Principal Hidden Unit Analysis: Generation and Interpretation of Principal Networks by Minimum Entropy Method

Ryotaro Kamimura
Information Science Laboratory
Tokai University
1117 Kitakaname Hiratsuka Kanagawa 259-12, Japan
RYO@cc.u-tokai.ac.jp

Abstract

In the present paper, a principal hidden unit analysis with entropy minimization is proposed to obtain a simple or fundamental structure from original complex structures. The principal hidden unit analysis is composed of four steps. First, entropy, defined with respect to the hidden unit activity, is minimized. Second, several principal hidden units are selected, according to R-index, representing the strength of the response of hidden units to input patterns. Third, the performance of the obtained principal network is examined with respect to the error or generalization. Finally, the internal representation of the obtained principal network must appropriately be interpreted. Applied to a rule-plus-exception, a symmetry problem and an autoencoder, it was confirmed in all cases that by using entropy method, a small number of principal hidden units were selected. With these principal hidden units, principal networks were constructed, producing targets almost perfectly. The internal representation could easily be interpreted especially for simple problems.

Introduction

There have been many attempts to obtain networks with a suitable or optimal network size(e.g. Chauvin, 1989; Chung & Lee, 1992; Mozer & Smolensky, 1989). These attempts aim at the improvement of the generalization performance and the explicit interpretation of the internal representation (Jacobs & Jordan, 1992). With too many parameters, the generalization performance can not be improved and too complex structures prevent us to interpret the internal representation (Rumelhart et al., 1986) or coding strategies (Gorman & Sejnowski, 1988). If original oversized networks can automatically be reduced to a network with an appropriate or optimal size, it is much easier to interpret or examine the internal representation which networks can create.

To extract a simple structure, we have used a method of entropy minimization (Kamimura, in press). Entropy H is defined with respect to the hidden unit activity,

$$H = -\alpha \sum_{i}^{M} p_i \log p_i, \tag{1}$$

where p_i is a normalized activity of ith hidden unit and α is a parameter and the summation is only over all the hidden units (M hidden units). If this entropy is minimized, only one hidden unit is turned on, while all the other hidden units are turned off by multiple strong inhibitory connections (Kamimura, in press). On the other hand, if entropy is maximized, all the hidden units are equally activated. If entropy is sufficiently decreased, only a small number of hidden units are turned on, while all the other units are off and not used for producing outputs. Thus, this entropy function can be used to detect unnecessary hidden units to be eliminated, and to construct simple networks.

A principal hidden unit analysis proposed in this paper, takes advantage of the localization of activation by entropy minimization. The analysis can be used to extract simple structures almost automatically from original complex structures. This analysis consists of four steps. First, entropy (H) must be minimized.

By minimizing entropy, only a small number of hidden units are activated and used to produce targets. Other units are completely inhibited through the effect of entropy minimization. This means that multiple strong negative connections are cooperated to turn off unnecessary units. Second, several principal hidden units must be determined. A principal hidden unit is defined as a unit which contributes significantly to the performance of networks, for example, the production of targets. A measure, called R-index, is introduced to evaluate the effectiveness of hidden units. R-index for ith hidden unit is defined by

$$R_i = \frac{1}{K}\sum_k^K v_i^k, \tag{2}$$

where v_i^k is an activity of ith hidden unit for kth input pattern. This index represents how strongly a hidden unit responds to input patterns. If this R-index is greater than a threshold ϵ, then a hidden unit with that R-index must be selected as a principal hidden unit. Third, the performance of the principal network must be examined, for example with respect to the error rate or generalization. Finally, we must interpret the internal representation of a principal network, constructed by selected principal hidden units. Since the obtained principal network is much simpler than an original network, it is easy to interpret the internal representation of networks or coding mechanism, that is, what kind of coding strategies are adopted by networks to solve problems.

Theory and Computational Methods

Entropy Minimization Method

We have applied entropy minimization method to recurrent back-propagation (Kamimura, in press). In this section, we formulate the entropy method for standard back-propagation.

Suppose that a network is composed of three layers: input, competitive hidden and output layers. Hidden units are denoted by v_i and input terminals by ξ_j. Then, connections from inputs to hidden units are denoted by w_{ij} and connections from hidden units to output units are denoted by W_{ij}.

A hidden unit produces an output

$$v_i = f(u_i),$$

where

$$u_i = \sum_j^N w_{ij}\xi_j.$$

where ξ_i is a ith element of an input pattern and N is the number of elements in the pattern. An entropy function at competitive hidden layer is defined by

$$H = -\alpha\sum_i^M p_i \log p_i, \tag{3}$$

where

$$p_i = \frac{v_i}{\sum_r^M v_r},$$

and M is the number of competitive hidden units. Differentiating entropy function with respect to connections from input to hidden layer, we have

$$\begin{aligned} -\frac{\partial H}{\partial w_{ij}} &= -\frac{\partial H}{\partial v_i}\frac{\partial v_i}{\partial w_{ij}} \\ &= \phi_i \xi_j, \end{aligned} \tag{4}$$

where

$$\phi_i = (\log p_i + 1)\frac{\sum_r v_r - v_i}{(\sum_r v_r)^2} f'(u_i). \tag{5}$$

By using phi rule, update rules can be summarized as follows. First, for connections from competitive hidden units to output units, only delta rule must be used. Thus, weights are updated by an equation:

$$\begin{aligned} \Delta w_{ij} &= -\beta\frac{\partial E}{\partial W_{ij}} \\ &= \beta\delta_i v_j. \end{aligned} \tag{6}$$

For connections from input units to competitive hidden units, in addition to delta rule, phi rule must be incorporated as

$$\begin{aligned} \Delta w_{ij} &= -\alpha\frac{\partial H}{\partial w_{ij}} - \beta\frac{\partial E}{\partial w_{ij}} \\ &= \alpha\phi_i\xi_j + \beta\delta_i\xi_j. \end{aligned} \tag{7}$$

This update rule means that in addition to the error minimization, entropy must be minimized in the course of the learning.

R-index

To evaluate the effectiveness of hidden units, R-index is introduced. R-index for ith hidden

unit is defined by

$$R_i = \frac{1}{K} \sum_{k}^{K} v_i^k, \qquad (8)$$

where K is the number of input patterns. This index is bounded between zero and one:

$$0 < R_i < 1.$$

When the index is close to one, the hidden unit responds strongly to all the input patterns. On the other hand, if the index is close to zero, the hidden unit responds to no input patterns, and thus the hidden unit is unnecessary for a given problem.

Results and Discussion

Rule-Plus-Exception Problem

Principal hidden unit analysis was first applied to a rule-plus-exception problem (Mozer & Smolensky, 1989). In this problem, one output unit and four inputs (for example, A, B, C, D unit) are used. The output unit is turned on, if two input units are on, for example, A and B are on. In an exceptional case, if all the four input units are off, the output is also on. The task to be learned is described by a function: $AB + \overline{A}\,\overline{B}\,\overline{C}\,\overline{D}$. Fifteen patterns of total 16 input patterns can be explained by the so-called *rule*: AB. On the other hand, only one case is explained by an *exception*, that is, $\overline{A}\,\overline{B}\,\overline{C}\,\overline{D}$. Networks are expected to learn the rule, ignoring the exceptional case. We employed extremely redundant eight hidden units to evaluate the performance of the principal hidden unit analysis.

First, we examined to what extent entropy can be minimized. Figure 1 shows entropy as a function of the parameter α. Entropy is decreased as the parameter α is gradually increased. A minimum value of entropy was 0.061. Entropy was averaged over all the input patterns and was divided by the maximum entropy. Since values of entropy ranged between zero and one, this value was close to the minimum value.

To determine principal hidden units, we have introduced R-index which shows how strongly hidden units respond to input patterns. Figure 2 shows the R-index for all the hidden units, computed with standard back-propagation (white) and with entropy method

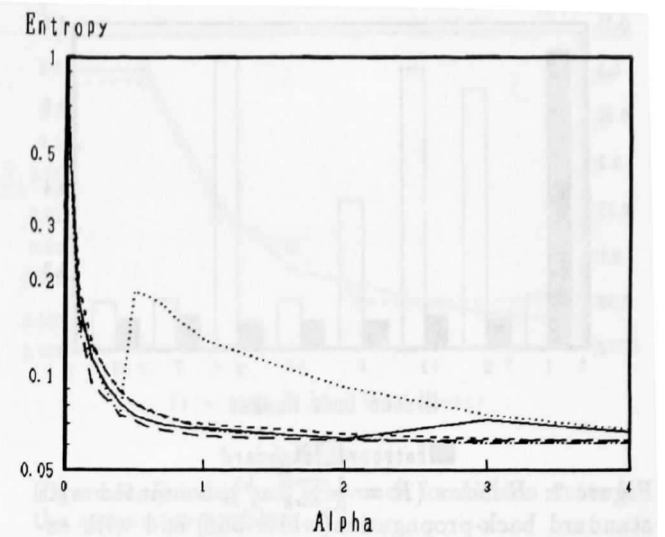

Figure 1: Entropies computed with five different initial values, ranging from -0.5 to 0.5 as a function of the parameter α for the rule-plus-exception problem. Entropy was normalized, ranging between zero and one, and the parameter was divided by the maximum entropy: $\log M$.

(black bar). Immediately, one principal hidden unit is extracted, because R-index of the first hidden unit is much higher than that of all the other hidden units, when entropy method is used. Thus, a principal network can be constructed only with this principal hidden unit. This principal network with one hidden unit can produce 15 target patterns of total 16 patterns. The strategy of this principal network was clear. That is, only when two input units were on, the activity of the hidden units could exceed the bias of the hidden unit. Otherwise, the activity could not exceed the bias, and the hidden unit was turned off.

Symmetry Problem

We applied our method to the so-called *symmetry problem*(8 bits) (Rumelhart et al., 1986). Because we have already known its suitable network size, that is, a network with two hidden units (Rumelhart et al., 1986). In addition, it has been well known that typical symmetric connections are generated as input-hidden connections.

First, we decreased entropy as much as possible. Entropy reached a lowest point of 0.068, when the parameter was 0.0005. Entropy values were normalized between zero and one.

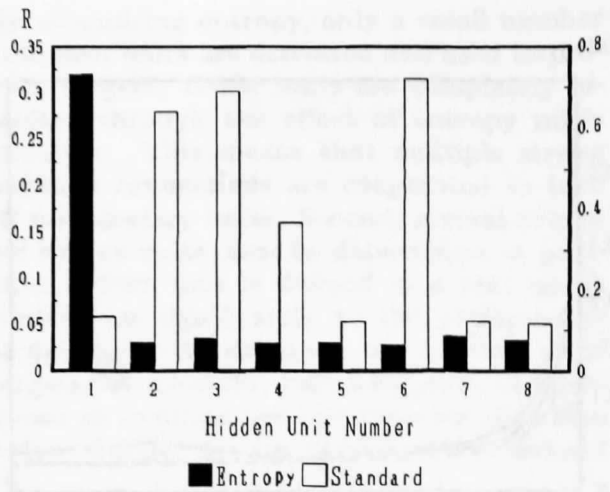

Figure 2: R-index ($R = \frac{1}{K}\sum_k^K v_i^k$), computed with standard back-propagation(white bar) and with entropy method (black bar), when the parameter α was 1.0, divided by the maximum entropy: $\log M$.

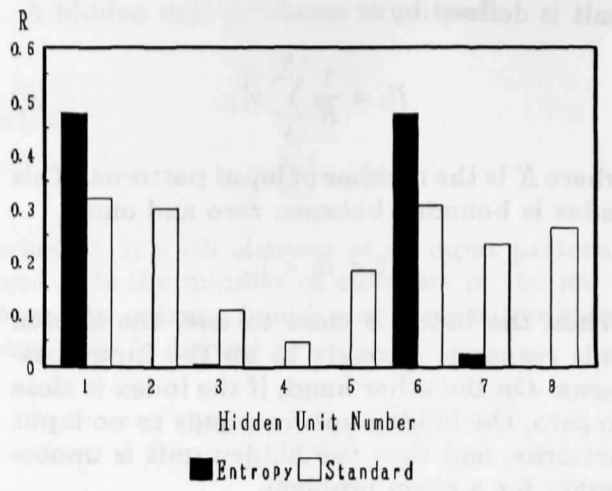

Figure 3: R-index for the symmetry problem, computed with standard back-propagation (white) and entropy method(black).

Thus, this state is significantly close to a final state of minimum entropy.

After having minimized entropy, principal hidden units must be determined, which contribute mainly to the mechanism of networks. Figure 3 shows R-indices for all the hidden units. Let us see black bars, representing R-index by entropy method (α = 0.0005). We can immediately detect two major hidden units. R-index for the first and sixth hidden units are much larger. Thus, the first principal unit is the sixth hidden units and the second principal unit is the first hidden units. In this case, a principal network can clearly be constructed by the first and the sixth hidden units. On the other hand, by using standard back-propagation, R-indices are more evenly distributed over many hidden units(white bars). These results show that by using entropy method, a small number of principal hidden units can be selected.

We have seen that by entropy method, two principal units can be detected, and a principal network is constructed with them. Let us examine the performance of the principal network. When only two hidden units, that is, two principal hidden units, were used, the error rate was completely zero, meaning that an obtained principal network can produce outputs as correctly as the original network. On the other hand, the error rate with standard back-propagation did not easily decrease. For the error rate to be zero, the network must use as many as six hidden units of total eight hidden units.

Concerning the internal representation obtained by the principal hidden unit analysis, it was observed that connections symmetric about the middle were equal in magnitude and opposite in sign, as described by Rumelhart et al. (Rumelhart et al., 1986). The solution was slightly different from the solution obtained by standard back-propagation with two hidden units. However, the main point was completely the same.

Autoencoder

We applied the method to a larger network in which 35 input, hidden and output units were employed. The network must exactly reproduce five alphabet letters: *B, C, D, E, F, G* at output units. Since the difference between these letters are small, compared with the difference between other letters, these letters are expected to be compressed into a smaller number of hidden units.

A minimum entropy was searched by changing the parameter α. Entropy decreased gradually as the parameter increased. For example, when the parameter α was 0.28, the network could reach a final minimum entropy: 0.069 for a set of initial values.

Figure 4 shows R-indices for 35 hidden units. As can be seen in the figure, only three major hidden units can immediately be

pointed out for entropy method. For example, 6th, 17th, 28th hidden unit, can perfectly be considered to be principal hidden units, because their R-indices are considerably higher than those of other hidden units. On the other hand, by using standard backpropagation, many hidden units are activated, and thus the information upon input patterns are distributed over many hidden units.

Let us examine the performance of network with principal hidden units. For entropy method, when the number of principal hidden units was increased to three units, the error became completely zero, meaning that networks can produce the original alphabet letters as perfectly as the original network only with three hidden units. However, by using the standard method, the error decreased very slowly, and could reach zero error with 16 hidden units.

Let us interpret the function of hidden units. To see clearly the meaning of hidden units, networks were constructed only with principal hidden units, and the outputs generated by the networks were carefully examined. The first principal hidden unit tended to produce B. Letters: D,F are also produced as B. were also produced as B. The second principal hidden unit was concerned with a letter C and G. The distance between these two letters is small enough to be unified into one hidden unit. A letter D was also produced as G. The third principal hidden unit could produce letters E and F, because the distance between these two letters is very small. Finally, a letter D was observed to be produced with the first and the second principal hidden unit.

R-index, Relevance and Variance

R-index, measuring the strength of the response of hidden units to input patterns, have been used to determine the effectiveness of hidden units. Relevance, proposed by Mozer and Smolensky (Mozer & Smolensky, 1989), has also been useful to evaluate the effectiveness of hidden units. The relevance ρ is defined by

$$\rho_i = E_{without\ i} - E_{with\ i},$$

where $E_{without\ i}$ means the error between target and outputs without ith hidden unit. Let us compare the relevance with our R-index. Figure 5 shows the relevance (dotted line) and

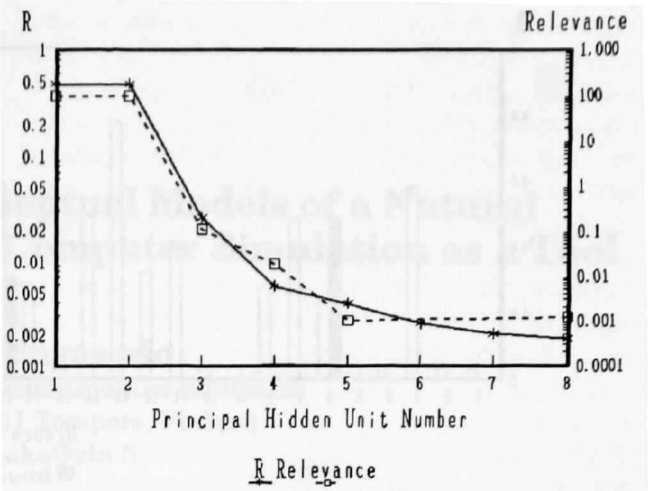

Figure 5: R-index and relevance for hidden units for the symmetry problem.

R-index (solid line) for the symmetry problem, discussed in the previous section. The relevance and R-index is clearly correlated, as shown in the figure. The correlation coefficient was 0.99 for the symmetry problem. Thus, both R-index and the relevance can be used to evaluate the utility of hidden units. However, to compute the R-index is much simpler than to compute the relevance.

The variance of input-hidden connections is also used to show the strength of the response of hidden units to input patterns. The variance (s_i^2) of ith hidden unit is defined by

$$s_i^2 = \frac{1}{M-1}\sum_j^M (w_{ij} - \overline{w_i})^2,$$

where M is the number of hidden units and $\overline{w_i}$ is an average over all the connections into ith hidden units. Figure 6 show the variance and the R-index for the symmetry problem. As shown in the figure, the variance and R-index is clearly correlated. The correlation coefficient was about 0.99. This means that principal hidden units are considered to be units with larger variance of input-hidden connections.

Conclusion

In this paper, we have proposed a method of principal hidden unit analysis by entropy minimization. By minimizing entropy, only a small number of hidden units are turned

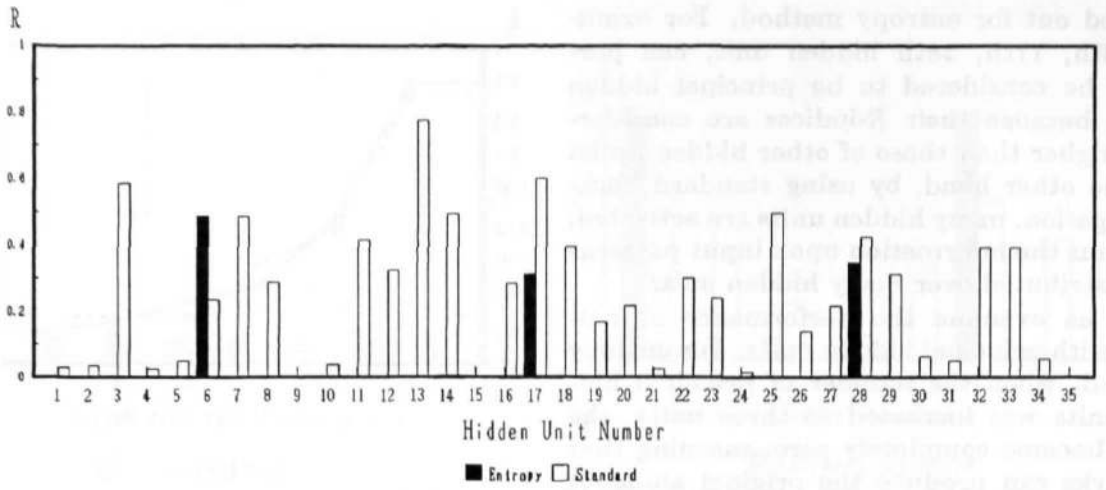

Figure 4: R-index for 35 hidden units for standard method(white) and entropy method(black).

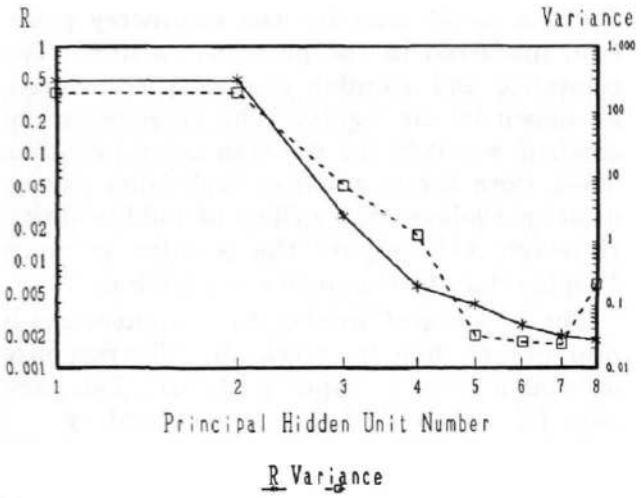

Figure 6: R-index and variance of input-hidden connections for the symmetry problem.

on, while all the other hidden units are off by strong inhibitory connections. Principal hidden units are considered to be units which responds quite strongly to input patterns. With these principal hidden units, principal networks can be constructed. These principal networks can produce outputs as correctly as original oversized networks. Thus, the internal representation networks can create is easy to interpret. In addition, without using arbitrary criteria to eliminate hidden units, it is possible to obtain an optimal size for a given problem by using principal hidden unit analysis, that is, the automatic determination of optimal size without retraining is possible. Finally, we think that our method can easily be extended to unsupervised learning to extract some features hidden in input patterns.

References

Chauvin. Y. 1989. A backpropagation algorithm with optimal use of hidden units. in D. S. Touretzky (Ed.), *Advances in Neural Information Processing Systems 1*, CA; Morgan Kaufmann; 519-526.

Chung F.L. and Lee T. 1992. A node pruning algorithm for back-propagation networks, *International Journal of Neural Systems* 3: 301-314.

Gorman .R.P. and Sejnowski T. J. 1988. Analysis of hidden units in a layered network trained to classify sonar targets, *Neural Networks* 1: 75-89.

Jacobs R.A. and Jordan M.I. 1992. Computational consequences of a bias toward short connections, *Journal of cognitive neuroscience* 4: 323-336.

Kamimura R. (in press). Minimum entropy method in neural networks, *Proceeding of 1993 IEEE International Conference on Neural Networks.*

Mozer M.C. and Smolensky P. 1989. Using relevance to reduce network size automatically, *Connection Science* 1:3-16.

Rumelhart D.E., Hinton G.E. and Williams R.J. 1986. Learning internal representation by error propagation in D. E. Rumelhart, J. L. McClelland (Ed.), *Parallel Distributed Processing* 1: 318-362. Cambridge, Massachusetts: the MIT Press.