NEURAL NETWORKS AND OPEN TEXTURE

Trevor Bench-Capon

Department of Computer Science
The University of Liverpool
Liverpool L69 3BX

ABSTRACT.

In this paper some experiments designed to explore the suitability of using neural nets to tackle problems of open texture in law are described. Three key questions are investigated: can a net classify cases successfully; can an acceptable rationale be uncovered by an examination of the net; and can we derive rules describing the problem from an examination of the net?

1. Introduction

Open texture represents one of the most challenging aspects of constructing an adjudication system in AI and law. A number of approaches have been tried, including: simply requiring the user to resolve open textured questions (Sergot et al 1986); encoding expert judgement (Capper and Susskind 1988); using competing rules (Bench-Capon and Sergot 1988); case based reasoning (Ashley 1990); mathematical analysis (Greenleaf et al 1987); and using deep conceptual models (McCarty 1989). Absolute success has not, however, been attained by any of these techniques, and all depend ultimately on the availability of an expert who can perform the appropriate analysis of the concept in question, and a great deal of skill intensive labour. Currently a fashionable approach - as in many other areas of AI - is based on artificial neural networks ("neural nets") which are intended to obviate the need for much of the manual analysis.

This is not the place to give a full account of neural networks, but a brief description is in order. More detail can be found in any of the many books on neural nets such as Zurada (1992). Basically neural nets consist of nodes grouped into layers. There is an input layer, an output layer, and one or more hidden layers. The nodes in a layer are connected to the nodes in the next layer, as shown in figure 1.

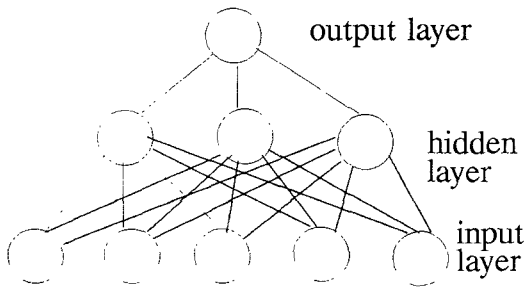


Figure 1

The output from a node is determined by the bias of that node, and by the weight of the link to the next layer. A net is trained by supplying a set of inputs to the input layer, and propagating them to get a value for the node or nodes in the output layer. These are compared with the desired values for these outputs: if they are not correct the weights and biases are adjusted. A net converges when the output is sufficiently close to the desired value for a predefined number of training examples.

The essential idea is that a neural net, when presented with a set of training cases, can be trained to perform the desired classification. Note that this, in its purest form, does not require that anyone understand the domain: all that is necessary is that a body of classified examples be available. This is clearly an attractive idea: if it holds good it removes a thorny problem without needing much skill or effort on the part of the system builder. Of course, it is necessary to feed the correct factors into the neural net. The least skill-dependent strategy is to include every factor that is available. This raises the question of whether the performance degrades if irrelevant "noise" factors are included.

Some critics of neural nets have simply expressed scepticism that neural nets can achieve what must be regarded as a task requiring extensive judgement, but substantial criticisms have centered on the lack of transparency. Neural nets act as a black box, giving oracular pronouncements, but offering no explanation. Since explanations are held to be of particular importance in law - unlike domains such as image and speech recognition where neural nets are well established, in which a correct answer is sufficient justification - this lack of explanation might preclude their use in the field of law.

This criticism has been answered in several ways: one suggests that we can discover the rationale used by the net by an examination of the weights and links developed during the training process (Walker 1992), another uses an analysis of the internals of the net to construct symbolic rules (Sestito and Dillon 1992, Bocherau et al 1991). Alternative forms of legitimation include presenting the training set to the user.

Neural nets are, if the claims that are made for them are correct, of such potential, that it is important that these claims are investigated. I wish to investigate three in particular:

- That neural nets can achieve a high degree of success in the classification of cases in an open textured domain, unguided except by past cases.
- That examination of the internals of the trained network can provide an acceptable rationale for the classifications produced by the network
- That neural nets can provide a useful tool in acquiring symbolic knowledge of an open textured domain.

This paper describes some experiments which cast some light on these claims.

2. The experiments

For the purposes of the experiments, I chose to use an artificial problem. A real domain would not provide the best test since, if *ex hypothesi* the domain is not well understood, there is no way of evaluating the rationale arrived at by the network. Moreover by using an artificial domain, training and test sets possessing specified properties can be readily generated. None the less the problem should exhibit the kinds of feature found in legal domains. The problem selected was based on a fictional welfare benefit paid to pensioners to defray expenses for visiting a spouse in hospital. The conditions were:

1 The person should be of pensionable age (60 for a woman, 65 for a man);

2 The person should have paid contributions in four out of the last five relevant contribution years;

3 The person should be a spouse of the patient;

4 The person should not be absent from the UK;

5 The person should have capital resources not amounting to more than £3,000;

6 If the relative is an in-patient the hospital should be within a certain distance: if an out-patient, beyond that distance.

These conditions represent a range of typical condition types: 3 and 4 are Boolean necessary conditions, 5 is a threshold on a continuous variable representing a necessary condition, and 2 relates five variables, not all of which need be satisfied. 1 and 6 are more interesting since the relevance of a variable depends on the value of another: in 1 sex is relevant only for ages between 60 and 65, and in 6 the effect of the distance variable depends on the Boolean saying whether the patient is an in-patient or an out-patient. Of course, as set up, the example is not an open-textured problem, it only appears as such while we are in ignorance of the conditions. This makes the analysis clearer, but is not crucial: similar experiments were conducted in which the sharp distinctions made by the conditions were blurred, with broadly identical results.

The training sets and test data were generated from a LISP program. The initial training set consisted of 50% satisfying cases, where the outcome of each condition was generated randomly within the range which would satisfy the relevant condition, and other values generated randomly across the full range. The other cases were generated so that an equal number would specifically fail on each of the conditions, other values being generated randomly. To see whether the inclusion of irrelevant factors was a problem, a number of "noise" attributes with no effect on the outcome were also included: in the experiments there were 52 such noise attributes, giving a total of 64 input factors. All training sets comprised 2400 cases. The test set was generated using the same program, but with a different random number seed so that different examples would be produced.

Several neural nets were constructed, using one, two and three hidden layers, and with various shapes. All the networks were fully connected and used back propagation. The Aspirin (Leighton 1991) software was used to implement the networks. In practice, the "shape" of the networks made little difference, and so the networks discussed will be of a conventional triangular shape, with each layer reducing the number of nodes. Thus each of the networks had an input layer of 64 nodes and an output layer on one node, 1 signifying entitlement and 0 non-entitlement. The one hidden layer network had 12 nodes in the hidden layer, the two hidden layer net 24 nodes in the first hidden layer

and 6 in the second, and the three hidden layer net 24 in the first, 10 in the second and 3 in the third.

Can the Net Classify?

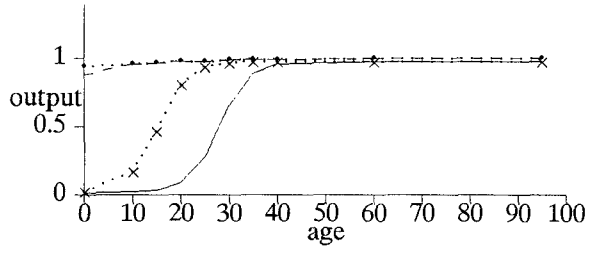
All three of the test networks converged. When run on the test set, they produced uniformly good results. Thus when run on 2000 test cases the following success rates were achieved:

One hidden layer:	99.25%
Two hidden layers:	98.90%
Three hidden layers:	98.75%

This was a very encouraging level of performance, and might be considered acceptable, even in a legal application.

Can we justify the classifications?

Thus the performance of the nets was good. Analysis of the networks, however, proved disappointing. For example one of the things we wished to discover, was what the net made of the pensionable age condition. This the reader should recall depends on age and sex. The effect of these two factors was discovered by constructing a set of test cases in which all the conditions except pensionable age were satisfied, and in which the age varied from 0 to 100 in steps of five, for both men and women. We can now plot the effect of age, for men and for women, on entitlement. The graph in figure 2 results.



•....• 1 Hidden Layer
- - - - 2 Hidden Layers
------- 3 Hidden Layers - men
x....x 3 Hidden Layers - women

Figure 2

What we would like is a plot which gives 0 as the entilement for men under 65 and women under 60, and 1 for those over these ages. In fact, as we can see, that in the case of the one and two hidden layer networks the age condition is simply assumed to be satisfied. Sex makes no difference, and even for very low ages entitlement is close to 1. Only in the case of the net with three hidden layers are age and sex given any attention, and then the assessment of significance is highly imprecise. The condition is considered satisfied at far too low an age, and sex is considered to make 15 years difference rather than the actual 5. Moreover, it is this net which exhibits the least successful performance.

The graph for the combination of distance and inpatient status is also a straight line close to 1 in the case of all three networks: effectively this condition is always assumed to be satisfied. What this indicates is that the level of performance is achieved without any proper consideration of two of the six conditions: these conditions are simply guessed to be satisfied.

How can this degree of performance be achieved without considering two important conditions? The answer is given by seeing that if one condition is not satisfied, the chances are strongly in favour of some other condition being unsatisfied also. Consider 1200 cases which fail to satisfy the conditions. Suppose that only the spouse condition were known, and all other cases were guessed satisfied. The net would correctly solve the 200 cases specifically designed to fail the spouse condition - and half the rest, because they incidentally happen to fail that condition. Thus knowledge of a single condition would serve to correctly classify 700 cases, or 58.3%. Knowledge of two conditions would serve to correctly classify the 400 cases specifically directed at those two conditions, plus half the remainder, plus half the resulting remainder - a total of 1000 cases, or 83.3%. Knowledge of 3 conditions serves to classify 600 + 300 + 150 + 75 cases, 1125 or 93.7%. Knowledge of four conditions classifies 800 + 200 + 100 + 50 + 25 cases, 1175 or 97.91%. In addition all satisfying cases will be classified correctly. Thus knowledge of 4 conditions would be expected to solve all but 25 out of 2400 cases in the test data described - 98.95% - very close indeed to the figures in fact achieved. Running the nets on a set of test data designed to fail only one of the conditions supported the analysis by producing the following rather poorer results:

One hidden layer:	72.25%
Two hidden layers:	76.67%
Three hidden layers:	74.33%

which goes some way to showing how important it is that all factors are given their due consideration.

This, of course, proves little. It relies on assumptions about the distribution of unsatisfied conditions in the training and test cases, and perhaps also depends on the precise nature of the net. What it does show, however, is that an apparently acceptable level of performance can be achieved by the networks without any identification of some of the significant features of the problem. That these features have been missed, however, can only be detected if we have some prior knowledge of the domain which allows us to say this: otherwise we would have no way of telling that the

four conditions that were discovered were not in fact the whole story. Thus the rationale for the classification is inadequate.

If we make different assumptions about the data we get a somewhat different picture. The nets were retrained on a training set comprising 1200 cases satisfying the conditions, and 1200 failing cases, comprising equal numbers failing each on the conditions, but failing *only* on that condition.

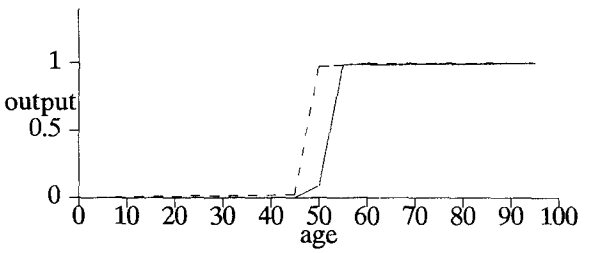
In this case only the nets with one and two layers converged, the net with three hidden layers never attaining a performance of much above 80%. The other two, run against the same test data as the first set achieved the following results.

One hidden layer: 99.25%
Two hidden layers 99.00%

On the set of test data in which all failing cases failed on only a single condition the results were:

One hidden layer: 97.91%
Two hidden layers: 98.08%

Although the performance on the original test data is not significantly better than the nets trained on the first training set. the internal examination of the net was much more encouraging. The graph of outcome as age varies for this net, constructed in the same way as that for the previous net shown in figure 2, is shown in Figure 3:

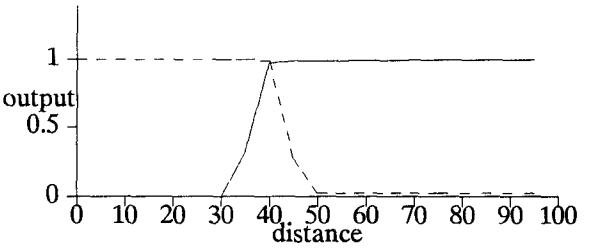


------- Men
- - - - Women

Figure 3

This is still a little generous, in terms of age, women satisfying the condition at 45 and men at 50, fifteen years early in both cases. In this case, however, the five year difference made by sex is correctly identified, and the gradient is much closer to the desired vertical. This then is a considerable improvement. Encouragement can also be gained from considering the condition relating distance and in-patient status. Recall that as originally trained the net could attribute no significance to the conditions. If, however, on the retrained net we vary distance from 0 to 100, for both in-patients and out-patients, with the other five conditions satisfied, we get the graph shown in Figure 4.

This is again a little generous, in that for 40 miles



------- out-patients
- - - - in-patients

Figure 4

distance both in-patients and out-patients are deemed to satisfy the condition, whereas the condition requires that there be a change in both cases at 50 miles distance. The picture is, however, acceptably close to the truth, and almost exact in the case of in-patients.

The nets trained on cases failing only a single condition, are thus using a much more acceptable, if still imperfect, rationale. That the rationale is acceptable can again, however, be seen only from a standpoint of knowing what the rationale should be. Examination of the raw weights produced by the net fails to suggest, for example, that sex is more important than some of the noise variables. It is only by knowing that it should influence age, and so examining its influence on the contribution of this factor that its subtle effect within the net can be determined.

What this experiment with the second training set shows is that neural nets can produce an acceptable rationale - provided they have the right training examples to work on. Preparation of such a training set depends critically, however, on some prior knowledge of the conditions which enables us to produce sets of cases with the right properties. In a situation where the conditions were unknown, we would not know whether our training cases exhibited the properties of the first training set or the second, and so would not know whether the rationale was acceptable or not. Significantly the actual performances of the nets are not significantly different when tested against data with similar properties to that on which they were trained, and so success is no guide as to the correctness of the rationale.

We could now pose the question as to whether a real set of case data would be more likely to resemble the first training set or the second. It could be argued that claims for a benefit are likely to be made only by those with some chance of qualifying. We would thus expect cases which fail to do so on one or two conditions rather than all six, and to fail with the factors quite close to the required threshold. This would suggest that the real data would tend to resemble the second training set more closely than the first. The argument

is not, however, conclusive, and requires some understanding of the conditions on the part of claimants. Again this is not an implausible assumption, but one which we need to be wary of before placing too much confidence in the rationale produced by a net in a given case where we have no prior understanding of the domain.

Can we derive rules from the nets?

In order to derive rules from the nets, it must be possible to determine both which factors are significant, and what their significance is. Suppose we invert the network so that the desired outcome becomes the input and the input factors the output from the net, and train the net of data expressed in this form. If the net told us nothing we would expect all the outputs to be approximately 0.5, and most of them are. There are, however, some significant deviations. If we list the outputs in order of their deviation from 0.5 we get the following top twelve attributes:

Spouse?	0.995
Absent?	0.016
Contribution 5	0.920
Capital	0.118
Contribution 1	0.875
Contribution 4	0.819
Contribution 2	0.809
Contribution 3	0.797
Age	0.779
Noise 8	0.776
Noise 16	0.720
Sex	0.354

Again this is relatively encouraging. The two Boolean conditions show up very clearly, the five contribution inputs (four of which you recall must be true) are shown to have a positive influence; capital a strong negative influence, and age a somewhat weaker positive influence. Sex also appears as a factor with some negative influence (male was represented as 1 and female as 0), although the effect is somewhat spoiled by the intrusion of two noise attributes, which suggests some degree of spurious correlation. The two significant factors which do not appear, distance and inpatient status, could not be expected to be detected since they can take any value at all, provided that they are in the right combination. The values given by the net are in fact 0.56 and 0.409 respectively. With the exception of this, admittedly somewhat peculiar condition, it seems that we can indeed identify the most significant attributes.

Determining the nature of their significance is, however, a different matter. The Boolean conditions present no problems, nor does capital, since we can graph the effect of varying this input and obtain something close to the desired threshold. Where, however, it is *combinations* of attributes that have significance, it is far less clear as to how we are to determine the attributes to combine, or the manner in which to combine them. Other techniques, such as those described in Sestito and Dillon 1991 also failed to make the illuminating connections. For this reason we should remain sceptical about the potential for deriving rules from attributes, unless we have some alternative domain analysis to guide us. If we have hypotheses, the evidence from the neural net may be capable of use to evaluate and refine them, but it is not necessarily so useful in forming the required hypotheses.

Discussion

Neural networks have had their most significant successes in areas involving the interpretation of sense data: character recognition, speech recognition, evaluation of sonar data (Gorman and Sejnowski 1988) and the like. In these cases the inputs are *homogeneous* - all of a like kind - *independent* - the significance of an input does not depend on the value taken by some other input - and contribute to the problem in a broadly similar fashion. Law is a rather different domain: it is the product of rational reflection, rather than "naturally occurring". In consequence the input factors tend to be heterogeneous, interact with each other in a variety of ways, so that a factor may be significant only for certain values for other factors, as in the case of sex in the example, or even have the nature of their influence altered, as in the case of distance in the example. Finally they tend to contribute to the problem in a variety of different ways. These are significant differences and we should be wary about trying to replicate the success of neural nets in sense orientated domains in the domain of law. Moreover in the domains where neural nets have proved themselves the rationale is interesting, but not critical: in contrast in the legal domain the rationale is of great importance. It is not enough to perform well, the performance must also be justified.

Conclusion

In this short paper I have reported some experiments with neural nets. I believe the following to have been shown:

- Neural nets are capable of producing a high degree of success in classifying cases in domains where the factors involved in the classification are unknown
- The inclusion of irrelevant factors has little impact, either on the performance or on the rationale produced.
- A rationale can be derived by examining the net: however the level of performance is no

proof that the rationale is sound.

- It is not the case that the net with the best performance has the soundest rationale: therefore the rationale can be evaluated only with respect to independent knowledge of the domain.
- Rules expressing the rationale are hard to discover where they involve non-straightforward combinations of factors
- Analysis of the net may (but may not) help in evaluating hypotheses concerning the rationale for decisions in the domain.

Thus we can see that neural nets may give good results as a treatment of open texture, particularly in terms of performance. The quality of the rationale will, however, depend, on the nature of training set. The greater the proportion of boundary cases, the more likely it is that conditions falling on this boundary will be correctly identified. If we have a nascent understanding of the question which we can use to direct our construction of test cases, and evaluation of the rationale, we are more likely we are to be produce an effective net. The use of such understanding, however, goes away from the attractive idea of answering the question without skill and effort, and the process begins to resemble the generate, test and refine methodology associated with conventional rule based systems. The main aim of this paper is to caution against an over sanguine acceptance of neural nets in the legal domain: in particular the possibility of a net which can give high quality performance without basing its classification on all the significant factors needs to be kept in mind at all times.

References

Ashley, K.D. [1990] *Modelling legal argument: Reasoning with cases and hypotheticals.* MIT Press.

Bench-Capon, T.J.M., Sergot, M.J. [1985] *Towards a rule-based representation of open texture in law.* In Computing Power and Legal Language (Walter, C., Ed). Greenwood/Quorum Press, New York, 1988, pp 39-60.

Bocherau, L., Bourcier, D., and Bourgine, P., [1991] *Extracting Legal Knowledge by Means of a Multilayer Neural Network Application to Municipal Jurisprudence* in Proceedings of the Third International Conference on AI and Law, Oxford, 1991. ACM Press, pp 297-306.

Capper, P.N., Susskind, R.E. [1988] *Latent Damage Law - The Expert System.* Butterworths, London, 1988.

Greenleaf, G., Mowbray, A., Tyree, A.L. [1987] *Expert systems in law: The DATALEX Project.* Proc. First International Conference on Artificial Intelligence and Law, Boston, May 1987 (ACM Press), pp 9-17.

Gorman, R.P., and Sejnowski, T.J., [1988], *Analysis of Hidden Units in a Layered Network Trained to Classify Sonar* Targets Neural Networks 1 pp75-89.

Leighton, R.L., [1991], *The Aspirin/MIGRAINES Software Tools User Manual, Release V5.* The Mitre Corporation.

McCarty, L.T. [1989] *A Language for Legal Discourse I. Basic features.* Proc. Second International Conference on Artificial Intelligence and Law, Vancouver, (ACM Press), pp 180-189.

Sergot, M.J., Sadri, F., Kowalski, R.A., Kriwaczek, F., Hammond, P., Cory, H.T. [1986a] *The British Nationality Act as a logic program.* Communications of the ACM 29, 5 (May 1986), pp 370-386.

Sestito, S., and Dillon, T., [1992] *Automated Knowledge Acquisition of Rules with Continuously Valued Attributes,* Proceedings of the 12th International Conference on Expert Systems, Avignon, 1992. EC2, pp645-56.

Walker, R., [1992] *An Expert Systems Architecture for Heterogeneous Domains,* Doctoral thesis, Vrige Univeriteit of Amsterdam.

Zurada, J.M., [1992] *Introduction to Artificial Neural Systems,* West Publishing Company.