1995 IEEE INTERNATIONAL CONFERENCE ON

NEURAL NETWORKS

PROCEEDINGS

Perth
Western Australia
27 November-1 December
1995

1995 IEEE INTERNATIONAL CONFERENCE ON

EVOLUTIONARY COMPUTATION

Perth
Western Australia
29 November-1 December
1995

ICNN '95 / ICEC '95 ' ANZIIS '95 HOMEPAGE

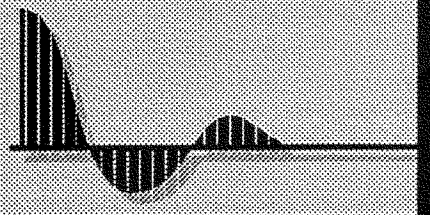
Select the Topic of Interest

PROCEEDINGS OF THE THIRD AUSTRALIAN AND NEW ZEALAND CONFERENCE ON

INTELLIGENT INFORMATION SYSTEMS

Perth
Western Australia
27 November
1995

ABOUT CAUSAL PRODUCTIONS



Pattern Classification of EEG Signals Using a Log-linearized Gaussian Mixture Neural Network

Osamu FUKUDA, Toshio TSUJI and Makoto KANEKO
Faculty of Engineering, Hiroshima University
email: Fukuda@huis.hiroshima-u.ac.jp

ABSTRACT

In this paper, we propose a pattern classification method of EEG (electroencephalogram) signals measured by a simple and handy electroence phalograph to evaluate possibility of the EEG signals as a human interface tool. Subjects are asked to switch their eye states or exposed a flash light turning on and off alternatively according to pseudo-random series for 450 seconds. The EEG signals are measured during experiments and used for classification. Each EEG signal may have different distribution depending on two states of the stimulation such as eye opening/closing and presence/absence of the flash light. Therefore a Log-linearized Gaussian Mixture Neural Network (LLGMN) incorporated a statistical model is used. It is shown from the experiments that the EEG signals can be classified sufficiently and classification rates change depending on the number of training data and the dimension of feature vectors.

1. Introduction

An EEG signal pattern is changed by external or internal factors such as photic stimulation, auditory stimulation, and intentions of movements, which may be used as an interface in a virtual reality and a teleoperation, or a communication tool for handicapped person, if an operator's intended movement is estimated from the EEG patterns. Up to the present time, some investigations of EEG pattern classification using neural networks have been carried out [1][2]. Most of them, however, were motivated to develop an automatic diagnosis in a clinic, and few studies to develop a new interface tool were carried out [3]. In case of the pattern classification of unclear EEG signals using back propagation neural networks (BPN) [4], the networks need a large number of training data, learning iterations, and a large scale of structure. Therefore, it is very difficult to attain high classification performance. Recently, the neural networks with incorporated probability density function (*pdf*) has attracted considerable attention.

In this paper, a pattern classification method of EEG signals using neural networks with an incorporated *pdf* model is proposed. The network used here can acquire the *pdf* information of sample data through training. Using the networks, the pattern classification of EEG signals are carried out under photic stimulation by eye opening/closing and an artificial light.

2. Log-linearized Gaussian Mixture Neural Network

2.1. A Network Structure

Generally, an input for a pattern classification problem can be considered as a stochastic variable with a certain distribution. In this case, the pattern classification problem usually reduces to an estimation problem of a *pdf*, since the classification can be performed according to the Bayes decision rule if a posteriori probability of the input pattern is obtained accurately. Consequently, incorporating an estimation procedure of a *pdf* into neural networks improves a generalization ability, thus high classification performance can be expected.

Tsuji et al. [5] proposes a *Log-linearized Gaussian Mixture Neural Network* (LLGMN) based on the Gaussian Mixture Model (GMM) for pattern classification problems. By applying the log-linear model to a product of a mixture coefficient and a mixture component of the GMM, a model of a *pdf* can be incorporated into the feedforward neural network and a simple learning algorithm based on the back propagation is still applicable. Based on the *pdf* model incorporated by learning, the LLGMN can estimate posteriori probabilities of the input data that are not used in learning. Also, the network structure such as an activation function of each unit, a number of layers and a number of units can be determined by the corresponding structure of the GMM incorporated in the network.

Figure 1 shows the structure of the LLGMN, which is of a feedforward one with four layers. First,

0-7803-2768-3/95/$4.00

the input feature vector $\boldsymbol{x} = [x_1, x_2, \cdots, x_d]^T \in \Re^d$ is preprocessed and converted into the modified input vector $\boldsymbol{X} \in \Re^H$ $(H = 1 + d(d+3)/2)$ according to (1) in order to represent the *pdf* corresponding to each component of the GMM as a linear operation of $\boldsymbol{X}$ [5]:

$$\begin{aligned} \boldsymbol{X} &= [1, \boldsymbol{x}^{\mathrm{T}}, x_1^2, x_1x_2, \cdots, x_1x_d, x_2^2, x_2x_3, \\ &\quad \cdots, x_2x_d, \cdots, x_d^2]^{\mathrm{T}}. \end{aligned} \tag{1}$$

The first layer consists of H units corresponding to the dimension of $\boldsymbol{X}$ and the identity function is used for the activation function of each unit. The second layer receives the output of the first layer weighted by a coefficient $w_h^{(k,m)}$. The input to the unit $\{k, m\}$ in the second layer, $Y_{k,m}$, and the output, ${}^{(2)}O_{k,m}$, are defined as

$$Y_{k,m} = \sum_{h=1}^{H} {}^{(1)}O_h w_h^{(k,m)}, \tag{2}$$

$${}^{(2)}O_{k,m} = \frac{\exp[Y_{k,m}]}{\sum_{k'=1}^{K} \sum_{m'=1}^{M_K} \exp[Y_{k',m'}]}, \tag{3}$$

where ${}^{(1)}O_j$ denotes the output of the j-th unit in the first layer and $w_h^{(K,M_K)} = 0$ $(h = 1, 2, \cdots, H)$. It should be noted that (3) can be considered as a kind of generalized sigmoid functions.

Finally, the unit k integrates the outputs of M_k units $\{k, m\}(m = 1, \cdots, M_K)$ in the second layer. The relationship between the input and the output is defined as

$${}^{(3)}I_k = \sum_{m=1}^{M_k} {}^{(2)}O_{k,m}, \tag{4}$$

$${}^{(3)}O_k = {}^{(3)}I_k. \tag{5}$$

Each unit in the second layer of the LLGMN corresponds to each component of the GMM incorporated in the network [5]. Therefore the second layer consists of the same number of units as the total component number of the GMM and its outputs (3) represent the posteriori probability of each component. The third layer consists of k units corresponding to the number of classes and outputs the posteriori probability of the class k $(k = 1, \cdots, K)$. In the LLGMN defined above, the posteriori probability of each class can be calculated based on the log-linearized Gaussian mixture structure incorporated in the network by learning only the weight coefficients $w_h^{(k,m)}$ between the first layer and the second layer.

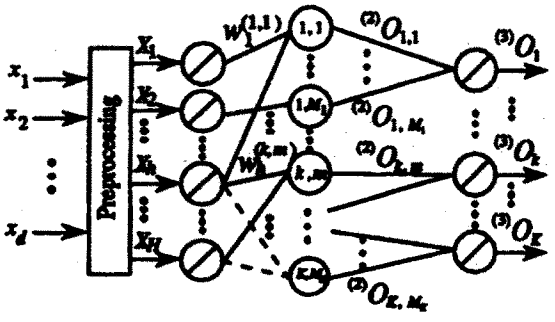


Fig. 1: Structure of an LLGMN

2.2. Learning Rule

Now, let us consider a supervised learning with a teacher vector $\boldsymbol{T}^{(n)} = [T_1^{(n)}, \cdots, T_k^{(n)}, \cdots, T_K^{(n)}]^{\mathrm{T}}$ for the n-th feature vector $\boldsymbol{X}^{(n)}$. If a teacher provides a perfect classification, $T_k^{(n)} = 1$ for the particular class k and $T_k^{(n)} = 0$ for all other classes.

The network is trained using a given set including N data $\boldsymbol{X}^{(n)}$ $(n = 1, \cdots, N)$, where the output ${}^{(3)}O_k$ of the LLGMN corresponds to $P(k|\boldsymbol{X}^{(n)})$. As an energy function for the network, we use

$$J = \sum_{n=1}^{N} J_n = -\sum_{n=1}^{N} \sum_{k=1}^{K} T_k^{(n)} \log {}^{(3)}O_k \tag{6}$$

and the learning is performed to minimize it, that is, maximize the likelihood.

For $\boldsymbol{X}^{(n)}$, a weight modification $\Delta w_h^{(k,m)}$ of the corresponding weight $w_h^{(k,m)}$ $(h = 1, \cdots, H)$ is defined as

$$\Delta w_h^{(k,m)} = -\eta_1 \sum_{n=1}^{N} \frac{\partial J_n}{\partial w_h^{(k,m)}}, \tag{7}$$

$$\begin{aligned} \frac{\partial J_n}{\partial w_h^{(k,m)}} &= \frac{\partial}{\partial w_h^{(k,m)}} \left(-\sum_{k=1}^{K} T_k^{(n)} \log {}^{(3)}O_k\right) \\ &= \left({}^{(2)}O_{k,m} - \frac{{}^{(2)}O_{k,m}}{{}^{(3)}O_k} T_k^{(n)}\right) X_h^{(n)} \end{aligned} \tag{8}$$

in a collective learning scheme, where $\eta_1 > 0$ is a learning rate.

It can be seen from the learning rule (8) that the teacher signals, which are given for classes, are back-propagated to each component according to the ratio of the posteriori probability ${}^{(2)}O_{k,m}$ of each component to the posteriori probability ${}^{(3)}O_k$ of the class k.

In the present paper, dynamics of a terminal attractor [6] is incorporated in the learning rule in order to speed it up. The terminal attractor is based on a concept that the Lipschitz conditions are violated at the equilibrium point. The dynamics converges to the equilibrium point in a finite specified time.

A weight $w_h^{(k,m)}$ is considered as a time dependent continuous variable and its time derivative is defined as

$$\dot{w}_h^{(k,m)} = -\eta_2 \gamma \frac{\partial J}{\partial w_h^{(k,m)}}, \tag{9}$$

$$\gamma = \frac{J^\beta}{\sum_{h=1}^{H}\sum_{k=1}^{K}\sum_{m=1}^{M_h}\left(\frac{\partial J}{\partial w_h^{(k,m)}}\right)^2}, \tag{10}$$

where $\eta_2 > 0$ is a positive learning rate and β ($0 < \beta < 1$) is a constant. The time derivative of the energy function J can be calculated as

$$\begin{aligned} \dot{J} &= \sum_{h=1}^{H}\sum_{k=1}^{K}\sum_{m=1}^{M_h}\left(\frac{\partial J}{\partial w_h^{(k,m)}}\dot{w}_h^{(k,m)}\right) \\ &= -\eta_2 J^\beta \leq 0. \end{aligned} \tag{11}$$

From (11), it can be seen that J is a monotonically non-increasing function, and always converges stably to the equilibrium point (the global minimum or one of local minima). In this case, the convergence time can be calculated as

$$\begin{aligned} t_f &= \int_0^{t_f} dt = \int_{J_*}^{J_f} \frac{dJ}{\dot{J}} = \frac{J_0^{1-\beta} - J_f^{1-\beta}}{\eta_2(1-\beta)} \\ &\leq \frac{J_0^{1-\beta}}{\eta_2(1-\beta)}, \end{aligned} \tag{12}$$

where J_0 is the initial value of the energy function J calculated using initial weights, and J_f is the final value of J at the equilibrium point. In the case of $J_f = 0$, the equal sign of (12) is held. Thus, the convergence time can be specified by learning rate η_2. On the other hand, in the case of $J_f \neq 0$, the convergence time is always less than the upper limit of (12). In this paper, the learning is carried out by a discrete form of (13) derived from (9):

$$\begin{aligned} w_h^{(k,m)}(t+\Delta t) &= w_h^{(k,m)}(t) + \frac{\Delta t}{2}(\dot{w}_h^{(k,m)}(t) \\ &\quad + \dot{w}_h^{(k,m)}(t+\Delta t)), \end{aligned} \tag{13}$$

where Δt denotes a sampling time.

3. Pattern Classification Method of EEG Signals

3.1. Experimental Apparatus

To evaluate possibility of the EEG signals as a human interface tool, a simple and handy electroence phalograph (IBVA, Random ELECTRONICS DESIGN) is used, which enables us to measure EEG signals in usual environments. This system consists of a head band, a transmitter, and a receiver.

The transmitter is attached to the head band. The EEG signals measured from the electrodes are digitised by an A/D converter (the sampling frequency =120Hz, quantisation = 8bits) after they are amplified and filtered out through high-pass (3Hz) and low-pass (40Hz) analogue filters. The size of the transmitter is quite compact (93mm × 51mm × 25mm). The personal computer, which is connected to the receiver, collects data. The surface electrodes are located at Fp1 and Fp2 that are specified by the International 10-20 Electrode System. Some noise contained within the EEG can be removed significantly by bipolar derivation between two electrodes at Fp1 and Fp2.

3.2. Experimental Conditions

In this paper, the EEG signals are measured under two kinds of conditions as follows:

[1] Photic stimulation by opening and closing eyes

Subjects have a rest on a seat in a computer room. First, EEG signals are measured during both eye opening and closing (60 seconds for each). The measured signals are used for training data. Next, subjects are asked to switch their eye states alternatively according to the pseudo-random series for 450 seconds.

[2] Photic Stimulation by an artificial light

Subjects have a rest on a seat in a darken computer room, and open their eyes. A flash light (xenon, illuminating power:0.176[J]) is set at a distance 50 cm apart from their eyes, which turns on and off with 4Hz.

Table: 1: Frequency range used in the classification experiments

Dimension of the input vector	Frequency ranges (Hz)					
d =2	0~8	9~35	-	-	-	-
d =3	0~8	9~20	21~35	-	-	-
d =4	0~8	9~12	13~20	21~35	-	-
d =5	0~4	5~8	9~12	13~20	21~35	-
d =6	0~2	3~4	5~8	9~12	13~20	21~35
Elements of the input vector	x_1	x_2	x_3	x_4	x_5	x_6

A power spectral density function of the measured EEG signal is estimated using FFT of every 128 sampled data. The power spectral density function (from 0 to 35Hz) is divided into several ranges, frequency bands of which are prepared based on the clinical use of the brain wave (delta, theta, alpha, beta). Time series of mean values of the power spectral density function within each frequency ranges are calculated, and normalized between [0, 1] in each range. Thus dimensional data $(x_1, x_2, \cdots, x_d)$ are obtained and used as the input vector of the networks, where d denotes the number of frequency

ranges. The frequency ranges used in this paper are shown in Table 1.

3.3. Pattern Classification Using Neural Networks

To compare the LLGMN with other neural networks, the pattern classification experiments are carried out using four types of the networks: the LLGMN, the MLANS (Maximum Likelihood Artificial Neural System) [7] which was developed by the direct use of the GMM, and two types of the BPNs with one or two hidden layers.

Learning for the LLGMN is carried out using (9), (10), (13) ($\beta = 0.5, \Delta t = 0.01, t_f = 5.0$). In the MLANS, learning procedure is continued until a Bhattacharyya distance [7] of the posteriori probabilities with an iteration becomes less than 0.0001. On the other hand, in BPNs, learning procedure is continued until the mean squared error becomes less than 0.1. However, if the mean squared error after 50000 iterations does not reach less than 0.2, learning procedure is stopped.

4. Evaluation of Classification Ability

4.1. EEG Classification of Eye States

A) Classification ability of the LLGMN

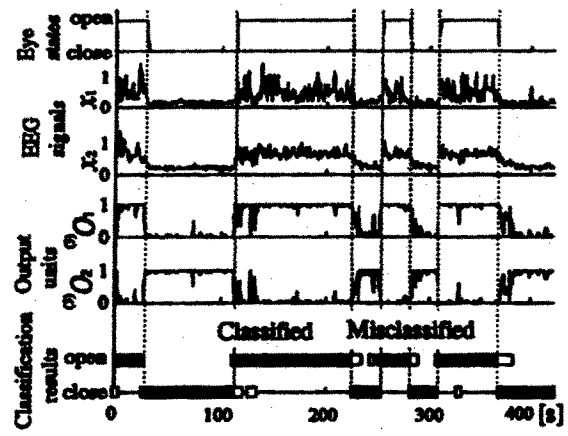


Fig. 2: An example of classification results for EEG signals by the *LLGMN*

To examine the classification ability of the networks, experiments are performed for five subjects (A, B, C, D: male, E: female). Each network is trained using 112 data (56 for each class). Then, the ratio of the correct classification to 422 data that are not used in learning are computed.

The input vector is two dimensional data for two classes shown in Table 1 (d=2, H=6, K=2). The LLGMN includes six units in the second layer which is corresponding to the total component number of GMM; six in the input layer; and two in the output layer. Figure 2 shows the classification result by the LLGMN (subject A).

Table: 2: Classification results of eye states

Subject	Performance	BPN with 1 hidden layer	BPN with 2 hidden layers	MLANS	LLGM
A (male)	Classification rate(%)	72.6	84.3	85.5	91.1
	Standard deviation	12.1	3.4	2.1	0.4
	Convergence rate(%)	53.3	86.7	100.0	100.0
B (male)	Classification rate(%)	76.2	84.4	83.7	83.3
	Standard deviation	6.1	3.1	0.7	0.6
	Convergence rate(%)	73.3	83.3	100.0	100.0
C (male)	Classification rate(%)	81.6	88.7	89.9	88.6
	Standard deviation	5.4	2.7	0.4	1.4
	Convergence rate(%)	80.0	83.3	100.0	100.0
D (male)	Classification rate(%)	73.4	78.4	80.6	81.3
	Standard deviation	5.7	3.7	0.5	1.1
	Convergence rate(%)	56.7	60.0	100.0	100.0
E (female)	Classification rate(%)	86.3	89.7	90.6	93.2
	Standard deviation	6.7	3.1	1.6	0.6
	Convergence rate(%)	56.7	60.0	100.0	100.0

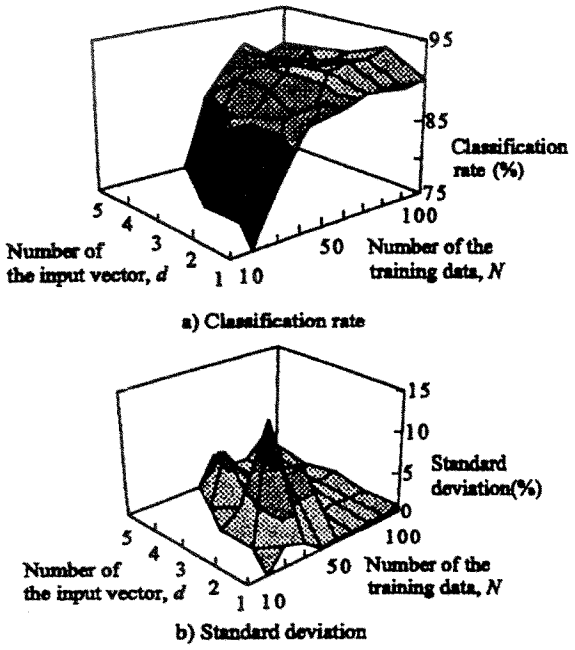
In the figure, the timing of switching eye states, the input pattern of the LLGMN (x_1, x_2), the output of the network ($^{(3)}O_1, ^{(3)}O_2$) and the classification results are shown. In this case, the LLGMN achieves considerably high performance with 91.1 percent of the classification rate. The misclassified data are observed immediately after switching eye states from opening to closing.

Table 2 shows classification results for five subjects. The BPNs with one or two hidden layers include two units with the identity activation functions in the input layer corresponding to the dimension of x; fifteen units with the sigmoid functions in each hidden layer; and two units with the sigmoid functions in the output layer. The mean values and the standard deviations of the classification rate for 30 kinds of initial weights which are randomly chosen are shown. The convergence rate is defined as the ratio of the number of converged learning to 30 trials.

As a common result of all subjects, the convergence rates of the MLANS and the LLGMN are greater than the ones of the BPNs. In the BPNs with one or two hidden layers, the mean values of the convergence rate are always less than the one of the LLGMN. In this experiment, the convergence rates of the MLANS and the LLGMN are 100 percent. Also, the standard deviations of the classification rates of the LLGMN are quite small.

B) Changes of the classification rates with the training data

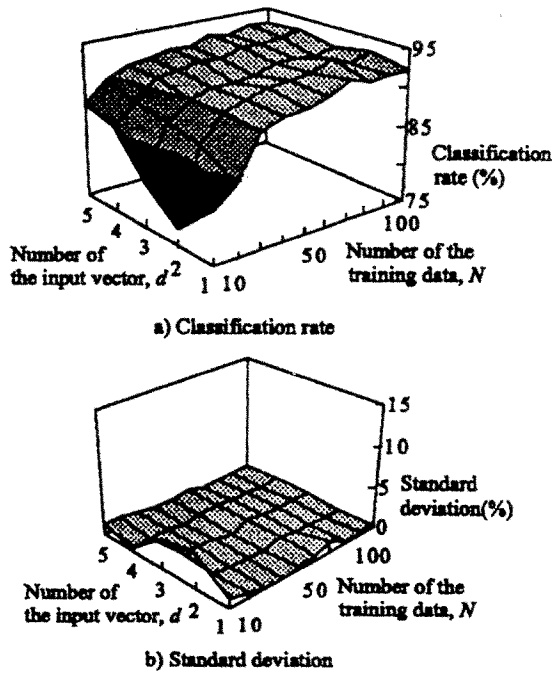
Next, we examined the changes of the classification rates with the number of the training data N and the dimension of the input vector d shown in Table 1. For each input vector, the number of training data N are changed from 10 to 100.



a) Classification rate

b) Standard deviation

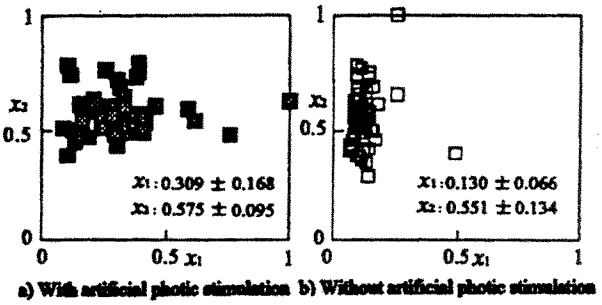
Fig. 3: Effect of the training data on classification results of eye states by MLANS



a) Classification rate

b) Standard deviation

Fig. 4: Effect of the training data on classification results of eye states by LLGMN



a) With artificial photic stimulation b) Without artificial photic stimulation

Fig. 5: Scatter diagrams of EEG data under the photic stimulation. The mean values and the standard deviations are also shown in the figures

Here, the pattern classification experiments are carried out using the LLGMN and the MLANS . Both networks are trained using fifty sets of the training data ($N = 10, 20, \cdots, 100$, $d = 1, 2, \cdots, 5$). Then the ratio of the correct classification to 422 data, which are not used in learning, is computed. Figure 3 and 4 show the mean values and the standard deviations of the classification rate for ten kinds of initial weights by using LLGMN and MLANS, respectively. Although both the networks can achieve high classification rate for large number of training data, the difference becomes clear as the number of the training data decreases. The LLGMN keeps the classification rate high even for small sample size of the training data, whereas the classification rate of the MLANS decreases. Note that the covariance matrices included in the MLANS could not be estimated in some cases ($N = 10, \cdots, 40$), because the number of the data belonging to each component decreases remarkably when the number of the training data is small. The statistical structure incorporated in the LLGMN realises considerably high classification ability for a small sample size of the training data.

Also, the classification rates of the LLGMN with sufficient number of the input vector d show a tendency to be high, even if the number of the training data decreases. On the other hand, the classification rate of the MLANS decrease considerably in those cases, and the standard deviations of the classification rates are much greater than those of the LLGMN.

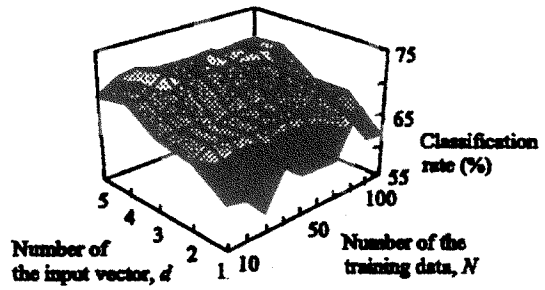
4.2. EEG Classification of the Artificial Photic Stimulation

Next, the pattern classification experiments are carried out under the artificial photic stimulation. Two dimensional input data (d=2 in Table 1) are shown in Figure 5, in which 100 data (50 for each class) are plotted. The distribution of the input data is changed according to presence or absence of the artificial photic stimulation and it seems to be difficult to classify the data into the different classes because of overlapping between classes.

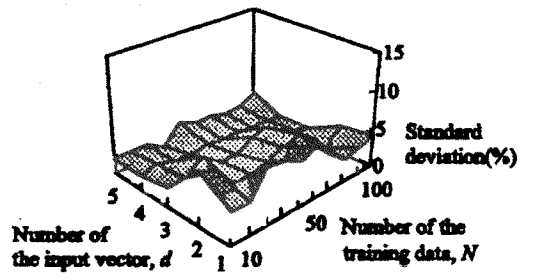
Table 3 shows experimental results for five subjects. The dimension of the input vector $d = 2, 6$ and the number of the training data $N = 50, 100$ are used. Then, the ratio of the correct classification to 422 data, which are not used in learning, is computed. Compared to the classification

Table: 3: The classification results under the artificial photic stimulation

Number of the learning data		N =50		N =100	
Dimension of the input vector		d =1	d =5	d =1	d =5
subject A	Classification rate(%)	60.1	67.6	62.4	69.9
(male)	Standard deviation	6.0	2.9	4.4	2.6
subject B	Classification rate(%)	81.1	82.6	83.7	84.5
(male)	Standard deviation	3.2	3.2	3.4	1.7
subject C	Classification rate(%)	62.5	64.9	70.7	71.9
(male)	Standard deviation	1.8	2.2	1.2	1.7
subject D	Classification rate(%)	67.4	72.4	74.4	76.7
(male)	Standard deviation	3.0	1.2	2.1	1.7
subject E	Classification rate(%)	67.7	71.6	73.8	75.5
(male)	Standard deviation	4.2	1.7	2.5	1.2



a) Classification rate



b) Standard deviation

Fig. 6: Effect of the training data on classification results of artificial photic stimulation

result of eye states, the classification rates under the artificial photic stimulation decrease. Although the difference among individuals can be observed, the classification rates tend to improve with the increase of the number of the training data from $N = 50$ to $N = 100$ and the dimension of the input vector from $d = 2$ to $d = 6$. Also, the standard deviations of the classification rates tend to decrease.

Figure 6 shows the effect of the training data on the classification results of subject A. The classification rates improve with the increase of the dimension of the input vector. On the other hand, any improvement of the classification rates depending on the number of the training data are not observed. It may be understood that a small sample size of the training data is enough to construct the statistical model, because the EEG pattern of subject A is quite stable during the experiment.

5. Conclusion

In this paper, experiments were carried out to evaluate classification ability of the LLGMN for EEG signals. The results obtained here are summarized as follows:

- The EEG signals measured by the handy electroence phalograph can be classified by the LLGMN with sufficient accuracy.
- The neural networks that are incorporated a kind of the statistical model can improve the classification rate and the convergence rate.
- The classification rates tend to improve with the increase of the dimension of the input vector and the number of the training data.

Future research will be directed to developing some techniques to incorporate dynamic changes of the EEG characteristics into the neural network.

Acknowledgments

This work was supported in part by Tateisi Science and Technology Foundation.

References

[1] F. Y. Wu, J. D. Slater, L. S. Honih and R. E. Ramsay, "A neural network for event-related potential diagnosis," Computer in biology and machine, vol. 23, no. 3, pp251-264, 1993.

[2] N. Schaltenbrand, R. Lengelle and J. P. Macher, "Neural network model : Application to autmatic analysis of human sleep," Computer and biomedical reserch, vol. 26, no. 2, pp157-171, 1993.

[3] A. Hiraiwa, K. Shimomura, Y. Tokunaga, "Pattern Recognition of Readiness Potentials and EMG by Neural Networks Proceeding Voluntary Movement," proc. 5th Symposium on *Human Interface*, pp209-214, 1989 (in Japanese).

[4] D. E. Rumelhart, J. L. McClelland and R. J. Williams, "Learning Internal Representations by Error propagation," in *Parallel Distributed Processing vol. I*, pp318-362, MIT Press, 1986.

[5] T. Tsuji, H. Ichinobe, O. Fukuda, and M. Kaneko, "A Maximum Likelihood Neural Network Based on a Log-Linearized Gaussian Mixture Model," *ICNN*, 1995 (in press).

[6] M. Zak, "Terminal Attractors for Addressable Memory in Neural Networks," Physics Letters A, Vol.133, pp218-222, 1988. pp3-189, 1991.

[7] Perlovsky L. I. and McManus M. M. ; "Maximum Likelihood Neural Networks for Sensor Fusion and Adaptive Classification," Neural Networks, 4, pp89-102, 1991.