Classification of Sleep-Waking States using Modular Neural Networks

P. A. Estévez, M. E. Fernández, C. M. Held, C. A. Holzmann, C. A. Pérez and J. P. Pérez
Department of Electrical Engineering – University of Chile
Casilla 412-3, Santiago, CHILE

Abstract

In this paper we apply modular neural network models to classify sleep-waking states in infants. The performances of three connectionist models are compared: a) multilayer perceptron (MLP), b) mixture of experts (ME) and c) fuzzy ganglionar lattice (FGL). We propose a new methodology for enhancing neural classifiers based on input variable selection and confusion error analysis using expert criteria. The ME model resulted more robust than MLP and FGL models in presence of inconsistent or noisy data. Input variable selection and confusion error analysis using expert criteria, led to parsimonious models with less parameters and better classification rates.

1. Introduction

The identification of sleep-waking states in infants allows to study different aspects of child neurofunctional development, as well as pathologies associated to such states [8,15]. The sleep-waking states are usually recognized by visual inspection of paper polysomnographic recordings of EEG (electroencephalogram), EOG (electrooculogram), EMG (electromyogram) and other physiological signals. This procedure is time consuming and requires the presence of a highly trained expert. Computarized polysomnographic systems have recently been developed to reduce paper consumption (an 8 hours paper record consists of over 1,000 pages), decrease storage needs, reduce expert analysis time and allow for high-quality data presentation [24].

Holzmann et al. [9] developed an expert system for classifying sleep-waking states in infants. This system is based on Fuzzy Ganglionar Lattices (FGLs), which intend to emulate the human expert reasoning [8]. The FGL aims to grasp the expert's mental model and to generate explanations [8,15].

Artificial neural networks (ANN) excel at pattern recognition applications, since they are universal approximators and can constitute optimal Bayesian classifiers [16]. ANN are robust in the presence of inconsistent and noisy data, and generalize well to unknown patterns. Moreover, there exist powerful learning algorithms associated to ANN models [21]. A drawback of ANN compared to knowledge-based systems is their lacking of symbolic reasoning and semantic representation. A well-known example of ANN is the feedforward multilayer perceptron (MLP) [16]. The MLP corresponds to the standard global neural network approach.

On the other hand, modular neural networks decompose a complex function or task into a set of simpler functions or subtasks [1,11,17,20]. They present overall advantages over single MLPs in terms of generalization, interpretability of the models and learning speed [11]. An example of modular neural network is the so-called Mixture of Experts (ME) [10-13]. The underlying idea is simple: instead of using a global neural network, the ME model learns several local models called experts, and simultaneously learns to partition the input space into regions.

In this paper we apply modular neural network models to classify sleep-waking states in infants. In particular we compare the performances of three connectionist models: a) MLP, b) ME and c) FGL. In addition we propose a new methodology to enhance the interpretability of neural classifiers based on input variable selection and confusion error analysis using expert criteria.

2. Three Connectionist Models

2.1 Fuzzy Ganglionar Lattices

The fuzzy glangionar lattice model [8] intends to emulate the human expert reasoning, representing it as a sequence of elementary reasoning units (ERs). Such ERs are realized by a nonlinear operator, denoted by the function pER, where p is the number of antecedents [8,15]. Fig. 1 illustrates this operator, which is characterized as follows:

$$pER(x) = \sum_{i=1}^{2^p-1} \omega_i \prod_{j=1}^{p} x_j^{b_{ij}} + \omega_0, \tag{1}$$

where x_j corresponds to the j-th component of the input vector $\mathbf{x} \in \Re^p$, ω_i is the weight associated to the i-th term in the sum and b_{ij} is a binary index that determines the participation of the j-th input in the i-th term. The total number of parameters in a pER is 2^p. The nonlinearity of this operator is given by the products between two or more antecedents. Such products allow to incorporate the expert's evaluation of the conjunction of groups of antecedents [15].

0-7803-6583-6/00/$10.00 © 2000 IEEE

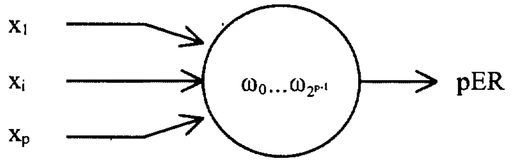

Fig. 1 Elementary reasoning unit in a FGL.

The expansion of the original input space to include cross-product terms, allows the pER operator to have a simpler architecture. For example, it is well known that the exclusive-OR Boolean function is not linearly separable and cannot be solved by a single perceptron. A hidden layer is required to solve this problem. Alternatively, if the two-dimensional input space is expanded to include the product of both inputs, the problem becomes linearly separable [14]. Besides FGLs, there are several neural network models based on the idea of including high order terms in their elementary units [14]. A drawback of this approach is the combinatorial explosion of input components. One way to deal with this problem is to limit the number of inputs (fan-in) to each elementary unit and use a hierarchical lattice [8,15]. The pER coefficients are adjusted by gradient descent method with momentum [15].

2.2 Global Neural Networks

A MLP is a collection of processing elements called neurons, interconnected by links of different strengths or weights. It has been shown that a three-layered MLP, with an input layer, a hidden layer and an output layer, is a universal approximator, if there are enough hidden units [16]. This result establishes that a MLP with one hidden layer can build any mapping from the input space to the output space (classes).

The most popular learning algorithm is error backpropagation based on gradient descent. However, the rate of convergence of gradient descent methods is often very slow, since the successive descent directions have a tendency to interfer, i.e. a minimization in one direction can spoil the minimization previously achieved in another direction [21]. For this reason second order quasi Newton or conjugate gradient methods are preferred, which are usually one or two orders of magnitude faster than first order methods [21]. In this paper we use a quasi-Newton second order algorithm called BPQ for training MLPs [18].

Standard neural network models are considered black boxes, and it is difficult to explain why certain patterns are classified as members of one class or another. In part this is due to the complexity of the resulting nets and to the distributed representation of knowledge. However, it is possible to extract rules from a trained global neural network through pruning and discretizing operations [19]. In this paper we use an input pruning method based on the weight elimination method developed in [2], expanded in [5] to the elimination of inputs.

2.3 Mixture of Experts

The ME model automatically decomposes complex tasks into simpler subtasks, partitioning the input space into local regions and assigning expert networks to each region [10,12,22]. The so-called gating networks allows this partition to happen. Fig. 2 illustrates the ME model for K experts (or local neural networks) and a single gating network.

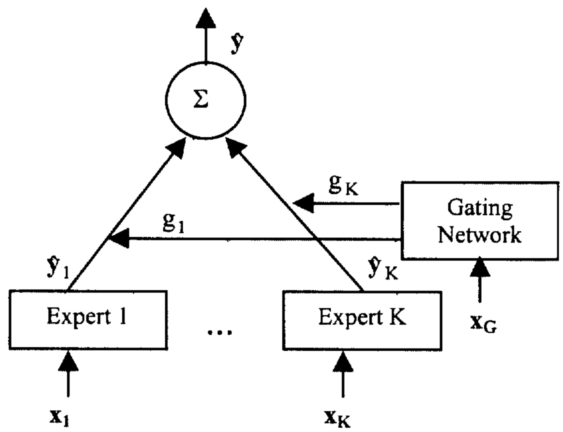

Fig. 2 Mixture of experts. The system output is the weighted sum of the K expert outputs, where the weights are the gating network outputs.

For classification problems, softmax activation functions are used in both expert and gating network units. The softmax function is defined as follows:

$$y_\ell = \frac{\exp(u_\ell)}{\sum_{i=1}^{L} \exp(u_i)}, \tag{2}$$

where u_i represents the $i-th$ linear combination of inputs. It has been shown that y_ℓ can be interpreted as the probability that the estimation belongs to the $\ell-th$ class, under a multinomial distribution for L classes [12].

In the ME model, learning is treated as a maximum likelihood problem. In particular the Expectation-Maximization (EM) algorithm is used [3]. The EM algorithm consists of two steps, the E-step (expectation) and the M-step (maximization). In the E-step the expected values of the so-called hidden variables are calculated, using the current estimation of parameters and the observed data. With this information, in the M-step the likelihood is maximized, and the parameters of the model are updated. There are several ways to implement the M-step. In this work we use an extension of the BPQ algorithm [18]. A detailed description of the EM algorithm for a mixture of experts and examples of its application to pattern classification can be found in [4,6,7,22,23].

3. Classification of Sleep-Waking States

Sleep-waking states are classified as wakefulness (WA), rapid-eye movement sleep (REMS) and quiet or non-REM sleep (NREM). In turn, NREM is subdivided into four distinct stages, from NREM1 to NREM4.

Holzmann et al. [9] have developed a model of sleep-waking states based on signal patterns found in EEG, EOG and EMG activities that allows the identification of the following five sleep-waking states: WA, NREM1, NREM2, NREM3&4, and REMS. The model considers the following five fundamental characteristics: predominance of slow-delta (SD) waves in EEG, predominance of theta (TH) waves in EEG, presence of sleep spindles (SS) in EEG, presence of rapid eye movements (REM) in EOG and predominance of muscle tone (MT) in EMG.

Fig. 3 shows examples of the five kinds of signal patterns searched for in the polysomnographic recordings: SD, TH, SS, REM and MT. Signals are analyzed on a page basis, with each page lasting 20 or 30 s. The detection of specific patterns in the EEG, EOG and EMG signals is difficult due to the presence of noise and artifacts. Artifacts occur due to body movements, electrode displacements and other interfering signals, such as EEG presence in the EOG channel. The predominant background activity in the EEG is established whenever the pattern is present more than a certain percentage of the page, which implies the setting of thresholds. In [9] a full description of the detection algorithms for signal patterns is given. These algorithms determine a value in the [0, 1] interval for each of the five signal patterns. In the case of SS and REM a single event within a time window suffices to establish its presence. Following the standard accepted by sleep experts a criterion of minimum state duration of 1 min. is adopted.

In [9] an expert characterization model of sleep-waking states is presented. This model is summarized in Table 1, and relates the characteristics described above to the states.

The sleep database built in [9] was available for this study. Each input-output pattern consists of the five characteristics mentioned as inputs, and one of the five states considered as output. A total of 880 pages from three recordings were included, with one input-output pattern per page. The recordings were taken during children naps lasting two to three hours. For comparison purposes, in this paper we use the same data partition than in [9]. The training set was composed only of data without artifacts (241 pages). Two validation-test sets were prepared, one without artifacts, SNA (135 pages for validation and 139 pages for testing), and one with artifacts, SWA (165 pages for validation and 177 pages for testing). 23 pages marked as indeterminate sleep were discarded in this study, because the total number of examples for this class was too small for training.

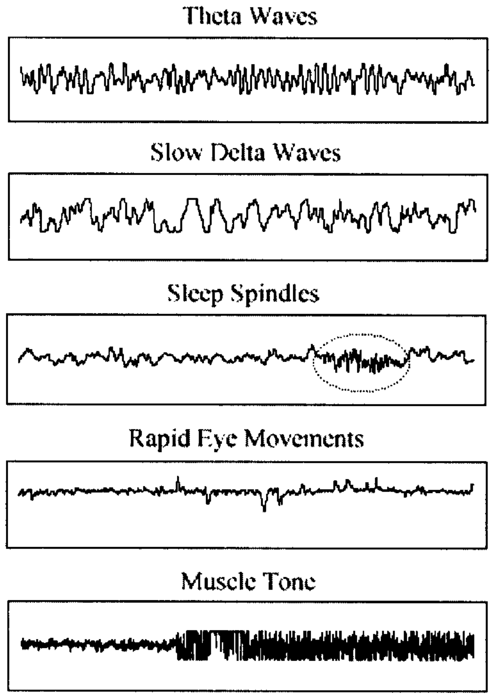

Fig. 3 Samples of signal patterns searched for in EEG (theta waves, slow delta waves and sleep spindles), in EOG (rapid eye movements) and in EMG (muscle tone).

Table 1
Expert characterization model for sleep-waking states in infants. YES/NO indicates which characteristics must be present or absent in each state, and X stands for don't care.

State/Charact.	**SD**	**TH**	**SS**	**REM**	**MT**
WA	NO	X	NO	YES	YES
NREM1	NO	YES	NO	NO	X
NREM2	NO	X	YES	NO	X
NREM3&4	YES	NO	X	NO	X
REMS	X	YES	NO	YES	NO

4. Results

For the FGL model 5 elementary reasoning units were used, one for each state. The expert characterization model given in Table 1 was used to determine the antecedents for each unit as follows:

WA = WA (SD, SS, REM, MT)
NREM1 = NREM1 (SD, TH, SS, REM)
NREM2 = NREM2 (SD, SS, REM)
NREM3&4 = NREM3&4 (SD, TH, REM)
REMS = REMS (TH, SS, REM, MT)

The resulting FGL had 80 parameters, including cross-products up to third or fourth order, depending on the number of antecedents. The FGL model was trained using gradient descent. The learning rate and the momentum factor were independently determined for each elementary reasoning unit. The combination of learning rate and momentum used were as follows: WA: 0.01, 0.005; NREM1: 0.005, 0.0025; NREM2: 0.04, 0.02; NREM3&4: 0.04, 0.02; REMS: 0.02, 0.0005.

The MLP architecture consisted of 5 inputs units, N hidden units and 5 output units. To determine the optimal number of hidden units, N was varied from 3 to 9. It resulted that the best architecture was 6 hidden units for the SNA dataset and 7 hidden units for the SWA dataset. Thus, the total number of weights associated to each model is 72 for the SNA case and 82 for the SWA case. The MLP models were trained using BPQ for 300 epochs.

A ME architecture of two expert networks and a single gating network, all without hidden units, was used. The 5 original inputs were fed into all modules. The total number of weights was 54. The ME model was trained using the EM algorithm, with the BPQ algorithm in the M-step.

Table 2 shows the percentage of correct classifications obtained with the SNA dataset, for each one of the three connectionist models considered. Table 3 shows the results obtained with the SWA dataset. The numbers given correspond to the average and standard deviation of ten trials with random weight initializations. Pairwise two-tailed t-tests show that there are no statistical differences at the level of significance 0.01 on the test set, among the three connectionist models, for both datasets. Note however that the ME model gave the minimum variance, with the minimum number of parameters.

After learning convergence, the raw outputs of all models were filtered using the criterion of minimum state duration. Pattern variations lasting less than 1 min. were discarded, thus the sleep state from the previous interval prevailed. Tables 4 and 5 show the results of filtering outputs for the models obtained with the SNA dataset and the SWA dataset, respectively. The beneficial effect of filtering becomes clear by comparing Tables 2, 3, 4 and 5. All results are enhanced, e.g. the ME model performance rises from 81.4% to 90.1% of correct classifications in the SWA test set. Once again no statistical differences were found at the level of significance 0.01 on the test set, among the three connectionist models after filtering, for both datasets.

As a second step in comparing the three connectionist models, input selection was applied. The FGL model already includes variable selection, since inputs to the elementary reasoning units were selected according to the expert characterization model of Table 1. For the MLP model, each one of the five inputs was eliminated in turn and the resulting nets were trained again. Results showed that the five combinations of pruned MLPs performed worse than the non-pruned MLP, confirming the expert criteria that all five inputs are relevant.

Table 2
Percentage of correct classifications for the SNA dataset.

Model	Training	Validation	Test
MLP	100.0 ± 0.0	96.6 ± 0.7	92.0 ± 0.9
FGL	96.7 ± 2.2	94.0 ± 2.5	93.4 ± 1.2
ME	100.0 ± 0.0	97.3 ± 0.3	92.7 ± 0.7

Table 3
Percentage of correct classifications for the SWA dataset.

Model	Training	Validation	Test
MLP	100.0 ± 0.0	88.8 ± 0.7	80.4 ± 1.1
RG	96.6 ± 2.2	86.8 ± 2.4	81.1 ± 1.7
ME	99.5 ± 0.5	89.6 ± 0.6	81.4 ± 1.1

Table 4
Percentage of correct classifications for the SNA dataset after filtering outputs by the criterion of 1 min. state duration.

Model	Training	Validation	Test
MLP	100.0 ± 0.0	98.7 ± 2.5	94.1 ± 1.9
FGL	97.7 ± 1.7	97.3 ± 2.0	97.1 ± 0.7
ME	100.0 ± 0.0	100.0 ± 0.0	93.9 ± 1.5

Table 5
Percentage of correct classifications for the SWA dataset after filtering outputs by the criterion of 1 min. state duration.

Model	Training	Validation	Test
MLP	100.0 ± 0.0	93.0 ± 1.7	89.9 ± 1.5
FGL	97.7 ± 1.7	92.6 ± 2.0	87.9 ± 2.2
ME	100.0 ± 0.0	93.4 ± 0.2	90.1 ± 1.7

For the ME model, since the same inputs are applied to the expert and gating networks, it is feasible to eliminate some inputs in each network. After pruning the ME model for the SWA dataset, the following inputs were selected:

Expert1: SS, REM and MT
Expert2: SD, TH, REM and MT
Gating Network: SD, SS and REM

Fig. 4 shows the outputs of the gating network over the training set for a given simulation. The upper and lower graphs correspond to the outputs associated to experts 1 and 2, respectively. In both graphs the x-axis corresponds to the training pattern number (e.g. training samples 1-70 correspond to the WA state). It can be clearly observed that expert 1 is associated to the WA and REMS states, and expert 2 to all NREM stages. This partition is in agreement with the expert characterization model given in Table 1. The WA and REMS states can be distinguished from all NREM stages by the presence/absence of REM in the EOG signal. The partition further shows that the sleep-waking states can be decomposed into two groups of linearly separable classes, one group of two classes (WA and REMS) and another of three classes (NREM1, NREM2 and NREM3&4).

Table 6 shows the percentage of correct classifications obtained for the ME model with the SWA dataset, after applying input variable selection. The resulting model called pruned ME-SWA was then trained with both datasets. No statistical differences were found at the level of significance 0.01 on the test set, between the pruned and non-pruned ME models. Table 7 shows the results for the pruned ME-SWA model after filtering outputs by the criterion of 1 min. state duration.

As a next step, an analysis of the confusion errors of the pruned ME model was performed. Most confusions were between NREM1 and REMS states. From the expert characterization model of Table 1, it can be seen that the main difference between both states is the presence or absence of REM. This analysis led to the binarization of the REM input using a threshold close to zero. Table 8 shows the percentage of correct classifications for the pruned SWA model after binarizing the REM input. Comparing these results with those shown in Table 6, it can be observed that binarization of REM led to a gain of about 5 percentage points on the test set for both datasets. The number of confusions between NREM1 and REMS was reduced by half. Table 9 shows the results for the ME model after pruning, binarizing and filtering.

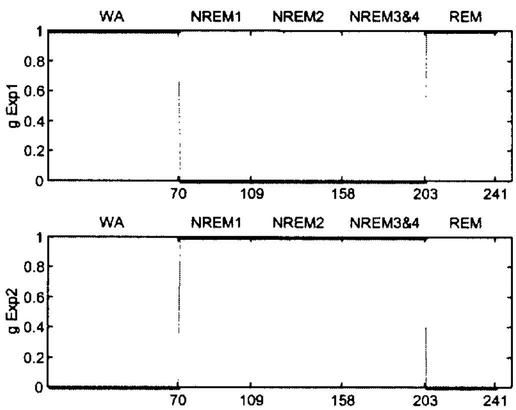

Fig. 4. Outputs of the gating network of the ME model for the SWA dataset, after applying input selection. The x-axis corresponds to training pattern numbers (e.g. training samples 1-70 correspond to the WA state). The y-axis indicates the output value in the [0, 1] interval.

Table 6
Percentage of correct classifications for the pruned ME-SWA model on both datasets.

Dataset	Training	Validation	Test
SWA	95.9 ± 0.2	89.0 ± 0.7	82.4 ± 1.1
SNA	96.2 ± 1.3	95.0 ± 0.6	88.8 ± 1.0

Table 7
Results for the pruned ME-SWA model after filtering outputs by the criterion of 1 min. state duration.

Dataset	Training	Validation	Test
SWA	99.2 ± 0.1	93.2 ± 1.2	93.7 ± 0.8
SNA	99.3 ± 0.2	98.7 ± 0.4	93.9 ± 1.6

Table 8
Percentage of correct classifications for the pruned ME-SWA model after binarizing the REM input.

Dataset	Training	Validation	Test
SWA	97.1 ± 1.9	91.5 ± 1.0	87.1 ± 1.3
SNA	96.7 ± 1.6	96.1 ± 1.0	94.6 ± 1.2

Table 9
Results for the pruned & binarized ME-SWA model after filtering by the criterion of 1 min. state duration.

Dataset	Training	Validation	Test
SWA	99.4 ± 0.4	96.6 ± 0.7	94.3 ± 1.6
SNA	99.3 ± 1.6	98.8 ± 0.6	98.6 ± 0.0

Tables 10 and 11 show the final results for the three connectionist models on both the SWA and SNA datasets. Since FGL was mainly used as a reference from previous work, the REM binarization was not applied to it. The MLP results include REM binarization. The ME results include REM binarization and input variable selection. There are statistical differences at the level of significance 0.05 on the SWA test set, between the ME model and the MLP model, and between the ME model and the FGL model. No significant statistical differences were found among the three models for the SNA dataset.

Table 10
Final results for FGL, MLP and ME models on the SWA dataset.

Model	Training	Validation	Test
FGL	97.7 ± 1.7	92.6 ± 2.0	87.9 ± 2.2
MLP	100.0 ± 0.0	96.7 ± 0.8	90.7 ± 0.4
ME	99.4 ± 0.4	96.6 ± 0.7	94.3 ± 1.6

Table 11
Final results for FGL, MLP and ME models on the SNA dataset.

Model	Training	Validation	Test
FGL	97.7 ± 1.7	97.3 ± 2.0	97.1 ± 0.7
MLP	100.0 ± 0.0	99.9 ± 0.2	98.6 ± 0.0
ME	99.3 ± 1.6	98.8 ± 0.6	98.6 ± 0.0

5. Conclusions

A new methodology for building modular neural network models based on selection of input variables and confusion error analysis has been successfully applied to the classification of infant sleep-waking states. The resulting model has better classification rates and less parameters.

The performances of FLG, MLP and ME were compared on two datasets. For the data set without artifacts, the error performances of all three models were similar. For the dataset with artifacts, the ME model showed better classification rates on an independent test set than the MLP and FGL models. Moreover, the ME model obtained has no hidden units nor high order terms, and is easily interpretable in terms of the expert characterization model of sleep-waking states in infants. The ME model resulted more robust than the other models in the presence of inconsistent or noisy data.

Acknowledgements

This research was partly supported by Conicyt-Chile under grant Fondecyt 1980909 and the Department of Electrical Engineering, University of Chile.

References

[1] Caelli, T., Guan L. and Wen W., "Modularity in Neural Computing", *Proc. IEEE*, pp. 1497-1517, Sept. 1999.
[2] Cottrell, M., Girard B., Girard Y., Mangeas M., and Muller C., "Neural Modeling for Time Series: A Statistical Stepwise Method for Weight Elimination", *IEEE Trans. Neural Networks*, Vol. 6, N° 6, pp. 1355-1364, 1995.
[3] Dempster, A. P., Laird N. M. and Rubin D. B., "Maximun Likelihood from Incomplete Data via the EM algorithm", *J. R. Statist. Soc.* B 39, pp. 1-38, 1977.
[4] Estévez, P., Paugam-Moisy H., Puzenat D., and Ugarte M., "A Scalable Parallel Algorithm for Training a Hierarchical Mixture of Neural Experts", Parallel Computing, 2000 (submitted).
[5] Estévez, P. and Caballero R., "A Niching Genetic Algorithm for Selecting Features for Neural Network Classifiers", ICANN'98, Skovde, Sweden, Vol. I., pp. 311-316, 1998.
[6] Estévez, P., Ugarte M., Paugam-Moisy H., and Puzenat D., "Modular Parallel Implementation for HME Neural Networks", International Conference on Parallel and Distributed Processing Techniques and Applications, PDTA-98, Las Vegas, USA, Vol. III, pp. 1365-1372, July 1998.
[7] Estévez, P. and Nakano R., "Hierarchical Mixture of Experts and Max-Min Propagation Neural Networks", IEEE Int. Conf. on Neural Networks, Perth, Australia, Vol. I., pp. 651-656, Nov. 1995.
[8] Holzmann, C., Ehijo A. and Pérez C. "Methodology for an Expert System on Fuzzy Analog Ganglionar Lattices", *Medical Progress through Technology* 21, pp. 147-158, 1996.
[9] Holzmann, C., Pérez C., Held C., San Martín M., Pizarro F., Pérez J., Garrido M. and Peirano P.. "Expert System Classification of Sleep-Waking States in Infants". *Medical & Biological Engineering & Computing. Journal*, UK, vol. 37, pp. 446-476, 1999.
[10] Jacobs, R. A., Jordan M. I., Nowlan S. J. and Hinton G. E., "Adaptive Mixtures of Local Experts", *Neural Computation* 3(1), pp. 79-87, 1991.
[11] Jacobs, R. A., Jordan M. I. and Barto A. G., "Task Decomposition through Competition in a Modular Connectionist Architecture – the What and Where Vision Tasks", *Cognitive Science* 15(2), pp. 219-250, 1991.
[12] Jordan, M. I., and Jacobs R. A., "Hierarchical Mixtures of Experts and the EM algorithm", *Neural Computation* 6(2), pp. 181-214, 1994.
[13] Jordan, M. I., "A Statistical Approach to Decision Tree Modelling", *Proceedings of the Seventh Anual ACM Conference on Computational Learning Theory*, ACM Press, 1994.
[14] Pao, Y.H., *Adaptive Pattern Recognition and Neural Networks*, Addison-Wesley, 1989.
[15] Pizarro, F. "Desarrollo de un Sistema Experto para Clasificar los Estados de Sueño y la Vigilia", Master Thesis, Dept. Computer Science, University of Chile, Diciembre, 1998.
[16] Reed, R.D. and Marks, R.J., *Neural Smithing*, MIT Press, Cambridge, Massachusetts, 1999.
[17] Ronco, E., and Gawthrop P., "Modular Neural Networks: a State of the Art", *Technical Report CSC-95026, Centre for System and Control*, University of Glasgow, Glasgow, UK, 1995.
[18] Saito, K., and Nakano R., "Partial BFGS Update and Efficient Step-length Calculation for Three-layer Neural Networks", *Neural Computation* 9, pp. 123-141, 1997.
[19] Setiono, R., "Extracting Rules from Neural Networks by Pruning and Hidden-Unit Splitting", *Neural Computation* 9, pp. 205-225, 1997.
[20] Sharkey, A., "Multi-net Systems", in *Combining Artificial Neural Nets*, Amanda Sharkey (ed.), Springer-Verlag, London, pp. 1-30, 1999.
[21] Sheperd, A.J., *Second-order Methods for Neural Networks*, Springer-Verlag, London, 1997.
[22] Waterhouse S. R., "Classification and Regression Using Mixtures of Experts", *Ph.D. Thesis*, University of Cambridge. 1998.
[23] Weigend, A. S., Mangeas M. and Srivastava, A. N., "Nonlinear Gated Experts for Time Series – Discovering Regimes and Avoiding Overfitting", *Int. J. Neural Syst.* Vol. 6, N° 4, pp. 373-399, 1995.
[24] White, D. P. and Gibb, T. J., "Evaluation of a Computarized Polysomnographic System", *Sleep*, Vol. 21, N°2, pp. 188-196, 1998.