Information-Theoretic Active SOM for Improving Generalization Performance

Ryotaro Kamimura
IT Education Center and School of Science and Technology, Tokai Univerisity
4-1-1 Kitakaname Hiratsuka Kanagawa, 259-1292, Japan

***Abstract*—In this paper, we introduce a new type of information-theoretic method called "information-theoretic active SOM", based on the self-organizing maps (SOM) for training multi-layered neural networks. The SOM is one of the most important techniques in unsupervised learning. However, SOM knowledge is sometimes ambiguous and cannot be easily interpreted. Thus, we introduce the information-theoretic method to produce clearer and interpretable representations. The present method extends this information-theoretic approach into supervised learning. The main contribution can be summarized by three points. First, it is shown that clear representations by the information-theoretic method can be effective in training supervised learning. Second, the method is sufficiently simple where there are two separated components, namely, information maximization and error minimization component. Usually, two components are mixed in one framework, and it is difficult to compromise between them. In addition, the knowledge obtained by this information-theoretic SOM can be used to solve the shortage of unlabeled data, because the information maximization component is unsupervised and can process all input data with and without labels. The method was applied to the well-known image segmentation datasets. Experimental results showed that clear weights were produced and generalization performance was improved by using the information-theoretic SOM. In addition, the final results were stable, almost independent of the parameter values.**

***Keywords*—SOM; Labeled and Unlabeled; Supervised and Unsupervised; Generalization; Interpretation**

I. Introduction

The present paper aims to introduce a new type of information-theoretic method called "information-theoretic active self-organizing maps (SOM)" to improve generalization performance. The novelty and contribution of the new method can be summarized by three points, namely, the utility of information-theoretic SOM for supervised learning, simple and separated computation, and application to the target shortage problem.

A. Explicit Knowledge for Supervised Learning

First, the present paper aims to show the utility of using the information-theoretic SOM for supervised learning. Self-organizing maps (SOM) have been established as one of the most important unsupervised methods in neural networks [1], [2]. Knowledge obtained by the SOM and represented over connection weights has been exclusively used for the visualization of input patterns [3], [4], [5], [6], [7], [8]. However, one of the main problems is that SOM knowledge is sometimes ambiguous and hard to interpret [9], [10], [11], [12], [13], [14], [15], [16], [17], [18], [19]. Thus, contrary to its good reputation for visualization, practically it has been difficult to use and visualize SOM knowledge.

The information-theoretic SOM has been introduced to improve and clarify SOM knowledge [20], [21]. In this method, information on input patterns is increased while maintaining neighborhood relations between neurons. By controlling the information content of input patterns, connection weights can be modified for better visualization. When this information content is increased, a smaller number of hidden neurons tend to represent input patterns. Because many input patterns are compressed into a smaller number of hidden neurons, it becomes easier to interpret the final activities of hidden neurons. It has been observed that increased information content can improve the interpretation of neurons' behaviors.

The present paper tries to show that this knowledge by the information-theoretic method can be used to train neural networks in supervised ways. While the SOM was originally developed for unsupervised learning, the rich knowledge obtained by this method has stimulated a number of attempts to use it for supervised learning as well [22], [23], [24],[25]. However, they were not necessarily successful and it can be said that they could not reach the performance level of the conventional supervised learning methods. This is because SOM knowledge is itself created in unsupervised ways and not necessarily suited for training supervised neural networks. For this, the information-theoretic SOM has good potentiality, because the knowledge obtained by the method is much clearer than that by the conventional methods. The present paper tries to show the effectiveness of this clear representation for training supervised neural networks.

B. Simple and Separated Computing

The present method is well suited for the supervised SOM [24] [25] with simple and separated computing components. As above mentioned, the SOM knowledge has been used for supervised learning. However, these attempts have not necessarily been successful, because it is difficult to compromise between error minimization and competition processes. Though information-theoretic methods have been applied to supervised learning, one of the major problems is that information maximization is sometimes contradictory to error minimization between targets and outputs. Thus, it becomes difficult to compromise between those two contradictory procedures, especially when the problems become more complex. The present method solves this problem by separating the information maximization and error minimization components.

Borrowing procedures from the fiel of deep learning [26], [27], the present method separates the information maximization or unsupervised phase from the supervised information use phase. By virtue of this separation, each phase, unsupervised or supervised, can focus on its own main task of information maximization or error minimization.

C. Application to Label-Shortage Problem

Then, the present method can be applied to the so-called "label-shortage problem" [28], [29]. As is frequently pointed out, there is little labeled data, while unlabeled data are abundant. A variety of methods have been developed to handle the shortage of labeled data. Among them, the most important methods are semi-supervised learning and active learning. Both methods try to utilize the knowledge of unlabeled data to ameliorate the shortage of labeled data. Active learning tries to recruit the most informative unlabeled patterns to reinforce supervised learning [29]. On the other hand, in semi-supervised learning, information on unlabeled data is used to estimate the targets in explicit or implicit ways [28].

To cope with the "labeled data shortage" problem, we can use the knowledge generated by the information-theoretic SOM, since it can be produced in unsupervised ways. As mentioned above, the two phases of learning, namely, the information maximization and error minimization phases, are separated. In the firs information maximization phase, the information-theoretic SOM is applied to obtained knowledge on input patterns with and without labels. Then, in the supervised phase, this knowledge is used to train connection weights for supervised learning. Similar methods have been proposed for semi-supervised learning, for example, the use of generative models to gain features for classificatio [30], [31]. The present method can use the rich knowledge through the information-theoretic SOM, which can be expected to produce much information on the entire input patterns.

D. Paper Organization

In Section 2, we present how to compute connection weights in both unsupervised and supervised ways. In the unsupervised phase, collective outputs from multiple hidden or competitive neurons are computed. Information maximization processes are realized in terms of decreasing Kullback-Leibler divergence between collected an individual outputs. In the supervised phase, the softmax learning procedures are used to produce update rules. In Section 3, the experimental results of the image segmentation data sets are shown from the well-known machine learning database. First, the most explicit connection weights are obtained by changing the number of winners. Then, generalization errors and the number of epochs are examined. Experimental results for the dataset show that improved generalization could be obtained with clearer connection weights. Compared with generalization by the conventional BP and support vector machines (SVM), the present method gave the better performance. In addition, these results were more stable than those by the conventional method.

II. Theory and Computational Methods

A. Information-Theoretic Supervised SOM

Figure 1 shows how the information-theoretic SOM is applied. In Figure 1(a), there is a small number of labeled data, while unlabeled data are abundant. In the unsupervised phase in Figure 1(b), all data (both labeled and unlabeled) are used for training the information-theoretic SOM. Then, in Figure 1(c), the connection weights by the unsupervised phase are transferred to the supervised learning phase. Taking those connection weights as initial weights, supervised learning is performed. Naturally, the supervised learning is conducted only with labeled data. The problem is whether SOM knowledge by all data (labeled and unlabeled) can be effective in improving performance.

B. Basic Components

As shown in Figure 1, a network is composed of an input layer, competitive layer and output layer. Let us explain how to compute the output from the 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

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

where σ denotes the spread parameter.

In the output layer, we use the sofmax output computed by

$$o_i^s = f\left(\sum_{j=1}^{M} W_{ji} v_j^s\right) \tag{2}$$

where W_{ji} are connection weights from the competitive neurons of the last competitive layer to the output ones.

C. Unsupervised Phase

In the unsupervised phase, the individual neurons try to imitate the outputs by multiple winners to realize self-organization. By normalizing the output, we have the firin probability

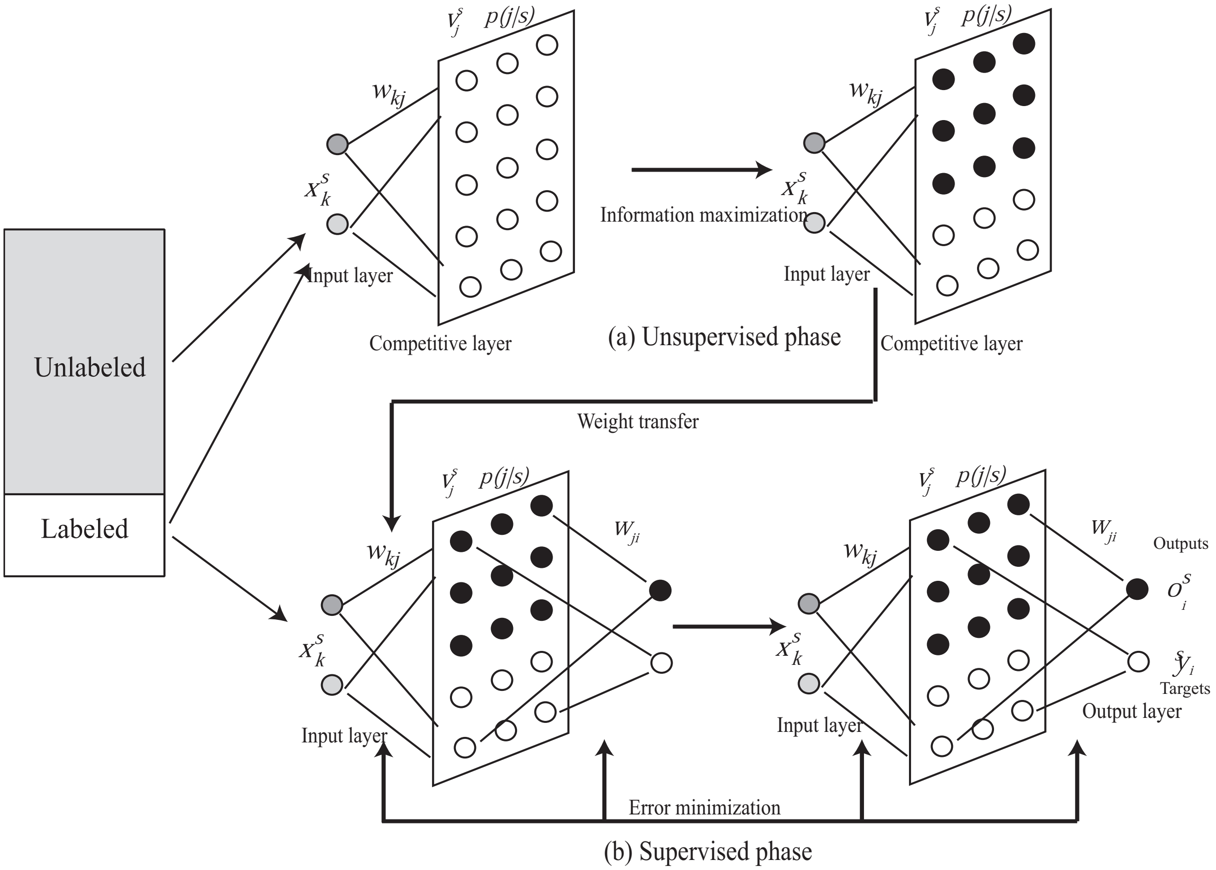
$$p(j \mid s) = \frac{v_j^s}{\sum_{m=1}^{M} v_m^s}. \tag{3}$$

In addition to this firin probability, the output by multiple neurons or winners is used to realize cooperation between neurons as done in the SOM. Now, suppose that the neurons c_1, c_2 are the firs and the second winners, and so on. Then, the corresponding outputs can be ranked as follows:

$$v_{c_1} > v_{c_2} > \ldots > v_{c_M}. \tag{4}$$

Following the formulation of SOM, the distance between the winner and the other neurons is computed by

$$\phi_{jc_1} = \exp\left(-\frac{\parallel \mathbf{r}_j - \mathbf{r}_{c_1} \parallel^2}{2\sigma_{ngh}^2}\right), \tag{5}$$



(a) Unsupervised phase

(b) Supervised phase

Fig. 1. Learning processes for the information-theoretic supervised SOM with the unsupervised (a) and supervised (b) phase.

where $\mathbf{r}_j$ denotes the position of the jth neuron on the output map and σ_{ngh} is the spread parameter. The jth neuron's output is the weighted sum of R winners' outputs and computed by

$$z_j^s(R) = \sum_{m=1}^{R} \phi_{jc_m} v_{c_m}. \tag{6}$$

The firin probability by the multiple winners is define by

$$q(j \mid s; R) = \frac{z_j^s(R)}{\sum_{j=1}^{M} z_m^s(R)}. \tag{7}$$

Learning should be performed to reduce the difference between these outputs. This difference using the Kullback-Leibler divergence is computed by

$$KL = \sum_{s=1}^{S} p(s) \sum_{j=1}^{M} p(j \mid s) \log \frac{p(j \mid s)}{q(j \mid s; R)}. \tag{8}$$

In addition to the KL divergence, there are the other errors which must be minimized, namely quantization errors between connection weights and input patterns

$$Q \quad = \quad \sum_{s=1}^{S} p(s) \sum_{j=1}^{M} p(j \mid s) \|\mathbf{x}^s - \mathbf{w}_j\|^2. \tag{9}$$

Fixing this quantization error and minimizing the KL-divergence, the optimal firin rates are computed by

$$p^*(j \mid s) = \frac{q(j \mid s; R) \exp\left(-\frac{\|\mathbf{x}^s - \mathbf{w}_j\|^2}{2\sigma^2}\right)}{\sum_{m=1}^{M} q(m \mid s; R) \exp\left(-\frac{\|\mathbf{x}^s - \mathbf{w}_m\|^2}{2\sigma^2}\right)}. \tag{10}$$

In addition, for connection weights, the re-estimation formula [20] are obtained by

$$\mathbf{w}_j = \frac{\sum_{s=1}^{S} p^*(j \mid s) \mathbf{x}^s}{\sum_{s=1}^{S} p^*(j \mid s)}. \tag{11}$$

D. Supervised Fine Tuning

In the output layer, the sofmax output is computed by

$$o_i^s = \frac{\exp\left(\sum_{j=1}^{M} W_{ji} v_j^s\right)}{\sum_{m=1}^{N} \exp\left(\sum_{j=1}^{M} W_{jm} v_j^s\right)}, \tag{12}$$

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

$$E = -\sum_{s=1}^{S} \sum_{i=1}^{N} y_i^s \log o_i^s, \tag{13}$$

where y is the target and N is the number of output neurons. The error function is differentiated with respect to connection weights in the competitive and output layer. The update formula for the firs competitive layer is shown by

$$\Delta w_{kj} \quad = \quad \frac{\eta}{S} \sum_{s=1}^{S} \delta_j^s (x_k^s - w_{kj}), \tag{14}$$

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

III. Results and Discussion

A. Image Segmentation Data

1) Experiment Outline: The dataset of the image segmentation was taken from the well-known machine learning database [32]. The dataset was drawn randomly from a database of 7 outdoor images. The images were hand-segmented to create a classificatio for every pixel. The number of input patterns was 2310. The number of input neurons was 19 and the number of output neurons was seven, corresponding to the seven outdoor images. The number of competitive neurons was 5 by 12 neurons, as shown in Figure 2. The number of training patterns was increased from 10 to 100 to demonstrate the effect of the unsupervised information-theoretic method. The number of patterns for the validation set was 500, and the remaining patterns were used for testing (1710 patterns).

2) Improved Interpretation: First, we showed that the proposed method could produce more interpretable connection weights. Figure 3 shows U-matrices (1) and the corresponding labels (2) by the conventional method (a) and the information-theoretic method (b). By the conventional method in Figure 3(a), a clear class boundary in warmer colors could be seen on the upper side of the matrix. Another boundary, though weaker, was seen on the lower side of the matrix. However, by using the information-theoretic method with three winners in Figure 3(b), those class boundaries became stronger in warmer colors.

The same tendency was obtained for the connection weights, meaning that stronger characteristics could be seen by the information-theoretic method. Comparing connection weights by the SOM, some, in particular those in Figure 4(b6), (b-9)-(b-12) were accentuated by the information-theoretic method .

3) Improved Generalization : We compared generalization performance of three methods, namely, the conventional BP, SVM and the information-theoretic method. For fair comparison, the SVM was fine-tuned the box constraint and kernel scale parameters were extensively changed to have the best possible results. As seen in Table I, the generalization errors by the information-theoretic method, including the average, minimum and maximum ones, were much lower than those by the conventional methods. For example, when the number of training patterns was the smallest (10 patterns), the average, minimum and maximum values were 0.620, 0.473 and 0.753 by the conventional method, and 0.732, 0.474 and 0.909 by the SVM, respectively. Those values decreased to 0.438, 0.280 and 0.620 by the information-theoretic method. The differences between them decreased when the number of input patterns increased. However, even if the number of input patterns increased to 100 patterns, the average, minimum and maximum values decreased from 0.185, 0.161 and 0.222 by the conventional method to 0.123, 0.110 and 0.144 by the information-theoretic method. Interestingly, the SVM gave the worst errors when the number of input patterns was 100. In addition, the standard deviation of the generalization errors was smaller by the information-theoretic method. The number of learning epochs was similar across both methods.

TABLE I. Summary of experimental results by the conventional BP, SVM and the information-theoretic method for the segmentation data set. The notation CNV and INF represent the conventional and information-theoretic methods, respectively.

		Generalization error				
Methods	Patterns	Average	Std dev	Min	Max	Epochs
CNV	10	0.620	0.098	0.473	0.753	468
	20	0.543	0.093	0.372	0.695	470
	30	0.433	0.100	0.264	0.581	445
	40	0.352	0.078	0.261	0.486	492
	50	0.291	0.080	0.187	0.451	480
	100	0.185	0.019	0.161	0.222	466
INF	10	0.438	0.098	0.280	0.620	316
	20	0.310	0.069	0.209	0.430	444
	30	0.249	0.062	0.168	0.337	396
	40	0.209	0.059	0.133	0.282	390
	50	0.181	0.041	0.131	0.260	466
	100	0.123	0.011	0.110	0.144	460
SVM	10	0.732	0.165	0.474	0.909	
	20	0.499	0.156	0.311	0.805	
	30	0.426	0.172	0.247	0.836	
	40	0.314	0.057	0.212	0.392	
	50	0.276	0.042	0.213	0.336	
	100	0.201	0.021	0.161	0.236	

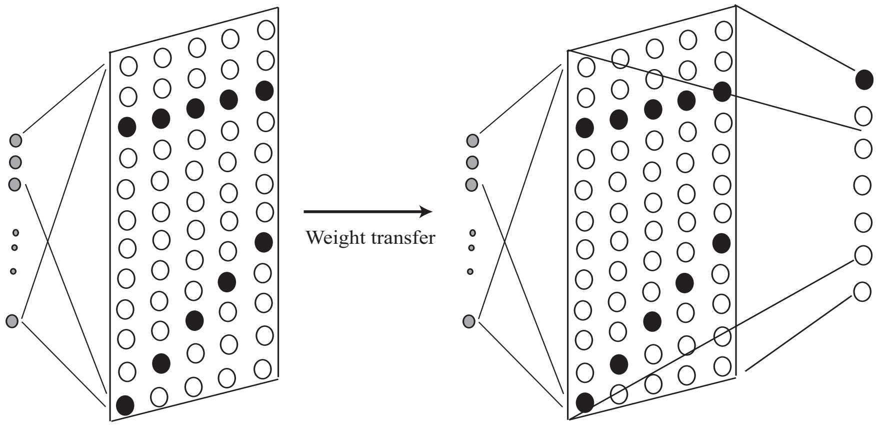
4) Improved Stability: Our analysis showed that the fina values by the information-theoretic method were relatively stable, meaning that the generalization errors and the number of epochs were not relatively affected by the change in the parameter values.

Figure 5 shows the generalization errors by the conventional method in blue and by the information-theoretic method in red as a function of the parameter σ when the number of input patterns increased from 10 (a) to 100 (d). The generalization errors by the information-theoretic method in red were well lower than those by the conventional method in blue, particularly when the number of input patterns was smaller. Even if the number of input pattern was larger, the generalization errors by the information-theoretic method were lower than those by the conventional method, in particular when the parameter σ became smaller.

One of the most important things to note is that the generalization errors generated by the information-theoretic method were more stable than those by the conventional method. When the number of input patterns was ten, the generalization errors were higher for smaller values of the parameter σ. However, the generalization errors were almost constant for all values of the parameter σ. This tendency of stability became more apparent when the number of input patterns increased.

Figure 6 shows that the number of epochs produced the lowest validation errors when the number of training patterns increased from 10(a) to 100(d). By the conventional method, the number of epochs gradually increased when the parameter σ increased. On the other hand, the number of epochs by the information-theoretic method remained stable, almost independent of the parameter σ.

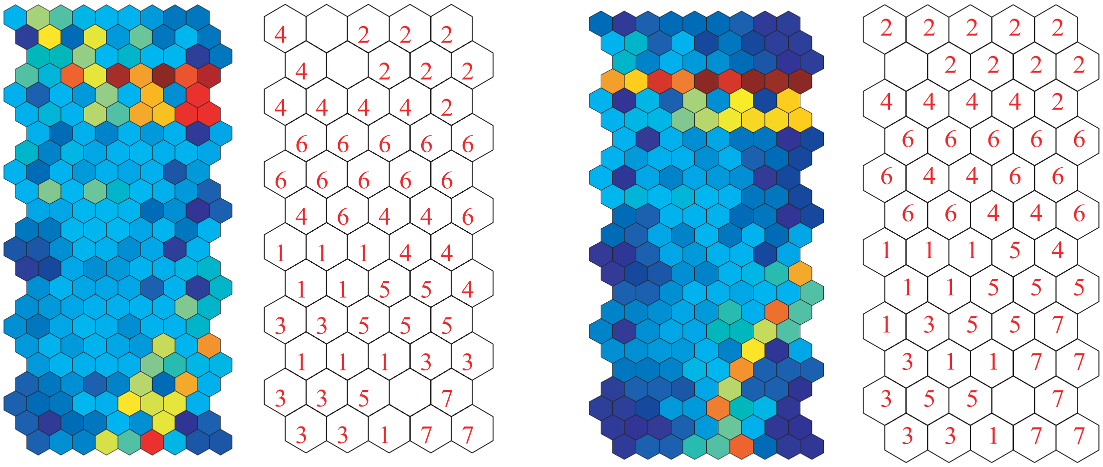
Finally, our method was able to effectively reduce the



(a) Unsuperivsed

(b) Superivsed

Fig. 2. Network architecture with 19 input neurons, 5 by 12 competitive and 7 output neurons for the image segmentation data.



(a1) U-matrix (a2) Labels

(a) Conventional

(b1) U-matrix (b2) U-matrix

(b) Information-theoretic

Fig. 3. U-matrices and labels by the conventional SOM (a) and when the number of winners was three (b).

number of errors in cases where the conventional method failed to do so. Figure 7 shows an example of learning processes. Without knowledge in Figure 7(a), learning was impossible. On the other hand, all errors became immediately smaller by the information-theoretic method.

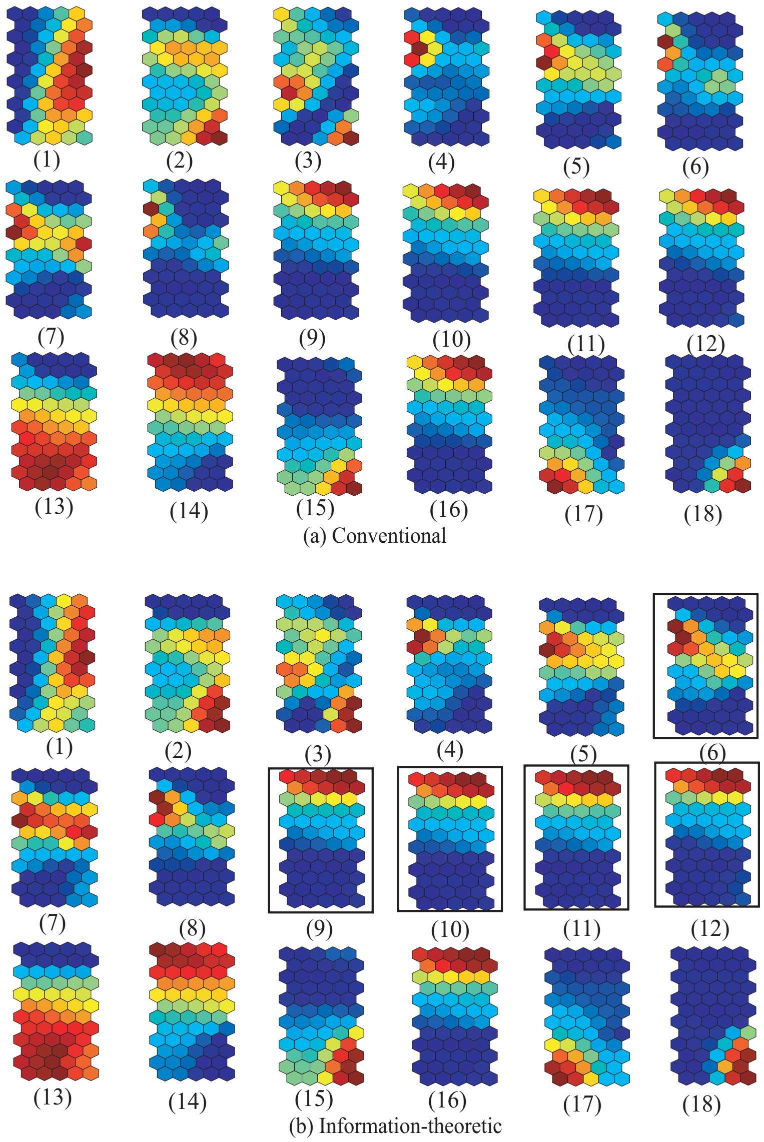
B. Discussion

1) Validity of the Method and Experimental Results: In this paper, we showed that information obtained by the information-theoretic SOM can be used to improve generalization performance with a relatively smaller number of labeled input patterns. In the field of active and semi-supervised learning [30], [31], [33], there have been many attempts to use information content in unlabeled data for training neural networks. The present method suggests that the information-theoretic SOM can be used to train neural networks with information in unlabeled data.

The main results can be summarized by the following three points, namely, improved interpretability, generalization and stability. First, final representations were easier to interpret when using the information-theoretic method. By appropriately increasing the number of winners, final connection weights were well visualized by using the well-known U-matrix in Figure 3, though the number of winners had to be heuristically determined. The final visualized weights showed much clearer maps than those by the conventional method.

Second, the clearer weights could be used to train multi-layered neural networks with better generalization performance. In particular, when the number of training patterns was smaller, improved generalization performance could be more explicitly observed. Figure 7 shows that learning was accelerated even when the learning itself was impossible by the conventional SOM. This suggests that knowledge obtained by the information-theoretic method can be used to train multi-layered neural networks.

Third, the final results were obtainable almost independently of the parameter values. As shown in the experimental results in Figure 5, generalization errors were almost unchanged when the parameter σ was increased. On the other hand, by using the conventional multi-layered neural networks, drastic changes were observed when the parameter σ was changed. The present method, thus, could be used to stabilize learning processes via easy tuning of the parameters. In addition, in Figure 6, the number of training epochs to reach the lowest validation error showed the stable number of epochs by the present method. On the other hand, the conventional



(a) Conventional

(b) Information-theoretic

Fig. 4. Connection weights by the conventional SOM(a) and the information-theoretic method (b).

method showed drastic changes in the number of epochs.

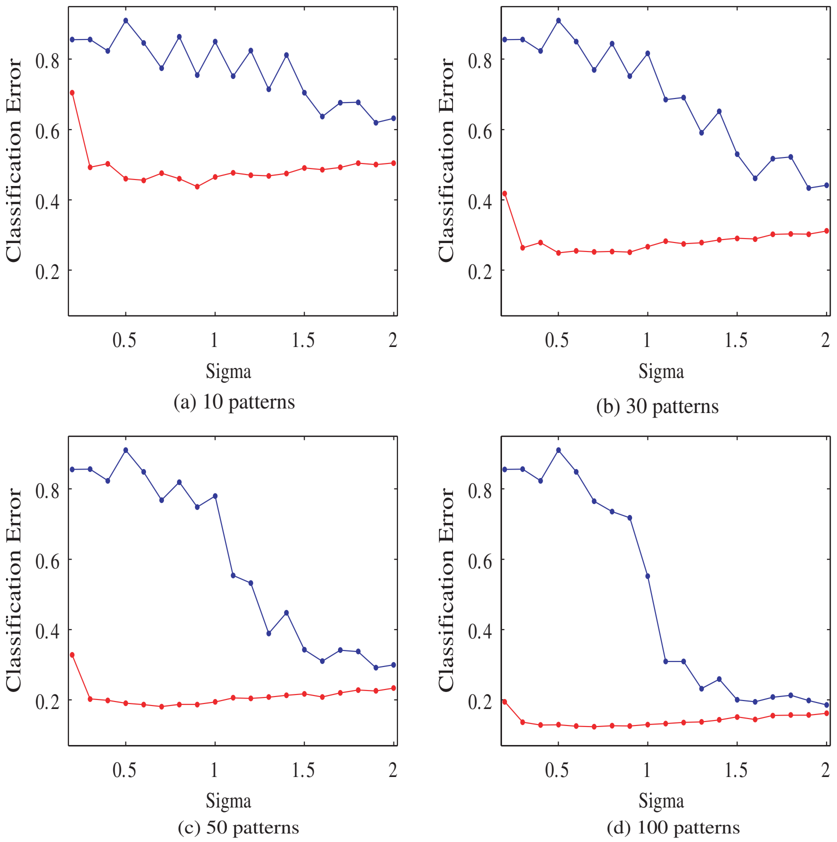
The experimental results showed that the present method could utilize unlabeled data to train supervised networks. In addition, the results obtained by the present method were accompanied by explicit internal representations, permitting possible interpretation. The reason for this improved performance is due to the fact that self-organizing maps, based on competitive learning, aim to separate input patterns into several classes with an equal number of input patterns. Thus, the information-theoretic method classifies input patterns into several classes, and in the final supervised phase, only minor adjustments need to be made to the connection weights for input patterns.

C. Problems of the Method

Though the present method demonstrated relatively greater stability and generalization, it has two problems, namely, the number of winners and relations between interpretability and generalization. Both are due to the absence of any explicit measure of interpretability.

First, the number of winners needed to determine the outputs is uncertain. As mentioned, the number of neurons is critically related to the clarity of final internal representations. Thus, the number of neurons should be increased appropriately. However, because there are no explicit criteria to quantify the clarity of representations, the number of neurons was chosen very heuristically. Accordingly, some criteria for clear representations are needed for detecting the number of winners.

Second, relations between interpretability and generalization are also uncertain, because there are no criteria to ensure the clarity of representations. In the present paper, the clearest possible representations were intuitively chosen at the outset; then, the relations between them were examined. However, the intuitively clearest possible representations did not necessarily produce the best possible generalization performance. To examine the exact relations between interpretability and generalization, some criteria are needed to determine the clarity



(a) 10 patterns

(b) 30 patterns

(c) 50 patterns

(d) 100 patterns

Fig. 5. The generalization errors by the conventional method in blue and by the information-theoretic method in red as a function of the parameter σ when the number of input patterns increased from 10 (a) to 100 (d).

of internal representations.

D. Possibility of the Method

The possibilities of the present method can be summarized by the following four points: interpretation, active learning, deep learning and self-organizing maps. First, the present method provides neural networks with improved interpretation performance. One of the main problems of neural networks is that it is impossible to interpret the final representations obtained through learning [34], [35], [36], [37], [38], [39] [40], [41]. Even if novel new machine learning methods such as SVM, active and semi-supervised learning and deep learning show better performance in particular for generalization, it is practically impossible to interpret the final results. The present method aims mainly to produce interpretable representations and to relate these representations to improved generalization. The method will be the first step towards interpretation-oriented neural networks.

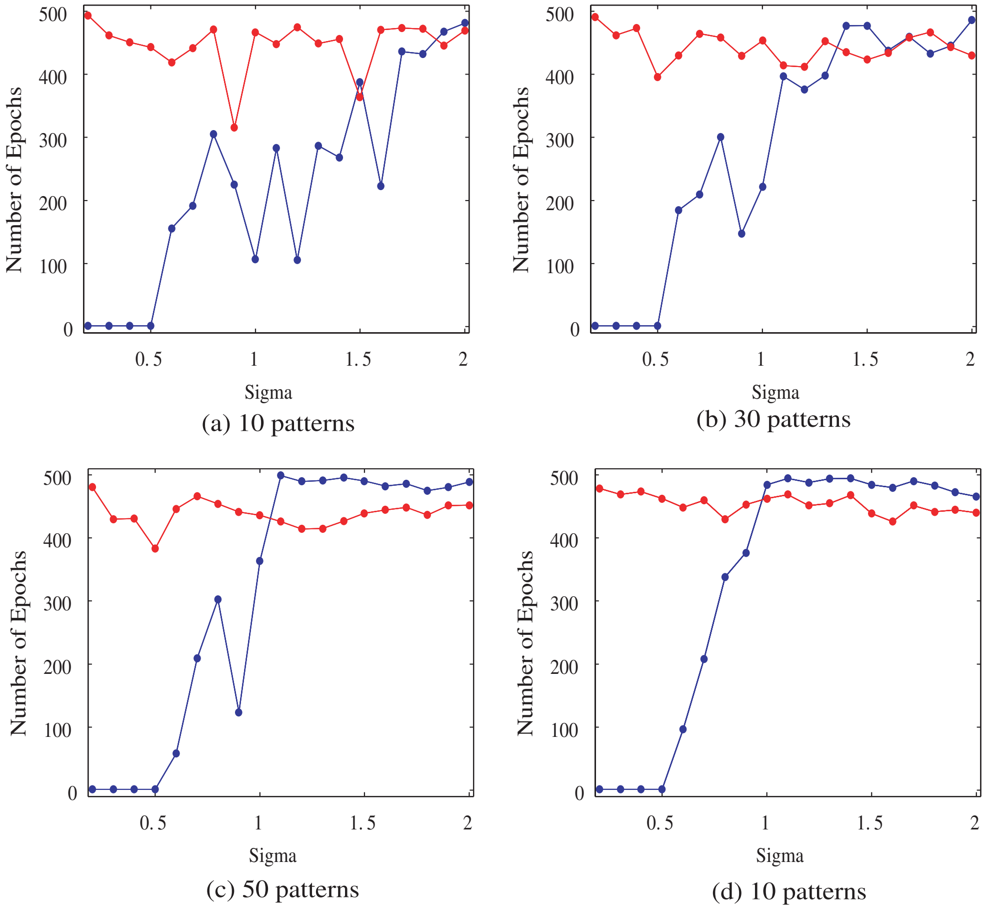
Second, in terms of active learning, the present method does not actively recruit input patterns to be labeled. It can thus be called "passive" learning. The next stage is to actively recruit the patterns to be labeled as done in active learning. In this case, the information content accumulated by the information-theoretic SOM can be used to choose candidate patterns to be labeled. This will be a new form of active learning which considers the information content stored in competitive neurons.

Third, in terms of deep learning, the present method is a form of shallow learning with only one hidden (competitive) layer. However, it is easy to extend this shallow model to a hierarchical deep model by adding multiple competitive layers. In this case, each layer added can be interpreted because the information-theoretic SOM has been developed to explicitly visualize connection weights. Thus, this is a new type of multi-layered network architecture for deep learning in which all hidden layers can be explicitly interpreted.

Finally, this paper suggests a new use for self-organizing maps. It has been shown that the information content by the information-theoretic SOM can be used to visualize only connection weights. However, in addition to visualization, the information obtained by the SOM can be used for many different purposes, such as training. Thus, the present method opens up a new possibility for using the SOM for different purposes.

IV. Conclusion

In this paper, it has been shown that the information-theoretic method can produce clear representations, and that the knowledge obtained by the information-theoretic SOM can be used to train supervised neural networks. Though the SOM



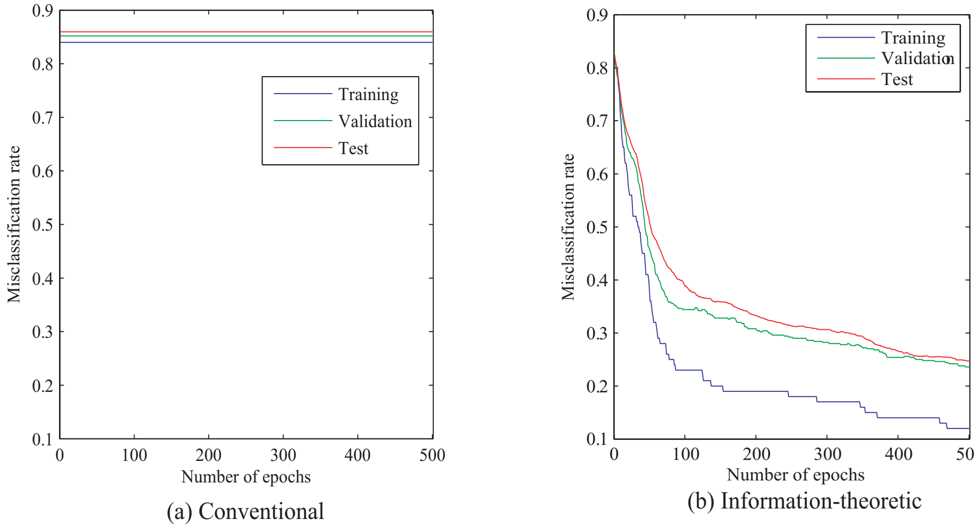
(a) 10 patterns

(b) 30 patterns

(c) 50 patterns

(d) 10 patterns

Fig. 6. The number of epochs by the method without knowledge in blue and with knowledge in red when the number of input patterns increased from 10 (a) to 100(d).



(a) Conventional

(b) Information-theoretic

Fig. 7. Training (blue), validation (green) and testing (red) error rates by the method without knowledge by unsupervised learning (a) and with the knowledge (b), where the parameter σ was 0.2 and 100 patterns.

was developed to create more interpretable representations, it sometimes produces very ambiguous representations. The information-theoretic SOM was introduced in this paper to obtain more explicit and interpretable knowledge on input patterns. In addition, the method aimed to solve the shortage of labeled data problem. In actual situations, the amount of labeled date is scarce, and it is difficul to label unlabeled data. On the other hand, unlabeled data is abundant. Thus, the problem is to fin a method which can maximize the use of abundant unlabeled data. In the present method, the information-theoretic SOM acquires information content on input patterns in unsupervised ways, and can be used to overcome the shortage of labeled data. Finally, the method aimed to solve the problem of compromising between error minimization of targets and outputs, and information maximization. It has been shown that error minimization is not necessarily used to increase information content in the information-theoretic sense. To solve this problem, the information acquisition and use phases are separated. Information was firs maximized in the acquisition phase, and then error was minimized in the information use phase. This separation showed better results for generalization and interpretation.

By applying the method to the image segmentation data sets from the machine learning database, favorable results were obtained and summarized by three points. The information-theoretic methods could produce much clearer internal representations which were accompanied by improved generalization. In particular, when the number of training patterns became smaller, the difference between our method and conventional ones become clearer. In addition, for all experimental results, the stabilization of learning processes was observed. This means that the number of learning epochs and generalization errors tended to be almost independent of the different values of the parameter. Applying the method to the image segmentation data sets revealed that it was able to produce more interpretable representations which were accompanied by improved generalization performance and stability.

However, one of the problems is that the relations between interpretability and generalization are uncertain. This means that though interpretability is roughly related to generalization performance, it is not necessarily accompanied by better generalization. To better relate interpretability to generalization, a method should be developed to unify the two concepts. In addition, the present method should be combined with some active learning techniques to recruit new input patterns to be labeled. Though the problems mentioned above should be solved, the present method nevertheless opened up new possibilities for using SOM knowledge.

References

[1] T. Kohonen, *Self-organization and associative memory*, vol. 8. Springer Science & Business Media, 2012.

[2] T. Kohonen and P. Somervuo, “Self-organizing maps of symbol strings,” *Neurocomputing*, vol. 21, no. 1, pp. 19–30, 1998.

[3] S. Kaski, J. Nikkila, M. Oja, J. Venna, P. Toronen, and E. Castren, “Trustworthiness and metrics in visualizing similarity of gene expression,” *BMC Bioinformatics*, vol. 4, no. 48, 2003.

[4] J. Venna and S. Kaski, “Neighborhood preservation in nonlinear projection methods: an experimental study,” in *Lecture Notes in Computer Science*, vol. 2130, pp. 485–491, 2001.

[5] G. Polzlbauer, “Survey and comparison of quality measures for self-organizing maps,” in *Proceedings of the fifth workshop on Data Analysis (WDA04)*, pp. 67–82, 2004.

[6] J. A. Lee and M. Verleysen, “Quality assessment of nonlinear dimensionality reduction based on K-ary neighborhoods,” in *JMLR: Workshop and conference proceedings*, vol. 4, pp. 21–35, 2008.

[7] T. Villmann, R. D. M. Herrmann, and T. Martinez, “Topology preservation in self-organizing feature maps: exact definitio and measurment,” *IEEE Transactions on Neural Networks*, vol. 8, no. 2, pp. 256–266, 1997.

[8] H.-U. Bauer and K. Pawelzik, “Quantifying the neighborhood preservation of self-organizing maps,” *IEEE Transactions on Neural Networks*, vol. 3, no. 4, pp. 570–578, 1992.

[9] J. Vesanto, “SOM-based data visualization methods,” *Intelligent Data Analysis*, vol. 3, pp. 111–126, 1999.

[10] S. Kaski, J. Nikkila, and T. Kohonen, “Methods for interpreting a self-organized map in data analysis,” in *Proceedings of European Symposium on Artificial Neural Networks*, (Bruges, Belgium), 1998.

[11] I. Mao and A. K. Jain, “Artificia neural networks for feature extraction and multivariate data projection,” *IEEE Transactions on Neural Networks*, vol. 6, no. 2, pp. 296–317, 1995.

[12] C. De Runz, E. Desjardin, and M. Herbin, “Unsupervised visual data mining using self-organizing maps and a data-driven color mapping,” in *Information Visualisation (IV), 2012 16th International Conference on*, pp. 241–245, IEEE, 2012.

[13] S.-L. Shieh and I.-E. Liao, “A new approach for data clustering and visualization using self-organizing maps,” *Expert Systems with Applications*, vol. 39, no. 15, pp. 11924–11933, 2012.

[14] H. Yin, “ViSOM-a novel method for multivariate data projection and structure visualization,” *IEEE Transactions on Neural Networks*, vol. 13, no. 1, pp. 237–243, 2002.

[15] M.-C. Su and H.-T. Chang, “A new model of self-organizing neural networks and its application in data projection,” *IEEE Transactions on Neural Networks*, vol. 123, no. 1, pp. 153–158, 2001.

[16] S. Wu and T. W. Chow, “Prsom: a new visualization method by hybridizing multidimensional scaling and self-organizing map,” *Neural Networks, IEEE Transactions on*, vol. 16, no. 6, pp. 1362–1380, 2005.

[17] L. Xu, Y. Xu, and T. W. Chow, “PolSOM-a new method for multidimentional data visualization,” *Pattern Recognition*, vol. 43, pp. 1668–1675, 2010.

[18] Y. Xu, L. Xu, and T. W. Chow, “Pposom: A new variant of polsom by using probabilistic assignment for multidimensional data visualization,” *Neurocomputing*, vol. 74, no. 11, pp. 2018–2027, 2011.

[19] L. Xu and T. Chow, “Multivariate data classificatio using PolSOM,” in *Prognostics and System Health Management Conference (PHM-Shenzhen), 2011*, pp. 1–4, IEEE, 2011.

[20] R. Kamimura, “Self-enhancement learning: target-creating learning and its application to self-organizing maps,” *Biological cybernetics*, pp. 1–34, 2011.

[21] R. Kamimura, “Constrained information maximization by free energy minimization,” *International Journal of General Systems*, vol. 40, no. 7, pp. 701–725, 2011.

[22] T. Kohonen, *Self-Organization and Associative Memory*. New York: Springer-Verlag, 1988.

[23] T. Kohonen, *Self-Organizing Maps*. Springer-Verlag, 1995.

[24] S. Ohno, S. Kidera, and T. Kirimoto, “Efficien automatic target recognition method for aircraft SAR image using supervised som clustering,” in *Synthetic Aperture Radar (APSAR), 2013 Asia-Pacific Conference on*, pp. 601–604, IEEE, 2013.

[25] J. I. Titapiccolo, M. Ferrario, S. Cerutti, C. Barbieri, F. Mari, E. Gatti, and M. Signorini, “A supervised SOM approach to stratify cardiovascular risk in dialysis patients,” in *XIII Mediterranean Conference on Medical and Biological Engineering and Computing 2013*, pp. 1233–1236, Springer, 2014.

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

[27] 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.

[28] X. Zhu, “Semi-supervised learning literature survey,” Tech. Rep. 1530, Compuer sciences, University of Wisconsin-Madison, 2005.

[29] B. Settles, “Active learning literature survey,” *University of Wisconsin, Madison*, vol. 52, pp. 55–66, 2010.

[30] T. Jaakkola, D. Haussler, *et al.*, “Exploiting generative models in discriminative classifiers” *Advances in neural information processing systems*, pp. 487–493, 1999.

[31] A. Holub, M. Welling, and P. Perona, “Exploiting unlabelled data for hybrid object classification” in *Proc. Neural Information Processing Systems, Workshop Inter-Class Transfer*, vol. 7, 2005.

[32] K. Bache and M. Lichman, “UCI machine learning repository,” 2013.

[33] 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.

[34] L. I. Nord and S. P. Jacobsson, “A novel method for examination of the variable contribution to computational neural network models,” *Chemometrics and Intelligent Laboratory Systems*, vol. 44, pp. 153–160, 1998.

[35] A. Micheli, A. Sperduti, and A. Starita, “Analysis of the internal representations developed by neural networks for structures applied to quantitative structure-activity relationship studies of benzodiazepines,” *Journal of Chemical Information and Computer Sciences*, vol. 41, pp. 202–218, 2001.

[36] D. E. Rumelhart, G. E. Hinton, and R. Williams, “Learning internal representations by error progagation,” in *Parallel Distributed Processing* (D. E. Rumelhart and G. E. H. et al., eds.), vol. 1, pp. 318–362, Cambridge: MIT Press, 1986.

[37] M. Ishikawa, “Structural learning with forgetting,” *Neural Networks*, vol. 9, no. 3, pp. 509–521, 1996.

[38] M. Ishikawa, “Rule extraction by successive regularization,” *Neural Networks*, vol. 13, pp. 1171–1183, 2000.

[39] J. A. Alexander and M. C. Mozer, “Template-based procedures for neural network interpretation,” *Neural Networks*, vol. 12, pp. 479–498, 1999.

[40] G. G. Towell and J. W. Shavlik, “Extracting refine rules from knowledge-based neural networks,” *Machine learning*, vol. 13, pp. 71–101, 1993.

[41] R. Feraud and F. Clerot, “A methodology to explain neural network classification” *Neural Networks*, vol. 15, pp. 237–246, 2002.