On generalization of multilayer neural network applied to predicting protein secondary structure

著者	Nakayama Kenji, Hirano Akihiro, Fukumura K.
journal or publication title	IEEE International Conference on Neural Networks - Conference Proceedings
volume	2
page range	1209-1213
year	2004-07-01
URL	http://hdl.handle.net/2297/6794

On Generalization of Multilayer Neural Network Applied to Predicting Protein Secondary Structure

Kenji Nakayama Akihiro Hirano Ken-ichi Fukumura
Dept. of Information and Systems Eng., Faculty of Eng., Kanazawa Univ.
2-40-20, Kodatsuno, Kanazawa, 920-8667, JAPAN
e-mail:nakayama@t.kanazawa-u.ac.jp

***Abstract*— A learning process of a single neural network (SNN) to improve prediction accuracy of protein secondary structure is optimized. The protein secondary structures are predicted using a multiple alignment of amino acid as the input data. A multi-modal neural network (MNN) has been proposed to improve the precision of prediction. This method uses five independent neural networks, and the final decision is made by averaging all outputs of five SNNs. In the proposed method, the same prediction accuracy can be achieved by using only a single NN and optimizing a learning process. In a learning process of protein structure prediction, over learning is easily occurred. So, the learning process is optimized so as to avoid the over learning. For this purpose, small learning rates, adding small random noise to the input data, and updating the connection weights by the average in some group are useful. The prediction accuracy 58% obtained by using the conventional SNN is improved to 66%, which is the same accuracy of the MNN, which needs five SNNs.**

I. Introduction

Hundred thousands different kinds of proteins are synthesized in human body. They are large, complex molecules made up of long chains of subunits called amino acids that attached in a linear string. The amino acid sequence of a protein chain is called the primary structure. Different regions of the sequence form secondary structures, α-helices and β-sheets. The three dimensional structures are formed by combining the secondary structure into several domains [1]. Since a function of the protein is determined by its three-dimensional structures, it is useful to predict the secondary structure of proteins. It is also helpful to understand its role and responsibilities in the cell. Therefore, the prediction of secondary structure of protein is an important theme in genome science. Computational predictive tools have been developed. Multilayer neural networks have been applied to this field[2]-[5]. The three-dimensional structure of a protein is uniquely determined by its sequence of amino acids. These traditional method is based on a local input window of amino acids with orthogonal encoding. The output layer consists of three units, which represent the secondary structure classes for the amino acid located at the center of the window.

In the prediction, the multiple sequence alignment is used to replace the orthogonal encoding of amino acids. The evolutionary information is important to significantly improve predictions, and also the multiple sequence alignment is useful. The multi-modal neural network (MNN) has been proposed to improve the precision accuracy with a multilayer NN [6]. In MNN, several neural networks do the precision in parallel and the results are determined by ballot. The prediction accuracy is improved about 7% than that of a single neural networks, and 66% prediction accuracy is achieved. However, it requires five neural networks.

In this paper, the learning process of a single neural network (SNN) is optimized in order to improve the prediction accuracy. A main problem of training the neural networks for predicting the protein secondary structure is ’over learning’. The learning process and the parameters are optimized so as to avoid ’over learning’, and improve generalization.

II. Data Information for Prediction

The data information recorded in HSS files [7], which can be got from ftp://ftp.ebi.ac.ukk/pub/databases/hssp/. Each HSSP file includes the contents used in prediction.

Secondary structure is most often assigned based on the hydrogen bond pattern between the backbone carbonyl and NH groups [8]. By DSSP [9], eight kinds of secondary structures are distinguished. These eight classes are often grouped into three classes, that is helix (H), strand (E), and non regular structure (L). The task of neural networks is to predict what kind of state of secondary structures (H, E, L) is correspond to the amino acid sequence.

III. Neural Network for Predition

A. Multilayer Neural Network

In this paper, a single multilayer neural network (SNN) shown in Fig.1 is used.

B. Input and Output Data Assignment

Input and output data assignments are shown in Fig.2 [6]. The input layer expresses the amino acid sequence pattern. WTKC. . . indicate amino acids. The window width is 9, that is from $c-4$ to $c+4$, c is the central, the corresponding secondary structure will be predicted at the output. Profile generated from a multiple alignment is used to represent the amino acids. The alignment is obtained from HSSP files. There are 20 kinds of amino acids. Two kinds of additional information are used. Therefore, 22 input units are assigned to one amino acid, and 9 amino acids are used at the same time for prediction. Totally, $(20 + 2) \times 9 = 198$ input units are prepared.

The output layer requires three units to discriminate three kinds of the secondary structures. The targets are assigned as

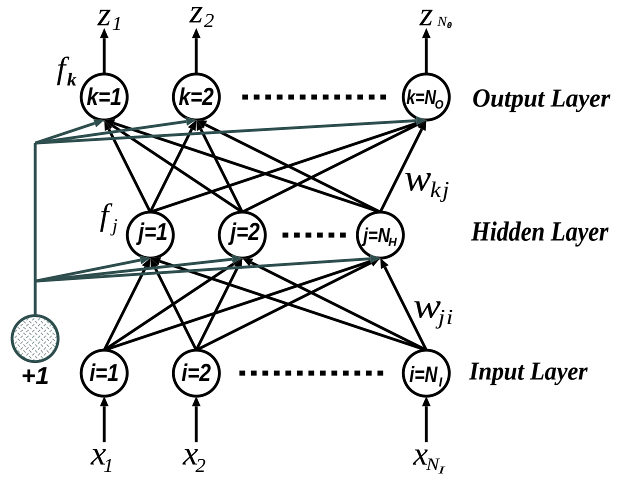

Fig. 1. Multilayer neural network used for predicting protein secondary structure.

follows: (H, E, L)=(100, 010, 001). The target represents the secondary structure of the amino acid located at the central of the sequence of 9 amino acids.

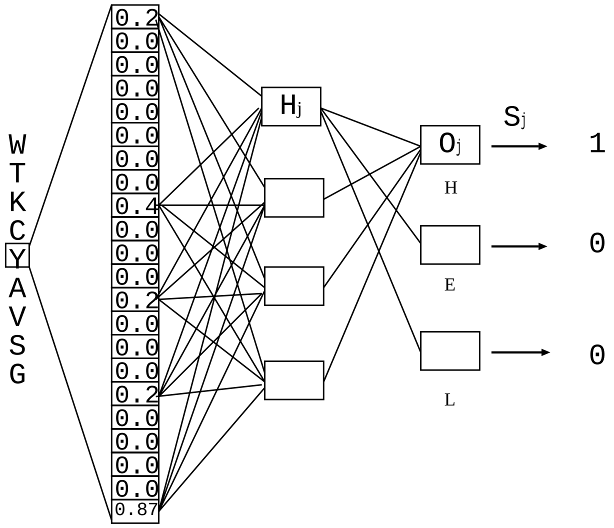

Fig. 2. Input and output data assignments for neural network.

C. Evaluation of Prediction Accuracy

The widely used score to measure the prediction accuracy is used.

$$Q_3 = 100\sum_{i=1}^{3}\frac{c_i}{N} \qquad (1)$$

$$Q_H = 100\frac{c_H}{N_H} \qquad (2)$$

$$Q_E = 100\frac{c_E}{N_E} \qquad (3)$$

$$Q_L = 100\frac{c_L}{N_L} \qquad (4)$$

c_i is the number of the residues correctly predicted in state i, that is H, E, L. N is the total number of residues in the protein. Q_3 expresses the percentage of correctly predicted residues in all of the three states, H, E, L. Q_H, Q_E and Q_L are the percentage of correctly predicted residues in each state.

IV. Optimization of Learning Process

A. Training and Test Data

35 proteins, who have the alignments, similarity is less than 25%, are obtained from the HSSP file [6]. The number of amino acids is 5964. 596 amino acids are randomly selected for test, and the rest amino acids are used for training the neural network.

B. Number of Iterations

5 hidden units, a learning rate $\eta = 0.001$ and no momentum term are used. The learning curve and the prediction accuracy are shown in Figs.3, 4 and 5. The output error in MSE is monotonously decreased. Proportionally, the prediction accuracy for the training data is increased. However, it increases to some point, after that, it is gradually decreased. This is 'over learning'. However, it is usually difficult to determine the stopping point using only the output error.

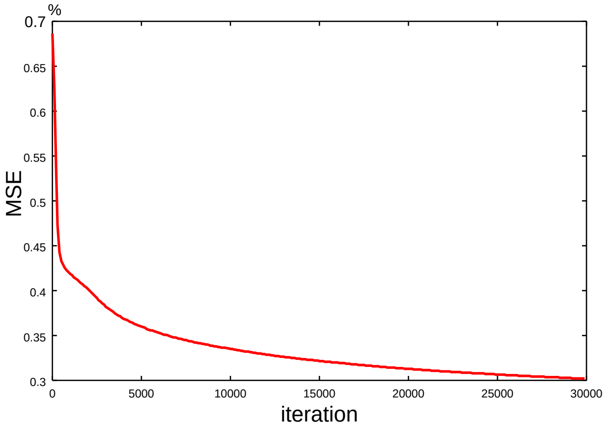

Fig. 3. Output error in MSE using 5 hidden units and learning rate $\eta = 0.001$.

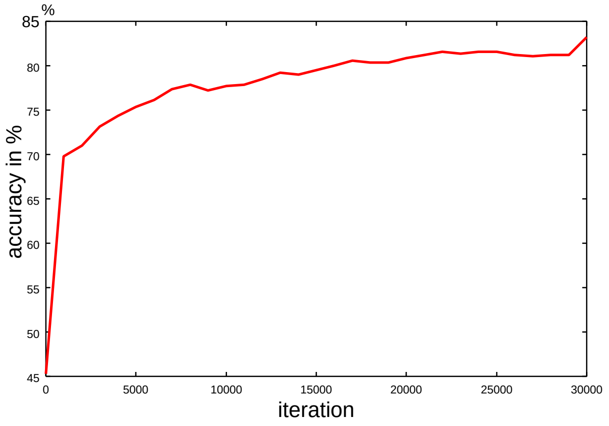

Fig. 4. Prediction accuracy in % for training data.

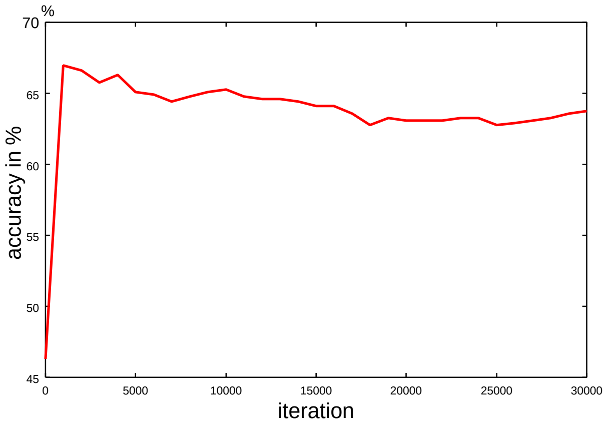

Fig. 5. Prediction accuracy in % for test data.

C. Number of Hidden Units

The number of hidden units is changed from 1 to 10. The output error in MSE and the total prediction accuracy Q_3 for the test data are shown in Figs.6 and 7. The output error is well reduced by using 10 hidden units. However, the prediction accuracy Q_3 for the test data is not good in the 10 hidden unit case. 5 or 3 hidden units are better from a generalization view point.

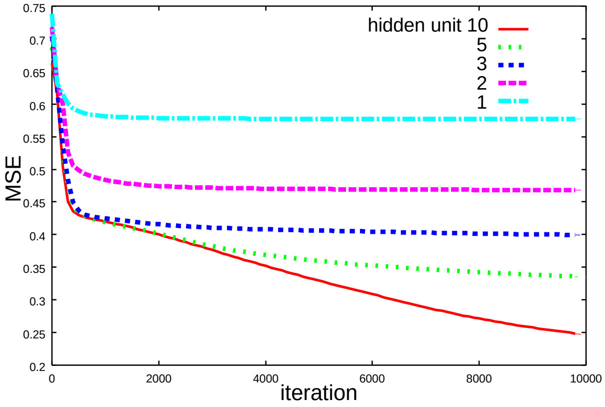

Fig. 6. Output error in MSE for different number of hidden units.

D. Momentum Term

The momentum term is taken into account. Figures8, 9 and 10 shown the output error, the prediction accuracy for the training data and the test deat, respectively. The weight of the momentum is 0.9. In this case, also 'over learning' occurs.

E. Learning Rate

In order to avoide 'over learning', the learning rate η is controlled. The output error and the total prediction accuracy Q_3 for the test data are shown in Figs.11 and 12, respectively. Using $\eta = 0.00001$, reduction of the output error is very slow, and the MSE is not well reduced. However, the prediction accuracy of $\eta = 0.0001$ is the base after 35,000 iterations.

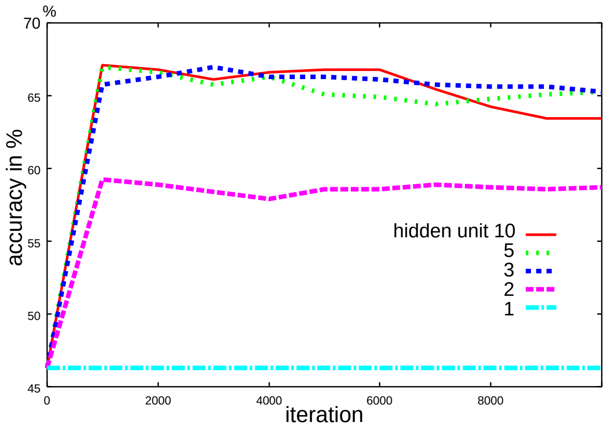

Fig. 7. Total prediction accuracy Q_3 in % for test data.

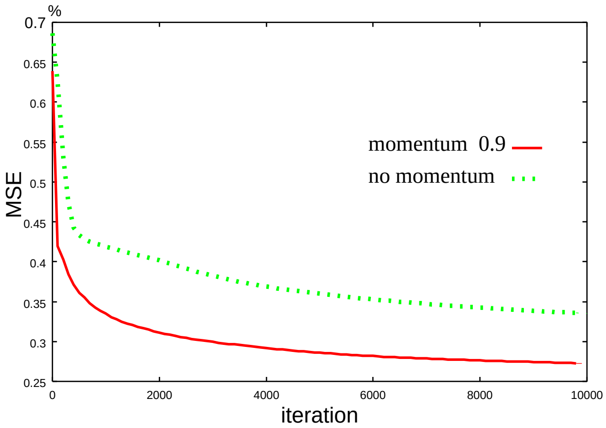

Fig. 8. Output error in MSE for different number of hidden units.

F. Adding Random Noise

In order to improve the prediction accuracy with a relatively large learning rate, that is in a short convergence time, random noises are added to the input data. Since the input data are distributed from 0 to 1, the random noise is generated during $-0.1 \sim 0.1$. Different random numbers are added in each iteration. The output error and the total prediction accuracy for the test data are shown in Figs.13 and 14, respectively. By adding the random noise for the input data, 'margin', The MSE is not well reduced, while the prediction accuracy is good for the test data.

Even though omitted in this paper, averaging the corrections of the connection weights over some number of data is also useful, which can avoid falling into 'over learning', and improve generalization.

V. Comparison beween Multi-Modal NN and Optimized Single NN

the multi-modal NN (MNN) has been proposed to improve the prediction accuracy [6]. It achieves 66% accuracy, which is the same as the optimized single NN, as described in this paper. However, it requires 5 NNs. This means the network size is 5 times as large as the single NN. The prediction accuracy of the single NN shown in [6] and the optimized

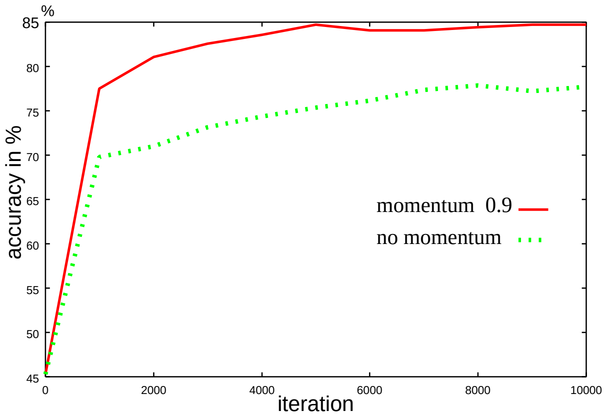

Fig. 9. Total prediction accuracy Q_3 in % for training data.

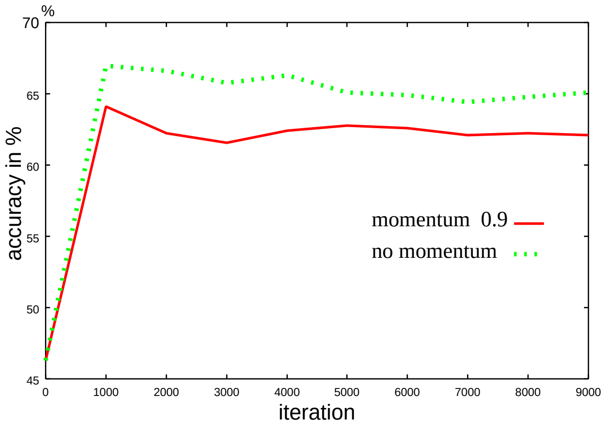

Fig. 10. Total prediction accuracy Q_3 in % for test data.

learning in this paper are compared in Fig.15. Three cases are (A) a single NN in [6], (B) a single NN with $\eta = 0.00001$, and (C) a single NN with random noise and $\eta = 0.001$. In the single NN base, the optimized learning can improve the prediction accuracy from 58% to 66%.

In the multi-modal NN, 5 NNs are independently trained, and the final decision is made by ballot. This process is equivalent to improve generalization by averaging. Therefore, this can be done in a single NN by using a small learning rate, adding random noise to the input data, and averaging the correction of the connection weights.

VI. Conclusions

Learning neural networks applied to predicting protein secondary structure easily falls into 'over learning'. One useful method to improve the prediction accuracy is to avoid 'over learning'. In this paper, this problem has been investigated from several points. A small number of hidden units is good. A very small learning rate can also avoid 'over learning'. Furthermore, adding small random noises to the input data can achieve high accuracy with a short time convergence. Updating the connection weights with average over some number of data can also avoid 'over learning'. By these optimization, a single NN can achieve 66% accuracy, which is accomplished by the multi-modal NN, which needs 5 NNs.

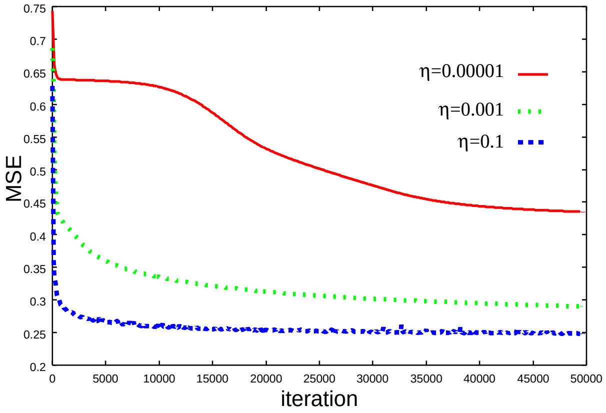

Fig. 11. Output error using several learning rates.

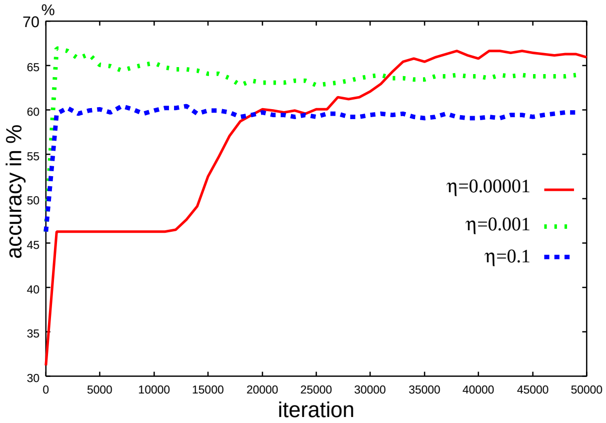

Fig. 12. Total prediction accuracy Q_3 in % for test data.

References

[1] D.M.Mount, Bioinformatics: Sequence and Genome Analysis, Cold spring Harbor Laboratory Press, New York 2001.

[2] TM Yi and S.Lander, "Protein secondary structure prediction using neirest-neighbor methods", J Mol Biol, 232, pp.117-1129, 1993

[3] AA Salamov and VVSolovey, "Prediction of protein secondary structure by combining nearest-eighbor algorithms and multiple sequence alignment", JMol Biol, 247, pp.11-15, 1995.

[4] R.Burkhard and S.Chris, "Prediction of protein secondary structure at better than 70% accuracy", Acadimic Press Lomited, pp.587-599, April 1993.

[5] D.G.Kneller, F.E.Cohen and R.Langridge, "Improvements in protein secondary structure prediction by enhanced neural networks", J Mol Biol, 214, pp.171-182, 1990.

[6] H.Zhu, I.Yoshihara and K.Yamamori, "Prediction of protein secondary structure by multi-modal neural networks", IEEE&INNS, Proc. IJCNN2002, pp.280-285, May, 2002.

[7] C.Sander and R.Schneider, "Database of homology-derived structures and the structural meaning of sequence alignment", Proteins Struct. Funct. Genet, 9, pp.56-68, 1991.

[8] B.Rost and C.Sander, "Third generation prediction of secondary structures", Webster D. M. (ed.): 'Predicting protein structure'. Humana Press, 1998.

[9] W.Kabsch and C.Sander, "Dictionary of protwin secondary structure: Pattern recognition of Hydrogen bonded and geometrical features", Biopolymers, 22, pp.2577-2637, 1983.

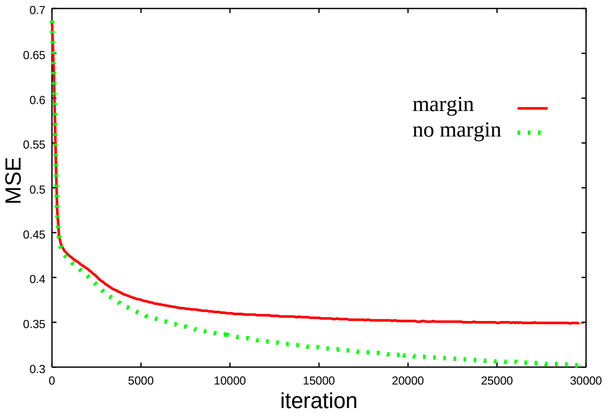

Fig. 13. Output error by adding random noise to input data, denoted ’margin’.

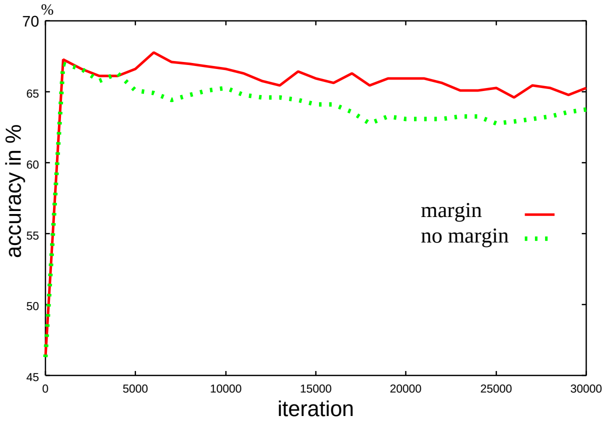

Fig. 14. Total prediction accuracy Q_3 in % for test data. Random noise is added to input data, denoted ’margin’.

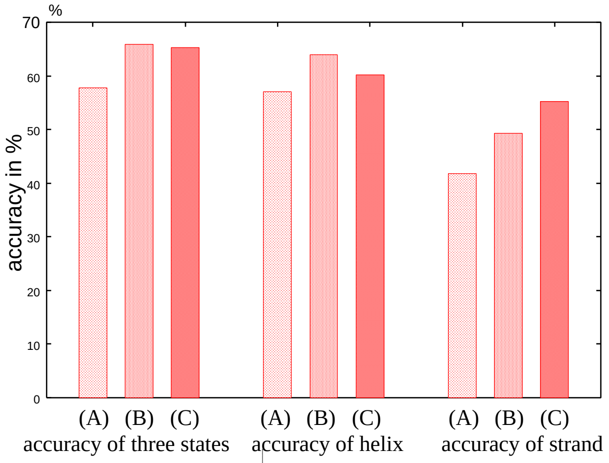

Fig. 15. Comparison of prediction accuracy of (A)single NN in [6], (B)single NN with $\eta = 0.00001$ and (C)single NN with random noise and $\eta = 0.001$.