Chapter VIII
An Approach to Artificial Concept Learning Based on Human Concept Learning by Using Artificial Neural Networks

Enrique Mérida-Casermeiro
University of Málaga, Spain

Domingo López-Rodríguez
University of Málaga, Spain

J.M. Ortiz-de-Lazcano-Lobato
University of Málaga, Spain

ABSTRACT

In this chapter, two important issues concerning associative memory by neural networks are studied: a new model of hebbian learning, as well as the effect of the network capacity when retrieving patterns and performing clustering tasks. Particularly, an explanation of the energy function when the capacity is exceeded: the limitation in pattern storage implies that similar patterns are going to be identified by the network, therefore forming different clusters. This ability can be translated as an unsupervised learning of pattern clusters, with one major advantage over most clustering algorithms: the number of data classes is automatically learned, as confirmed by the experiments. Two methods to reinforce learning are proposed to improve the quality of the clustering, by enhancing the learning of patterns relationships. As a related issue, a study on the net capacity, depending on the number of neurons and possible outputs, is presented, and some interesting conclusions are commented.

INTRODUCTION

Hebb (1949) introduced a physiological learning method based on the reinforcement of the interconnection strength between neurons. It was explained in the following terms:

When an axon of cell A is near enough to excite a cell B and repeatedly or persistently takes part in firing it, some growth process or metabolic change takes place in one or both cells such that A's efficiency, as one of the cells firing B, is increased.

This kind of learning method has been widely applied to recurrent networks in order to store and retrieve patterns in terms of their similarity. Models that used this learning rule were the bipolar model (BH) presented by J. J. Hopfield in 1982 (Hopfield, 1982) representing a powerful neural model for content addressable memory, or its analogical version, among others. These networks, although successful in solving many combinatorial optimization problems, present two main problems when used as content-addressable memory: their low capacity and the apparition of spurious patterns.

The capacity parameter α is usually defined as the quotient between the maximum number of patterns to load into the network and the number of used neurons that obtains an acceptable error probability (usually p_{error}=0.05 or 0.01). It has been shown that this constant is approximately α=0.15 for BH.

This value means that, in order to load K patterns, more than K/α neurons will be needed to achieve an error probability lower than or equal to p_{error}. Or equivalently, if the net is formed by N neurons, the maximum number of patterns that can be loaded in the net (with that error constraint) is $K<\alpha N$.

Since patterns are associated to states of the network with minimal energy, we wonder about what happens with these states if the network capacity is exceeded.

The main idea of this chapter holds that when patterns are very close each other, or if the net capacity is exceeded, then local minima corresponding to similar patterns tend to be combined, forming one unique local minimum. So, although considered as a limitation of the net as associative memory, this fact can explain the way in which the human brain form concepts: several patterns, all of them similar to a common typical representative, are associated and form a group in which particular features are not distinguishable.

Obviously, enough samples are needed to generalize and not to distinguish their particular features in both cases: artificial and natural (human) concept learning. If there are few samples from some class, they will still be retrieved by the net individually, that is, as an associative memory.

NEURAL BACKGROUND

Associative memory has received much attention for the last two decades. Though numerous models have been developed and investigated, the most influential is Hopfield's Associative Memory, based on his bipolar model (Hopfield, 1982). This kind of memory arises as a result of his studies on collective computation in neural networks.

Hopfield's model consists in a fully-interconnected series of bi-valued neurons (outputs are either -1 or +1). Neural connection strength is expressed in terms of weight matrix $W=(w_{i,j})$, where $w_{i,j}$ represents the synaptic connection between neurons i and j. This matrix is determined in the learning phase by applying Hebb's postulate of learning Hebb, and no further synaptic modification is considered later.

Two main problems arise in this model: the apparition of spurious patterns and its low capacity.

Spurious patterns are stable states, that is, local minima of the corresponding energy function of the network, not associated to any stored (input) pattern. The simplest, but not the least important, case of apparition of spurious patterns is the fact of storing, given a pattern, its opposite, i.e. both X and $-X$ are stable states for the net, but only one of them has been introduced as an input pattern.

The problem of spurious patterns is very fundamental for cognitive modelers as well as practical users of neural networks. Many solutions have been suggested in the literature. Some of them (Parisi, 1986) (Hertz et al., 1987) are based on introducing asymmetry in synaptic connections. However, it has been demonstrated that synaptic asymmetry does not provide by itself a satisfactory solution to the problem of spurious patterns, see (Treves et al., 1988) (Singh et al., 1995). Athitan et al. (1997) provided a solution based on neural self-interactions with a suitably chosen magnitude, if Hebb's learning rule is used, leading to the near (but not) total suppression of spurious patterns.

Crick (1983) suggested the idea of unlearning the spurious patterns as a biologically plausible solution to suppress them. With a physiological explanation, they suggest that spurious patterns are unlearned randomly by human brain during sleep, by means of a process that is the reverse of Hebb's learning rule. This may result in the suppression of many spurious patterns with large basins of attraction. Experiments have shown that their idea leads to an enlargening of the basins for correct patterns along with the elimination of a significant fraction of spurious patterns (van Hemmen et al., 1991). However, a great number of spurious patterns with small basins of attraction do survive. Also, in the process of undiscriminate reverse learning, there is a finite probability of unlearning correct patterns, what makes this strategy unacceptable.

On the other hand, the capacity parameter α is usually defined as the quotient between the maximum number of patterns to load into the network, and the number of used neurons that achieve an acceptable error probability in the retrieving phase, usually $p_e = 0.01$ or $p_e = 0.05$. It was empirically shown that this constant is approximately $\alpha = 0.15$ for BH (very close to its actual value, $\alpha = 0.1847$, see (Hertz et al., 1991)). The meaning of this capacity parameter is that, if the net is formed by N neurons, a maximum of $K \leq \alpha N$ patterns can be stored and retrieved with little error probability.

McElliece et al. (1987) showed that an upper bound for the asymptotic capacity of the network is $\frac{1}{2\log N}$, if most of the input (prototype) patterns are to remain as fixed points. This capacity decreases to $\frac{1}{4\log N}$ if every pattern must be a fixed point of the net.

By using Markov chains to study capacity and the recall error probability, Ho et al. (1992) showed results very similar to those obtained by McEliece, since for them it is $\alpha = 0.12$ for small values of N, and the asymptotical capacity is given by $\frac{1}{4\log N}$.

Kuh et al. (1989) manifested roughly similar estimations by making use of normal approximation theory and the theorems about exchangeables random variables.

Hopfield's Model

Hopfield's bipolar model consists in a network formed by N neurons, whose outputs (states) belong to the set {-1,1}. Thus, the state of the net at time t is completely defined by a N-dimensional state vector $\mathbf{V}(t) = (V_1(t), V_2(t), \ldots, V_N(t)) \in \{-1,1\}^N$.

Associated to every state vector there is an energy function, expressed in the following terms:

$$E(\mathbf{V}) = -\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} w_{i,j} V_i V_j + \sum_{i=1}^{N} \theta_i V_i \qquad (1.1)$$

where $w_{i,j}$ is the connection weight between neurons i and j, and θ_i is the threshold corresponding to i-th neuron (since thresholds are not used in the case of associative memory, from now on all of them will be considered 0). This energy function determines the behavior of the net.

Hopfield's Associative Memory

Let us consider $\{X^{(k)} : k = 1, \ldots, K\}$, a set of bipolar patterns to be loaded into the network. In order to store these patterns, weight matrix W must be determined. This is achieved by applying Hebb's classical rule for learning. So, the change of the weights, when pattern $X = (X_i)$ is introduced into the network, is given by $\Delta w_{i,j} = X_i X_j$. Thus, the final expression for the weights is:

$$w_{i,j} = \sum_{k=1}^{K} X_i^{(k)} X_j^{(k)} \tag{1.2}$$

In this case, the energy function that is minimized by the network can be expressed in the following terms:

$$E(\vec{V}) = -\frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \sum_{k=1}^{K} X_i^{(k)} X_j^{(k)} V_i V_j \tag{1.3}$$

In order to retrieve a pattern, once the learning phase has finished, the net is initialized with the known part of the pattern (called probe). Then, the dynamics makes the network converge to a stable state (due to the decrease of the energy function), corresponding to a local minimum. Usually this stable state is close to the initial probe.

If all input patterns form an orthogonal set, then they are correctly retrieved, otherwise some errors may happen in the recall procedure. But, as the dimension of the pattern space is N, there is no orthogonal set with cardinality greater than N. This implies that if the number of patterns exceed the number of neurons, errors may occur. Thus, capacity can not be greater than 1 in this model.

MREM Model with Semi-Parallel Dynamics

The Multivalued REcurrent Model (MREM) consists of a recurrent neural network formed by N neurons, where the state of each neuron i is defined by its output V_i (i=1,…,N), taking values in any finite set $\mathcal{M} = \{m_1, m_2, \ldots, m_L\}$. This set does not need to be numerical.

The state of the network, at time t, is given by a N-dimensional vector, $\mathbf{V}(t) = (V_1(t), V_2(t), \ldots, V_N(t)) \in \mathcal{M}^N$. Associated to every state vector, an energy function, characterizing the behaviour of the net, is defined:

$$E(V) = -\frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} w_{i,j} f(V_i, V_j) + \sum_{i=1}^{N} \theta_i(V_i) \tag{2.1}$$

where $w_{i,j}$ is the weight of the connection from the j-th neuron to the i-th neuron, and $f : \mathcal{M} \times \mathcal{M} \to \mathbb{R}$ can be considered as a measure of similarity between the outputs of two neurons, usually verifying the following similarity conditions:

1. For all $x \in \mathcal{M}$, $f(x,x) = c \in \mathbb{R}$.
2. f is a symmetric function: for every $x,y \in \mathcal{M}$, $f(x,y) = f(y,x)$.
3. If $x \neq y$, then $f(x,y) \leq c$.

and $\theta_i: \mathcal{M} \to \mathbb{R}$ are the threshold functions. Since thresholds will not be used for content addressable memory, henceforth we will consider θ_i be the zero function for all $i \in I$.

The introduction of this similarity function provides, to the network, of a wide range of possibilities to represent different problems (Mérida et al., 2001) (Mérida et al., 2002). So, it leads to a better representation than other multi-valued models, like SOAR and MAREN (Erdem et al., 1996) (Ozturk et al., 1997), since in those models most of the information enclosed in the multi-valued representation is lost by the use of the sign function that only produces values in {-1,0,1}.

It is clear that MREM, using bipolar ($\mathcal{M}$ = {1,1}) or bi-valued ($\mathcal{M}$ = {0,1}) neurons, along with the similarity function given by $f(x, y) = xy$ and constant bias functions, reduces to Hopfield's model. So, this model can be considered a powerful generalization of Hopfield's model.

In every instant, the net evolves to reach a state of lower energy than the current one.

It has been proved that the MREM model with its associated dynamics always converges to a minimal state. This result is particularly important when dealing with combinatorial optimization problems, where the application of MREM has been very fruitful (López-Rodríguez et al, 2006) (Mérida-Casermeiro et al., 2001a) (Mérida-Casermeiro et al., 2001b) (Mérida-Casermeiro et al., 2002a) (Mérida-Casermeiro et al., 2002b) (Mérida-Casermeiro et al., 2003) (Mérida-Casermeiro et al., 2004) (Mérida-Casermeiro et al., 2005).

MREM as Auto-Associative Memory

Now, let $\{X^{(k)} : k = 1,\dots,K\}$ be a set of patterns to be loaded into the neural network. Then, in order to store a pattern, $X=(X_1,X_2,\dots,X_N)$, components of the W matrix must be modified in order to make $\mathbf{X}$ the state of the network with minimal energy.

Since energy function is defined, we modify the components of matrix W in order to reduce the energy of state $V=\mathbf{X}$ by the rule:

$$\Delta w_{i,j} = -2\frac{\partial E}{\partial w_{i,j}} = f(X_i, X_j).$$

The coefficient 2 does not produce any effect on the storage of the patterns, and it is here chosen for simplicity. Considering that, at first, $W=0$, that is, all the states of the network have the same energy and adding over all the patterns, the next expression is obtained:

$$w_{i,j} = \sum_{k=1}^{K} f(X_i^{(k)}, X_j^{(k)}) \tag{3.1}$$

Equation is a generalization of *Hebb's postulate of learning*, because the weight $w_{i,j}$ between neurons is increased in correspondence with their similarity.

It must be pointed out that, when bipolar neurons and the product function are used, $f(x,y)=xy$, the well-known learning rule of patterns in the Hopfield's network is obtained. In fact, this is equivalent to choose $f(x,y)=1$ if $x=y$, and otherwise $f(x,y)=0$.

In what follows, we will consider the similarity function given by: $f(x, y)=1$ if $x=y$ and -1, otherwise.

In order to recover a loaded pattern, the network is initialized with the known part of that pattern. The network dynamics will converge to a stable state (due to the decreasing of the energy function), that is, a minimum of the energy function, and it will be the answer of the network. Usually this stable state is next to the initial one.

How to Avoid Spurious States

When a pattern X is loaded into the network, by modifying weight matrix W, not only the energy corresponding to state $\mathbf{V} = X$ is decreased. This fact can be explained in terms of the so-called associated vectors.

Given a state **V**, its associated matrix is defined as $G_{\mathbf{V}} = (g_{i,j})$ such that $g_{i,j} = f(V_i, V_j)$.

Its associated vector is $A_{\mathbf{V}} = (a_k)$, with $a_{(i-1)N+j} = g_{i,j}$, that is, it is built by expanding the associated matrix as a vector of N^2 components.

With this notation, the energy function can be expressed as:

$$E(\mathbf{V}) = -\frac{1}{2}\sum_{k=1}^{K} < A_{X^{(k)}}, A_{\mathbf{V}} > \tag{3.2}$$

where $<\cdot,\cdot>$ denotes the usual inner product.

Lemma 1. The increment of energy of a state **V** when pattern X is loaded into the network, by using Equation , is given by:

$$\Delta E(\mathbf{V}) = -\frac{1}{2} < A_X, A_{\mathbf{V}} > \tag{3.3}$$

Lemma 2. Given a state vector **v**, we have $A_{\mathbf{V}} = A_{-\mathbf{V}}$. So $E(\mathbf{V}) = E(-\mathbf{V})$.

These two results explain why spurious patterns are loaded into the network. Let us see it with an example:

Suppose that we want to get pattern $X = (-1, 1, -1, -1, 1)$ loaded into a BH network. Then, its associated matrix will be:

$$G_X = \begin{pmatrix} 1 & -1 & 1 & 1 & -1 \\ -1 & 1 & -1 & -1 & 1 \\ 1 & -1 & 1 & 1 & -1 \\ 1 & -1 & 1 & 1 & -1 \\ -1 & 1 & -1 & -1 & 1 \end{pmatrix}$$

and therefore the associated vector will be:

$${}_X = (1,-1,1,1,-1,-1,1,-1,-1,1,1,$$
$$-1,1,1,-1,1,-1,1,1,-1,-1,1,-1,-1,1)$$

But, as can be easily verified, the vector $-X = (1,-1,1,1,-1)$ has the same associated matrix and vector than the original pattern X, that is, $A_{-X} = A_X$. By using the previous lemmas, we obtain that the increment of energy of both X and $-X$ is the same, so these two vectors are those whose energy decreases most:

$$\Delta E(X) = -\frac{1}{2} < A_X, A_X >=$$
$$-\frac{1}{2} < A_X, A_{-X} >= \Delta E(-X)$$

This fact also implies that the corresponding ΔW is the same for both vectors.

These results explain the well-known problem of loading the opposite pattern of Hopfield's associative memory.

When using MREM, spurious patterns are generated by the network in the same way. For example, when we load the pattern $X = (3,3,2,1,4,2)$, also the pattern $X_1 = (4,4,3,2,1,3)$ is loaded, but also $X_2 = (1,1,4,3,2,4)$, since all of them have the same associated vector, and produce the same decrease in the energy function. So, in MREM, the number of spurious patterns appearing after the load of a vector into the net is greater than the corresponding in BH.

Since all associated vectors are vectors of N^2 components taking value in $\{-1, 1\}$, their norms are equal, $\| A_{\mathbf{v}} \|_k = N$ for all $\mathbf{v}$. This result implies that what is actually stored in the network is the orientation of the vectors associated to loaded patterns.

From the above expression for the increment of energy, and using that components of associated vectors are either -1 or 1, the following expression for the decrease of energy when a pattern is loaded is obtained:

$$-\Delta E(\mathbf{V}) = \frac{1}{2}(N - 2d_H(\mathbf{V}, X))^2 \qquad (3.4)$$

where $d_H(\mathbf{V}, X)$ is the Hamming distance between vectors $\mathbf{v}$ and X.

After this explanation, we propose a solution for this problem:

The augmented pattern $\hat{X}$, associated to X, is defined by appending to X the possible values of its components, that is, if $\mathcal{M} = \{m_1, \ldots, m_L\}$, then $\hat{X} = (X_1, \ldots, X_N, m_1, \ldots, m_L)$.

Particularly:

- In case of bipolar outputs, $\mathcal{M} = \{-1, 1\}$, and consequently it is $\hat{X} = (X_1, \ldots, X_N, -1, 1)$.
- If $\mathcal{M} = \{1, \ldots, L\}$, then $\hat{X} = (X_1, \ldots, X_N, 1, 2, \ldots, L)$.

By making use of augmented patterns, the problem of spurious patterns is solved, as stated in the next result:

Lemma 3. The function Ψ that associates an augmented pattern to its corresponding associated vector is injective.

It can be shown that if augmented patterns are used, the state $\mathbf{V}$ whose energy decreases most when pattern X is introduced in the net, is $\mathbf{V} = X$. This result is deduced from Equation , applied to the case of $N + L$ components:

$$-\Delta E(\hat{\mathbf{V}}) = \frac{1}{2}(N + L - 2d_H(\hat{\mathbf{V}}, \hat{X}))^2 \cdot$$

So, if $\mathbf{V} \neq X$ then $\hat{\mathbf{V}} \neq \hat{X}$ and $1 \leq d_H(\mathbf{V}, X) = d_H(\hat{\mathbf{V}}, \hat{X}) \leq N$, and the next inequality holds:

$$L - N = N + L - 2N \leq N + L$$
$$- 2d_H(\hat{\mathbf{V}}, \hat{X}) \leq N + L - 2$$

Therefore

$$-2\Delta E(\hat{\mathbf{V}}) = (N + L - 2d_H(\hat{\mathbf{V}}, \hat{X}))^2 \leq \max\{(N - L)^2, (N + L - 2)^2\} =$$
$$= (N + L - 2)^2 < (N + L)^2 = -2\Delta E(\hat{X})$$

which demonstrates our statement.

Then, in order to load a pattern X, it will suffice to load its augmented version, which will be the unique state maximizing the decrease of energy.

In this example, we can see how this method works. Consider, at first, that W=0, that is $E(V)$=0 for all state vector V. Then, in the original model MREM, the pattern X=(3,3,2,1,4,2) is loaded, and matrix W is updated. Note that, in this case, N=6 and L=4. Then, if Y=(4,4,3,2,1,3), we can compute:

$$E(X) = -\frac{1}{2}(N - 2d_H(X, X))^2 = -\frac{1}{2}6^2 = -18$$
$$E(Y) = -\frac{1}{2}(N - 2d_H(Y, X))^2 = -\frac{1}{2}(6 - 2 \cdot 6)^2 = -\frac{1}{2}(-6)^2 = -18$$

Therefore, Y has been also loaded into the network, since X is one global minimum of the energy function, and $E(Y)$=$E(X)$. With the original model, Y is a spurious pattern. Let us apply the technique of augmented patterns to solve this problem. In this case, the augmented patterns are: $\widehat{X} = (3,3,2,1,4,2,1,2,3,4)$ and $\widehat{Y} = (4,4,3,2,1,3,1,2,3,4)$. We can now compute the energy value associated to X and Y:

$$E(X) = -\frac{1}{2}(N + L - 2d_H(\widehat{X}, \widehat{X}))^2 = -\frac{1}{2}(6 + 4 - 0)^2 = -50$$
$$E(Y) = -\frac{1}{2}(N + L - 2d_H(\widehat{Y}, \widehat{X}))^2 = -\frac{1}{2}(6 + 4 - 2 \cdot 6)^2 = -2$$

This result ($E(X)$<$E(Y)$) implies that Y is not stored in the net, since it is not a minimum of the energy function. So, this technique is able to avoid the apparition of spurious patterns.

It must be noted that it will only be necessary to consider N neurons, their weights, and the weights corresponding to the last L neurons, that remain fixed, and do not need to be implemented.

SOME REMARKS ON THE CAPACITY OF THE NET

In Mérida et al. (2002), authors find an expression for the capacity parameter α for MREM model in terms of the number of neurons N and the number of possible states for each neuron, L, for the case in which N is big enough to apply the Central Limit Theorem ($N \geq 30$):

$$\alpha(N, L) \approx \frac{1}{N} + \frac{\frac{A^2}{z_\alpha^2} - B}{NC} \qquad (4.1)$$

where

$$A = N + 3 + (N - 1)\frac{4 - L}{L}, \quad B = 8(N - 1)\frac{L - 2}{L^2}, \quad C = \frac{8N}{L}$$

and z_α is obtained by imposing the condition that the maximum allowed error probability in retrieving patterns is p_{error}. For p_{error}=0.01, we get z_α≈2.326.

Some facts can be extracted from the above expression.

For a fixed number of neurons, capacity is not bounded above:

Suppose N fixed. Equation (4.2) can be rewritten in the following form:

$$\alpha(N,L) \approx \frac{1}{N^2 z_\alpha^2}\left(2L + 4(N-1) + z_\alpha^2 + 2(N-1)\frac{N-1+z_\alpha^2}{L}\right) \tag{4.2}$$

If we make L tend to ∞, we get $\lim_{L\to\infty} \alpha(N,L) = \infty$, since the coefficient of L in this expression is positive.

What actually happens is that $\alpha(N,\cdot)$, as a function of L, has a minimum at the point $L_0(N) = \sqrt{(N-1)(N-1+z_\alpha^2)} \approx N$ for z_α=2.326. It is a decreasing function for $L<L_0(N)$ and increasing for $L\geq L_0(N)$.

One consequence of this result is that, for appropriate choice of N and L, the capacity of the net can be α (N,L)>1.

This fact can be interpreted as a adequate representation of the multi-valued information, because, to represent the same patterns as MREM with N and L fixed, BH needs NL binary neurons and therefore the maximum number of stored patterns may be greater than N. So it is not a strange thing that the capacity can reach values greater than 1, if the patterns are multi-valued, MREM needs much less neurons to represent the pattern than BH.

For a fixed number of possible outputs, capacity is bounded below by a positive constant:

Suppose L is fixed. Equation can be rewritten as follows:

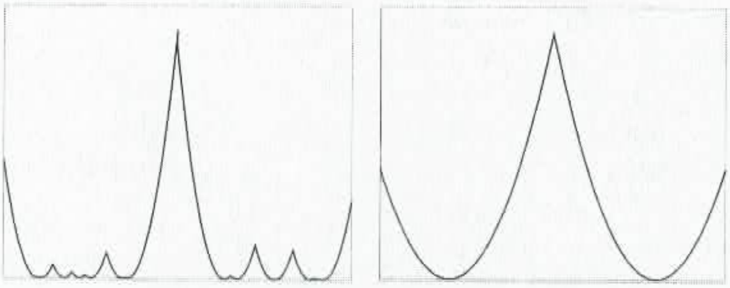
$$\alpha(N,L) \approx \frac{1}{L z_\alpha^2}\left(2 + \frac{4(L-1)+2z_\alpha^2}{N} + \frac{2(L-1)^2+(L-2)z_\alpha^2}{N^2}\right) \tag{4.3}$$

It can be easily seen that this expression represents a function whose value decreases as N grows. So, a net with more neurons than other, and the same possible states, will present less capacity than the second one.

Thus, a minimum positive capacity can be computed for each possible value of L, verifying $\alpha_{\min}(L) = \lim_{N\to+\infty} \alpha(N,L) = \frac{2}{L z_\alpha^2} > 0$.

$\alpha_{\min}(L)$ coincides with the asymptotic capacity for the net with L possible neuron outputs. For example, if L=2 (as in BH), an asymptotic capacity of $\alpha_{\min}(2)$=0.1847 is obtained, exactly the capacity for BH provided in other works (Hertz et al., 1991).

Figure 1. Many individual patterns are loaded into the net, forming a group of local minima of the energy function (left). When the number of patterns to be loaded is greater than the capacity of the network, the corresponding local minima are merged (right). The formation of these new local optima can be interpreted as the apparition and learning of the associated concepts.



WHEN CAPACITY IS EXCEEDED

This work tries to explain what may happen psychologically in the human brain. When a reduced number of patterns has to be memorized, the brain is able to remember all of them when necessary. Similarly, when the capacity of the net is not exceeded, the net is able to retrieve exactly the same patterns that were loaded into it. But when the brain receives a great amount of data to be recognized or classified, it distinguishes between some groups of data (in an unsupervised way) and thus forming concepts. This kind of behaviour is also simulated by neural networks, as we will show next.

Then, learning rules as Hebb's (or the more general given by equation (3.1), where connection between neurons is reinforced by the similarity of their expected outputs), may produce classifiers that discover some knowledge from the input patterns, like the actual number of groups in which the data are divided. Then, an unsupervised clustering of the input pattern space is automatically performed. This unsupervised clustering generates the concept space, formed by the equivalence classes found for the input patterns.

If a pattern, say X, is to be loaded in the net, by applying equation (3.1), a local minimum of the energy function E is created at $V=X$. If another pattern X' is apart from X, its load will create another local minimum. But, if X and X' are close each other, these two local minima created by the learning rule will be merged, forming one local minima instead.

Then, if a group of patterns is loaded into the net (overflowing its capacity), and all of them are close each other, only one local minimum will be formed, and at the moment of retrieving these data, the unique pattern to be retrieved will be associated to the state of minimum energy. So, patterns can be classified by the stable state of the net which they converge to. This stable state can be considered as a representative of the concept associated to that group of patterns.

This way, the learning of new concepts corresponds to the saturation of individuals, that is, the only way to learn a general concept is by presenting to the learner (the network, or the human brain, both cases have the same behaviour) a great number of individuals satisfying that concept.

For example, if a boy has to learn the concept 'chair', he will be presented a series of chairs, of many styles and sizes, until he is able to recognize a chair, and distinguish it from a table or a bed.

RELATIONSHIPS LEARNING

Equation for the learning rule, generalization of Hebb's one, shows that the only thing taking part in updating weight matrix is the pattern to be loaded into the net at that time. So, it represents a very 'local' information, and does not take account of the possible relationships that pattern could have with the already stored ones. So, it is convenient to introduce an additional mechanism in the learning phase, such that the information concerning to relationships between patterns is incorporated in the update of the weight matrix. In what follows, we will consider that the similarity function is $f(x, y) = 2\delta_{x,y} - 1$, that is, its value is 1 if $x = y$ and -1 otherwise.

Relationships learning method: Suppose that we have the (augmented) pattern X_1 stored in the net. So, we have the weight matrix $W = (w_{i,j})$. If pattern X_2 is to be loaded into the network, by applying equation (3.1), components of matrix ΔW are obtained.

If $w_{i,j}$ and $\Delta W_{i,j}$ have positive signum (both values equal 1), it means that $X_{1i} = X_{1j}$ and $X_{2i} = X_{2j}$, indicating the relationship between components i and j of X_1 and X_2. If both are negative valued, something similar happens, but with inequalities instead of equalities.

Figure 2. Scheme of the relationship between RL and overflowing network capacity

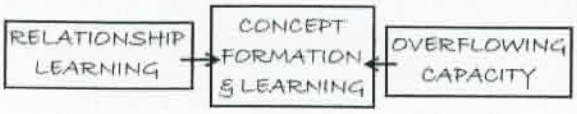
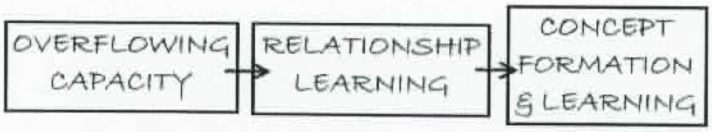


Figure 3. Scheme for the formation and learning of concepts



So, the fact of $w_{i,j}$ and $\Delta W_{i,j}$ having the same signum is a clue of a relationship that is repeated between components i and j of patterns X_1 and X_2. In order to reinforce the learning of this relationship, we propose a novel technique, presenting also another kind of desirable behavior: The model proposed before, given by equation (3.1), is totally independent of the order in which patterns are presented to the net. This fact does not actually happen in the human brain, since every new information is analyzed and compared to data and concepts previously learned and stored.

So, to simulate this kind of learning, a method named RL is presented:

Let us multiply by a constant, $\beta > 1$, the components of matrices W and ΔW where the equality of signum is verified, i.e., the components verifying $w_{i,j} \cdot \Delta W_{i,j} > 0$. Hence the weight matrix learned by the network is, after loading pattern X_2:

$$w'_{i,j} = \begin{cases} w_{i,j} + \Delta W_{i,j} & \text{if } w_{i,j} \cdot \Delta W_{i,j} < 0 \\ \beta[w_{i,j} + \Delta W_{i,j}] & \text{if } w_{i,j} \cdot \Delta W_{i,j} > 0 \end{cases} \tag{5.1}$$

Similarly, if there are some patterns $\{X_1, X_2, \ldots, X_R\}$ already stored in the network, in terms of matrix W, and pattern X_{R+1} is to be loaded, matrix ΔW (corresponding to X_{R+1}) is computed and then the new learning rule given by equation (5.1) is applied.

It must be noted that this method satisfy Hebb's postulate of learning quoted before.

This learning reinforcement technique, RL, has the advantage that it is also possible to learn patterns one by one or by blocks, by analyzing at a time a whole set of patterns, and comparing the resulting ΔW to the already stored in the net. Then, for instance, if $\{X_1, \ldots, X_R\}$ has already been loaded into the net in terms of matrix W, we can load a whole set $\{Y_1, \ldots, Y_M\}$ by computing $\Delta W = (\sum_{k=1}^{M} f(y_{ki}, y_{kj}))_{i,j}$ and then applying equation (5.1).

As shown in Figure 2, RL and the capacity of the network are related via their ability to make the learner (the neural network, or the human brain, whatever the case) form and learn concepts.

With this idea in mind, one can think that what the capacity overflow actually induces is the learning of the existing relationships between patterns: as explained in Figure 1, several local minima, representing stored patterns, are merged when network capacity is overflowed. This implies a better recognition of the relationship between similar patterns, which usually represent the same

Table 1. Average clustering results on 10 runs of these algorithms, where n_a indicates the number of obtained clusters, p_{cc} is the correct cassification percentage (that is, the percentage of simulations in which $n_a = n$) and Err. is the average error percentage.

K	n=3						n=4					
	RL			IRL			RL			IRL		
	n_a	p_{cc}	Err.	n_a	p_{cc}	Err.	n_a	p_{cc}	Err.	n_a	p_{cc}	Err.
75	3.1	90	0.53	3.1	90	0.53	4.7	60	3.46	4.1	90	0.26
150	3.3	70	0.33	3.1	90	0.13	4.8	70	1.06	4.4	80	0.73
300	3.7	70	0.60	3.3	80	0.43	5.5	30	3.06	4.6	60	4.80
450	3.5	70	0.37	3.1	90	0.17	4.9	60	0.55	4.4	80	0.31
600	3.2	80	0.25	3.2	80	0.25	5.4	50	0.61	4.6	50	0.31
750	3.3	80	0.06	3.0	100	0.00	5.4	30	3.49	4.6	60	3.12
900	3.2	90	3.23	3.0	100	0.00	5.5	20	0.56	4.8	40	0.42
Av.	3.3	78	0.76	3.1	90	0.21	5.2	46	1.82	4.5	66	1.42

concept. This concept is thus formed and learned. So, the process of formation and learning concepts can be seen in Figure 3.

Iterative relationships learning method: RL can be improved in many ways. In this sense, an iterative approach, IRL, to enhance the solution given by RL, is presented.

Suppose that, by using equation (5.1) of RL, matrix W_X related to pattern set $X = \{X^{(k)} : k \in \mathcal{K}\}$ has been learned and denote by $Y^{(k)}$ the stable state reached by the network (with weight matrix W_X) when beginning from the initial state given by $\mathbf{V} = X^{(k)}$.

Then, the cardinal of $\{Y^{(k)} : k \in \mathcal{K}\}$ is (no multiplicities) the number of classes that RL finds, n_a.

$Y := \{Y^{(k)} : k \in \mathcal{K}\}$ can be considered (with all multiplicities included) as a new pattern set, formed by a noiseless version of patterns in X. So, if applying a second time RL to Y, by using equation (5.1) to build a new matrix W_Y, better results are expected than in the first iteration, since the algorithm is working with a more refined pattern set.

This whole process can be repeated iteratively until a given stop criterion is satisfied. For example, when two consecutive classifications assign each pattern to the same cluster.

SIMULATIONS

a. In order to show the ability of RL and IRL to perform a clustering task, as mentioned above, several simulations have been made whose purpose is the clustering of discrete data.

Several datasets have been created, each of them formed by K 50-dimensional patterns randomly generated around n centroids, whose components were integers in the interval [1,10]. That is, the n centroids were first generated and input patterns were formed from them by introducing some random noise modifying one component of the centroid with probability 0.018. So, the Hamming distance between input patterns and the corresponding centroids is a binomial distribution B(50,0.018). Patterns are equally distributed among the n clusters. It must be noted that patterns may have Hamming distance even 5 or 6 from their respective centroid, and new clusters can be formed by this kind of patterns.

So, a network with $N = 50$ neurons taking value in the set $\mathcal{M} = \{1,\ldots,10\}$ has been considered. The parameter of learning reinforcement has been chosen $\beta = 1.5$. It has been observed that similar results are obtained for a wide range of values of β.

The results obtained in our experiments are shown in Table 1. It can be observed not only the

Figure 4. The cameraman image

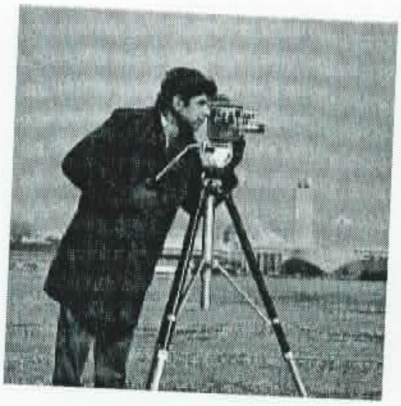
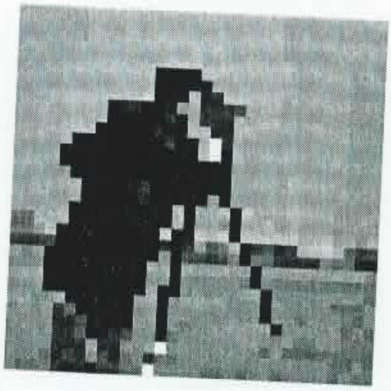


Figure 5. The resulting image after performing the unsupervised clustering



low classification error (from 0% to 4.80% on average), but in addition these new techniques get the exact, or very approximate, number of groups in which the pattern set is actually divided in almost every simulation. In fact, whenever the number n_n of discovered clusters equals n, an error percentage of 0% is obtained, retrieving in those cases the initial centroids. It can also be verified that IRL clearly outperforms RL in most cases, getting a more accurate classification and improving the estimation of the number of clusters, as seen in the last row of the table.

b. A more practical example is given below. In this case, it has been studied how, by overflowing the network capacity, the net is able to extract relationships between patterns and form clusters or groups representing concepts.

Here, the problem is to cluster an image, attending to its contents. It's a rather simple example, but it has many applications in image segmentation and understanding. For more details on image processing techniques, see Egmont-Petersen et al. (2002) and references therein.

In our case, the well-known cameraman image (Figure 4) has been used as benchmark.

This image (256x256 pixels) was divided in windows of size 8x8 pixels. Each window represents a pattern, rearranging its values columnwise from top to bottom. So, there were 1024 patterns of dimension 64.

These patterns were loaded into a MREM network with 64 neurons. Then, the net was initialized with every pattern, and the corresponding stable state (a new window) was used to substitute the original window. The resulting clustering can be viewed in Figure 5. It should be noted that there were 14 different clusters, detected automatically.

The stable states achieved by the net correspond to concepts. Since pattern components were gray values of pixels, the network has learnt concepts related to these graylevels.

The interesting point is that all this work has been done in an unsupervised way. Other clustering methods need to adjust the number of clusters before the training, and are based on Euclidean distance, which implies that patterns are embedded in the Euclidean space. However, discrete patterns like the ones described in this work can be non-numerical (qualitative) and may have a different topology from the usual one in the Euclidean space.

FUTURE TRENDS

Recently, concept learning has received much attention from the Artificial Intelligence community. There are sectors of this community interested in using techniques such as fuzzy and classical logics to study this process.

Regarding associative memories and neural networks, the current trend is to adapt or develop learning rules in order to increase the capacity of the network.

CONCLUSION

In this work, we have explained that the limitation in capacity of storing patterns in a recurrent network has not to be considered as determinant, but it can be used for the unsupervised classification of discrete patterns, acting as a concept learning system.

The neural model MREM, based on multivalued neurons, has been developed, as a generalization of the discrete Hopfield model and as an auto-associative memory, improving some of the undesirable aspects of the original Hopfield model: some methods and results for the net not to store spurious patterns in the learning phase have been shown, reducing so the number of local minima of the energy function not associated to an input pattern.

By applying a slight modification to Hebb's learning rule, a mechanism to reinforce the learning of the relationships between different patterns has been introduced to the first part of the process, incorporating knowledge corresponding to several patterns simultaneously. This mechanism can be repeated iteratively to enhance the relationship learning procedure.

One of the main facts expressed in this work is that network capacity, which can be viewed as a restriction for the network as associative memory, becomes a powerful ally when forming and learning concepts (groups of similar patterns), since it implies the learning of relationships between patterns mentioned above.

In addition, simulations confirm the idea expressed in this work, since, by overflowing the capacity of the net, we can get optimal results of classification in many cases. This technique is therefore very useful when tackling classification and learning problems, with the advantage of being unsupervised.

FUTURE RESEARCH DIRECTIONS

This chapter presents a new open research line, from the fact that some new modifications of the learning rule, reinforcing other aspects of the relationships among patterns, may be developed.

An interesting point to be studied is the possible generalization of Hebb's learning rule such that the capacity of a neural network increases, since it is crucial for many applications.

REFERENCES

Athithan, G., & Dasgupta, C. (1997). *On the Problem of Spurious Patterns in Neural Associative Models*, IEEE Transactions on Neural Networks, vol. 8, no. 6, 1483-1491.

Crick, F., & Mitchinson, G. (1983). *The Function of Dream Sleep*, Nature, vol. 304, 111-114.

Egmont-Petersen, M., de Ridder, D., & Handels, H. (2002). *Image processing with neural networks—a review*, Pattern Recognition, vol. 35, 2279-2301.

Erdem, M. H., & Ozturk, Y. (1996). *A New family of Multivalued Networks*, Neural Networks 9,6, 979-89.

Hebb, D. O. (1949). *The Organization of Behavior*. New York, Wiley.

Hemmen, J. L. van, & Kuhn, R. (1991). *Collective Phenomena in Neural Networks*, E. Domany, J. L. van Hemmen and K. Shulten, eds. Berlin: Springer-Verlag, 1-105.

Hertz, J., Krogh, A., & Palmer, R. G. (1991). *Introduction to the Theory of Neural Computation*. Lecture Notes, Volume 1, Addison Wesley.

Hertz, J. A., Grinstein, G., & Solla, S. A. (1987). *Heidelberg Colloquium on Glassy Dynamics*, J. L. van Hemmen and I. Morgenstern, eds. Berlin: Springer-Verlag, 538-546.

Ho, C. Y., Sasase, I. and Mori, S. (1992). On *the Capacity of the Hopfield Associative Memory*. In Proceedings of IJCNN 1992, II196-II201.

Hopfield, J. J. (1982). *Neural Networks and Physical Systems with Emergent Collective Computational Abilities*. In Proceedings of the National Academy of Science, USA, 79, 2554-2558.

Hopfield, J. J. (1984). *Neurons with graded response have collective computational properties like those of two-state neurons*, Proceedings of the National Academy of Sciences USA, 81, 3088-3092.

Kuh, A., & Dickinson, B. W. (1989). *Information Capacity of Associative Memory*. IEEE Transactions on Information Theory, vol. IT-35, 59-68.

López-Rodríguez, D., Mérida-Casermeiro, E., Ortiz-de-Lazcano-Lobato, J. M., & López-Rubio, E. (2006). *Image Compression by Vector Quantization with Recurrent Discrete Networks*. Lecture Notes in Computer Science, 4132, 595-605.

McEliece, R. J., Posner, E. C., Rodemich, E. R., & Venkatesh, S. S. (1990). *The Capacity of the Hopfield Associative Memory*. IEEE Transactions on Information Theory, vol. IT-33, no. 4, 461-482.

Mérida-Casermeiro, E., Muñoz-Pérez, J., & Benítez-Rochel, R. (2001). *A recurrent multivalued neural network for the N-queens problem*, Lecture Notes in Computer Science 2084, 522-529.

Mérida-Casermeiro, E., Galán-Marín, G., & Muñoz-Pérez, J. (2001). *An Efficient Multivalued Hopfield Network for the Travelling Salesman Problem*. Neural Processing Letters, 14:203-216.

Mérida-Casermeiro, E., & Muñoz-Pérez, J. (2002). *MREM: An Associative Autonomous Recurrent Network*. Journal of Intelligent and Fuzzy Systems, 12 (3-4), 163-173.

Mérida Casermeiro, E., Muñoz-Pérez, J. & García-Bernal, M.A. (2002). *An Associative Multivalued Recurrent Network*, IBERAMIA 2002, 509-518.

Mérida-Casermeiro, E., Muñoz-Pérez, J., & Domínguez-Merino, E. (2003). *An N-parallel Multivalued Network: Applications to the Travelling Salesman Problem*. Computational Methods in Neural Modelling, Lecture Notes in Computer Science, 2686, 406-413.

Mérida-Casermeiro, E., & López-Rodríguez, D. (2004). *Multivalued Neural Network for Graph MaxCut Problem*, ICCMSE, 1, 375-378.

Mérida-Casermeiro, E., & López-Rodríguez, D. (2005). *Graph Partitioning via Recurrent Multivalued Neural Networks*. Lecture Notes in Computer Science, 3512:1149-1156.

Ozturk, Y., & Abut, H. (1997). *System of associative relationships (SOAR)*, In Proceedings of ASILOMAR.

Parisi, G. (1986). *Asymmetric Neural Networks and the Process of Learning*, J. Phys. A: Math. and Gen., vol 19, L675-L680.

Singh, M. P., Chengxiang, Z., & Dasgupta, C. (1995). *Analytic Study of the Effects of Synaptic Asymmetry*, Phys. Rev. E, vol. 52, 5261-5272.

Treves, A. & Amit, D. J. (1988). *Metastable States in Asymmetrically Diluted Hopfield Networks*, J. Phys A: Math. and Gen., vol. 21, 3155-3169.

ADDITIONAL READING

Abu-Mustafa, Y.S. & St. Jacques, J-M. (1985). *Information capacity of the Hopfield model*, IEEE Trans. on Information Theory, Vol. IT-31, No. 4, 461-464.

Amari, S. (1977). *Neural theory of association and concept formation*, Biological Cybernetics, vol. 26, pp. 175-185.

Amit, D. J. (1989). *Modeling Brain Functions. Cambridge*, U.K.: Cambridge Univ. Press.

Amit, D.J., Gutfreund, H., & Sompolinsky, H. (1987). *Statistical mechanics of neural networks near saturation*, Annals Physics, New York, vol. 173, 30–67.

Athithan, G. (1995). *Associative storage of complex sequences in recurrent neural networks*, in Proceedings IEEE International Conference on Neural Networks, vol. 4, 1971–1976.

Athithan, G. (1995). *A comparative study of two learning rules*, Pramana Journal Physics, vol. 45, no. 6, pp. 569–582.

Athithan, G. (1995). *Neural-network models for spatio-temporal associative memory*, Current Science, 68(9), 917–929.

Bowsher, D. (1979). *Introduction to the Anatomy and Physiology of the Nervous System*, 4th ed. Oxford, U.K.: Blackwell, 31.

Cernuschi-Was, B.(1989). *Partial Simultaneous Updating in Hopfield Memories*, IEEE Trans. on Systems, Man and Cybernetics, Vol. 19, No. 4, 887-888.

Chuan Kuo, I. (1994). *Capacity of Associative Memory*, Ph.D. Thesis, University of Southern California.

Eccles, J. G. (1953). *The Neurophysiological Basis of Mind*. Oxford: Clarendon.

Forrest, B. M., & Wallace, D. J. (1991). *Storage capacity of learning in Ising-spin neural networks*, in Models of Neural Networks, E. Domany, J. L. van Hemmen, and K. Shulten, Eds. Berlin: Springer-Verlag, 121–148.

Gardner, E. (1988). *The space of interactions in neural-network models*, Journal Physics A: Math. and Gen., vol. 21, pp. 257–270.

Gross, D. J., & Mezard, M. (1984). *The simplest spin glass*, Nuclear Phys., vol. B 240 IFS121. DD. 431-452.

Grossberg, S. (1982). *Studies of Mind and Brain*. Boston: Reidel.

Grossberg, S. (ed.), (1986). *The Adaptive Brain*; Vol. I: Cognition Learning, Reinforcement, and Rhythm: Vol. II: Vision. Speech, Language, and Motor Control. Amsterdam, The Netherlands, North-Holland.

Guyon, I., Personnaz, L., Nadal, J.P., & Dreyfus, G. (1988). *Storage and retrieval of complex sequences in neural networks*, Physics Review A, vol. 38, p. 6365.

Hassoun, M.H., & Watta, P.B. (1995). *Alternatives to energy function-based analysis of recurrent neural networks*, in Computational Intelligence: A Dynamical System Perspective, M. Palaniswami et al., Eds. New York: IEEE Press, pp. 46–67.

Hertz, J., Krogh, A., & Palmer, R. G. (1991). *Introduction to the Theory of Neural Computation*. Reading, MA: Addison-Wesley, 40.

Herz, A. V. M., Li, Z., & van Hemmen, J. L. (1991). *Statistical mechanics of temporal association in neural networks with transmission delays*, Physics Review Letters, vol. 66, no. 10, p. 1370.

Hinton, G. E., & Anderson, J. A. (eds.), (1981). *Parallel Models of Associative Memory*, Hillsdale, NJ: Erlbaum.

Hopfield, J. J., & Tank, D. W. (1985). *Neural computation of decisions in optimization problems*, Biological Cybernetics, vol. 52, 141-152.

Kohonen, T. (1977). *Associative Memory: A System-Theoretic Approach.* Berlin: Springer-Verlag.

Krauth, W., & Mezard, M., (1987). *Learning algorithms with optimal stability in neural networks*, Journal Physics A: Math. and Gen., vol. 20, L745–L752.

Lee, B. W., & Sheu, B. J. (1991). *Modified Hopfield neural networks for retrieving the optimal solution*, IEEE Transactions on Neural Networks, vol. 2, p. 137.

Little, W. A., & Shaw, G. L. (1978). *Analytic study of the memory storage capacity of a neural network*, Math. Biosci., vol. 39, 281-290.

Little, W. A. (1974). *The existence of persistent states in the brain*, Muth. Biosci., vol. 19, pp. 101-120.

Loève, M. (1977). *Probability Theory*, Vol. I, Springer-Verlag.

Marr, D. (1969). *A theory of cerebellar cortex*, Journal Physiology, vol. 202, p. 437.

McEliece, R. J. (1977). *The Theory of Information and Coding*, vol. 3 of Encyclopedia of Mathematics and Its Application. Reading, MA: Addison-Wesley.

McEliece, R.J., & Posner, E. C. (1985). *The number of stable points of an infinite-range spin glass memory*, Telecommunications and Data Acquisition Progress Report, vol. 42-83, Jet Propulsion Lab., California Inst. Technol. Pasadena, 209-215.

Nakano, K. (1972). *Associatron-A model of associative memory*, IEEE Transactions on Systems, Man and Cybernetics, vol. SMC-2, 380-388.

Noback, C. R.. & Demarest, R. J. (1975). *The Human Nervous System: Basic Principles of Neurobiology.* New York: McGraw-Hill, 87.

Palm, G. (1980). *On associative memory*, Biol. Cybern., vol. 36, 19-31.

Parisi, G. (1986). *Asymmetric neural networks and the process of learning*, Journal Physics A: Math. and Gen., vol. 19, pp. L675–L680.

Psaltis, D., Hoon Park, C., & Hong, J. (1988). *High order memories and their optical implementation*, Neural Networks, vol. 1, 149-163.

Smetanin, Y. (1994). *A Las Vegas method of region-of-attraction enlargement in neural networks*, in Proceedings 5th International Workshop Image Processing Computer Opt. (DIP-94), SPIE, vol. 2363, 77–81.

Tank, D. W., & Hopfield, J. J. (1987). *Collective computation in neuron like circuits*, Scientific American, pp. 62–70.

Tank, D. W., & Hopfield, J. J. (1986). *Simple optimization networks: An A/D converter and a linear programming circuit*, IEEE Transactions on Circuits Systems. vol. CAS-33. DD. 533-541.

Venkataraman, G., & Athithan, G. (1991). *Spin glass, the travelling salesman problem, neural networks, and all that*, Pramana Journal of Physics, vol. 36, no. 1, 1–77.

Verleysen, M., Sirletti, B., Vandemeulebroecke, A., & Jespers, P. G. A. (1989). *A high-storage capacity content-addressable memory and its learning algorithm*, IEEE Transactions on Circuits Systems, vol. 36, p. 762.

Wozencraft, J. M., & Jacobs, I. M. (1965). *Principles of Communication Engineering.* New York: Wiley.