INTERNATIONAL COMPUTER SCIENCE INSTITUTE

1947 Center St. • Suite 600 • Berkeley, California 94704-1198 • (510) 643-9153 • FAX (510) 643-7684

Simplified ART: A new class of ART algorithms

Andrea Baraldi* and Ethem Alpaydın†

baraldi@icsi.berkeley.edu, ethem@icsi.berkeley.edu

TR-98-004

February 1998

Abstract. The Simplified Adaptive Resonance Theory (SART) class of networks is proposed to handle problems encountered in Adaptive Resonance Theory 1 (ART 1)-based algorithms when detection of binary and analog patterns is performed. The basic idea of SART is to substitute ART 1-based "unidirectional" (asymmetric) activation and match functions with "bidirectional" (symmetric) function pairs. This substitution makes the class of SART algorithms potentially more robust and less time-consuming than ART 1-based systems. One SART algorithm, termed Fuzzy SART, is discussed. Fuzzy SART employs probabilistic and possibilistic fuzzy membership functions to combine soft competitive learning with outlier detection. Its soft competitive strategy relates Fuzzy SART to the well-known Self-Organizing Map and Neural Gas clustering algorithm. A new Normalized Vector Distance, which can be employed by Fuzzy SART, is also presented. Fuzzy SART performs better than ART 1-based Carpenter-Grossberg-Rosen Fuzzy ART in the clustering of a simple two-dimensional data set and the standard four-dimensional IRIS data set. As expected, Fuzzy SART is less sensitive than Fuzzy ART to small changes in input parameters and in the order of the presentation sequence. In the clustering of the IRIS data set, performances of Fuzzy SART are analogous to or better than those of several clustering models found in the literature.

Keywords: hard and soft competitive learning, cluster detection, ART 1-based systems, Self-Organizing Map, Neural Gas algorithm, fuzzy set theory, fuzzy clustering.

*On leave from IMGA-CNR, Bologna 40129 Italy.

†On leave from Department of Computer Engineering, Boğaziçi University, Istanbul TR-80815 Turkey.

Nomenclature

$\mathbf{W}$	template vector
$\mathbf{X}$	input vector
FS	feature space
d	dimensionality of the input space ($d \in \mathcal{I}^+$)
t	index of time ($t \in \mathcal{I}^+$)
c	current number of output categories (processing elements) ($c \in \mathcal{I}^+$)
n	number of input patterns ($n \in \mathcal{I}^+$)
j	index of template vector or output category ($j \in \{1, c\}$)
k	index of input vector component ($k \in \{1, d\}$)
i	index of input vector ($i \in \{1, n\}$)
E_j	jth output processing element
F_k	kth input unit
$\mathbf{W_j^{(t)}}$	jth bottom-up template vector, equivalent to weights of input-to-output connections converging on processing element E_j at time t
$\mathbf{X^{(t)}}$	input vector at time t
$W_{k,j}^{(t)}$	connection weight from the kth input unit to the jth output processing element
$X_k^{(t)}$	kth scalar component of $\mathbf{X^{(t)}}$
$\mu_j^{(t)}$	activation value of unit E_j at time t
$\arg\max_{j=1,\ldots,c} \left\{ \mu_j^{(t)} \right\}$	picks index j for which $\mu_j^{(t)}$ is maximum
ρ	vigilance threshold ($\rho \in [0, 1]$)
α	Fuzzy ART choice parameter
$\|\mathbf{X^{(t)}}\|$	norm of $\mathbf{X^{(t)}}$
$\mid \mathbf{X^{(t)}} \mid$	length (modulus) of $\mathbf{X^{(t)}}$
θ_j	angle between $\mathbf{X^{(t)}}$ and $\mathbf{W_j^{(t)}}$
β	learning rate ($\beta \in [0, 1]$)
$p(C_j)$	*a priori* probability of category (pattern's state) C_j
$p(C_j \mid \mathbf{X_i})$	*a posteriori* conditional probability that the pattern's state is C_j, given that the pattern is $\mathbf{X_i}$
$p(\mathbf{X_i} \mid C_j)$	class conditional probability that the pattern is $\mathbf{X_i}$, given that the pattern's state is C_j
$R_{i,j}$	relative or probabilstic fuzzy membership (typicality) value of $\mathbf{X_i}$ with respect to fuzzy concept C_j ($R_{i,j} \in [0, 1]$)
$A_{i,j}$	absolute or possibilistic fuzzy membership (typicality) value of $\mathbf{X_i}$ with respect to fuzzy concept C_j ($A_{i,j} > 0$)
$d_{i,j}$	Euclidean distance between $\mathbf{X_i}$ and $\mathbf{W_j}$
σ_j	resolution parameter of output unit E_j
η_j	resolution parameter of output unit E_j
p_j	resolution parameter of output unit E_j

e_j	local time counter of output unit E_j
β_j	learning rate of output unit E_j
$\epsilon_j(e_j, R_{i,j})$	learning coefficient which is monotonically nonincreasing with e_j and monotonically nondecreasing with $R_{i,j}$ ($\epsilon_j \in [0,1]$)
r_j	neighborhood-ranking ($r_j \in \{0, c-1\}$)
$\sigma_j(e_j)$	spread value defined as a monotonically decreasing function of e_j ($\sigma_j > 0$)
$h_j[r_j, \sigma_j(e_j)]$	learning coefficient defined as a monotonically decreasing function of r_j and e_j ($h_j \in [0,1]$)
PE	Processing Element
$\vec{M}$	Unidirectional Degree of Match
$\overline{M}$	Bidirectional Degree of Match
$\vec{AF}$	Unidirectional Activation Function
$\overline{AF}$	Bidirectional Activation Function
$\vec{MF}$	Unidirectional Match Function
$\overline{MF}$	Bidirectional Match Function
NVD	Normalized Vector Distance
MDM	Modulus Degree of Match
ADM	Angle Degree of Match
MDM_{max}	Modulus Degree of Match Threshold
ADM_{max}	Angle Degree of Match Threshold
VDM	Vector Degree of Match

1 Introduction

All natural systems provided with cognitive capabilities feature feedback interaction with their external environment. Owing to this environmental feedback, natural systems weaken or reinforce their behaviors as a function of their success (Serra & Zanarini, 1990; Parisi, 1991). Mimicking the real world, an artificial cognitive system employing reinforcement learning "is allowed to react to each training case; it is then told whether its reaction was good or bad" (Masters, 1994), "but no actual desired values are given" (Bishop, 1995). One example of artificial reinforcement learning is that provided by the Adaptive Resonance Theory (ART)-based clustering algorithms, where an *orienting subsystem* models some external evaluation of the pattern-matching reaction of the *attentional subsystem* to an input stimulus (Carpenter & Grossberg, 1987a, 1987b; Carpenter, Grossberg & Rosen, 1991; Carpenter, Grossberg, Markuzon, Reynolds & Rosen, 1992). The *a priori* knowledge exploited by the ART orienting subsystem consists of a user-defined vigilance threshold which provides an upper limit on the size of the nodes' receptive field in the input space, such that coarser grouping of input patterns is obtained when the vigilance parameter is lowered.

In recent years, several ART-based models have been presented. ART 1 categorizes binary patterns but features sensitivity to the order of presentation of the random sequence (Carpenter & Grossberg, 1987a). This finding led to the development of the Improved ART 1 system (IART 1), which is less dependent than ART 1 on the order of presentation of the input sequence (Shih, Moh & Chang, 1992). The Adaptive Hamming Net (AHN), which is functionally equivalent to ART 1, optimizes ART 1 both in terms of computation time and storage requirement (Hung & Lin, 1995). ART 2, designed to detect regularities in analog random sequences, employs a computationally expensive architecture which presents difficulties in parameter selection (Carpenter & Grossberg, 1987b). To overcome these difficulties, the Fuzzy ART system was developed as a generalization of ART 1 (Carpenter, Grossberg & Rosen, 1991; Carpenter, Grossberg, Markuzon, Reynolds & Rosen, 1992). This means however that ART 1-based structural problems may also affect Fuzzy ART. Our goal is to provide a new synthesis between properties of Fuzzy ART and other successful clustering algorithms such as the Self-Organizing Map (SOM) (Kohonen, 1990, 1995) and Neural Gas (NG) (Martinetz, Berkovich & Schulten, 1993), to extend the abilities of these separate approaches.

This paper is organized as follows. In Section 2, a general template for ART 1-based algorithms is presented. In Section 3 Fuzzy ART is discussed, and improvements are recommended. Based on these recommendations, the Simplified ART (SART) class of algorithms is defined in Section 4. Section 5 presents interpattern similarity measures that can be employed in SART implementations. Section 6 proposes exploitation of probabilistic and possibilistic fuzzy membership functions to combine soft competitive learning and outlier detection in SART architectures. Section 7 presents a soft competitive SART implementation, termed Fuzzy SART. This section gives a brief review of Kohonen's constraints, developed for the Kohonen Vector Quantization (VQ) and SOM algorithms and then adopted by the well-known NG algorithm. In Section 8 the performance of Fuzzy SART is compared with that of Fuzzy ART on a simple two-dimensional data set and the standard four-dimensional IRIS data set. Conclusions are reported in Section 9.

2 The class of ART 1-based models

At least three ART 1-based clustering algorithms can be found in the literature: i) IART 1, which employs a slightly modified ART 1 architecture to extract statistical regularities from binary samples; ii) Adaptive Hamming Net (AHN), which is a feed-forward network that optimizes ART 1 in terms of computation time and storage requirement (Hung & Lin, 1995); and iii) Fuzzy ART, which extracts statistical regularities from random samples of binary as well as analog pattern distributions.

In this section we describe a general framework that covers these algorithms.

2.1 Definitions

Let us consider two generic feature vectors (patterns) $\mathbf{W}$ and $\mathbf{X}$ belonging to feature space FS.

Definition 1. The "unidirectional" or "asymmetric" degree to which $\mathbf{W}$ matches $\mathbf{X}$ ($\mathbf{W} \rightarrow \mathbf{X}$), $\forall \mathbf{W}, \mathbf{X} \in FS$, is a mapping

$$\vec{M}(\mathbf{W}, \mathbf{X}) : FS \times FS \rightarrow [0, 1],$$

where $\vec{M}$ stands for Unidirectional Degree of Match. Scalar function $\vec{M}(\mathbf{W}, \mathbf{X})$ is such that:
(a) it provides an interpattern similarity value, $\forall \mathbf{W}, \mathbf{X} \in FS$;
(b) it provides a relative number, belonging to range $[0, 1]$, $\forall \mathbf{W}, \mathbf{X} \in FS$, such that $\vec{M}(\mathbf{X}, \mathbf{X}) = 1$ $\forall \mathbf{X} \in FS$; and
(c) it is a "unidirectional" (asymmetric) measure, i.e., at least one pattern pair ($\mathbf{W}$,$\mathbf{X}$) exists such that: $\vec{M}(\mathbf{W}, \mathbf{X}) \neq \vec{M}(\mathbf{X}, \mathbf{W})$.

Definition 2. The "bidirectional" or "symmetric" degree of match between $\mathbf{W}$ and $\mathbf{X}$ ($\mathbf{W} \leftrightarrow \mathbf{X}$), $\forall \mathbf{W}, \mathbf{X} \in FS$, is a mapping

$$\overline{M}(\mathbf{W}, \mathbf{X}) : FS \times FS \rightarrow [0, 1],$$

where $\overline{M}$ stands for Bidirectional Degree of Match. Scalar function $\overline{M}(\mathbf{W}, \mathbf{X})$ is such that:
(a) it provides an interpattern similarity value, $\forall \mathbf{W}, \mathbf{X} \in FS$;
(b) it provides a relative number, belonging to range $[0, 1]$, $\forall \mathbf{W}, \mathbf{X} \in FS$, such that $\overline{M}(\mathbf{W}, \mathbf{X}) = 1$ iff $\mathbf{W} = \mathbf{X}$; and
(c) it is a "bidirectional" (symmetric) measure such that $\overline{M}(\mathbf{W}, \mathbf{X}) = \overline{M}(\mathbf{X}, \mathbf{W})$, $\forall \mathbf{W}, \mathbf{X} \in FS$.

2.2 Processing scheme

Although in its original form the ART 1 attentional subsystem employs bottom-up (feed-forward) and top-down (feed-backward) connections, it is easy to prove that this module is mathematically equivalent to an attentional subsystem where feed-forward connections are adopted exclusively (Baraldi & Parmiggiani, 1995a). For example, the Adaptive Hamming Net (AHN), shown in Fig. 1, is a feed-forward network functionally equivalent to ART 1 (Hung & Lin, 1995).

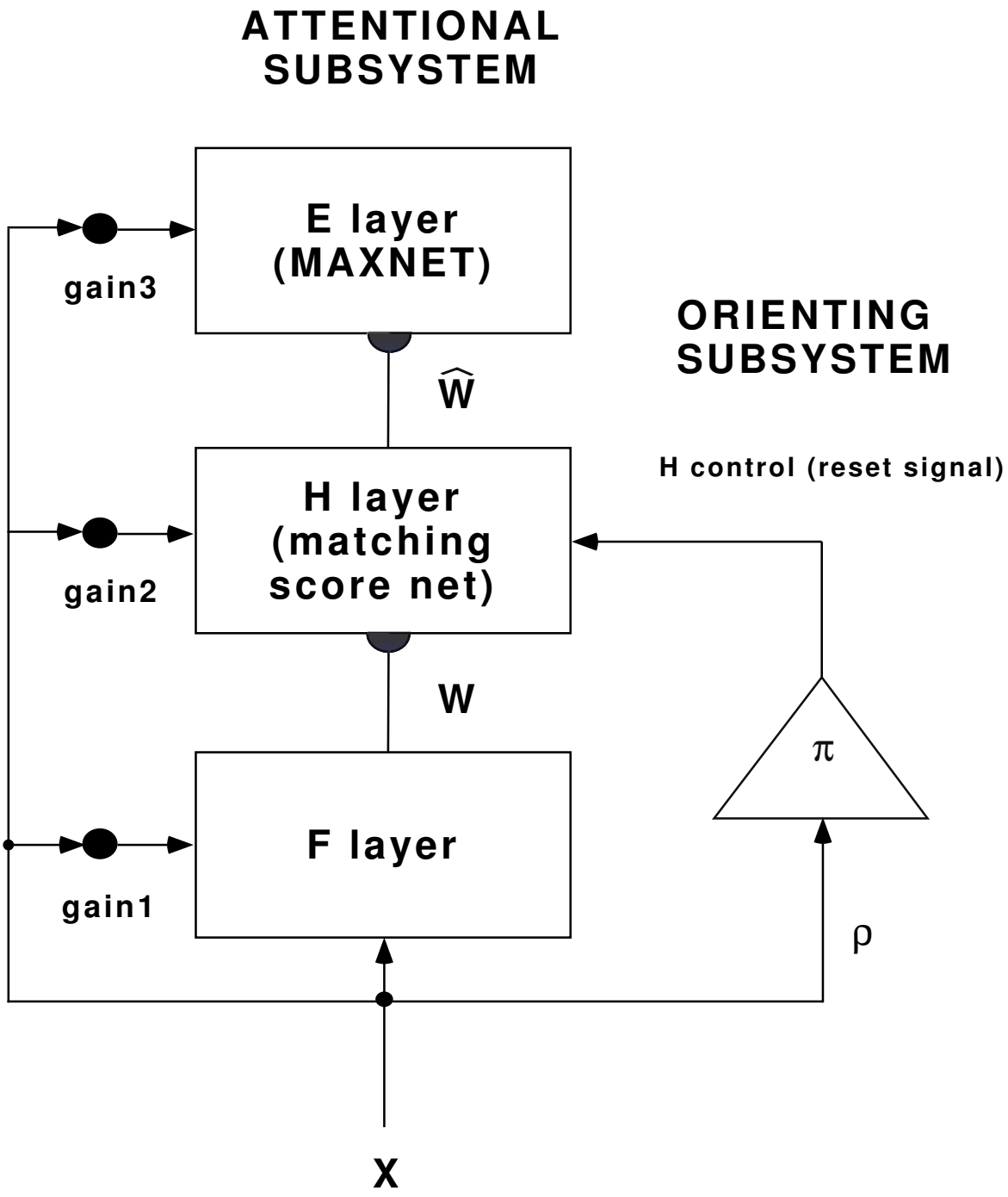


Figure 1: AHN system. For detail on the meaning of threshold π and weights $\mathbf{W}$ and $\widehat{\mathbf{W}}$, refer to Hung & Lin (1995).

We can generalize this result by stating that the attentional module of all ART 1-based systems is functionally equivalent to a feed-forward network featuring no top-down connection. This simplified view is acknowledged by Carpenter, Grossberg & Rosen when they state (1991, p. 763): "[in ART 1-based Fuzzy ART] weight vectors subsume both the bottom-up and top-down weight vectors of ART 1".

This simplification yields, as a major consequence, a change in the meaning of the term "resonance" as traditionally applied to ART 1-based systems. This term should no longer indicate "the basic feature of all ART systems, notably, pattern-matching between bottom-up input and top-down learned prototype vectors" (Carpenter, Grossberg & Rosen, 1991, p. 760), just as the term "resonance" has never been used with reference to pattern matching performed by a feed-forward Kohonen network.

In our view, the term "resonance", as employed in ART, means rather that all ART 1-based algorithms share the same modular architecture, consisting of:
i) a completely generic (unsupervised), flat (without hidden layers), feed-forward (bottom-up) network performing pattern recognition (as Kohonen's networks, e.g., VQ and SOM), termed *attentional subsystem*; and
ii) a supervised/unsupervised knowledge interface unit, termed *orienting subsystem*, where the quality of unsupervised bottom-up pattern recognition is compared to top-down requirements (expectations, or prior knowledge) provided by the external environment (supervisor). In the orienting subsystem, if unsupervised knowledge matches external expectations, then "resonance" occurs. This means that the unsupervised pattern recognition activity of the attentional module is reinforced according to a reinforcement learning mechanism, i.e., prototype adaptation takes place. If resonance does not occur, the orienting subsystem allows the attentional module to increase its resources (processing elements) to match external requirements.

To describe the class of ART 1-based clustering algorithms, let us consider the attentional subsystem as a two-layer network. The first layer is termed *Feature representation field* (F layer), and consists of input units F_k, $k = 1, ..., d$, where d is the dimensionality of the input space. An input vector, identified as $\mathbf{X^{(t)}} = (X_1^{(t)}, \ldots, X_d^{(t)})$, is presented to the input layer at time t, where, in the binary case $X_k \in \{0, 1\}$, or, in the analog case $X_k \in \mathcal{R}$, $k = 1, ..., d$. The second layer, termed *Exemplar representation field* (E layer), stores an arbitrary number c of Processing Elements (PEs). These PEs, also termed *output nodes*, *exemplars*, *categories*, *components*, or *clusters*, are identified as E_j, $j = 1, ..., c$.

At time t, each processing unit E_j computes an activation value as an interpattern similarity measure between vectors $\mathbf{X^{(t)}}$ and $\mathbf{W_j^{(t)}}$, where $\mathbf{W_j^{(t)}} = (W_{1,j}^{(t)}, \ldots, W_{d,j}^{(t)})$, termed *template vector*, *reference vector* or *cluster prototype*, is the *receptive field center* in the input space of output unit E_j. In a feed-forward (bottom-up) network structure, reference vector components $W_{k,j}^{(t)}$, $k = 1, ..., d$, correspond to weights of input-to-output connections converging on output node E_j from input units F_k, $k = 1, ..., d$, at time t.

Orienting subsystem supervision of attentional pattern matching causes coarser partitions of the input space when a user-defined *vigilance parameter* $\rho \in [0, 1]$ is lowered. This is equivalent to considering this vigilance threshold as an upper limit on the size of the nodes' receptive field in the input space.

The modular architecture of ART 1-based systems is shown in Fig. 2. The following

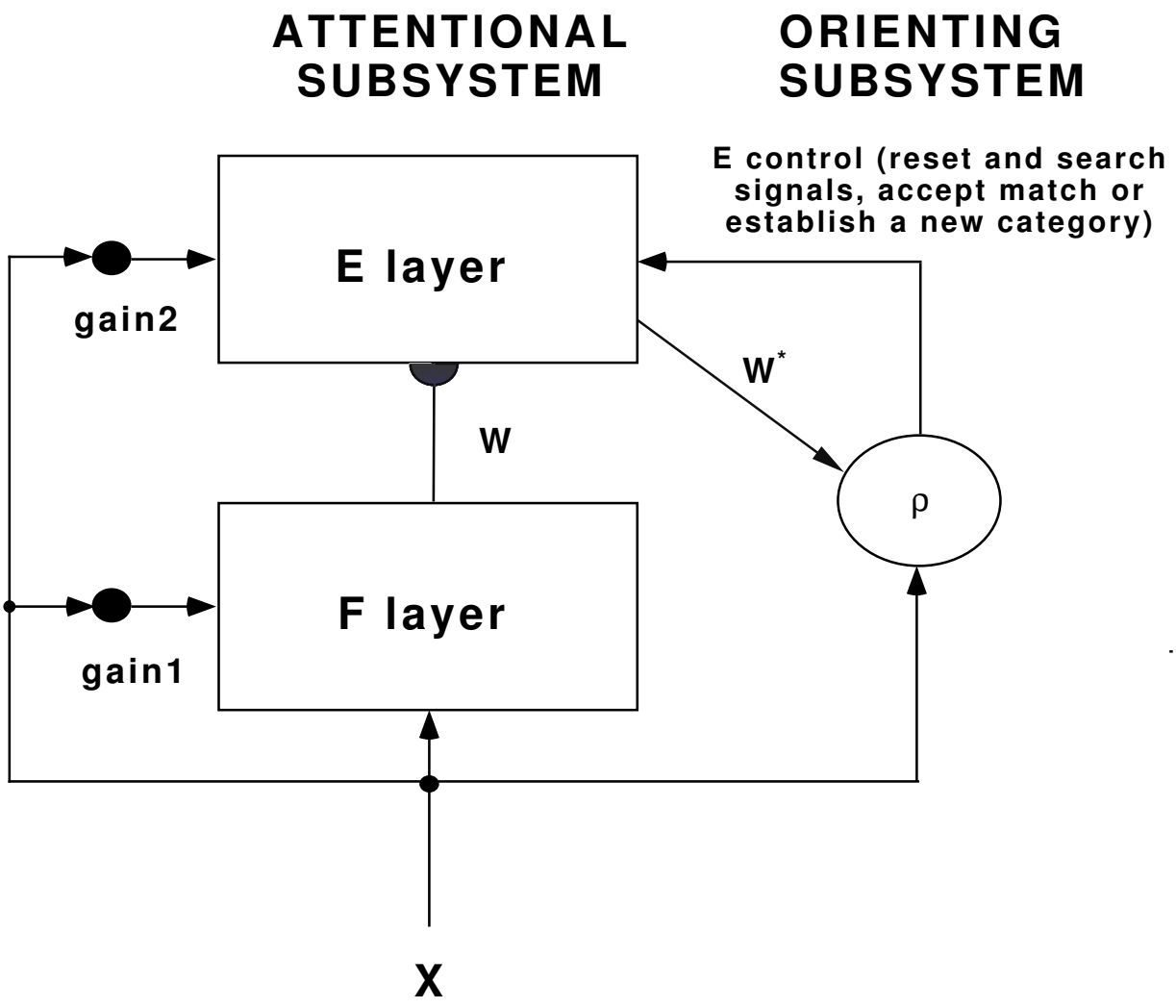


Figure 2: ART 1-based system architecture, where $\mathbf{W}$ identifies a matrix of bottom-up connections and $\mathbf{W}^*$ is the best-matching template. For more detail, refer to the text.

sequential algorithm implements the basic idea behind ART 1-based models.

Step 0. Initialization. PE counter c and pattern counter t are set to 0.

Step 1. Input pattern presentation. The pattern counter is increased by one as $t = t + 1$, and a new pattern $\mathbf{X^{(t)}}$ is presented to the input nodes.

Step 2. Activation value computation. Activation values of output PEs are computed as

$$\mu_j^{(t)} = \vec{AF}(\mathbf{X^{(t)}}, \mathbf{W_j^{(t)}}), \quad j = 1, ..., c, \tag{1}$$

where $\vec{AF}(\mathbf{X^{(t)}}, \mathbf{W_j^{(t)}})$ is a Unidirectional Activation Function, also called *choice function*. Function $\vec{AF}(\mathbf{X^{(t)}}, \mathbf{W_j^{(t)}})$ belongs to the class of $\vec{M}$ functions (see Subsection 2.1): it measures the degree to which $\mathbf{X^{(t)}} \rightarrow \mathbf{W_j^{(t)}}$ ($\mathbf{X^{(t)}}$ matches $\mathbf{W_j^{(t)}}$), but it does not assess the reverse situation, i.e., the degree to which $\mathbf{X^{(t)}} \leftarrow \mathbf{W_j^{(t)}}$ ($\mathbf{W_j^{(t)}}$ matches $\mathbf{X^{(t)}}$).

Step 3. Detection of processing units eligible for being resonant. Processing units that are affected by the arrival of an input pattern at a given time are said to belong to the *resonance domain*. ART 1-based models enforce a hard competitive learning mechanism, otherwise called Winner-Take-All (WTA) strategy (Martinetz, Berkovich & Schulten, 1993; Fritzke, 1997a). This means that ART 1-based systems allow only the best-matching prototype to be attracted by $\mathbf{X^{(t)}}$, i.e., only the best-matching unit, also termed *recognition category*, can belong to the resonace domain. The best-matching unit, identified as E_{j^*}, is the solution, if any, to the maximization problem (Healy, Caudell & Smith, 1993)

$$j^* = \arg \max_{j=1,...,c} \left\{ \vec{AF}(\mathbf{X^{(t)}}, \mathbf{W_j^{(t)}}) \right\}, \tag{2}$$

subject to the vigilance constraint described below. The best-matching weight vector converging on E_{j^*} is identified as $\mathbf{W_{j^*}^{(t)}}$.

Step 4. Resonance domain detection. The orienting subsystem selects among processing units candidated by the attentional subsystem for being resonant those that match external requirements. Only these units are said to belong to the resonance domain. To select these units, the orienting subsystem employs a *vigilance constraint*, also termed *vigilance test* or *hypothesis test*, combined with a *mismatch reset condition* and a *search process*. The vigilance test applied to the pattern matching activity of the attentional module is

$$\vec{MF}(\mathbf{W_{j^*}^{(t)}}, \mathbf{X^{(t)}}) \geq \rho, \quad \rho \in [0, 1], \tag{3}$$

where user-defined vigilance parameter ρ provides a model of top-down external expectation, while $\vec{MF}(\mathbf{W_{j^*}^{(t)}}, \mathbf{X^{(t)}})$ is a Unidirectional Match Function. This function belongs to the class of $\vec{M}$ functions (see Section 2.1): it measures the degree to which $\mathbf{W_{j^*}^{(t)}} \rightarrow \mathbf{X^{(t)}}$, but it does not assess the reverse situation, i.e., the degree to which $\mathbf{X^{(t)}} \rightarrow \mathbf{W_{j^*}^{(t)}}$.

If the vigilance test is not satisfied (i.e., "resonance" does not occur), the mismatch reset condition is enforced. It inhibits the best-matching node and searches for the second

best-matching node, which is then submitted to the vigilance test. Note that vigilance test computes value $\vec{MF}(\mathbf{W_j^{(t)}}, \mathbf{X^{(t)}})$ while units eligible for being resonant are searched according to their $\vec{AF}(\mathbf{X^{(t)}}, \mathbf{W_j^{(t)}})$ value. Since the best-matching template $\mathbf{W_{j^*}^{(t)}}$ in terms of $\vec{AF}(\mathbf{X^{(t)}}, \mathbf{W_j^{(t)}})$ is not necessarily the best-matching template in terms of $\vec{MF}(\mathbf{W_j^{(t)}}, \mathbf{X^{(t)}})$, this justifies the search, to be repeated until either the vigilance test is passed or no more nodes are available for testing.

Step 5(a). Resonance condition: reinforcement learning. If the vigilance test is satisfied, i.e., "resonance" occurs, the attentional subsystem is allowed to reinforce its pattern-matching activity by adjusting template vectors of units belonging to the resonance domain. These prototypes are modified according to an ART 1-based model-dependent *weight adaptation law*, so as to be moved closer to input pattern $\mathbf{X^{(t)}}$ (i.e., the input pattern is used as an attractor). In hard competitive ART 1-based systems, only the best-matching prototype $\mathbf{W_{j^*}^{(t)}}$ is updated by $\mathbf{X^{(t)}}$.

Step 5(b). Non-resonance condition: new processing element allocation. If no solution exists to the maximization problem described above, i.e., the resonance domain is an empty set, then "resonance" does not occur, and one new processing unit is dynamically allocated to match external expectations. Thus, the PE counter is increased as $c = c + 1$, and a new node E_c is allocated to match input pattern $\mathbf{X^{(t)}}$, such that $\mathbf{W_c^{(t+1)}} = \mathbf{X^{(t)}}$. As a consequence, ART 1-based models require no randomization of initial templates since initial values are data-driven.

Step 6. Goto step 1.

3 Fuzzy ART

Fuzzy ART requires a preprocessing stage where either input pattern normalization or complement coding is used to prevent category proliferation. The latter technique normalizes input vectors while preserving their amplitude information but it doubles the number of connections (Carpenter, Grossberg & Rosen, 1991; Carpenter, Grossberg, Markuzon, Reynolds & Rosen, 1992). To simplify our discussion, let us identify the analog input pattern presented to Fuzzy ART at time t after normalization preprocessing as $\mathbf{X^{(t)}} = (X_1^{(t)}, \ldots, X_d^{(t)})$, where $X_k \in [0, 1]$, $k = 1, ..., d$.

3.1 Fuzzy ART-specific features

To fit the processing scheme described in Section 2.2, Fuzzy ART implementation details are now presented. Whenever necessary, relationships with ART 1 and IART 1 are also highlighted.

Activation function. The activation function is defined as

$$\vec{AF}_1(\mathbf{X^{(t)}}, \mathbf{W_j^{(t)}}) = \frac{\sum_{k=1}^{d} \min\{X_k^{(t)}, W_{k,j}^{(t)}\}}{\alpha + \sum_{k=1}^{d} W_{k,j}^{(t)}}, \quad j = 1, ..., c, \quad X_k, W_{k,j}^{(t)} \in [0, 1], \tag{4}$$

where parameter α, ranging in [0.001, 1) (Carpenter, Grossberg, Markuzon, Reynolds & Rosen, 1992), breaks ties in favor of the longer of two template vectors. It is to be noted that:
i) $\vec{AF}_1(\mathbf{X^{(t)}}, \mathbf{W_j^{(t)}})$ belongs to the class of $\vec{M}$ functions (see Section 2.1).
ii) In ART 1 and IART 1, the "unidirectional" activation function applied to binary vector pairs is (Hung & Lin, 1995)

$$\vec{AF}_2(\mathbf{X^{(t)}}, \mathbf{W_j^{(t)}}) = \frac{\sum_{k=1}^{d} W_{k,j}^{(t)} \cdot X_k^{(t)}}{\alpha + \sum_{k=1}^{d} W_{k,j}^{(t)}}, \quad j = 1, ..., c; \quad X_k^{(t)}, W_{k,j}^{(t)} \in \{0, 1\}. \tag{5}$$

Equation (4) generalizes Equation (5) by substituting the product and norm operators with operations that resemble those employed in fuzzy set theory (e.g., intersection and cardinality). As Simpson observed (1993, p. 37): "for these operations to be correctly interpreted as fuzzy operations, they would have to be applied to membership values, not to the parameters of the activation function." This also means that the "degree of fuzzification" of Fuzzy ART with respect to ART 1 is questionable.

Match function. The match function is defined as

$$\vec{MF}_1(\mathbf{W_{j^*}^{(t)}}, \mathbf{X^{(t)}}) = \frac{\sum_{k=1}^{d} \min\{W_{k,j^*}^{(t)}, X_k^{(t)}\}}{\sum_{k=1}^{d} X_k^{(t)}}, \quad X_k, W_{k,j^*}^{(t)} \in [0, 1]. \tag{6}$$

It is to be noted that:
i) $\vec{MF}_1(\mathbf{W_{j^*}^{(t)}}, \mathbf{X^{(t)}})$ belongs to the class of $\vec{M}$ functions (see Section 2.1).
ii) Parameters ρ in Equation (3) and α in Equation (4) are interrelated as illustrated by Huang, Georgiopoulos & Heileman (1995). For example, if $\alpha \leq \rho/(1-\rho)$, then Fuzzy ART completes its learning in one list presentation when complement coding is employed for preprocessing.
iii) In ART 1 and IART 1, the "unidirectional" match function applied to binary vector pairs is (Hung & Lin, 1995)

$$\vec{MF}_2(\mathbf{W_{j^*}^{(t)}}, \mathbf{X^{(t)}}) = \frac{\sum_{k=1}^{d} W_{k,j^*}^{(t)} \cdot X_k^{(t)}}{\sum_{k=1}^{d} X_k^{(t)}}, \quad X_k^{(t)}, W_{k,j^*}^{(t)} \in \{0, 1\}. \tag{7}$$

Also Equation (6) generalizes Equation (7) by substituting the product and norm operators with fuzzy-like operators. Equation (7) provides a normalized measure of how many unit-valued (informative) components of $\mathbf{X^{(t)}}$ are matched by those of $\mathbf{W_{j^*}^{(t)}}$, i.e., it measures the degree to which $\mathbf{W_{j^*}^{(t)}} \rightarrow \mathbf{X^{(t)}}$. Since Equation (7) applies to binary vectors, it can be written as (see Appendix A)

$$\vec{MF}_2(\mathbf{W_{j^*}^{(t)}}, \mathbf{X}^{(t)}) = \frac{\mid \mathbf{W_{j^*}^{(t)}} \mid \cdot \cos\theta_{j^*}}{\mid \mathbf{X^{(t)}} \mid}, \tag{8}$$

where θ_{j^*} is the angle between $\mathbf{X^{(t)}}$ and $\mathbf{W_{j^*}^{(t)}}$ and operator $|\cdot|$ is the vector length. The numerator in Equation (8) is the projection of $\mathbf{W_{j^*}^{(t)}}$ along the direction of $\mathbf{X^{(t)}}$.
iv) The only difference between ART 1 and IART 1 is that the latter model applies two unidirectional and complementary vigilance tests. This is equivalent to stating that IART 1 adopts one bidirectional vigilance test. Besides Equation (7), IART 1 employs match function (Hung & Lin, 1995)

$$\vec{M}F_3(\mathbf{X^{(t)}}, \mathbf{W_{j^*}^{(t)}}) = \frac{\sum_{k=1}^{d} W_{k,j^*}^{(t)} \cdot X_k^{(t)}}{\sum_{k=1}^{d} W_{k,j^*}^{(t)}}, \quad X_k^{(t)}, W_{k,j^*}^{(t)} \in \{0,1\}. \tag{9}$$

Equation (9) provides a normalized measure of how many unit-valued components of $\mathbf{W_{j^*}^{(t)}}$ are matched by those of $\mathbf{X^{(t)}}$, i.e., it measures the degree to which $\mathbf{X^{(t)}} \rightarrow \mathbf{W_{j^*}^{(t)}}$. In analogy with Equation (8), Equation (9) can be written as

$$\vec{M}F_3(\mathbf{X^{(t)}}, \mathbf{W_{j^*}^{(t)}}) = \frac{|\ \mathbf{X^{(t)}}\ | \cdot \cos\theta_{j^*}}{|\ \mathbf{W_{j^*}^{(t)}}\ |}. \tag{10}$$

The numerator in Equation (10) is the projection of $\mathbf{X^{(t)}}$ along the direction of $\mathbf{W_{j^*}^{(t)}}$.

Resonance condition. Weight adaptation law is

$$W_{k,j}^{(t+1)} = \begin{cases} (1-\beta) \cdot W_{k,j}^{(t)} + \beta \cdot \min\{W_{k,j}^{(t)}, X_k^{(t)}\}, & \text{if } j = j^*,\ k = 1, ..., d; \\ W_{k,j}^{(t)}, & \text{if } j \neq j^*,\ k = 1, ..., d, \end{cases} \tag{11}$$

with learning rate $\beta \in [0,1]$. In the fast-learning case, β is taken as 1. Equation (11) stresses the fact that only winner template $\mathbf{W_{j^*}^{(t)}}$ is allowed to be attracted by input pattern $\mathbf{X^{(t)}}$, which makes the model hard competitive.

3.2 Weaknesses of Fuzzy ART

From the analysis of the model and the processing example proposed in Appendix B, we conclude that:

i) Fuzzy ART is sensitive to the order of presentation of the random sequence. This finding is consistent with the results of Shih, Moh & Chang (1992) about ART 1.

ii) Fuzzy ART is time-consuming, its search process requiring up to $c \log c$ steps to sort all activation values.

iii) Fuzzy ART may be affected by overfitting, since a single poorly mapped pattern suffices to initiate the creation of a new unit, and no noise category removal mechanism is employed by the system. In other words, Fuzzy ART may fit the noise and not just the data.

iv) Since its learning rate is independent of time, Fuzzy ART lacks stability because of

excessive plasticity, i.e., the processing of a new data set can move templates located by the system during the previous learning phase.

v) Owing to its hard competitive implementation, Fuzzy ART is expected to be less robust and more likely to generate dead units than its possible soft competitive implementations (Fritzke, 1997a; Martinetz, Berkovich & Schulten, 1993; Davè & Krishnapuram, 1997).

vi) It does not preserve topological information (Martinetz, Berkovich & Schulten, 1994).

vii) Termination is not based on optimizing any model of the process or its data.

3.3 Improvements to Fuzzy ART

Possible solutions to the above weaknesses are proposed:

i) ART 1, IART 1 and Fuzzy ART are all affected by the same structural problem: they all employ an inherently asymmetric design to perform an inherently symmetric task. The assessment of the interpattern degree of match is an inherently symmetric task. Nonetheless, ART 1-based systems break this measure into two steps, such that activation and match functions compute two unidirectional and complementary similarity values. The problem is that these two complementary similarity values are employed separately by two different (asymmetric) specialized tasks. The first task selects the best-matching template. The second task constrains the input pattern to fall within a bounded hypervolume of acceptance centered on the best-matching template.

As IART 1 improves ART 1 by replacing the unidirectional match function with a bidirectional match function (Shih, Moh & Chang, 1992), the next step in the evolution of ART 1-based architectures should replace the pair of unidirectional activation and match functions with a pair of bidirectional functions.

ii) ART 1-based systems require no searching when an Adaptive Hamming Net (AHN) approach is applied (Hung & Lin, 1995). The key idea is to convert the sequential search procedure of ART 1 into an optimization problem (see Section 2.2), which can be solved by parallel implementation. In a first stage, match function values are computed in parallel and all the indexes of reference vectors that do not satisfy the vigilance constraint are filtered out at once. Next, activation values of all surviving candidate categories are computed in parallel and passed to a MAXNET output layer, which detects the best-matching unit E_{j^*} at once.

iii) Noise point and outlier detection capability must be combined with a noise category removal mechanism to avoid overfitting.

iv) Learning rates should decrease monotonically with time, according to a cooling schedule, hereafter referred to as the *first Kohonen constraint*. When a node loses its plasticity (i.e., its learning rate tends to zero), then it becomes stable, since its response to the same stimulus does not change with time.

v) A model transition from soft competitive to hard competitive learning should be driven through time, e.g., according to what is hereafter referred to as the *second Kohonen constraint*. It requires the degree of overlap among node receptive fields to decrease monotonically with time until it becomes zero, as receptive fields become Voronoi polyhedra (Fritzke, 1997a).

vi) To guarantee topologically correct mapping, dynamic generation/removal of synaptic links between node pairs should be enforced according to the competitive Hebbian learning mechanism (Martinetz, Berkovich & Schulten, 1994; Fritzke, 1997a).

vii) Stronger relationships with other clustering algorithms capable of minimizing a cost function, such as the Neural Gas algorithm (NG) (Martinetz, Berkovich & Schulten, 1993), should be pursued.

4 SART framework

To overcome limitations of ART 1-based models, we propose a new class of ART processing schemes, hereafter referred to as Simplified ART (SART), consisting of a pair of attentional and orienting subsystems and capable of processing real-valued multidimensional patterns (Fig. 2). Using the ART 1-based sequential processing scheme of Section 2.2 as the general framework, a SART algorithm can be summarized as follows.

Given a real and multidimensional input pattern $\mathbf{X^{(t)}} \in \mathcal{R}^d$ (i.e., no preprocessing constraint must be enforced), the best-matching unit E_{j^*} is the solution, if any, to the maximization problem

$$j^* = \arg \max_{j=1,\ldots,c} \left\{ \overline{AF}(\mathbf{W_j^{(t)}}, \mathbf{X^{(t)}}) \right\}, \tag{12}$$

subject to vigilance constraint

$$\overline{MF}(\mathbf{W_{j^*}^{(t)}}, \mathbf{X^{(t)}}) \geq \rho, \quad \rho \in [0,1], \tag{13}$$

where $\overline{AF}(\mathbf{W_j^{(t)}}, \mathbf{X^{(t)}})$ is a Bidirectional Activation Function and $\overline{MF}(\mathbf{W_j^{(t)}}, \mathbf{X^{(t)}})$ is a Bidirectional Match Function, both belonging to the class of $\overline{M}$ functions (see Section 2.1).

From this definition stems one important corollary.

If $\overline{AF}(\mathbf{W_j^{(t)}}, \mathbf{X^{(t)}})$ and $\overline{MF}(\mathbf{W_j^{(t)}}, \mathbf{X^{(t)}})$ are chosen such that $\overline{MF}(\mathbf{W_1^{(t)}}, \mathbf{X^{(t)}}) > \overline{MF}(\mathbf{W_2^{(t)}}, \mathbf{X^{(t)}})$ implies that $\overline{AF}(\mathbf{W_1^{(t)}}, \mathbf{X^{(t)}}) > \overline{AF}(\mathbf{W_2^{(t)}}, \mathbf{X^{(t)}})$, and vice versa, then no mismatch reset condition and search process (or search-like procedure, in case of parallel implementation) are required to detect the resonance domain (see Step 4 in Section 2.2). This means that when best-matching template $\mathbf{W_{j^*}^{(t)}}$, selected according to $\overline{AF}(\mathbf{W_j^{(t)}}, \mathbf{X^{(t)}})$, does not satisfy the vigilance criterion, a new processing unit can be immediately allocated to match $\mathbf{X^{(t)}}$. For example, if $\overline{AF}(\mathbf{W_j^{(t)}}, \mathbf{X^{(t)}}) \equiv \overline{MF}(\mathbf{W_j^{(t)}}, \mathbf{X^{(t)}})$, this corollary is obviously true.

5 Interpattern similarity measures as relative numbers

Three possible examples of interpattern similarity measures belonging to the class of $\overline{M}$ functions are discussed. The first example proposes a new measure termed Vector Degree of Match (VDM). The latter two examples are rather driven from similarity equations already discussed in Section 3. If they apply to analog patterns, then these functions can be employed as $\overline{AF}$s and/or $\overline{MF}$s in SART implementations, according to the SART general framework proposed in Section 4.

5.1 VDM measure

Two analog vectors $\mathbf{W}$ and $\mathbf{X} \in \mathcal{R}^d$ are equal iff they feature the same vector length, direction and orientation, i.e., (i) the angle between them, identified as θ, and (ii) their vector length difference, are equal to zero. Let us consider these two conditions separately, then their combination.

i) The Modulus Degree of Match (MDM) is a relative number, belonging to range (0,1], defined as

$$MDM(\mathbf{W}, \mathbf{X}) = min\{\mid \mathbf{W} \mid / \mid \mathbf{X} \mid, \mid \mathbf{X} \mid / \mid \mathbf{W} \mid\}, \tag{14}$$

where $\mid \mathbf{W} \mid$ and $\mid \mathbf{X} \mid$ are the moduli of $\mathbf{W}$ and $\mathbf{X}$ respectively. Equation (14), which is independent of multiplicative noise, was developed for SAR image processing where speckle is modeled as multiplicative noise (Baraldi & Parmiggiani, 1995c). Alternative MDM expressions independent of additive noise may be developed as well.

ii) We can write that

$$\gamma = \cos\theta = (\mathbf{X} \circ \mathbf{W})/(\mid \mathbf{X} \mid \cdot \mid \mathbf{W} \mid), \tag{15}$$

where $(\mathbf{X} \circ \mathbf{W})$ is the scalar product between $\mathbf{X}$ and $\mathbf{W}$, with γ ranging from -1 to $+1$. Thus, $\theta = \arccos(\gamma)$, where θ belongs to range $[0, \pi]$. The Angle Degree of Match (ADM) is a relative number, also belonging to range [0,1], defined as

$$ADM(\mathbf{W}, \mathbf{X}) = (\pi - \theta)/\pi. \tag{16}$$

In line with the two constraints required by the criterion of vector pair equivalence presented above, a possible nonlinear expression for VDM combines variables MDM and ADM as

$$VDM(\mathbf{W}, \mathbf{X}) = MDM(\mathbf{W}, \mathbf{X}) \cdot ADM(\mathbf{W}, \mathbf{X}), \tag{17}$$

so that $0 < VDM \leq \min\{\ MDM,\ ADM\} \leq 1$. Equation (17) implies that $\mathbf{W}$ and $\mathbf{X}$ are the same vector, i.e., they are identical, such that $VDM = 1$, iff (i) their in-between angle is zero ($ADM = 1$); and (ii) their moduli are the same ($MDM = 1$).

Note that $VDM(\mathbf{W}, \mathbf{X}) = VDM(\mathbf{X}, \mathbf{W})$. Then VDM satisfies all conditions required to belong to the class of $\overline{M}$ functions (see Section 2.1), i.e.,

$$\overline{M}_1(\mathbf{W}, \mathbf{X}) \equiv VDM(\mathbf{W}, \mathbf{X}). \tag{18}$$

Since it applies to analog vector pairs, this equation can be employed as $\overline{AF}$ and/or $\overline{MF}$ in SART models (see Section 4). An example of PE employing Equation (18) as its $\overline{AF}$

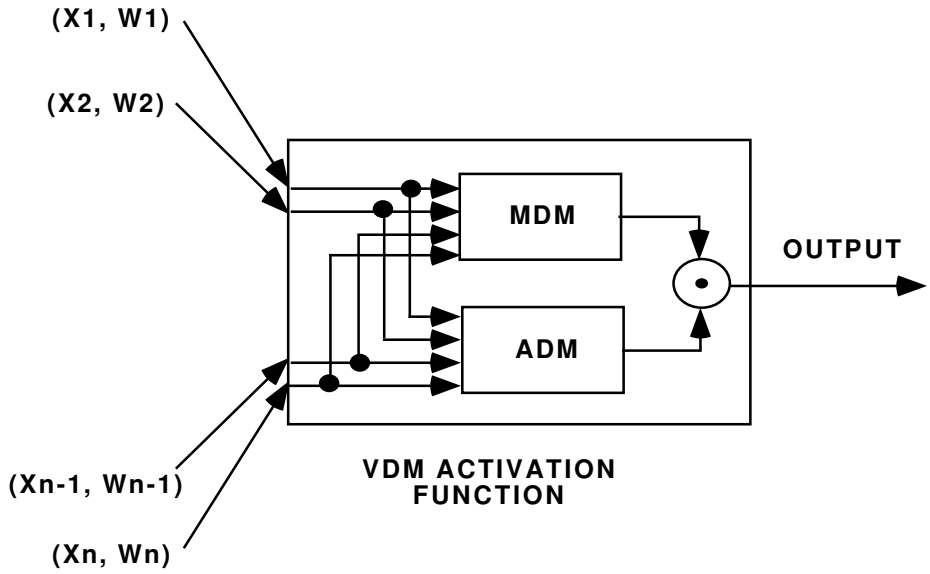


Figure 3: A processing unit employing Equation (17) as its activation function.

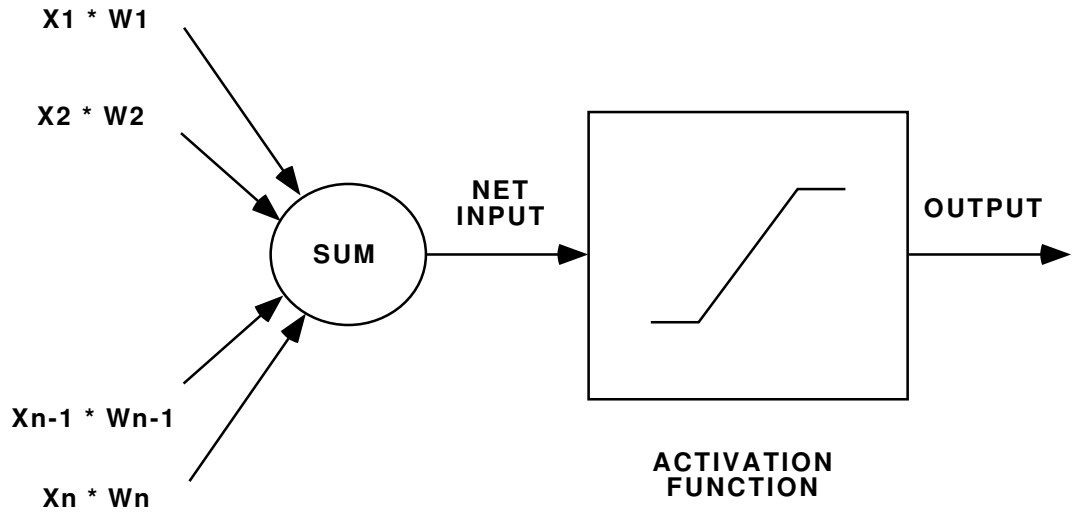


Figure 4: The "traditional" perceptron model (Masters, 1994; Pao, 1989).

is shown in Fig. 3, to be compared with Fig. 4 showing a traditional perceptron which computes its output activation level as a two-step operation: first, the scalar (dot) product between the set of synaptic weights and the input pattern is computed as the *net input to the node*, then the net input is acted on by an activation function (Masters, 1994; Pao, 1989).

5.1.1 Normalized Vector Distance metric

The intervector difference (dissimilarity or contrast) between $\mathbf{W}$ and $\mathbf{X}$, $\forall \mathbf{W}, \mathbf{X} \in \mathcal{R}^d$, is traditionally computed by means of the $\mathcal{L}^1$ or $\mathcal{L}^2$ norm whose range is $(0, +\infty)$. The Normalized Vector Distance metric (NVD) rather assesses the distance between $\mathbf{W}$ and $\mathbf{X}$ as a relative number, $\forall \mathbf{W}, \mathbf{X} \in \mathcal{R}^d$. NVD is defined as

$$NVD(\mathbf{W}, \mathbf{X}) = 1 - VDM(\mathbf{W}, \mathbf{X}), \tag{19}$$

where $VDM(\mathbf{W}, \mathbf{X})$ is computed according to Equation (17). If $NVD = 0$ ($VDM = 1$), then $\mathbf{W}$ and $\mathbf{X}$ are the same vector. If $NVD \to 1$ ($VDM \to 0$), then the two vectors are maximally different. It can be demonstrated that $\mathcal{R}^d$ is a metric space (Pao, 1989) with metric NVD, where the following relationships hold true (Baraldi & Parmiggiani, 1995a, b, c, 1996)

1. NVD is a mapping from the metric space $\mathcal{R}^d$ to range [0,1], i.e., $NVD : \mathcal{R}^d \times \mathcal{R}^d \to [0, 1]$ (normalized positivity);
2. $NVD(\mathbf{X}, \mathbf{Y}) = 0$ iff $\mathbf{X} = \mathbf{Y}$;
3. $NVD(\mathbf{X}, \mathbf{W}) = NVD(\mathbf{W}, \mathbf{X})$, $\forall \mathbf{X}, \mathbf{W} \in \mathcal{R}^d$ (symmetry);
4. $NVD(\mathbf{X}, \mathbf{W}) \leq NVD(\mathbf{X}, \mathbf{G}) + NVD(\mathbf{G}, \mathbf{W})$, $\forall \mathbf{X}, \mathbf{G}, \mathbf{W} \in \mathcal{R}^d$ (triangle inequality).

5.1.2 VDM and NVD properties

It is important to stress that, while the Euclidean distance is invariant with respect to both translation and rotation, NVD (VDM) is: i) invariant with respect to rotation; ii) not invariant with respect to change of scale; and iii) not invariant with respect to translation. As an elementary example of these properties, let us consider, in a 1-D space, two pair of points, e.g., 5 and 10 vs. 20 and 25, where the second pair is obtained by translation of the first. The Euclidean distance between the two points of each pair is 5, while the NVD value is 0.5 and 0.2 respectively. This means that NVD (VDM) is not appropriate for assessing interpattern dissimilarities (similarities) in the Euclidean space (for an example, refer to Section 9). [1]

[1] An additional interesting relationship can be established between NVD computation and the way in which the mammalian visual system performs independent detection of achromatic and chromatic color contrasts (Boynton, 1990); in fact, MDM is inversely related to achromatic color differences, while ADM is inversely related to chromatic color differences.

5.2 More examples of $\overline{M}$ functions

A second example of $\overline{M}$ function can be obtained by applying a binary operator, e.g., sum or product, to the unidirectional activation and match function pair of Fuzzy ART. For example, the product between Equations (4) and (6) can be applied to analog vector pairs such that

$$\overline{M}_2(\mathbf{W}, \mathbf{X}) = \frac{(\sum_{k=1}^{d} \min\{W_k, X_k\})^2}{\sum_{k=1}^{d} W_k \cdot \sum_{k=1}^{d} X_k}, \quad W_k, X_k \in \mathcal{R}. \tag{20}$$

Note that, unlike Equation (20), Equation (18) processes vector-length and vector-angle information independently. In ART 1, the product between unidirectional activation and match functions, computed as Equations (8) and (10), gives

$$\overline{M}_3(\mathbf{W}, \mathbf{X}) = \frac{(\sum_{k=1}^{d} W_k \cdot X_k)^2}{\sum_{k=1}^{d} W_k \cdot \sum_{k=1}^{d} X_k} = \cos^2(\theta), \quad W_k, X_k \in \{0, 1\}, \tag{21}$$

where θ is the angle between binary vectors $\mathbf{W}$ and $\mathbf{X}$. Equation (21) states that two binary vectors are the same vector iff their in-between angle θ is zero, regardless of their moduli.

6 Fuzzy memberships, mixture probabilities and outlier detection in SART systems

This section proposes exploitation of probabilistic and possibilistic fuzzy membership functions to combine soft competitive learning and outlier detection in SART models. These two properties refer to recommendations iii) and v) proposed in Section 3.3 to improve ART 1-based systems.

6.1 Absolute and relative fuzzy memberships

Let $\mathbf{X_i}$ be instance i of input variable $\mathbf{X}$, $i = 1, ..., n$, where n is the total number of input instances, such that $\mathbf{X_i}$ may belong to a generic *state* (also termed *category* or *component*) C_j, $j = 1, ..., c$, where c is the total number of possible states. The extent to which $\mathbf{X_i}$ is compatible with a vague (fuzzy) concept associated with generic state C_j can be interpreted "more in terms of a *possibility (compatibility) distribution* rather than in terms of a *probability distribution*" (Pao, 1989, p. 58). This legitimizes some possibility distributions, called fuzzy membership functions, that "we believe are *useful*, but might find difficult to justify on the basis of objective probabilities" (Pao, 1989, p. 57). Depending on the conditions required to state that c fuzzy states C_j, $j = 1, ..., c$, are a fuzzy c-partition of the input data set, membership functions can be divided into two categories (Krishnapuram & Keller, 1993; Davè & Krishnapuram, 1997):

1. *relative or probabilistic or constrained fuzzy membership (typicality) values* $R_{i,j}$; and
2. *absolute or possibilistic fuzzy membership (typicality) values* $A_{i,j}$,

where index i ranges over patterns and index j over concepts. Absolute and relative membership types are related by the following equation:

$$R_{i,j} = \frac{A_{i,j}}{\sum_{h=1}^{c} A_{i,h}}, \quad i = 1, ..., n, \quad j = 1, ..., c. \tag{22}$$

Relative typicality values, $R_{i,j}$, must satisfy the following three conditions (Tsao, Bezdek & Pal, 1994; Pao, 1989):

i) $R_{i,j} \in [0,1]$, $i = 1, ..., n$, $j = 1, ..., c$;

ii) $\sum_{j=1}^{c} R_{i,j} = 1$, $i = 1, ..., n$; and

iii) $0 < \sum_{i=1}^{n} R_{i,j} < n$, $j = 1, ..., c$.

Constraint (ii) is an inherently probabilistic constraint (Krishnapuram & Keller, 1993), relating $R_{i,j}$ values to posterior probability estimates in a Bayesian framework. Because of condition (ii), $R_{i,j}$ values are relative numbers dependent on the absolute membership of the pattern in all other classes, thus indirectly on the total number of classes. This also means that PEs exploiting a relative membership function as their activation function are context-sensitive, i.e., $R_{i,j}$ provides a tool for modeling network-wide internode communication by subsuming that PEs are coupled through feed-sideways (lateral) connections (Ancona, Ridella, Rovetta & Zunino, 1997).

Possibilistic membership functions relax condition (ii) to satisfy the following constraints (Krishnapuram and Keller, 1993):

iv) $A_{i,j} \in [0,1]$, $i = 1, ..., n$, $j = 1, ..., c$;

v) $\max_j \{A_{i,j}\} > 0$, $i = 1, ..., n$; and

vi) $0 < \sum_{i=1}^{n} A_{i,j} < n$, $j = 1, ..., c$.

Owing to condition (v), the sum of absolute memberships of a noise point in all the "good" categories need not be equal to one. In this paper the definition of absolute membership function is further relaxed to satisfy constraint (v) exclusively, i.e., a fuzzy set employing absolute membership values may not be normal as its membership values may feature no least upper bound equal to one (Pao, 1989). Term $A_{i,j}$ is an absolute similarity value depending on fuzzy state C_j exclusively, given input pattern $\mathbf{X_i}$. In other words, $A_{i,j}$ is context-insensitive, since it is not affected by any other state. Thus, PEs exploiting an absolute membership as their activation function are independent, i.e., they feature no lateral connection.

Both probabilistic and possibilistic fuzzy clustering are affected by some well-known drawbacks. On one hand in probabilistic fuzzy clustering, owing to condition (ii), noise points and outliers, featuring low possibilistic typicalities with respect to all templates, may have significantly high probabilistic membership values and may severely affect the prototype parameter estimate (e.g., refer to Davè & Krishnapuram, 1997). On the other hand in possibilistic fuzzy clustering, learning rates computed from absolute typicalities tend to produce coincident clusters (Barni, Cappellini & Mecocci, 1996; Davè and Krishnapuram, 1997). This poor behavior can be explained by the fact that cluster prototype are uncoupled in possibilistic clustering, i.e., possibilistic clustering algorithms try to minimize an objective function by operating on each cluster independently. This leads to an increase in the number of local minima.

Different $A_{i,j}$ expressions, consistent with the definition provided above, were found to be useful in the existing literature. These include the following:

$$A_{i,j} = \begin{cases} \frac{1}{1+d_{i,j}^2/\eta_j} \in (0,1] \text{ (Krishnapuram \& Keller, 1993)}, & (23) \\ Gaussian_{i,j} = e^{-\frac{d_{i,j}^2}{2\sigma_j^2}} \in (0,1] \text{ (Gaussian mixtures; Williamson, 1996, 1997)}, & (24) \\ \frac{1}{(d_{i,j}^2)^{p_j}} \in (0,\infty) \text{ (Bezdek \& Pal, 1993)}, & (25) \\ \frac{1}{(1-Gaussian_{i,j})^2} \in (1,\infty) \text{ (Baraldi \& Parmiggiani, 1997)}, & (26) \end{cases}$$

where $d_{i,j} = d(\mathbf{X_i}, \mathbf{W_j})$ is assumed to be the Euclidean distance between input pattern $\mathbf{X_i}$ and prototype (receptive field center) $\mathbf{W_j}$ of the j-th category. Variables σ_j, η_j and p_j are all resolution parameters belonging to range $(0,\infty)$ (refer to Davè & Krishnapuram, 1997). It is to be noted that Equations (23) and (24) belong to the class of $\overline{M}$ functions (see Section 2.1). For any $A_{i,j}$ expression chosen among Equations (23) to (26), relative membership function (22) also belongs to the class of $\overline{M}$ functions.

6.2 Fuzzy memberships and mixture probabilities

Note that absolute membership function (24) relates probabilistic membership (22) to Gaussian mixture models, which are widely employed in the framework of optimization problems featuring a firm statistic foundation (Dempster, Laird & Rubin, 1977; Martinetz, Berkovich & Schulten, 1993; Buhmann, 1995; Alpaydın, 1998). In a mixture probability model consisting of c mixture components C_j, $j = 1, ..., c$, let $p(C_j)$ be the *a priori* probability that a patterns belongs to mixture component C_j, and $p(\mathbf{X_i}|C_j)$ be the class conditional probability that the pattern is $\mathbf{X_i}$, given that the pattern's state is C_j. If these statistics are known, *a posteriori* conditional probability $p(C_j|\mathbf{X_i})$ can be estimated using Bayes' rule as

$$p(C_j|\mathbf{X_i}) = \frac{p(\mathbf{X_i}|C_j) \cdot p(C_j)}{\sum_{h=1}^{c} p(\mathbf{X_i}|C_h) \cdot p(C_h)}, \quad i = 1, ..., n, \quad j = 1, ..., c. \tag{27}$$

If $p(C_h) = 1/c, \forall h \in \{1, c\}$, i.e., all states are assumed to be equally likely, then Equation (27) becomes

$$p(C_j|\mathbf{X_i}) = \frac{p(\mathbf{X_i}|C_j)}{\sum_{h=1}^{c} p(\mathbf{X_i}|C_h)}, \quad i = 1, ..., n, \quad j = 1, ..., c. \tag{28}$$

The following relationships hold true:

i) $p(C_j|\mathbf{X_i})$, $p(\mathbf{X_i}|C_j)$ and $p(C_j)$ belong to range $[0, 1]$;

ii) $\sum_{h=1}^{c} p(C_h|\mathbf{X_i}) = 1$, $i = 1, ..., n$, i.e., mixture components C_j, $j = 1, ..., c$, provide a complete partition of the input space; and

iii) $\sum_{h=1}^{c} p(C_h) = 1$.

From the comparison of Equation (22) with Equation (28) and properties (i)-(vi) in Section 6.1 with properties (i)-(iii) above we can write that,

if priors are considered the same (i.e., they are ignored), then $\{p(\mathbf{X_i}|C_j)\} \subset \{A_{i,j}\}$, thus, $\{p(C_j|\mathbf{X_i})\} \subset \{R_{i,j}\}$; in other words, *(objective) probability distributions are a subset of (useful) possibility distributions.*

6.3 Useful properties of fuzzy memberhips

Depending on the application, one absolute membership among Equations (23) to (26) may be preferred in the computation of Equation (22). As an example, let us consider a fuzzy clustering algorithm, where the distributed system consists of c PEs employing a relative membership function as their activation function. We intend to demonstrate that, under the hypothesis that learning rate $\beta_j(R_{i,j}) \in [0,1]$, $i = 1, ..., n$, $j = 1, ..., c$, is a monotonically increasing function of $R_{i,j}$ (e.g., Bezdek & Pal, 1993), if Equation (22) computes absolute typicality as either Equation (25) or Equation (26) then the network *may reach convergence faster* than by exploiting a relative membership function featuring Equation (23) or Equation (24) as its absolute typicality. Substituting Equation (25) in Equation (22) we obtain

$$R_{i,j} = \frac{1/d_{i,j}^2}{\sum_{h=1}^{c} 1/d_{i,h}^2}, \tag{29}$$

where resolution parameters p_j are ignored to simplify the discussion. Substituting Equation (26) in Equation (22) we obtain

$$R_{i,j} = \frac{1/(1 - Gaussian_{i,j})^2}{\sum_{h=1}^{c} 1/(1 - Gaussian_{i,h})^2)} = \frac{1/(1 - e^{-d_{i,j}^2})}{\sum_{h=1}^{c} 1/(1 - e^{-d_{i,h}^2})}, \tag{30}$$

where resolution parameters σ_j are ignored to simplify the discussion. Let us consider an initial situation in which, by randomization, template vectors (receptive field centers) of PEs match input patterns perfectly: this situation is characterized by an ideal requantization error equal to a global minimum (zero), i.e., the clustering system is expected to leave cluster centers unchanged from their initial position and reach termination after one processing cycle of the input data set (epoch). According to Equations (29) and (30), if for a given pattern $\mathbf{X_i}$ condition ($d_{i,j} = 0$) holds true, then $R_{i,j} = 1$ while $R_{i,h} = 0$, $\forall h \neq j$. Therefore, no template vector $\mathbf{W_h} \neq \mathbf{W_j}$ is moved by a fraction $\beta_h \propto R_{i,h}$ toward attractor $\mathbf{X_i}$, since this is already perfectly matched by template $\mathbf{W_j}$, i.e., no adaptation of receptive field centers takes place, as expected (since the initial condition is already optimal). Conversely, it is easy to demonstrate that receptive field centers are moved when the clustering net employs a relative membership (22) computing absolute typicality as either Equation (23) or Equation (24).

6.4 Fuzzy memberships and outlier detection in SART systems

To satisfy the soft competitive learning requirement (v) in Section 3.3, a clustering algorithm can compute activation values according to Equation (22), that guarantees context sensitivity (see Section 6.1). To pursue fast convergence, Equation (22) can be implemented as either Equation (29) or Equation (30) (see Section 6.3). However, Equations (29) and (30) are unable to detect noise points and outliers (see Section 6.1), as required by constraint (iii) in Section 3.3.

To summarize, our problem is: how can a clustering algorithm compute Equation (29) or Equation (30) while noise point and outlier detection is simultaneously guaranteed?

One possible solution consists of validating one node's activation $R_{i,j}$ iff its absolute membership term $A_{i,j}$ has passed a SART vigilance test, i.e., $A_{i,j}$ is above a given vigilance

threshold. In other words, "a node's activation represents its credit for the current input, and match provides a criterion for deciding if an input is an outlier for that category, and should be ignored" (anonymous referee).

For example, when Equation (30) is employed, outliers are detected by requiring that

$$A_{i,j} = 1/(1 - e^{-d_{i,j}^2}) \geq \rho^*, \quad \rho^* \in (1, \infty). \tag{31}$$

This is equivalent to constraining

$$Gaussian_{i,j} = e^{-d_{i,j}^2} \geq \rho = (\rho^* - 1)/\rho^*, \quad \rho \in (0, 1), \quad \rho^* \in (1, \infty). \tag{32}$$

Since $Gaussian_{i,j}$ belongs to the class of $\overline{M}$ functions and $\rho \in (0, 1)$, then vigilance test (32) is consistent with inequality (13) employed in SART framework (see Section 4).

One example of ART-based network that exploits a Gaussian mixture model of the input space and provides an *a posteriori* probability estimate iff class conditional likelihood satisfies a (S)ART-based bidirectional vigilance test is the Gaussian ARTMAP (GAM) model (Williamson, 1996, 1997).

7 Fuzzy SART

Two SART implementations can be found in the literature, employing a hard (WTA) and a soft competitive learning strategy respectively (Baraldi & Parmiggiani, 1995a, 1995b). In agreement with theoretical expectations (see Section 3.2), the soft competitive version performed better than the hard competitive one (Baraldi & Parmiggiani, 1995b). This development is quite similar to that regarding GAM, which was originally proposed as a hard competitive incremental algorithm (Williamson, 1996), then as a soft competitive (distributed learning) incremental algorithm (Williamson, 1997).

In this section we present a soft competitive SART implementation, termed Fuzzy SART, based on recommendations suggested in Section 3.3 to improve Fuzzy ART. This also means that Fuzzy SART combines SART architecture with probabilistic and possibilistic fuzzy membership functions as outlined in Section 6.4. This "fuzzification" process justifies exploitation of the name Fuzzy SART. With regard to learning strategy, Fuzzy SART is intended to combine useful properties driven from successful clustering algorithms, such as SOM and Neural Gas (NG, Martinetz, Berkovich & Schulten, 1993). To summarize, Fuzzy SART aims to provide a new synthesis between properties of ART, SOM and NG, to extend abilities of these separate approaches. While limitations of Fuzzy ART and improvements to this algorithm have been detailed in Sections 3.2 and 3.3, a brief review of SOM and NG is presented below.

7.1 Review of SOM and NG

Both SOM and NG satisfy the two Kohonen's constraints, introduced in Section 3.3, that should be met by the proposed Fuzzy SART algorithm as well. These two contraints derive from neurophysiological studies and provide an annealing schedule (Martinetz, Berkovich & Schulten, 1993; Ancona, Ridella, Rovetta & Zunino, 1997). They consist of two empirical functions of time, which must be user defined (Kohonen, 1990, 1995). The first Kohonen

heuristic rule requires that learning rates decrease monotonically with time according to a cooling scheme, i.e., as the number of processing epochs increases, all learning rates (winner as well as non-winner) decrease towards zero. Important properties of this cooling schedule have been analyzed by Bezdek & Pal (1993), Karayannis, Bezdek, Pal, Hathaway & Pai (1996), Ritter, Martinetz & Schulten (1992), Fritzke (1997a), Mulier & Cherkassky (1995a).

The second Kohonen heuristic rule requires that the size of the update (resonance) neighborhood centered on the winner node must decrease monotonically with time, such that a soft competitive learning strategy converges into a hard competitive (WTA) learning paradigm. This model transition is equivalent to stating that the initial overlap between nodes' receptive fields must decrease monotonically with time until it is reduced to zero, as hard competitive learning renders receptive fields equivalent to Voronoi polyhedra (Fritzke, 1997a). Interpretations of this second Kohonen heuristic rule, and relationships between SOM and other optimization techniques such as deterministic annealing (Rose, Guerewitz & Fox, 1995) and the Expectation-Maximization (EM) approach (Dempster, Laird & Rubin, 1977) are proposed in Luttrell (1990), Martinetz, Berkovich & Schulten (1993), Buhmann (1995), Mulier & Cherkassky (1995b), Alpaydın (1998). From a general perspective, it is important to remember that, compared to hard competitive learning, soft competitive learning not only decreases dependency on initialization (Martinetz, Berkovich & Schulten, 1993), but also reduces the presence of dead units (Fritzke, 1997a).

Despite its many successes in practical applications, SOM contains some major deficiencies (many of which are acknowledged in Kohonen, 1995), as listed below:
i) Termination is not based on optimizing any model of the process or its data (Tsao, Bezdek & Pal, 1994). Indeed, it has been shown that an objective function cannot exist for the SOM algorithm, i.e., there exists no cost function yielding Kohonen's adaptation rule as its gradient (Erwin, Obermayer & Schulten, 1992; Bishop, Svensen & C. Williams, 1996). SOM instead features a set of potential functions, one for each node, to be independently minimized following a stochastic (on-line) gradient descent (Erwin, Obermayer & Schulten, 1992).
ii) The size of the output lattice, the learning rate and the size of the resonance neighborhood must be varied empirically from one data set to another to achieve useful results (Tsao, Bezdek & Pal, 1994).
iii) Topology preserving mapping as defined by Martinetz, Berkovich & Schulten (1994) is not guaranteed.
iv) Prototype parameter estimates may be severely affected by noise points and outliers. This is due to the fact that learning rates in SOM are computed as a function of the number of processing epochs and node position in the grid, while they are independent of the actual distance separating the input pattern from the cluster template.

It is important to stress that while Kohonen's Vector Quantization (VQ) and SOM represent two important paradigms for information representation both in theory and in practice (Kohonen, 1995), another clustering algorithm, termed Neural Gas (NG, Martinetz, Berkovich & Schulten, 1993), has quickly gained popularity as a successful on-line vector quantizer (Fritzke, 1997a). NG implements a stochastic gradient descent of an analytical cost function, as opposed to SOM. Moreover, NG satisfies Kohonen's two constraints. In detail, NG implements model transitions from soft to hard competitive learning by: (a) employing metrical neighbors in the input space rather than topological neighbors belonging

to an output lattice as in SOM; and (b) sorting the activation values of processing elements as the only network-wide internode communication (Ancona, Ridella, Rovetta & Zunino, 1997). The main deficiencies of NG are that:
i) It employs a fixed user-defined number of clusters (this problem is closely linked to that of robustness, Davè & Krishnapuram, 1997).
ii) It does not preserve topological information.
iii) Prototype parameter estimates may be severely affected by noise points and outliers since learning rates in NG are computed as a function of the number of processing epochs and the rank of the reference vector, while they are independent of the actual distance separating the input pattern from the cluster template.

7.2 Fuzzy SART-specific features

To simplify the discussion, our hypothesis is to deal with a finite data set $\{\mathbf{X}\}$, consisting of n input patterns $\mathbf{X_i}$, $i = 1, ..., n$, where $\mathbf{X_i} \in \mathcal{R}^d$, which is repeatedly presented to the network until a termination criterion is satisfied. Each presentation sequence is termed a training epoch.

To run Fuzzy SART, the user specifies vigilance threshold $\rho \in [0, 1]$, and a lower limit for the number of epochs each node has to live through, $e_{min} \geq 1$, this parameter affecting the time required by the algorithm to reach termination.

The following steps characterize Fuzzy SART implementation, to be fit into the general SART framework proposed in Section 4.

Initialization. When output unit E_c is generated, its local (PE-based) time counter e_c is initialized to 0. Fuzzy SART employs PE-based time counters to compute PE-based plasticities (learning rates). In Fuzzy SART, the "age" (local time) of processing unit E_c is an integer value $e_c \geq 0$, equal to the number of times the finite input data set has been iteratively presented to the system while E_c exists. Although the presence of PE-based variables has never been stressed in the development of clustering algorithms featuring a fixed number of units, e.g., SOM and NG, it has been employed in Kohonen-based growing networks (Fritzke, 1994, 1995), as well as in GAM (to estimate priors, Williamson, 1997).

Activation function. The activation function is a relative membership defined as (22), i.e.,

$$\overline{AF}_1(\mathbf{W_j^{(t)}}, \mathbf{X_i^{(t)}}) = R_{i,j}^{(t)} = \frac{A_{i,j}^{(t)}}{\sum_{h=1}^{c} A_{i,h}^{(t)}}, \quad i = 1, ..., n, \quad j = 1, ..., c, \tag{33}$$

where absolute membership $A_{i,j}^{(t)}$ employs Equation (18) according to Equations (25) and (26) as

$$A_{i,j}^{(t)} = \frac{1}{[NVD(\mathbf{W_j^{(t)}}, \mathbf{X_i^{(t)}})]^2} = \frac{1}{[1 - VDM(\mathbf{W_j^{(t)}}, \mathbf{X_i^{(t)}})]^2}, \quad i = 1, ..., n, \quad j = 1, ..., c, \tag{34}$$

such that $A_{i,j}^{(t)} \in (1, \infty)$ since $VDM(\mathbf{W_j^{(t)}}, \mathbf{X_i^{(t)}}) \in (0, 1]$ (see Section 5.1).

Detection of processing units eligible for being resonant. In line with NG, Fuzzy SART applies a soft competitive learning mechanism based on *neighborhood-ranking* of metrical neighbors in the input space (Martinetz, Berkovich & Schulten, 1993; Fritzke, 1997a). Best ranking $r_{j^*}^{(t)} = 0$ is assigned to the best-matching unit E_{j^*} detected as (see Section 4)

$$j^* = \arg \max_{h=1,...,c} \{\overline{AF}_1(\mathbf{W_h^{(t)}}, \mathbf{X_i^{(t)}})\} = \arg \max_{h=1,...,c} \left\{R_{i,h}^{(t)}\right\},$$

which is equivalent to conditions

$$j^* = \arg \max_{h=1,...,c} \left\{A_{i,h}^{(t)}\right\} = \arg \max_{h=1,...,c} \left\{VDM(\mathbf{W_h^{(t)}}, \mathbf{X_i^{(t)}})\right\}. \tag{35}$$

Next, $r_j^{(t)} = 1$ if $\overline{AF}_1(\mathbf{W_j^{(t)}}, \mathbf{X_i^{(t)}})$ is second largest, etc. Both empirical evidence and theoretical predictions indicate that a few (five to ten) "top" positions in the list of sorted activation values (i.e., $r_j^{(t)} \in \{0, 9\}$) are sufficient to attain almost ideal results (Ancona, Ridella, Rovetta & Zunino, 1997).

Resonance domain detection. Given Equation (34), Equation (31) applied to the winner unit E_{j^*} becomes

$$A_{i,j^*}^{(t)} = \frac{1}{[1 - VDM(\mathbf{W_{j^*}^{(t)}}, \mathbf{X_i^{(t)}})]^2} \geq \rho^*, \quad \rho^* \in (1, \infty),$$

which is equivalent to constraint (see also Equation (32))

$$VDM(\mathbf{W_{j^*}^{(t)}}, \mathbf{X_i^{(t)}}) \geq \rho = (\rho^* - 1)/\rho^*, \quad \rho \in (0, 1), \quad \rho^* \in (1, \infty). \tag{36}$$

Owing to Equations (13) and (36), Fuzzy SART match function is defined as Equation (17), i.e.,

$$\overline{MF}_1(\mathbf{W_{j^*}^{(t)}}, \mathbf{X_i^{(t)}}) \equiv VDM(\mathbf{W_{j^*}^{(t)}}, \mathbf{X_i^{(t)}}).$$

This value has already been computed when Equation (34) is processed. Note that condition $\overline{MF}_1(\mathbf{W_1}, \mathbf{X}) > \overline{MF}_1(\mathbf{W_2}, \mathbf{X})$ implies that $\overline{AF}_1(\mathbf{W_1}, \mathbf{X_i}) > \overline{AF}_1(\mathbf{W_2}, \mathbf{X_i})$, and vice versa, $\forall \mathbf{W_1}, \mathbf{W_2}$ and $\mathbf{X} \in \mathcal{R}^d$. This means that the corollary presented in Section 4 applies to Fuzzy SART, i.e., to detect the resonance domain, Fuzzy SART requires no mismatch reset condition and search process.

Resonance condition: weight updating. A Kohonen weight adaptation law is applied to all processing units belonging to the resonance domain as

$$W_{k,j}^{(t+1)} = W_{k,j}^{(t)} + \beta_j^{(t)} \cdot (X_{k,i}^{(t)} - W_{k,j}^{(t)}), \quad k = 1, ..., d, \quad i = 1, ..., n, \quad \forall j \in \{1, c\} : r_j^{(t)} \in \{0, 9\}. \tag{37}$$

In line with the NG algorithm, processing unit E_j, whose local epoch counter is e_j, features learning rate $\beta_j^{(t)}$ defined as

$$\beta_j^{(t)} = \epsilon_j^{(t)} \cdot h_j^{(t)}, \quad \forall j \in \{1, c\} : r_j^{(t)} \in \{0, 9\}, \tag{38}$$

where $\beta_j^{(t)}$, $\epsilon_j^{(t)}$, and $h_j^{(t)}$ all belong to range [0,1], such that $\beta_j^{(t)} \leq \min\{\epsilon_j^{(t)}, h_j^{(t)}\}$. In detail,

$$\epsilon_j^{(t)} = \epsilon_j^{(t)}(e_j, R_{i,j}^{(t)}) = \epsilon_{ini}(\epsilon_{fin}/\epsilon_{ini})^{e_j/e_{min}}, \tag{39}$$

where e_{min} is the user-defined decay parameter described above, while variables ϵ_{ini} and ϵ_{fin} are computed as

$$1 \geq \epsilon_{ini} = \max\{R_{i,j}^{(t)}, \epsilon\} \geq \epsilon_{fin} = \min\{R_{i,j}^{(t)}, \epsilon\} > 0, \tag{40}$$

where parameter ϵ is the maximum lower limit for the learning rate (e.g., $\epsilon = 0.005$ is fixed by the application developer). Coefficient $\epsilon_j^{(t)}$ is monotonically nonincreasing with e_j and monotonically nondecreasing with $R_{i,j}^{(t)}$. Owing to the exploitation of $R_{i,j}^{(t)}$ in Equation (40), $\epsilon_j^{(t)}$ depends on the entire set of distances between prototypes and the input pattern. Experimental evidence shows that if the term $R_{i,j}^{(t)}$ is replaced by $A_{i,j}^{(t)}$ in the computation of ϵ_{ini} and ϵ_{fin}, then Fuzzy SART shows a tendency to produce coincident clusters, in line with possibilistic clustering algorithms (see Section 6.1). In Equation (38), term $h_j^{(t)}$ reduces the overlap between node receptive fields according to the following expression

$$h_j^{(t)} = h_j^{(t)}[r_j^{(t)}, \sigma_j(e_j)] = e^{-r_j^{(t)}/\sigma_j(e_j)}, \tag{41}$$

where $r_j^{(t)}$ is the neighborhood ranking of node E_j and $\sigma_j(e_j)$ is a spread value computed as a monotonically decreasing function of time, e.g.,

$$\sigma_j(e_j) = \sigma_{ini}(\sigma_{fin}/\sigma_{ini})^{e_j/e_{min}}, \tag{42}$$

where $\sigma_{ini} \geq \sigma_{fin}$. Widely employed settings for these parameters are $\sigma_{ini} = 5$, and $\sigma_{fin} = 0.01$ (Martinetz, Berkovich & Schulten, 1993; Ancona, Ridella, Rovetta & Zunino, 1997). Thus, learning coefficient $h_j^{(t)}$ is monotonically decreasing with neighborhood ranking $r_j^{(t)}$ and time e_j if $j \neq j^*$.

Controls at epoch termination. When the entire input data set is presented to the system, i.e, if $[(t\%n) = 0]$, where operator % computes the remainder of t divided by n, then the following operations occur: a) PE-based time (epoch) counters are incremented by one as $e_j = e_j + 1$, $j = 1, ..., c$; and b) superfluous cells are removed, such that, $\forall j \in \{1, c\}$, if processing element E_j has not been the best-matching unit for any pattern assignment during the last processing epoch, then it is removed, and PE counter is decreased as $c = c-1$.

7.3 Fuzzy SART complexity

As with the NG algorithm, the computationally expensive part of Fuzzy SART is the determination of "neighborhood-ranking", whose computation time increases as $c \log c$ (Martinetz, Berkovich & Schulten, 1993). The same ranking mechanism is employed by the Fuzzy ART mismatch reset condition and search process to detect the winner node (see Section 3.1). We conclude that serial implementations of NG, Fuzzy SART and Fuzzy ART share the same computational complexity. About this conclusion, several considerations

can be made. If Fuzzy ART employs normalization preprocessing, it loses vector-length information. If Fuzzy ART employs complement coding, it doubles its storage requirement and computation time with respect to that of Fuzzy SART. Finally, Fuzzy ART is hard competitive while Fuzzy SART is soft competitive, i.e., the latter system is more efficient, performing more complex learning activities in the same number of steps.

7.4 Advantages of Fuzzy SART

The simple example given in Appendix C shows that Fuzzy SART features several properties of interest:

i) It is less sensitive than Fuzzy ART to the order of presentation of the random sequence.

ii) It improves processing time of Fuzzy ART by employing no search process.

iii) Its vigilance scale is more sensitive than the one employed by Fuzzy ART (to obtain the same number of clusters, Fuzzy ART requires a larger value of the vigilance threshold).

iv) It is both stable and plastic, owing to the combination of its update strategy and dynamic allocation and removal of processing resources (once a template has reached termination, it is not moved by subsequent training sessions).

v) It avoids input data pre-processing such as normalization or complement coding.

vi) In line with recommendation (iv) of Section 3.3, it satisfies the first Kohonen constraint as learning coefficient $\epsilon_j^{(t)}(e_j, R_{i,j})$ decreases monotonically with time e_j, see Equations (38) and (39).

vii) In line with recommendation (v) of Section 3.3, it satisfies the second Kohonen constraint, so that a model transition scheme from soft to hard competitive learning is pursued, as learning rate coefficient $h_j^{(t)}[r_j^{(t)}, \sigma_j(e_j)]$ decreases monotonically with e_j and $r_j^{(t)}$ if $j \neq j^*$, see Equations (38), (41) and (42).

viii) It features enhanced robustness against noise by means of coordinated actions which are summarized as follows. a) Noise points do not affect existing prototypes. b) Since a detected noise pattern (i.e., a pattern that does not pass the vigilance test) is sufficient to generate one new category, a category removal mechanism is provided to avoid overfitting.

7.5 Weaknesses and possible developments of Fuzzy SART

The main deficiencies of Fuzzy SART are that:

i) Since metric NVD is not invariant to translation (see Section 5.1.2), it does not apply successfully to the Euclidean space, as shown in Section 9. If Fuzzy SART is applied to the Euclidean space, then its PEs should not employ Equation (17) in their activation func-

tion, but rather employ the Euclidean norm, as in Equation (26) (Baraldi & Parmiggiani, 1997).

ii) It does not preserve topological information. To overcome this problem, Fuzzy SART can be provided with the *Competitive Hebbian Learning* mechanism (CHL) that introduces competition among synaptic links (Martinetz, Berkovich & Schulten, 1994). According to CHL, during on-line processing of an input pattern *the two template vectors closest to the pattern must be detected, then the two output nodes providing these templates must be connected by an edge (synaptic link).* This rule generates an output graph, termed the *induced Delaunay triangulation*, which preserves topology optimally in a very general sense. Fuzzy SART provided with CHL is termed Fully self-Organizing SART (FOSART). Note that FOSART requires no ranking of the set of output values provided by the battery of attentional nodes, since the update neighborhood is identified as the set of nodes topologically connected to the winner unit (Fritzke, 1997a). To summarize, exploitation of the competitive Hebbian learning mechanism allows FOSART to: (a) develop an output map (grid) providing a topologically correct mapping of input patterns; and (b) reduce the computational complexity of Fuzzy SART by employing no ranking mechanism to perform soft competitive learning (Baraldi & Parmiggiani, 1997).

iii) Termination is not based on optimizing any model of the process or its data, although some relationships with the learning policy of the NG algorithm, which performs stochastic gradient descent of an analytical cost function, have been established.

8 Experimental comparison of Fuzzy ART and Fuzzy SART

Our aim is to assess the functional consequences of the different Fuzzy ART and Fuzzy SART architectures. Applications to simple data sets are sufficient to let these functional differences emerge naturally.

The Fuzzy SART and Fuzzy ART systems belong to the class of squared-error clustering programs (Dubes & Jain, 1976). To compare their performances, these two systems must be applied to several data sets to detect the same number of output categories while the following features are considered (Dubes & Jain, 1976): i) the value of their squared error criterion; ii) the number of misclassified input patterns; iii) the number of iterations before reaching termination (epochs); iv) the sensitivity (stability) of the two algorithms to the order of the training sequence; and v) the sensitivity (stability) of the two algorithms to their input parameters.

Since squared-error clustering algorithms detect groups of patterns that are hyperspherical or hyperellipsoidal in shape, two simple data sets are selected from the literature. The first data set, shown in Fig. 5, is two-dimensional and consists of 24 points (Simpson, 1993). The second data set is the 4-dimensional standard IRIS data set, consisting of 50 vectors for each of 3 classes (Fisher, 1936). Exploitation of the IRIS data set allows comparison of Fuzzy SART and Fuzzy ART with other clustering models found in the literature. Typical error rates for unsupervised categorization of the IRIS data set are 10-16 mistakes (Bezdek & Pal, 1995; Tsao, Bezdek & Pal, 1994). Despite their simplicity, these two data sets are sufficient to reveal the different functional properties characterizing Fuzzy ART and Fuzzy

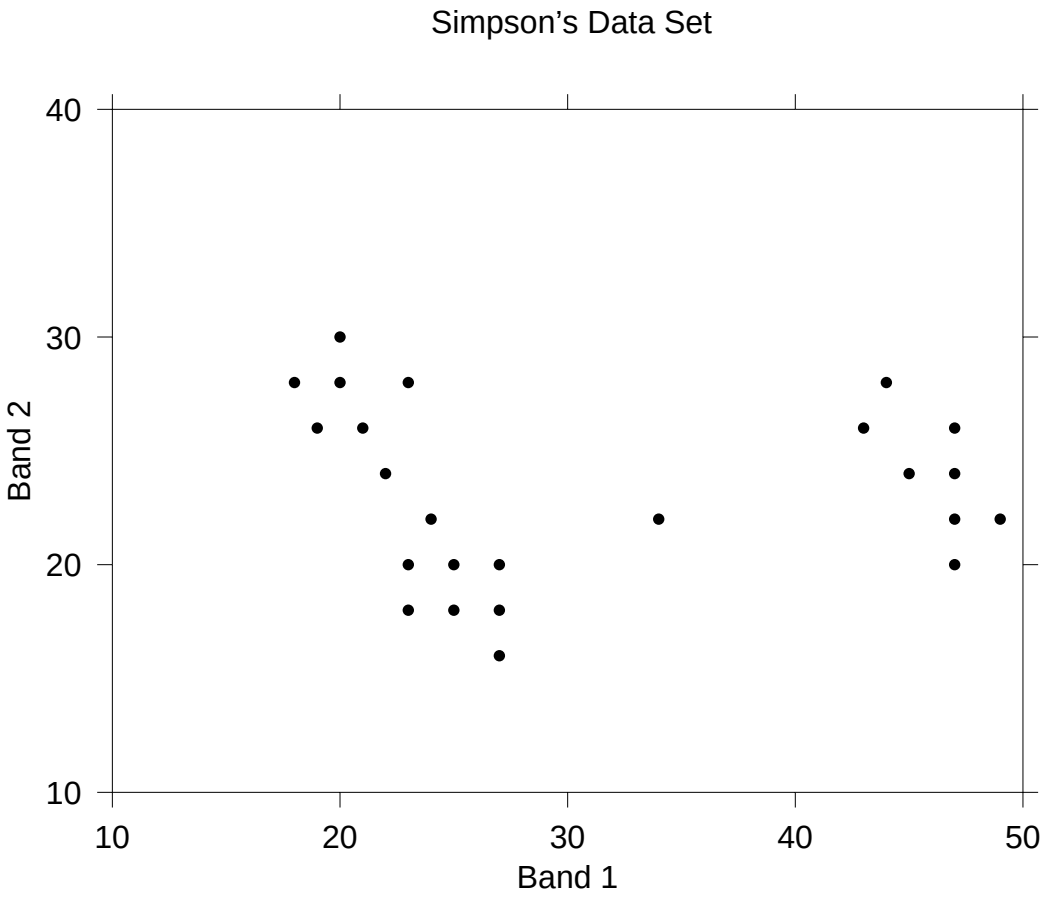


Figure 5: Two-dimensional data set employed for testing (Simpson, 1993).

SART.

8.1 Experimental set up

Although Fuzzy ART requires normalization of input patterns to avoid category proliferation (see Section 3), we employ Fuzzy ART without any normalization step because: i) no category proliferation affects Fuzzy ART in the clustering of either the Simpson or the IRIS data set; and ii) no normalization has been employed in several clustering examples of the IRIS data set to be found in the literature for performance comparison.

When Fuzzy ART is employed with no normalization preprocessing, experimental evidence reveals that weight adaptation law (11) becomes completely inadequate and is then replaced by the Kohonen law (37), where learning rate β_j is derived from Equation (39) as

$$\beta_j = \epsilon_{ini}(\epsilon_{fin}/\epsilon_{ini})^{e_j/e_{min}}, \tag{43}$$

where variable e_j and parameter e_{min} are defined as in Equation (39), parameter ϵ_{fin} is fixed to 0.005 as parameter ϵ in Equation (39), and parameter ϵ_{ini} is user-defined such that inequality $\epsilon_{ini} > \epsilon_{fin}$ holds true. Owing to the proposed substitutions, the Mean

Square Quantization Error (MSE) of the Fuzzy ART algorithm in the clustering of both the Simpson and IRIS data set decreased by a factor of up to 70 %. At the same time, in line with Fuzzy ART, we force Fuzzy SART to employ a hard competitive strategy exclusively by setting parameters $\sigma_{ini} = \sigma_{fin} = 0.0001$ in (42), so that, in (38), coefficient $h_{j^*} = 1$ for the winner node E_{j^*}, while $h_j^{(t)} = 0$, $\forall j \neq j^*$. Since these choices reduce the degree of difference between learning and termination strategies adopted by Fuzzy ART and Fuzzy SART, we expect functional differences between these two clustering models, if any, to emerge naturally as a consequence of their alternative solutions in the choice of the activation and match function pair.

Each system was tested several times with different values for free parameters chosen within reasonable ranges. The Fuzzy SART input parameters are $\rho \in (0, 1)$ and $e_{min} \in [1, +\infty)$. The three Fuzzy ART input parameters are vigilance threshold $\rho \in (0, 1)$, $\epsilon_{ini} \in (0.0005, 1]$ (see above) and $e_{min} \in [1, +\infty)$. Parameter α in Equation (4) is set to 0.001 (Huang, Georgiopoulos & Heileman, 1995).

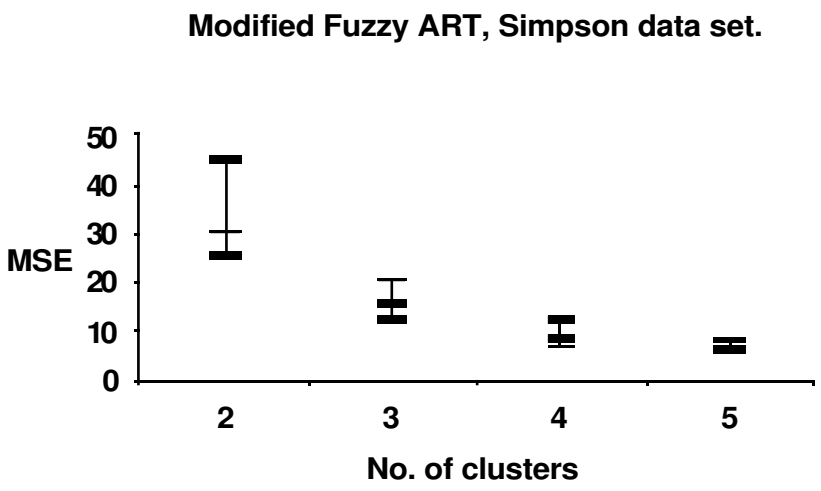
8.2 Two-dimensional data clustering

Ten different sequences of the Simpson data set are iteratively presented to the Fuzzy SART and the modified Fuzzy ART systems. The number of detected clusters is constrained to vary from 2 to 5, values considered reasonable in light of a visual inspection of Fig. 5, while e_{min} is sequentially fixed at 10, 20 and 100. Parameter ϵ_{ini} employed by Fuzzy ART is set to 0.05 and 0.2. Overall, Fuzzy SART was run 30 times for each number of clusters, while Fuzzy ART was run 60 times. Fig. 6 shows the average, best and worst MSE values obtained with the two clustering systems. For the same combination of input parameters and by changing the list presentation, Fuzzy ART detected a varying number of clusters in 30.5 % of the measured cases, with an average standard deviation of 0.217 clusters, while Fuzzy SART detected a varying number of clusters in 25 % of the measured cases, with an average standard deviation of 0.125 clusters. Our conclusion is that Fuzzy ART is more sensitive than Fuzzy SART to small changes in input parameters and in the order of the presentation sequence in the clustering of the Simpson data set.

8.3 IRIS data clustering

To relate unsupervised categories to the IRIS labeled classes of patterns, a relabeling algorithm is employed as follows: each category is associated with the class providing the majority of category activation patterns. Input parameters are adjusted until the number of detected categories is equal to 3, 5, 8 and 12 respectively. Tables 1 and 2 shows the best performance of Fuzzy ART and Fuzzy SART respectively in terms of MSE minimization.

Overall, Fuzzy SART is superior to Fuzzy ART with respect to both MSE minimization and pattern misclassification. When the number of detected clusters is 3, the value of misclassified patterns makes Fuzzy SART competitive with other clustering models found in the literature (Bezdek & Pal, 1995), (Kim & Mitra, 1993), while Fuzzy ART is by no means competitive. For example, Fuzzy SART performs better than: i) the Fuzzy Min-Max clustering model (see Fig. 10 in Simpson (1993), where the smallest number of misclassified patterns is 18 when the number of clusters is 3); ii) the Fuzzy c-means algorithm, affected



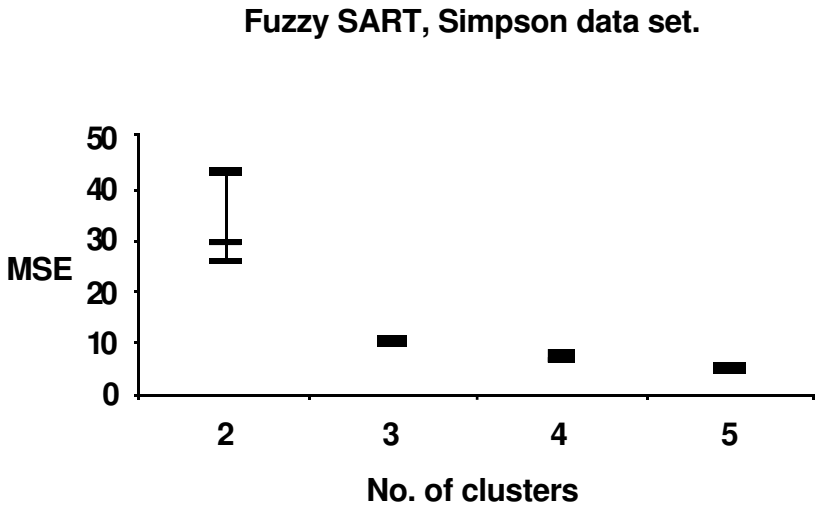


Figure 6: Clustering of the Simpson data set. (a) Fuzzy ART performance with average, maximum and minimum MSE values. Codebooks from size 2 to 5 were generated. Sixty simulations were performed for each codebook size. (b) Fuzzy SART performance with average, maximum and minimum MSE values. Codebooks from size 2 to 5 were generated. Thirdy simulations were performed for each codebook size.

Table 1: Input parameters and output values of the best Fuzzy ART performance in the categorization of the IRIS data set.

Free parameter values*	**No. of clusters**	**MSE**	**Misclassified patterns**	**No. of iterations**
$\rho = 0.82,\ \epsilon_{ini} = 0.05,\ e_{min} = 50$	3	0.804	43	50
$\rho = 0.90,\ \epsilon_{ini} = 0.20,\ e_{min} = 10$	5	0.377	16	10
$\rho = 0.935,\ \epsilon_{ini} = 0.20,\ e_{min} = 10$	8	0.246	17	10
$\rho = 0.944,\ \epsilon_{ini} = 0.20,\ e_{min} = 50$	12	0.210	16	50

* $\alpha = 0.001$; $\epsilon_{fin} = 0.005$.

Table 2: Input parameters and output values of the best Fuzzy SART performance in the categorization of the IRIS data set.

Free parameter values*	**No. of clusters**	**MSE**	**Misclassified patterns**	**No. of iterations**
$\rho = 0.36,\ \epsilon_{min} = 50$	3	0.555	11	50
$\rho = 0.60,\ \epsilon_{min} = 10$	5	0.333	18	10
$\rho = 0.75,\ \epsilon_{min} = 10$	8	0.204	4	10
$\rho = 0.78,\ \epsilon_{min} = 10$	12	0.174	3	10

* $\sigma_{ini} = \sigma_{fin} = 0.0001$; $\epsilon_{fin} = 0.005$.

by 15 misclassifications (Kim & Mitra, 1993); and iii) the Kohonen VQ algorithm, affected by 17 misclassifications (Kim & Mitra, 1993). Table 3 gives the numerical values of the physically labeled IRIS subsample means. Tables 4 and 5 report Fuzzy SART and Fuzzy ART terminal centroids when the number of detected categories is equal to 3. Tables 6 and 7 present the confusion matrices of Fuzzy ART and Fuzzy SART respectively, when the number of detected categories is equal to 3.

9 Conclusions

When applied to simple data sets, the ART 1-based Fuzzy ART system is not stable with respect to small changes in input parameters and in the order of the training sequence. This experimental evidence is in line with what is found in the literature concerning the ART 1 system. Our work is focused on detecting structural problems that affect ART 1-based system design. One potential problem is identified in the sequential exploitation of two complementary unidirectional (asymmetric) functions to compute an inherently symmetric interpattern similarity value. An alternative ART-based general framework, termed SART, is then presented to extract statistical regularities from analog samples. This new frame-

Table 3: Numerical values of the centers of the IRIS classes.

	Band 1	**Band 2**	**Band 3**	**Band 4**
Class 1	5.006	3.428	1.462	0.246
Class 2	5.936	2.770	4.260	1.326
Class 3	6.588	2.974	5.552	2.026

Table 4: Numerical values of the reference vectors detected by Fuzzy ART in the processing of the IRIS data set. $\rho = 0.82$, $\epsilon ini = 0.05$, $e_{min} = 50$. No. of clusters = 3.

	Band 1	**Band 2**	**Band 3**	**Band 4**
Class 1	5.004	3.426	1.462	0.246
Class 2	6.185	2.857	4.854	1.680
Class 3	7.619	3.222	6.484	2.196

Table 5: Numerical values of the reference vectors detected by Fuzzy SART in the processing of the IRIS data set. $\rho = 0.36$, $e_{min} = 50$. No. of clusters = 3.

	Band 1	**Band 2**	**Band 3**	**Band 4**
Class 1	5.003	3.425	1.461	0.245
Class 2	5.802	2.733	4.230	1.348
Class 3	6.762	3.036	5.613	2.020

work employs bidirectional activation and match functions computing interpattern similarity values as relative numbers. A specific SART implementation, termed Fuzzy SART, is developed to take advantage of the combination between the SART framework and useful absolute and relative fuzzy membership functions.

According to theoretical and experimental considerations, Fuzzy SART features several interesting properties when compared to existing clustering algorithms: i) the system is easy to use, requiring only two main parameters having an intuitive physical meaning; ii) unlike Fuzzy ART, the system requires no input data preprocessing; iii) unlike SOM and NG, the system requires no *a priori* knowledge of the size of the network; iv) unlike SOM, the system requires no *a priori* knowledge of the topology of the network; v) unlike SOM and NG, the system requires no randomization of the initial templates; vi) unlike SOM and NG, the system is capable of detecting outliers; vii) unlike Fuzzy ART, the system is capable of removing noise categories to avoid overfitting; viii) unlike Fuzzy ART, the system is fairly stable with respect to small changes in input parameters and in the order of the presentation sequence; and ix) the number of misclassified patterns detected by Fuzzy SART in the processing of the IRIS data set is competitive with that of other clustering models found in the literature.

Failure modes of the proposed algorithm are that: i) Fuzzy SART does not manage

Table 6: Confusion matrix generated by Fuzzy ART clustering of the IRIS data set. $\rho = 0.82$, $e_{min} = 50$, $\epsilon_{ini} = 0.05$. No. of clusters = 3.

	Category 1	**Category 2**	**Category 3**
Class 1	50		
Class 2		50	
Class 3		43	7

Table 7: Confusion matrix generated by Fuzzy SART clustering of the IRIS data set. $\rho = 0.36$, $e_{min} = 50$. No. of clusters = 3.

	Category 1	**Category 2**	**Category 3**
Class 1	50		
Class 2		42	8
Class 3		3	47

topological information, i.e., it cannot provide topologically correct mapping, although this property is incorporated in FOSART (see Section 7.5); ii) if it employs similarity measure VDM, then Fuzzy SART should not be applied to data sets belonging to the Euclidean space, as shown in Figs. 7 and 8, where an input 3-D digitized human face (Borghese, Ferrigno, Baroni, Savarè, Ferrari & Pedotti, 1998), and the output resampled data set are shown respectively; and ii) Fuzzy SART does not minimize any known objective function.

Preliminary results of the new version of Fuzzy SART, termed FOSART (see Section 7.5), are encouraging, as shown in Figs. 9 and 10 (Baraldi & Parmiggiani, 1997).

Both Fuzzy SART and FOSART can be employed as the hidden layer in any two-layer supervised system computing function approximation through scatter-partitioning (Fritzke, 1994, 1997b). In these systems, supervised errors at the output are taken into account to determine number and scatter position of PEs belonging to the hidden clustering layer (Alpaydın, 1998; Bishop, 1995; Fritzke, 1994, 1997b).

Acknowledgments

We are grateful to R. Savarè, G. Ferrigno, N. A. Borghese and S. Ferrari for providing us with the 3-D digitized human face data set. Andrea Baraldi thanks F. Parmiggiani for his support. Both authors wish to thank Prof. J. Feldman for the stimulating environment at ICSI, and the anonymous referees for their valuable comments. Ethem Alpaydın is a Fullbright Scholar.

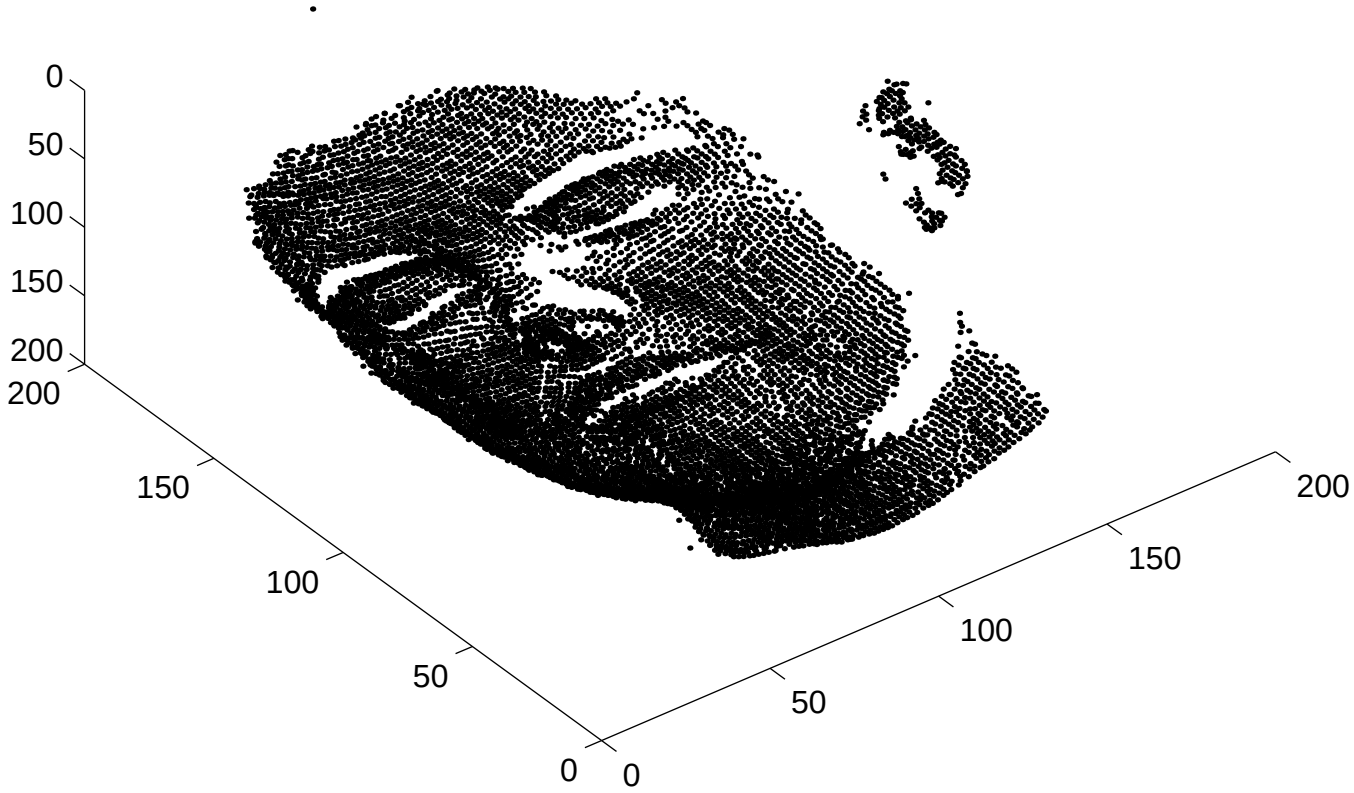


Figure 7: 3-D digitized human face consisting of 9371 vectors.

Appendix A

Let us consider two binary vectors $\mathbf{X^{(t)}}$ and $\mathbf{W_{j^*}^{(t)}}$. Since $X_k^{(t)} \in \{0,1\}$, $k = 1, ..., d$, we can write

$$\|\mathbf{X^{(t)}}\| = \sum_{k=1}^{d} X_k^{(t)} = \sum_{k=1}^{d} (X_k^{(t)})^2 = \mid \mathbf{X^{(t)}} \mid^2, \tag{44}$$

where $\|\mathbf{X^{(t)}}\|$ is the norm of the input vector and

$$\mid \mathbf{X^{(t)}} \mid = \sqrt{\sum_{k=1}^{d} (X_k^{(t)})^2} \tag{45}$$

is the modulus of $\mathbf{X^{(t)}}$. In line with Equation (44) we can also write

$$\|\mathbf{W_{j^*}^{(t)}}\| = \mid \mathbf{W_{j^*}^{(t)}} \mid^2. \tag{46}$$

Substituting (44) and (46) in (7), then

$$\vec{MF}(\mathbf{W_{j^*}^{(t)}}, \mathbf{X}^{(t)}) = \frac{\mathbf{X^{(t)}} \circ \mathbf{W_{j^*}^{(t)}}}{\mid \mathbf{X^{(t)}} \mid^2} = \frac{\mid \mathbf{X^{(t)}} \mid \cdot \mid \mathbf{W_{j^*}^{(t)}} \mid \cdot \cos\theta_{j^*}}{\mid \mathbf{X^{(t)}} \mid^2} = \frac{\mid \mathbf{W_{j^*}^{(t)}} \mid \cdot \cos\theta_{j^*}}{\mid \mathbf{X^{(t)}} \mid}, \tag{47}$$

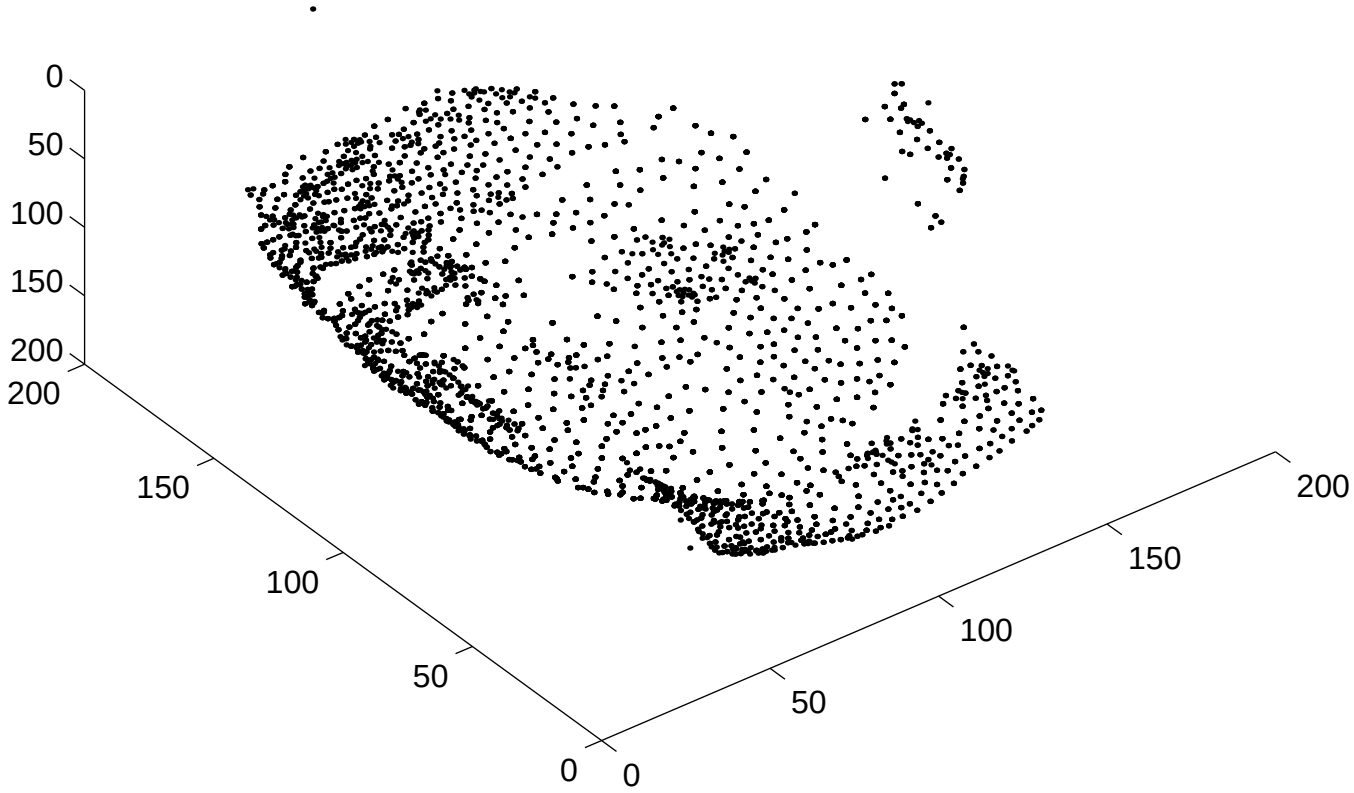


Figure 8: Reference vectors detected by Fuzzy SART when the digitized human face data set is processed. Output information is: No. of nodes = 1775, MSE = 5.810, No. of iterations = 69.

where $\mathbf{X^{(t)}} \circ \mathbf{W_{j^*}^{(t)}}$ is the scalar (dot) product between $\mathbf{X^{(t)}}$ and $\mathbf{W_{j^*}^{(t)}}$, and θ_{j^*} is the angle between $\mathbf{X^{(t)}}$ and $\mathbf{W_{j^*}^{(t)}}$.

Appendix B

As is true for ART 1 (see Section 1), Fuzzy ART, which is ART 1-based, is also expected to be sensitive to changes in the order of presentation of the random sequence.

Let us consider the following example. The input parameters are $\rho = 0.55$, $\alpha = 0.001$, $\beta = 1$ (see Section 3.1). The presentation list is $\mathbf{X^{(1)}} = (0, 0, 1, 0, 0)$, $\mathbf{X^{(2)}} = (0, 1, 1, 1, 1)$, and $\mathbf{X^{(3)}} = (1, 1, 1, 0, 0)$. This is submitted to the Fuzzy ART preprocessing normalization step. The presentation list becomes $\mathbf{X^{(1)}} = (0, 0, 1, 0, 0)$, $\mathbf{X^{(2)}} = (0, 1/2, 1/2, 1/2, 1/2)$, and $\mathbf{X^{(3)}} = (1/\sqrt{3}, 1/\sqrt{3}, 1/\sqrt{3}, 0, 0)$.

Patterns $\mathbf{X^{(1)}}$ and $\mathbf{X^{(2)}}$ generate two categories $\mathbf{W_1^{(3)}} = \mathbf{W_1^{(2)}} = \mathbf{X^{(1)}}$ and $\mathbf{W_2^{(3)}} = \mathbf{X^{(2)}}$ respectively (since vigilance test (3) is such that $\vec{MF}_1(\mathbf{W_1^{(2)}}, \mathbf{X^{(2)}}) = 0.5/\ (4 \cdot 0.5) = 0.25 < \rho$). The winner template for pattern $\mathbf{X^{(3)}}$ is chosen according to Equations (2) and (4) as $j^* = \arg\max\{\frac{1/\sqrt{3}}{0.001+1}, \frac{2 \cdot 0.5}{0.001+4 \cdot 0.5}\} = \arg\max\{0.576, 0.499\} = 1$, i.e., $\mathbf{W_{j^*}^{(3)}} = \mathbf{W_1^{(3)}}$. The vigilance test (3) is such that: $\vec{MF}_1(\mathbf{W_1^{(3)}}, \mathbf{X^{(3)}}) = (1/\sqrt{3})/(3 \cdot (1/\sqrt{3})) = 0.33 < \rho$. Thus,

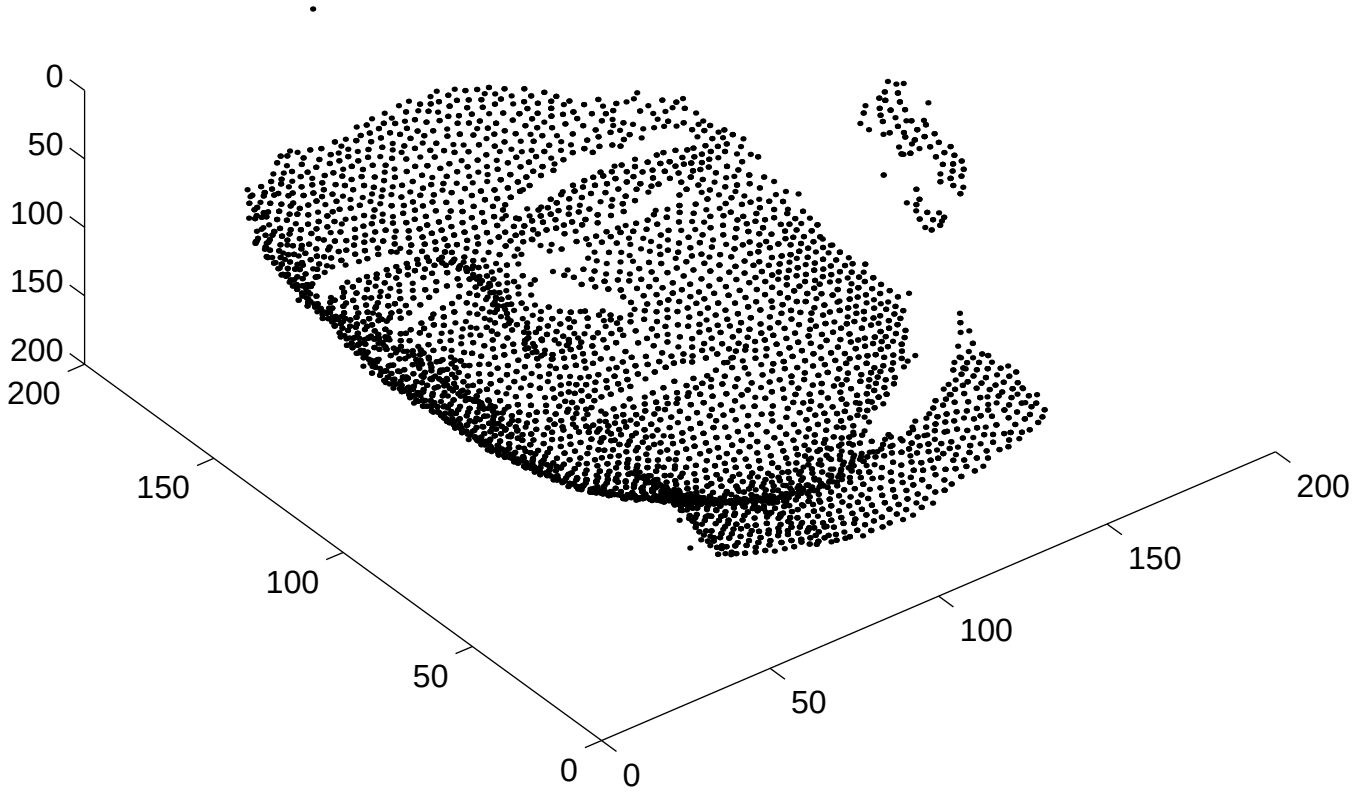


Figure 9: Reference vectors detected by FOSART when the digitized human face data set is processed. Output information is: No. of nodes = 3370, No. of submaps = 60, $MSE = 1.44$, No. of iterations = 10.

the reset condition and search process is started. The second-best template is $\mathbf{W_2^{(3)}}$. Then, $\vec{MF}_1(\mathbf{W_2^{(3)}}, \mathbf{X^{(3)}}) = (2 \cdot 0.5)/(3 \cdot (1/\sqrt{3})) = 0.577 > \rho$. Since the vigilance test is satisfied, then fast category adaptation (11) leads to $\mathbf{W_2^{(4)}} = (0, 0.5, 0.5, 0, 0)$. Thus, final templates are $\mathbf{W_1^{(4)}} = \mathbf{X^{(1)}}$, while $\mathbf{W_2^{(4)}} = (0, 0.5,\ 0.5, 0, 0)$.

Let us consider a different order of the input sequence where the input vectors described above are presented as follows: $\mathbf{X^{(1)}} = (0,\ 0,\ 1,\ 0,\ 0)$, $\mathbf{X^{(2)}} = (1, 1, 1, 0, 0)$, and $\mathbf{X^{(3)}} = (0, 1, 1, 1, 1)$. Due to input pattern normalization, the presentation list becomes $\mathbf{X^{(1)}} = (0, 0, 1, 0, 0)$, $\mathbf{X^{(2)}} = (1/\sqrt{3}, 1/\sqrt{3}, 1/\sqrt{3}, 0, 0)$, and $\mathbf{X^{(3)}} = (0,\ 1/2,\ 1/2,\ 1/2,\ 1/2)$. Patterns $\mathbf{X^{(1)}}$ and $\mathbf{X^{(2)}}$ generate two categories $\mathbf{W_1^{(3)}} = \mathbf{W_1^{(2)}} = \mathbf{X^{(1)}}$ and $\mathbf{W_2^{(3)}} = \mathbf{X^{(2)}}$ respectively (since vigilance test (3) is such that $\vec{MF}_1(\mathbf{W_1^{(2)}}, \mathbf{X^{(2)}}) = (1/\sqrt{3})/(3/\sqrt{3}) = 0.33 < \rho$). The winner template for pattern $\mathbf{X^{(3)}}$ is chosen according to (2) and (4) as $j^* = \arg\max\{\frac{0.5}{0.001+1}, \frac{2 \cdot 0.5}{0.001+3 \cdot (1/\sqrt{3})}\} = \arg\max\{0.499, 0.576\} = 2$, i.e., $\mathbf{W_{j^*}^{(3)}} = \mathbf{W_2^{(3)}}$. Vigilance test (3) is such that: $\vec{MF}_1(\mathbf{W_2^{(3)}}, \mathbf{X^{(3)}}) = (2 \cdot 0.5)/(4 \cdot 0.5) = 0.5 < \rho$. Thus, the reset condition and search process are started. The second-best template is $\mathbf{W_1^{(3)}}$. In this case, $\vec{MF}_1(\mathbf{W_1^{(3)}}, \mathbf{X^{(3)}}) = 0.5/(4 \cdot 0.5) = 0.25 < \rho$. Since the vigilance test is not satisfied, then a new category is dynamically allocated so that final templates are $\mathbf{W_1^{(4)}} = \mathbf{X^{(1)}}$, $\mathbf{W_2^{(4)}}$

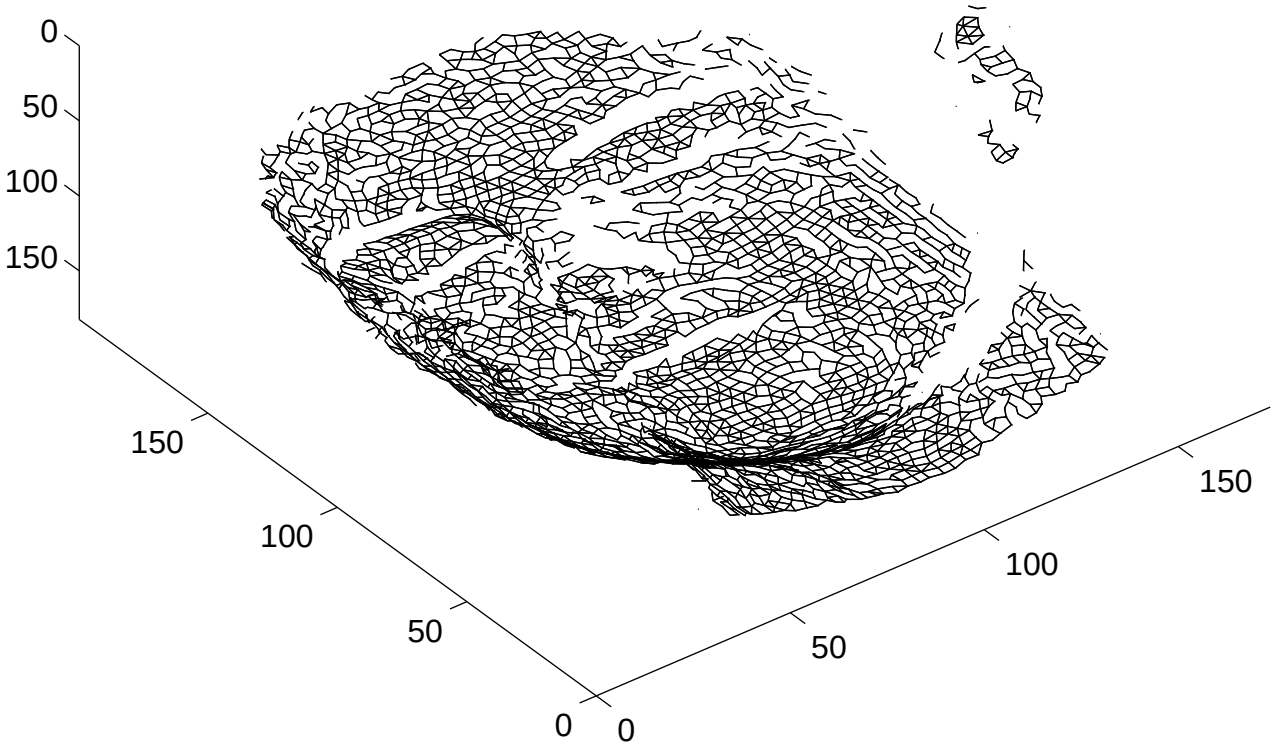


Figure 10: Projection onto input space of synaptic links detected by FOSART in the output map of the digitized human face data set.

$= \mathbf{X^{(2)}}$, and $\mathbf{W_3^{(4)}} = \mathbf{X^{(3)}}$.

Clustering results obtained when Fuzzy ART processes the two presentation sequences are inconsistent in terms of number of clusters.

Appendix C

Let us consider the same example employed to test Fuzzy ART in Appendix B. Note that since Fuzzy ART has employed a normalized input data set, then differences in vector length between input pattern pairs are lost, i.e., term MDM, which is equal to 1 according to Equation (14), is useless in the computation of similarity value VDM employed in Equations (34) and (36). The input parameter is set to $\rho = 0.4$. This value is sufficiently large to guarantee detection of more than one cluster. Nonetheless, the results of this experiment will be easy to generalize.

The presentation list is $\mathbf{X^{(1)}} = (0,\ 0,\ 1,\ 0,\ 0)$, $\mathbf{X^{(2)}} = (0,\ 1/2,\ 1/2,\ 1/2,\ 1/2)$, and $\mathbf{X^{(3)}} = (1/\sqrt{3},\ 1/\sqrt{3},\ 1/\sqrt{3},\ 0,\ 0)$. Patterns $\mathbf{X^{(1)}}$ and $\mathbf{X^{(2)}}$ generate two categories, $\mathbf{W_1^{(3)}} = \mathbf{W_1^{(2)}} = \mathbf{X^{(1)}}$ and $\mathbf{W_2^{(3)}} = \mathbf{X^{(2)}}$ respectively (since $\theta(\mathbf{W_1^{(2)}},\ \mathbf{X^{(2)}}) = \text{arc cos}\ (0.5) = 60^o$, then $ADM(\mathbf{W_1^{(2)}}, \mathbf{X^{(2)}}) = (90 - 60)/90 = 0.333 = \overline{MF}_1(\mathbf{W_1^{(2)}},\ \mathbf{X^{(2)}}) < \rho$). The winner template for pattern $\mathbf{X^{(3)}}$ is chosen according to Equation (35) as $j^* = \arg\max\{\frac{90 - arccos(1/\sqrt{3})}{90}, \frac{90 - arccos(1/\sqrt{3})}{90}\} = \arg\max\{0.392, 0.392\} = \{1, 2\}$, i.e., $\mathbf{W_{j^*}^{(3)}} =$

$\mathbf{W_1^{(3)}}$ or $\mathbf{W_{j^*}^{(3)}} = \mathbf{W_2^{(3)}}$. Let us consider that $\mathbf{W_{j^*}^{(3)}} = \mathbf{W_1^{(3)}}$. Since $\overline{MF}_1(\mathbf{W_1^{(3)}}, \mathbf{X^{(3)}}) = 0.392 < \rho$, then a new template is allocated (see Section 7.2), such that final templates are $\mathbf{W_1^{(4)}} = \mathbf{X^{(1)}}$, $\mathbf{W_2^{(4)}} = \mathbf{X^{(2)}}$ and $\mathbf{W_3^{(4)}} = \mathbf{X^{(3)}}$.

Let us consider, as in the Fuzzy ART example proposed in Appendix B, a different order of the input sequence where the input vectors described above are presented as follows: $\mathbf{X^{(1)}}$ = (0, 0, 1, 0, 0), $\mathbf{X^{(2)}}$ = $(1/\sqrt{3}, 1/\sqrt{3}, 1/\sqrt{3}, 0, 0)$, and $\mathbf{X^{(3)}}$ = (0, 1/2, 1/2, 1/2, 1/2). Patterns $\mathbf{X^{(1)}}$ and $\mathbf{X^{(2)}}$ generate two categories $\mathbf{W_1^{(3)}} = \mathbf{W_1^{(2)}} = \mathbf{X^{(1)}}$ and $\mathbf{W_2^{(3)}} = \mathbf{X^{(2)}}$ respectively (since $\theta(\mathbf{W_1^{(2)}}, \mathbf{X^{(2)}}) = \arccos(1/\sqrt{3}) = 54.73^o$, then $ADM(\mathbf{W_1^{(2)}}, \mathbf{X^{(2)}}) = (90 - 54.73)/90 = 0.392 = \overline{MF}_1(\mathbf{W_1^{(2)}}, \mathbf{X^{(2)}}) < \rho$). The winner template for pattern $\mathbf{X^{(3)}}$ is chosen according to Equation (35) as $j^* = \arg\max\{\frac{90-arccos 0.5}{90}, \frac{90-arccos(1/\sqrt{3})}{90}\} = \arg\max\{0.333, 0.392\} = 2$, i.e., $\mathbf{W_{j^*}^{(3)}} = \mathbf{W_2^{(3)}}$. Since $\overline{MF}_1(\mathbf{W_2^{(3)}}, \mathbf{X^{(3)}}) = 0.392 < \rho$, then a new template is allocated (see Section 7.2), such that final templates are $\mathbf{W_1^{(4)}} = \mathbf{X^{(1)}}$, $\mathbf{W_2^{(4)}} = \mathbf{X^{(2)}}$ and $\mathbf{W_3^{(4)}} = \mathbf{X^{(3)}}$.

Clustering results obtained when Fuzzy SART processes the two presentation sequences are consistent in terms of number of clusters. Note that when the third input instance is processed, the match value computed by match function (17) is the same for the two input sequences (equal to 0.392). We can conclude that in this example Fuzzy SART is insensitive to the change in the presentation order of the input sequence.

References

Alpaydın, E. (1998). Soft vector quantization and the EM algorithm. *Neural Networks*, in press.

Ancona, F., Ridella, S., Rovetta, S., & Zunino, R. (1997). On the importance of sorting in "Neural Gas" training of vector quantizers. *Proc. International Conference on Neural Networks '97*, Houston, TX, June 1997, vol. 3, pp. 1804-1808.

Barni, M., Cappellini, V., & Mecocci, A. (1996). Comments on "A possibilistic approach to clustering". *IEEE Trans. Fuzzy Systems*, **4**(3), 393-396.

Baraldi, A., & Parmiggiani, F. (1995a). A neural network for unsupervised categorization of multivalued input patterns: an application to satellite image clustering. *IEEE Trans. Geosci. Remote Sensing*, **33**(2), 305-316.

Baraldi, A., & Parmiggiani, F. (1995b). A self-organizing neural network merging Kohonen's and ART models. *Proc. International Conference on Neural Networks '95*, Perth, Australia, December 1995, vol. 5, pp. 2444-2449.

Baraldi, A., & Parmiggiani, F. (1995c). A refined Gamma MAP SAR speckle filter with improved geometrical adaptivity. *IEEE Trans. Geosci. Remote Sensing*, **33**(5), 1245-1257.

Baraldi, A., & Parmiggiani, F. (1996). Combined detection of intensity and chromatic contours in color images. *Optical Engineering*, **35**(5), 1413-1439.

Baraldi, A., & Parmiggiani, F. (1997). Novel neural network model combining radial basis function, competitive Hebbian learning rule, and fuzzy simplified adaptive resonance theory. *Proc. SPIE's Optical Science, Engineering and Instrumentation '97: Applications of Fuzzy Logic Technology IV*, San Diego, CA, July 1997, vol. 3165, 98-112.

Bezdek, J. C., & Pal, N. R. (1993). Generalized clustering networks and Kohonen's self-organizing scheme. *IEEE Trans. on Neural Networks*, **4**(4), 549-557.

Bezdek, J. C., & Pal, N. R. (1995). Two soft relatives of learning vector quantization. *Neural Networks*, **8**(5), 729-743.

Bishop, C. (1995). *Neural Networks for Pattern Recognition*. Oxford University Press.

Bishop, C., Svensen, M., & Williams, C. (1996). "GTM: a principled alternative to the self-organizing map," *Proc. Int. Conf. on Artificial Neural Networks, ICANN'96*, Springer-Verlag, 164-170.

Borghese, N. A., Ferrigno, G., Baroni, G., Savarè, R., Ferrari, S., & Pedotti, A. (1998) AUTOSCAN: A flexible and portable scanner of 3D surfaces. *IEEE Computer Graphics & Applications*, in press.

Boynton, R. M. (1990). Human color perception. In *Science of Vision*, K. N. Leibovic, Ed., 211- 253, Springer-Verlag, New York.

Buhmann, J. (1995). Learning and data clustering. In M. Arbib (Ed.), *Handbook of Brain Theory and Neural Networks*, Bradford Books / MIT Press.

Carpenter, G. A., & Grossberg, S. (1987a). 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. (1987b). ART2: Self-organization of stable category recognition codes for analog input patterns. *Applied Optics*, **26**(21), 4919-4930.

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., 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**(5), 698-713.

Davè R. N., & Krishnapuram R. (1997). Robust clustering method: a unified view. *IEEE Transactions on Fuzzy Systems*, **5**(2), 270-293.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. Royal Statist. Soc. Ser. B*, **39**, 1-38.

Dubes, R., & Jain, A. K. (1976). Clustering techniques: the user's dilemma. *Pattern Recognition*, **8**, 247-260.

Erwin, E., Obermayer, K., & Schulten, K. . (1992). Self-organizing maps: ordering, convergence properties and energy functions. *Biol. Cybernetics*, **67**, 47-55.

Fritzke, B. (1994). Growing cell structures - A self-organizing network for unsupervised and supervised learning. *Neural Networks*, **7**(9), 1441-1460.

Fritzke, B. (1995). A growing neural gas network learns topologies. In G. Tesauro, D. S. Touretzky & T. K. Leen (Eds.), *Advances in Neural Information Processing Systems 7* (pp. 625-632). Cambridge, MA: MIT Press.

Fritzke, B. (1997a). Some competitive learning methods. Draft document, *http://www.neuroinformatik.ruhr-uni-bochum.de/ini/VDM/ research/gsn/DemoGNG.*

Fritzke, B. (1997b). Incremental neuro-fuzzy systems. *Proc. SPIE's Optical Science, Engineering and Instrumentation '97: Applications of Fuzzy Logic Technology IV*, San Diego, CA, July 1997.

Healy, M. J., Caudell, T. P., & Smith, D. G. (1993). A neural architecture for pattern sequence verification through inferencing. *IEEE Transactions on Neural Networks*, **4**(1), 9-20.

Huang, J., Georgiopoulos, M., & Heileman, G. L. (1995). Fuzzy ART properties. *Neural Networks*, **8**(2), 203-213.

Hung C., & Lin, S. (1995). Adaptive Hamming Net: a fast-learning ART 1 model without searching. *Neural Networks*, **8**(4), 605-618.

Luttrell, S. P. (1990). Derivation of a class of training algorithms. *IEEE Trans. on Neural Networks*, **1**, 229-232.

Karayannis, N. B., Bezdek, J. C., Pal, N. R., Hathaway, R. J., & Pai, P. (1996). Repair to GLVQ: A new family of competitive learning schemes. *IEEE Trans. on Neural Networks*, **7**(5), 1062-1071.

Kim, Y. S., & Mitra, S. (1993). Integrated Adaptive Fuzzy Clustering (IAFC) algorithm. *Proceedings of the Second IEEE International Conference on Fuzzy Systems*, bf 2, 1264-1268.

Kohonen, T. (1990). The self-organizing map. *Proceedings of the IEEE*, **78**(9), 1464-1480.

Kohonen, T. (1995). Self-Organizing Maps, Berlin: Springer Verlag.

Krishnapuram, R., & Keller, J. M. (1993). A possibilistic approach to clustering. *IEEE Transactions on Fuzzy Systems*, **1**(2), 98-110.

Martinetz, T., Berkovich, G., & Schulten, K. (1993). Neural-Gas network for quantization and its application to time-series predictions. *IEEE Transactions on Neural Networks*, **4**(4), 558-569.

Martinetz, T., Berkovich, G., & Schulten, K. (1994). Topology representing networks. *Neural Networks*, **7**(3), 507-522.

Masters, T. (1994). Signal and image processing with neural networks - A C++ sourcebook. New York: Wiley.

Mulier, F. M., & Cherkassky V. S. (1995a). Statistical analysis of self-organization. *Neural Networks*, **8**(5), 717-727.

Mulier, F. M., & Cherkassky V. S. (1995b). Self-organization as an iterative kernel smoothing process. *Neural Computation*, **7**, 1165-1177.

Pao Y. (1989). Adaptive pattern recognition and neural networks. Reading, MA: Addison-Wesley.

Parisi, D. (1991). La scienza cognitiva tra intelligenza artificiale e vita artificiale. In E. Biondi, P. Morasso & V. Tagliasco (Eds.), *Neuroscienze e scienze dell'artificiale: dal neurone all'intelligenza* (pp. 321-341). Bologna, Italy: Patron.

Rose, K., Guerewitz, F., & Fox, G. (1990). A deterministic approach to clustering. *Pattern Recognition Letters*, **11**(11), 589-594.

Ritter, H., Martinetz, T., & Schulten, K. (1992). Neural computation and self-organizing maps. Reading, MA: Addison-Wesley.

Serra, R., & Zanarini, G. (1990). Complex systems and cognitive processes. Berlin: Springer-Verlag.

Shih, F. Y., Moh, J., & Chang, F. (1992). A new ART-based neural architecture for pattern classification and image enhancement without prior knowledge. *Pattern Recognition*, **25**(5), 533-542.

Simpson, P. K. (1993). Fuzzy min-max neural networks - Part 2: clustering. *IEEE Transactions on Fuzzy Systems*, **1**(1), 32-45.

Tsao, E. C., Bezdek, J. C., & Pal, N. R. (1994). Fuzzy Kohonen clustering network. *Pattern Recognition*, **27**(5), 757-764.

Williamson, J. R. (1996). Gaussian ARTMAP: a neural network for fast incremental learning of noisy multidimensional maps. *Neural Networks*, **9**(5), 881-897.

Williamson, J. R. (1997). A constructive, incremental-learning network for mixture modeling and classification. *Neural Computation*, **9**, 1517-1543.