# Avoiding overfitting in multilayer perceptrons with feeling-of-knowing using self-organizing maps

Kazushi Murakoshi *

*Department of Knowledge-based Information Engineering, Toyohashi University of Technology, 1-1 Hibarigaoka, Tenpaku-cho, Toyohashi 441-8580, Japan*

Received 14 June 2004; revised 24 September 2004; accepted 28 September 2004

**Abstract**

Overfitting in multilayer perceptron (MLP) training is a serious problem. The purpose of this study is to avoid overfitting in on-line learning. To overcome the overfitting problem, we have investigated feeling-of-knowing (FOK) using self-organizing maps (SOMs). We propose MLPs with FOK using the SOMs method to overcome the overfitting problem. In this method, the learning process advances according to the degree of FOK calculated using SOMs. The mean square error obtained for the test set using the proposed method is significantly less than that in a conventional MLP method. Consequently, the proposed method avoids overfitting.

*Key words:*
feeling-of-knowing; self-organizing maps; multilayer perceptrons; reliability; similarity; on-line learning

## 1 Introduction

Overfitting in multilayer perceptron (MLP) training is a serious problem. The purpose of this study is to avoid overfitting in on-line learning.

Overfitting is the phenomenon in which a learning algorithm adapts too well to a training set. The performance for the test set suffers due to the application of techniques that have learnt the training set too well. Bagging and bootstrap

* Corresponding author. phone: +81-532-44-6899; fax: +81-532-44-6873.
*Email address:* mura@tutkie.tut.ac.jp (Kazushi Murakoshi).

*Article published in Biosystems, 80(1): 37-40, 2005. doi:10.1016/j.biosystems.2004.09.031*

approaches (Breiman, 1994; Schapire, 1990; Tibshirani and Knight, 1995) have been developed to cope with this problem. These techniques, however, are ineffective in on-line learning, where a training set is not provided beforehand.

The adaptive natural gradient learning (ANGL) algorithm (Amari et al., 2000; Park et al., 2000) is known to enable ideal performances for the on-line learning of MLPs. The ANGL algorithm avoids plateaus, while the backpropagation (BP) type learning algorithms (Rumelhart et al., 1986) do not. However, the ANGL algorithm still has an overfitting problem. For the overfitting problem, Park et al. (2004) have proposed a method using optimized regularization. This method, however, is not designed for on-line learning but for batch learning.

To overcome the overfitting problem for on-line learning, we have investigated feeling-of-knowing (FOK). The FOK is a subjective sense of knowing a thing before recalling it. We hypothesized that the overfitting problem can be solved using the FOK: we do not execute the learning of a training data at a high FOK, while we execute the learning of a training data at a low FOK.

The characteristics of the FOK have been investigated in psychological experiments (Hart, 1965; Metcalfe, 1986; Reder and Ritter, 1992) and recently in functional magnetic resonance imaging (fMRI) experiments (Kikyo et al., 2002; Maril et al., 2003). Even from these experimental data, the neuronal mechanism of the FOK has not been clarified to the extent that we can formulate a plausible biological model. The results of these experiments, however, suggest that the processes of FOK are different from those of knowing the answer: FOK is part of an unsupervised learning process while knowing the answer is part of a supervised learning process. Thus, we have selected self-organizing maps (SOMs) (Kohonen, 1995), one of the unsupervised learning methods, as a model of FOK.

In Section 2, we propose an algorithm for avoiding overfitting in MLPs with FOK using SOMs. Section 3 shows the results of computer experiments. Section 4 concludes this paper.

## 2 MLPs with FOK using SOMs

We illustrate the structure and information flow of MLPs with FOK using SOMs in Fig. 1. The left-hand part of Figure 1 indicates MLP learning while the right-hand part indicates SOM learning. The MLP part learns the input data vector $\boldsymbol{x}$ using an ANGL algorithm (Amari et al., 2000; Park et al., 2000). In parallel with the MLP learning, the SOM part learns the input vector $\boldsymbol{x}$ as

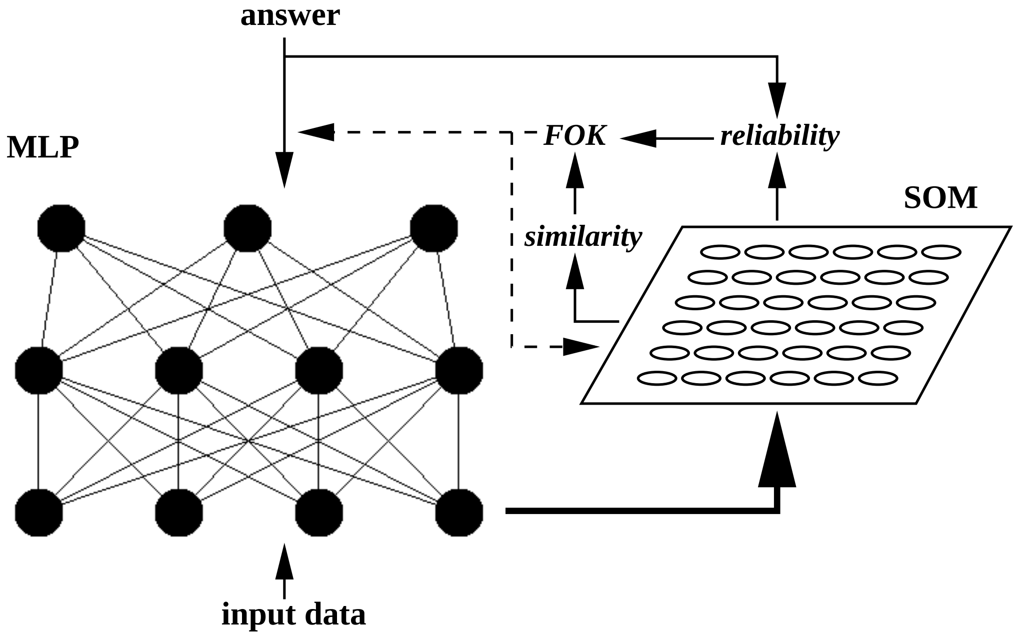


Fig. 1. Structure and information flow of MLP with FOK using SOM.

follows (Kohonen, 1995):

$$\frac{\boldsymbol{x}^{\mathrm{T}}(t)\boldsymbol{m}_c(t)}{\|\boldsymbol{x}(t)\|\|\boldsymbol{m}_c(t)\|} = \max_i \left\{ \frac{\boldsymbol{x}^{\mathrm{T}}(t)\boldsymbol{m}_i(t)}{\|\boldsymbol{x}(t)\|\|\boldsymbol{m}_i(t)\|} \right\}, \tag{1}$$

$$\boldsymbol{m}_i(t+1) = \frac{\boldsymbol{m}_i(t) + \alpha'\boldsymbol{x}}{\|\boldsymbol{m}_i(t) + \alpha'\boldsymbol{x}\|}, \tag{2}$$

$$\alpha' = \alpha \exp\left( -\frac{\|r_c - r_i\|^2}{2\sigma^2} \right), \tag{3}$$

where $\boldsymbol{m}_i$ is the weight vector of the $i$-th node in the SOM, $\alpha$ is the learning rate, $t$ is a discrete time index, $r_c$ and $r_i$ are position vectors at the winner node $c$ and the $i$-th node, respectively, and $\sigma$ is variance of the Gaussian function. In Equation (1) the inner products of $\boldsymbol{x}$ and $\boldsymbol{m_i}$ are divided by their norms for standardization.

We define the left-hand side of Equation (1) as $similarity(t)$:

$$similarity(t) = \frac{\boldsymbol{x}^{\mathrm{T}}(t)\boldsymbol{m}_c(t)}{\|\boldsymbol{x}(t)\|\|\boldsymbol{m}_c(t)\|}. \tag{4}$$

While $similarity(t)$ does not need the answer, $reliability(t)$ which needs the answer is calculated as follows.

For each time step, the class counter of the winning node $c$ is incremented by one. Afterwards, the largest class counter of each node is used as its label. In each node, $reliability$ $(t)$ is calculated by dividing the largest class counter by the summation of all class counters.

As mentioned in Section 1, the neuronal mechanism of the FOK has not been clarified to the extent that we formulate a plausible biological model. Here, we formulate a simple computational model of FOK on the basis of the consideration that FOK is a sense of knowing a thing before recalling it. Using $similarity(t)$ and $reliability(t-1)$ of the winning node, $FOK(t)$ is calculated as the geometric average between them, as follows:

$$FOK(t) = \sqrt{similarity(t) \times reliability(t-1)}. \tag{5}$$

The reason $reliability(t-1)$ is used instead of $reliability(t)$ is that $reliability(t)$ requires the answer for the input vector at time $t$, whereas FOK does not need the answer (Hart, 1965; Metcalfe, 1986; Reder and Ritter, 1992; Schunn et al., 1992).

Along with the hypothesis described in Section 1, we use $FOK(t)$ as a learning controller in MLPs and SOMs. Concretely, neither MLPs nor SOMs learn with probability $FOK(t)$ when the class for the input $\boldsymbol{x}$ agrees with the label of the winning node, because learning at high $FOK(t)$ includes the risk of overfitting.

## 3 Computer experiments

We conducted computer experiments to compare the mean square error (MSE) obtained for the test set using the conventional MLP method and that obtained using the proposed MLP method with FOK. A number of learning problems, namely, the iris problem, the glass problem, and the vehicle problem, were selected from the UCI machine learning database (UCI Machine Learning Group, 2003). The numbers and parameters used in the experiments are shown in Table 1. For every independent paired run of the two methods, we randomly selected approximately two thirds of all instances in the benchmark problem for training, and the remaining data were used for testing. In each experiment, the initial values of the weights in MLP and SOM were randomly selected subject to the uniform distribution on the intervals $[-0.1, 0.1]$ and $[0.45, 0.55]$, respectively. We conducted twenty runs for 100000 steps, with different initial values for each algorithm and each problem. At each step, the input $\boldsymbol{x}$ was randomly selected.

The obtained results are shown in Table 2. In each learning problem, there was a significant difference using the paired t-test ($p < 0.01$) in terms of the mean $\pm$ standard deviation (s.d.) between that in the proposed MLP with FOK using SOMs method and that in the conventional MLP method, as shown in Table 2.

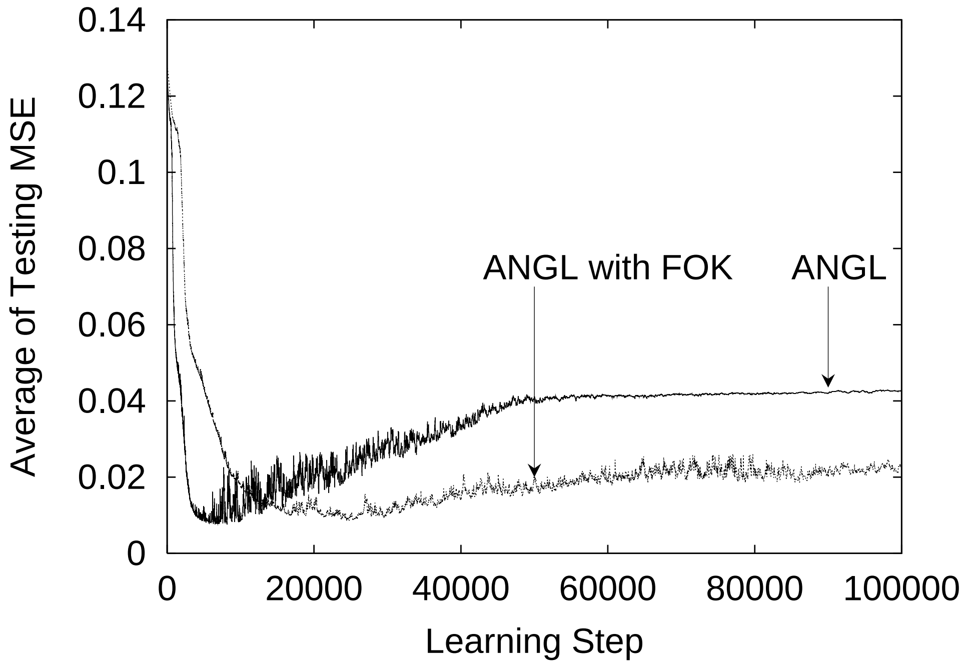


Fig. 2. Learning curves of test set for the iris problem.

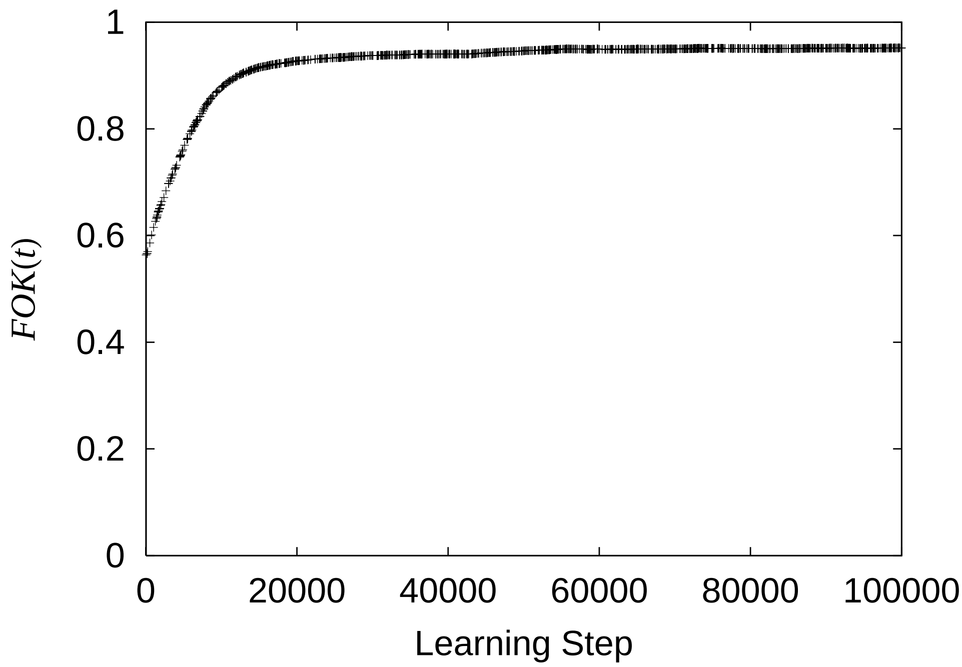


Fig. 3. Learning curve of $FOK(t)$ of an input instance for the iris problem.

We show an example of a learning curve of testing MSE for the iris problem in Fig. 2. We can confirm that the learning curve obtained with the conventional ANGL method reveals overfitting while those obtained with the proposed

ANGL with FOK method avoids overfitting. The learning curves of testing MSE for the other problems were similar to that shown in Fig. 2.

Figure 3 shows a learning curve of $FOK(t)$ of an input instance $\boldsymbol{x}$ for the iris problem. As the learning step progresses, $FOK(t)$ increases, which results in the avoidance of the overfitting exhibited in Fig. 2.

## 4 Conclusion and discussion

We propose a learning method to overcome the overfitting problem in on-line learning. In this method, the learning process advances according to the degree of FOK calculated using SOMs. Consequently, the proposed MLPs with FOK using SOMs method avoids overfitting.

In another method, BP-SOM (Weijters et al., 1997; Weijters, 1995; Weijters et al., 1998), SOMs were also used to avoid overfitting. The input of the SOM in the BP-SOM method is the hidden layer output of the MLP while the input of the SOM in our proposed method is the input layer output of the MLP. The MLP using BP-SOM method learns its weights according to the learning of the SOM. That is, the SOM of the BP-SOM method directly affects the learning of the MLP. Such learning carries the risk of destroying convergence in the learning of MLPs, while our proposed method does not. Actually, the ANGL-SOM, a method using the ANGL algorithm instead of the BP algorithm in the BP-SOM method was not able to learn well. Such problems, on the other hand, do not arise in our method, because the ANGL algorithm and SOMs perform learning in parallel.

The map size of the SOM in this study was fixed. The size affects the capacity of the FOK on our proposed method. On the other hand, capacity of human FOK seems to be adequately adjusted. Realizing a method of adjusting the map size is a future work.

## References

Amari, S., Park, H., and Fukumizu, K., 2000. Adaptive method of realizing natural gradient learning for mulilayer perceptrons. Neural Computation, 12, 1399–1409.

Breiman, L., 1994. Bagging predictors. Technical Report, Department of Statics, University of Calfornia, Berkeley.

Hart, J. T., 1965. Memory and the feeling-of-knowing experience. Journal of Educational Psychology, 56, 208–216.

Kikyo, H., Ohki, K., and Miyashita, Y., 2002. Neural correletes for feeling-of-knowing: An fMRI parametric analysis. Neuron, 36, 177–186.

Kohonen, T. 1995. Self-organizing maps. Berlin; New York: Springer-Verlag.

Maril, A., Simons, J. S., Mitchell, J. P., and Schwartz, B. L., 2003. Feeling-of-knowing in episodic memory: an event-related fMRI study. Neuroimage, 18, 827–836.

Metcalfe, J., 1986. Feeling of knowing in memory and problem solving. Journal of Experimental Psychology: Learning, Memory, and Cognition, 12, 288–294.

Park, H., Amari, S., and Fukumizu, K., 2000. Adaptive natural gradient learning algorithms for various stochastic models. Neural Networks, 13, 755–764.

Park, H., Murata, N., and Amari, S., 2004. Improving generalization performance of natural gradient learning using optimized regularization by NIC. Neural Computation, 16, 355–382.

Reder, L. M., and Ritter, F. E., 1992. What determines initial feeling of knowing? familiarity with question terms, not with the answer. Journal of Experimental Psychology: Learning, Memory, and Cognition, 18, 435–452.

Rumelhart, D. E., Hinton, G. E., and Williams, R. J. 1986. Learning internal representations by error propagation. In D. E. Rumelhart, J. L. McClelland, and the PDP Research Group (Eds.), Parallel distributed processing: Explorations in the microstructure of cognition (pp. 318–362). Cambridge, Massachusettes: MIT Press.

Schapire, R. E., 1990. The strength of weak learnability. Machine Learning, 5, 197–227.

Schunn, C. D., Reder, L. M., Nhouyvanisvong, A., Richards, D. R., and Stroffolino, P. J., 1992. To calculate or not calculate: A source activation confusion (SAC) model of problem-familiarity's role in strategy selection. Journal of Experimental Psychology: Learning, Memory, and Cognition, 23, 3–29.

Tibshirani, R., and Knight, K., 1995. Model search and inference by bootstrap "bumping". Technical Report, Department of Statics and Department of Preventive Medicine and Biostatics, University of Calfornia, Berkeley.

UCI Machine Learning Group., 2003. UCI machine learning home page. (Retrieved in 2003 from http://www.ics.uci.edu/˜mlearn/)

Weijters, A., 1995. The BP-SOM architecture and learning rule. Neural Processing Letters, 2(6), 13–16.

Weijters, T., van den Bosch, A., and van den Herik, J., 1998. Interpretable neural networks with BP-SOM. Lecture Notes in Artificial Intelligence, 1416, 564–573.

Weijters, T., van den Herik, H. J., van den Bosch, A., and Postma, E., 1997. Avoiding overfitting with bp-som. Proceeding of 15-th International Joint Conference on AI, 2, 1140–1145.

Table 1
Numbers and parameters used in experiments.

| data name | iris | glass | vehicle |
|---:|---:|---:|---:|
| number of instances | 150 | 214 | 846 |
| number of attributes | 4 | 9 | 18 |
| (=number of MLP input units) | | | |
| number of classes | 3 | 7 | 4 |
| (=number of MLP output units) | | | |
| map size of SOM | 6×6 | 6×6 | 6×6 |
| number of training data | 100 | 150 | 550 |
| number of testing data | 50 | 64 | 296 |
| learning rate of MLPs | 0.001 | 0.001 | 0.001 |
| learning rate of SOMs | 0.001 | 0.001 | 0.001 |
| variance of Gaussian function in SOMs | 1.0 | 1.0 | 1.0 |

Table 2
Results.

| data name | average ± s.d. of last MSE for testing data with ANGL | average ± s.d. of last MSE for testing data with ANGL-FOK |
|---:|---:|---:|
| iris | $0.0166 \pm 0.01114$ | * $0.0143 \pm 0.01049$ |
| glass | $0.0506 \pm 0.00738$ | * $0.0433 \pm 0.00373$ |
| vehicle | $0.0524 \pm 0.01133$ | * $0.0443 \pm 0.00512$ |

*There was a significant difference from the other method using paired t-test ($p < 0.01$).