Controlled Information Maximization for SOM Knowledge Induced Learning

Ryotaro Kamimura
IT Education Cente and Graduate School of Science and Technology, Tokai Univerisity
1117 Kitakaname, Hiratsuka, Kanagawa 259-1292, Japan
ryo@keyaki.cc.u-tokai.ac.jp

Abstract— *The present paper aims to control information content in multi-layered neural networks to improve generalization performance. Following Linsker's maximum information principle, information should be increased as much as possible in multi-layered neural networks. However, it is needed to control information increase appropriately to improve the performance. Thus, the present paper proposes a method to control information content so as to increase generalization performance. Experimental results on an artificial data and the spam data set showed that improved generalization performance could be obtained by appropriately controlling information content. In particular, better performance could be observed for complex problems. Compared with the results by the conventional methods such as the support vector machine, better performance could be obtained when the information was larger. Thus,, the present results certainly show a possibility of SOM knowledge in training multi-layered networks.*

Keywords: Maximum information, controlling information, multi-layered networks, SOM, deep learning

1. Introduction

1.1 Maximum Information

Information-theoretic methods have received due attention since Linsker [1], [2], [3], [4] tried to describe information processing in living systems by the maximum information principle. In this principle, information content in multi-layered neural networks should be increased as mush as possible for each processing stage. Linsker demonstrated the generation of feature detecting neurons by maximizing information content for simple and linear neural networks. However, because difficulty have existed in training multi-layered neural networks, few results on this performance of fully multi-layered neural networks have been reported.

Recently, multi-layered neural networks has received much attention because several methods to facilitate the learning of multi-layered neural networks have been proposed in the deep learning [5], [6], [7], [8]. Thus, the time has come to examine the effectiveness of the maximum information principle in training multi-layered neural networks. In the deep learning, unsupervised feature detection is realized by the auto-encoder and the restricted Boltzmann machines. However, they are not necessarily good at detecting main features of input patterns, because they have not been developed as feature detectors. Thus, it is needed to use more efficient feature detecting methods for multi-layered neural networks.

1.2 SOM Knowledge

In training multi-layered neural networks, it is important to extract the main features of input patterns. In the present paper, the self-organizing maps (SOM) is used to detect the features for training multi-layered neural networks. As is well known, the SOM has been developed to extract important features and in addition to visualize those features. If it is possible to use the features detected by the SOM for training multi-layered neural networks, the training can be more facilitated, and in addition, final results can be visualized for easy interpretation.

Recently, the SOM was found to be effective in training multi-layered neural networks under the condition that information content of each hidden layer is maximized or increased as much as possible [9]. This means that Linsker's principle of maximum information preservation is effective in training multi-layered neural networks with the SOM. Meantime, it has been observed that information should not be simply increased. The information increase or maximization should be appropriately controlled to have better performance, in particular, better generalization performance. Thus, the objective of the paper is to control appropriately the process of information maximization and to explore to what extent generalization performance can be improved.

1.3 Outline

In Section 2, the SOM knowledge induced learning is introduced, which is composed of SOM and supervised multi-layered neural networks. Then, the information content is defined as decrease of uncertainty of neurons. This information is controlled by using the number of layers multiplied by the other parameter r. The parameter r is introduced to adjust the information content for given problems. In Section 3, the artificial and spam data are used to examine to what extent information can be increased and generalization performance

can be improved. Experimental results show that information can be increased and correspondingly generalization errors can be decreased by the present method.

2. Theory and Computational Methods

2.1 SOM Knowledge Induced Learning

The SOM knowledge induced learning is a method to use the knowledge by the SOM to train multi-layered neural networks. Figure 1 shows a network architecture for the learning. As shown in the figure, the learning is composed of two phases, namely, the information acquisition (a) and use (b) phase. In the information acquisition phase in Figure 1(a), each competitive layer is trained with SOM to produce weights. These weights are used to train multi-layered neural networks in Figure 1(b). In the information use phase, the ordinary back-propagation learning is applied with the early stopping criteria. The problem is whether the weights by the SOM are effective in improving generalization performance.

2.2 Information Content

The SOM knowledge is effective only with the maximum information principle. Thus, this section deals with how to increase information content. As shown in Figure 1, a network is composed of the input layer, multiple competitive layers and an output layer. Let us explain how to compute output from competitive and output neurons. Now, the sth input pattern can be represented by $\mathbf{x}^s = [x_1^s, x_2^s, \cdots, x_L^s]^T,\ s = 1, 2, \cdots, S$. Connection weights into the jth competitive neuron are denoted by $\mathbf{w}_j = [w_{1j}, w_{2j}, \cdots, w_{Lj}]^T,\ j = 1, 2, \ldots, M$. The output from an output neuron is computed by

$$v_j^s = \exp\left(-\frac{\| \mathbf{x}^s - \mathbf{w}_j \|^2}{\sigma^2}\right), \tag{1}$$

where σ denotes a spread parameter or Gaussian width. The output from the jth neuron is defined by

$$v_j = \frac{1}{S}\sum_{j=1}^{M} v_j^s. \tag{2}$$

The firing probabilities are computed by

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

The uncertainty or entropy of this neuron is

$$H = -\sum_{j=1}^{M} p(j)\log p(j). \tag{4}$$

The information content is defined by difference between maximum and observed uncertainty

$$\begin{aligned} I &= H^{max} - H \\ &= \log M + \sum_{j=1}^{M} p(j)\log p(j). \end{aligned} \tag{5}$$

2.3 Controlled Information Maximization

This information can be increased by decreasing the Gaussian width σ. The width is here defined by

$$\sigma(t) = \frac{1}{t^r}, \tag{6}$$

where t denotes the layer number. When the number of layers increases, the spread parameter σ decreases and the corresponding information tends to increase. In addition, the parameter r is needed to control the spread parameter. When the parameter r increases, the spread parameter σ decreases and correspondingly the information tends to increase.

Figure 2 shows the spread parameter σ as a function of the number of layers t when the parameter r increases from 0.1 to 2.5. As shown in the figure, the spread parameter decreases when the the number of layers increases. In addition, the spread parameter decreases when the parameter t increases. When the layer number is higher, the spread parameter gradually decreases and information increases. In this case, the number of strongly firing neurons in black gradually diminishes as shown in Figure 1. This means that the number of effective competitive neurons gradually diminishes and features can be gradually compressed into a smaller number of competitive neurons.

3. Results and Discussion

3.1 Application to Artificial Data

3.1.1 Experimental Outline

To show the effectiveness of the information maximization, an artificial data set was made, which could be divided into two classes as shown in Figure 3(a). The total number of input patterns was 2000. Among them, only 100 patterns were for training ones. Even if the number of training pattern increased, the tendency here reported was observed. The remaining 900 and 1000 patterns were for the validation and testing ones, respectively. The number of input, competitive and output neurons were 2, 25 (5 by 5) and 2, respectively.

Then, to make the problem more complex, the standard deviation of the data increased gradually. When the standard deviation increased from one in Figure 3(a) to five in Figure 3(b), the boundary between two classes became ambiguous and the problem of classification became more difficult.

3.1.2 Weights by SOM

The SOM trys to imitate input patterns as much as possible. This means that connection weights tend to be expanded to include all input patterns. Figures 4(a) and (b) show connection weights in blue and data in green by the self-organizing maps. In Figure 4(a) and (b), weights in blue were expanded to cover all data points in green. This means that the SOM tried to acquire information over connection weights on input patterns as much as possible. The problem is whether these weights are effective in training multi-layered neural networks.

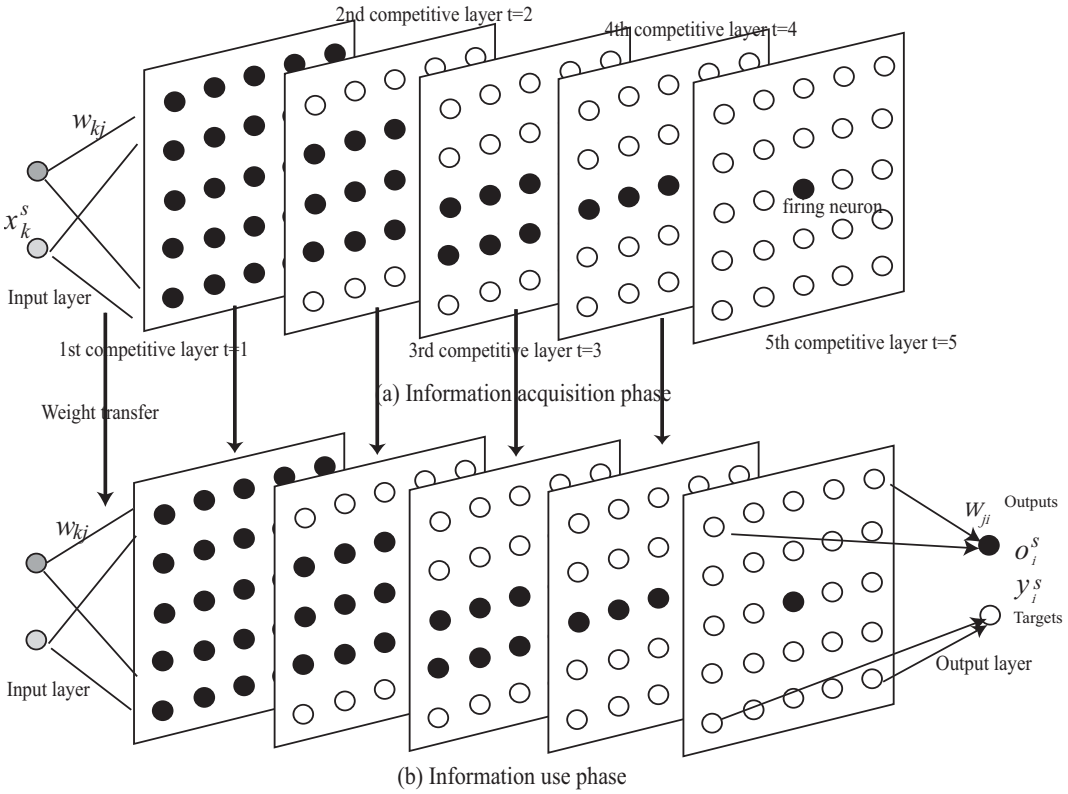


Fig. 1: Network architecture with two components of SOM knowledge induced learning where black neurons fire strongly.

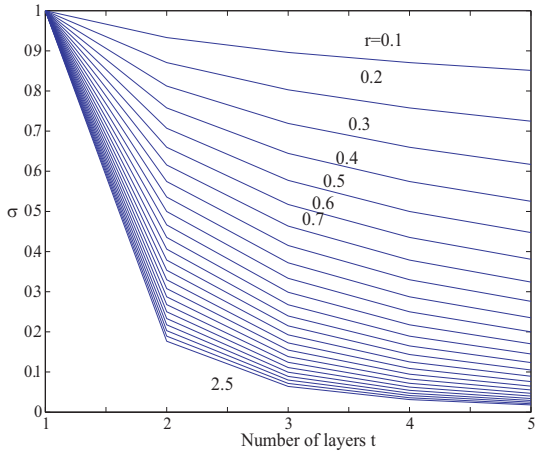


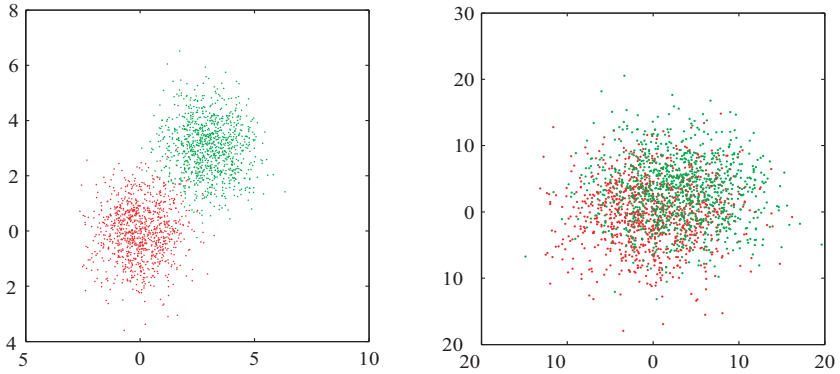
Fig. 2: Spread parameter σ as a function of the number of layers t for different valued of the parameter r.

3.1.3 Results with Information Maximization

Figure 5 shows information and generalization errors when the parameter r increased from 0.5 to 2.5. As shown in Figure 5(a), when the standard deviation was one, information increased when the parameter increased. However, the generalization errors did not decrease and in the end, they increased rapidly. Figure 5(b) shows the results when the standard deviation was three. As shown in Figure 5(b1), information increased gradually. Then, generalization errors decreased gradually as shown in Figure 5(b2). Figure 5(c) shows the results when the standard deviation was five. As shown in Figure 5(c1), information increased and the generalization errors decreased, though some fluctuations could be seen in Figure 5(c2).

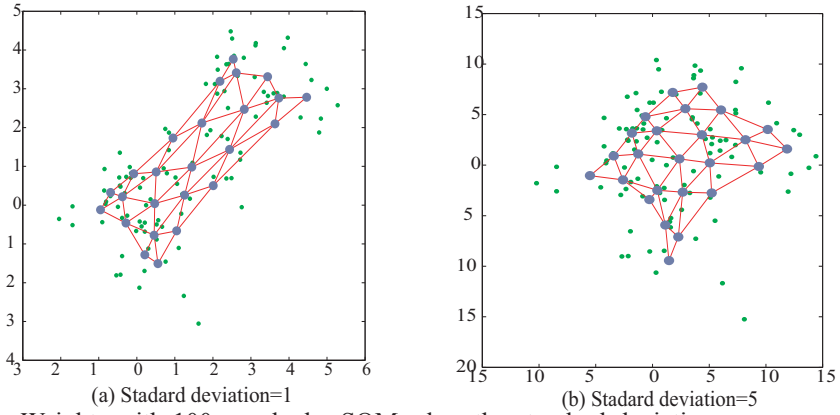
3.1.4 Summary of Results on Generalization

Table I shows the summary of generalization errors. The best average errors in bold faces were obtained by the information maximization. Only when the standard deviation was one, the support vector machine (SVM) showed the performance equivalent to that by the information maximization. When the standard deviation was one, the best error of 0.017 by information maximization was obtained for $r = 2.0$. When the standard deviation was two, the best error was obtained with $r = 1.6$. When the standard deviation was three, the best error was obtained for $r = 2.4$. When the standard deviation was four, the best error was with $r = 2.0$. Finally, when the standard deviation was five, the best error was obtained for $r = 1.7$. Thus, when the parameter r was higher, and correspondingly information was higher, the better performance could be obtained.



(a) Standard deviation=1 (b) Standard deviation=5

Fig. 3: Data with five different values of standard deviation.



(a) Stadard deviation=1 (b) Stadard deviation=5

Fig. 4: Weights with 100 epochs by SOM when the standard deviation was one and five.

Table 1: Summary of experimental results for the artificial data, where "Conv", "With" and "Without" represent the conventional multi-layered networks and networks with and without information maximization, respectively. The values of the parameter r denotes networks with the best generalization performance.

			SOM induced learning			
Std dev		Conv	Without	With	r	SVM
1	Avg	0.022	0.145	**0.017**	2.0	**0.017**
	Std dev	0.007	0.166	0.004		0.005
2	Avg	0.172	0.338	**0.153**	1.6	0.160
	Std dev	0.021	0.112	0.010		0.012
3	Avg	0.272	0.432	**0.250**	2.4	0.266
	Std dev	0.025	0.098	0.022		0.017
4	Avg	0.329	0.483	**0.312**	2.0	0.326
	Std dev	0.024	0.090	0.021		0.015
5	Avg	0.365	0.466	**0.348**	1.7	0.370
	Std dev	0.027	0.061	0.017		0.016

3.1.5 Results without Information Maximization

Figure 6 shows information as a function of the number of layers by the method without information maximization. As can be seen in the figure, information tended to increase gradually when the number of layers increased, though the amount of information was smaller than that by the information maximization. The present method is successfully used to increase the information, because this natural tendency of information increase can be accentuated by the present method. However, when the layer number was three, the information decreased in Figure 6. In Figure 5(b), the information increased when the standard deviation was three. Thus, the present method can increase the information in spite of the absence of natural tendency of information increase. In addition, in Figure 5, information increase seems to be correlated with improved generalization when the standard deviation is larger. This means that when the problem becomes more complex, the present method will be more effective in increasing generalization performance.

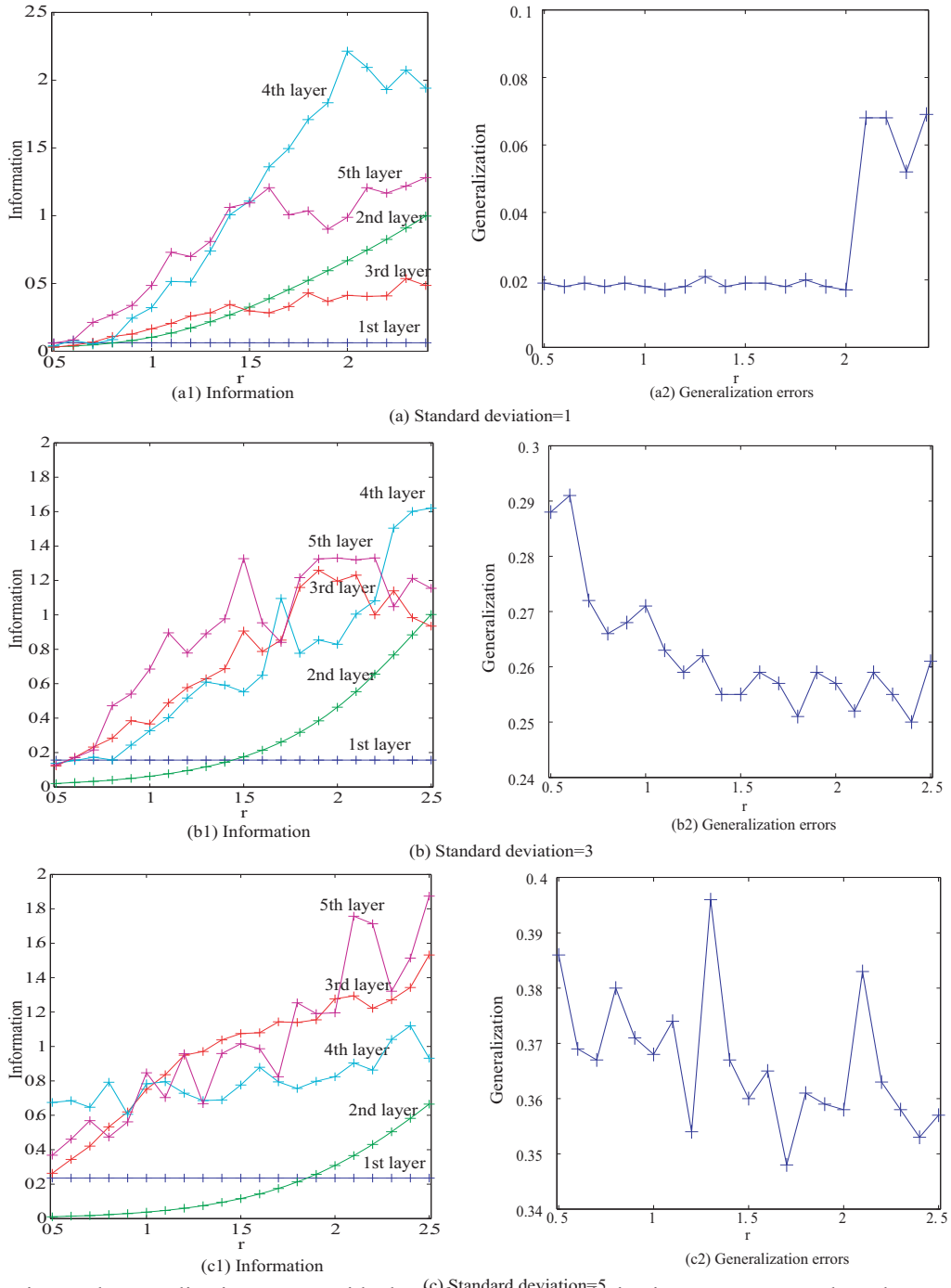


Fig. 5: Information and generalization errors with the information maximization component when the standard deviation increased from one to five.

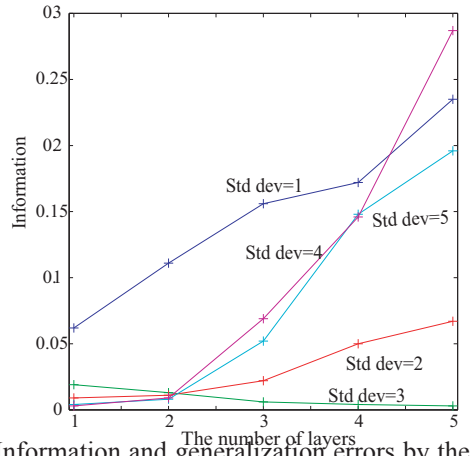


Fig. 6: Information and generalization errors by the method without the information maximization component.

3.2 Application to Spam Data Set

3.2.1 Experimental Outline

The spam data set from the machine learning database [10] was used to examine the performance of the present method. The number of patterns was 4601 with 57 variables and 1000 of them were for training data. The number of validation data set was 1000 and the remaining ones were for testing. The number of input, competitive and output neurons were 57, 25 (5 by 5) and 2, respectively.

3.2.2 Information and Generalization

Figure 7 shows information and generalization errors by the present method. Information content increased gradually when the parameter r increased from 0.5 to 2.0 in Figure 7(a), though in the fourth layer information decreased. Figure 7(b) shows generalization when the parameter r increased from 0.5 to 2.0. The generalization errors decreased gradually and the lowest error was obtained when the parameter r was 1.5. Those results show that when information increased, generalization errors tended to decrease accordingly.

As mentioned, for the fourth layer, the information increased when the parameter r increased from 0.5 to 1.4 in Figure 7(a). However, the information then decreased when the parameter r increased from 1.5 to 2.0 in Figure 7(a). As shown in Figure 7(b), the generalization errors fluctuated when the parameter r increased from 1.5 to 2.0. This fluctuation may be explained by the decrease in information for the fourth layer.

3.2.3 Summary of Generalization Performance

Table II shows the summary of generalization errors. The lowest error of 0.142 was obtained by the present method. The second best error of 0.150 was by the support vector machine. Then, the conventional multi-layered networks shows the third best error of 0.183. The worst error of 0.363 was by the method without the maximum information component.

Table 2: Summary of experimental results for the spam data, where "Conv", "With" and "Without" represent the conventional multi-layered networks and networks with and without information maximization, respectively. The values of the parameter r denotes networks with the best generalization performance.

		SOM induced learning			
	Conv	Without	With	r	SVM
Avg	0.183	0.363	**0.142**	1.5	0.150
Std dev	0.019	0.089	0.035		0.010

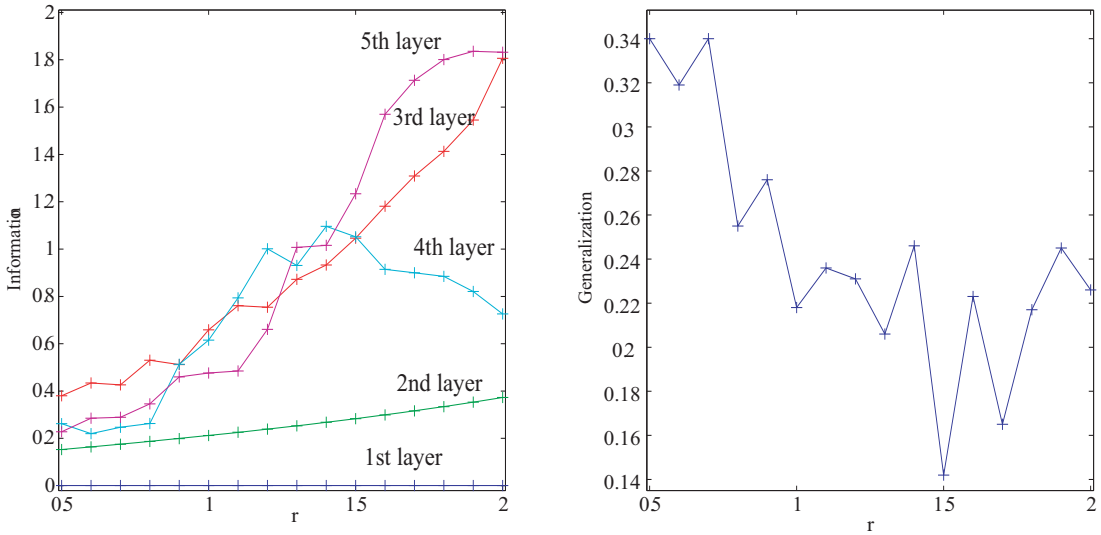
As shown in the table, the largest standard deviation of 0.089 was obtained by the method without the maximum information component. By the maximum information component, the standard deviation decreased from 0.089 to 0.035, which was however the second largest value. Thus, the present method produced results with larger standard deviation and these large values can be decreased by the information maximization component. However, by the present method, the standard deviation was still larger. Thus, it is necessary to examine why such large standard deviation is produced and to develop a method to stabilize the learning by multi-layered neural networks with SOM knowledge.

3.2.4 Comparison of Information Increase

Figure 8 shows the information increase by the method without the maximum information component. The information content increased when the layer number increased from one to three, and then it decreased when the layer number increased from four and five. The information maximization component could increase information in spite of the tendency of information decrease for the higher layers.

3.3 Conclusion

The present paper has shown that it is important to control information content in training multi-layered neural networks. Linsker stated that information content should be maximized for each processing stage. However, simple information maximization does not necessarily imply better performance in multi-layered neural networks. Information content is increased appropriately for each processing stage. Experimental results on the artificial data and spam data set showed that the appropriate control of information increase was essential in increasing better generalization performance. One of the main problems is that the present method sometimes tended to produce the larger variances of results. Thus, it is needed to develop a method to stabilize learning. Though there are some problems to be solved, the present results certainly show that the appropriate control of information content is one of the most important



(a) Information (b) Generalization

Fig. 7: Information and generalization errors by the method with the information maximization component for the spam data set.

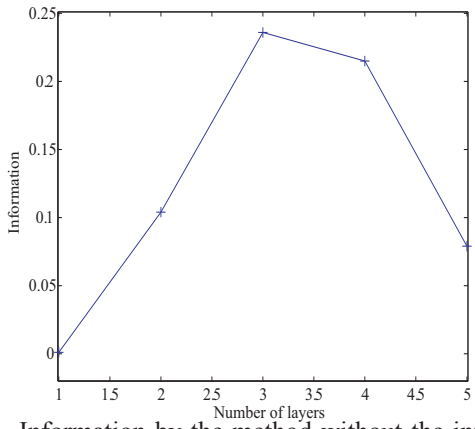


Fig. 8: Information by the method without the information maximization component.

factors in training multi-layered neural networks with SOM knowledge.

References

[1] R. Linsker, "Self-organization in a perceptual network," *Computer*, vol. 21, pp. 105–117, 1988.

[2] R. Linsker, "How to generate ordered maps by maximizing the mutual information between input and output," *Neural Computation*, vol. 1, pp. 402–411, 1989.

[3] R. Linsker, "Local synaptic rules suffice to maximize mutual information in a linear network," *Neural Computation*, vol. 4, pp. 691–702, 1992.

[4] R. Linsker, "Improved local learning rule for information maximization and related applications," *Neural Networks*, vol. 18, pp. 261–265, 2005.

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

[6] G. 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.

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

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

[9] R. Kamimura and R. Kitajima, "Som knowledge induced learning with maximum information principle to improve multi-layered neural networks," in *Proceedings of computational intelligence conferences*, 2015.

[10] K. Bache and M. Lichman, "UCI machine learning repository," 2013.