A Comparison Between Classical Unsupervised Classifiers and ART3 Neural Networks

Michael J. Berchtold
Carl Zeiss Germany, Postfach 1380, D-73444 Oberkochen, FRG
e-mail: berchtold@cae007.v-ppb.zeiss.de

Abstract

This contribution reports a comparative study concerning connectionist and classical unsupervised classifiers. In order to understand the main concepts behind the ART3 model, a mathematical analysis is carried out that makes it possible to categorise the investigated connectionist model in the framework of well-established pattern recognition methods. We concentrate our attention on the clustering algorithm and the search mechanism used in ART3 neural networks. New results on two topics will be presented: (1) There exists a formal equivalence between ART3 and the well-known cosine classifier with respect to their clustering algorithm. (2) In the batch adaptation mode, the ART3 model implements a gradient descent on an energy function.

1 Introduction

One of the most complex neural models is ART3, proposed by Carpenter & Grossberg ([1]). We analyse this model mathematically to understand its basic concepts. Our attention is focused on the clustering algorithm and the search mechanism, to be able to categorise ART3 with respect to classical clustering algorithms, especially the so-called cosine classifier (CC, [2]).

2 Mathematical Analysis

We want to describe mathematically two major results of our research on two layer ART3-networks: (1) There exists a formal equivalence between the ART3 clustering algorithm and the well-known classical unsupervised CC. (2) In the batch adaptation mode, ART3 implements a gradient descent on an energy function, like many other neural and classical classifiers.

2.1 Equivalence of clustering algorithms

We show that the ART3 similarity measure $s^{ART3}$ is equivalent to the genuine cosine measure $\cos \alpha$. $\alpha$ is the angle between an active bottom-up prototype $\psi_J$ (or a top-down prototype $\lambda_J$) and the current input vector $i$. As usual, we will write vectors in lower-case bold-faced type and matrices in upper-case bold-faced type. Let $c$ and $d$ denote two positive scalar parameters. $\eta$ is the learning rate and $\mu = 1 - \eta$ the decay rate. Carpenter & Grossberg proposed a learning rate $\eta = 0.9$. $q^{[2]}$ denotes the second computation of the activity vector $q$, and $y^{[2]}$ represents the binary output vector of the $F_2$-layer. After bottom-up and top-down propagation of the input vector $i$ through the network, one finally gets a normalised activity vector:

$$q^{[2]} = \frac{p^{[1]} + c \lambda y^{[2]}}{\|p^{[1]} + c \lambda y^{[2]}\|} = \frac{p^{[1]} + c \lambda y^{[2]}}{\|p^{[1]} + c \lambda y^{[2]}\|}.$$ (1)
The similarity measure $s^{ART3}$ is given by the $L_2$-norm of the vector:

$$r = \frac{u_{[1]} + dq_{[2]}}{\|u_{[1]}\| + d \|q_{[2]}\|}.$$  \hspace{1cm} (2)

Applying the identity $u_{[1]} \lambda_J = \|\lambda_J\| \cos \alpha$ and simple arithmetics one gets:

$$s^{ART3}(\cos \alpha, \|\lambda_J\|) = \|r\| = \frac{\left(\sum_{i=1}^{K} (u_{[i]} + dq_{[i]}r)\right)^{\frac{1}{2}}}{\|u_{[1]}\| + d \|q_{[2]}\|}$$

$$= \frac{1}{1 + d} \left[ 1 + \frac{d}{2} \frac{1 + c \|\lambda_J\| \cos \alpha + c^2 \|\lambda_J\|^2 + d^2}{(1 + 2c \|\lambda_J\| \cos \alpha + c^2 \|\lambda_J\|^2 + d^2)^{\frac{1}{2}}} \right]^{\frac{1}{2}}. \hspace{1cm} (3)$$

The following Rule expresses the condition for which two similarity measures are equivalent. Equivalence here means that one can get the same clustering result with two similar clustering devices 1 and 2, which differ only in their similarity measures $s_1$ and $s_2$.

Rule: (Equivalence of similarity measures) Two similarity measures $s_1$ and $s_2$ are equivalent, if $s_2$ is a strictly monotonically increasing function of $s_1$.

Therefore, we have to prove that the similarity measure of equation (3) is a strictly monotonically increasing function of the genuine cosine measure $\cos \alpha$. For this reason, we calculate the first partial derivation of $s^{ART3}$ with respect to $\cos \alpha$. In equation (4), we substituted $\cos \alpha$ by $z$ and $\|\lambda_J\|$ by $y$, respectively.

$$\frac{\partial s^{ART3}(x,y)}{\partial x} = \frac{c d y}{(1 + 2c z y + c^2 y^2)^{\frac{1}{2}}} \left(1 - \frac{1 + c z}{1 + 2c z y + c^2 y^2}\right)^{\frac{1}{2}} > 0 \hspace{1cm} (4)$$

Equation (4) satisfies the above Rule, because all parameters are larger than zero and $x > 0$ for $0 < \alpha < \pi/2$. This guarantees the equivalence of the two similarity measures. We now examine the adaptation rules for both methods. The ART3 adaptation rule for the bottom-up weight-matrix $\Psi$ is given for the discrete case and for the $k$-th input presentation by:

$$\Delta \psi_{j}^{(k)} = \psi_{j}^{(k+1)} - \psi_{j}^{(k)} = \eta u_{(k+1)} - \mu \psi_{j}^{(k)}.$$  \hspace{1cm} (5)

This can be easily transformed to:

$$\psi_{j}^{(k)} = (1 - \mu)^k \psi_{j}^{(0)} + \eta \sum_{i=1}^{k} (1 - \mu)^{k-i} u_{(i)}^{(k-\infty)} \eta \sum_{i=1}^{k} (1 - \mu)^{k-i} u_{(i)}.$$  \hspace{1cm} (6)

For large $k$ equation (6) becomes $\eta \sum_{i=1}^{k} (1 - \mu)^{k-i} u_{(i)}^{(0)}$, because in this case $(1 - \mu)^k$ vanishes for $0 < \mu < 1$. Note, that the same result holds for $\lambda_J$, and that the vector norm of all $\psi_J$ and $\lambda_J$ asymptotically increases towards $g = 1/(1 - \eta)$. For the proposed parameter values, we get a $g$-value of 10. The fact that not all the prototypes are equally long during the adaptation phase causes some difficulties in the search for the most similar prototype to the current input vector ($\lambda_J$).

The CC adaptation rule for normalised input vectors $x$ and learning rate $\eta_J = 1/n_J$ is given by ($n_J$ is the number of assigned input vectors):

$$\psi_{j}^{(k+1)} = (1 - \eta_J^{(k+1)}) \psi_{j}^{(k)} + \eta_J^{(k+1)} x^{(k+1)}.$$  \hspace{1cm} (7)