Working Paper 94-03
Statistics and Econometrics Series 02
January 1994

Departamento de Estadística y Econometría
Universidad Carlos III de Madrid
Calle Madrid, 126
28903 Getafe (Spain)
Fax (341) 624-9849

DIFFUSE PATTERN LEARNING WITH FUZZY ARTMAP AND PASS

Jorge Muruzábal and Alberto Muñoz*

Abstract

Fuzzy ARTMAP is compared to a classifier system (CS) called PASS (predictive adaptive sequential system). Previously reported results in a benchmark classification task suggest that Fuzzy ARTMAP systems perform better and are more parsimonious than systems based on the CS architecture. The tasks considered here differ from ordinary classificatory tasks in the amount of output uncertainty associated with input categories. To be successful, learning systems must identify not only correct input categories, but also the most likely outputs for those categories. Performance under various types of diffuse patterns is investigated using a simulated scenario.

Key Words
Diffuseness; Competitive learning; Prediction.

*Muruzábal, Departamento de Estadística y Econometría, Universidad Carlos III de Madrid; Muñoz, Departamento de Física Aplicada, Universidad de Salamanca.

Diffuse pattern learning with Fuzzy ARTMAP and PASS[1].

Jorge Muruzábal
Department of Statistics and Econometrics
University Carlos III, 28903 Getafe, Spain

and

Alberto Muñoz
Department of Applied Physics
University of Salamanca, 37007 Salamanca, Spain

Abstract

Fuzzy ARTMAP is compared to a classifier system (CS) called PASS (predictive adaptive sequential system). Previously reported results in a benchmark classification task suggest that Fuzzy ARTMAP systems perform better and are more parsimonious than systems based on the CS architecture. The tasks considered here differ from ordinary classificatory tasks in the amount of output uncertainty associated with input categories. To be successful, learning systems must identify not only correct input categories, but also the most likely outputs for those categories. Performance under various types of diffuse patterns is investigated using a simulated scenario.

1 Introduction

Carpenter, Grossberg, Markuzon, Reynolds and Rosen [3] present results in a letter recognition task indicating that Fuzzy ARTMAP systems perform better and use fewer resources than the classifier system (CS) schemes considered by Frey and Slate [4]. In this paper, we propose various pattern learning tasks and analyze the behavior of Fuzzy ARTMAP and a different CS implementation called PASS (predictive adaptive sequential system) [6]. The tasks considered here involve learning the association between a binary input vector and an output scalar, but they differ from ordinary classificatory tasks in the amount of output uncertainty associated with input categories. Thus, these patterns reflect more statistical regularities than function-like assignments.

[1] Correspondence to the first author. Support from DGICYT grant PB92-0246 (Spain) and CICYT grant TIC93-0702-C02-02 (Spain) is gratefully acknowledged.

The patterns we consider can be made increasingly diffuse in various ways. We first focus on the effect of raising the output uncertainty associated with input categories. High-uncertainty patterns are interesting in that they reflect situations in which the output is read with considerable noise and/or the chosen input vector misses important variance-explaining features or predictors. Patterns can also be diffuse in the sense of presenting a low signal-to-noise ratio. This will be of interest when only a relatively small fraction of the data is expected to contain useful regularities. Finally, we consider patterns with rather general (large) input categories, that is, patterns where only a few input coordinates are actually relevant to determine the most likely output. To gain some initial understanding of the strengths and weaknesses of the two families of systems, each of these diffuse patterns will be studied separately in this paper: the mixed case in which resources must be shared with more obvious (sharper) patterns is postponed for future work.

The organization is as follows. Section 2 introduces the data-generating mechanism and sets up the language to define each type of diffuseness. Sections 3 and 4 briefly summarize the main aspects of the algorithms. Section 5 reports on the empirical results and presents some preliminary conclusions.

2 The data source

We consider a simple (stochastic) pattern learning task in which data pairs (x,y) are independently drawn from a mixture distribution on the joint sampling space $\{0.1\}^n \times (0.1)$. This distribution involves $c \geq 1$ elementary components or patterns specifying particular regularities to be observed from time to time by the systems. Each pattern is defined by a triple (θ, η, υ), where $0 < \theta \leq 1$ is the mixing proportion, η is a schema (a subspace formed by fixing some coordinates), and υ is a probability distribution on $(0,1)$. A minimum coherency requirement is enforced by considering disjoint η_i only. which also permits straightforward calculation of the optimal level of performance attainable by the systems. For simplicity, υ distributions are always taken from the beta family, so they are identified by their parameters α and β. Further. we only consider unimodal densities $(\alpha, \beta \geq 1)$ here.

If possible. the previous set of elementary patterns is automatically augmented with *noise*; noise is devised as the triple $(\theta_0, \eta_0, \upsilon_0)$, where u_0 denotes the uniform distribution over the unit interval and θ_0 and η_0 represent respectively the complement to 1 of the sum of θ_i and the complement of the union of η_i. Thus, the resulting distribution is a particular case of the "signal vs. noise" paradigm.

Sampling proceeds then as follows: a triple is selected at random according to the relative frequencies θ_i, $(i=0,1, ..., c)$, x is obtained by either randomly filling up the undefined coordinates in η_i $(i\geq1)$ or simply choosing a string at random from η_0, and y is taken as an independent realization of υ_i. A wide array of situations can be obtained by varying the amount of noise, the number of elementary patterns, the specificity of schemata η and the sharpness of distributions υ. For example, using the standard "don't care" or "wildcard" symbol #, the pattern defined by $\theta=.7$, $\eta=(00011\#)$ and $\alpha=8$, $\beta=1$ is very different from the (rather diffuse) pattern $\theta=.02$, $\eta=(0\#\#\#\#\#)$ and $\alpha=2$, $\beta=1$.

3 Fuzzy ARTMAP

ARTMAP systems are based on the long-introduced Adaptive Resonance Theory (ART) in neural network modelling [1,2]. Given the present nature of the data, we consider Fuzzy ARTMAP systems only [3].

In a nutshell. Fuzzy ARTMAP systems learn by simultaneously (i) establishing suitable categories in both input and output space (tasks carried out within the so-called A and B modules respectively), and (ii) linking input and output categories according to joint occurrence and predictive success (the linkages being stored in a special unit called the map field or AB module). Modules are made out of fields and fields are made out of neurons (nodes). All categorization and learning are achieved by sequentially modifying three sets of neuron weights, one in each module. The number of weights in the A and B modules are system parameters determining the number and dimension of the weights in the AB module. During training, both x and y are provided as input to the A and B modules, which causes activation to flow from the excited neurons (categories) in A and B into AB, and then (potentially) back from AB into A (see below). During testing, a given input vector typically activates (predicts) a single category in the B and AB modules.

In ARTMAP systems, data can be processed with either natural or complement coding [2,3]: if natural coding is used, a data item d is processed "as is", otherwise, d is augmented with d^c, the (coordinatewise) complement to 1. Thus, if d is a n-dimensional binary vector x, then the system actually works on a 2n-dimensional binary vector containing n ones and n zeros, whereas, if d is a scalar y, then the vector (y,1-y) is supplied. Natural coding introduces an asymmetry in the treatment of zeros and ones which does not correspond with the symmetric role played in the task of interest here. Also, under complement coding, the weight vectors in the A module can be related to schemata like the η_i in section 2.

Within a given A or B module, the tendency of the system to commit new neurons (as opposed to using previously commited neurons) is controlled by the so-called vigilance parameter $0 \leq \rho \leq 1$. When ρ is large, the system tends to commit neurons more easily: otherwise, relatively fewer (and therefore larger) categories are constructed. Vigilance parameters in the A and B modules are denoted by ρ_a and ρ_b respectively.

During training in the A module, a first decision is made on the basis of a similarity measure between the binary input I and the existing categories u_j ; this is defined as $\frac{|I \wedge u_j|}{b+|u_j|}$. where b>0 is a system parameter, $\wedge$ is the fuzzy AND operator (defined as coordinatewise minimum), and $| \, |$ stands for the sum of coordinates of its argument. The winning category maximizes this measure. Parameter b controls the degree to which categories that match I exactly tend to win over partial matches.

One of the peculiarities of ARTMAP systems is the fact that the winning category is selected only if a second similarity measure, defined as $\frac{|I \wedge u_j|}{|I|} = \frac{|I \wedge u_j|}{n}$, surpasses ρ_a. Otherwise. the best candidates from the first test are tried out in turn until one succeeds or a fresh neuron is committed.

Let J and K denote respectively the (overall) winning categories in modules A and B. When learning is triggered. the associated weight vector in A (say) is updated as $u_J^{(new)} = (1-\lambda_a)\, u_J^{(old)} + \lambda_a\, I \wedge u_J^{(old)}$ ($0<\lambda_a \leq 1$; all weight vectors are initialized with ones). Thus. under "fast-learning" ($\lambda_a=1$), the designed schema u_J in module A generalizes to the schema $I \wedge u_J$. Under the "fast-learn/slow-recode" option (recommended for noisy data). $\lambda_a=1$ only when a node is first committed, thereafter it is fixed and strictly lower than 1. As regards module B, $\lambda_b=1$ throughout this paper.

If there is disagreement during training between the system's prediction (determined by J) and the observed response category (K). the system revises its prediction by raising the vigilance parameter ρ_a by the minimal amount needed so J no longer passes the second similarity test (and is therefore turned off). The process continues in the A module as described earlier: new winners J' are tested until perhaps an agreement is reached in AB, in which case new learning occurs. Parameter ρ_a returns in either case to its "baseline" value before the next training pair is presented.

During testing, input is fed into the A module, where two things may happen: either a winner J is excited or not. If it is, then the system's predicted category in output space may be read off the associated AB weight vector; otherwise, no prediction is offered and the system issues an "I don't know" flag. Of course, a high baseline value for ρ_a increases the frequency of nonresponse during testing.

4 PASS

We now turn our attention to PASS [6]. Like "fast-learning" Fuzzy ARTMAP, PASS also expresses its predictions as links between schemata in input space and certain regions in output space. yet it does not build categories in output space, only in input space. In output space, it constructs a conditional probability distribution given an input vector. In this paper, however. we simply replace this distribution by its (bounded) support. so we can compare predictive success on the same grounds.

Like any other classifier system (CS) [5], PASS consists essentially of performance system and learning operators. The performance system is based on an unstructured population of elementary predictive rules called classifiers. Each classifier in PASS has the form

IF s THEN PREDICT d
(WITH STRENGTH S, RECALLING E),

where s is a schema, d is a subinterval of (0,1) of bounded length, S is a scalar quantity called strength. and E is a small list of observed pairs (x,y). Strength reflects the classifier's previous success. The exception list E contains a few recent cases where the classifier proved wrong: heuristic operators act on these lists when they reach a (small) threshold length. with the result that new classifiers are formed and/or old ones modified [6]. Thus. PASS classifiers incorporate the additional structures S and E representing two forms of memory not available in ARTMAP systems.

Learning in PASS is "slow" in the sense that no decisions on schemata or predictions are made on the basis of single data items, every action is based on accumulated data. On the other hand, classifiers may be and often are discarded altogether whenever the system decides to try some other alternatives. As in Fuzzy

ARTMAP, no overall check for consistency or completeness is contemplated, and no provision for the emergence of structure within the population is made explicitly.

Predictions in both PASS and Fuzzy ARTMAP are based on competitive processes (called auctions in CS parlance). A number of winners are selected and predictions follow the opinion of these winners. In PASS, however, winners are considered simultaneously, and the outcome of the competition is stochastic. The elementary predictions read off the winners are combined (weighted by strength) to yield the system's predictive distribution (from which both bounded-length convex and non-convex predictive supports can be formed). PASS enjoys then a potentially higher level of communication among classifiers, at the price that decisions do not necessarily reflect the "best" knowledge currently available in the system.

As in Fuzzy ARTMAP, the proportion of specified coordinates or specificity D of competing schemata participates explicitly in the auction, this time together with S. PASS adheres to the traditional auction in which matched classifiers place bids $B=\kappa SD^{\gamma}$. and effective bids $B^{*}=B^{\phi}D^{\psi}$. the probability of winning being then proportional to B^{*} (all system parameters are nonnegative). This auction can be made highly dependent on D alone. yet competition is usually (and here in particular) restricted to perfect-matching schemata. The number of winners to draw from the list of matched classifiers. say m, is a critical parameter controlling both the amount of mixing prior to prediction and the relative frequency at which classifiers are tested out.

The auction is one but the situations in PASS where stochasticity is present. A version of the genetic algorithm performs in the background, though it has not been found to contribute much in its present form [6]. The set of procedures acting on the exception lists contribute more to learning and are partly randomized as well. Reward itself is stochastic: the strength of a "correct" *winning* classifier is updated as $S^{(new)}=(1-tax)\,S^{(old)}-B+R$, where tax is a small fraction of strength usefully collected from matched and not matched classifiers at every time step, and $R=+R_0$ or $-\pi S$ depending on its prediction's "coverage" of the associated pattern distribution(s) υ. The recurring presence of stochasticity in PASS is in sharp contrast with the strict determinism found in Fuzzy ARTMAP.

5 Experimental results

We now present a summary of our experiments. We have tried the algorithms just described in three different pattern learning tasks. In each task, we provide the systems with a training sample of size 500. Performance is measured as the proportion of correct predictions on an independent test set of 500 observations (no *voting strategies*, as suggested in [2], are considered, although they are probably quite powerful here as well). Data are processed using complement coding in Fuzzy ARTMAP and natural coding in PASS. While no further learning occurs in either system during testing (no new categories or classifiers are created nor old ones modified), strength continues to be updated as usual.

For the sake of comparison, the number of neurons in module A of Fuzzy ARTMAP and the number of classifiers in PASS are both set to a maximum of μ=50. To understand the full effect of this constrain is an interesting research area in both systems: in ART-based systems, vigilance parameters act critically on the number of categories finally created by the system, while dynamic manipulation of μ (along with m) may promote some form of "knowledge condensation" in CS and seems useful in PASS (see below). Note that fixing μ does not make the systems equally demanding, as classifiers in PASS require additional memory to implement their exception lists. Also, the bound set by μ in PASS plays more the role of an attractor rather than a hard limit.

The systems must also be granted the same scope in their predictive effort. Both Fuzzy ARTMAP and PASS include system parameters that bound the length of their predictions. For our current purpose, a bound of .3 seems appropriate and is used in all runs (the most direct way, though not the only one, of achieving this in Fuzzy ARTMAP is to set ρ_b to .7, preserving the original spirit of the architecture). This bound determines the maximum level of performance attainable by either system at any given task, which provides a useful reference value.

The versions we have investigated differ in system parameters ρ_a, b and λ_a in Fuzzy ARTMAP, and m, μ and the activity rate of the rule-generating procedures (ω) in PASS. Results are reported below on the performance of "slow-recode" Fuzzy ARTMAP with (baseline) ρ_a between 0 and .2, b between .05 and .1 and λ_a between .05 and .1. Auction/reinforcement parameters in PASS are $\phi=\psi=\gamma=1$, $\pi=10\%$, tax=1%, $\kappa=.1$, $R_0=150$ and an initial strength of 200. Other system parameters are kept at values discussed in previous work [6], although the version used here is different in that (i) the initial strength of most newly-created classifiers is now set to the current median of the population, and (ii) certain (namely, `explain-N`) modifications replacing the classifiers that suggested them).

Training in PASS was manually split into two epochs: for the first two presentations (cycles), m=5, μ=50 and ω was "high", while for the remaining two, m=3, μ=40 and ω was "low" (in test mode, m=3 and ω=0). Four is a "small" number of presentations for PASS, for the system usually benefits from an additional four or six cycles. In contrast, the number of training cycles in Fuzzy ARTMAP varied between 10 and 40. Indeed, Fuzzy ARTMAP's high speed of processing allows for a much larger number of training cycles in considerably less time. However, the ultimate comparison in terms of processing time is hopeless until parallel versions of the algorithms are confronted.

The systems were tested in three (toy) tasks described in Table 1. Task I presents six high-uncertainty patterns intertwined with noise at a 10% rate; since patterns occur relatively often and their schemata are sharply defined, the main difficulty resides in the slight departure from uniformity. The same schemata define task II, except now noise occurring at a 52% rate makes it hard to detect the otherwise obvious departure from uniformity. In task III, moderate noise joins a moderate departure from uniformity, but the number of irrelevant parameters is large.

Table 1. The problems

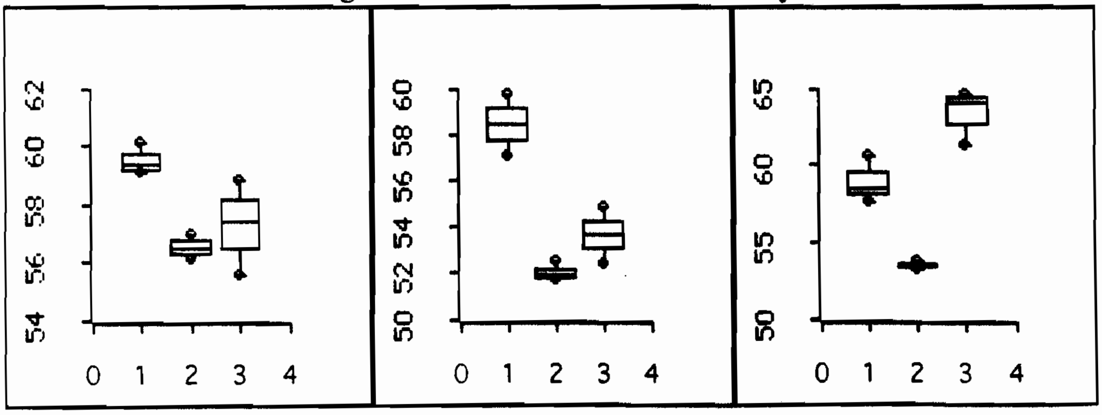
(a) Task I (optimal performance rate=62.0%)
θ=.15, η=(0##1#00#0##1), α=1, β=3
θ=.15, η=(1###01###010), α=1, β=3
θ=.15, η=(#00##10##1#1), α=1, β=3
θ=.15, η=(##01#11#0#0#), α=3, β=1
θ=.15, η=(0##01#1#11##), α=3, β=1
θ=.15, η=(1##1#0#10##1), α=3, β=1

(b) Task II (optimal performance rate=59.6%)
θ=.08, η=(0##1#00#0##1), α=1, β=7
θ=.08, η=(1###01###010), α=1, β=7
θ=.08, η=(#00##10##1#1), α=1, β=7
θ=.08, η=(##01#11#0#0#), α=7, β=1
θ=.08, η=(0##01#1#11##), α=7, β=1
θ=.08, η=(1##1#0#10##1), α=7, β=1

(c) Task III (optimal performance rate=72.6%)
$\theta=.2$, $\eta=$(0#######0###0##), $\alpha=1$, $\beta=5$
$\theta=.2$, $\eta=$(0#######1######), $\alpha=5$, $\beta=1$
$\theta=.2$, $\eta=$(1###########1##), $\alpha=1$, $\beta=5$
$\theta=.2$, $\eta=$(1#######1###0##), $\alpha=5$, $\beta=1$

We found "fast-learning" Fuzzy ARTMAP not to be competitive in these tasks, which is in contrast with the previously reported success in less diffuse problems [3]. We also found "slow-recode" Fuzzy ARTMAP and PASS to reach comparable levels of performance in tasks I and II (see figure 1), a surprising fact given the nature of the architectures and learning mechanisms. The high sensitivity of Fuzzy ARTMAP with respect to ρ_b is also manifest. In Task III, PASS proves somewhat superior. Fuzzy ARTMAP's relative lack of success in Task III suggests a type of regularity that it might find hard to detect in general. In the same direction, we plan to investigate a "contaminated" pattern learning task where some training x vectors have crucial coordinates flipped.

Figure 1. Performance summary



Each boxplot is based on five independent runs; while the same version of PASS was used in the three tasks, Fuzzy ARTMAP parameters were slightly tuned in each case. Frames correspond (left to right) to tasks in Table 1. Within each frame, the first two boxplots correspond to "slow-recode" ARTMAP con ρ_b=2/3 and ρ_b=.7 respectively, the third corresponds to PASS. The first and last boxplots in each frame are not of course really comparable, as they refer to different optima).

We also note that neither system is completely successful at recovering all defining schemata. It appears that PASS categories tend to be larger than needed (sometimes ingeniously exploiting "hidden" aspects of the set of patterns), whereas Fuzzy ARTMAP categories tend to be finer (under the "slow-recode" option, category interpretation is of course complicated by weight coordinates far from 0 or 1).

We conclude by pointing out some further directions for research. It seems to us, for example, that the joint election of several winners, along with the associated combination of beliefs, may improve Fuzzy ARTMAP's performance dramatically. It would also be very interesting to develop adaptive schemes for ρ_b (and ρ_a). As regards PASS, automatic manipulation of both the exploration rate (ω) and the number of winners (m) during training seems crucial to attain additional stability and convergence; natural heuristics to guide such manipulations may be obtained from the slope of the learning curve.

References

[1] Carpenter, G. A. and Grossberg, S. (1987). A massively parallel architecture for a self-organizing neural pattern recognition machine. Computer Vision, Graphics, and Image Processing, Vol. 37, pp. 54-115.

[2] Carpenter, G. A., Grossberg, S., and Reynolds, J. H. (1991). ARTMAP: supervised real-time learning and classification of nonstationary data by a self-organizing neural network. Neural Networks, Vol. 4, pp. 565-588.

[3] Carpenter, G. A., Grossberg, S., Markuzon, N., Reynolds, J. H. and Rosen, D. B. (1992). FUZZY ARTMAP: a neural network architecture for incremental supervised learning of analog multidimensional maps. IEEE Transactions on Neural Networks, Vol. 3, No. 5, pp. 698-713.

[4] Frey and Slate (1991). Letter recognition using Holland-style adaptive classifiers. Machine Learning, Vol. 6, pp. 161-182.

[5] Holland, J. H., Holyoak, K. J., Nisbett, R. E., & Thagard, P. R. (1986). Induction: Processes of Inference, Learning and Discovery. Cambridge, MA: MIT Press.

[6] Muruzábal, J. (1993). PASS: a simple classifier system for data analysis. Tech. Rep. 93-20, Statistics and Econometrics Dept., University Carlos III, Madrid, Spain.