[8] M. S. Bazaraa, H. D. Sherali, and C. M. Shetty, *Nonlinear Programming: Theory and Algorithms*, 2nd ed. New York: Wiley, 1993.
[9] J. M. Ortega and W. C. Rheinboldt, *Iterative Solution of Nonlinear Equations in Several Variables*. New York: Academic, 1970.
[10] B. C. Eaves, "On the basic theorem of complementarity," *Math. Program.*, vol. 1, pp. 68–75, 1971.
[11] F. Facchinei, "Structural and stability properties of P_0 nonlinear complementarity problems," *Math. Operations Res.*, vol. 23, no. 3, pp. 735–745, 1998.
[12] J. S. Pang, "A posterior error bounds for the linearly-constrained variational inequality problem," *Math. Operations Res.*, vol. 12, no. 3, pp. 474–484, 1987.
[13] M. Fukushima, "Equivalent differentiable optimization problems and descent methods for asymmetric variational inequality problems," *Math. Program.*, vol. 53, pp. 99–110, 1992.
[14] G. Auchmuty, "Variational principles for variational inequalities," *Numerical Functional Analysis and Optimization*, vol. 10, pp. 863–874, 1989.
[15] D. Kinderlehrer and G. Stampcchia, *An Introduction to Variational Inequalities and Their Applications*. New York: Academic, 1980.
[16] J.-J. E. Slotine and W. Li, *Applied Nonlinear Control*. Englewood Cliffs, NJ: Prentice-Hall, 1991.
[17] L. Perko, *Differential Equations and Dynamical Systems*. New York: Springer-Verlag, 1991.
[18] J. K. Hale, *Ordinary Differential Equations*. New York: Wiley, 1969.

VEP Optimal Channel Selection Using Genetic Algorithm for Neural Network Classification of Alcoholics

Ramaswamy Palaniappan, Paramesran Raveendran, and Sigeru Omatu

***Abstract*—In this letter, neural networks (NNs) classify alcoholics and nonalcoholics using features extracted from visual evoked potential (VEP). A genetic algorithm (GA) is used to select the minimum number of channels that maximize classification performance. GA population fitness is evaluated using fuzzy ARTMAP (FA) NN, instead of the widely used multilayer perceptron (MLP). MLP, despite its effective classification, requires long training time (on the order of 10^3 times compared to FA). This causes it to be unsuitable to be used with GA, especially for on-line training. It is shown empirically that the optimal channel configuration selected by the proposed method is unbiased, i.e., it is optimal not only for FA but also for MLP classification. Therefore, it is proposed that for future experiments, these optimal channels could be considered for applications that involve classification of alcoholics.**

***Index Terms*—Alcoholism, digital filter, fuzzy ARTMAP (FA), multilayer perceptron (MLP), visual evoked potential (VEP).**

I. Introduction

Evoked potential (EP) is typically generated by the nervous system in response to motor or sensory stimulation. The stimulus modality can be of visual, auditory, somatosensory, or motor, of which the visual stimulus gives rise to visual evoked potential (VEP) [11]. In recent years, EP analysis has become very useful for neuropsychological studies and clinical purposes [3], [4], [11], [13], [16]. Specifically, the effects of alcohol on the central nervous system of humans have been studied using evoked responses [3]. Evoked response has also been used to determine genetic predisposition toward alcoholism [4]. Several reports have shown that alcohol increases the latency of VEP in humans [3], [11]. Latency estimation requires signal averaging from a number of trials to reduce contamination from background electroencephalogram (EEG) activity [1], [9], [11]. However, signal averaging requires many trials, which lead to system complexity and higher computational time as compared to single trial analysis.

In this work, single trials of VEP signals are analyzed [6], [13]. A band-pass digital filter with center frequency at 40 Hz is designed to extract signals in gamma band. Parseval's theorem is used to obtain the spectral power of the filtered signal. Gamma band is particularly chosen since it is reported that gamma band spectra centered at 40 Hz is evoked during the application of sensory simulation [2], [13]. The extracted spectral power values are used to classify alcoholics and nonalcoholic subjects. This method is more efficient than methods requiring power spectrum computation like the periodogram analysis [10].

A problem encountered in analyzing VEP is the determination of channels, or electrodes, that carry significant information for classification purposes. Although modern VEP measuring instruments provide many electrodes for measuring data, in general not all are necessary to study the particular task. Different tasks such as auditory or visual may require different channel configurations to capture the information from the brain. Prior knowledge of the configuration of these channels will allow a reduction in the required hardware and computation time. Therefore, methods of identifying these channels should be devised.

In this study, we propose the use of a GA [8] to select the optimal channel configuration to maximize classification of alcoholics and nonalcoholics using VEP. The use of GA to select features for EEG classification of a brain computer interface has been investigated [7]. This method requires two classifiers: a k-nearest neighbor classifier to evaluate the GA population fitness and linear vector quantization-3 algorithm to classify the different mental thought processes represented by EEG. Another method to select relevant electrodes for multivariate EEG classification of hand movements using principal component analysis (PCA) has been proposed [12]. However, the latter maximizes signal representation with minimum features, which might not necessarily maximize classification performance. On the other hand, GA is useful in maximizing classification, as shown in this work.

GA requires a fitness or objective function, which provides a measure of performance of the population individuals. The evaluation function must be relatively fast since GA incurs the cost of evaluating all the population of potential solutions. For this reason we have used FA [5] neural network classification to evaluate the fitness function, and not other types of neural networks like multilayer perceptron (MLP) trained with the backpropagation algorithm (MLP-BP), which is relatively slow in training as compared to FA. The following explanation illustrates this fact further. Using results from our experimental study, MLP-BP takes approximately 1 h to evaluate a population. Therefore, it takes nearly 42 days to complete 1000 evaluations, which could be just 50 generations of a population size of 20 strings! In contrast, FA network takes only a second to complete an evaluation, so its fusion with GA can produce the same results within 17 mins as compared to FA fusion with MLP-BP. Although FA is used with GA, we show experimentally that the selected channels are unbiased, i.e., the channels are also optimal for MLP-BP classification. In other words, the selected channels are optimal irrespective of the classifier.

The proposed method can be considered to consist of two separate and independent systems. One system is MLP-BP, while the other is

Manuscript received August 8, 2001; revised October 18, 2001.

R. Palaniappan and P. Raveendran are with the Department of Electrical and Telecommunication, Faculty of Engineering, University of Malaya, Kuala Lumpur 50603, Malaysia.

S. Omatu is with the Department of Computer and System Science, College of Engineering, University of Osaka Prefecture, Sakai, Osaka 593, Japan.

Publisher Item Identifier S 1045-9227(02)01793-9.

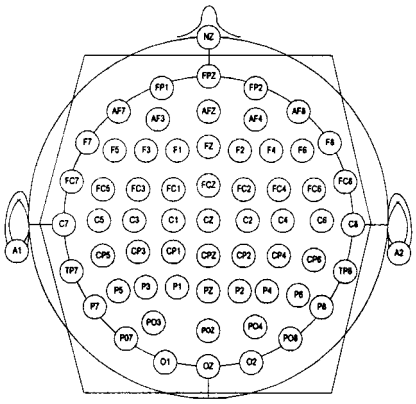


Fig. 1. Sixty-one channel electrode system (channels used outlined by the hexagon).

genetic algorithm and fuzzy ARTMAP (GAFA). GAFA is used to identify the optimal channels, i.e., channels that carry the most discriminatory information. MLP-BP is used to test the classification performance using all the available and the selected optimal VEP channels. This is to show that the performance of optimal channels is close to the performance of using all the available channels. Similar results are obtained when FA is used as a classifier. Since the optimal channels are fewer in number than all the available channels, this results in savings of hardware (electrodes, A/D cards, etc.) and a reduction in computation (i.e., classification) time.

The GAFA network uses 200 VEP patterns for training and 200 VEP patterns for validation. The training data is used to train the FA, and the validation data is used to test the classification accuracy given by the channels selected by GA. MLP-BP with GAFA is trained using 200 VEP training patterns while the MLP-BP without GAFA is trained with 200 VEP training patterns plus 200 VEP validation patterns. This is to alleviate bias and to ensure a fair comparison of performances between MLP-BP with GAFA and MLP-BP without GAFA. Both the MLP-BPs are tested with 400 VEP patterns.

The method can also be applied in any classification problem where feature set reduction is necessary without losing performance, for example, in selecting optimal features for image classification. Although our work concentrated on offline optimal channel selection and classification, the selected optimal channels can also be used in an online classification system to reduce classification time.

II. Method to Extract VEP Features

In this study, VEP data is recorded from 20 subjects from which ten are alcoholics and ten are nonalcoholics. The alcoholics tested have been abstinent for a minimum period of one month and have also been off all medication for the same period of time. Most of them had been drinking heavily for a minimum of 15 years and started drinking at approximately 20 years of age. The nonalcoholics are not alcohol or substance abusers. The subjects are seated in a reclining chair located in a sound attenuated RF shielded room. Measurements are taken for one second from 61 electrodes placed on the subject's scalp, which are sampled at 256 Hz. The signals are hardware band-pass filtered between 0.02 and 50 Hz. The electrode positions, as shown in Fig. 1, are located at sites using extension of Standard Electrode Position Nomenclature, American Encephalographic Association. The VEP signals are recorded from the subjects while being exposed to a single stimulus, which is a picture of an object chosen from the Snodgrass and Vanderwart picture set [15]. These pictures are common black and white line drawings like an airplane, a banana, a ball, etc., executed according to

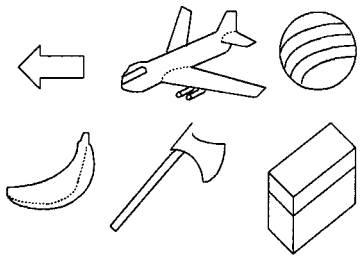


Fig. 2. Some objects from the Snodgrass and Vanderwart picture set.

a set of rules that provide consistency of pictorial representation. The pictures have been standardized on variables of central relevance to memory and cognitive processing. These pictures, as shown in Fig. 2, represent different concrete objects that are easily named, i.e., they have definite verbal labels. For further details of the data collection process, refer to Zhang *et al.* [16].

For this study, VEP signals with eye blink artifact contamination are removed in the preprocessing stage using a computer program written to detect VEP signals with magnitudes above 100 μV. These VEP signals detected with eye blinks are then discarded from the experimental study and additional trials are included as replacements. The threshold value of 100 μV is used since blinking produces 100–200 μV potential lasting 250 ms [9]. Each subject gave 40 artifact free trials, therefore producing a total of 800 VEP patterns for the experimental study.

VEP's are band-pass filtered to extract signals with energies in the gamma band centered at 40 Hz. A low-pass filter (LPF) and a high-pass filter (HPF) are used to construct the band-pass filter (BPF). Using Z transform notation, the transfer function of these filters are

$$\begin{aligned} \mathrm{LPF}(z) &= (1+z^{-1})^M \text{ and} \\ \mathrm{HPF}(z) &= (1-z^{-1})^N. \end{aligned} \tag{1}$$

The gain of this BPF is maximum at the frequency given by [13]

$$f_0 = \frac{f_s}{2\pi}\cos^{-1}\left(\frac{M-N}{M+N}\right). \tag{2}$$

A center frequency of 40 Hz corresponds to filter orders of $M = 7$ and $N = 2$, using a sampling frequency of 128 Hz. Integer multiples of these filter orders could be used to reduce the filter bandwidth. The gain can be obtained by replacing z with $e^{j2\pi fT}$, where T is the sampling time and $j = \sqrt{-1}$. It is given by

$$G_{M,N}(f) = (2\cos\pi fT)^M\,|2\sin\pi fT|^N. \tag{3}$$

In the time domain, the filter can be implemented by

$$\begin{aligned} y(n) &= \sum_{r=0}^{M} {}^{M}C_r\, x(n-r) \text{ and} \\ z(n) &= \sum_{r=0}^{N} (-1)^{r\,N}C_r\, y(n-r) \end{aligned} \tag{4}$$

where ${}^{Q}C_r = Q!/(r!(Q-r)!)$, $y(n)$ is the output of the LPF, $z(n)$ is the HPF output and $x(n)$ is the input signal. After some preliminary simulation trials, filter orders of $M = 28$ and $N = 8$ are used since this gave the 3 dB bandwidth from 32 to 48 Hz (rounded to the nearest integer), which is suitable for our purpose. Another advantage of this BPF filter in this study is the ability to attenuate power line contamination since the gain factor at 60 Hz is approximately -17.8 dB. The gain factor at 40 Hz is approximately 47.2. This value is divided later in the filtered signal so that the gain factor at 40 Hz will be unity. The gain response for this filter is shown in Fig. 3.

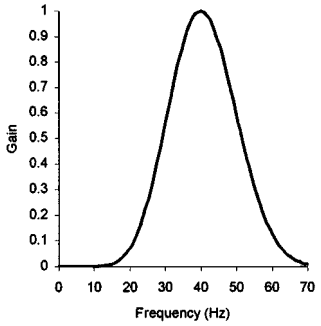


Fig. 3. Gain response of filter with orders of $M = 28$ and $N = 8$.

The filtered output, $y(n)$ contains signals mostly in the 32–48 Hz frequency range, and Parseval's theorem can now be applied to obtain the equivalent spectral power of the signal, using

$$\text{Spectral power} = \frac{1}{P}\sum_{n=1}^{P}[z(n)]^2 \tag{5}$$

where P is the total number of data in the filtered signal. The power values from each of the 61 channels are concatenated into one feature array representing the particular VEP pattern. Fig. 4 shows the process of extracting features from VEP signals.

III. Genetic Algorithm and Fuzzy ARTMAP to Select Optimal Channels

In this section, the method using GA and FA to select optimal channels will be discussed. As mentioned earlier, FA is chosen instead of other neural network (NN) like MLP-BP due to its fast training ability, which is very crucial to applications involving GA. In this study, GAFA use two different data sets, namely training and validation, consisting of 200 VEP feature arrays each. In other words, FA is trained with 200 VEP training feature arrays and tested with 200 VEP validation feature arrays. The VEP feature arrays from the testing data set is not used here to ensure unbiasness in the comparison of MLP-BP classification using all the 61 channels and the selected optimal channels.

Initially, a set of populations are generated as random binary strings (a sequence of 1's and 0's) with a certain number of bits used to represent the active/inactive state of the channel. A value of 1 denotes the activation of the channel feature (i.e., the channel feature is used) and a value of 0 denotes deactivation of the channel feature (i.e., the channel feature is not used). In our case, we have 61 channels; therefore we need 61 bits to represent each population. Following this convention, we generate 25 populations. Fig. 5 illustrates this situation.

Using these populations, VEP features of the active channels from the training data are fed into FA to be trained. Since the GA proposed here requires FA classification performance as a measure of fitness of the populations, we need to validate the performance of these populations. VEP features (of the same active channels as in the training data) from the validation data are now used to evaluate the FA performance in classifying alcoholics and nonalcoholics. This process of training and validation is repeated for all the populations. The fitness function for each population is

$$\text{fitness}_{\text{population}} = \frac{\text{VEP}_{\text{correct}}}{\text{VEP}_{\text{total}}} + \frac{\text{channels}_{\text{inactive}}}{\text{channels}_{\text{total}}}, \tag{6}$$

where $\text{VEP}_{\text{correct}}$ equals the correctly classified VEP patterns and $\text{VEP}_{\text{total}}$ equals the total number of VEP patterns in the validation set; $\text{channels}_{\text{inactive}}$ represents the inactive channels (represented by 0 in the populations) and the value of $\text{channels}_{\text{total}}$ is 61 to represent the total 61 channels. Half of the population fitness value comes from the FA classification ratio and the other half comes from the number of inactive channels. This fitness function given by (6) serves to maximize classification while minimizing the number of channels. From this procedure, optimal channels that maintain classification using fewer than the total 61 channels can be obtained. Using lower number of channels benefits from less hardware and lower computational time.

GA uses the performance of this validation data set to generate the populations in the next generation using reproduction, crossover and mutation operators [8]. Tournament selection is applied during reproduction from a pool of five populations chosen randomly among the total populations. A two-point crossover is used since they are able to wrap around at the end of the string. As such, it is better than a single point crossover. The crossover probability is set at 0.3 while the mutation is set at a lower probability of 0.03 to reduce excessive random perturbations. This entire cycle is then iterated for 500 generations, or until population convergence is reached where more than 20 of the total 25 populations are similar. FA is run with a vigilance parameter value of 0 and in the fast learning mode. This is to minimize the FA network size and to reduce the training time [5]. Fig. 6 illustrates the procedures involved in GAFA. Table I gives the eight optimal channels selected by GAFA and their locations. The configuration of these channels is stored for later use.

IV. MLP-BP Classification Using Eight Optimal Channels and 61 Channels

A MLP NN with single hidden layer trained by the backpropagation (BP) algorithm [14] is used to classify the VEP feature arrays as belonging to the alcoholic or nonalcoholic subjects class. Fig. 7 shows the architecture of the MLP-BP neural network used in this study. The output nodes are set at two so that the neural network can classify into either the alcoholics or nonalcoholics category. The hidden layer nodes are varied from ten to 50 in steps of ten.

The eight-channel MLP-BP is trained using 200 VEP training patterns while the 61-channel MLP-BP is trained with 200 VEP training patterns plus 200 VEP validation patterns. The 61-channel MLP-BP is trained with more VEP patterns since the 200 VEP validation patterns have been used in GAFA to identify the eight optimal channels. This ensures proper benchmarking comparison of performances between the eight-channel MLP-BP and 61-channel MLP-BP. Both the MLP-BPs are tested with 400 VEP patterns. The VEP feature arrays in each data set are chosen randomly, and training is conducted until the average error falls below 0.01 or reaches a maximum iteration limit of 50 000. The classification performance on the testing data set with the varying number of hidden nodes is given in Table II. The table also shows the classification time taken to classify a test VEP pattern. The entire simulation is written in C language and run on an Intel Pentium III 800 MHz PC with 128 MB RAM.

The results from this table show that the classification performance does not vary greatly with the variation in the number of hidden nodes. In general, the high classification performance shows that the proposed method of extracting features from VEP signals can be used with a neural network to successfully classify alcoholics and nonalcoholics. For all the varying hidden units, the classification performance of the eight-channel configuration as compared to the 61-channel configuration is lower about 2%. This is further indicated by the averaged classification of 94.30% for the eight-channel system and 96.10% for the 61-channel system. Using the optimal channels, the number of electrodes and hardware design are reduced considerably. The significant reduction of computation time for the eight-channel configuration can be noticed by comparing the time taken for classification of a VEP pattern. This is shown by the averaged classification time of 0.26 ms for the eight-channel system and 1.65 ms for the 61-channel system.

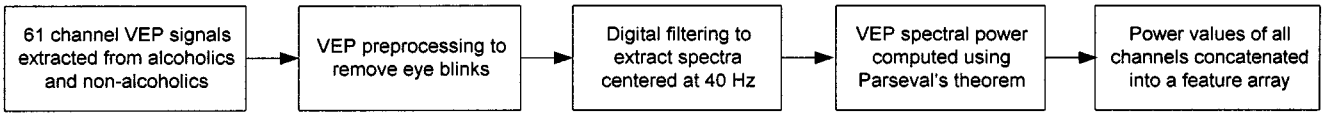


Fig. 4. VEP feature extraction to create feature arrays.

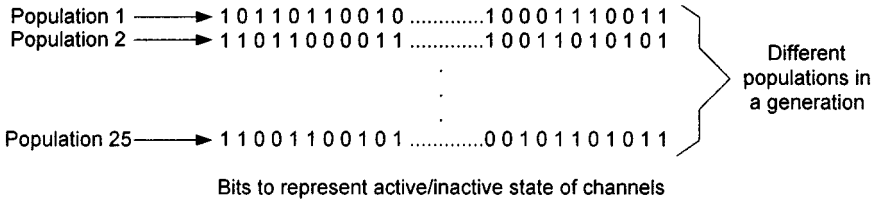


Fig. 5. Initial GA populations.

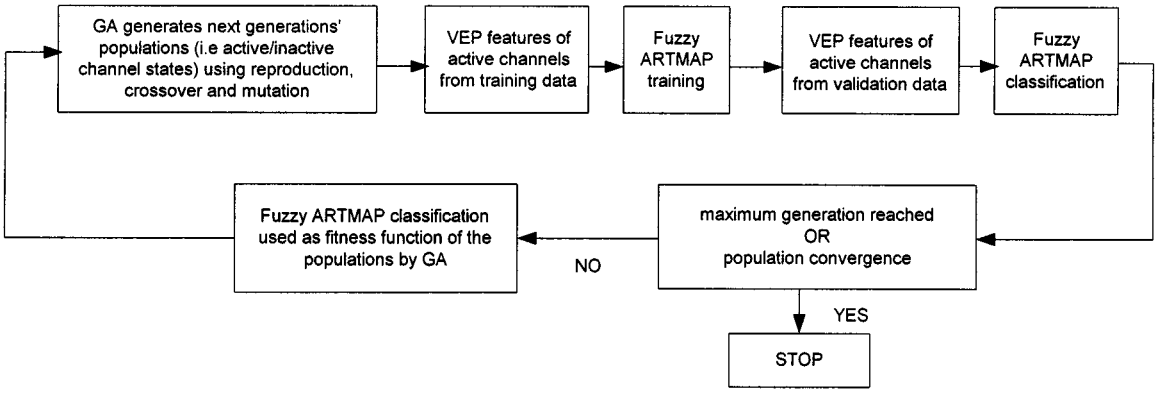
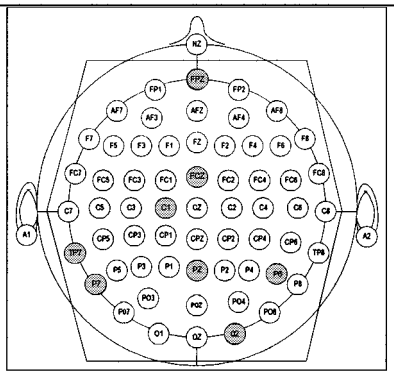


Fig. 6. GAFA method used to select optimal channels.

TABLE I
OPTIMAL CHANNELS SELECTED BY GAFA

Number of channels	8
Channel names	PZ P7 O2 FPZ TP7 P6 C1 FCZ
Channel locations (shadowed)	

V. FUZZY ARTMAP CLASSIFICATION USING 8 OPTIMAL CHANNELS AND 61 CHANNELS

To illustrate the advantage of the proposed method further, FA is used as a classifier to classify alcoholics and nonalcoholics using the eight optimal channels and all the available 61 channels. The training and testing patterns are same as those used in MLP-BP. Classification results and computation time for classifying a VEP pattern for varying vigilance parameter values are as shown in Table III. The performance obtained using the eight optimal channels are slightly poorer (on a margin of about 5%) to that achieved by using all the available 61 channels. Using averaged classification accuracy denotes that the eight-channel system performs at 86.20%, while for the 61-channel system is at 90.10%. But the computation time for the eight-channel classification (averaged classification time of 0.93 ms) is considerably reduced than the 61-channels (averaged classification time of 9.70 ms). This is true for all the varying vigilance parameter values.

Comparing the performance of MLP-BP and FA, it can be concluded that using the former gives better classification performance. Classification time using MLP-BP is less than FA since the MLP-BP ar-

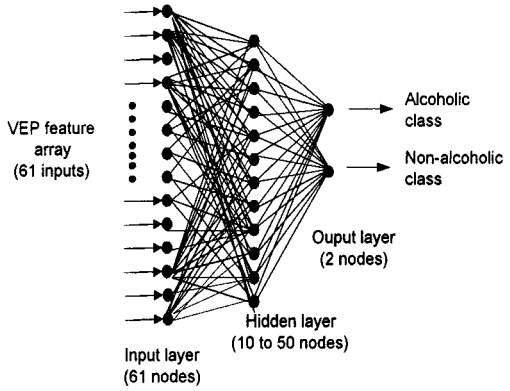


Fig. 7. MLP-BP NN architecture used in this study.

TABLE II
MLP-BP CLASSIFICATION RESULTS

	8 optimal channels		All 61 channels	
Hidden Nodes	Time (ms)	Performance (%)	Time (ms)	Performance (%)
10	0.15	94.00	0.80	95.75
20	0.25	94.75	1.10	96.25
30	0.28	94.25	1.65	95.75
40	0.30	94.00	1.95	96.25
50	0.31	94.50	2.75	96.50
Average	**0.26**	**94.30**	**1.65**	**96.10**

TABLE III
FA CLASSIFICATION RESULTS

Vigilance	8 optimal channels		All 61 channels	
parameter	Time (ms)	Performance(%)	Time (ms)	Performance(%)
0	0.83	85.00	6.18	90.25
0.1	0.83	85.75	6.18	90.25
0.2	0.83	85.75	6.18	90.25
0.3	0.70	86.50	7.03	90.25
0.4	0.70	86.50	6.85	88.25
0.5	0.83	85.75	7.70	90.25
0.6	1.23	86.50	8.53	90.75
0.7	1.23	87.50	9.20	88.50
0.8	1.08	86.00	12.23	90.75
0.9	1.10	86.75	26.90	91.50
Average	**0.93**	**86.20**	**9.70**	**90.10**

chitecture is less complex than FA. However, MLP-BP training takes longer time, requiring many more iterations than FA before achieving convergence. In addition, MLP-BP also has the possibility of getting stuck in local minima, which is not a problem for FA in fast learning mode. But these shortcomings of MLP-BP as compared to FA are less important in comparison to the advantages of improved classification percentage and lower classification time, which is especially true for off-line MLP-BP training. The experimental results also show that the eight optimal channels selected by GAFA are unbiased, i.e., the optimal channels can be used irrespective of the classifier. As such, GAFA can be used for selecting the optimal channels and MLP-BP can be used for classification with these optimal channels.

VI. CONCLUSION

In this letter, we have proposed the use of GA to select optimal channels for NN classification of alcoholics and nonalcoholics using single trial multichannel VEP signals. The method combines GA and FA to select optimal channels or electrodes, which can be used with MLP-BP for classification. FA is used with GA rather than MLP-BP due to its low training time. The independent classification results of MLP-BP show that the selected eight optimal channels significantly reduce computational time while maintaining the classification performance as obtained using all 61 channels. This is because the proposed method can pick up the discriminatory channels that are vital for classification from channels that impair or do not influence the classification. As such, it is proposed that these optimal channels could be used in future experiments. In addition to reducing computational time, optimal channels require less hardware. The research reported by Begleiter, *et al.* [4] states evidence of some irreversible alteration in the brain of alcoholics. Results obtained in our experimental study also confirm this fact. This is shown by the ability of the MLP-BP NN to successfully classify alcoholics and nonalcoholics using the VEP features, even though the alcoholics had quit alcohol for some time. In conclusion, the high MLP-BP classification rates indicate the validity of the proposed methods to optimally classify alcoholics and nonalcoholics.

ACKNOWLEDGMENT

The authors acknowledge the assistance of Prof. H. Begleiter at the Neurodynamics Laboratory at the State University of New York Health Center, Brooklyn, who recorded the raw VEP data, and P. Conlon of Sasco Hill Research for making the data available to us. The authors thank Dr. S. A. Rajeswary from the Language and Linguistic Faculty, University Malaya, for proofreading the manuscript and the reviewers for their comments.

REFERENCES

[1] J. I. Aunon, C. D. McGillem, and D. G. Childers, "Signal processing in event potential research: Averaging and modeling," *CRC Crit. Rev. Bioeng.*, vol. 5, pp. 323–367, 1981.

[2] E. Basar, C. B. Eroglu, T. Demiralp, and M. Schurman, "Time and frequency analysis of the brain's distributed gamma-band system," *IEEE Eng. Med. Biol. Mag.*, pp. 400–410, July/Aug. 1995.

[3] H. Begleiter and A. Platz, "The effects of alcohol on the central nervous system in humans," in *The Biology of Alcoholism*, B. Kissin and H. Begleiter, Eds. New York: Plenum, 1972, vol. 2, Physiology and Behavior.

[4] H. Begleiter *et al.*, "Quantitative trait loci analysis of human event-related brain potentials," *Electroencephalogr. Clin. Neurophysiol.*, vol. 108, pp. 244–250, 1998.

[5] G. A. Carpenter, S. Grossberg, and J. H. Reynolds, "A fuzzy ARTMAP nonparametric probability estimator for nonstationary pattern recognition problems," *IEEE Trans. Neural Networks*, vol. 6, pp. 1330–1336, Nov. 1995.

[6] D. G. Childers, I. S. Fischler, T. L. Boaz, N. W. Perry, and A. A. Arroyo, "Multichannel, single trial event related potential classification," *IEEE Trans. Biomed. Eng.*, vol. 33, pp. 1069–1075, 1986.

[7] D. Flotzinger, M. Pregenzer, and G. Pfurtscheller, "Feature selection with distinction sensitive learning vector quantization and genetic algorithm," in *Proc. IEEE Int. Conf. World Congr. Comput. Intell.*, vol. 6, 1994, pp. 3448–3458.

[8] D. E. Goldberg, *Genetic Algorithms in Search, Optimization and Machine Learning*. Reading, MA: Addison-Wesley, 1989.

[9] A. Kriss, "Recording technique," in *Evoked Potentials in Clinical Testing*, A. M. Halliday, Ed: Churchill Livingstone, 1993.

[10] A. P. Liavas, G. V. Moustakides, G. Henning, E. Z. Psarakis, and P. Husar, "A periodogram-based method for the detection of steady-state visually evoked potentials," *IEEE Trans. Biomed. Eng.*, vol. 45, pp. 242–248, 1998.

[11] K. E. Misulis, *Spehlmann's Evoked Potential Primer: Visual, Auditory and Somatosensory Evoked Potentials in Clinical Diagnosis*: Butterworth-Heinemann, 1994.

[12] T. Muller, T. Ball, R. Kristeva-Feige, T. Mergner, and J. Timmer, "Selecting relevant electrode positions for classification tasks based on the electro-encephalogram," *Med. Biol. Eng. Comput.*, vol. 38, pp. 62–67, 2000.

[13] R. Palaniappan and P. Raveendran, "Single trial 40 Hz visual evoked potential detection and classification," in *Proc. IEEE 11th Workshop Statist. Signal Processing (SSP)*, Singapore, Aug. 6–8, 2001, pp. 249–252.

[14] D. E. Rumelhart and J. L. McCelland, *Parallel Distributed Processing: Exploration in the Microstructure of Cognition*. Cambridge, MA: MIT Press, 1986, vol. 1.

[15] J. G. Snodgrass and M. Vanderwart, "A standardized set of 260 pictures: Norms for name agreement, image agreement, familiarity, and visual complexity," *J. Experimental Psych: Human Learning and Memory*, vol. 6, no. 2, pp. 174–215, 1980.

[16] X. L. Zhang, H. Begleiter, B. Porjesz, W. Wang, and A. Litke, "Event related potentials during object recognition tasks," *Brain Res. Bull.*, vol. 38, no. 6, pp. 531–538, 1995.

On Learning Context-Free and Context-Sensitive Languages

Mikael Bodén and Janet Wiles

***Abstract*—The long short-term memory (LSTM) is not the only neural network which learns a context sensitive language. Second-order sequential cascaded networks (SCNs) are able to induce means from a finite fragment of a context-sensitive language for processing strings outside the training set. The dynamical behavior of the SCN is qualitatively distinct from that observed in LSTM networks. Differences in performance and dynamics are discussed.**

***Index Terms*—Language, prediction, recurrent neural network (RNN).**

I. Introduction

Gers and Schmidhuber [9] present a set of simulations with the so-called long short-term memory (LSTM) network on learning and generalizing to a couple of context free and a context sensitive language. The successful result is profound for at least two reasons: First, Gold [10] showed that, under certain assumptions, no super-finite languages can be learned from positive (grammatically correct) examples only. The possibilities are thus that the network (and its environment) enforces a learning bias which enables the network to capture the language, or the predictive learning task implicitly incorporates information of what is ungrammatical (cf. [14]). Second, the network establishes the necessary means for processing embedded sentences without requiring potentially infinite memory (e.g., by using stacks). Instead the network relies on the analog nature of its state space.

Contrary to what is claimed in [9], the LSTM is not the only network architecture which has been shown able to learn and generalize to a context sensitive language. Specifically, second-order sequential cascaded networks (SCNs) [12] are able to learn to process strings well outside the training set in a manner which naturally scales with the length of strings [4]. First-order simple recurrent networks (SRNs) [8] induce similar mechanisms to SCNs but have not been observed to generalize [6].

Manuscript received September 13, 2001.

M. Bodén is with the School of Information Science, Computer and Electrical Engineering, Halmstad University, 30118 Halmstad, Sweden (e-mail: mikael.boden@ide.hh.se).

J. Wiles is with the School of Information Technology and Electrical Engineering, University of Queensland, St. Lucia 4072, Australia (e-mail: janetw@itee.uq.edu.au).

Publisher Item Identifier S 1045-9227(02)01792-7.

TABLE I
RESULTS FOR RECURRENT NETWORKS ON THE CSL $a^n b^n c^n$, SHOWING (FROM LEFT TO RIGHT) THE NUMBER OF HIDDEN (STATE) UNITS, THE VALUES OF n USED DURING TRAINING, THE NUMBER OF SEQUENCES USED DURING TRAINING, THE NUMBER OF FOUND SOLUTIONS/TRIALS, AND THE LARGEST ACCEPTED TEST SET

Network	Hidden Units	Train. Set[n]	Train. Str. [10^3]	Sol./ Tri.	Best Test [n]
SCN/BPTT[4]	2	1, ..., 10	max 20	23/1000	1, ..., 18
SRN/IHC[6]	3	1, ..., 12	na	na	1, ..., 12
LSTM[9]	na	1, ..., 10	avg 54	10/10	1, ..., 52

As reported, the LSTM network exhibit impressive generalization performance by simply adding a fixed amount to a linear counter in its state space [9]. Notably, both SRNs and SCNs process these recursive languages in a qualitatively different manner compared to LSTM. The difference in dynamics is highlighted as it provides insight into the issue of generalization but also as it may have implications for application and modeling purposes. Moreover, complementary results to Gers and Schmidhuber's [9] treatment are supplied. We focus on the SCN since it clearly demonstrates generalization beyond training data.

II. Learning Ability

Similar to [9] networks are trained using a set of strings, called S, generated from $a^n b^n c^n$ where $n \in 1, \ldots, 10$. Strings from S are presented consecutively, the network is trained to predict the next letter, and n is selected randomly for each string.[1] Contrary to [9] we do not employ *start-of-string* or *end-of-string* symbols (the language is still context-sensitive). The crucial test for successfully processing a string is based on predicting the first letter of the next string.

The SCN has three input and three output units (one for each symbol; a, b and c). Two sigmoidal state units[2] are sufficient. Consequently, the SCN has a small and bounded state space $[0, 1]$ in contrast with the LSTM which is equipped with several specialized units of which some are unbounded and some bounded.

Backpropagation through time (BPTT) is used for training the SCN. The best SCN generalizes to all strings $n \in 1, \ldots, 18$ (see Table I) and the best LSTM manages all strings $n \in 1, \ldots, 52$ [9]. BPTT suffers from a "vanishing gradient" and is thus prone to miss long-term dependencies [1], [11]. This may to some extent explain the low proportion of SCNs successfully learning the language (see Table I). In a related study the low success rate and observed instability during learning are partly explained by the radical shifts in dynamics employed by the network [5]. Chalup and Blair [6] trained a three hidden unit SRN to predict $a^n b^n c^n$ using an incremental version of hill-climbing (IHC; see Table I).

III. Generalization

A. Infinite Languages

It is important to note that neither Gers and Schmidhuber [9] nor we are actually training our networks using a nonregular context sensitive language. A finite fragment of what a context sensitive grammar generates is a straightforwardly regular language—there are only ten

[1]The target is only the next letter of the current string and not (as in [9]) all possible letters according to the grammar. According to [9] LSTM performs similarly in both cases.

[2]The logistic activation function was used.