A New Information Processing Measure for Adaptive Complex Systems

Manuel A. Sánchez-Montañés and Fernando J. Corbacho

***Abstract*—This paper presents an implementation-independent measure of the amount of information processing performed by (part of) an adaptive system which depends on the goal to be performed by the overall system. This new measure gives rise to a theoretical framework under which several classical supervised and unsupervised learning algorithms fall and, additionally, new efficient learning algorithms can be derived. In the context of neural networks, the framework of information theory strives to design neurally inspired structures from which complex functionality should emerge. Yet, classical measures of information have not taken an explicit account of some of the fundamental concepts in brain theory and neural computation, namely that optimal coding depends on the specific task(s) to be solved by the system and that goal orientedness also depends on extracting relevant information from the environment to be able to affect it in the desired way. We present a new information processing measure that takes into account both the extraction of relevant information and the reduction of spurious information for the task to be solved by the system. This measure is implementation-independent and therefore can be used to analyze and design different adaptive systems. Specifically, we show its application for learning perceptrons, decision trees and linear autoencoders.**

***Index Terms*—Adaptive systems, information theory, unsupervised and supervised learning.**

I. Introduction

THE main objective of this work consists in the definition of a new general (i.e., implementation-independent) measure of the amount of information processing performed by (part of) an adaptive system which depends on the goal to be performed by the overall system. As a consequence the degree of processing cannot be completely defined if the goal of the system is unknown.

Fig. 1 summarizes the overall process. The task to be performed by the system is to achieve the goal (g) starting from the input (x). We would like to quantify the amount of information processing the sub-system performs, given that it transforms x into y (the output), when the goal of the overall system is g. We shall denote this amount of information processing by ΔP. Some of the concepts presented in this paper have been published in previous work [1].

An important aspect we would like to emphasize is that the proposed measure must not depend on the implementation specifics but on the global properties of the system. Therefore, it must depend on the relations between the global states of the system in different steps of processing and not on the local structural properties that capture relations between states of the system components (e.g., Barlow's notion of redundancy, see Fig. 2). As a consequence, it can be shown that the different processing units can be implemented by different means much in the same way that schemas correspond to behavioral specifications [2], [3] and can be implemented in a variety of ways such as neural networks, fuzzy logic, etc.

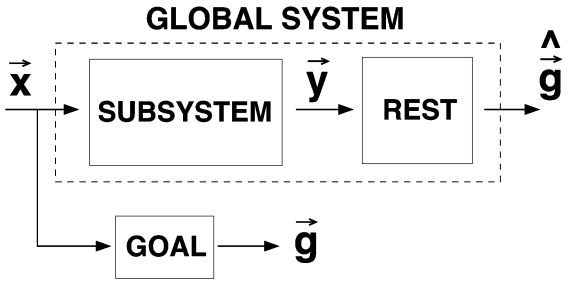

Fig. 1. Global schema of the overall process. The task to be performed by the system is to achieve g starting from x.

A

X	Y	G
h	0001	a
j	0000	b
k	0001	a
m	0000	b

B

X	Y	G
h	0101	a
j	1010	b
k	0101	a
m	1010	b

C

X	Y	G
h	Φ	a
j	Γ	b
k	Φ	a
m	Γ	b

Fig. 2. (A, B) Two different systems with four internal processing units and the same global properties. Each of the bits in Y corresponds to the activity of a single processing unit. While system A only requires a single processing unit to be active to determine G, system B uses most of the processing units to be active. Barlow's redundancy of system B is greater than A since the value of any of the local processing units determines the others. However the number of global internal states and statistical correspondence to G are in both cases equivalent (C). Therefore, they are equivalent from a global statistical point of view.

To place this work in context we should point out that classical information theory schools [4], [5] search for optimal ways of coding information. It is not the aim of this paper to provide a detailed comparison of the different approaches. The authors refer the interested reader to [6] for detailed expositions on this topic. More specifically, information theory has received widespread attention in the neural computation arena [7]–[13] to cite a few examples. In this regard, we fully agree with Atick [9] in the use of information theory as a basis for a first principles approach to neural computation. The relevance derives, as Atick points out, from the fact that the nervous system possesses a multitude of subsystems that acquire, process and communicate information. To bring to bear the more general problem of information, we follow Weaver's [14] classification at three

Manuscript received March 15, 2003; revised October 30, 2003. This work was supported by the MCyT under Grant BFI2003-07276.

The authors are with the Escuela Politécnica Superior, Universidad Autónoma de Madrid, Madrid 28049, Spain and Cognodata Consulting, Madrid 28010, Spain (e-mail: manuel@ii.uam.es).

Digital Object Identifier 10.1109/TNN.2004.828768

different levels: 1) technical problems: how accurately can the symbols of communication be transmitted? 2) semantic problems: how precisely do the transmitted symbols convey the desired meaning? 3) effectiveness problems: how effectively does the received meaning affect the receiver's conduct in the desired way?

We claim that any adaptive system (including the brain) living in an active environment must solve these three problems. Yet, as Weaver [14] already pointed out, classical information theory deals mainly with the technical problem. We claim that even today a shift of view is necessary to take into proper consideration the semantic and the effectiveness levels. This paper provides a step towards dealing with the semantic and the effectiveness problems by making optimal coding depend on the specific task(s) to be solved by the system. In the classical approach, the emphasis is on the maximization of information transfer. That is, processing is passive instead of being active (i.e., elaborating the data and approaching the goal) hence, posing a paradox for an information processing system. Notice, in this regard, that a perfect communication channel has maximal-mutual information yet minimal information processing.

The information processing measure presented in this paper is implementation independent and therefore, can be used to analyze and design different adaptive systems. Several classical supervised and unsupervised learning algorithms are obtained as the optimal solutions for special cases. Specifically, we show its application for learning perceptrons, decision trees, and linear autoencoders.

II. Requirements for the New Information Processing Measure

Next, we would like to impose a set of requirements that this new measure should have in order to be regarded as a candidate for an active general-information processing measure. From these requirements a specific family of measures emerges and we shall select a specific measure for specificity reasons and derive the properties this specific measure has. Thus, let us now derive a specific information processing measure ΔP that meets the following requirements.

1) It must be a measurable quantity that does not depend on the specific system implementation, that is, it should depend on the statistical properties of the states of the system and not on local properties dependent on implementation details.
2) It should take into account the task(s) to be solved by the system. The input to the system can be statistically rich and complex, yet it may be mostly useless if it is not related to the task (nonreversibility property). Thus, it should take into account how much the data has to be processed (number of transformations) in order to extract the relevant information for the task. Therefore the information processing measure should penalize both the loss of relevant information and the introduction of spurious information.
3) It must be an effective processing measure, that is, ΔP must depend on x (the input), y (the output) and g (the goal) but it must not depend on the information processing path taken to go from x to y $(x \rightarrow y)$, that is

$$\Delta P(x \rightarrow y|g) = \Delta P(x \rightarrow w|g) + \Delta P(w \rightarrow y|g) \quad (1)$$

for all x, y, w, g, that is

$$\Delta P(x \rightarrow y|g) = d(x,g) - d(y,g) \quad (2)$$

where the function $d(a,b)$ defines a sort of distance function that should depend on the global statistical relationships between the states of a and the states of b.
4) The maximum value for $\Delta P(x \rightarrow y|g)$, when x and g are fixed, must occur when $y = g$ and as a consequence $d(y,g) \geq d(g,g)$. So that d can be chosen, without loss of generality, such that

$$d(y,g) \geq 0; \quad d(g,g) = 0 \quad (3)$$

for all y, g.

Then, ΔP is null for a perfect communication channel (in the classical sense, i.e., an exact copy of the input message is produced) and maximal for the case of perfect transformation to the objective alphabet (active property).
5) The maximum value for $\Delta P(x \rightarrow y|g)$ when g is allowed to vary, for all x, y and g is

$$\Delta P(x \rightarrow y|g) \leq \Delta P(x \rightarrow y|y) \quad (4)$$

and as a consequence $d(x,g) \leq d(x,y) + d(y,g)$ which corresponds to the triangular inequality. So that taking into account the triangular inequality and (3), it can be concluded that d is a *pseudodistance* function.
6) It should account for uncertainties introduced by different means, such as: loss of meaningful information, environmental noise, stochasticity of the processing elements, and so on, while preserving the relevant part of the information.

A. Selection of the Pseudodistance Function for the Information Processing Measure

As it was expressed before, one of the requirements for the new information processing measure is that it must not depend on the specific architecture in which the overall system is implemented. This implies that the new measure must capture the global statistical relations which take into account the relations between the global states of the system in different steps of processing and not the local structural properties which capture relations between states of the system components. From this point of view, the Shannon's conditional entropy function [6] meets the desired properties for $d(x,g)$ for discrete systems. For continuous systems the representation should be first discretized. Another possible candidate would be the Bayes error, whereas the mean square error would not satisfy the implementation-independence requirement as it depends on the local structural properties of the system.

So let us consider Shannon's conditioned entropy, since it meets the desired properties for $d(x,g)$ and hence $d(x,g) = H(G|X)$ and using (2)

$$\Delta P(x \rightarrow y|g) = H(G|X) - H(G|Y) \quad (5)$$

A

X	Y	G
h	Φ	1
j	Γ	2
k	△	3
m	Φ	1

H(G|Y)=0
H(Y|G)=0

B

X	Y	G
h	Φ	1
j	Γ	2
k	△	3
m	Σ	1

H(G|Y)=0
H(Y|G)>0

C

X	Y	G
h	Φ	1
j	Γ	2
k	Γ	3
m	Φ	1

H(G|Y)>0
H(Y|G)=0

Fig. 3. State tables for three different systems (A), (B), and (C). X represents the input space, Y represents the output to the rest of the whole system and G represents the corresponding goal space.

so that ΔP corresponds to the difference in uncertainty before and after the information processing is performed. When the entropy is considered, $\Delta P \leq 0$ (we refer the reader to the generalized data processing inequality theorem in the Appendix 7.1). Additionally, the maximum value occurs when $\Delta P = 0$ and it is achieved in a perfect communication channel (i.e., when $x = y$). Hence, it would be better for the system not to do any processing which is a paradox when dealing with a measure for information processing. The same problem occurs when Fisher information or the Bayes error is utilized. This is due to the fact that in the more classical approaches the amount of information never increases when it undergoes any processing, that is, processing is passive instead of being active (i.e., elaborating the data and approaching the goal). In this regard, a perfect communication channel has maximal mutual information yet minimal information processing. Hence, it can be concluded that information processing is more than avoiding the creation of uncertainty and it must also take into account the reduction of spurious information. In the next section, a new information processing measure will be presented that meets all the aforementioned requirements.

III. Definition of the New Information Processing Measure

As previously stated, the new information processing system must keep all the relevant information, but at the same time it must reduce the spurious information. Hence, we define a function d that also takes into account the reduction of spurious information, which can be seen as the uncertainty in y given g.

An example may well illustrate this point. Consider the three systems depicted in Fig. 3. System A has 0 uncertainty and 0 spurious information about the goal. On the other hand, system B has larger spurious information $H(Y|G) > 0$ since there is a spurious state at Y for $G = 1$. Contrarily, system C has 0 spurious information but it has larger uncertainty $H(G|Y)$ since the same state of Y gives rise to two different states at G for $Y = \Phi$.

Thus, we present a new function d

$$d(x, g) = H(X|G) + \alpha H(G|X) \tag{6}$$

where $\alpha > 0$ and weights the creation of uncertainty versus the creation of spurious information. This function d meets the required properties, namely $d(x, g) \geq 0$, $d(g, g) = 0$, $d(x, g) \leq d(x, y) + d(y, g)$. Additionally, spurious states and loss of relevant information cause d to grow, so that $d(y, g) = 0$ if and only if y is a bijection of g (see Appendix 7.2 for the corresponding proof). Hence, with d as in (6) the information processing measure ΔP can be expressed as

$$\Delta P(x \to y|g) = [H(X|G) - H(Y|G)] - \alpha [H(G|Y) - H(G|X)] \tag{7}$$

where the first term corresponds to *complexity reduction*, that is, minimization of spurious information, whereas the second term corresponds to *uncertainty creation*, that is, loss of relevant information. ΔP becomes larger as Y gets closer to G reflecting the fact that the system processing is taking the output of the system closer to the goal. This can be achieved by minimizing complexity and/or uncertainty. ΔP may take positive and negative values whereas the term of loss of information is always zero or positive (see Appendix 7.1 for a proof). Note that $\Delta P = 0$ for a perfect communication channel (see Appendix 7.3 for a proof).

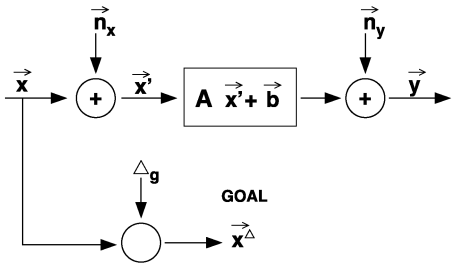

Fig. 4. Schema for a linear autoencoder. The goal of the system is to communicate the input $\vec{x}$ with a degree of precision given by Δ_g.

IV. Results

To validate the new information processing measure, several test cases are used as a proof of concept. The first test case describes a specific instance of a linear system, namely an autoencoder. The second, third and fourth set of test cases are based on learning the optimal structure of a network of nonlinear units. The first set deals with a synthetic classification problem and the second set deals with several classification problems from the *Proben*1 benchmark archive [15]. In both cases we have compared the solutions that optimize the new measure with respect to the solutions that optimize mutual information. The new measure proves to be clearly superior under conditions of noise, overfitting and allocation of optimal number of resources. The fourth test case deals with stochastic neurons and gives rise to population coding for the Gaussian classification task introduced in the second test case. And lastly the fifth test case deals with the induction of decision trees using the new information processing measure.

A. A Case of a Linear System: An Autoencoder

Consider the system in Fig. 4 where a layer of noisy linear neurons responds to the stimulus $\vec{x}$ as

$$\vec{y} = A(\vec{x} + \vec{n}_{\vec{x}}) + \vec{b} + \vec{n}_{\vec{y}} \tag{8}$$

where $\vec{y}$ is the vector of the responses in the layer, $\vec{n}_{\vec{x}}$ is the noise in the input (due to noisy receptors for instance) and $\vec{n}_{\vec{y}}$ is the noise intrinsic to the neurons. For simplicity reasons we assume both kinds of noise are zero-mean normal distributed with

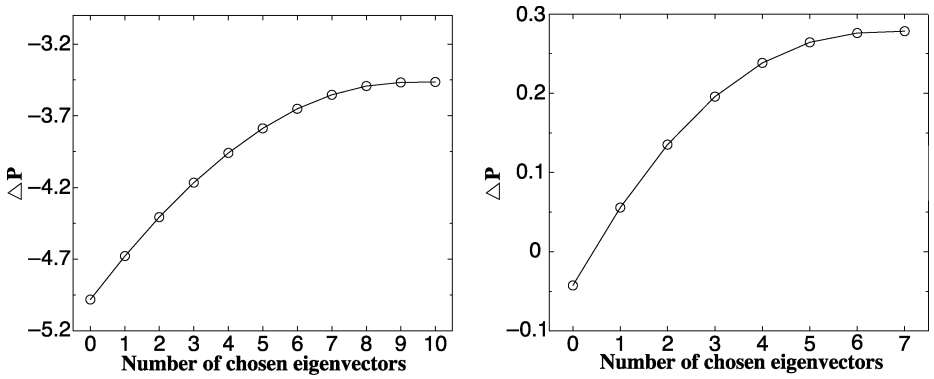

Fig. 5. Contribution of the chosen eigenvectors in the optimal configuration of the linear autoencoder. (a) $b = 1/2$. (b) $b = 2$.

covariance matrices N_x and N_y respectively. The goal of the system is to communicate the input $\vec{x}$ with a degree of precision given by Δ_g. The input statistics are assumed to be well represented by a multidimensional Gaussian of covariance matrix C.

Now we need to discretize $\vec{x}'$ and $\vec{y}$ in order to evaluate our processing measure. In general, the entropy of the discretization of a continuous variable $\vec{a}$ ($\vec{a}^{\Delta}$) can be expressed as $H(\vec{a}^{\Delta}) = H(\vec{a}) - \log(\Delta_a)$, where Δ_a is the level of discretization in $\vec{a}$, for sufficiently small Δa [6]. Note that the term $\log(\Delta_a)$ is mathematically equivalent to the entropy of a Gaussian noise with variance equal to $(2\pi)^{-1}\Delta_a^2$. In our system, the noise $n_{\vec{x}}$ provides a natural discretization of $\vec{x}'$ and on the other hand $\vec{n}_y$ provides a natural discretization to $\vec{y}$. Then, following an approximation introduced in [16], $H(\vec{x}'^{\Delta}) = H(\vec{x}') - H(\vec{n}_x)$ and $H(\vec{y}'^{\Delta}) = H(\vec{y}') - H(\vec{n}_y)$. Moreover, from a mathematical point of view $H(\vec{y}|\vec{x}^{\Delta}) = H(\vec{y}|\vec{x} + \vec{n}_d)$ where $\vec{n}_d$ is a Gaussian noise of variance equal to $(2\pi)^{-1}\Delta_x^2$ and Δx is the discretization bin in $\vec{x}$. Thus, the system is characterized by the entropies $H(\vec{x}'^{\Delta}|\vec{g}^{\Delta}) = 1/2 \ln\det(I + N_x^{-1}(2\pi\Delta_g^{-2} \cdot I + C^{-1})^{-1})$, $H(\vec{y}'^{\Delta}|\vec{g}^{\Delta}) = 1/2 \ln\det(I + N_y^{-1}A(N_x + (2\pi\Delta_g^{-2} \cdot I + C^{-1})^{-1})A^T)$, $H(\vec{y}^{\Delta}) = 1/2 \ln\det(I + N_y^{-1}A(C + N_x)A^T)$ and $H(\vec{x}'^{\Delta}) = 1/2 \ln\det(I + N_x^{-1}C)$.

Using the property $H(\vec{a}) - H(\vec{a}|\vec{b}) = H(\vec{b}) - H(\vec{b}|\vec{a})$ [6] in (7), and after simplifications, the ΔP in our system can be expressed as

$$\Delta P(\vec{x}'^{\Delta} \to \vec{y}^{\Delta}|\vec{x}^{\Delta}) = \frac{1}{2}\ln\frac{(\det(I+D))^{1+\alpha}}{(\det(I+S))^{\alpha}} + \frac{1}{2}\ln\frac{\left(\det(I+V\Phi V^T)\right)^{\alpha}}{(\det(I+VV^T))^{\alpha+1}} \quad (9)$$

where $D = N_x^{-1}(2\pi\Delta_g^{-2} \cdot I + C^{-1})^{-1}$, $S = N_x^{-1}C$, $\Phi = (N_x + N_xD)^{-1/2}(N_x + C)(N_x + N_xD)^{-1/2}$ and $V = N_y^{-1/2}A(N_x + N_xD)^{1/2}$.

The second term in the summation 9 is the one which determines the maximization of ΔP since the other one does not depend on A. It can be easily proven that if R is a rotation in the space of neurons, then the solution $\hat{V} = R \cdot V$ has exactly the same ΔP than V. Thus, there does not exist a unique optimal configuration but a family of optimal solutions.

In the Appendix 7.4 we derive the optimal family of solutions, which can be summarized as follows:

- Let us define $a_i \equiv \sigma_{Ci}^2/\sigma_x^2$ and $b \equiv \Delta_g^2/\Delta_x^2$, where σ_{Ci}^2 are the eigenvalues of C. Take the eigenvectors of C which satisfy

$$a_i > \frac{1 + b + \sqrt{(1+b)^2 + 4\alpha b}}{2\alpha}. \quad (10)$$

  In case no eigenvalues satisfy this inequality, take $A = 0$. If the number of neurons is less than the number of eigenvectors satisfying the requirement, take the eigenvectors with greatest eigenvalues
- Assign **only** one neuron to one of the selected eigenvectors making its receptive field proportional to the eigenvector using gain

$$\frac{\sigma_y}{\sigma_x}\sqrt{\frac{\alpha a_i^2 - (a_i + b + a_i b)}{(a_i + 1)(a_i + b + a_i b)}}. \quad (11)$$

Any optimal solution is then a rotation in the space of neurons of this basic solution.

Similar results are obtained when maximizing the mutual information between $\vec{x}$ and $\vec{y}$ and imposing additional constraints to the system such as the ones introduced in [17]. Note that we need to impose no constraint at all in order to obtain the results reported in this section.

In Fig. 5, we show the optimal configuration when x has 10 components with the set a_i homogeneously distributed between 2 and 8, and $\alpha = 1$. If the required discretization is chosen to be smaller than the discretization induced by the input noise n_x in x ($b = \Delta_g^2/\Delta_x^2 = 1/2$), then the system performance ΔP is negative (Fig. 5 left). Although all the eigenvectors are chosen and contribute to make ΔP larger the system cannot communicate the input with the desired precision. On the contrary, when $b = 2$ the ΔP of the optimal system is positive, and only includes seven eigenvectors [Fig. 5(b)].

The optimal configuration is thus equivalent to principal component analysis (PCA) [18] where the number of eigenvectors is determined by the input and noise statistics as well as by the desired precision. Moreover, PCA can be seen as a special case of independent component analysis (ICA) [19], [20] where the statistics of the input sources are Gaussian. Therefore, we expect to obtain similar results to ICA when applying the new information processing measure to the non-Gaussian statistics case.

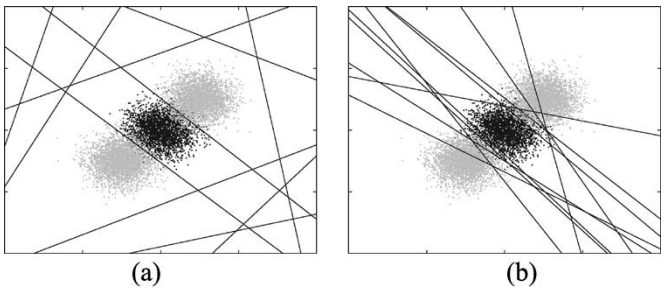

Fig. 6. Comparison of the solutions that optimize the new measure with respect to the solutions that optimize mutual information for the perceptron case. (a) Results when using ten processing elements as the maximum number of resources with the new information processing measure. Notice that the new measure needs to use only two out of the ten maximum number. (b) Similarly for mutual information. Notice that mutual information uses the ten classifiers.

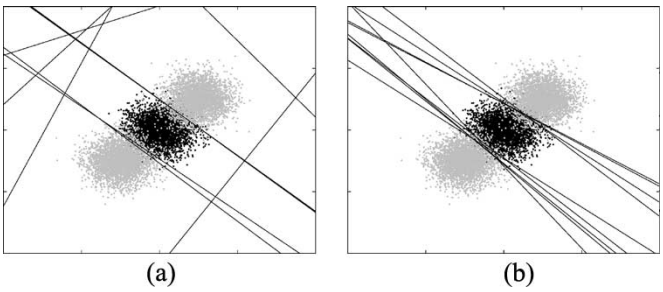

Fig. 7. Results when using ten stochastic processing elements as the maximum number of resources with the new information processing measure. (a) Level of noise $= 2\%$. (b) level of noise $= 20\%$.

B. Perceptron Learning

1) Perceptron for a Simple Classification Task: Here we investigate learning the optimal structure of a network of nonlinear units in a simple classification task. The dataset *Gaussians* consists of three equiprobable clusters of data elements belonging to two different classes. There are three mutually exclusive processes which generate vectors ($x1$, $x2$) following Gaussian overlapping distributions (see Fig. 6).

Two of the processes are considered of class "grey" while one of them is considered of class "black". The goal of the global system is thus, to predict, given a new example ($x1$, $x2$), to which of the two classes it belongs to.

We consider that in our global system the first processing step is a layer of nonlinear neurons. The output of the ith classifier (y_i) is 1 in case $\vec{m}_i\vec{x} + b_i > 0$, 0 otherwise, where $\vec{x}$ is the input pattern. The binary vector composed by all the classifiers outputs $\vec{Y}$ determines the achievable accuracy of the rest of the system as well as the amount of processing it has to do.

The adaptive system must find the configuration that maximizes ΔP. The optimal classifiers configurations have been generated by searching the parameter space by means of a genetic algorithm [21] due to its global search properties. The α parameter is 4, the number of examples used in the optimization is 10 000. We have performed several computer experiments with different random seeds and initial conditions leading to the same results.

We have compared the solutions that optimize the new measure with respect to the solutions that optimize mutual information. For the case of the new measure, processing is equal to the reduction of spurious information minus loss of the relevant information. Yet, mutual information only takes into account uncertainty minimization ignoring the reduction of complexity. Fig. 6(a) displays the configuration selected when using a pool of ten nonlinear units. Note that the optimal configuration only uses two of them since the output of the rest is kept constant. However, if mutual information $I(Y;G)$ is chosen as the objective function to maximize, a configuration where all the resources are used is obtained [Fig. 6(b)]. This is due to the fact that mutual information only takes into account uncertainty minimization ignoring extraction of spurious information for this simple task.

TABLE I
TEST ERRORS FOR THE DIFFERENT DATABASES. C IS THE NUMBER OF USED CLASSIFIERS OUT OF THE MAXIMUM ALLOWED (15 FOR THE *HEART1* DATABASE, 10 FOR THE OTHERS)

	ΔP		Mutual Information		Proben1
Data Set	Test error (%)	C	Test error (%)	C	Test error (%)
gaussians	5.2	2	5.4	10	-
cancer1	0.57	1	11.49	10	1.149
cancer2	4.598	1	14.94	10	4.52
heart1	21.74	2	57.4	15	19.72

2) Stochastic Neurons and Population Coding: In this section we study how intrinsic noise in the processing elements affects the optimal system. Thus, we consider the same classification problem as in the previous section but now the neurons are stochastic. The output of each neuron is computed as previously, but then each neuron switches its output with certain probability. The optimal system is again calculated using a genetic algorithm. Notice that for a level of noise of 2% the system uses more than two classifiers [Fig. 7(a)]. Also notice that the new measure begins to use more resources to account for the noise in the input data. When the noise is increased up to 20% we see that the optimal system uses ten classifiers [Fig. 7(b)].

Thus, we observe that the maximization of ΔP adjusts the number of resources used to the level of inherent noise in the processing elements. This is related to population coding as described by [22].

3) Perceptron for Proben1 Tasks: In this section, we will use the *cancer1*, *cancer2*, and *heart1* databases taken from the *Proben1* archive [15] to validate the derived perceptron learning algorithm for deterministic neurons.

The α parameter has been tuned using a separate validation set. In all cases we have compared the solutions that optimize the new measure with respect to the solutions that optimize mutual information and the best result reported by [15]. Table I displays the results obtained and includes C as the number of resources utilized by each system.

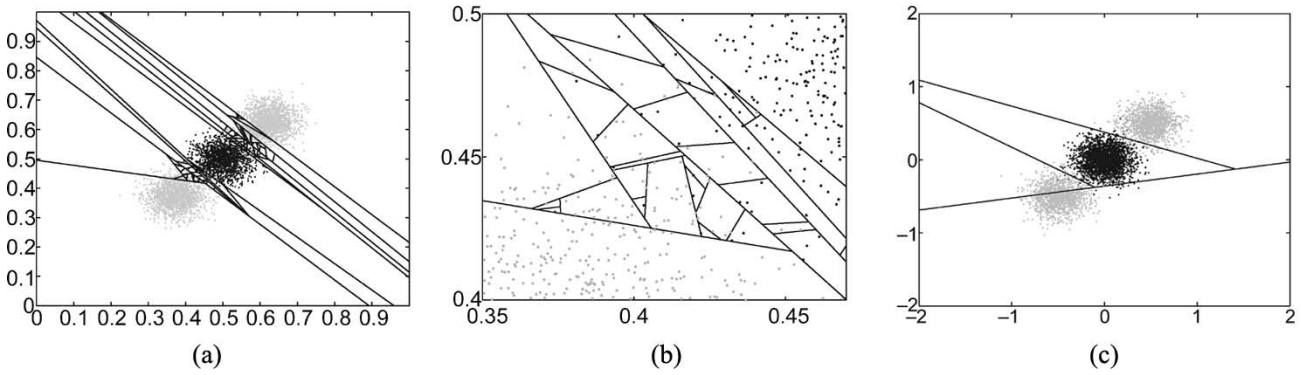

Fig. 8. Overfitting with a classical inductive decision tree. (a) Results of applying ID3 to the Gaussians database without pruning. (b) Zoom in of (a). (c) Result of applying the new measure to the induction of the tree.

C. Construction of Decision Trees

In this section, we apply the new information processing measure to the induction of decision trees. The output of a decision tree for an input pattern is the terminal node that classifies that pattern [23]. We would like $\Delta P(X \rightarrow Y|G)$ to be maximized for the induced tree. Since $\Delta P(X \rightarrow Y|G) = d(X,G) - d(Y,G)$, the maximization of this quantity is equivalent to the minimization of $d(Y,G)$ since the input statistics are constant in this context. It follows that $d(Y,G)$ can be written as (see Appendix 7.5 for the details)

$$d(Y,G) = d(Y_{noN},G) + p(N)\left(H_N(Y|G) - \alpha I_N(Y;G)\right) \tag{12}$$

where $d(\vec{Y}_{noN},G)$ is the distance to the goal from the tree without the subtree N and $H_N(Y)$ and $I_N(Y;G)$ are the entropy and mutual information computed using the local statistics in N.

Therefore, it is natural to define a greedy construction algorithm that starts with a root node, choosing the expansion A that maximizes $\alpha I_N(A;G) - H_N(A|G)$, and then use it recursively in the children subtrees. Note that, if this quantity is negative, it will contribute to make (12) greater. Therefore, if we reach a node where all possible expansions make $\alpha I_N(A;G) - H_N(A|G)$ negative we stop expanding that branch. As it can be seen, the new measure provides a method for constructing decision trees as well as a natural stopping criteria to avoid overfitting. This is in contrast with many other algorithms which use a local information gain measure such as

$$Gain_N(A) \equiv H_N(G) - H_N(G|A) = I_N(G;A) \tag{13}$$

in order to evaluate the goodness of an expansion A in node N (such as ID3 [24]). Since the information gain (13) is always ≥ 0, it achieves the value zero when the number of examples in the node to expand is one or all the examples at the node have the very same class [24]. Therefore, a recursive application of the gain information criteria would make the tree expand since all the examples at a terminal node have the same class. This produces very complex trees in general that need to be post-pruned [23]. Additionally, this procedure has difficulties with attributes with many possible values [24]. For this reason in the literature the gain ratio is defined [24] as

$$\text{Gain ratio}_N(A) \equiv \frac{I_N(G;A)}{H_N(A)} \tag{14}$$

to overcome this problem, but still it has the problem of being always a positive quantity.

On the other hand, the greedy maximization of the new proposed measure for the induction of decision trees can be seen as a technique which combines the good features from the information gain, the gain ratio and early stopping. Fig. 8(a) displays the results of applying ID3 to the Gaussians database without pruning. Fig. 8(b) corresponds to the zoom in of Fig. 8(a). Fig. 8(c) shows the result of applying the new measure to the induction of the tree. As it can be observed no pruning is needed.

V. Discussion

Many of the classical information based techniques are implementation-dependent, whereas the measure of information processing proposed in this paper is based on the global statistical properties of the system independent of any particular implementation. In this regard, the notion of spurious information exposed in this paper is different from the notion of redundancy exposed by Barlow [7], [9], [25] (see Section I). Barlow's redundancy depends on the system implementation, (redundancy between the elements of processing), whereas our notion of spurious information is based on the global activity of the system and depends on the statistical relations between the global states of the system.

In this regard, the process performed by the replication of one deterministic neuron many times is equivalent to the process done by one of such neurons. Therefore, our concept of spurious information is not a matter of independence between processing elements, but of unnecessary information in the global activity of the system with respect to the goal. Therefore, population codings can be studied within this framework since the measure does not explicitly punish this kind of coding; moreover, it considers they are appropriate in order to deal with noise (cf. von Neuman's redundancy scheme [26]). Next we will describe the information bottleneck (IB) method [27] since it shares some similarities with the work exposed in this paper.

A. Comparison With the Information Bottleneck Method

The IB method [27] has some commonalities with the framework presented in this paper since it also allows for the construction of learning systems by searching for an optimal internal rep-

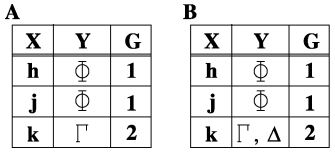

Fig. 9. The systems (A) and (B) are equivalent for the IB functional. In B the state Γ has been split in two states Γ and Δ which are randomly activated. Therefore, the IB method does not determine the number of internal states. However, our processing measure penalizes the introduction of this spurious information: since $H(Y|G)_A < H(Y|G)_B$ and $H(G|Y)_A = H(G|Y)_B$ then $\Delta P_A > \Delta P_B$.

resentation. The IB method is derived from an interpretation of rate distortion theory [6]. Following the notation in this paper, the IB method attempts to minimize the functional

$$L = I(Y;X) - \beta I(Y;G) \tag{15}$$

with β a constant. Using the property $I(a;b) = H(a) - H(a|b) = H(b) - H(b|a)$ [6] the previous equation can be rewritten as

$$L = \beta H(Y|G) - (\beta - 1)H(Y) - H(Y|X). \tag{16}$$

On the other hand, with the new information processing measure proposed in this paper the following expression must be minimized

$$\begin{aligned} d(X,Y) &= H(Y|G) + \alpha H(G|Y) \\ &= (1+\alpha)H(Y|G) - \alpha H(Y) + \alpha H(G) \end{aligned} \tag{17}$$

where we have used again the property $H(a) - H(a|b) = H(b) - H(b|a)$. Note the last term does not take part in the optimization since it is constant for a given problem. The main difference between our approach and IB is that $H(Y|X)$ plays a role in L (being negligible only when $\beta \to \infty$). That is, in IB the introduction of noise is not penalized but quite on the contrary. However, in the framework proposed in this paper the introduction of noise is always penalized due to the introduction of spurious states. In order to illustrate this point, let us consider a system with internal states Y. If one of such states, e.g., y_1, is split into y_1^a and y_1^b randomly (therefore $p(g|y_1^a) = p(g|y_1^b) = p(g|y_1)$ and $p(x|y_1^a) = p(x|y_1^b) = p(x|y_1)$) then it is straightforward to show that $I(X;Y) = I(X;Y^*)$ and $I(Y;G) = I(Y^*;G)$ where Y^* are the internal states after the splitting. As a consequence, the IB functional is the same in both situations (Fig. 9). However, our measure penalizes the introduction of spurious states and therefore the system in Fig. 9(a) has always greater ΔP than the one in Fig. 9(b).

Another consequence of the term $H(Y|X)$ in the IB functional is its preference in some situations to solutions intrinsically noisy. These solutions, apart from introducing spurious noisy information in the representation, produce a loss of relevant information [Fig. 10(a) and (b)]. However, ΔP penalizes both the spurious information and the loss of relevant information, having a preference for deterministic solutions [Fig. 10(c) and (d)].

VI. Conclusion

This paper presents a new information processing measure that allows the optimal construction of adaptive complex systems. We have presented a general framework under which we have shown that several classical supervised and unsupervised algorithms fall and new efficient algorithms are developed. In particular we have shown how PCA is a special case of the linear autoencoder under the proposed framework. We would also expect to obtain ICA when the Gaussianity restriction is not imposed. The framework proposed also works for nonlinear systems. In this regard we have shown several examples based on networks of nonlinear neurons and we are currently elaborating more cases.

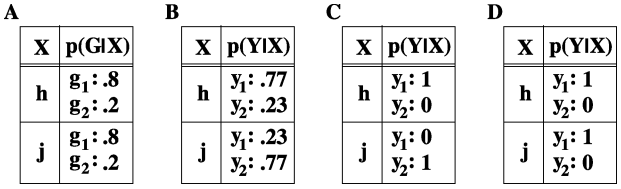

Fig. 10. (A) A problem where the goal g is not completely determined by the input x. (B) Solution which optimizes the IB functional when the number of internal states is constrained to be two and $\beta = 3$. The optimal internal representation is stochastic, giving rise to the creation of spurious information ($H(Y|G) > H(X|G)$) and the increment of uncertainty about the goal ($H(G|Y) > H(G|X)$). Therefore $\Delta P_A < 0$. (C) and (D) Solutions which maximize ΔP. The number of internal states is automatically determined by the maximization of ΔP as two. When $\alpha \geq 2.6$ the optimal solution is a perfect communication channel of x (figure c) whereas if $\alpha < 2.6$ the optimal solution is a system with a constant internal state (D). Note that in both cases the optimal representation is deterministic.

While many of the information theoretic frameworks are more closely related to unsupervised learning, the proposed framework is able to naturally cope with supervised learning problems as well, since the dependency of the optimal coding on the goal to be obtained is central to the whole theory. For all the adaptive systems presented in this paper the new measure has naturally given rise to the optimal use of resources hence leading to the avoidance of the overfitting problems that occur in many adaptive systems, e.g., perceptrons and inductive decision trees.

To prove that the proposed framework is independent of the language of representation we have also applied the new information processing measure to the induction of decision trees. As a result the obtained decision trees have shown good classification performance while automatically avoiding the use of an excessive number of nodes without the need for post-pruning. Future work includes validation of the general framework using other representation languages such as recurrent neural networks and hidden Markov models. This work can also be extended to the problem of building reliable systems with unreliable components (cf. [28], [29]) since it naturally allows for population coding as mentioned in the discussion section.

Appendix

A. Generalized Data Processing Inequality

Let us consider the Markov chain in Fig. 11. $\vec{y}_1$ and $\vec{g}$ are (possibly) stochastic functions of $\vec{x}$, and $\vec{y}_2$ is a (possibly) stochastic function of $\vec{y}_1$. That is, $\vec{y}_2$ and $\vec{g}$ are conditionally independent

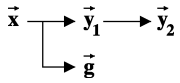

Fig. 11. Flow of information in a closed system. The second processing step is statistically independent of g given y_1.

given $\vec{y}_1$. The joint probability function of $\vec{y}_1$, $\vec{y}_2$ and $\vec{g}$ can be then described as

$$p(\vec{y}_1, \vec{y}_2, \vec{g}) = p(\vec{y}_1) \cdot p(\vec{g}|\vec{y}_1) \cdot p(\vec{y}_2|\vec{y}_1). \tag{18}$$

Then it is easy to prove that $I(\vec{y}_2; \vec{g}|\vec{y}_1) = 0$ [6]. On the other hand, $I(\vec{g}; \vec{y}_1, \vec{y}_2)$ can be described in two equivalent manners [6]

$$I(\vec{g}; \vec{y}_1, \vec{y}_2) = I(\vec{y}_1; \vec{g}) + I(\vec{y}_2; \vec{g}|\vec{y}_1) = I(\vec{y}_2; \vec{g}) + I(\vec{y}_1; \vec{g}|\vec{y}_2) \tag{19}$$

where, using the fact that $I(\vec{y}_2; \vec{g}|\vec{y}_1) = 0$ and $I(\vec{y}_1; \vec{g}|\vec{y}_2) \geq 0$ it follows:

$$I(\vec{y}_2; \vec{g}) \leq I(\vec{y}_1; \vec{g}) \tag{20}$$

that is, no any processing can increase the information of a closed system about the objective. In particular, if we choose $y_1 = x$

$$I(\vec{y}; \vec{g}) \leq I(\vec{x}; \vec{g}). \tag{21}$$

Finally, since $I(a; b) = H(a) - H(a|b)$ [6] this expression is equivalent to

$$H(\vec{g}; \vec{x}) \leq H(\vec{g}; \vec{y}). \tag{22}$$

B. Properties of the Measure of Distance Between Two Random Variables

Given the definition of distance measure between A and the desired goal G for discrete variables

$$d(A, G) = \alpha H(G|A) + H(A|G). \tag{23}$$

It is semidefinite positive.

The discrete entropies are always positive or 0 [6]. For continuous entropies, once they are discretized, they are always positive. Since $\alpha > 0$, this makes $d \geq 0$.

It is reflexive.

This is immediate since $H(A|A) = 0$ [6].

It satisfies the triangular inequality

First we will prove that

$$H(U|W) \leq H(U|Z) + H(Z|W) \tag{24}$$

for any random variables U, W and Z. Using the equalities [6]

$$\begin{aligned} I(V; W|Z) &= H(V|Z) - H(V|W, Z) \\ &= H(W|Z) - H(W|V, Z) \end{aligned} \tag{25}$$

where we get $H(V|Z) = H(V|W, Z) + H(W|Z) - H(W|V, Z)$. Because $H(W|V, Z) \geq 0$ and $H(V|W, Z) \leq H(V|W)$ [6] we prove (24). If we use this in (23) together with the fact that $\alpha \geq 0$

$$\begin{aligned} d(A, G) &= H(A|G) + \alpha H(G|A) \\ &\leq H(A|B) + H(B|G) + \alpha \left(H(G|B) + H(B|A)\right) \\ &= d(A, B) + d(B, G). \end{aligned} \tag{26}$$

The system is optimal $(d(A, G) = 0)$ **if and only if the output of the system is the objective or a relabeling of it**.

We say that the output A of the system is a relabeling of the objective G when there is a one-to-one mapping between all the elements with nonzero probability in A and all elements with nonzero probability in G. For $d(A, G)$ to be 0 it is needed to satisfy simultaneously $H(A|G) = 0$ and $H(G|A) = 0$. Let us first consider $H(A|G) = 0$. Using its definition we have $H(A|G) = -\sum_{i,j} p(g_i)p(a_j|g_i) \log(p(a_j|g_i)) = 0$. Because $0 \leq p \leq 1$, all the terms in the summatory are greater or equal to zero. Therefore, $H(A|G) = 0$ is only possible if $p(a_j|g_i) = 0$ or 1 for every a_j which $p(g_i) > 0$. Thus, each symbol with nonzero probability in G is associated with only one element in A. This, together with analog considerations for the case $H(G|A) = 0$, we show that $d(A, G) = 0$ implies that there is a one to one correspondence between the nonzero probability symbols in A and the nonzero probability symbols in G.

On the other hand, if there is a bijection between A and G, for any symbol a_i in A with nonzero probability $p(g_j|a_i)$ can only be 0 or 1, since in other case there would be more symbols in g corresponding to a_i. The same can be argued for $p(a_i|p_j)$, which all together imply $d(A, G) = 0$.

C. $\Delta P = 0$ for a Perfect Communication Channel

A communication channel where the transmitter sends the signal X and the receiver takes Y is called "perfect" if the transmission occurs without any loss, that is, $I(x; y)$ is maximum given the statistic distribution $p(x)$. Because $I(x; y) = H(x) - H(x|y) = H(y) - H(y|x)$ and the entropies are positive, a perfect communication channel satisfies $H(y) = H(x)$ and $H(y|x) = H(x|y) = 0$ [6].

In our formalism, the "goal" in this case is to recover the original signal X, so $G = X$. Using this fact in (7) and using the fact that $H(Y|X) = 0$, $H(X|Y) = 0$ in a perfect communication channel, and $H(X|X) = 0$ in general, we get

$$\begin{aligned} \Delta P &= \left(H(X|G) - H(Y|G)\right) - \alpha \left(H(G|Y) - H(G|X)\right) \\ &= \left(H(X|X) - H(Y|X)\right) - \alpha \left(H(X|Y) - H(X|X)\right) \\ &= (0 - 0) - \alpha(0 - 0) \\ &= 0. \end{aligned}$$

D. Maximization of ΔP for the Linear Autoencoder

It is straightforward to show that if R_y is a rotation matrix of $n_y \times n_y$ components (rotation in the space of neurons), then the change $V \rightarrow R_y \cdot V$ keeps ΔP (9) unaltered. Therefore there is not a unique maximum but a *family* of optimal configurations. On the other hand, it can be easily proven that if the absolute

value of any component of V tends to ∞ then the functional (9) goes to $-\infty$. Thus, the global maximum of the functional occurs in a fixed point where the gradient of the function is null

$$\alpha(I+V\Phi V^T)^{-1}V\Phi-(\alpha+1)(I+VV^T)^{-1}V=0. \quad (27)$$

Let us consider $\hat{V}$ as one of such fixed points. Since $\hat{V}\hat{V}^T$ is a symmetric matrix, there exists a rotation R_y such that $R_y\hat{V}\hat{V}^TR_y^T$ is diagonal. That is, the point $\tilde{V}=R_y\cdot\hat{V}$ makes $\tilde{V}\tilde{V}^T$ diagonal. It is straightforward to show that $\tilde{V}$ also satisfies (27), being another fixed point of ΔP. If now we consider this equation evaluated in $\tilde{V}$ and right multiply by $\tilde{V}$, we get after rearranging terms

$$\alpha(I+\tilde{V}\Phi\tilde{V}^T)^{-1}=\alpha I-(\alpha+1)\cdot(I+\tilde{V}\tilde{V}^T)^{-1}\cdot\tilde{V}\tilde{V}^T. \quad (28)$$

Since $\tilde{V}\tilde{V}^T\equiv D_1$ is diagonal, $\alpha>0$, and Φ is definite–positive, then from this expression we get that $\tilde{V}\Phi\tilde{V}^T\equiv D_2$ must also be diagonal. Then (27) evaluated in $\tilde{V}$ is equivalent to $\Phi\vec{\tilde{v}}_i=\gamma_i\vec{\tilde{v}}_i$ where $\vec{\tilde{v}}_i$ is the ith row of $\tilde{V}$ $(i=1\ldots n_y)$ and $\gamma_i\equiv(1+1/\alpha)[(I+D_1)^{-1}(I+D_2)]_{ii}$. That is, the rows of the fixed point $\tilde{V}^T$ $(\vec{\tilde{v}}_i)$ are either eigenvectors of Φ or null vectors. Note that since $\tilde{V}\tilde{V}^T$ is diagonal, then $\vec{\tilde{v}}_i^T\vec{\tilde{v}}_j=0$ for any pair $i\neq j$. Therefore, the set of non null rows of $\tilde{V}$ forms an orthogonal set of vectors. This eliminates the possibility of repeated eigenvectors in the rows.

Since both $\tilde{V}\tilde{V}^T$ and $\tilde{V}\Phi\tilde{V}^T$ are diagonal, the contribution of $\tilde{V}$ to the functional (second term in (9)) can be easily rewritten as

$$\frac{1}{2}\sum_{i=1}^{n_y}\ln\frac{(1+t_i\lambda_i)^\alpha}{(1+t_i)^{\alpha+1}} \quad (29)$$

where t_i is the squared norm of row i. In case row i is an eigenvector of Φ, λ_i is defined as its eigenvalue. In other case row i is a null vector and then λ_i is defined as 0. Notice that the functional is invariant to a global change in the sign of a row and to permutations between rows. Up to now we have proved that any fixed point of ΔP is a rotation of a "diagonal" solution $\tilde{V}$ whose rows are formed by either null vectors or different eigenvectors of Φ. Since the global optimal family is composed by fixed points, then the diagonal solution which maximizes (29) will belong to the family. This family is then composed by any rotation of this diagonal solution.

Now we will determine which eigenvectors of Φ form part of this solution and what is their norm. Note that since $\log(1)=0$, the contribution of the null rows of $\tilde{V}$ to the functional (29) is null. Suppose that $\vec{a}$ is an eigenvector of Φ with eigenvalue λ, and squared norm t (therefore $t\geq 0$). If it were included in the optimal configuration, it would contribute to the functional with

$$\frac{1}{2}\ln\frac{(1+t\lambda)^\alpha}{(1+t)^{\alpha+1}} \quad (30)$$

and the squared norm t should maximize this contribution. It is straightforward to calculate the optimal t, obtaining two different cases:

1) $\alpha-(\alpha+1)/\lambda\leq 0$: the optimal value for t is 0 and therefore $\vec{a}$ is null. Thus, this eigenvector cannot exist in the optimal solution.
2) $\alpha-(\alpha+1)/\lambda)0$: in case this eigenvector exists in the optimal configuration, it has squared norm t equal to $\alpha-(\alpha+1)/\lambda$ and its contribution to the functional is greater than zero.

It is straightforward to show that the eigenvectors of Φ and C are the same, and the eigenvalues of Φ are

$$\lambda_i=\frac{(a_i+b)(1+a_i)}{a_i+b+a_ib} \quad (31)$$

where $a_i\equiv\sigma_{Ci}^2/\sigma_x^2$, $b\equiv\Delta_g^2/\Delta_x^2$, and $\sigma_{c_i}^2$ are the eigenvalues of C. This together with $a_i>0$ leads us to express the optimal t as:

1) $a_i\leq(1+b+\sqrt{(1+b)^2+4\alpha b}/2\alpha)$: the optimal value for t is 0 and therefore $\vec{a}$ is null. Thus, this eigenvector cannot exist in the optimal solution.
2) $a_i>(1+b+\sqrt{(1+b)^2+4\alpha b}/2\alpha)$: in case this eigenvector exists in the optimal configuration, it has squared norm t equal to $[\alpha a_i^2-(a_i+b+a_ib)]/[(a_i+1)(a_i+b)]$ and its contribution to the functional is greater than zero.

Considering the change of variables described previously $(V=N_y^{-1/2}A(N_x+N_xD)^{1/2})$, the square norm of the corresponding row in A will be $(\sigma_y^2/\sigma_x^2)\cdot[\alpha a_i^2-(1_i+b+a_ib)]/[(a_i+1)(a_i+b+a_ib)]$. Let us introduce n_c as the number of eigenvectors which satisfy the condition 2. In case $n_c\leq n_y$ (number of total neurons) the optimal diagonal configuration will include all of them, the other neurons being null. In case $n_c>n_y$ only the n_y eigenvectors with greatest contribution to ΔP will be part of the optimal diagonal configuration. Now we will show that these are the eigenvectors with highest eigenvalue a_i.

Let us consider two eigenvectors of Φ, $\vec{a}$ and $\vec{b}$, with eigenvalues λ_a and λ_b respectively. Let us assume that they satisfy $\lambda_a>1+(1/\alpha)$, $\lambda_b>1+(1/\alpha)$ (they are candidates to be in the optimal configuration), and $\lambda_a>\lambda_b$. Since the optimal contribution of a candidate eigenvector occurs at $t=\alpha-(\alpha+1)/\lambda$, using (30) we can calculate its contribution to ΔP as

$$\frac{1}{2}\ln\frac{\alpha^\alpha}{(\alpha+1)^{\alpha+1}}+\frac{1}{2}\ln\frac{\lambda^{\alpha+1}}{(\lambda-1)}. \quad (32)$$

Note that all the terms are well defined since $\alpha-(\alpha+1)/\lambda>0$ implies $\lambda>1$. Finally, let us calculate the derivative of the eigenvector contribution with respect to λ

$$\frac{d}{d\lambda}\left(\frac{1}{2}\ln\frac{\alpha^\alpha}{(\alpha+1)^{\alpha+1}}+\frac{1}{2}\ln\frac{\lambda^{\alpha+1}}{(\lambda-1)}\right)=\frac{\alpha\lambda-\alpha-1}{\lambda(\lambda-1)} \quad (33)$$

which, if $\lambda>1+(1/\alpha)$, is always positive. Therefore, if $\lambda_a>\lambda_b$, the contribution of $\vec{a}$ to ΔP is greater than that of $\vec{b}$. As a conclusion, in case $n_c>n_y$ then the optimal diagonal solution is formed by those eigenvectors of Φ with greatest eigenvalues.

E. Construction of Decision Trees

The output of a tree for a pattern is the terminal node that classifies it. Then $d(Y,G)=H(Y|G)+\alpha H(G|Y)$. Consider

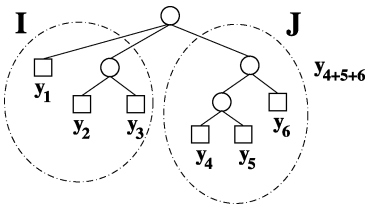

Fig. 12. General schema of a decision subtree. Decision nodes are drawn as circles whereas classification nodes are squares. Each terminal node corresponds to a different state of the system. We call y_{14+5+6} to the node resulting of replacing J by a single terminal node.

the tree in Fig. 12(a) where the subtree I is a child of the root node, and J are the rest of the children of the root. Because a choice can be broken down into several successive choices, the global entropy is the weighted sum of the individual values of H

$$H(Y) = H_r(Y) + p(I)H_I(Y) + (1 - p(I))\, H_J(Y) \quad (34)$$

where $H_I(Y)$ and $H_J(Y)$ are the entropies calculated with the local statistics respectively, and $H_r(Y)$ is the entropy of the tree replacing I and J by leaf nodes (therefore $H_r(Y) = -p(I)\ln p(I) - (1 - p(I)\ln(1 - p(I)))$. Note that $H_r(Y) + p(I)H_I(Y)$ is just the entropy of the same tree replacing J by a single terminal node $(H_{noJ}(Y))$ (Fig. 12).

Thus $H(Y) = H_{noJ}(Y) + p(J)H_J(Y)$, where $p(J)$ is the probability of a pattern to reach J. Analogously, $H(Y|G) = H_{noJ}(Y|G) + p(J)H_J(Y|G)$. Then $d(\vec{Y}, \vec{G})$ can be written as

$$d(\vec{Y}, \vec{G}) = d(\vec{Y}_{no\,J}, G) + p(J)\,(H_J(Y|G) - \alpha I_J(Y;G)) \quad (35)$$

where $d(\vec{Y}_{noJ}, \vec{G})$ is the distance of the tree without the subtree J, and $H_J(Y)$ and $I_J(Y;G)$ are computed using the local statistics in J.

ACKNOWLEDGMENT

The authors would like to thank R. Huerta, L. F. Lago, J. Otterpohl, F. de Borja Rodríguez, and T. Pearce for helpful discussions.

REFERENCES

[1] M. Sánchez-Montañés and F. Corbacho, "Toward a new information processing measure for neural computation," in *Proc Int. Conf. Artificial Neural Networks (ICANN '02)*, vol. 2415, Madrid, Spain, 2002, p. 637.
[2] M. A. Arbib, *The Encyclopedia of Artificial Intelligence*, 2nd ed. New York: Wiley, 1992, vol. 2, ch. Schema Theory, pp. 1427–1443.
[3] F. Corbacho, "Schema-based learning," *Artif. Intell.*, vol. 101, no. 1–2, pp. 370–373, 1998.
[4] C. E. Shannon, "A mathematical theory of communication," *Bell Sys. Tech. J*, vol. 27, pp. 379–423, 1948.
[5] R. A. Fisher, *Statistical Methods and Scientific Inference*, 2nd ed. London, U.K.: Oliver and Boyd, 1959.
[6] T. Cover and J. Thomas, *Elements of Information Theory*. New York: Wiley, 1991.
[7] H. Barlow, "Unsupervised learning," *Neur. Comput.*, vol. 1, pp. 295–311, 1989.
[8] R. Linsker, "Self-organization in a perceptual network," *IEEE Computer*, vol. 21, pp. 105–117, Mar. 1988.
[9] J. Atick, "Could information theory provide an ecological theory of sensory processing?," *Network*, vol. 3, pp. 213–251, 1992.
[10] A. Borst and F. Theunissen, "Information theory and neural coding," *Nat. Neuroscience*, vol. 2, no. 11, pp. 947–957, 1999.
[11] D. MacKay, *Information Theory, Inference and Learning Algorithms*. Cambridge, U.K.: Cambridge Univ. Press, 2003.
[12] D. L. Ruderman, "The statistics of natural images," *Network*, vol. 5, pp. 517–548, 1994.
[13] J. Schmidhuber, "Learning factorial codes by predictability minimization," *Neural Computat.*, vol. 4, no. 6, pp. 863–879, 1992.
[14] C. E. Shannon and W. Weaver, *The Mathematical Theory of Communication*. Urbana, IL: Univ. of Illinois Press, 1949.
[15] L. Prechelt, "PROBEN1: A Set of Neural Network Benchmark Problems and Benchmarking Rules," Fakultat fur Informatik, Univ. Karlsruhe, Germany Internal Report, Max-Planck-Inst. of Biophysical Chemistry, Gottingen, West Germany, Technical Report 21/94, 1994.
[16] M. Sánchez-Montañés, "A Theory of Information Processing for Adaptive Systems: Inspiration From Biology, Formal Analysis and Application to Artificial Systems," Ph.D. dissertation, Univ. Autónoma de Madrid, Madrid, Spain, 2003.
[17] A. Campa, P. D. Giudice, N. Parga, and J.-P. Nadal, "Maximization of mutual information in a linear noisy network: a detailed study," *Network: Computat. Neural Systems*, vol. 6, pp. 449–468, 1995.
[18] E. Oja, "A simplified neuron model as a principal component analyzer," *J. Math. Biology*, vol. 15, pp. 267–273, 1982.
[19] A. Bell and T. Sejnowski, "An information-maximization approach to blind separation and blind de-convolution," *Neural Comput.*, vol. 7, no. 6, pp. 1129–1159, 1995.
[20] S. Amari, A. Cichochi, and H. H. Yang, *A New Learning Algorithm for Blind Signal Separation*, ser. Advances in Neural Information Processing Systems. Cambridge, MA: MIT Press, 1996, vol. 8.
[21] D. Levine, PGAPack Parallel Genetic Algorithm Library, 1998.
[22] A. Pouget, P. Dayan, and R. Zemel, "Information processing with population codes," *Nature Rev. Neuroscience*, vol. 1, no. 2, pp. 125–132, 2000.
[23] J. R. Quinlan, "Induction of decision trees," *Mach. Learn.*, vol. 1, no. 1, pp. 81–106, 1986.
[24] T. M. Mitchell, *Machine Learning*. New York: McGraw-Hill, 1997.
[25] A. Redlich, "Redundancy extraction as a strategy for unsupervised learning," *Neural Computat.*, vol. 5, pp. 289–304, 1993.
[26] J. von Neumann, "Probabilistic logics and the synthesis of reliable organizms from unreliable components," in *Automata Studies*, C. Shannon and C. McCarthy, Eds. Princeton, NJ: Princeton Univ. Press, 1956, pp. 43–98.
[27] N. Tishby, F. Pereira, and W. Bialek, "The information bottleneck method," in *Proc. 37th Annu. Allerton Conf. Communication, Control, and Computing*, IL, 1999, pp. 368–377.
[28] S. Winograd and J. D. Cowan, *Reliable Computation in the Presence of Noise*. Cambridge, MA: MIT Press, 1963.
[29] P. Elías, "Computation in the presence of noise," *IBM J. Res. Develop.*, vol. 2, p. 346, 1958.

Manuel A. Sánchez-Montañés received the B.Sc. degree (with honors) in physics from the Universidad Complutense de Madrid, Spain, in 1997, and the Ph.D. degree (*cum laude*) in computer science from the Universidad Autonoma de Madrid, Spain, in 2003.

He is currently an Assistant Professor in the Computer Science Department, the Universidad Autónoma de Madrid, Spain, and a Scientific Collaborator for the data mining company Cognodata. Madrid, Spain. His main research interest is the search of general principles of organization of both biological and artificial adaptive systems.

Fernando J. Corbacho received the B.Sc. degree (*magna cum laude*) from the University of Minnesota, MN, in 1990, and the M.Sc. and Ph.D. degrees in computer science from the University of Southern California, Los Angeles, CA, in 1993 and 1997, respectively.

He is currently an Ad Honorem Professor in the Computer Science Departmen, Universidad Autónoma de Madrid, Spain and Co-founder and Chief Technology Officer of Cognodata, Madrid, Spain. Cognodata is a firm specialized in the use of data mining and artificial intelligence technics to solve business problems specially in the area of marketing intelligence. He is engaged in the development of a theory of organization for adaptive autonomous agents. His main reserach interests include machine learning, schema-based learning, and the emergence of intelligence. He is a member of several computer and neuroscience associations.