Third European Congress on Intelligent Techniques and Soft Computing

Aachen, Germany,
August 28–31, 1995
Proceedings
Volume 3

EUFIT '95

ELITE-Foundation
Promenade 9
D-52076 Aachen, Germany
Phone: +49-2408-6969
Fax: +49-2408-94582
Email: eufit@mitgmbh.de
http://www.mitgmbh.de/elite/eufit.html

Acknowledgement:
Partial Support for EUFIT '95
has been provided by
the European Commission
(ESPRIT Basic Research)
and by the DGOR Working Group
"Fuzzy Sets and Neural Networks"

Herausgeber:

ELITE - European Laboratory for Intelligent Techniques Engineering
Prof. Dr. Dr. h.c. Hans-Jürgen Zimmermann

Third European Congress on Intelligent Techniques and Soft Computing
EUFIT '95, Aachen, Germany - August 28 - 31, 1995

- Proceedings Volume I: Page 0001 - 0687
- Proceedings Volume II: Page 0689 - 1310
- **Proceedings Volume III: Page 1311 - 1916**

1. Edition Aachen:
Verlag Mainz, Wissenschaftsverlag, Aachen, 1995

ISBN 3-930911-67-1 (three volumes paperback)

Verlag Mainz
Süsterfeldstraße 83
D-52072 Aachen
Tel.: +49 (0) 241-873434
Fax: +49 (0) 241-875577

Druck: Fotodruck Mainz GmbH, Aachen

Scene Analysis and Object Recognition

Image Segmentation and Feature Extraction

Image Analysis with Innovative Techniques

Fuzzy Classification Rules I

Pattern Recognition and Document Image Analysis

A NEURO-FUZZY SYSTEM FOR BAYESIAN CLASSIFICATION

Francesco Masulli
Assistant Professor
Department of Physics - University of Genoa
Via Dodecaneso 33 - 16146 Genova, Italy
Phone: +39 10 353 6297 - Fax: +39 10 362 2790
E-Mail: masulli@genova.infn.it

Franco Casalino
Associate Researcher
INFM - National Institute for Matter Physics
Via Dodecaneso 33 - 16146 Genova, Italy
Phone: +39 10 353 6314 - Fax: +39 10 362 2790
E-Mail: casalino@nngl.ge.infn.it

ABSTRACT

The *adaptive fuzzy system* (AFS) is a simple neuro-fuzzy system with the property of function approximation and the capability of learning from examples. The AFS permits one to build a non-parametric classifier able to approximate the Bayes discriminant function, in the large training set limit. Theoretical and experimental results are very satisfactory. The peculiarity of the AFS resides in the fact that, if a linguistic description of classes is available, in addition to a numerical training set, the AFS make it possible to combine the linguistic and statistical approaches to pattern recognition.

Keywords: fuzzy systems, Bayes classifiers, learning techniques.

1. Introduction

In the last few years, some interesting theoretical results on some feedforward connectionist systems have been obtained. In particular, it is has been shown that multilayer perceptrons (MLPs), radial basis function networks (RBFs), and fuzzy systems are able to perform function approximation[5,21,12,9,15] . Moreover, it has been demonstrated that a classifier based on an MLP can approximate the Bayes optimal discriminant function, for suitable choices of the cost function to be approximated during the training phase, and in the large training set limit[22,7,11,18,1] . As pointed out by Ruck et al.[22] , this property is not based on any assumption about the particular feedforward system used; only the system's function approximation capability is required.

In this paper, we present a simple neuro-fuzzy system, previously studied in[10,15] , with the function approximation property and the capability of learning from examples. In the following we shall refer to this fuzzy system as an *adaptive fuzzy system* (AFS). The AFS permits one to build a non-parametric classifier able to approximate the Bayes discriminant function. In this paper we describe of this classifier.

In the next section the AFS is presented. In Section 3, we show the relation between the parameter estimation problem and the Bayesian classification property. In Section 4, the problem of structure identification is presented. In Section 5, some experimental results are reported and discussed. Conclusions are presented in Section 6.

2. The Adaptive Fuzzy System

The AFS[10,15] is a multi-input-multi-output (MIMO) fuzzy logic system based on the following assumptions: center average defuzzifier, product-inference rule, singleton fuzzifier, and Gaussian membership function. More specifically, if there are K units in the input layer, J fuzzy inference rules and I outputs, the rule activations can be written as:

$$r_j = \prod_k \mu_{jk}(x_k), \tag{1}$$

EUFIT '95, August 28-31,1995
Promenade 9, D-52076 Aachen, Germany

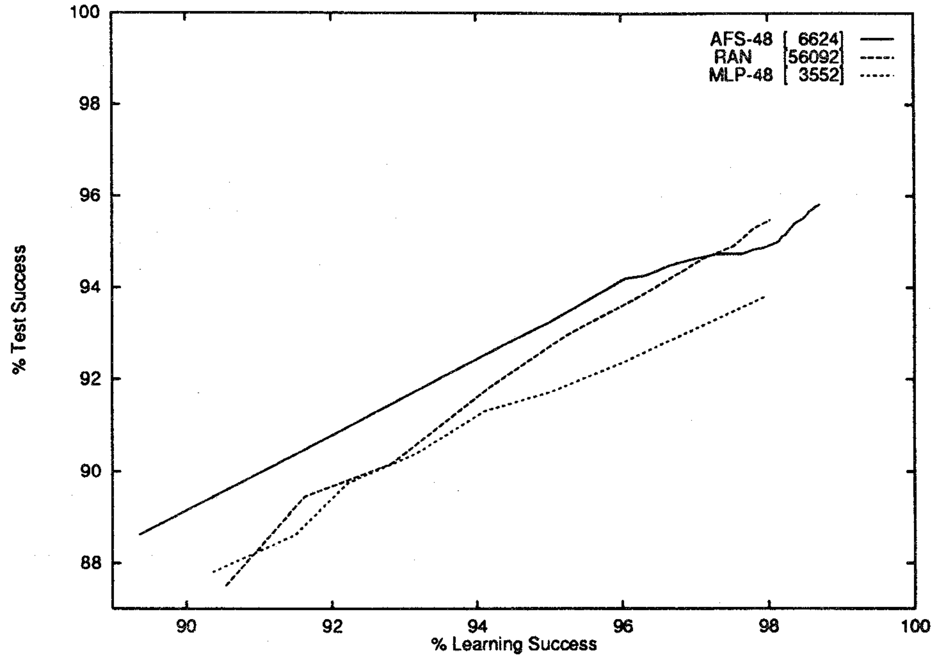

Figure 1: *Comparison among the performances of the MLP, of the AFS and of the RAN. The numbers within brackets refer to the parameters used by each of the three systems.*

The quantity $\mu_{jk}(x_k)$ represents the value of the membership function of the component x_k of the input vector for the j-th rule, and is defined as:

$$\mu_{jk}(x_k) = \exp\left(-\frac{(x_k - m_{jk})^2}{2\sigma_{jk}^2}\right), \tag{2}$$

where m_{jk} and σ_{jk}^2 are the means and the variances. The values of the output units are:

$$y_i = \frac{\sum_j r_j s_{ij}}{\sum_j r_j}, \tag{3}$$

and s_{ij} is the fuzzy singleton of the $j-th$ rule associated with the output y_i.

The AFS can be regarded as a feedforward connectionist system with just one hidden layer whose units correspond to the fuzzy MIMO rules.

In[9,15], on the basis of the Stone-Weierstrass Theorem[23] the Universal Approximation Theorem was demonstrated, that guarantees that an AFS can perform function approximation at an assigned precision. As is well known, similar results on function approximation have been obtained by other feedforward connectionist systems, such as MLPs and RBF networks[5,21].

The AFS can be identified both by exploiting the linguistic knowledge available (***structure identification problem***)[13] and by using the information contained in a data set (***parameter estimation problem***)[13].

In the next sections, we shall consider AFS applications within the context of supervised pattern recognition.

3. AFS Parameter Estimation and the Asymptotic Bayesian Classification Property

A purely "fuzzy" classifier can be realized by training an AFS by an algorithm able to find the values of the parameters (or *weights*) that minimize a suitable cost function, like the *mean square error* (MSE):

$$MSE = \frac{\sum_{i,n}(y_i^n - t_i^n)^2}{N}, \tag{4}$$

where N is the size of the training set, $\mathbf{y}^n = (y_i^n)$ is the network output, and $\mathbf{t}^n = (t_i^n)$ is the n-th label of the associative pair of the training set. The components of $\mathbf{t}^n$ are defined as follows:

$$t_i = \begin{cases} 1 & \text{if the pattern belongs to class } i, \\ 0 & \text{otherwise.} \end{cases} \tag{5}$$

Rules	Parameters	$(\%L, \%T)$	epochs	(%L1, %T1)	Time (min)
24	3312	(97.55, 94.00)	30	(91.0, 88.0)	3.5
32	4416	(98.05, 95.00)	30	(89.0, 87.0)	4.0
48	6624	(98.65, 95.81)	30	(91.5, 88.5)	16.0
64	8832	(98.32, 96.02)	30	(93.5, 92.0)	23.0
128	17664	(98.97, 96.20)	30	(94.5, 93.5)	46.0

Table 1: AFS performances. The columns represent, in order, the number of rules forming the AFS , the number of adaptive parameters, the percentages of learning success (%L) and of test success (%T) at the end of the training phase, the number of epochs required by the training phase, the percentages of learning success (%L1) and of test success (%T1) at end of the first epoch, and the duration of each epoch (in min on a Sun 10/20).

The cost function (4) can be minimized by many different techniques, among which the gradient descent technique, clustering methods[15] , Kalman filters[8] , genetic algorithms[4] , etc. In our experiments, the AFS parameters (i.e., m_{jk}, σ_{jk} and s_{ij}) were obtained by performing a gradient descent with respect to the MSE across the training set. The learning formulas are as follows[10,15] :

$$\Delta s_{ij} = \eta_s [t_i - y_i] \psi_j \tag{6}$$

$$\Delta m_{jk} = \eta_m \psi_j \sum_i [t_i - y_i][s_{ij} - y_i][x_k - m_{jk}]/\sigma_{jk}^2 \tag{7}$$

$$\Delta \sigma_{jk} = \eta_\sigma \psi_j \sum_i [t_i - y_i][s_{ij} - y_i][x_k - m_{jk}]^2/\sigma_{jk}^3 \tag{8}$$

where

$$\psi_j = \frac{\prod_k \mu_{jk}(x_k)}{\sum_j \prod_k \mu_{jk}(x_k)} \tag{9}$$

represents the normalized activation of rule j, and η_s, η_m, and η_σ are the learning rates of s_{ij}, m_{jk}, and σ_{jk}. In our experiments, we adopted an adaptive learning-rate scheme, as proposed in[24] , and we noticed a considerable speed-up of the training phase[2] .

An important theoretical result in the neural networks literature, is the demonstration that the MLP approximates the Bayes optimal discriminant function, in the large training set limit[22,7,11,18,1] . As pointed out by Ruck et al.[22] , the demonstration that MLPs can approximate the Bayes optimal discriminant function is not based on any assumption about the particular feedforward system used; only the system's capability for function approximation is assumed. The extension of this result to the AFS is immediate[16]

4. AFS Structure Identification and Linguistic Knowledge

If a linguistic description of classes is available, in addition to the numerical training set, the AFS permits one to integrate the two kinds of information. On the contrary, if a linguistic description is not available, as in the case with our experiments, the structure identification must be achieved experimentally according to a performance-based criterion.

5. Empirical results

5.1. Data Set and Preprocessing

We used a training and a test sets extracted from the NIST-3 data-base[6] . Each set contained 10,000 associative pairs of segmented handwritten characters. The NIST-3 data-base, distributed on a cd-rom, contains 313389 characters coded as 128×128 binary matrix images and labeled by the corresponding ASCII codes. The preprocessing required the following steps: a character image was extracted from the cd-rom and normalized to a 32×32 binary matrix; a low-pass filter was applied in order to remove some small spots and holes from the image; a shear transform was performed on the character image to straighten the axis joining the first upper-left point of the character image to the last lower-right point; the image was then skeletonized by using a thinning algorithm[19] ; finally, the character representation was transformed into a 64-element vector, each vector element representing the number of black pixels contained in adjacent 4×4 squares. It is worth noting that the obtained character representation exhibits sufficient degrees of invariance to both the scale and small image shifts or rotations.

EUFIT '95, August 28-31,1995
Promenade 9, D-52076 Aachen, Germany

Training set		AFS_1	AFS_2	AFS_3	AFS_4	AFS_5	AFS_6	AFS_7	AFS_8	AFS_9	AFS_{10}
Class	Examples	%	%	%	%	%	%	%	%	%	%
0	1052	100	100	100	100	100	1.615	1.711	1.711	1.711	1.806
1	1134	100	100	100	1.234	1.410	1.410	1.410	1.499	1.499	1.587
2	966	100	100	100	100	100	100.	4.968	5.590	6.004	6.107
3	1059	100	100	100	100	100	100	100	100	3.966	4.438
4	967	100	3.177	1.964	2.378	3.309	3.826	3.826	3.826	3.826	3.826
5	842	100	100	100	100	100	100	100	100	100	11.63
6	948	100	100	100	100	4.00	4.00	4.113	4.113	4.113	4.535
7	1052	100	100	100	100	100	100	100	2.471	2.471	2.471
8	978	0.0	1.87	5.725	6.032	6.748	6.952	7.668	7.873	8.077	8.691
9	1002	100	100	2.295	2.694	2.694	2.894	3.093	3.393	3.493	3.592

Table 2: Training set error rates on 10,000 patterns by ten different AFSs, ranging from 1 to 10 rules.

5.2. Experimental Results 1: Learning and Classification Performances

In[16] a comparison is made between the performances of three connectionist feed-forward classifiers, namely an MLP, a RAN, and an AFS. The resource allocating network (RAN) is a radial-basis function neural network characterized by the growth of its architecture during the training phase[20] .The structure of each of the three systems was made up by 64 input units, 48 hidden nodes, and 10 output units (one for each class to be recognized). As shown in Figure 1, the three nets exhibit similar generalization properties, as we expected from a theoretical standpoint.

The MLP reaches a smaller generalization value than the other two systems. This may be due to the problem of *false positive*, as discussed by Lee.[14] The training phase of the AFS is faster that that of the MLP. On the contrary the RAN is the slowest one during the training phase; this depends on the growth of its architecture up to more than $50,000$ parameters (each of them must be optimized !). Moreover, it is interesting to note that the RAN shows the highest derivative of the Test Success, as compared with the Training Success. This depends on the possibility of allocating new units dynamically during the learning phase for such as system.

Table 1 gives the results of experiments on AFSs with different number of rules. Some results are quite impressive; for instance, in the case of an AFS with 128 rules, a single epoch of 46 min is enough to obtain a percentage of test success %T1 = 93.5.

5.3. Experimental Results 2: Structure Identification and Semantic Phase Transition

Ten different AFS have been trained on the training set described in Section 5.1. The AFSs differ in the number of rules, ranging from 1 up to 10. In Table 2, for each class of digit the numbers of patterns in the training set and the percentage of patterns not correctly recognized by each AFS at end of the training phase, are shown. It worth noting a close relationship between the number of rules in the AFS and the number of classes which are recognized by the AFS: $10-\beta$ classes are not recognized at all for AFS with β rules, while the other classes are well recognized. On the contrary, in the case of AFS_{10} (or of AFSs with more than 10 rules) the error rate is uniformly distributed along all the classes. This behavior of the AFS, that we call *semantic phase transition*, has been confirmed by further series of simulations, and and is not usual in other feed-forward connectionist systems. The semantic phase transition phenomenon, gives a lower bound to the AFS structure: the AFS must contain a number of rules at least equal to the number of classes to be discriminated.

6. Conclusions

In this paper, we have studied the behavior of a classifier based on an AFS. The AFS permits one to build a non-parametric classifier able to approximate the Bayes discriminant function, in the large training set limit. Theoretical and experimental results are very satisfactory.

As discussed in the previous sections, the AFS shares several features with the MLP: a feed-forward architecture, learning from a numerical data set by a gradient descent technique, function approximation capabilities, and an asymptotic approximation to the Bayes classifier. On the contrary some specific behavior, such as the semantic phase transition can be ascribed to the derivation of the studied system from a fuzzy logic framework. The peculiarity of the AFS resides in the fact that, if a linguistic description of classes is available, in addition to a numerical training set, the AFS permits the linguistic and statistical approaches to pattern recognition to be combined.

7. Acknowledgments

This work was supported by grants from CNR-Progetto Strategico Reti Neurali, GNCB-CNR, INFM, and MURST. Part of this work was carried out in the Summer of 1994, while F. Masulli was a Visiting Scientist at the International Computer Science Institute at Berkeley (USA). We thank Massimo Riani, Alessandro Sperduti, and Fabrizio Vannucci for helpful discussions.

8. References

1. E. Barnard, F. Kanaya, and S. Miyake. Comments on 'Bayes statistical behavior and valid generalization of pattern classifying neural networks' (with reply). *IEEE Transactions on Neural Networks*, 3:1026–7, 1992.

2. F. Casalino. Fuzzy systems for handwriting recognition. Laurea Thesis in Computer Science, University of Genoa, Genova - Italy, 1993. in Italian.

3. F. Casalino, F. Masulli, Sperduti, and F. A., Vannucci. Semantic phase transition in a classifier based on an adaptive fuzzy system. In *Proceedings of the Third IEEE International Conference on Fuzzy Systems, IEEE-FUZZ94*, pages 808- 812, Orlando, Florida, 1994.

4. R. Caviglia. Soft-computing methods of time series forecasting. Laurea Thesis in Computer Science, University of Genoa, Genova - Italy, 1994. in Italian.

5. G. Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals, and Systems*, 2:303–314, 1989.

6. M.D. Garris and R.A. Wilkinson. *NIST Special Database3 Handwritten Segmented Characters*. National Institute of Standard and Technology, Gaithesburg, MD , USA, 1992.

7. J. Hampshire and B. Pearlmutter. Equivalence proofs for multi-layer perceptron classifiers and the Bayesian discriminant function. In D.S. Touretzky, G. Hinton, and T. Sejnowski, editors, *Connectionist models : Proceedings of the 1990 Summer school*, pages 13–19, Denver, 1990. Morgan Kaufmann, San Mateo.

8. J.S.R. Jang. ANFIS: Adaptive-network-based fuzzy inference system. *IEEE Trans. on Systems, Man, and Cybernetics*, 23:655–684, 1993.

9. C.C. Jou. On the mapping capabilities of fuzzy inference systems. In *IJCNN International Joint Conference on Neural Networks*, volume 2, pages 703–713, Baltimore, MD, USA, 7-11 June 1992, 1992. IEEE, New York, NY.

10. C.C. Jou. Comparing learning performance of neural networks and fuzzy systems. In *IEEE International Conference on Fuzzy Systems*, pages 1028–1033, San Francisco, 1993. IEEE, New York, NY.

11. F. Kanaya and S. Miyake. Bayes statistical behavior and valid generalization of pattern classifying neural networks. *IEEE Transactions on Neural Networks*, 2:471–475, 1991.

12. B. Kosko. Fuzzy systems as universal approximators. In *IEEE International Joint Conference on Fuzzy Systems*, pages 1153–1162, San Diego, CA, USA, 8-12 March 1992, 1992. IEEE, New York, NY, USA.

13. C.C. Lee. Fuzzy logic in control systems: fuzzy logic controller. I. *IEEE Transactions on Systems, Man and Cybernetics*, 20:404–418, 1990.

14. Y. Lee. Handwritten digit recognition using k nearest-neighbor, radial-basis function, and backpropagation neural networks. *Neural Computation*, 3:440–449, 1991.

15. Wang Li-Xing. *Adaptive Fuzzy Systems and Control*. Prentice Hall, Englewood Cliffs, New Jersey, 1994.

16. F. Masulli. Bayesian classification by feedforward connectionist systems. In F. Masulli, P. G. Morasso, and A. Schenone, editors, *Neural Networks in Biomedicine - Proceedings of the Advanced School of the Italian Biomedical Physics Association - Como, Italy , 1993*, pages 145–162, Singapore, 1994. World Scientific.

17. F. Masulli, F. Casalino, and F. Vannucci. Bayesian properties and performances of adaptive fuzzy systems in pattern recognition problems. In M. Marinaro and P.G. Morasso, editors, *Proceedings of the European Conference on Artificial Neural Networks, ICANN-94*, pages 189–192, Sorrento, Italy, 1994. Springer.

18. S. Miyake and F. Kanaya. A neural network approach to a bayesian statistical decision problem. *IEEE Transactions on Neural Networks*, 2:538–540, 1991.

19. Theo Pavlidis. *Algorithms for Graphics and Image Processing*. Springer-Verlag, 1982.

20. J. Platt. A resource-allocating network for function interpolation. *Neural Computation*, 3:213–225, 1991.

21. T. Poggio and F. Girosi. Regularization algorithms for learning that are equivalent to multilayer networks. *Science*, 247:978–982, 1990.

22. D.W. Ruck, S.K. Rogers, M. Kabrisky, M.E. Oxley, and B.W. Suther. The multilayer perceptron as an approximation to a Bayes optimal discriminant function. *IEEE Transactions on Neural Networks*, 1:296–298, 1990.

23. W. Rudin. *Principles of Mathematical Analysis*. McGraw-Hill, Inc., New York, 1960.

24. T.P. Vogl, J.K. Mangis, A.K. Rigler, W.T. Zink, and D.L. Alkon. Accelerating the convergence of the back-propagation method. *Biological Cybernetics*, 59:257–263, 1988.