ARTMAP-DS: Pattern Discrimination by Discounting Similarities

Gail A. Carpenter[1] and Frank D. M. Wilson[2]

[1] Boston University, Boston MA 02215, USA
[2] Raytheon Company, Marlborough MA 01752, USA

Abstract. ARTMAP–DS extends fuzzy ARTMAP to discriminate between similar inputs by discounting similarities. When two or more candidate category representations are activated by a given input, features that the candidate representations have in common are ignored prior to determining the winning category. Simulations illustrate the network's ability to recognize similar inputs, such as STAR and START, in a noisy environment.

1 Focusing Attention on Small Differences

ARTMAP–DS is a supervised neural network for learning and recognition. The network extends fuzzy ARTMAP (Carpenter et al., 1992) to discriminate between similar inputs by discounting similarities. The network functions by focusing attention on differences between candidate category representations activated by a given input, and then checking to see which features are in fact present in the input, ignoring features that the candidate representations have in common. Attentional focusing is particularly needed in syllable and word recognition applications, where a primacy gradient input representation (Grossberg, 1978) may cause low–amplitude feature representations (in the later parts of sequences) that are vulnerable to input error and processing noise. A high value of the vigilance parameter, ρ, is needed to ensure that a fuzzy ART network can distinguish between similar input sequences such as STAR and START (Wilson, 1996; Carpenter & Wilson, 1997); but a high value also prevents the system from correctly classifying noisy inputs. The complement-coded input representation used in fuzzy ART (Carpenter, Grossberg, & Rosen, 1991) exacerbates this problem, since the contribution in the input from the phonemes or syllables that are present may be largely masked by the contribution from the larger number that are absent. With ARTMAP-DS, a difference in the later part of the input sequence is not much harder to detect than an earlier one.

2 Fuzzy ARTMAP

Fuzzy ARTMAP is a supervised neural network for learning, recognition and prediction. Figure 1 illustrates a fuzzy ARTMAP system for classification problems, where each input $\mathbf{a}$ learns to predict an output class K. The network creates internal recognition categories during training. The input vector $\mathbf{a}$ is scaled so that each $a_i \in [0,1]$ $(i = 1 \cdots M)$. Complement coding doubles the number of components in the input vector, which becomes $\mathbf{I} = (\mathbf{a}, \mathbf{a}^c)$, where the ith component of $\mathbf{a}^c$ is $a_i^c = (1 - a_i)$. With fast learning, the ART$_a$ weight vector $\mathbf{w}_j \equiv (w_{j1}, \ldots, w_{j,2M})$ records the largest and smallest component values of input vectors placed in the jth category.

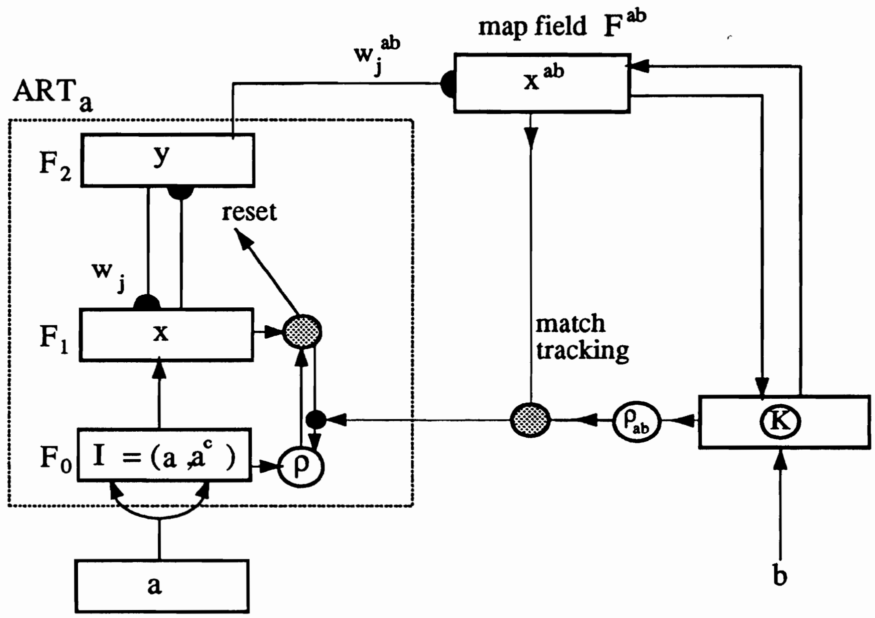

Fig. 1. A fuzzy ARTMAP network for classification.

The $F_1 \rightarrow F_2$ input T_j is given by the Weber law function:

$$T_j(\mathbf{I}) = \frac{|\mathbf{I} \wedge \mathbf{w}_j|}{\alpha + |\mathbf{w}_j|}, \tag{1}$$

where $(\mathbf{P} \wedge \mathbf{Q})_i \equiv \min(P_i, Q_i)$ and $|\mathbf{P}| = \sum_{i=1}^{2M} |P_i|$. Activity at F_2 is denoted by $\mathbf{y} \equiv (y_1, \ldots, y_N)$. With winner-take-all coding, only the F_2 node J that receives the largest $F_1 \rightarrow F_2$ input T_j becomes active. Node J remains active if it satisfies the matching criterion:

$$\frac{|\mathbf{I} \wedge \mathbf{w}_J|}{|\mathbf{I}|} = \frac{|\mathbf{I} \wedge \mathbf{w}_J|}{M} > \rho, \tag{2}$$

where $\rho \in [0,1]$ is the dimensionless ART$_a$ *vigilance parameter*. Otherwise, the network resets the active F_2 node and searches until J satisfies (2). At the start of each input presentation, ρ equals a baseline vigilance, $\overline{\rho}$. When node J is active at F_2 and class label K is active in the training input $\mathbf{b} \equiv (b_1, \ldots, b_L)$, activity at the map field F^{ab} is $\mathbf{x}^{ab} = \mathbf{b} \wedge \mathbf{w}_J^{ab}$, where $\mathbf{w}_j^{ab} \equiv (w_{j1}^{ab}, \ldots, w_{jL}^{ab})$ denotes the weight vector from the jth F_2 node to F^{ab}. If node J then makes an incorrect class prediction (i.e., if $\mathbf{x}^{ab} \neq \mathbf{b}$), a *match tracking* signal raises ART$_a$ vigilance ρ to $|\mathbf{I} \wedge \mathbf{w}_J^a| \;/\; |\mathbf{I}| + \epsilon$, where ϵ is vanishingly small. This increase is just enough to induce a search, which continues until either some F_2 node becomes active for the first time, in which case the weight vector $\mathbf{w}_J^{ab}$ is set equal to $\mathbf{x}^{ab}$, so that J learns the correct output class label $k(J) = K$; or until a node J that has previously learned to predict K becomes active. During testing, a pattern $\mathbf{a}$ that activates node J is predicted to belong to the class $K = k(J)$.

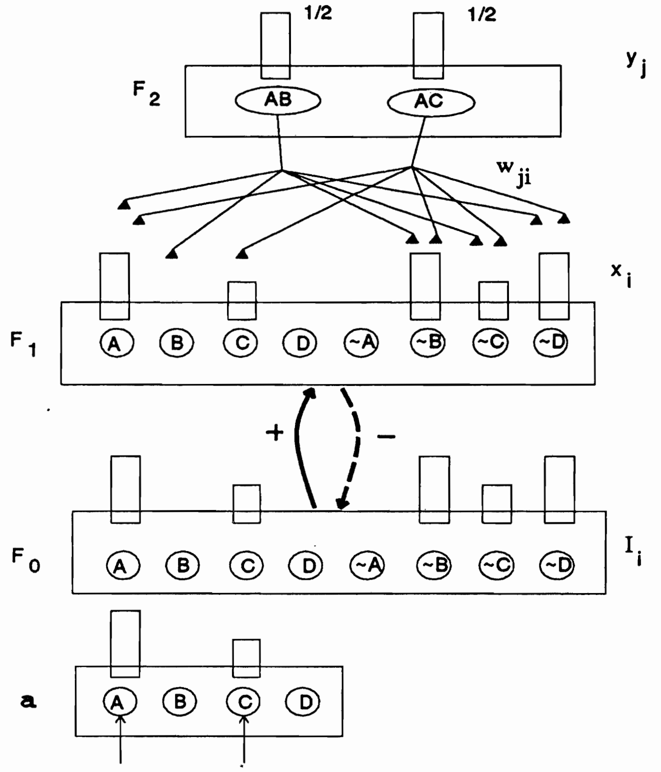

Fig. 2. The ARTMAP–DS network. When more than one F_2 node remains strongly active after contrast–enhancement, F_1 first registers features that are present both in the input and in all strongly active templates. Inhibitory $F_1 \rightarrow F_0$ connections then mask out these input features. Thus, if input AC activates both AB and AC at F_2, A and "not D" are inhibited and "not B" and "not C" are partially inhibited at F_0. After renormalization, F_0 sends a masked input back to F_1 and F_2, which allows F_2 to choose among the partially active nodes. The figure shows the initial situation, in which activity at F_1 is equal to $\mathbf{I}^{(norm)}$, and AB and AC have just become active at F_2. After F_1 has inhibited F_0, only nodes C and "not B" will remain active at F_0.

3 The ARTMAP–DS Network

ARTMAP–DS (Figure 2) replaces the ART$_a$ subsystem of fuzzy ARTMAP. During training, ARTMAP–DS functions identically to fuzzy ARTMAP. During subsequent recognition tests, input to F_0 is the complement–coded vector $\mathbf{I} = (\mathbf{a}, \mathbf{a}^c)$. F_0 activity is normalized, $I_i^{(norm)} = I_i/|\mathbf{I}|_\infty$, where $|\mathbf{I}|_\infty = \max_i \{I_i\}$ is the L^∞ norm ($\lim_{p\to\infty}$ $(\sum_i I_i^p)^{(1/p)}$). Bottom-up $F_0 \to F_1$ input results in initial activity $\mathbf{x}^{(init)}$ at F_1 that is equal to $\mathbf{I}^{(norm)}$. Initial input to F_2 is given by $T_j(\mathbf{x}^{(init)})$, where T_j is as defined in (1). Activity at F_2 is contrast-enhanced:

$$y_j = \frac{T_j^p}{\sum_r T_r^p}, \tag{3}$$

where p is significantly greater than 1, but is not so large that F_2 activity approximates choice. If there is then a clear winner at F_2, i.e., if $y_J \geq \theta$ for some J, where $0.5 \leq \theta < 1$, then the input is predicted to belong to class $K = k(J)$ as in vanilla fuzzy ARTMAP. Otherwise ARTMAP discrimination by discounting similarities is invoked, as follows.

The field F_1 first registers features that are present both in the input and in all strongly active templates. The matched activity $\mathbf{x}^{(match)}$ at F_1 is:

$$x_i^{(match)} = I_i^{(norm)} \wedge \left[\bigwedge_{j:y_j>\phi} w_{ji} \right], \tag{4}$$

where $0 < \phi < \theta$. For example, if the input **a** is STIR, and if STAR and STIR nodes are both significantly active at F_2 after contrast-enhancement, F_1 will register S, T and R. (In all simulations, $\phi = 0.1$ and $\theta = 0.7$.) Inhibitory connections from F_1 to F_0 then mask out these common features, as follows:

$$I_i^{(inh)} = [I_i^{(norm)} - x_i^{(match)}]^+ = [I_i^{(norm)} - \bigwedge_{j:y_j>\phi} w_{ji}]^+, \tag{5}$$

where $[w]^+ \equiv \max\,(w, 0)$. Activity at F_0 is then renormalized, F_0 sends a masked input ("I, not A") back to F_1 and F_2, and F_2 makes a choice (STIR) from among the partially active nodes. That is, the network replaces the former value of $\mathbf{I}$ with $\mathbf{I}^{(inh)}$, and recomputes $\mathbf{I}^{(norm)}$, $\mathbf{x}^{(init)}$ and T_j. At F_2, contrast-enhancement and thresholding are repeated, but only among nodes that had been active in the previous iteration. If there is still no clear winner at F_2, this discrimination process is repeated until a clear winner does emerge. As in vanilla fuzzy ARTMAP, $\rho_a = 0$ during recognition tests, so the matching criterion (2) is automatically satisfied, once an F_2 node J that is a clear winner emerges.

4 Syllable Recognition with Input Noise

This section describes simulations that illustrate the ability of ARTMAP-DS to give better recognition performance than fuzzy ARTMAP for noisy inputs. The inputs are words that are represented as phoneme sequences using a primacy gradient with steepness 0.5, i.e., the input for the nth phoneme in the sequence has amplitude $0.5^{(n-1)}$. First, the five notional phoneme sequences STAR, STARK, STIR, SHARK and SHIRK are presented to ARTMAP-DS and to fuzzy ARTMAP for learning.

Learning in ARTMAP–DS and in fuzzy ARTMAP functions identically. Supervised learning ensures that each input is coded by a separate F_2 node. During learning, inputs $\mathbf{a}$ are noise–free and a correct representation of each input is learned. During subsequent performance, Gaussian noise with amplitude proportional to the component magnitudes is added to each input component, i.e., $a_i^{(noisy)} = a_i\ (1 + n_i)$, where n_i is Gaussian with mean 0 and standard deviation σ_{noise}. Each word was tested 20 times at each noise level. Further simulations illustrate recognition performance using a set of 50 monosyllabic words (see Appendix) constructed from 22 phonemes. After noise–free learning, each word is tested 10 times at each noise level. Table 1 shows that for both simulations, ARTMAP–DS achieves better recognition performance than fuzzy ARTMAP.

Standard deviation of noise (σ_{noise})		0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0	# test set inputs
# test set A recognition errors	Fuzzy ARTMAP	0	0	0	3	3	4	7	11	18	21	100
	ARTMAP–DS	0	0	0	1	1	2	4	8	15	16	100
# test set B recognition errors	Fuzzy ARTMAP	0	2	8	24	56	88	112	134	159	176	500
	ARTMAP–DS	0	2	8	22	53	83	108	131	149	163	500

Table 1. Recognition error rates when (A) the inputs STAR, STARK, STIR, SHARK and SHIRK, and (B) the 50 words listed in the Appendix are repeatedly presented to fuzzy ARTMAP and to ARTMAP–DS, with noise added to the inputs. Noise is Gaussian, with amplitude proportional to the magnitude of each input component. Parameters $\alpha = 0.001$, $\phi = 0.1$ and $\theta = 0.7$. For test set (A), $p = 20$, and for test set (B), $p = 400$.

5 Syllable Recognition with Network Noise

This section describes simulations that illustrate the ability of ARTMAP–DS to give better recognition performance than fuzzy ARTMAP when endogenous noise perturbs the match field F_1 during recognition tests. First, the input phoneme sequences are learned by each network as in the previous section. During learning, F_1 is noise–free. During subsequent performance, initial activity $x_i^{(init)}$ at F_1 is calculated by adding Gaussian noise n_i to each normalized input component $I_i^{(norm)}$, bounding the result so that it lies within the range $[0, 1]$, and then renormalizing at F_1 to ensure that $|\mathbf{x}^{(init)}| = M$. After noise–free learning, each word from the 50–word lexicon (see Appendix) is tested 10 times at each noise level. ARTMAP–DS achieves a performance improvement over fuzzy ARTMAP across a broad range of values of p $(100 \leq p \leq 400)$. Low values of p do not provide enough contrast enhancement at F_2, resulting in a higher error rate. For values of p greater than 400, F_2 dynamics approximate choice, so ARTMAP–DS reduces to fuzzy ARTMAP. Table 2 shows that ARTMAP–DS again achieves better recognition performance than fuzzy ARTMAP.

Standard deviation of noise (σ_{noise})		0.05	0.1	0.15	0.2	# test set inputs
# test set	Fuzzy ARTMAP	8	45	103	139	500
recognition errors	ARTMAP–DS	2	27	63	121	500

Table 2. Number of recognition errors when the 50 words listed in the Appendix are each presented 10 times to the ARTMAP–DS network and to fuzzy ARTMAP, with noise added at F_1. Parameters $p = 400$, $\phi = 0.1$ and $\theta = 0.7$.

References

Carpenter, G. A., Grossberg, S., Markuzon, N., Reynolds, J. H., & Rosen, D. B. (1992). Fuzzy ARTMAP: A neural network architecture for incremental supervised learning of analog multidimensional maps. *IEEE Transactions on Neural Networks, 3*, 698–713.

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

Carpenter, G. A., & Wilson, F. D. M. (1997). Segmentation-ART: A neural network for word recognition from continuous speech. *In preparation.*

Grossberg, S. (1978). A theory of human memory: self–organization and performance of sensory–motor codes, maps and plans. *Progress in Theoretical Biology, 5*, 233–374.

Wilson, F. D. M. (1996). *Neural networks for noise–tolerant category discrimination with application to continuous speech segmentation.* Ph.D. thesis, Boston, MA: Boston University.

Appendix

In the 50–word simulations, the phoneme set is: */p/*, */b/*, */t/*, */d/*, */k/*, */g/*, */s/*, */S/*, */r/*, */l/*, */a/*, */i/*, */u/*, */$\wedge$/*, */f/*, */m/*, */n/*, */ϵ/*, */o/*, */θ/*, */h/* and silence. The words and their transcriptions are: star */star/*, stark */stark/*, stir */stir/*, shark*/shark/*, shirk */shirk/*, odd */ad/*, are */ar/*, ark */ark/*, art */art/*, box */baks/*, bar */bar/*, bark */bark/*, be */bi/*, beast*/bist/*, beat */bit/*, breed */brid/*, brood */brud/*, boot */but/*, dark */dark/*, dart */dart/*, drop */drap/*, drew */dru/*, friend */frend/*, Greek */grik/*, greet */grit/*, grew */gru/*, group */grup/*, car */kar/*, key */ki/*, keep */kip/*, crop */krap/*, creep */krip/*, lead */lid/*, leap */lip/*, least */list/*, par */par/*, park */park/*, part */part/*, pat */pat/*, see */si/*, seek */sik/*, seal */sil/*, seat */sit/*, spark */spark/*, spot */spat/*, stop */stap/*, struck */str$\wedge$k/*, streak */strik/*, sue */su/*, tar */tar/*. Although the correct phonetic representations of the English words SHARK and SHIRK are */Sak/* and */S$\wedge$k/* respectively, the nonword phoneme sequences */shark/* and */shirk/* are used instead, in order to allow the 50–word lexicon to incorporate the 5–word lexicon as a subset.

Acknowledgements: This research was supported in part by the National Science Foundation (NSF–IRI–94–01659) and the Office of Naval Research (ONR N00014–95–1–0409 and ONR N00014–95–1–0657)