SUPERVISED LEARNING BY ADAPTIVE RESONANCE NEURAL NETWORKS

Gail A. Carpenter†, Stephen Grossberg‡, Natalya Markuzon§, John H. Reynolds¶, and David B. Rosen¶

Center for Adaptive Systems and Department of Cognitive and Neural Systems, Boston University, 111 Cummington Street, Boston, Massachusetts 02215 USA

1. Introduction

ARTMAP is a class of neural network architectures that perform incremental supervised learning of recognition categories and multidimensional maps in response to input vectors presented in arbitrary order. The first ARTMAP system (Carpenter, Grossberg, and Reynolds, 1991) was used to classify binary vectors. This article describes a more general ARTMAP system that learns to classify analog as well as binary vectors (Carpenter, Grossberg, Markuzon, Reynolds, and Rosen, 1992). This generalization is accomplished by replacing the ART 1 modules (Carpenter and Grossberg, 1987) of the binary ARTMAP system with Fuzzy ART modules (Carpenter, Grossberg, and Rosen, 1991). Where ART 1 dynamics are described in terms of set-theoretic operations, Fuzzy ART dynamics are described in terms of fuzzy set-theoretic operations (Zadeh, 1965). Hence the new system is called Fuzzy ARTMAP. Also introduced is an ARTMAP *voting strategy*.

† Supported in part by British Petroleum (89-A-1204), DARPA (AFOSR 90-0083), the National Science Foundation (NSF IRI 90-00530) and the Office of Naval Research (ONR N00014-91-J-4100).

‡ Supported in part by the Air Force Office of Scientific Research (AFOSR 90-0175), DARPA (AFOSR 90-0083) and the Office of Naval Research (ONR N00014-91-J-4100).

§ Supported in part by National Science Foundation (NSF IRI 90-00530) and British Petroleum (89-A-1204).

¶ Supported in part by DARPA (AFOSR 90-0083).

Acknowledgements: The authors wish to thank Cynthia E. Bradford for her valuable assistance in the preparation of the manuscript.

This voting strategy is based on the observation that ARTMAP fast learning typically leads to different adaptive weights and recognition categories for different orderings of a given training set, even when overall predictive accuracy of all simulations is similar. The different category structures cause the set of test set items where errors occur to vary from one simulation to the next. The voting strategy uses an ARTMAP system that is trained several times on input sets with different orderings. The final prediction for a given test set item is the one made by the largest number of simulations. Since the set of items making erroneous predictions varies from one simulation to the next, voting cancels many of the errors. Further, the voting strategy can be used to assign probability estimates to competing predictions given small, noisy, or incomplete training sets.

Simulations illustrate Fuzzy ARTMAP performance as compared to benchmark back propagation and genetic algorithm systems. In all cases, Fuzzy ARTMAP simulations lead to favorable levels of learned predictive accuracy, speed, and code compression in both on-line and off-line settings. Two simulations are described below. Fuzzy ARTMAP is also easy to use. It has a small number of parameters, requires no problem-specific system crafting or choice of initial weight values, and does not get trapped in local minima.

Each ARTMAP system includes a pair of Adaptive Resonance Theory modules (ART_a and ART_b) that create stable recognition categories in response to arbitrary sequences of input patterns (Figure 1). During supervised learning, the ART_a module receives a stream $\{\mathbf{a}^{(p)}\}$ of input patterns and ART_b receives a stream $\{\mathbf{b}^{(p)}\}$ of input patterns, where $\mathbf{b}^{(p)}$ is the correct prediction given $\mathbf{a}^{(p)}$. These modules are linked by an associative learning network and an internal controller that ensures autonomous system operation in real time. The controller is designed to create the minimal number of ART_a recognition categories, or "hidden units," needed to meet accuracy criteria. It does this by realizing a Minimax Learning Rule that enables an ARTMAP system to learn quickly, efficiently, and accurately as it conjointly *minimizes* predictive error and *maximizes* predictive generalization. This scheme automatically links predictive success to category size on a trial-by-trial basis using only local operations. It works by increasing set theory can be incorporated naturally into ART systems. For example, the intersection ($\cap$) operator that describes ART 1 dynamics is replaced by the AND operator ($\wedge$) of fuzzy set theory (Zadeh, 1965) in the choice, search, and learning laws of ART 1 (Figure 2). Especially noteworthy is the close relationship between the computation that defines fuzzy subsethood (Kosko, 1986) and the computation that defines category choice in ART 1. Replacing operation $\cap$ by operation $\wedge$ leads to a more powerful version of ART 1. Whereas ART 1 can learn stable categories only in response to binary input vectors, Fuzzy ART can learn stable categories in response to either analog or binary input vectors. Moreover, Fuzzy ART reduces to ART 1 in response to binary input vectors.

In Fuzzy ART, learning always converges because all adaptive weights are monotone nonincreasing. Without additional processing, this useful stability property could lead to the unattractive property of category proliferation as too many adaptive weights converge to zero. A preprocessing step, called complement coding, uses on-cell and off-cell responses to prevent category proliferation. Complement coding normalizes input vectors while preserving the amplitudes of individual feature activations. Without complement coding, an ART category memory encodes the degree to which critical features are consistently present in the training exemplars of that category. With complement coding, both the degree of absence and the degree of presence of features are represented by the category weight vector. The corresponding computations employ fuzzy OR ($\vee$, maximum) operators, as well as fuzzy AND ($\wedge$, minimum) operators.

2. Simulation: Circle-in-the-Square

The circle-in-the square problem requires a system to identify which points of a square lie inside and which lie outside a circle whose area equals half that of the square. This task was specified as a benchmark problem for system performance evaluation in the DARPA Artificial Neural Network Technology (ANNT) Program (Wilensky, 1990). Wilensky examined the performance of 2–n–1 back propagation systems on this problem. He studied systems where the number (n) of hidden units ranged from 5 to 100, and the corresponding number of weights ranged from 21 to 401. Training sets ranged in size from 150 to 14,000. To avoid over-fitting, training was stopped when accuracy on the training set reached 90%. This criterion level was reached most quickly (5,000 epochs) in systems with 20 to 40 hidden units. In this condition, approximately 90% of test set points, as well as training set points, were correctly classified.

Fuzzy ARTMAP performance on this task in 1 training epoch is illustrated in Figures 3 and 4. As training set size increased from 100 exemplars (Figure 3(a)) to 100,000 exemplars (Figure 3(d)) the rate of correct test set predictions increased from 88.6% to 98.0% while the number of ART_a category nodes increased from 12 to 121. Each category node j required four learned weights $\mathbf{w}_j^a$ in ART_a plus one map field weight $\mathbf{w}_j$ to record whether category j predicts that a point lies inside or outside the circle. Thus, for example, 1-epoch training on 100 exemplars used 60 weights to achieve 88.6% test set accuracy. The map can be made arbitrarily accurate provided the number of ART_a nodes is allowed to increase as needed.

Figure 3 shows how a test set error rate is reduced from 11.4% to 2.0% as training set size increases from 100 to 100,000 in 1-epoch simulations. Test set error rate can be further reduced if exemplars are presented for as many epochs as necessary to reach 100% accuracy on the training set. The ARTMAP voting strategy provides a third way to eliminate test set errors. Recall that the voting strategy assumes a fixed set of training exemplars. Before each individual simulation the input ordering is randomly assembled. After each simulation the prediction of each test set item is recorded. Voting selects the outcome predicted by the largest number of individual simulations. In case of a tie, one outcome is selected at random. The number of votes cast for a given outcome provides a measure of predictive confidence at each test set point. Given a limited training set, voting across a few simulations can improve predictive accuracy by a factor that is comparable to the improvement that could be attained by an order of magnitude more training set inputs, as shown in the following example.

A fixed set of 1,000 randomly chosen exemplars was presented to a Fuzzy ARTMAP system on five independent 1-epoch circle-in-the-square simulations. After each simulation, inside/outside predictions

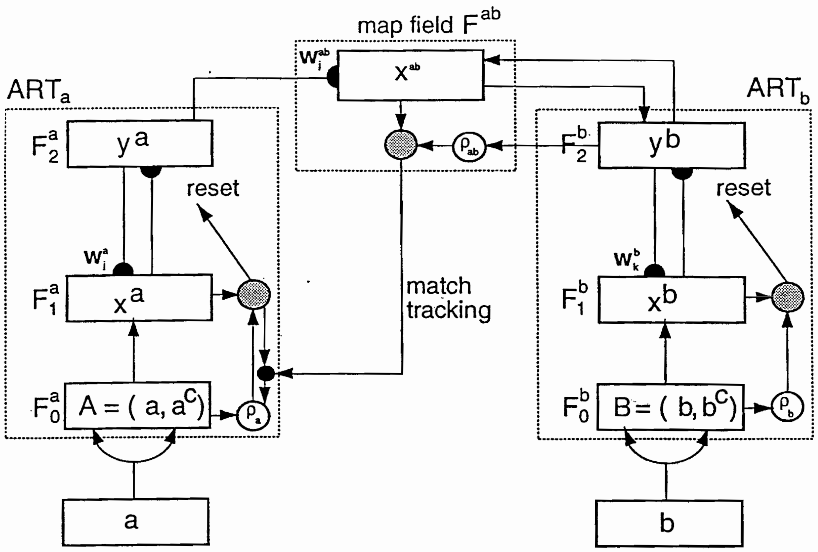

Figure 1. Fuzzy ARTMAP architecture. The ART$_a$ complement coding preprocessor transforms the M_a-vector $\mathbf{a}$ into the $2M_a$-vector $\mathbf{A} = (\mathbf{a}, \mathbf{a}^c)$ at the ART$_a$ field F_0^a. $\mathbf{A}$ is the input vector to the ART$_a$ field F_1^a. Similarly, the input to F_1^b is the $2M_b$-vector $(\mathbf{b}, \mathbf{b}^c)$. When a prediction by ART$_a$ is disconfirmed at ART$_b$, inhibition of map field activation induces the match tracking process. Match tracking raises the ART$_a$ vigilance (ρ_a) to just above the F_1^a to F_0^a match ratio $|\mathbf{x}^a|/|\mathbf{A}|$. This triggers an ART$_a$ search which leads to activation of either an ART$_a$ category that correctly predicts $\mathbf{b}$ or to a previously uncommitted ART$_a$ category node.

the vigilance parameter ρ_a of ART$_a$ by the minimal amount needed to correct a predictive error at ART$_b$.

Parameter ρ_a calibrates the minimum confidence that ART$_a$ must have in a recognition category, or hypothesis, activated by an input $\mathbf{a}^{(p)}$ in order for ART$_a$ to accept that category, rather than search for a better one through an automatically controlled process of hypothesis testing. Lower values of ρ_a enable larger categories to form. These lower ρ_a values lead to broader generalization and higher code compression. A predictive failure at ART$_b$ increases ρ_a by the minimum amount needed to trigger hypothesis testing at ART$_a$, using a mechanism called *match tracking* (Carpenter, Grossberg, and Reynolds, 1991). Match tracking sacrifices the minimum amount of generalization necessary to correct a predictive error. Hypothesis testing leads to the selection of a new ART$_a$ category, which focuses attention on a new cluster of $\mathbf{a}^{(p)}$ input features that is better able to predict $\mathbf{b}^{(p)}$. Due to the combination of match tracking and fast learning, a single ARTMAP system can learn a different prediction for a rare event than for a cloud of similar frequent events in which it is embedded.

ART 1 (BINARY)	FUZZY ART (ANALOG)
CATEGORY CHOICE	
$T_j = \dfrac{\lvert \mathbf{I} \cap \mathbf{w}_j \rvert}{\alpha + \lvert \mathbf{w}_j \rvert}$	$T_j = \dfrac{\lvert \mathbf{I} \wedge \mathbf{w}_j \rvert}{\alpha + \lvert \mathbf{w}_j \rvert}$
MATCH CRITERION	
$\dfrac{\lvert \mathbf{I} \cap \mathbf{w} \rvert}{\lvert \mathbf{I} \rvert} \geq \rho$	$\dfrac{\lvert \mathbf{I} \wedge \mathbf{w} \rvert}{\lvert \mathbf{I} \rvert} \geq \rho$
FAST LEARNING	
$\mathbf{w}_j^{(\text{new})} = \mathbf{I} \cap \mathbf{w}_j^{(\text{old})}$	$\mathbf{w}_j^{(\text{new})} = \mathbf{I} \wedge \mathbf{w}_j^{(\text{old})}$
$\cap$ = logical AND intersection	$\wedge$ = fuzzy AND minimum

Figure 2. Comparison of ART 1 and Fuzzy ART.

Whereas binary ARTMAP employs ART 1 systems for the ART$_a$ and ART$_b$ modules, Fuzzy ARTMAP substitutes Fuzzy ART systems for these modules. Fuzzy ART shows how computations from fuzzy

TABLE 1

		% Correct Test Set Predictions	No. ART_a Categories	No. Epochs
(a)	Average	91.8%	786	1
	Range	91.2%–92.6%	763-805	1
	Voting	95.3%		
(b)	Average	93.9%	1,021	5
	Range	93.4%–94.6%	990-1,070	5
	Voting	96.0%		

Table 1. Voting strategy applied to sets of 5 Fuzzy ARTMAP simulations of the Frey-Slate character recognition task, with training on 1 epoch (a) or 5 epochs (b). (a) Voting eliminated 43% of the errors, which dropped from 8.2% to 4.7%. (b) Voting eliminated 34% of the errors, which dropped from 6.1% to 4.0%.

were recorded on a 1,000-item test set. Accuracy on individual simulations ranged from 85.9% to 92.4%, averaging 90.5%; and the system used from 15 to 23 ART_a nodes. Voting by the five simulations improved test set accuracy to 93.9% (Figure 4(c)). In other words, test set errors were reduced from an average individual rate of 9.5% to a voting rate of 6.1%. Figure 4(d) indicates the number of votes cast for each test set point, and hence reflects variations in predictive confidence across different regions. Voting by more than five simulations maintained an error rate between 5.8% and 6.1%. This limit on further improvement by voting appears to be due to random gaps in the fixed 1,000-item training set. By comparison, a ten-fold increase in the size of the training set reduced the error by an amount similar to that achieved by five-simulation voting. For example, in Figure 3(b), 1-epoch training on 1,000 items yielded a test set error rate of 7.5%; while increasing the size of the training set to 10,000 reduced the test set error rate to 3.3% (Figure 3(c)).

3. Simulation: Letter Image Recognition

Frey and Slate (1991) recently developed a benchmark machine

learning task that they describe as a "difficult categorization problem" (p. 161). The task requires a system to identify an input exemplar as one of 26 capital letters A–Z. The database was derived from 20,000 unique black-and-white pixel images. The difficulty of the task is due to the wide variety of letter types represented: the twenty "fonts represent five different stroke styles (simplex, duplex, complex, and Gothic) and six different letter styles (block, script, italic, English, Italian, and German)" (p. 162). In addition each image was randomly distorted, leaving many of the characters misshapen. Sixteen numerical feature attributes were then obtained from each character image, and each attribute value was scaled to a range of 0 to 15. The resulting Letter Image Recognition file is archived in the UCI Repository of Machine Learning Databases and Domain Theories, maintained by David Aha and Patrick Murphy (ml_repository@ics.uci.edu).

Frey and Slate used this database to test performance of a family of classifiers based on Holland's genetic algorithms (Holland, 1980). The training set consisted of 16,000 exemplars, with the remaining 4,000 exemplars used for testing. Genetic algorithm classifiers having different input representations, weight update and rule creation schemes, and system parameters were systematically compared. Training was carried out for 5 epochs, plus a sixth "verification" pass during which no new rules were created but a large number of unsatisfactory rules were discarded. In Frey and Slate's comparative study, these systems had correct prediction rates that ranged from 24.5% to 80.8% on the 4,000-item test set. The best performance (80.8%) was obtained using an integer input representation, a reward sharing weight update, an exemplar method of rule creation, and a parameter setting that allowed an unused or erroneous rule to stay in the system for a long time before being discarded. After training, the optimal case, that had 80.8% performance rate, ended with 1,302 rules and 8 attributes per rule, plus over 35,000 more rules that were discarded during verification. (For purposes of comparison, a rule is somewhat analogous to an ART_a category in ARTMAP, and the number of attributes per rule is analogous to the size of ART_a category weight vectors.) Building on the results of their comparative study, Frey and Slate investigated two types of alternative algorithms, namely an accuracy-utility bidding system, that had slightly improved performance (81.6%) in the best case; and an exemplar/hybrid rule creation scheme that further improved performance, to a maximum of 82.7%, but that required the creation of over 100,000 rules prior to the verification step.

Fuzzy ARTMAP had an error rate on the letter recognition task that was consistently less than one third that of the three best Frey-Slate genetic algorithm classifiers described above. Moreover Fuzzy ARTMAP simulations each created fewer than 1,070 ART_a categories, compared to the 1,040–1,302 final rules of the three genetic classifiers with the best performance rates. With voting, Fuzzy ARTMAP reduced the error rate to 4.0% (Table 1). Most Fuzzy ARTMAP learning occurred on the first epoch, with test set performance on systems trained for one epoch typically over 97% that of systems exposed to inputs for the five epochs.

Table 1 shows how voting consistently improves performance. With 1 or 5 training epochs, Fuzzy ARTMAP was run for 5 independent simulations, each with a different input order. In all these, and in all other cases tested, voting performance was significantly better than performance of any of the individual simulations in a given group. In Table 1(a), for example, voting caused the error rate to drop to 4.7%, from a 5-simulation average of 8.2%. Hence with 1 training epoch, 5-simulation voting eliminated about 43% of the test set errors. In the 5-epoch simulations, where individual training set performance was close to 100%, 5-simulation voting still reduced the error rate by about 34% (Table 1(b)), where voting reduced the average error rate of 6.1% to a voting error rate of 4.0%.

REFERENCES

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*, **37**, 54–115.

Carpenter, G.A., Grossberg, S., Markuzon, N., Reynolds, J.H., and Rosen, D.B. (1991). Fuzzy ARTMAP: A neural network architecture for incremental supervised learning of analog multidimensional maps. *IEEE Transactions on Neural Networks*, in press.

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*, **4**, 565–588.

Carpenter, G.A., Grossberg, S., and Rosen, D.B. (1991). Fuzzy ART: Fast stable learning and categorization of analog patterns by an adaptive resonance system. *Neural Networks*, **4**, 759–771.

Frey, P.W. and Slate, D.J. (1991). Letter recognition using Holland-style adaptive classifiers. *Machine Learning*, **6**, 161–182.

Holland, J.H. (1980). Adaptive algorithms for discovering and using general patterns in growing knowledge bases. *International Journal of Policy Analysis and Information Systems*, **4**, 217–240.

Kosko, B. (1986). Fuzzy entropy and conditioning. *Information Sciences*, **40**, 165–174.

Wilensky, G. (1990). Analysis of neural network issues: Scaling, enhanced nodal processing, comparison with standard classification. DARPA Neural Network Program Review, October 29–30, 1990.

Zadeh, L. (1965). Fuzzy sets. *Information Control*, **8**, 338–353.