Pattern Recognition Letters 15 (1994) 215–217
North-Holland

March 1994

PATREC 1197

IAPR pages

Superlearning and neural network magic

Robert P.W. Duin

Pattern Recognition Group, Faculty of Applied Physics, Delft University of Technology, P.O. Box 5046, 2600 GA Delft, Netherlands

Received 28 January 1993

Over the past five years neural networks have generated substantial interest in the application of complex adaptive systems to various fields of science and technology. It appears easy to raise enthusiasm, to fill numerous conferences each year, to start new journals, and to publish dozens of books and monographs. Researchers in fields such as psychology, physiology, biology, theoretical physics, mathematical statistics, artificial intelligence, computer science and electrical engineering as well as some people in pattern recognition have contributed to the area. It has been applied in many more areas.

Only a small minority of all these scientists is able to demonstrate that they fully understand what is going on. For many it seems to be magic and this is probably part of the attraction. There are also people, however, with serious doubts as to the worth of what is gained by the neural network wave. I believe that there is a generally silent pattern recognition community who have such doubts. To support this statement, I will try to clarify the difference between some neural networks and traditional pattern recognition techniques. Finally, I will argue that there are certainly some things to do and to become less silent as either there is a not well understood phenomenon that should be studied, or neural network learning fits well in the traditional statistical pattern recognition paradigm from which seemingly surprising results can be explained.

E-mail: duin@ph.tn.tudelft.nl

The first wave of interest in neural network-like structures ended around 1970 with the publication of Minsky and Papert's book on perceptrons [4]. The revival of interest in neural networks started somewhere around 1985 and has assumed the characteristics of a fad. After 1987 it was suddenly discovered by many researchers outside the primary field of research as a potentially strong concept with a large applicability. For some period the words 'neural network' guaranteed to open doors to financial support. The words 'artificial neural network' and a superficial explanation of the relation with the human nervous system had a magic attraction to many people. It proved to be very easy to interest students for this, while at the same time a related topic such as Parzen estimators did not raise any enthusiasm when buzz words like neuron and fuzzy were avoided.

Between 1988 and 1990 more than 10 new international journals were started in the field of neural networks. In 1990 there were also at least 10 international conferences organized in this field. This resulted in several thousands of papers per year on neural network research and applications. In the same period, in the closely related field of pattern recognition, there were five papers published in total during the years 1986–1989 in the leading journals (*IEEE-PAMI*, *Pattern Recognition Letters* and *Pattern Rec-*

SSDI 0167-8655(93)E0060-2

ognition) with 'neural' or 'neuron' as a keyword or in the title, followed by seven in 1990 and 12 in 1991. These journals publish more than 300 papers per year in total.

Not only from these numbers, but also from statements made by leading researchers in the pattern recognition field, it can be concluded that many have serious doubts about the usefulness of the topic or are not convinced of its importance. In this context it is remarkable that in the heavily extended second edition of Fukunaga's *Statistical Pattern Recognition* (1990) neural networks are not discussed.

This attitude may have changed in 1992. There was a special issue of *Pattern Recognition Letters* on neural networks. Altogether the three above-mentioned journals published last year more than 35 papers on this subject. During the 11th International Conference on Pattern Recognition in September 1992 there was a highly successful tutorial on the topic and about 45 papers were presented concerning neural networks. Moreover, it appears that the general interest in statistical pattern recognition techniques has grown after a dip of almost 10 years. From these observations, however, it might also be inferred that the area of pattern recognition has attracted many people with a primary interest in neural networks that are looking for a fundamental understanding of the methodology. This makes it even more important that experienced researchers in pattern recognition with either a positive or negative opinion on the usefulness of neural networks contribute openly to the discussion. A journal such as *Pattern Recognition Letters* seems to be an appropriate forum for this through the controversy pages such as these or by short regular contributions. In order to stimulate an open discussion on the understanding and potential usefulness of neural networks for the field of pattern recognition, I would like to raise some issues that I believe to be critical.

What is intuitively wrong with the use of neural networks as learning machines? The answer lies in the fact that they are complex, non-linear systems with (too) many (hundreds, thousands, or more) free parameters. Optimization is, therefore, not only difficult but may yield non-generalizable results for small and moderate sizes of the learning set. From the traditional pattern recognition literature on the error rate as a function of the numbers of learning samples, features, and free parameters, starting with Cover [2], summarized by Jain and Chandrasekaran [3], and with recent contributions by Raudys and Jain [5,6] one may expect a very low apparent (resubstitution) error but a much higher true error for many of the neural network applications. These studies typically show that the sample size should be much larger than the dimensionality or feature size. For non-linear classifiers with more parameters than features, the sample size should be larger than the number of parameters. Baum [1] has shown how multi-layer perceptrons (a special case of neural networks) can perfectly classify (apparent error of zero) an arbitrary learning set if the number of learning samples is less than about half the number of parameters. In such a case generalization is not to be expected. Too few learning samples will result in an almost arbitrary classifier with a high, (that is, unacceptable) true error rate.

In contradiction to the above a priori observation is the vast number of successful neural network applications that have been reported, often with small learning sets and large numbers of parameters. Even if the positive bias inherent in published literature is taken into account (positive results are more likely to be published than negative ones), one has to admit that the reasoning concerning the size of the learning set vis à vis the number of free parameters does not seem to apply in a straightforward manner.

A typical example is the classical study in the neural network field by Sejnowski and Rosenberg [8] in which networks are used with many free parameters, for example, more than 20,000 and trained by some 5000 examples. Many more examples of this *superlearning* phenomenon can be found in the literature in which generalizable training results are found with more free parameters in the network than objects in the learning set. It is regrettable that in almost none of these studies is a thorough comparison made with classical pattern recognition methods. Another aspect of the study by Sejnowski and Rosenberg, and not uncommon for several other neural network applications is that the learning set is not a random selection of the universe of objects. They took the most frequently used words in English. In other applications the learning set is chosen in a systematic way, or selected by an expert.

Here are some possible hypotheses about why

neural networks might work in spite of the small sample size / parameter size ratio:

1. *The architecture.* It may be that the architecture of a neural network is such that not all free adjustable parameters are also independent free parameters. This hypothesis is supported by the observation that in some situations the network can successfully be trained even when large numbers of parameters have randomly chosen fixed values. E.g., see Schmidt et al. [7].

2. *The training rule.* It may be that neural networks are trained such that less than the entire parameter space is implicitly searched (and certainly not explicitly). Thus the effective number of free parameters may be less. This hypothesis is supported by the fact that the popular back-propagation training rule hardly ever ends in the zero error (global optimum) situations corresponding to Baum's construction mentioned above.

3. *The data.* It may be that the data in many real-world applications (or at least those that are reported) are such that they are insensitive for large subsets of the free parameters in a neural network. In other words: maybe the intrinsic dimensionality of the set of neural networks (by varying the parameters) is much larger than the intrinsic dimensionality of the resulting set of different classifiers in relation to the data. This is true if the data itself has a lower intrinsic dimension than the feature size. This hypothesis is supported by the observation that it is not to be expected that real-world object classes vary in tens or hundreds of independent free directions.

I admit that these three hypotheses may be considered as three different attempts to formulate a single, more basic interference between classifier and data.

These are some not yet fully worked out ideas on some aspects of neural networks viewed from the pattern recognition point of view. They may be amplified, modified, proved, or disproved. What is needed is the ability to define precisely in what sense neural networks are good for pattern recognition purposes. This would be most welcome not only because it would lead to a better understanding and application of neural networks, but it would also contribute to our own understanding of the possibilities for machine recognition and learning in general. Besides, people outside the pattern recognition field may benefit from this as they gain some new insights on the possibilities of neural networks.

I hope that those who have some understanding intuitively or well founded by mathematics or experiments will make their ideas known. In the pattern recognition field more comparative studies have to be made. In particular, the more successful neural network applications on real-world data should be compared with the classical pattern recognition techniques in order to find out whether the method or the data has caused the success. In relation to this, I hope that everyone will give free access to the data that he or she has used in published experiments. It could be stored on a public archive or kept readily available for colleagues that ask about it. A scientific observation that cannot be shared by colleagues is of no use.

References

[1] Baum, E.B., (1988). On the capabilities of multilayer perceptrons. *J. Complexity* 4, 193–215.

[2] Cover, T.M. (1965). Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition. *IEEE Trans. Elec. Comp.* 14, 326–334.

[3] Jain, A.K. and B. Chandrasekaran (1987). Dimensionality and sample size considerations in pattern recognition practice. In: P.R. Krishnaiah and L.N. Kanal, Eds., *Handbook of Statistics, Vol. 2.* North Holland, Amsterdam, 835–855.

[4] Minsky, M. and S. Papert (1969). *Perceptrons: An Introduction to Computational Geometry.* MIT Press, Cambridge, MA (1987).

[5] Raudys, S.J. and A.K. Jain (1991). Small sample size effects in statistical pattern recognition: recommendations for practitioners. *IEEE Trans. Pattern Anal. Machine Intell.* 13 (3), 252–264.

[6] Raudys, S.J. and A.K. Jain (1991). Small sample size problems in designing artificial neural networks. In: I.K. Sethi and A.K. Jain, Eds., *Artificial Neural Networks and Statistical Pattern Recognition.* North-Holland, Amsterdam.

[7] Schmidt, W.F., M.A. Kraaijveld and R.P.W. Duin (1992). Feed forward neural networks with random weights. In: *Proceedings 11th IAPR International Conference on Pattern Recognition, Volume II, Conference B: Pattern Recognition Methodology and Systems* (ICPR11, The Hague, The Netherlands, August 30–September 3, 1992). IEEE Computer Soc. Press, Los Alamitos, CA, 1–4.

[8] Sejnowski, T.J. and C.R. Rosenberg (1986). NETtalk: a parallel network that learns to read aloud. The Johns Hopkins University, Electrical Engineering and Computer Science.