Collective Interpretation Controlled by Simplified Selective Information-Driven Learning for Interpreting Multi-Layered Neural Networks

Ryotaro Kamimura
Kumamoto Drone Technology and Development Foundation
Techno Research Park, Techno Lab 203
1155-12 Tabaru Shimomashiki-Gun Kumamoto 861-2202
and IT Education Center, Tokai University
4-1-1 Kitakaname, Hiratsuka, Kanagawa 259-1292, Japan
Email: ryotarokami@gmail.com

***Abstract*—The present paper aims to interpret multi-layered neural networks by considering as many possible internal representations as possible, which is called "collective interpretation." The interpretation is performed in a syntagmatic and paradigmatic way. In the syntagmatic processing, all representations created in each step of the learning processes from the beginning to the final stage are considered. Then, in the paradigmatic approach, we try to deal with all possible representations by the syntagmatic processing. In addition, to make this collective interpretation easier, we control collective interpretation by the selective information, which is simplified to control the cost in terms of the strength of connection weights. The collective interpretation with the simplified selective information augmentation by the cost control was applied to three actual data sets: the traffic, facility for the elderly, and wine data sets. With the first two data sets, we could observe that the networks tried to extract simple and clear relations between inputs and outputs. For the wine data set, because the simple cost reduction could not be effective, the cost was first augmented to reduce the selective information, and then it was increased. The final compressed weights were also simplified for clearer interpretation. The results showed that the collective interpretation with the simple selective information control by the cost control could flexibly deal with input and output information for producing simple and interpretable representations.**

Keywords-collective interpretation; selective information; cost; partial compression; generalization

I. Introduction

The present paper aims to propose a new interpretation method composed of collective interpretation and selective information control [1], [2]. We discuss here several problems related to the conventional interpretation methods and then introduce a concept of collective interpretation. This interpretation tries to take into account as many internal representations as possible with a method of selective information to make the collective interpretation clearer and easier.

A. Interpretation Problem

As has been well known, neural networks have been notorious as one of the typical black-box models in machine learning, though there have been many attempts to interpret their internal representations from the beginning of the research [3]–[7]. Even if neural networks can show good performance in generalization, they have not been accepted as reliable models, because there have been serious risks we must face in unexpected ways. In addition, neural networks have been used to explain and understand human cognitive processes, as was done in the name of connectionism [3], [8], [9]. In this approach, the interpretation of internal representations obtained by neural networks is the objective of the research, and generalization performance, which is nowadays one of the main objectives in neural networks, is only one aspect among many to be explained.

Meanwhile, the massive invasion of neural networks as well as other machine learning techniques into our daily life has caused some concern about their use for our critical decision making. Then, due to the urgent need to respond to the right to explanation [10], there have been many different types of interpretation, in the field of convolutional neural networks (CNN) in particular. Those conventional interpretation methods can be classified into three types: conditional, individual, and intuitive.

First, the interpretation has been based only on a specific condition. Usually, we have tried to interpret an instance of network behavior only when an initial condition is applied. Actually, with a specific initial condition, for example, with a specific set of initial weights, learning is performed, followed by the interpretation of obtained representations. However, the final internal representations are greatly variable, depending on different initial conditions; it is almost impossible to give fixed and stable meanings to those different representations, and furthermore, some contradictory interpretations can be obtained. In particular, when we have tried to apply logical and linguistic rules to the interpretation [11]–[14], we have faced much difficulty in interpreting the different rules. With those formal methods for interpretation, we can produce a number of different formal and logical rules for interpretation. Certainly, we can determine a specific representation for interpretation. For example, we should interpret a representation related to the best generalization. Generalization is an important property to be pursued, but we need to consider many other factors for neural networks when we try to make them as close as possible to human intelligence. Those types of interpretation can be valid only under some specific conditions, such as specific initial conditions, best generalization, and so on. In the present paper, it is supposed that the interpretation should be

as independent as possible of any specific conditions. Second, the conventional methods tend to interpret network behaviors individually, which is closely related to the conditional interpretation. This means that the interpretation has been restricted to an interpretation responding to a specific input or a specific output. In particular, in the convolutional neural network (CNN), many individual interpretations or visualization methods have been developed with much success, for example, the activation maximization [15]–[20], the sensitivity detection [21]–[25], the layer-wise relevance propagation (LRP) [26]–[31], and so on. This is because the intuitive interpretation of image data sets to be discussed immediately below, dealt with by the CNN, has made it possible to understand an instance of network behavior seemingly where the interpretation has been replaced by the intuitive one for a specific image data set. This individual interpretation seems to be successfully applied to many data sets. However, one of the main problems is that the individual interpretation can produce a number of different types of interpretation on one data set, which can be contradictory from each other in some cases. We can say that the sum of individual interpretations cannot necessarily lead us to the full understanding of data sets, because there should be a number of cases in which some interpretations are contradictory to others [32].

The third one is also close to the first and second one, where the interpretation tends to be heavily dependent on our intuitive knowledge of data sets, in particular, when the method is applied to image data sets, as discussed above. Intuition is naturally one of the most important techniques in the explanation, because it is easy to persuade people how neural networks can understand inputs and produce outputs. However, this intuition has prevented us from understanding the true inference mechanism of neural networks. The inference mechanism of neural networks should be different from that of human beings, because the inference mechanism of human beings should be severely constrained culturally and physically so as to maintain their stability and existence [33]. Neural networks have been well known to produce unexpected final outputs, which have been called "adversarial examples" [34], [35]. The adversarial examples can be explained when we can interpret the inference mechanism without human intuition or human cultural bias toward or against the data sets. The neural network can deal with the data sets from a viewpoint that is different from that of human beings. More strongly, the viewpoint cannot be accepted by human beings due to the cultural and physical constraints on their inference mechanism [33], [36]. It should be repeated that human beings are strictly restricted by their physical or cultural conditions that might threaten their existence. Their inference mechanism has been acquired in those severe conditions, which naturally provides strong bias for the interpretation. Thus, the human inference can be only one of many different ways to deal with given inputs appropriately. In short, the adversarial attacks may show one truth about human intuition that the data sets cannot be necessarily well suited to interpret by the inference mechanism of neural networks.

Those limitations and conditions of interpretation seem to be related to the severe shortcomings of neural networks. However, when different types of explanations can be unified, neural networks can be ironically well suited for dealing with unstable and multiple interpretations, compared with the conventional statistical methods. As mentioned above, we have a problem of conditional interpretation, where one of the main problems of neural networks is that they are seriously dependent on initial conditions and where different initial conditions can produce completely different internal representations. Though this phenomenon seems to be one of the main drawbacks of neural networks, it can also be one of the merits of neural networks. This is because they can explain many different aspects of given tasks and data sets just by using different initial conditions. Conventional statistical methods have tried to obtain a representation fixed by a corresponding idealized model, while neural networks try to produce as many different types of representations as possible by using different initial conditions. At this point, all we have to do is to propose a method to unify those different types of representations created by neural networks.

B. Collective Interpretation

In this context, we present here a new type of interpretation method called "collective interpretation," aiming to consider all possible internal representations generated by neural networks. On the contrary, the conventional interpretation method in neural networks is an attempt to interpret only one internal representation. First, as mentioned above, we suppose that different results by different initial conditions should be considered one of the main merits of neural networks. The different results can be produced by an effort to see a given task or data set from a number of different viewpoints and in a number of different ways. All results by different initial conditions should have some meaning to explain the task. Our hypothesis is that all representations by different initial conditions should be taken into account to reach the full understanding of the inference mechanism of neural networks.

In collective interpretation, there are two components: network compression and selective information control. First, we introduce the network compression to simplify multi-layered neural networks. Our method of compression lies in compressing as many representations as possible into the simplest form for explaining the core knowledge obtained by neural networks. Model compression has received due attention recently to simplify multi-layered neural networks [37]–[44]. Those conventional methods have aimed to replace complicated multi-layered networks with simpler ones, keeping the same generalization performance as much as possible. Thus, the internal representations obtained by those methods cannot inherit the original representations of complex multi-layered neural networks. On the contrary, we have proposed a method to compress multi-layered neural networks [45], keeping information stored in weights in multi-layered neural networks as much as possible.

In addition to different types of internal representations, we consider connection weights in syntagmatic and paradigmatic ways. First, by restricting learning individually and conditionally, we train neural networks with a specific set of initial conditions and input patterns, and all representations in the course of learning are collected. This process is called "syntagmatic processing," where all representations, obtained in each learning step, are taken into account to produce collected representations. This syntagmatic processing should be performed for different initial conditions and input patterns, producing a number of different types of representations. Then, we should collect all those different compressed representations, which is called "paradigmatic processing." Collective interpretation is composed of the network compression where syntagmatic processing is first applied, followed by paradigmatic processing to deal with all possible representations.

An ideal collective interpretation should consider all instances obtained in neural learning, and it should extract some core structure by which all instances can be generated. The present paper uses a kind of partial conditional collective interpretation, where one condition is assumed for the collective interpretation. The condition is that information per cost should be maximized for simplification. Information-theoretic methods have been introduced from the beginning of research into neural networks, producing many principles affecting studies on neural information processing. For example, Linsker's maximum information preservation principle has had much influence on neural computing [46]–[49], in which some visual processing can be explained by the maximum information principle. Intuitively, we humans try to collect surrounding information to secure our existence and to keep it as secure as possible. Thus, though the principle should play more important roles in extracting some principles in neural computing, there have been few attempts made to use information-theoretic principles following important past studies [50]–[55]. In this context, we try to show how neural networks are transformed under the condition that information per cost is maximized. Then, we try to show that we can disentangle complicated representations into the simplest ones when the information per cost is maximized. For this, we introduce a method to increase the selective information for connection weights, expecting that those weights will be selected to be disentangled from each other.

However, when the information is formulated in the classical form of information measures such as entropy and mutual information, it is not so easy to understand how those measures are concretely related to the disentanglement of representation. This is because the abstract and ambiguous property of information, accompanied by the need for much computational resources, has prevented us from using them appropriately for the actual formulation. The present paper proposes a more simplified method to compute the selective information, which is not the abstract measure of information but which has the actual meaning of the number of important connection weights. Thus, the selective information can be applied to neural networks and to understanding how information can be stored in terms of the number of connection weights.

The selectivity has played important roles in neural networks, in particular, in generalization [56]–[61]. We should choose a small number of important connection weights, based on some criteria on the importance. However, it is impossible to know the importance of connection weights, and it has been stressed that the selectivity is of no use in generalization [56], [59], [61]. For coping with this problem of selectivity, we use the passive method to extract important ones. We use the concept of cost [62] in terms of strength of connection weights. We consider a connection weight important only when this weight remains strong by introducing the cost reduction method. This method is closely related to the conventional weights decay, but the fundamental difference is that the cost reduction is performed independently of error minimization. This simple and independent cost reduction method can eventually produce a small number of important weights.

Selective information can be maximized to simplify neural networks, but this simplification is not necessarily successful. The information on the given data set should be naturally obtained through inputs. However, those inputs are artificially prepared by our knowledge on the data set. When these inputs cannot be used to transmit information on inputs, simplified networks cannot necessarily represent information on relations between inputs and outputs. In this case, we must decrease information on inputs as much as possible. Thus, we first try to increase the selective information to simplify networks. Then, if it is impossible to simplify them, we try to decrease selective information in the first place and then increase selective information.

C. The Purpose of the Present Study

Considering the above problems, the present paper aims to propose a new interpretation method with three properties. First, connection weights produced by neural networks are exhaustively considered in syntagmatic and pragmatic ways. This tries to take into account all possible representations by the neural network. Second, the network simplification is performed not by the selective information maximization directly but by the corresponding cost minimization. Thus, the learning procedures are greatly simplified. Third, when the selective information maximization cannot give acceptable results, we first minimize the selective information by increasing the cost. Then, the ordinary selective information minimization is applied. This can be used to eliminate harmful information obtained through inputs.

D. Paper Organization

The paper has been organized as follows. In Section 2, after briefly explaining the concept of collective interpretation, we try to explain full compression with syntagmatic and paradigmatic compression. In addition, to see the intermediate states of learning, we introduce partial compression and how to partially compress intermediate layers. Then, we introduce the potentiality and corresponding selective information, followed

by the computational methods of cost reduction and augmentation. We applied the method to three data sets, namely, the traffic, facility for the elderly, and wine data sets. In all cases, we first tried to show that syntagmatic and paradigmatic compression could produce compressed weights close to the correlation coefficients between inputs and targets. In the first two data sets, by the cost reduction, we could increase the ratio of selective information to its cost. By the partial compression, we could see that the present method tried to deal with inputs from the lower hidden layers, while the other conventional methods could not consider inputs well. Because the wine data set could not produce reasonably good interpretation results, we first decreased the selective information by increasing the corresponding cost, and then, the selective information was increased by decreasing the cost. In all experimental results, the final collective interpretation showed that the main characteristics were based on the correlation coefficients between inputs and targets. The differences between compressed weights and correlation could be used to detect the effects of non-linear relations between inputs and outputs.

II. Theory and Computational Methods

We explain here the concept of collective interpretation, taking into account all internal representations by the syntagmatic and paradigmatic compression. In addition to the full network compression, we introduce the partial compression to see the states of intermediate layers. Then, we introduce the simplified information-theoretic method by the selective information, representing how many connection weights are combined with neurons. After formulating the potentiality and selective information, we introduce a practical method to control selective information, where instead of direct control of selective information, we try to control the cost in terms of strength of weights. Finally, we present a two-step learning method in which the cost is first increased, and then decreased when the simple cost reduction could not be applied.

A. Compression

1) Collective Interpretation: We introduce here a concept of collective interpretation in which we try to take into account all possible internal representations, assumed to have equal importance, created by the neural network. One of the main shortcomings of neural networks is that their learning behaviors are sometimes completely different when different initial conditions and different subsets of a data set are given, as shown on the left-hand side of Figure 1. However, we suppose here that this shortcoming of different learning behaviors should be one of the most important merits of neural networks. This means that a neural network tries to see a target object from many different points of view, corresponding to different initial conditions and different subsets of a data set. Then, we suppose that all representations created by the neural network should have some meaning related to the properties of the target objects. We more strongly assume that all possible representations should have the same importance, at least, in terms of interpretation, dealt with in this paper. As shown in Figure 1, a neural network can produce many different final representations by different initial conditions and different subsets of a data set. Then, we should interpret how a neural network tries to produce outputs, based on the corresponding inputs, considering all possible internal representations they create. The interpretation, taking into account all possible internal representations, can be called "collective interpretation" in this paper.

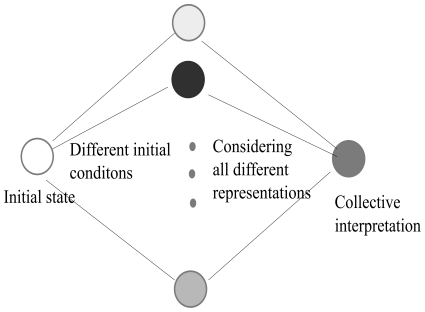

Fig. 1. Collective interpretation aiming to consider all possible internal representations created by a neural network.

2) Full Compression: For interpreting multi-layered neural networks, we first compress them into the simplest ones, as shown in Figure 2. We try here to trace all routes from inputs to the corresponding outputs by multiplying and summing all corresponding connection weights.

First, we compress connection weights from the first to the second layer, denoted by (1,2), and from the second to the third layer (2,3) for an initial condition and a subset of a data set. Then, we have the compressed weights between the first and the third layer, denoted by (1,3).

$$w_{ik}^{(1,3)} = \sum_{j=1}^{n_2} w_{ij}^{(1,2)} w_{jk}^{(2,3)} \tag{1}$$

Those compressed weights are further combined with weights from the third to the fourth layer (3,4), and we have the compressed weights between the first and the fourth layer (1,4).

$$w_{ik}^{(1,4)} = \sum_{k=1}^{n_3} w_{ik}^{(1,3)} w_{kl}^{(3,4)} \tag{2}$$

By repeating these processes, we have the compressed weights between the first and sixth layer, denoted by $w_{iq}^{(1,6)}$. Using those connection weights, we have the final and fully compressed weights (1,7).

$$w_{ir}^{(1,7)} = \sum_{q=1}^{n_6} w_{iq}^{(1,6)} w_{qr}^{(6,7)} \tag{3}$$

Because we consider all routes from the inputs to the outputs, the final connection weights should represent the overall characteristics of connection weights of the original multi-layered neural networks.

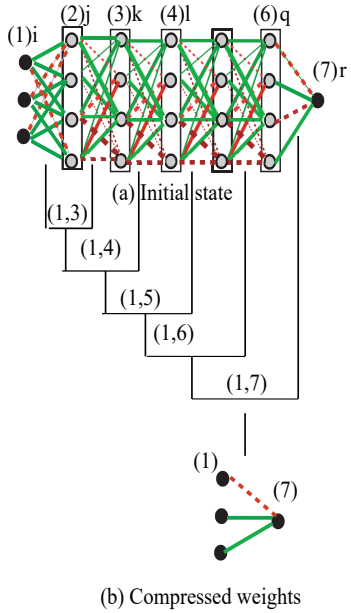

Fig. 2. Full compression for an initial condition and a subset of a data set from a seven-layered to a two-layered network without hidden layers.

3) Syntagmatic and Paradigmatic Compression: The full compression actually is composed of syntagmatic and paradigmatic compression in Figure 3. With an initial condition and a set of input patterns, we train a neural network, taking into account all internal representations by all possible conditions and subsets of a data set. For simplicity's sake, we suppose that only initial conditions are changed, but actually, the subset of the data set can be changed. Then, we average obtained connection weights over all weights obtained in a process of learning for the initial condition. Let us take an example of connection weights from the sixth to the seventh layer only and the maximum number of training steps for the sth initial condition in Figure 3(a4). Then, we can average all possible weights for all training epochs. For the weights from the sixth to the seventh weights $(6, 7; t)$ for the tth learning epoch, we can average all possible weights

$$\bar{w}_{qr}^{(6,7)} = \frac{1}{t_s} \sum_{t=1}^{t_s} w_{ir}^{(6,7;t)} \tag{4}$$

where t_s denotes the maximum number of learning steps for the sth initial condition. All other connection weights are averaged in the same way. Then, we compress those average weights in full compression.

$$\bar{w}_{ir}^{(1,7)} = \sum_{q=1}^{n_6} \bar{w}_{iq}^{(1,6)} \bar{w}_{qr}^{(6,7)} \tag{5}$$

where $\bar{w}_{iq}^{(1,6)}$ denote the compressed averaged weights up to the sixth layer. This compression can be called "syntagmatic compression" in Figure 3, because it tries to compress all connection weights obtained for all learning steps.

Finally, the syntagmatically compressed weights are averaged over all initial conditions and subsets of the data sets. For simplicity's sake, we restrict the compression for an initial condition, and we have the paradigmatic compression in Figure 3(b).

$$\bar{w}_{ir} = \frac{1}{s_m} \sum_{s=1}^{s_m} \bar{w}_{ir}^{(1,7)} \tag{6}$$

where s_m denotes the maximum number of initial conditions. We should repeat that we try to consider all possible representations created by neural networks. Thus, we can deal with all connection weights for all learning steps and by all different initial conditions and input patterns.

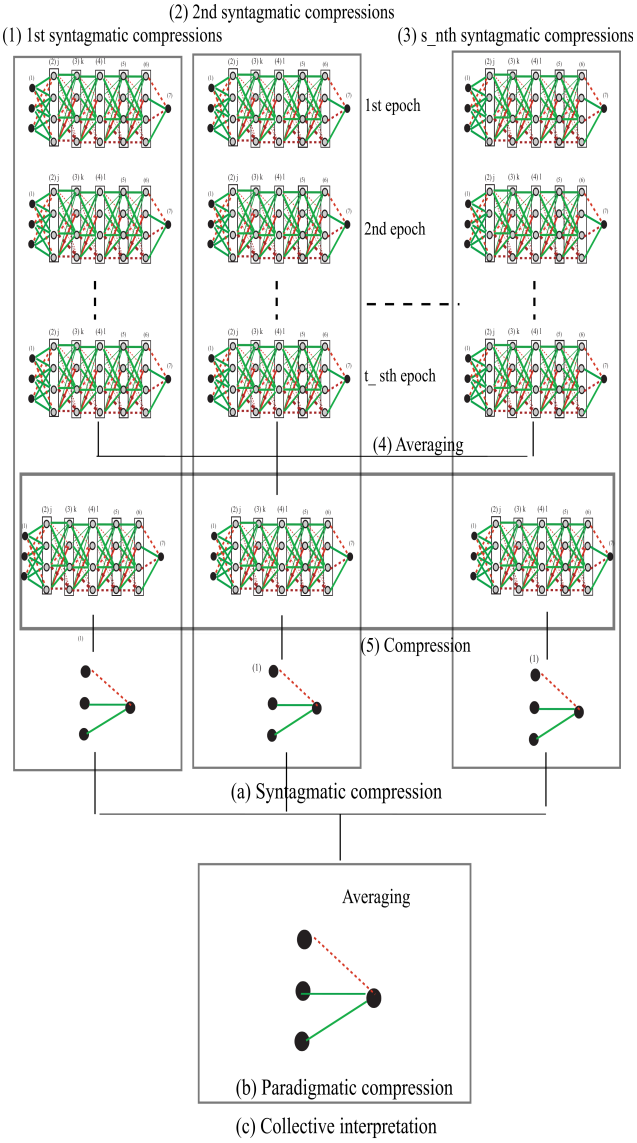

Fig. 3. Collective compression composed of syntagmatic (a) and paradigmatic (b) compression for collective interpretation (c).

4) Partial Compression : In addition to the full compression, we need to examine the outputs from the intermediate layers. For this purpose, we introduce the partial compression, in which compression is applied up to a specific layer. As shown in Figure 4, we illustrate the partial compression up

to the fourth layer. Now, let us assume that we have already compressed weights up to the fourth layer, denoted by $w_{il}^{(1,4)}$. In addition, the number of neurons in all hidden layers is supposed to be the same. The partially compressed weights up to the fourth layer can be computed by

$$w_{ir}^{(1,4,7)} = \sum_{q=1}^{n_6} w_{iq}^{(1,4)} w_{qr}^{(6,7)} \tag{7}$$

where $w_{iq}^{(1,4)}$ denote connection weights, compressed up to the fourth layer. For the other intermediate layers, we can compute the same partially compressed weights. The partial compression aims to examine to what degree the intermediate layers contain information on inputs as well as outputs.

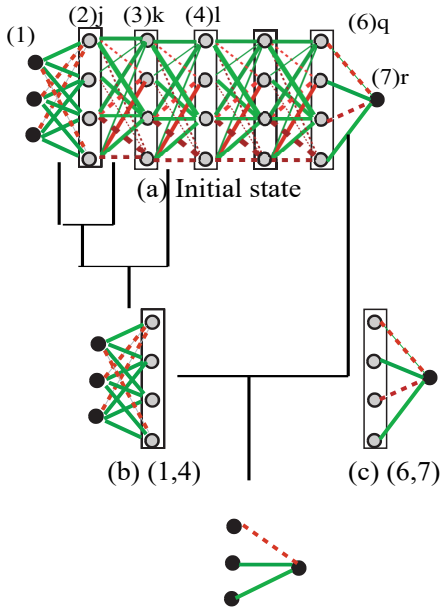

Fig. 4. An example of partial compression where only weights up to the fourth layer are compressed.

B. Reduction and Augmentation of Selective Information

1) Selective Information and Its Cost: The selective information can be defined by using the selective potentiality of connection weights. When the selective information increases, a small number of connection weights tends to be connected with some specific neurons. The individual potentiality of connection weights can be defined by the absolute values of weights, for example, from the second to the third layer, represented by (2,3), which is computed by

$$u_{jk}^{(2,3)} = \mid w_{jk}^{(2,3)} \mid \tag{8}$$

Then, we normalize these values by their maximum ones.

$$h_{jk}^{(2,3)} = \frac{u_{jk}^{(2,3)}}{\max_{j'k'} u_{j'k'}^{(2,3)}} \tag{9}$$

where the maximum operation is over all connection weights between two layers. Then, summing all these normalized values, the selective potentiality can be defined by

$$H^{(2,3)} = \beta_1 \sum_{j=1}^{n_2} \sum_{k=1}^{n_3} \left[\frac{u_{jk}^{(2,3)}}{\max_{j'k'} u_{j'k'}^{(2,3)}} \right] \tag{10}$$

where n_2 and n_3 denote the number of neurons in the second and the third layer, and β_1 is a parameter to control the strength. It should be larger than zero. Then, the complementary potentiality is defined by

$$g_{jk}^{(2,3)} = 1 - \frac{u_{jk}^{(2,3)}}{\max_{j'k'} u_{j'k'}^{(2,3)}} \tag{11}$$

Summing all these normalized values, the selective information can be defined by

$$G^{(2,3)} = \beta_2 \sum_{j=1}^{n_2} \sum_{k=1}^{n_3} \left[1 - \frac{u_{jk}^{(2,3)}}{\max_{j'k'} u_{j'k'}^{(2,3)}} \right] \tag{12}$$

In addition, we need to define the corresponding cost to represent the potentiality and information. In this paper, the cost is simply the sum of all the absolute weights.

$$C^{(2,3)} = \sum_{j=1}^{n_2} \sum_{k=1}^{n_3} u_{jk}^{(2,3)} \tag{13}$$

We suppose that the cost representing the information should be as small as possible, and then the final function to be controlled for the selective potentiality is

$$R^{(2,3)} = \frac{H^{(2,3)}}{C^{(2,3)}} \tag{14}$$

Then, for the selective information, the function to be controlled is

$$R^{(2,3)} = \frac{G^{(2,3)}}{C^{(2,3)}} \tag{15}$$

2) Cost Control for Sensitive Selective Information: The selective information should be augmented and the corresponding cost should be reduced in the majority of data sets. When we try to control the ratio of selective information to its cost, we have two possible ways to do so: selective information control or cost control. Because it is sometimes difficult to directly control the selective information, we focus on the cost and try to control it. In addition, when we try to increase the selective information, one of the major problems is that we cannot identify important connection weights or see whether a weight plays a major role in interpretation or generalization. Thus, we pay attention to the corresponding cost, and we try to reduce the cost as much as possible, which is expected to increase the selective information eventually. For this case, the connection weights at the $t+1$th learning step are simply computed by

$$w_{jk}^{(2,3)}(t+1) = \beta_1 \, w_{jk}^{(2,3)}(t) \tag{16}$$

where β_1 should range between zero and one, because we try to reduce the cost or the strength of connection weights.

However, for some data sets, we have found that the selective information augmentation and the corresponding cost reduction cannot be accompanied by disentangling connection weights into simplified ones for better interpretation. In those cases, we first reduce the selectivity at the expense of higher cost. Then, we try to increase the selective information and to decrease the corresponding cost. Figure 5 shows the process of a two-step method of selective information reduction and augmentation. In the initial state in Figure 5(a), connection weights are randomly initialized with the intermediate selectivity. Then, we try to decrease the selectivity at the expense of larger connection weights or higher cost in Figure 5(b). Finally, we try to decrease the cost and at the same time increase the selective information in Figure 5(c). In this case, for the initial learning steps, we have

$$w_{jk}^{(2,3)}(t+1) = \beta_2 \, w_{jk}^{(2,3)}(t) \tag{17}$$

The parameter β_2 should be larger than one. We try to increase the strength of connection weights. This leads us to the augmentation of selective potentiality at the expense of cost. We use this method because it is easy to decrease the selective information. Then, for the remaining learning steps, we have the same assimilation rule

$$w_{jk}^{(2,3)}(t+1) = \beta_1 \, w_{jk}^{(2,3)}(t) \tag{18}$$

However, the parameter β_1 should be between zero and one to reduce the cost and correspondingly to increase the selective information.

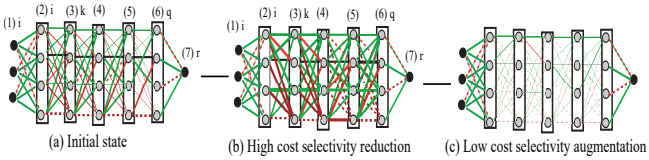

Fig. 5. Selective information augmentation (c) through higher cost selective information reduction (b).

3) Assimilation: Depending on their strength, weights are controlled to be smaller or larger. However, when the weights are controlled by the parameter β, we need to re-train a neural network to assimilate learning processes. We try to repeat this process of assimilation many times. One of the possible ways to do so is to use the dth sub-epoch t_d of the tth learning step, and it can be computed by

$$t_d = \theta_1 \left(\frac{t}{t_{max}} \right)^{\theta_2} + \theta_3 \tag{19}$$

where d is the dth sub-epoch of step of the tth learning step and t_{max} is the maximum number of learning steps with three parameters, $\theta_1, \theta_2, \theta_3$, to control the effect of assimilation.

Figure 6 shows a process of assimilation for a learning step. First, weights in an initial state in Figure 6(a) are multiplied by the parameter β (for example, smaller values) in Figure 6(b), and the strength of weights is reduced in proportion to the parameter β in Figure 6(c). Then, we repeat the assimilation steps for the learning step several times in Figure 6. Because the effect of the parameter β is weakened in this process of assimilation to reduce training errors, we must have weakened weights less than those at the initial stage of assimilation due to the effect of error minimization in Figure 6. Then, we repeat this process of assimilation for each learning step to obtain the final reduced weights. One of the important features of this assimilation method is that the assimilation (error minimization) and potentiality assignment (application of the parameter β) are performed separately. First, the strength of weights is reduced, and then, the effect of the parameter is assimilated (error minimization). This method, thus, can resolve the contradiction between error minimization and regularization, which are usually simultaneously performed.

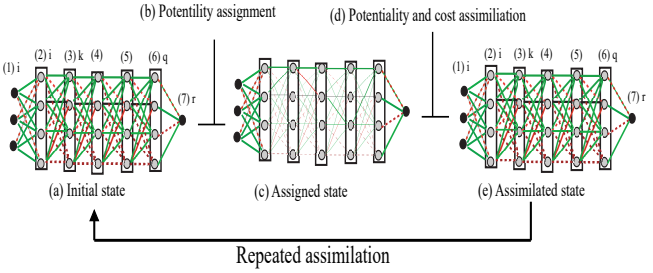

Fig. 6. A computational method to assimilate the effect of cost reduction.

III. Results and Discussion

We present here experimental results on three data sets: traffic, facility for the elderly, and wine. In the first two data sets, we used the simple cost reduction method to increase the selective information. For the third data set, the simple cost reduction could not produce reasonable results, so we first augmented the cost, and the usual cost reduction to increase the selective information was applied. With those three methods, we tried to show that we could compress networks syntagmatically and paradigmatically with the aid of cost or selective information control into simpler and clearer networks, whose connection weights could be closer to the correlation coefficients between inputs and targets of the original data sets. In addition, we could extract some properties due to the non-linear processing of neural networks.

A. Traffic Data Set

1) Experimental Outline: The database was created with records of behavior in urban traffic in the city of Sao Paulo in Brazil [63]. The number of inputs was 17, and the number of patterns was 135. Seventy percent of the data set was used for training, and the remainder was for testing. To make the reproduction of the present results easier, we tried to use the scikit-learning package with all default values except for the tangent-hyperbolic activation function and the number of epochs, which was changed according to the equation described above. Table I shows the parameter values for the experiments. In the following sections on experimental results, we used the same parameter values for the easy

TABLE I
SUMMARY OF PARAMETER VALUES FOR THE TRAFFIC DATA SET.

Parameters	Values
β_1	0.85
θ_1	5
θ_2	1
θ_3	5

reproduction of all results except for the third results, where a new parameter $\beta_2 = 1.3$ for augmentation of potentiality or information minimization was introduced.

2) Syntagmatic and Paradigmatic Compression : We compared compressed weights with correlation coefficients between inputs and targets of the original data set, supposing that the correlation coefficients were meaningful for describing the relations between inputs and outputs. The results show that the present method could produce syntagmatically and paradigmatically compressed weights close to the correlation coefficients between inputs and targets of the original data set. Though the weight decay and conventional method could produce reasonably high correlations, they were still lower and behind the correlations by the present method.

Figure 7 shows the syntagmatic (left) and paradigmatic (right) compression for the traffic data set for 100 different initial conditions and 100 different subsets of the data set. One of the main characteristics is that, when the parameter β_1 was 0.85 for the cost reduction, correlation coefficients between syntagmatically compressed weights and original correlations between inputs and targets of the original data set were much higher than those by any other method, and close to one (perfect correlation) in the box on the left-hand side of Figure 7(a). The box on the right-hand side of Figure 7 (a) shows the results of paradigmatic compression, and we could see that when the number of different initial conditions and different subsets of the data set increased, the correlation coefficients became close to the maximum of one. When the parameter α for the weight decay was set to 0.1 in Figure 7(b), the correlation coefficients for the syntagmatic compression became lower than those by the cost reduction in Figure 7(left, a). For the paradigmatic compression in Figure 7(right, b), the correlations became larger gradually, but the final correlations were lower than those by the present method in Figure 7(a). Finally, even without weight decay, the final results were quite similar to those with the weight decay in Figure 7(c).

The results confirmed that the collective interpretation could extract relations between inputs and outputs that were close to the original correlation coefficients between inputs and targets. Thus, neural networks, in particular, with the cost reduction, could disentangle connection weights that could be compressed to represent simple relations between inputs and outputs.

3) Selective Information, Cost, and Ratio: The results show that, though the new method could not increase selective information in the later stages of learning, the cost was reduced sufficiently to increase the ratio of information to its cost. On

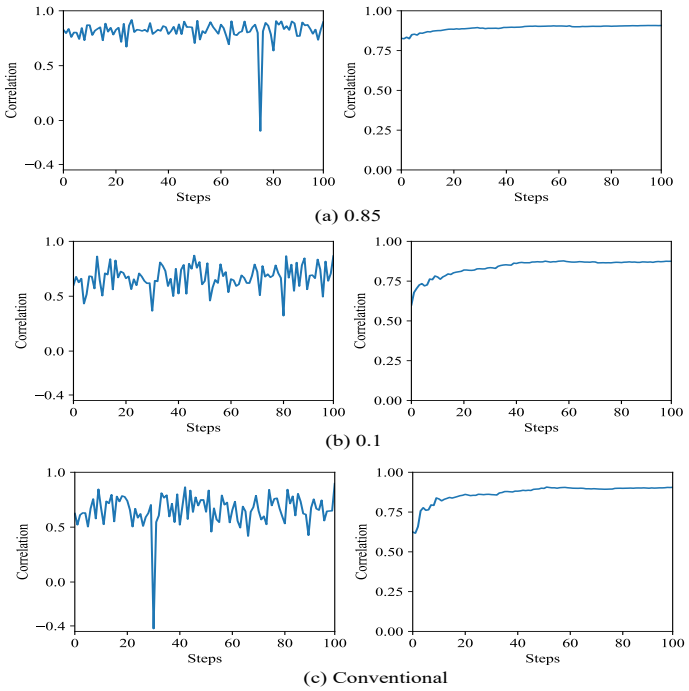

Fig. 7. Correlation coefficients between weights and original correlation coefficients by the syntagmatic compression (left) and by the paradigmatic compression (right), when the parameter β_1 was 0.85 (a), α was 0.1 for weight decay (b), and by the conventional method without weight decay (c) for the traffic data set.

the contrary, the weight decay and conventional method could not increase the ratio of information to its cost.

Figure 8 shows selective information (left), cost (middle), and the ratio of information to its cost (right). When the parameter for the cost reduction was 0.85 in Figure 8(a), the information first increased gradually, and then it decreased. On the other hand, the cost decreased gradually, and remained almost a constant in the later stages of learning. Naturally, the ratio of information to its cost increased and then decreased slightly in the end. When the weight decay was used and the parameter α was set to 0.1 in Figure 8(b), the information constantly increased, and the cost gradually decreased, though it did not decrease to the lower point attained by the present method. The ratio was much lower than that by the present method in Figure 8(right, b). Finally, when we used the conventional method without the weight decay in Figure 8(c), the information did not change, the cost remained higher, and finally, the ratio remained lower.

The results confirmed that the cost reduction could inhibit the generation of supposedly important connection weights. On the contrary, the weight decay constantly increased the selectivity of connection weights.

4) Weights: The results showed that the present method could produce weights where a small number of them became stronger, and we could also see that some groups of connection weights were identified. On the contrary, the weight decay and conventional method could not produce a similar result.

Figure 9(a) shows connection weights (1) and their individual potentiality (2). As can be seen in the figure, a

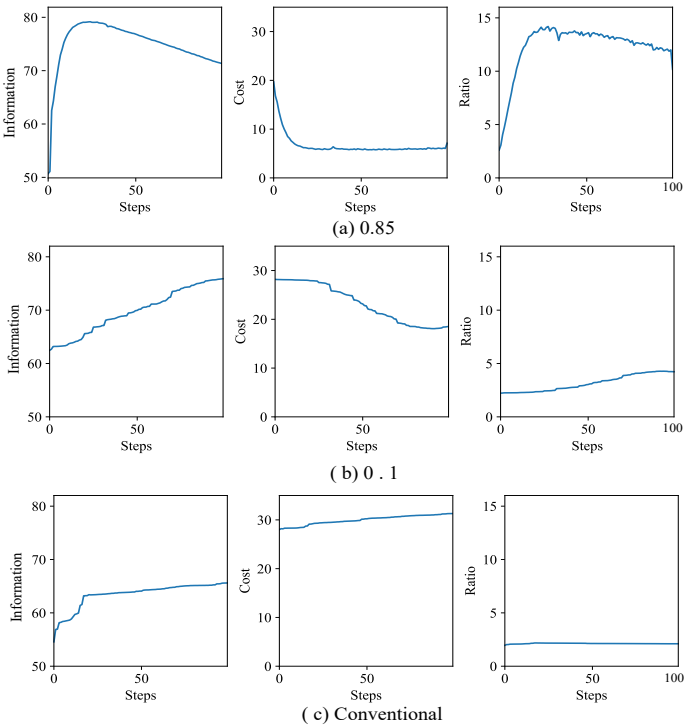

Fig. 8. Information (left), cost (middle), and ratio (right) when the parameter β_1 was 0.85 (a), α was 0.1 (b), and by the conventional method (c) for the traffic data set

small number of connection weights became stronger, and they responded to inputs in the precedent layers with clear regularity. This tendency was further enhanced over individual potentialities in Figure 9(2). Figure 9(b) shows weights and individual potentiality by the weight decay ($\alpha = 0.1$). Though we could not see any strong weights, a small number of weights could be seen by using the individual potentiality. Finally, when the conventional method was used in Figure 9(c), weights seemed to become randomly activated, though we could see a smaller number of individual potentiality.

5) Partial Compression: The results show that the present method tried to extract information from inputs, while the weight decay and conventional method tried to extract information from outputs.

Figure 10(a) shows partially compressed weights when the parameter β_1 was 0.85. As can be seen in the figure, only the initial partially compressed weights had higher connection weights, and the strength of weights remained small. This means that at the beginning the present method tried to acquire information on inputs and that it seemed to try to extract information on inputs as much as possible. Figures 10(b) and (c) show partially compressed weights by the weight decay and by the conventional method. Clear compressed weights could not be seen until the final compression was performed. This means that information from outputs played a critical role in creating the final connection weights.

6) Full Compression: The results show that the present method could produce compressed weights whose correlations with the original correlations between inputs and targets were

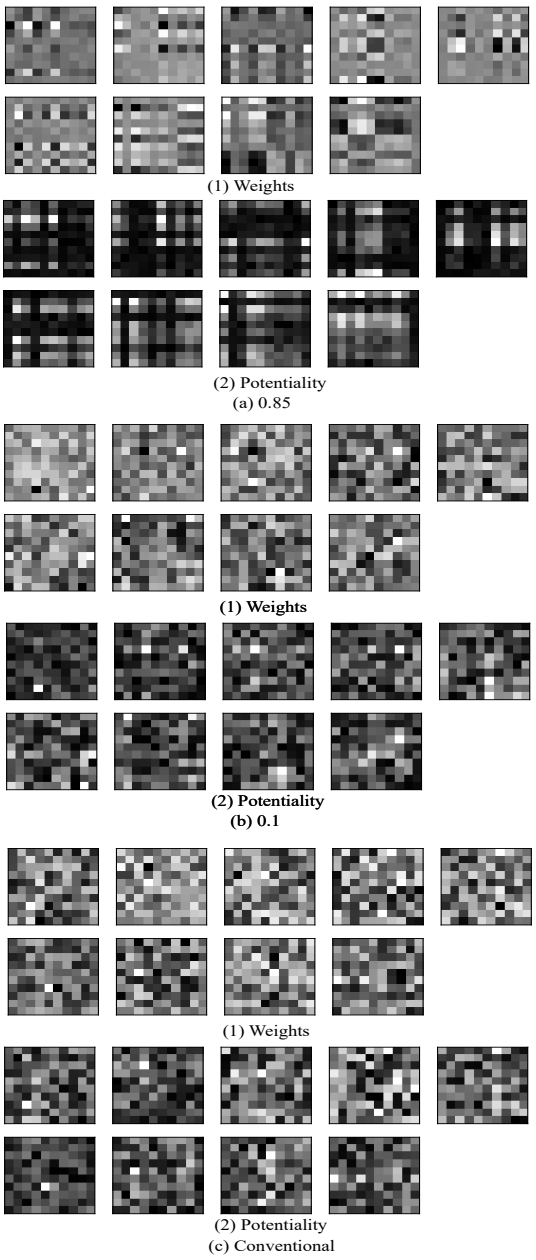

Fig. 9. Weights (1) and potentiality (2) when the parameter β_1 was 0.85 (a), α was 0.1 (b), and by the conventional method (c) for the traffic data set.

high and close to those by the logistic regression. In addition, the present method produced higher generalization accuracy.

Figure 11(1) shows the correlation coefficients between inputs and targets, and we could see that the first input (hour) played the most important role in traffic behavior. Figure 11(2) shows fully compressed weights by the paradigmatic compression when the parameter β_1 was 0.85. As can be seen in the figure, the correlation was 0.908, the second largest one behind the logistic regression analysis, and generalization accuracy was the highest at 0.812. When the weights decay was introduced, the correlation decreased to 0.875, and the accuracy also decreased to 0.808 in Figure 11(3). When the conventional method was used in Figure 11(4), the correlation

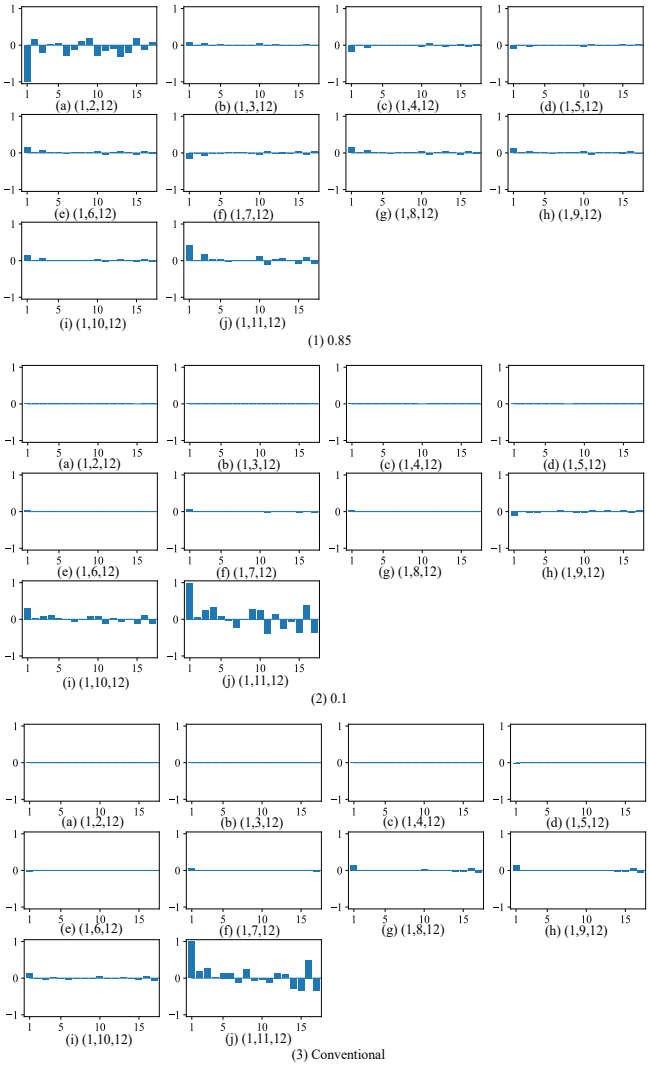

Fig. 10. Partially compressed weights when the parameter β_1 was 0.85 (1), α was 0.1 for weight decay (2), and by the conventional method (3) for the traffic data set.

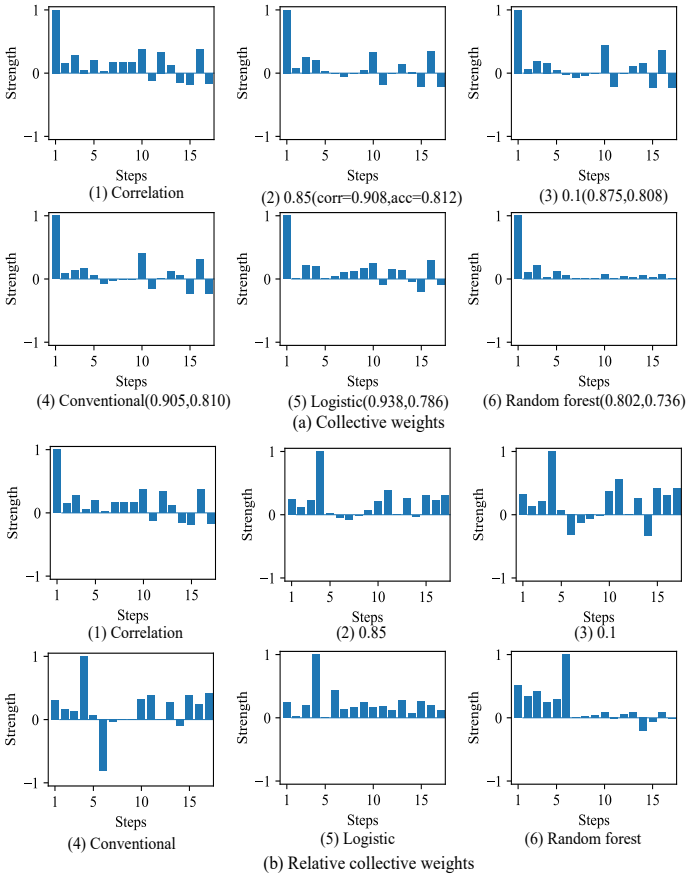

Fig. 11. Collective weights and related importance measures (a) and relative collective ones (b) for the present method (1) to the random forest (6) for the traffic data set. The numbers in the figure show the correlation coefficients (left) and generalization accuracy (right).

and accuracy were slightly larger than those by the weight decay. The logistic regression analysis in Figure 11(5) produced the largest correlation of 0.938, but the accuracy was lower, with the second worst value of 0.786, slightly better than the 0.736 of the random forest. Finally, when the random forest was used, the correlation and accuracy were the lowest in Figure 11(6).

When we used the relative correlation coefficients relative to the absolute original correlations between inputs and targets in Figure 11(b1)-(b5), the fourth input (vehicle excess) showed higher values for the cost reduction, weight decay, conventional method, and logistic regression analysis. This suggests that, in addition to the first input, the fourth input could play an important role in traffic behavior.

B. Facility for the Elderly Data Set

1) Experimental Outline: The second experiment used the data set of the facility for the elderly [64], in which we tried to distinguish between male and female residents and to identify the essential needs of residents for the facility. The objective of the experiment aimed to improve the services provided by the facility. The number of input variables was seven, and the number of patterns was 1,000. We used the same parameter values presented in the first experimental results on the traffic data set for easy reproduction of the results.

2) Syntagmatic and Paradigmatic Compression : The results show that the present method produced very high correlation coefficients, with almost perfect correlations with the original correlation coefficients between inputs and targets. The weight decay and conventional method could produce weights with higher correlations, but they were lower than those by the present method.

Figure 12(a) shows the syntagmatic (left) and paradigmatic (right) compression when 100 different initial conditions and 100 different subsets of data were used, where the parameter β_1 was set to 0.85 for cost reduction. As can be seen in the left-hand box on the syntagmatic compression, except for five low correlations between compressed weights and original correlations, the correlations became close to one. For the paradigmatic compression on the right-hand side, the correlations became immediately close to one, meaning that paradigmatic compression produced original correlations

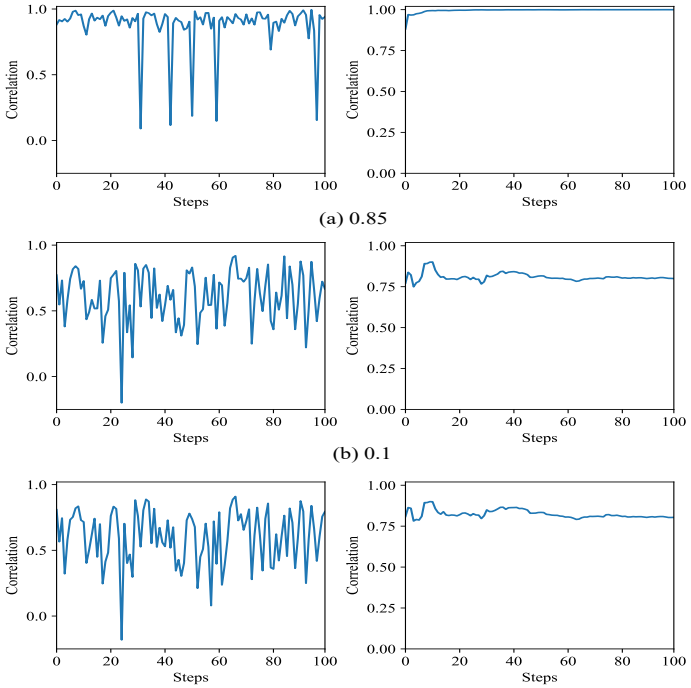

Fig. 12. Correlations by the syntagmatic compression (left) and by the paradigmatic compression (right) when the parameter β_1 was 0.85 (a), α was 0.1 for the weight decay (b), and by conventional method (c) for the facility for the elderly data set.

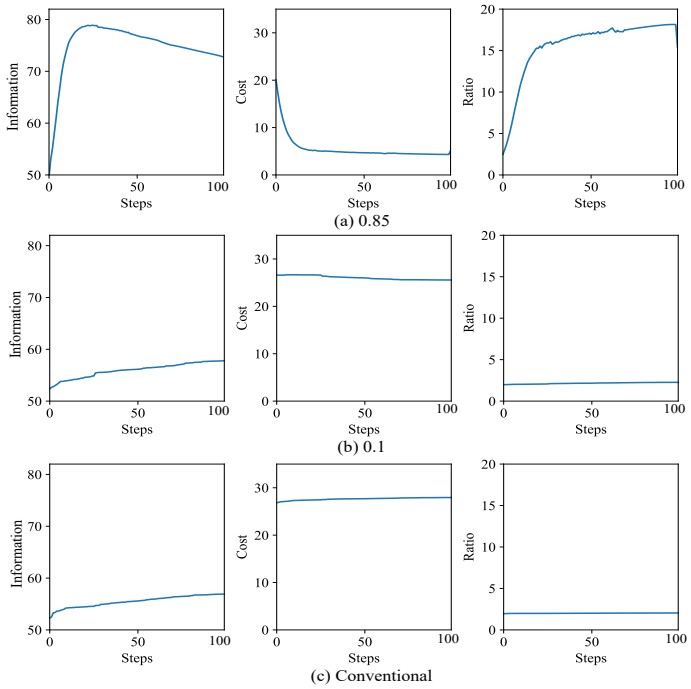

Fig. 13. Selective information (left), cost (middle), and ratio (right) when the parameter β_1 was 0.85 (a), α was 0.1 for weight decay (b), and by the conventional method (c) for the facility for the elderly data set

between inputs and targets over compressed weights. Figures 12(b) and (c) show syntagmatic (left) and paradigmatic compression (right) by the weight decay ($\alpha = 0.1$) and by the conventional method without weight decay. The two methods produced quite similar results for both types of compression. However, the correlations were lower than those by the present method. In particular, correlations with syntagmatically compressed weights fluctuated extensively.

These results showed that the present method could produce collective weights close to the original correlation coefficients between inputs and targets. We could obtain those results almost independently of different initial conditions and different inputs. On the contrary, the conventional methods produced lower correlations, and they fluctuated considerably.

3) Selective Information, Cost, and Ratio: The results show that the selective information increased up to a certain point, and then it decreased in the end. However, due to the smaller cost, the ratio increased gradually for all the learning steps. On the contrary, the weight decay and conventional method could not sufficiently increase selective information, and in addition, they could not decrease the cost. Then, ratios became smaller almost over all different runs.

Figure 13(a) shows selective information (left), cost (middle), and ratio of information to its cost (right) when the parameter β_1 was 0.85. The selective information increased, and then decreased gradually. Because information was not forced to be increased, the information could not naturally continue to be sufficiently increased. However, the cost constantly decreased when the number of learning steps increased. Then, the ratio of information to its cost increased rapidly. On the contrary, Figure 13(b) and (c) show the results by the weight decay and conventional method. The selective information slightly increased, but the cost remained large, and the ratios remained small for all the learning steps. The results confirmed that the present method could decrease the cost sufficiently to increase the selective information. Then, the ratio of information and cost increased gradually.

C. Weights and Individual Potentiality

The results show that the number of strong weights became smaller when the hidden layers became higher. On the contrary, the weight decay and conventional method could not produce explicit regularity over connection weights.

Figure 14(a) shows weights (1) and corresponding individual potentiality (2) when the parameter β_1 was 0.85. As can be seen in the figure, the number of strong connection weights gradually decreased when the hidden layers became higher. In addition, for the individual potentiality, we could see several groups of connection weights responding to the inputs in the same way. On the contrary, by the weight decay (b) and conventional method without weight decay (c), no regularity over connection weights and individual potentiality could be seen.

The results showed that the present method could decrease the number of strong connection weights, and connection weights cooperated with each other as several groups to transmit the information.

1) Partial Compression: The results show that the present method could extract information on inputs in the lower hidden

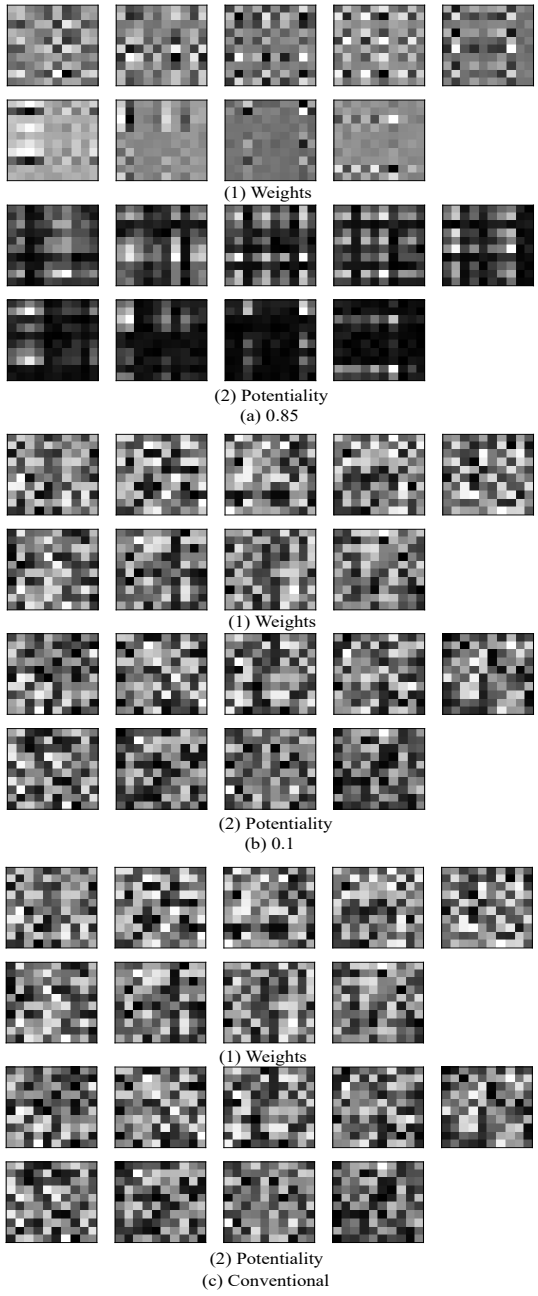

Fig. 14. Weights (a) and individual potentiality (b), when the parameter β_1 was 0.85 (a), α was 0.1 for the weight decay (b), and by the conventional method (c) for the facility for the elderly data set.

layers. On the contrary, the weight decay and conventional method could not extract the information in the hidden layers.

Figure 15 shows partially compressed weights by the present method (a), weight decay (b), and conventional method (c). The present method in Figure 15(a) produced strong partially compressed weights in the beginning, and the strength of compressed weights became smaller when the layers became higher. On the contrary, by the weight decay in Figure 15(b) and conventional method in Figure 15(c), the strength of partially compressed weights remained small until the final compression was applied.

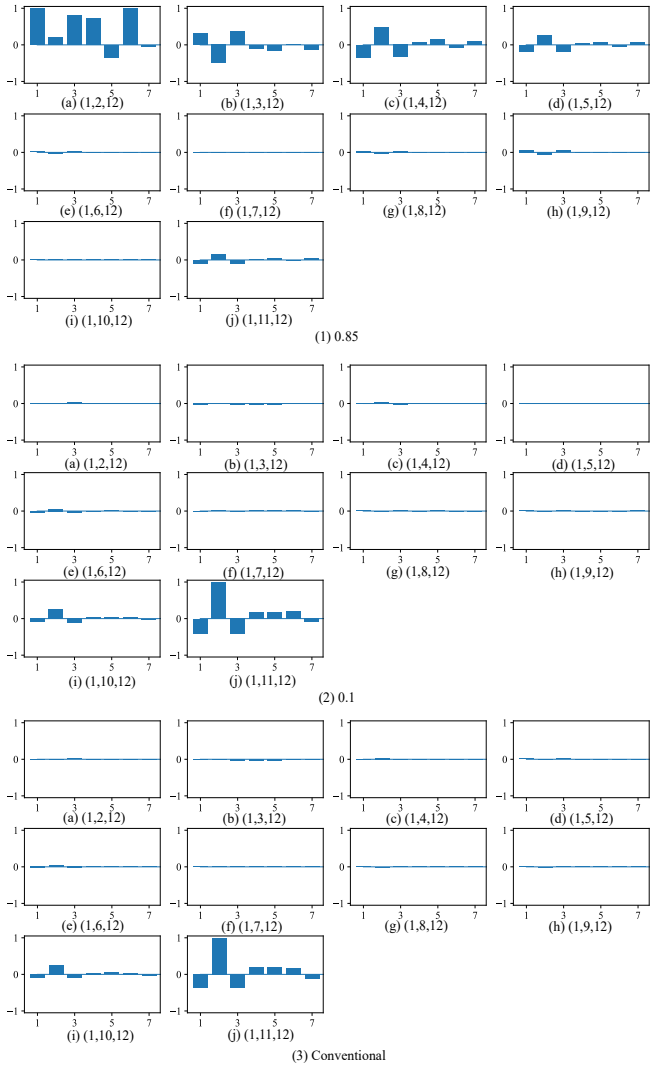

Fig. 15. Partially compressed weights, when the parameter β_1 was 0.85 (1), α was 0.1 for the weight decay (2), and by the conventional method (3) for the facility for the elderly data set.

These results show that the present method tried to acquire information content from inputs, and this information gradually decreased when going through many layers. On the contrary, the other conventional methods could not acquire enough information until we reached the final layer.

2) Full Compression: The results show that the present method could extract almost perfect correlations with higher generalization accuracy, compared with the weight decay and conventional method. The correlation coefficient was still higher than that obtained by the logistic regression.

Figure 16(a) shows correlation coefficients between inputs and targets of the original data set (1); collective weights by the present method, with the highest correlation coefficient (2), weight decay (3), and conventional method (4); regression coefficients by the logistic regression analysis (5); and prediction importance by the random forest (6). As can be seen in Figure 16(a2), the correlation was rounded to one (perfect correlation), and the generalization accuracy of 0.566 was the

second best, behind the 0.568 by the weight decay. Figure 16(a3) shows a case with the best generalization accuracy of 0.568 by the weight decay. The correlation coefficient decreased to 0.8. The conventional method in Figure 16(a4) produced the correlation of 0.804, and the accuracy was 0.566. The logistic regression in Figure 16(a5) produced a high correlation of 0.985, but the accuracy decreased to 0.551. Finally, the random forest produced the worst accuracy of 0.541 and the worst correlation of -0.294.

Figure 16(b) shows the relative collective weights. As shown in Figure 16(b2), the present method with the best correlation coefficient produced an almost even score over all inputs. On the contrary, the weight decay and conventional method in Figure 16(b3) and (b4) produced negative values for the latter three inputs. The logistic regression analysis in Figure 16(b5) produced evenly distributed and positive relative weights, but the strength varied considerably. Finally, the prediction importance in Figure 16(b6) by the random forest produced importance values completely different from other measures.

The results show that the present method with many hidden layers could produce connection weights close to the original correlation coefficients, keeping generalization sufficiently good. These results demonstrate that multi-layered neural networks could be transformed to identify individual correlation coefficients, and if differences between them and their original correlations were considerably large, neural networks tried to use non-linear and complicated connection weights.

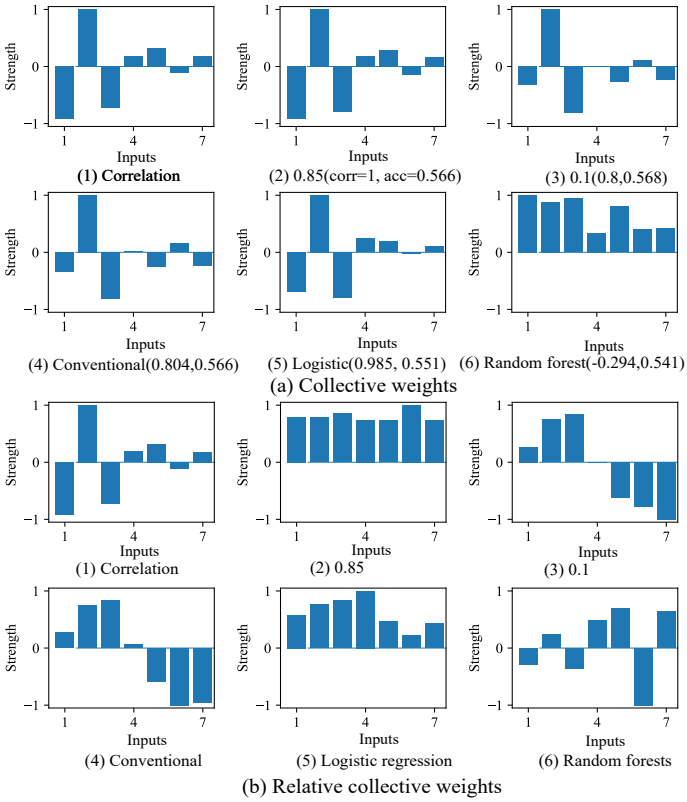

Fig. 16. Collective weights (a) and relative collective weights (b) for the facility for the elderly data set. Figures 1 to 6 denote the original correlation, compressed weights by the present method with best correlation, weight decay, conventional method, logistic regression, and random forest method.

D. Wine Data Set

1) Experimental Outline: The data set was composed of red and white wine samples from the north of Portugal, where we tried to distinguish between red and white ones based on 12 variables [65]. The number of samples was 6,497. Because the resultant correlation coefficients were lower than those in the above sections by the simple cost reduction, we tried to use the two-steps selectivity or cost control method. All the parameters used in this experiment were forced to be set to the same values as those in the above two experiments, except for the parameter β for the initial learning stage. The parameter β_2 was larger than one, actually, 1.3, in the beginning of learning (until one third of the total learning steps was reached). Then, the parameter was reduced to the normal 0.85. Thus, this method lay in cost augmentation in the first place, and then the cost was reduced.

2) Correlation Coefficients : The correlation coefficients between compressed weights and the original correlations computed by the data set were relatively high for all methods. However, the present method could produce higher correlations for all different runs.

Figure 17 shows correlation coefficients between compressed weights and the original correlations when the parameter β_2 was 1.3 (first part) and when β_1 was 0.85 (remaining part) (a), when the decay parameter α was 0.1 (b), and by the conventional method without the weight decay and selectivity control (c). As shown in the left-hand box in Figure 17(a), the correlation coefficients by the present method fluctuated in the processes of syntagmatic compression. However, in the processes of paradigmatic processing in the right-hand box in Figure 17(a), the correlation coefficients were very stable and close to those from the beginning. On the contrary, the correlation coefficients by the weight decay in Figure 17(b) and by the conventional method in Figure 17(c) tended to decrease gradually when the number of different runs increased. In addition, the correlation coefficients by the syntagmatic and paradigmatic compression were smaller than those by the present method.

The results show that the simplified two-step method could produce higher correlation coefficients for syntagmatic and paradigmatic compression.

3) Selective Information, Cost, and Ratio: The initial steps of learning by the simplified method could increase the cost considerably, keeping the selective information smaller. Then, in the subsequent steps, the selective information increased rapidly and, at the same time, the cost decreased considerably. Finally, the ratio of selective information to its cost increased in the subsequent steps. On the contrary, the other methods could not increase the selective information and decrease the cost.

Figure 18 shows the selective information (left), cost (middle), and the ratio of the information to the cost (right) by the present method (a) and the weight decay (b), and the ratio (c). Figure 18(a) shows the results by the present method when the

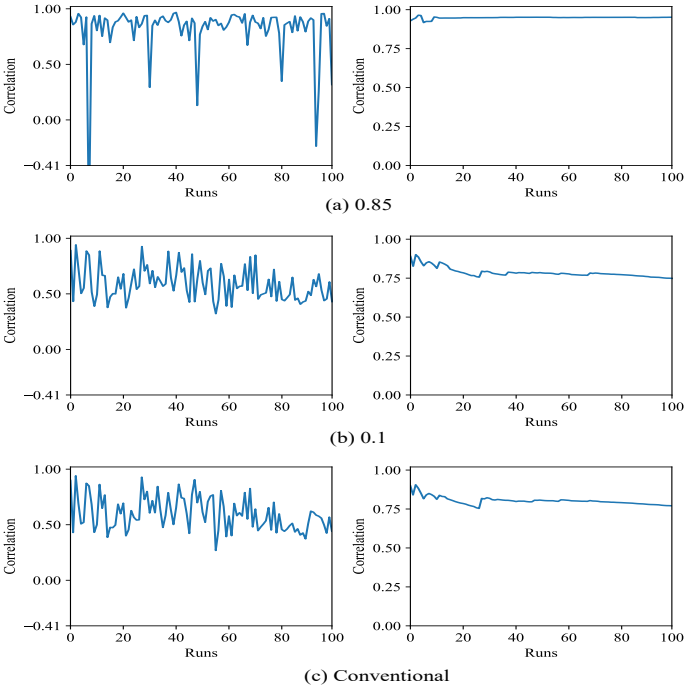

Fig. 17. Correlations between compressed weights and correlations of the original data set by the syntagmatic compression (left) and by the paradigmatic compression (right) when the parameter β_2 was 1.3 and β_1 was 0.85 (a), α was 0.1 for the weight decay (b), and by the conventional method (c) for the wine data set.

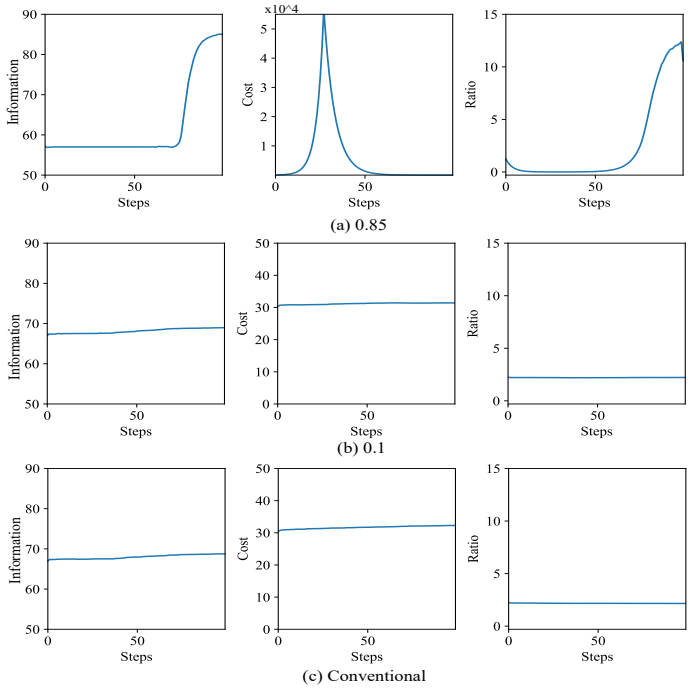

Fig. 18. Selective information (left), cost (middle), and ratio (right) when the parameter β_2 was 1.3 and β_1 was 0.85 (a), α was 0.1 for the weight decay (b), and by the conventional method (c) for the wine data set

parameter β_2 was 1.3 (initial) and β_1 was 0.85 (remaining). As can be seen in the figure, selective information was kept small in the initial steps of learning. Then, the selective information increased considerably in the remaining learning steps. The cost (middle) was forced to be increased up to a point where a further increase in the cost degraded the performance, and the cost was forced to be decreased considerably in the end. On the contrary, by using the weight decay in Figure 18(b), and the conventional method in Figure 18(c), the selective information had relatively high values without changes. The costs, shown in the figures in the middle, were larger than those by the present method. Finally, the ratio of selective information and its cost remained small for all learning steps.

The experimental results show that the present method could increase and then decrease the cost and correspondingly decrease and increase the selective information. On the other hand, the weight decay and conventional method could not well control the selective information and its cost.

4) Weights and Individual Potentiality: The weights for all hidden layers became relatively sparse by the present method, and in particular, the individual potentiality showed this sparsity tendency. However, the property of sparsity of the present method was not so different from that by the conventional methods.

Figure 19(a) shows connection weights (1) and the corresponding individual potentialities (2) by the present method. As can be seen in the figure, in particular, by seeing the individual potentialities, the number of stronger weights tended to be smaller, and weights became more selective by the present method. In the same way, by using the weights decay in Figure 19(b) and conventional method in Figure 19(c), the number of stronger weights seemed to be smaller. In particular, when we examined the individual potentialities, the sparse properties could be seen. However, we could not see large differences among the three methods. The results show that the final weights by the three methods seemed to be approximately the same in terms of their sparseness, though the present method could produce slightly more selective weights. This is due to the large parameter value β for the present method, and this large parameter value, accompanied by the large cost, prevented the present method from producing more selective states.

5) Partial Compression: The partially compressed weights produced a similar tendency for all three methods. The compressed weights by all the methods could not show explicit characteristics until the final and output layer was considered.

Figures 20(a), (b) and (c) show partially compressed weights by the present method, the weight decay, and the conventional method, respectively. Though the final compressed weights were different, all partially compressed weights were kept small. Only in the final compression step did compressed weights tend to be reasonably large. This can be explained by the fact that the selective information was forced to be smaller by increasing the cost. Then, selective information on inputs tended to disappear by the present method. This means that the information content in inputs could not be used to relate inputs and outputs.

6) Full Compression: The results by the full compression show that the present method could produce collective weights

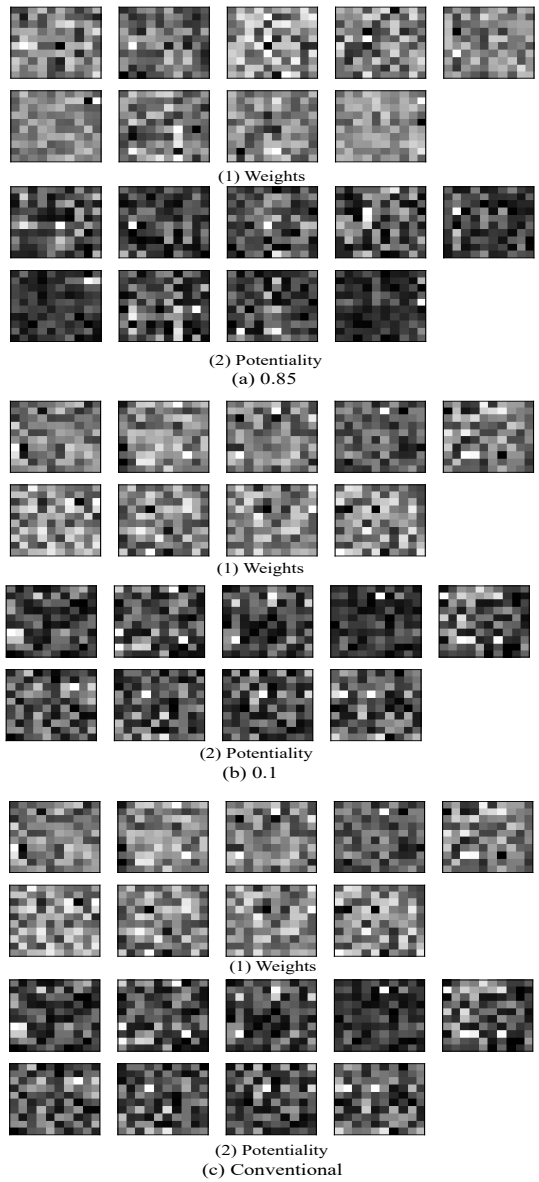

Fig. 19. Weights (a) and individual potentiality (b), when the parameter β_2 was 1.3 and β_1 was 0.85 (a), α was 0.1 for the weight decay (b), and by the conventional method (c) for the wine data set.

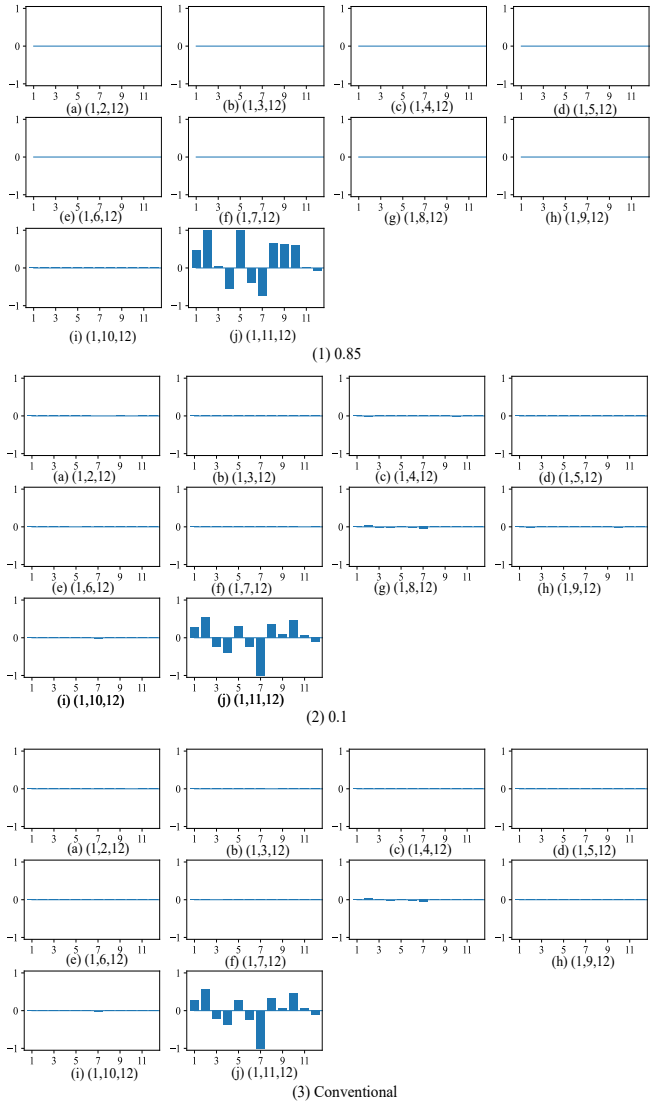

Fig. 20. Partially compressed weights when the parameter β_2 was 1.3 and β_1 was 0.85 (1), α was 0.1 for the weight decay (2), and by the conventional method (3) for the wine data set.

close to the original correlation coefficients. Though the conventional logistic regression analysis could produce similar correlation coefficients, its accuracy rate was smaller than that by the present method.

Figure 21 shows the correlation coefficients and the fully compressed weights by five methods. By using the present method, the correlation coefficient became 0.952, and the accuracy was 0.952 in Figure 21(a2). In addition, the similarity between the original correlation and compressed weights was observed in the positive relative weights for all inputs in Figure 21(b2). By using the weight decay, the correlation decreased to 0.749, and the accuracy rate was the highest one of 0.996 in Figure 21(a3). The conventional method could also produce the highest accuracy of 0.996, but the correlation coefficient decreased to 0.771 in Figure 21(a4). Though those methods produced lower correlation coefficients than those by the present method, the relative collective weights were positive except for input No.11 in Figure 21(b3) and (b4). Then, by the logistic regression analysis in Figure 21(a5), the correlation was 0.937, which was lower than the 0.952 by the present method. In addition, the accuracy by the present method was 0.995, larger than the 0.989 by the logistic regression analysis. Finally, the random forest in Figure 21(a6) produced the lowest correlation of -0.075, though the accuracy was the highest at 0.996. The random forest produced importance measures quite different from those by the other methods. The results show that the present method could produce the highest correlation coefficient, keeping high accuracy rates.

IV. Conclusion

The present paper aimed to propose a new type of interpretation method for multi-layered neural networks. The method

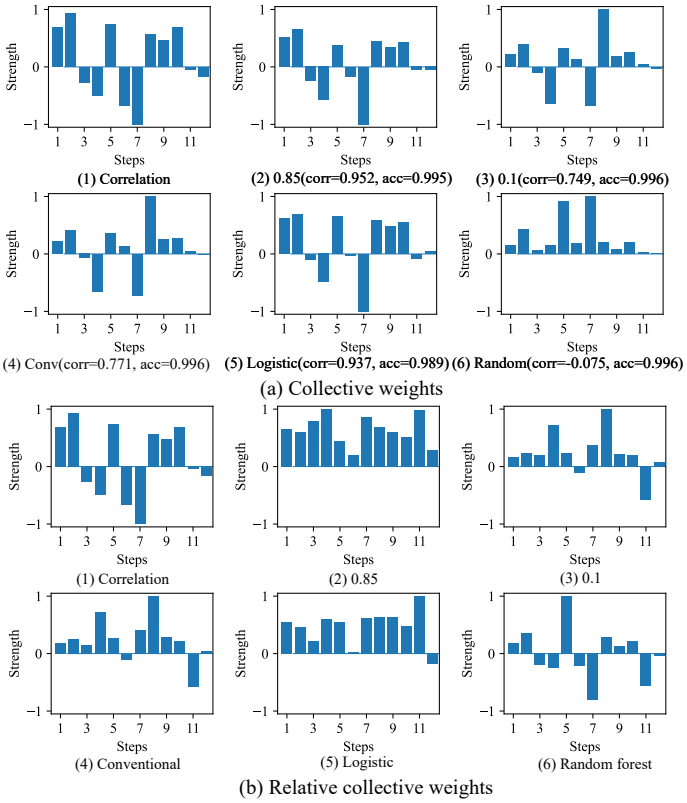

Fig. 21. Collective weights (a) and relative collective weights (b) by five methods for the wine data set. The numbers in the figure represent the correlation coefficients (left) and accuracy (right).

lies in considering all possible internal representations generated by multi-layered neural networks, in which we suppose that all representations by multi-layered neural network have the same status, meaning that many representations should be created by seeing a data set from different points of view.

One of the main shortcomings of interpretation methods of neural networks is that they try to understand only one aspect of representations. For example, they tried to show what components in a neural network can be responsible for a specific input. This type of individual interpretation has been extensively used in the present state of neural networks. In particular, in the CNN, dealing with image data sets, it has been extensively used, because it is easy to understand the specific input and the corresponding component intuitively. However, those corresponding components should be changed, sometimes drastically, by using different initial conditions, which is one of the main problems of neural networks. In our approach, we suppose that different representations created by different initial conditions can be used to explain the inference mechanism of neural networks. This means that we can interpret the representations from different points of view.

In actual learning, we have different internal representations in the course of learning. In addition, by different initial conditions and inputs, we have also different internal representations. We first take into account different representations in the course of learning, which can be called "syntagmatic compression." In the syntagmatic compression, all weights created in the course of learning with a specific initial condition are averaged and compressed. Then, all syntagmatically compressed weights are again averaged, which is called "paradigmatic compression." With syntagmatically and paradigmatically compressed representations, we can interpret neural networks in terms of collective interpretation, namely, from many viewpoints.

The collective compression was flexibly controlled by controlling the selective information. However, we proposed a more simplified method to control the selective information, that of controlling the cost in terms of weight strength. This means that the selective information control eventually corresponds to the cost control, which is much simpler to be implemented in actual learning. In addition, we proposed a new method to control the selective information by its cost, where the cost is first increased, and then it is decreased. This increase in the cost, corresponding to a decrease in selective information aimed to eliminate information on input patterns as much as possible. This is because the information represented by the inputs of neural networks cannot be used to relate the inputs to the corresponding targets.

The method was applied to three real data sets: the traffic, facility for the elderly, and wine data sets. In the first two cases, we could see that the selective information could be increased and, at the same time, the cost could be decreased in terms of the sum of absolute weights. The final collective weights by the present method were very close to the correlation coefficients between inputs and targets of the original data set. This could be explained by the fact that the present method could extract much information from inputs; on the contrary, the other conventional methods could not extract sufficient information from the inputs, but they were dependent exclusively on the outputs. In the experimental results of the third data set, the wine data set, the original information by the corresponding inputs was forced to be eliminated by the cost augmentation in the initial learning steps. This means that the input variables cannot represent well the information on the relations between inputs and outputs. This could be observed in the results of partial compression, where partially compressed weights could not show any regularity until the final output layer was included.

We should point out here two problems with the present method: extraction of features specific to input patterns, and how to control selective information. One of the main problems is how to identify differences between the original correlations and specific ones by the present method. We proposed a method to extract relative differences between them. However, we should develop a more refined method to distinguish between the original and new features dealt with by the present method. Second, we proposed a method to eliminate the selective information in the initial learning steps and applied it to the third data set. However, we did not know to what level we should eliminate the selective information. Thus, we need to examine more closely the exact effect of information reduction over information augmentation.

Finally, we should mention briefly some future work to be done on robustness and its relation to the selective information. First, while we focused on the interpretation in this paper, the collective concept described in this paper can be naturally applied to generalization accuracy. This is because the collective interpretation tries to interpret the inference mechanism, considering as many different internal representations as possible, including ones with higher and lower robustness. Our objective is to find some transformation rules from the collective and core ones to more concrete networks with different types of robustness [66], [67], [68]. We think that, for these transformation rules, the selective information control presented here can be of some use.

Though several problems should be solved for the present method to be applied to more practical data sets, the present study surely contributes to the problem of interpretation as well as the relations between selectivity and network performance.

V. Acknowledgments

We would like to thank the reviewers and editors for taking their time to read the drafts of the paper and give valuable comments on how to improve it. In addition, special thanks go to Mitali Das for reading and correcting the paper. Finally, this paper has been written, based on two papers: "Controlling Individual and Collective Information for Generating Interpretable Models of Multi-Layered Neural Networks" [2], presented in INTELLI2021 and "Selective Information-Driven Learning for Producing Interpretable Internal Representations in Multi-Layered Neural Networks" [1] in COGNITIVE2021.

References

[1] R. Kamimura, "Selective information-driven learning for producing interpretable internal representations in multi-layered neural networks," in *COGNITIVE 2021, The Thirteenth International Conference on Advanced Cognitive Technologies and Applications*, pp. 20–27, IARIA, 2021.

[2] R. Kamimura, "Controlling individual and collective information for generating interpretable models of multi-layered neural networks," in *INTELLI 2021, The Tenth International Conference on Intelligent Systems and Applications*, pp. 27–35, IARIA, 2021.

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

[4] R. Andrews, J. Diederich, and A. B. Tickle, "Survey and critique of techniques for extracting rules from trained artificial neural networks," *Knowledge-based systems*, vol. 8, no. 6, pp. 373–389, 1995.

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

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

[7] M. Ishikawa, "Rule extraction by successive regularization," *Neural Networks*, vol. 13, no. 10, pp. 1171–1183, 2000.

[8] D. E. Rumelhart and D. Zipser, "Feature discovery by competitive learning," in *Parallel Distributed Processing* (D. E. Rumelhart and G. E. H. et al., eds.), vol. 1, pp. 151–193, Cambridge: MIT Press, 1986.

[9] D. E. Rumelhart and J. L. McClelland, "On learning the past tenses of English verbs," in *Parallel Distributed Processing* (D. E. Rumelhart, G. E. Hinton, and R. J. Williams, eds.), vol. 2, pp. 216–271, Cambrige: MIT Press, 1986.

[10] B. Goodman and S. Flaxman, "European union regulations on algorithmic decision-making and a right to explanation," *arXiv preprint arXiv:1606.08813*, 2016.

[11] J. L. Castro, C. J. Mantas, and J. M. Benítez, "Interpretation of artificial neural networks by means of fuzzy rules," *IEEE Transactions on Neural Networks*, vol. 13, no. 1, pp. 101–116, 2002.

[12] T. Q. Huynh and J. A. Reggia, "Guiding hidden layer representations for improved rule extraction from neural networks," *IEEE Transactions on Neural Networks*, vol. 22, no. 2, pp. 264–275, 2011.

[13] F. Wang and C. Rudin, "Falling rule lists," in *Artificial Intelligence and Statistics*, pp. 1013–1022, 2015.

[14] B. Letham, C. Rudin, T. H. McCormick, D. Madigan, *et al.*, "Interpretable classifiers using rules and bayesian analysis: Building a better stroke prediction model," *The Annals of Applied Statistics*, vol. 9, no. 3, pp. 1350–1371, 2015.

[15] A. Nguyen, J. Yosinski, and J. Clune, "Understanding neural networks via feature visualization: A survey," in *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, pp. 55–76, Springer, 2019.

[16] D. Erhan, Y. Bengio, A. Courville, and P. Vincent, "Visualizing higher-layer features of a deep network," *University of Montreal*, vol. 1341, 2009.

[17] A. Nguyen, J. Clune, Y. Bengio, A. Dosovitskiy, and J. Yosinski, "Plug & play generative networks: Conditional iterative generation of images in latent space," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4467–4477, 2017.

[18] A. Mahendran and A. Vedaldi, "Understanding deep image representations by inverting them," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5188–5196, 2015.

[19] A. Nguyen, A. Dosovitskiy, J. Yosinski, T. Brox, and J. Clune, "Synthesizing the preferred inputs for neurons in neural networks via deep generator networks," in *Advances in neural information processing systems*, pp. 3387–3395, 2016.

[20] A. v. d. Oord, N. Kalchbrenner, and K. Kavukcuoglu, "Pixel recurrent neural networks," *arXiv preprint arXiv:1601.06759*, 2016.

[21] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep inside convolutional networks: Visualising image classification models and saliency maps," *arXiv preprint arXiv:1312.6034*, 2013.

[22] J. Khan, J. S. Wei, M. Ringner, L. H. Saal, M. Ladanyi, F. Westermann, F. Berthold, M. Schwab, C. R. Antonescu, C. Peterson, *et al.*, "Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks," *Nature medicine*, vol. 7, no. 6, pp. 673–679, 2001.

[23] D. Baehrens, T. Schroeter, S. Harmeling, M. Kawanabe, K. Hansen, and K.-R. MÃžller, "How to explain individual classification decisions," *Journal of Machine Learning Research*, vol. 11, no. Jun, pp. 1803–1831, 2010.

[24] D. Smilkov, N. Thorat, B. Kim, F. Viégas, and M. Wattenberg, "Smoothgrad: removing noise by adding noise," *arXiv preprint arXiv:1706.03825*, 2017.

[25] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," *arXiv preprint arXiv:1703.01365*, 2017.

[26] G. Montavon, A. Binder, S. Lapuschkin, W. Samek, and K.-R. Müller, "Layer-wise relevance propagation: an overview," in *Explainable AI: interpreting, explaining and visualizing deep learning*, pp. 193–209, Springer, 2019.

[27] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, and W. Samek, "On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation," *PloS one*, vol. 10, no. 7, p. e0130140, 2015.

[28] S. Lapuschkin, A. Binder, G. Montavon, K.-R. Muller, and W. Samek, "Analyzing classifiers: Fisher vectors and deep neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2912–2920, 2016.

[29] F. Arbabzadah, G. Montavon, K.-R. Müller, and W. Samek, "Identifying individual facial expressions by deconstructing a neural network," in *German Conference on Pattern Recognition*, pp. 344–354, Springer, 2016.

[30] I. Sturm, S. Lapuschkin, W. Samek, and K.-R. Müller, "Interpretable deep neural networks for single-trial eeg classification," *Journal of neuroscience methods*, vol. 274, pp. 141–145, 2016.

[31] A. Binder, G. Montavon, S. Lapuschkin, K.-R. Müller, and W. Samek, "Layer-wise relevance propagation for neural networks with local renormalization layers," in *International Conference on Artificial Neural Networks*, pp. 63–71, Springer, 2016.

[32] M. Polanyi, *The tacit dimension*. University of Chicago press, 2009.

[33] E. T. Hall, "Beyond culture. garden city, ny: Anchor," 1976.

[34] N. Carlini and D. Wagner, "Adversarial examples are not easily detected: Bypassing ten detection methods," in *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security*, pp. 3–14, 2017.

[35] A. Ilyas, S. Santurkar, D. Tsipras, L. Engstrom, B. Tran, and A. Madry, "Adversarial examples are not bugs, they are features," *arXiv preprint arXiv:1905.02175*, 2019.

[36] A. Kurakin, I. Goodfellow, and S. Bengio, "Adversarial machine learning at scale," *arXiv preprint arXiv:1611.01236*, 2016.

[37] C. Bucilu, R. Caruana, and A. Niculescu-Mizil, "Model compression," in *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 535–541, ACM, 2006.

[38] J. Ba and R. Caruana, "Do deep nets really need to be deep?," in *Advances in neural information processing systems*, pp. 2654–2662, 2014.

[39] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.

[40] R. Adriana, B. Nicolas, K. S. Ebrahimi, C. Antoine, G. Carlo, and B. Yoshua, "Fitnets: Hints for thin deep nets," *Proc. ICLR*, 2015.

[41] P. Luo, Z. Zhu, Z. Liu, X. Wang, and X. Tang, "Face model compression by distilling knowledge from neurons," in *Thirtieth AAAI Conference on Artificial Intelligence*, 2016.

[42] J. O. Neill, "An overview of neural network compression," *arXiv preprint arXiv:2006.03669*, 2020.

[43] J. Gou, B. Yu, S. J. Maybank, and D. Tao, "Knowledge distillation: A survey," 2020.

[44] Y. Cheng, D. Wang, P. Zhou, and T. Zhang, "A survey of model compression and acceleration for deep neural networks," 2020.

[45] R. Kamimura, "Neural self-compressor: Collective interpretation by compressing multi-layered neural networks into non-layered networks," *Neurocomputing*, vol. 323, pp. 12–36, 2019.

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

[47] 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.

[48] 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.

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

[50] K. Torkkola, "Nonlinear feature transform using maximum mutual information," in *Proceedings of International Joint Conference on Neural Networks*, pp. 2756–2761, 2001.

[51] K. Torkkola, "Feature extraction by non-parametric mutual information maximization," *Journal of Machine Learning Research*, vol. 3, pp. 1415–1438, 2003.

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

[53] M. M. Van Hulle, "The formation of topographic maps that maximize the average mutual information of the output responses to noiseless input signals," *Neural Computation*, vol. 9, no. 3, pp. 595–606, 1997.

[54] J. C. Principe, D. Xu, and J. Fisher, "Information theoretic learning," *Unsupervised adaptive filtering*, vol. 1, pp. 265–319, 2000.

[55] J. C. Principe, *Information theoretic learning: Renyi's entropy and kernel perspectives*. Springer Science & Business Media, 2010.

[56] A. S. Morcos, D. G. Barrett, N. C. Rabinowitz, and M. Botvinick, "On the importance of single directions for generalization," *stat*, vol. 1050, p. 15, 2018.

[57] I. Rafegas, M. Vanrell, L. A. Alexandre, and G. Arias, "Understanding trained cnns by indexing neuron selectivity," *Pattern Recognition Letters*, vol. 136, pp. 318–325, 2020.

[58] J. Ukita, "Causal importance of low-level feature selectivity for generalization in image recognition," *Neural Networks*, vol. 125, pp. 185–193, 2020.

[59] M. L. Leavitt and A. Morcos, "Selectivity considered harmful: evaluating the causal impact of class selectivity in dnns," *arXiv preprint arXiv:2003.01262*, 2020.

[60] W. J. Johnston, S. E. Palmer, and D. J. Freedman, "Nonlinear mixed selectivity supports reliable neural computation," *PLoS computational biology*, vol. 16, no. 2, p. e1007544, 2020.

[61] M. L. Leavitt and A. S. Morcos, "On the relationship between class selectivity, dimensionality, and robustness," *arXiv preprint arXiv:2007.04440*, 2020.

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

[63] C. Affonso, R. J. Sassi, and R. P. Ferreira, "Traffic flow breakdown prediction using feature reduction through rough-neuro fuzzy networks," in *The 2011 International Joint Conference on Neural Networks*, pp. 1943–1947, IEEE, 2011.

[64] U. Kenji, *Text mining (in Japanese)*. Asakura-shoten, 2021.

[65] P. Cortez, A. Cerdeira, F. Almeida, T. Matos, and J. Reis, "Modeling wine preferences by data mining from physicochemical properties," *Decision Support Systems*, vol. 47, no. 4, pp. 547–553, 2009.

[66] S. Zheng, Y. Song, T. Leung, and I. Goodfellow, "Improving the robustness of deep neural networks via stability training," in *Proceedings of the ieee conference on computer vision and pattern recognition*, pp. 4480–4488, 2016.

[67] N. Carlini and D. Wagner, "Towards evaluating the robustness of neural networks," in *2017 ieee symposium on security and privacy (sp)*, pp. 39–57, IEEE, 2017.

[68] B. Liu, C. Malon, L. Xue, and E. Kruus, "Improving neural network robustness through neighborhood preserving layers," in *25th International Conference on Pattern Recognition Workshops, ICPR 2020*, pp. 179–195, Springer Science and Business Media Deutschland GmbH, 2021.