MICHEL VERLEYSEN

CHAPTER 6

THE EXPLANATORY POWER OF ARTIFICIAL NEURAL NETWORKS

1. INTRODUCTION

Many engineering problems include some kind of recognition: from automatic character recognition to the control of steel quality in a steelworks, through the fault detection in nuclear plants or the prediction of financial rates, it is impossible to enumerate all domains where the key challenge is to identify an input-output relationship between variables or concepts. When the physical relationship is difficult to tackle, models are developed to approximate it.

There are many ways to develop such models. Linear ones are used in many cases, even if it known that the linearity limitation will make the model inadequate. Non-linear models are the solution, but they suffer from many limitations, related to the concept of recognition itself: what is the relation to be recognised if it is only known through examples? Artificial neural networks (ANN), i.e. models based on the remote analogy with the information processing in a human brain, try to answer to this question. ANN models are built (trained) on examples, the purpose being to keep the equilibrium between a correct training and a useful (in some cases meaningful) representation.

Despite the fact that ANN are known to be "blind", or non-explanatory, we intend to show that it is possible to feed to or to extract knowledge from these models; the step towards an explanatory power is then straightforward. But the real question is to know to what extend it is possible to interpret the results of such a "non-explanatory" model: what is the real difference between extracting representable knowledge from a computational model, and using a "blind" model to predict, classify or recognise some relationship?

2. BACKGROUND

Information technology is a keyword in our modern world. More than a fashion, computer science, artificial intelligence, and many other scientific breakthroughs transform our today life in a way that was unthinkable twenty years ago. Who was able to predict at the time that a pizza could be ordered through Internet and paid with electronic money, and that the same Internet network would be used to book a

R. Franck (ed.), The Explanatory Power of Models, 127-139.

plane ticket, to read a scientific article or to consult the weather forecasts in a holiday resort?

What electronics and computer science makes feasible is however limited by the inventiveness of engineers, who traditionally build machines able to efficiently achieve repetitive tasks. These tasks have to be described in terms of rules and sequences of operations, or software, and this is probably the main limitation of the state-of-the-art in computer science. In short, what is easily described and analysed is also easily programmed, and thus easily solved by the powerful machines built today. But problems that are more difficult to describe in terms of rules are hard to solve, just because the programming languages and the way how computers work are not adapted to non-rule based, perceptive tasks. For example, multiplying large matrices together, computing the trajectory of a space shuttle or drawing up the balance sheets of a company are tasks which may seem hard because they are computationally intensive, but which are in fact easily "solved" by a computer since they are easily described in terms or rules (mathematical, physical or legal ones respectively). On the other hand, recognising his/her neighbour is an easy task for everybody, but face recognition is a very hard problem for any computer, the reason being that the problem is hard (quite impossible) to describe by rules. Obviously, we don't recognise faces by looking at the hair cut and colour, the eyes colour and shape, the respective location of the neck and the eyes,…; our brain rather analyses a face image as a whole, and gives a decision (for example the name of the person) according to a global perception of this image.

Such comments make the background of the artificial neural networks (ANN) field. Already in the fifties, but more specifically in the early eighties, researchers tried to understand how the (human) brain works, or maybe more realistically how a few neurons can communicate together, exchange information and adapt their functionality to the past experience, in order to replicate their behaviour in computers of a new generation, which would in turn be more adapted to face recognition and other perception tasks such as speech and image recognition, sensory-motor control, fuzzy concepts association, etc.

Traditional artificial intelligence (AI) was not the right answer to this new challenge. Expert systems for example are now considered as tools able to integrate a qualitative dimension in the processing of information (Cottrell, 1995), rather than the miracle solution to "intelligent" problems. ANN may be considered as artificial intelligence too, while ANN techniques are radically different from "traditional" AI ones.

3. MODELLING

The domain of artificial neural networks is large and multidisciplinary, and this has two consequences. On the positive side, knowledge and experience acquired in each field concerned by ANN (from biology to electronics, through mathematics, statistics, control, computer science, etc.) is used to build new theories and concepts, as we will detail below. On the other side, transdisciplinary research is naturally less specific and thus less advanced in a particular field. The interest of ANN research is found in the transdisciplinary character itself, the purpose being to use

ideas from some fields (biology and neuroscience), tools from others (mathematics, statistics and computer science), to build new tools (neuroengineering) that can be used in many application areas (control, recognition, time series prediction, data analysis, etc.). We will limit our discussion to neuroengineering, an area that we will try to define below.

Modelling is a primary goal in many scientific domains. Science itself tries to create and define models whose aim is to be as generic as possible. Modern science assumes that fundamental laws exist and control all physical phenomena. A strong assumption is that these laws are observable, even if we don't know them; Maxwell's equations are a good example. Many electromagnetic phenomena were observed, understood and even modelled through equations before Maxwell; his unified theory has the merit to be compatible with previous observations and laws, but also to be more general and thus simpler in the concepts. One could find many other examples, such as Einstein's relativity, etc.

We generally assume that the laws are observable through experiences, but also that the experiences can be repeated indefinitely; this is an essential basic statistical concept: many experiences, or samples, are needed to build a theory, i.e. to discover a "model" and/or to fit its parameters. We will come back below on the dilemmas between building and fitting models, and also between fitting and generalisation.

It seems that science is built around the concept of models. But obviously, a human brain does not work in the same way. In the context of the face recognition problem mentioned above, nobody tries to create a (mental) model of face characteristics before recognising someone in the street. Our conceptualisation is fuzzier, more ambiguous, but richer too; we do not build "classical" models, but we build something else, much more difficult to describe. As stated above, this difficulty in the task or problem description is reflected into the inadequacy of traditional computers to handle the problem. How to "program" the solution of a problem that is itself difficult to describe? The pioneering ideas of neural networks were an attempt to answer to this deadlock: if a human brain can easily handle problems which seem complex for a traditional computer, building machines that "imitate" the human brain in some way could be the solution.

4. NEUROSCIENCE AND NEUROENGINEERING

To "imitate" the human brain could seem a science-fiction goal. Of course, the intention is not to build a super-computer having human capabilities… On the contrary, the goal of the ANN field is to get ideas from the brain biology and to use them in new computer architectures, in order to build machines that are more adapted to solve perception tasks.

Table 1 compares some pertinent characteristics of a "traditional computer" (based on Von Neumann's architecture) and of a human brain. It should be noticed that the items listed are indicative and qualitative only, and that each of them should merit a detailed discussion…

Table 1: a few indicative characteristics of traditional computers and human brains

traditional computer	human brain
single processor	massive parallelism (neurons and synapses)
high speed (100 MHz)	slow speed (100 Hz)
sequences of instructions (software)	learning (adaptation)
uniqueness of solutions	fuzzy behaviour and many possible solutions
very sensitive to errors	fault-tolerant
deterministic and ordered variables	fuzzy and non-quantified concepts

What makes a human brain so powerful (compared to our PCs…) in tackling perceptive problems, such as reading, speech recognition, sensory-motor coordination, face recognition, etc.? The two main ideas are first the massive parallelism of neurons and synapses contained in the brain, and secondly the adaptation of the huge number of parameters (synaptic coefficients) according to the experience. Based on these two concepts, researchers first tried to model how real neurons and synapses operate (see for example the pioneering work of MacCullogh and Pitts 1943), and then tried to imitate this mode of operation into artificial machines… and models! Most ANN models are far from the biological reality; modelling brain neurons and synapses, and developing computational tools able to perform efficiently in perceptive tasks are different businesses, despite a common inspiration. Werbos (1997) distinguishes between neuroscience and neuroengineering: the first research field aims at understanding how a brain works, the second one at mimicking its potentialities. Figure 1 illustrates this difference: "Neuroengineering tries to develop algorithms and architectures, inspired by what is known about brain functioning, to imitate brain capabilities which are not yet achieved by other means. By demonstrating algorithm capabilities and properties, it may raise issues which feed back to questions or hypotheses from neuroscience" (Werbos 1997).

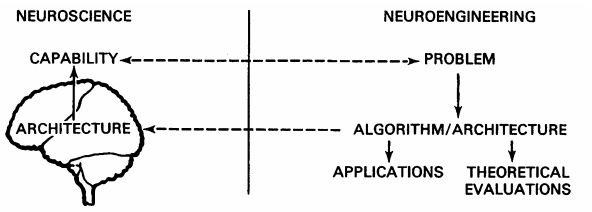


Figure 1: neuroscience and neuroengineering.
From P. Werbos, "What is a neural network?", in E. Fiesler & R. Beale (Eds.), "Handbook of Neural Computation." (p. A2.2:3, fig. A2.2.2), © IOP Publishing, 1996, reprinted with permission.

Neuroscience is thus a discipline where functional models (of the brain or parts of it) are developed, with the view of deducing some material knowledge of the biological system. On the contrary, neuroengineers do not even try to understand the process governing the behaviour of the brain (the process generating the outputs), but rather look to build *one* model, potentially very far from the process, which is *useful*. This is clearly a specific assumption: our interest is to build something that works and which can be *used*, whatever is the way to reach the goal. In that sense, neuroengineering is often far from neuroscience, because most ANNs are far from any biological plausibility...

5. DO WE HAVE TO MODEL REALITY?

In the context of artificial neural networks (neuroengineering), the question is ill-posed. Reality is not an objective: neuroengineering is nothing else than statistical analysis with specific mathematical tools. Statisticians (Mouchart 1998) consider that the starting point of any analysis consists in observations, and not in reality. In any situation, we have a (finite) set of observations, and we assume that these data represent reality. We could for example measure the tide at a specific coast location, each day during ten years, and try to guess (or to "predict") what will be the tide during the next two years. By limiting our observations to one value each day during ten years, we assume that the process that governs tides is entirely described by this finite set of observations. This is obviously not the case, nor it is in most modelling of natural or physical phenomena. However, we might be happy with our tide prediction, depending on its accuracy; it has no sense to expect an infinite precision in the forecast, first because we understand that we will never get it, but secondly because it is not useful too!

The law of gravity (alignment between the moon and the sun, etc.) governs the process of tides. In order to obtain a material or even a functional model of tides, one should therefore use the gravity laws, put into equations the position of the sun and the moon, etc. Nevertheless, fishermen are usually not interested in understanding how tides are created, but rather on predicting (estimating) at what time the tide will be high. Mathematical models, like artificial neural networks, may be used in that purpose; in such situation, the model developed is only a useful model but has no other more fundamental goal.

The fact that the reality is know or not, i.e. that the observations are sufficient to describe the reality or that they aren't, is thus not the right question here. The best that ANNs can do is adequately predict new observations! The reality behind the observations is not needed is such situation. Nevertheless, a further step in neural network analysis is somewhat contradictory to this argument. The dots in Figure 2 represent observations, and the plain lines three different models found by neural network (or other statistical) analysis. The plain line (a) is not acceptable, since it does not fit the observations adequately, while both plain lines (b) and (c) fit the data; however, everybody will agree that the plain line (b) is more acceptable than line (c). Why? Because even if we do not try to model reality, we still make the hypothesis that any model, including the three ones in Figure 2, should be reasonably close from the reality to be useful. It is intuitively obvious that the "expected reality behind the observations" is closer from model (b) than from model

(c); this intuitive concept is related to the smoothness of the model, which should be in accordance with the apparent smoothness of the observations. Choosing model (c) instead of model (b) is known as "overfitting": the expected result of a model is line (b), while mathematical criteria measuring how the observations are fitted will give the preference to line (c) if some precautions are not taken.

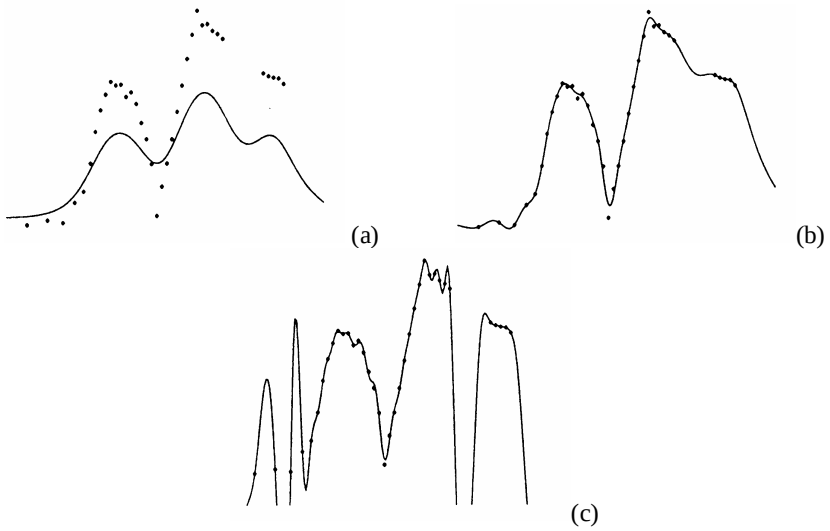


Figure 2: fitting and overfitting data.
From MacKay D., "Bayesian methods for supervised neural networks", in M.A. Arbib (Ed.), "Handbook of Brain Theory and Neural Networks" (p. 145, fig. 1), © MIT Press, 1995, reprinted with permission.

In fact, the usefulness of the model, which is the ultimate goal for engineers, will be higher for model (b), because model (b) will probably have a better *prediction* capability. In other words, it will probably better approximate unknown points, i.e. points that were not used to build the model.

6. WHAT ARE ARTIFICIAL NEURAL NETWORKS?

ANNs are tools invented to model observations, and not reality. This is a strong assumption, or limitation depending on the point-of-view. ANNs are *learning* models: they try to model something (a relation, a function, clusters, a dynamical process, etc.) by building a model which is as general as possible, but which includes a lot of parameters which are then fitted to achieve the expected goal. Some people will call them non-parametric models, since they have the universal approximation property (see below). This is not exactly true in a statistical context,

but the distinction between parametric and non-parametric models becomes senseless at this point.

ANNs are useful tools in data analysis and statistics; in fact, they differ from traditional data analysis or statistic techniques by their implementation, but not by their goals. One of the main characteristics of ANNs versus classical methods is that ANNs are essentially non-linear. They are thus inherently more powerful, since they can perform non-linear *and* linear analysis, where other methods are limited to find linear relationships between data. This advantage is however balanced by an increased complexity, both at the implementation and the computational point-of-views. Before going further in the discussion about ANNs and their explanatory power, we will briefly describe a few ANN models and their possible applications.

7. SUPERVISED NETWORKS

Supervised networks can be viewed as black boxes implementing a relation (a function) which is known through examples. By examples, we mean input-output pairs, as in the example of figure 2 where the X-axis represents the input and the Y-axis the output of a scalar function; in this example, the function is "known" through 37 examples, illustrated by dots. ANNs can of course cope with vector input and outputs instead of scalar ones. The principle of supervised networks is illustrated in figure 3. The neural network implements an input-output relationship, parameterised at random before learning. Learning consists in the following steps:

1. the inputs of an input-output pair are presented to the network;
2. the ANN computes the associated outputs;
3. the outputs computed by the ANN is compared to (subtracted from) the desired outputs, i.e. the outputs of the input-output pair;
4. the result of the comparison is used to slightly modify the network parameters (the "weights") in order to make the ANN better approximate the input-output pair;
5. operations 1 to 4 are repeated for all known input-output pairs (the observations), usually several times for each pair.

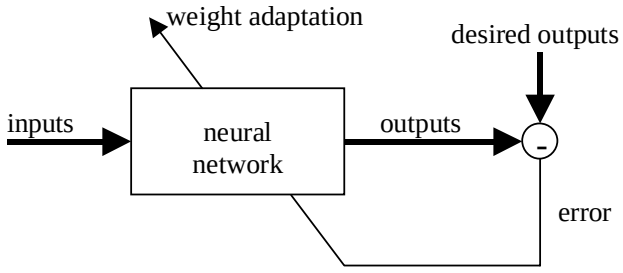


Figure 3: supervised neural network

If the learning is carefully realised, all observations are "learned" by the network. In a subsequent phase (which is called generalisation), new inputs can be presented to the ANN who will calculate corresponding outputs. In our Figure 2 example, the dots are learned while the plain line represents possible generalisation. Overfitting as in figure 2 (c) can be avoided through specific techniques beyond the scope of this chapter.

What makes neural networks different from other approximation techniques is the content of the black box called "neural network" in Figure 3. A widely known ANN is the multi-layer perceptron (MLP), sketched in Figure 4. Circles represent computing units ("neurons"), which implement a non-linear function of the sum their inputs. Each arrow represents a connection between the output of a neuron A and the input of a neuron B, and is associated to a parameter (weight) which is multiplied by the output of neuron A before entering neuron B. Each neuron in a layer is connected to all neurons in the next layer.

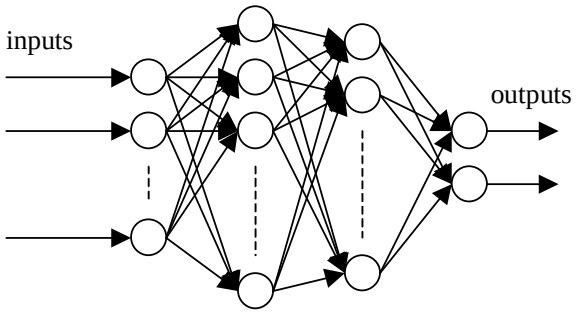


Figure 4: Multi-Layer Perceptron

The "model" is characterised by its size (number of neurons, layers, etc.), but also by the way how is parameters are set (learning rule) according to the input-output pairs.

MLPs have the "universal approximation" property: under weak conditions, MLPs of sufficient size are able to approximate any function from R^n to R^p, with an unlimited precision. This property makes the success of MLPs: in theory, any task formulated as an approximation problem can be solved!

It should be noticed that learning is really a complicated task. Learning in MLP is an optimisation procedure, which can be stuck in local minima, which is not guaranteed to convergence in practical situations. Nevertheless, despite these limitations, efficient learning rules have been proposed in the literature, and the MLP is widely used in many various application areas where some kind of approximation or classification is needed.

There exist many other supervised neural networks, devoted to approximation and classification tasks (radial-basis function networks, learning vector quantization, adaptive resonance theory,…). They slightly differ from MLP, but are used in the same way and for the same applications.

8. UNSUPERVISED NETWORKS

A radically different class of ANNs is the unsupervised network. Figure 5 shows that weights are adapted in unsupervised networks without using any external information about the quality of outputs (without "teacher").

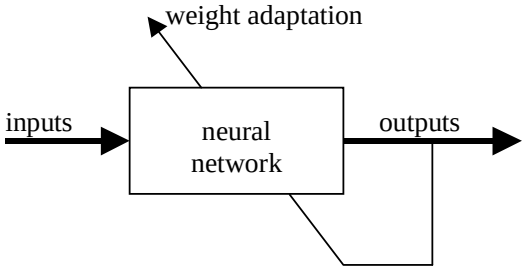


Figure 5: unsupervised neural network

Unsupervised learning is not a well-defined task: usually no criterion is used to evaluate the quality of learning with respect to a consign (or at least the criterions are less intuitive). However, unsupervised networks were found to have computational capabilities that can be used in many applications too.

We refer the reader to specialised literature for details on how unsupervised networks work and are used. In a few words, we can say that a priori information replaces a posteriori one for the learning: steps 3. and 4. of the learning in supervised networks are replaced by an adaptation of the parameters according to some property of input data, instead of a measure of the correctness of the model.

Unsupervised networks are used when data analysis must be performed without knowledge of true or measured output values. Grouping similar data, or partitioning data into small sets, are typical unsupervised tasks. Examples of Kohonen maps applications will be given in the next section.

9. NEURAL NETWORK APPLICATIONS

ANNs being mostly developed by engineers, it is not surprising to find most of the ANN applications in the engineering field. Nevertheless, ANN models can be and are used in all fields where some kind of approximation or analysis has to be performed on data collected from unknown processes. Applications fields include medicine, physical sciences, economics, business, computer science, arts, etc. Below is a list of application examples that can be found in (Fiesler et al., 1997) and (Kohonen 1997):

Supervised learning can be used for:

- intracardiac electrogram recognition in implantable cardioverter defibrillators;

- optimal robot trajectory planning;
- modelling of a polymerisation reactor;
- control of telescope adaptive optics;
- prediction of financial time series.

Unsupervised learning can be used for:

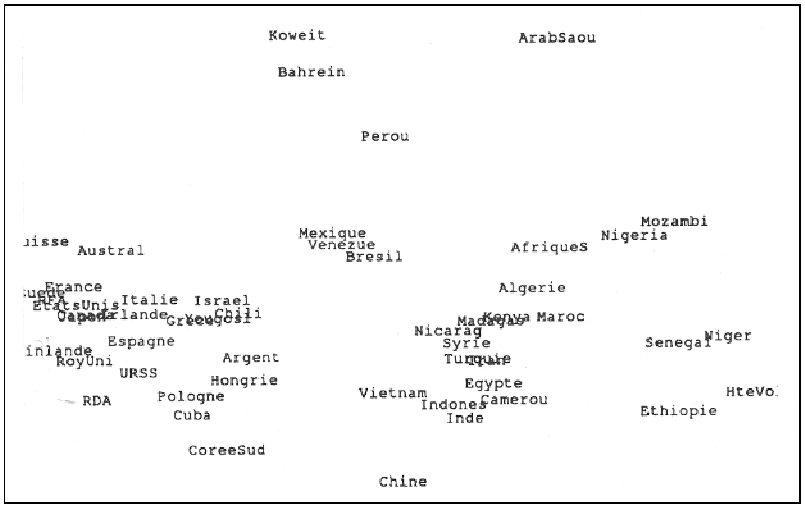
- analysis of socio-economic situations;
- classification of rock samples to determine archaeological origin;
- parsing of linguistic expressions;
- appraisal of land value of shore parcels;
- pitch classification of musical notes.

These applications were not chosen to be representative of all fields where ANN can be used, but to show the diversity of domains which are not restricted to engineering sciences. We will detail one of these applications, the analysis of socio-economic situations (Blayo et al. 1991).

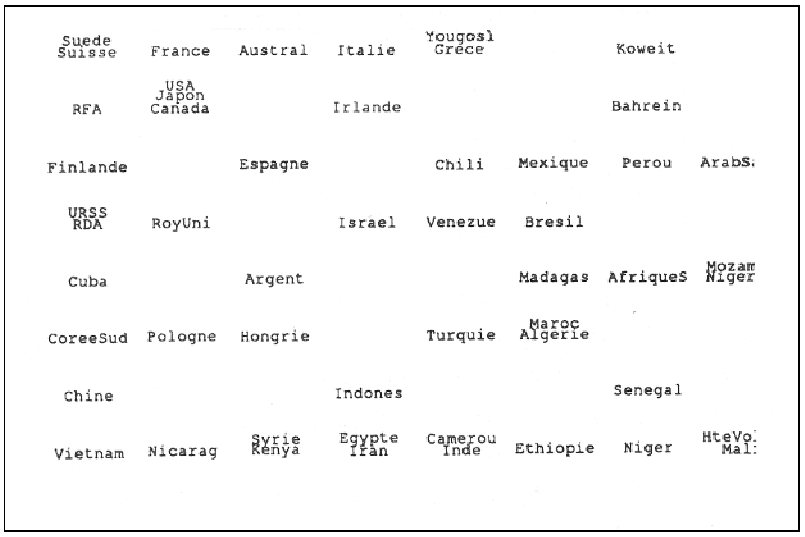
Non-linear dimension reduction is a typical task performed by supervised and unsupervised ANNs. When data are high-dimensional, such reduction can be interesting for two reasons. The first one is simply because low-dimensional data are easier to analyse by hand and to visualise. The second is that the solution of problems in high-dimensional spaces (classification for example) usually requires an exceeding number of data (observations) to reach acceptable performances, while the same level of performances can be reached with much less data in lower-dimensional space (this is a consequence of the empty-space phenomenon, known in data analysis).

Our example consists in analysing the socio-economic situation of 52 countries, according to six variables: the annual increase, the infant mortality, the illiteracy ration, the school attendance, the GIP (gross internal product per inhabitant), and the annual GIP increase.

Viewing or analysing points in a six-dimensional space is quite difficult. For this reason, a standard procedure is to use PCA (Principal Component Analysis) to project the six-dimensional space on a two-dimensional one. Projection means that some similarity criterion should be respected, i.e. that vectors close in the initial space will remain close in the resulting space. This condition is verified for PCA; however, PCA is a linear projection, and is thus able to cancel any linear relationship between variables, but not non-linear ones. Kohonen maps are able to capture non-linear relationships, the result being a better "unfolding" of the six-dimensional data in a two-dimensional plane. Figures 6 shows the six-dimensional database of 52 countries projected on a two-dimensional plane, respectively by the PCA (a) and the Kohonen maps (b) methods. Both have their advantages and drawbacks; nevertheless, it should be noticed that the most frequent use of this kind of projection is to facilitate a subsequent interpretation; Kohonen maps clearly outperform PCA in this context, because of the better unfolding which leads to a better repartition of countries on the map.



(a)



(b)

Figure 6. Socioeconomic situation of 52 countries projected by PCA (figure a) and Kohonen maps (figure b). From F. Blayo, P. Demartines, "Data analysis: How to compare Kohonen neural networks to other techniques?", in A. Prieto (Ed.), "Artificial Neural Networks" (p. 472, figs. 1-2), © Springer-Verlag, Lecture Notes in Computer Science 540, 1991, reprinted with permission.

10. THE EXPLANATORY POWER OF ANNS: COMMENTS AND CONCLUSION

Neural networks have the reputation to behave as black boxes; this can furthermore be understood as a lack of explanatory power. We would like to comment these two arguments, and to point out some explanatory capabilities of ANNs.

ANNs are first designed to model data, i.e. to model empirical evidence. The learning process is based on the empirical values or measurements. Obviously, if the learning in an ANN model is successful, then this model will represent the data (the empirical evidence) successfully. The evaluation of the model is an integral part of the learning. Since learning and evaluation are performed on the same data, the evaluation of the model will be good! (For the sake of completeness, the reader should be aware that statistical procedures to avoid overfitting –such as cross-validation, resampling, etc.– are used when training an ANN; when correctly used, this leads to a successful learning and a good evaluation of the model.)

In the context of ANNs, the right question is not does the model fits the reality, but rather is the model useful?

Reality is not the question here, as in any case the statistical analysis starts from the observations only. The fact that these observations faithfully represent (or not) reality is another (but important) question. ANNs will model the observations, though keeping in mind that the ultimate goal is not to model the observations, but to obtain a *useful* model: the best model is the one which gives the best predictions for new situations that were unknown when the model was built. Intuitively however, when we have to choose between several mathematical models with similar performances on the observations, common sense will lead to the choice of the smoother model, because smoothness is related to reality. Therefore, while the search for a material model of reality is not the question with ANNs, the reality is still taken into account at some stage of the modelling process.

Based on these comments, it is thus an obvious conclusion that ANNs do not provide any insight to reality, i.e. to a "true" model. This goes in the direction of the "black-box" reputation of ANNs. However, the point-of-view of people working in the field of or using ANNs is that this possible true model is irrelevant or simply not useful, at least in the applications they deal with. Even if a theoretical model, close to reality, would exist, it would probably be much too complicated to be used in practical engineering (or other) problems. Remember that ANNs were mostly developed by engineers!

Of course, everybody is not satisfied with such point of view. In many (other) situations (even engineering ones!), one of the purposes of building a model is to learn something about what is *behind* the data. Let us take the example of the human vision system. In the last decades, biologists and engineers developed models of the human vision system on the causal (data) level. During the development, there is no question about functional signification of the model (even if intuition during this development could be guided by some knowledge about the biological structure of the system). But of course the aim of developing this model is to understand how the human vision system works (and not what it does!). What was an aim (and a hope) during the development of an empirical model became a nice result afterwards: many models of the human visual system are valid at the material level (they do resemble –in some way– to the human system). But it must

be kept in mind that this remains an exception, and that most ANNs do not provide reasonable material models.

Moreover, the above conclusions about the non-explanatory characteristics of neural nets must be moderated. For example, most unsupervised networks are much easier to interpret than standard supervised Multi-Layer Perceptrons. Technical details on possible interpretation of the coefficients included in unsupervised models would go beyond the scope of this chapter. Nevertheless the concept can be illustrated by the socio-economic example above. If we look at figure 6 (b), regions in the map corresponding to particular situations (industrialised countries for example) will correspond to network parameters that can be interpreted as a "mean" value (or a typical situation) of these countries. The same applies to other groups, such as Latin-American countries, Eastern block states, etc.

Interpretation of the parameters in a neural network model is a major concern of ANN research. The above discussion pointed out that:

1. There exist some possibilities to interpret the network parameters, at least in specific ANN models.
2. It must be recognised that parameter interpretation is other models, as in MLPs, is difficult.
3. Researchers and users in the field of ANNs do not consider this broken link between empirical models and causal interpretation as a major drawback. On the contrary, in most situations, empirical models are considered as the only valid ones since the problem can only be known through empirical data.

The ANN field is characterised by a wide variety of models in competition. Most of them must be considered today as empirical models, despite the fact that they can help to interpretation of phenomena. But researchers begin to use emerging models as functional ones: the rapidly moving field of ANNs will certainly increase this trend in the near future!

REFERENCES

Abu-Mostafa Y., 1995, "Hints", Neural Computation 7, 639-671.

Blayo F. and Demartines P., 1991, Data analysis: how to compare Kohonen neural networks to other techniques?, in: A. Preito (ed.), Artificial neural networks, Lecture Notes in Computer Science 540, 469-476.

Cottrell M., 1995, Introduction aux réseaux de neurones artificiels et spécificités des méthodes neuronales, in: E. de Bodt, E.-F. Henrion eds., Les réseaux de neurones en finance: conception et applications (D-Facto, Brussels), 23-58.

Fiesler E. and Baele R. eds., 1997, Handbook of neural computation (IOP and University Press).

Kohonen T., 1997, Self-organizing maps (Springer-Verlag, Berlin Heidelberg), 2nd edition.

MacKay D., 1995, Bayesian methods for supervised neural networks, in: M.A. Arbid (ed.), The handbook of brain theory and neural networks (MIT Press, Cambridge MA), 144-149.

McCulloch W.S. and Pitts W.H., 1943, A logical calculus of the ideas immanent in nervous activity, Bull. Math. Biophys. 5, 115-133.

Mouchart M., 1999, Modelling, causality and exogeneity, this book.

Werbos P., 1997, What is a neural network?, in: E. Fiesler, R. Baele eds., Handbook of neural computation (IOP and University Press), A2.2:1-A2.2:4.