, "Finding minimum entropy codes," *Neural Computation*, vol. 1, no. 3, pp. 412–423, 1989.

[7] Z. Zenadic, "Information discriminant analysis: feature extraction with an information-theoretic objective," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 8, pp. 1394–1407, 2007.

[8] K. Torkkola, "Feature extraction by non-parametric mutual information maximization," *Journal of Machine Learning Research*, vol. 3, no. 7-8, pp. 1415–1438, 2003.

[9] J. M. Leiva-Murillo and A. Artés-Rodríguez, "Maximization of mutual information for supervised linear feature extraction," *IEEE Transactions on Neural Networks*, vol. 18, no. 5, pp. 1433–1441, 2007.

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

[11] R. Kamimura and F. Yoshida, "Teacher-directed learning: information-theoretic competitive learning in supervised multi-layered networks," *Connection Science*, vol. 15, no. 2-3, pp. 117–140, 2003.

[12] R. Kamimura, "Information theoretic competitive learning in self-adaptive multi-layered networks," *Connection Science*, vol. 13, no. 4, pp. 323–347, 2003.

[13] 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.

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

[15] R. Kamimura, "Information-theoretic competitive learning with inverse Euclidean distance output units," *Neural Processing Letters*, vol. 18, no. 3, pp. 163–184, 2003.

[16] G. E. Hinton, "Learning multiple layers of representation," *Trends in Cognitive Sciences*, vol. 11, no. 10, pp. 428–434, 2007.

[17] Y. Bengio, "Learning deep architectures for AI," *Foundations and Trends® in Machine Learning*, vol. 2, no. 1, pp. 1–127, 2009.

[18] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, 2006.

[19] G. E. Hinton, S. Osindero, and Y.-W. Teh, "A fast learning algorithm for deep belief nets," *Neural Computation*, vol. 18, no. 7, pp. 1527–1554, 2006.

[20] G. Hinton, "A practical guide to training restricted boltzmann machines," *Momentum*, vol. 9, no. 1, p. 926, 2010.

[21] C. Poultney, S. Chopra, Y. L. Cun et al., "Efficient learning of sparse representations with an energy-based model," in *Advances in Neural Information Processing Systems*, pp. 1137–1144, MIT Press, 2006.

[22] X. Glorot, A. Bordes, and Y. Bengio, "Deep sparse rectifier networks," in *Proceedings of the 14th International Conference on Artificial Intelligence and Statistics. JMLR W&CP Volume*, vol. 15, pp. 315–323, Fort Lauderdale, Fla, USA, April 2011.

[23] M. A. Ranzato, Y.-L. Boureau, and Y. L. Cun, "Sparse feature learning for deep belief networks," in *Proceedings of the Advances in Neural Information Processing Systems (NIPS '08)*, pp. 1185–1192, 2008.

[24] H. Lee, C. Ekanadham, and A. Y. Ng, "Sparse deep belief net model for visual area v2," in *Advances in Neural Information Processing Systems*, pp. 873–880, MIT Press, 2008.

[25] P. Lennie, "The cost of cortical computation," *Current Biology*, vol. 13, no. 6, pp. 493–497, 2003.

[26] E. P. Simoncelli, "4.7 Statistical modeling of photographic images," in *Handbook of Video and Image Processing*, Academic Press, 2005.

[27] 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.

[28] R. Kamimura and S. Nakanishi, "Hidden information maximization for feature detection and rule discovery," *Network: Computation in Neural Systems*, vol. 6, no. 4, pp. 577–602, 1995.

[29] R. Kamimura and T. Kamimura, "Structural information and linguistic rule extraction," in *Proceedings of the International Conference on Neural Information Processing (ICONIP '00)*, pp. 720–726, Taejon, Republic of Korea, November 2000.

[30] M. Lichman, *UCI Machine Learning Repository*, University of California, Irvine, Calif, USA, School of Information and Computer Sciences, 2013, http://archive.ics.uci.edu/ml/.

[31] A. Oyefusi, "Oil and the probability of rebel participation among youths in the Niger Delta of Nigeria," *Journal of Peace Research*, vol. 45, no. 4, pp. 539–555, 2008.