Connectionist, rule-based and Bayesian decision aids: an empirical comparison

20

S. Schwartz, J. Wiles, I. Gough and S. Phillips

INTRODUCTION

Diagnosis is one of the doctor's central tasks. As an intellectual activity, it is a variant of the more general skill of classification—assigning entities to different classes or categories. Classification is easy when each category (or, in the present case, each disease) has a specific, reliably detected, sign. Unfortunately, such 'pathognomonic' signs are rare. The common signs of illness (such as fever or pain) are shared by many different diseases and most laboratory test results have more than one possible cause. This non-specific relationship between signs and diseases ensures that there will always be an element of uncertainty in medical diagnosis. This uncertainty can be reduced by the discovery of more sensitive clinical signs, but it can never be eliminated. No test is perfectly accurate, signs can be misleading, and even the best treatments do not always succeed (see Schwartz and Griffin, 1986, for more on the probabilistic nature of medical decision-making).

Complicating matters even further is the frequent lack of any specific causal theory relating diagnostic signs to underlying pathophysiology. Consider, for example, patients who present at hospital casualty rooms complaining of acute abdominal pain. There are many possible causes: appendicitis, perforated ulcer, urinary tract infection, and so on. Doctors use a pattern of signs to discriminate among these conditions. For example, it is most common for appendicitis to occur in males, to begin with a central pain that moves to the right lower abdominal quadrant and to be accompanied by vomiting, loss of appetite, and so on. But this is not always the case. Females also get appendicitis and sometimes it begins with pain in the lower right quadrant. Some patients mimic the complete appendicitis pattern but turn out to have some other illness. Because there is no clear physiological theory to explain why appendicitis should produce a particular pattern of signs, we are left with purely empirical correlations—correlations with values distinctly less than 1.0 (de Dombal, 1984).

Given the uncertain relationship between signs and illnesses, it is often difficult for doctors to decide whether a patient's abdominal pain requires an immediate operation or whether it is safe to merely 'watch and wait'. There are risks either way. An infected appendix allowed to fester may perforate, creating a potentially lethal peritonitis. This situation can be avoided by an early operation. On the other hand, removing a healthy appendix needlessly exposes the patient to the risks of surgery. Because the risks of perforation are greater than those of an operation, all surgeons support the philosophy of 'when it doubt, cut it out'. An unavoidable result of this decision rule is that 25% or more of all appendix operations result in the removal of perfectly healthy organs (Adams *et al.*, 1986). Clearly, there would be considerable savings in money, time and surgical morbidity, if these unnecessary operations could be avoided (and if diseased appendixes could be removed as quickly as possible). It has often been suggested that computerized decision aids might be able to help doctors make better diagnoses. However, obtaining the necessary expertise, in a form suitable for computer coding, is often a major problem. System designers consult medical experts but, in many cases, these experts have considerable difficulty explaining how they go about diagnosing patients. In recent years, it has been suggested that this 'knowledge elicitation bottleneck' can be broken by using the machine learning techniques developed by cognitive scientists (Gallant, 1988; Schwartz, 1989; Schwartz *et al.*, 1989).

CONNECTIONIST NETWORKS

One increasingly popular approach uses connectionist networks to produce diagnostic advisors (Bounds *et al.*, 1988; Hart and Wyatt, 1989; Gallant, 1988). The argument in favour of connectionist networks derives from extensive research showing that much human expertise resides in complex pattern recognition. Chess grand masters, for example, excel because they have a large memory store of game patterns (de Groot, 1965). Similarly, in the medical domain, expert radiologists appear to differ from newly trained doctors mainly in their ability to recognize abnormal patterns quickly (Hillard *et al.*, 1985). Because connectionist networks can learn to recognize ill-defined patterns, they should—in principle at least—be able to learn to make difficult diagnoses even in the absence of a causal theory relating signs to diseases.

It should be kept in mind that the term 'connectionist' is a generic one that refers to many different types of network. Thus, although several attempts to use connectionist networks for diagnostic tasks have been published (Bounds *et al.*, 1988; Hart and Wyatt, 1989; Gallant, 1988), no two have used exactly the same structure or learning rules. Not surprisingly, therefore, the results have been equivocal. In general, studies using simple networks and artificial laboratory data find connectionist networks to work well (Gallant, 1988; Hunt, 1989), while those using actual patient data have found them to perform rather poorly (Hart and Wyatt, 1989). The reason for this difference is not entirely clear. It may result from differences in the specification of the various networks, or, possibly, in the relative completeness of the respective training sets. (Real patient data are never as complete, or 'clean', as artificial data.) In the present research, we used a feedforward network with one hidden layer which learned to classify patients using back-propagation (Rumelhart and McClelland, 1986).

RULE INDUCTION

Connectionist networks are not the only way that cognitive scientists have modelled learning from experience. Quinlan's (1983; 1986; 1988) ID3 algorithm tackles classification by breaking down the overall problem into a series of subclassifications. Specifically, ID3 constructs a decision tree. Each tree node represents the influence of the most diagnostic sign at that point in the sequence. At each node, the value of the sign is used to partition the cases into separate groups. The algorithm is then invoked recursively on the separate groups. The process continues until all the cases at a node fall into the same partition. When that happens a 'leaf node' is created and given a unique label.

ID3 is a particularly attractive alternative to a connectionist network because it is 'rule-based'. To convert the decision tree into a collection of conjunctive rules one simply traces each path from the root to a leaf (Quinlan, 1987). The rules generated by ID3 may produce new insights into the relationship between signs and diseases. In contrast, connectionist networks represent knowledge mathematically and are difficult to interpret in terms specific to a problem domain. ID3's 'divide and conquer' strategy is also consistent with the way doctors are actually taught to make diagnoses (Schwartz and Griffin, 1986). Indeed, branching algorithms, in the form of flow charts, are common in medical textbooks (Komaroff, 1982). In the present research, we directly compared ID3 and connectionist networks by applying both to the same set of patient data.

BAYESIAN PROBABILITY REVISION

A third approach to diagnosis, one that has been used extensively in designing decision aids in the domain of acute abdominal pain, is based on the probabilistic relationship between signs and diseases. Specifically, the conditional probabilities of the signs given the various diseases are combined using Bayes's formula to yield the posterior probability of the disease given a specific pattern of signs (see Sox *et al.*, 1988). This approach does not claim to model human learning or cognition. Nevertheless, it was included in our research because of the normative status of Bayesian probability revision and its widespread use in medicine. We applied Bayesian probability revision in a simplistic manner, making the unrealistic, but exceedingly common, assumption that the various signs were independent.

RESEARCH AIMS

This research project was conceived as an attempt to apply cognitive science techniques to a domain in which categorization carries 'life-or-death' implications. Specifically, the aims of the present research were: to compare the effectiveness of a back-propagation network, ID3 and Bayesian probability revision in classifying acute abdominal pain patients using a set of real patient data; to examine the 'practical value' of the three techniques by comparing their performance with the performance of trained doctors; and to ask whether the various techniques produce new insights concerning the relationships between signs and diseases.

METHOD

PROBLEM DOMAIN: DIFFERENTIAL DIAGNOSIS OF ACUTE ABDOMINAL PAIN

The data used in this study were collected prospectively from 276 patients over the age of 12 who presented to the casualty room at the Royal Brisbane Hospital complaining of acute abdominal pain (see Gough, 1988, for details). For our present purpose, the most important aspects of the data were the signs gathered for each patient, the doctor's initial diagnostic impression and the final diagnosis which served as the definitive criterion of accuracy or the 'gold standard'.

Although there were 41 diagnostic signs, each could take on at least two values. For example, there were six values for the sign 'aggravating factors' ('movement', 'coughing', 'food', and so on). We coded each value of a sign as either present or absent. Thus, the total number of inputs available for classification purposes was not 41 but 159. As is often the case with real patient data, some signs were not available for some patients. We assumed that such absent signs were distributed randomly across patients and diagnostic groups. For analytical purposes, we made the arbitrary decision to treat such signs as 'absent'.

As noted earlier, there are many possible causes of acute abdominal pain, but from the point of view of the casualty room doctor there are really only two important categories: either the patient needs an operation or the patient does not need an operation. Because most patients who need an operation are suffering from acute appendicitis, the differential diagnosis often boils down to whether the patient has appendicitis, some other serious illness or, for want of a better term, 'non-specific abdominal pain'. These were the three diagnostic categories used in the present research.

DESIGNATION OF TRAINING AND TESTING SETS

The **holdout** method (Weiss and Kapouleas, 1989) was used to partition the cases into training sets (which included approximately 90% of the cases) and testing sets of about 10% of the cases. To ensure that the training and testing sets were equally difficult, we compared doctors' performances on the two sets. Only those sets which were equally difficult for doctors (approximately the same percentage correct) were retained for further use. The holdout method was used in preference to other methods (bootstrapping, for example) because of the enormous length of time it takes to train a network on a new training set (several days of continuous computing, in some cases). To ensure any differences obtained were reliable, we ran both ID3 and the network many times, altering various parameters (for details, see below).

BACK-PROPAGATION PROCEDURES

The network consisted of three layers (input, hidden, output). Each input unit was completely connected to each hidden unit and each hidden unit was completely connected to each output unit. There were no lateral connections among the units. The input layer consisted of one unit for each value of a sign. For example, sex had two input units: 1 0 encoding a male, and 0 1 encoding female. At first glance it might seem odd to code sex separately for male and female; after all, the two units would be perfectly correlated (− 1.00). The reason for adopting the present approach is that the back-propagation algorithm can only learn on the presence of information and not on its absence. The amount of weight adjustment is proportional to a unit's activation. Encoding sex as a single unit (with, say, 0 standing for male and 1 for female) would mean that, for males, the activation would be 0, no error term could be calculated, and there would be no weight adjustment.[1] To enable fair comparisons, we coded the data in a similar manner for ID3 but not for Bayesian probability revision which, as already stated, was conducted under the assumption that the signs were independent. (Coding for Bayesian probability revision is described below.)

On the output side, we had a similar coding for appendicitis, non-specific abdominal pain and other illness. This method minimized the amount of decoding required and maximized the separation (or dissimilarity) between different inputs. It did mean, however, that there were a great number of weights to be updated. The approach taken was to use only a few diagnostic signs (a small, fast, network) as a starting point. We then added inputs checking performance at each stage. This allowed us to determine whether some subset of inputs produces optimum classification (or whether all diagnostic signs are necessary).

Using the implementation provided by McClelland and Rumelhart (1988), training involved repeated epochs (one forward and one backward pass through the network for all cases). For the purpose of calculating errors, the output with the highest weight was selected as the system's 'conclusion'. Processing continued until the network error (sum of the squares of the difference between the desired output and the actual output for each output unit) was minimized.

Back-propagation is not deterministic. The success of a network depends to a large extent on the starting weights and the number of hidden units. To optimize the network's classification performance, the number of hidden units was progressively varied. Also varied were the network parameters: *lgrain*, *lrate*, and *weights*. *lgrain* was set to either *pattern* (weights were adjusted after each pattern was presented), or *epoch* (errors were accumulated and the weights adjusted after all patterns were presented). *lrate* is the fraction of the error used when updating the weights. Nets with different numbers of hidden units were each run three times, each time with different random starting weights. After finding the best combination of factors and parameters, the data set was repartitioned and back-propagation rerun over the new training set using the new parameters.

ID3 PROCEDURES

Unlike back-propagation, ID3 is a deterministic algorithm. ID3 (C4.5) has various parameters with which to 'fine-tune' performance. Over 80 parameter combinations were tried altogether. The parameters used in the present research included:

1. Confidence Factor (CF). Under certain circumstances, ID3 may 'overfit' the training data. That is, the trees generated are specific to the particular training set and may not generalize well to the test set. Overfitting is usually caused by rules formulated to explain 'noise' in the data. The CF attempts to suppress noise by pruning the tree. To ensure that only noise, not relevant information, was eliminated, ID3 was run over a range of CFs.
2. Gain Criterion versus Gain Ratio Criterion. These are two methods of choosing which feature to place at the root of each decision tree (and subtree). The gain criterion tends to favour signs with large numbers of values, while the gain ratio is either neutral or slightly biased toward signs with few values.
3. Windowing versus No Windowing. If windowing is specified, then a tree is constructed from a subset of the training data called a 'window'. The process, called a **cycle**, of generating a tree from a window is repeated until all items not in the current window have been classified correctly. The size of the initial window and the window increment rate—the maximum number of items that can be added to a window at each cycle—may be specified by the researcher. If windowing is not specified then a single tree is constructed from the entire training set.
4. Subsetting versus No Subsetting. Subsetting partitions the training set into subsets which are examined separately. Subsetting, like the CF, is an attempt to improve generalization by restricting the tree's sensitivity to a specific training set.

BAYESIAN PROCEDURES

As noted earlier, the conditional probability of the signs given each of the three classifications was calculated for the training set cases, making the unrealistic, but nevertheless common, assumption that the signs were conditionally independent—the probability of a sign given a classification is not affected by the presence or absence of other signs. (To make this assumption as reasonable as possible, the data were recoded so that each sign could take on $n - 1$ values.) The test cases were classified into diagnostic groups by assigning each case to the group for which its signs produce the highest posterior probability. We also had available a more reliable set of conditional probabilities collected from more than 6000 cases by the Organisation Mondiale de Gastro-Enterologie (OMGE) and coded in a commercial computer program called MEDICL. For comparison purposes, we also used MEDICL to classify our patients.

RESULTS

BACK-PROPAGATION

Varying the network parameters produced dramatic effects. We also found considerable differences in performance depending on the criterion used to define a 'correct' classification. Table 20.1 summarizes system performance under a number of different conditions, using a relatively lenient criterion of 'correct'. A classification was considered to be correct if the output unit with the maximum activation corresponded

Table 20.1 Effects of varying the number of hidden units, *lgrain* and *lrate* on classification performance. The net had 159 input units

Hidden units	*lgrain*	*lrate*	Epochs	Error (sums of squares)	No. of correct test case classifications (out of 30)
30	*epoch*	0.5	100	246	0
30	*pattern*	0.5	100	246	0
30	*epoch*	0.1	100	194	16
30	*pattern*	0.02	100	13	15
30	*pattern*	0.02	400	10	15
30	*pattern*	0.02	1000	8	15
125	*epoch*	0.01	300	38	19
125	*epoch*	0.5	100	194	16
125	*pattern*	0.5	100	194	16
125	*pattern*	0.01	100	14	17
125	*pattern*	0.01	500	7	18
220	*epoch*	0.5	100	194	16
220	*pattern*	0.5	100	296	0
220	*epoch*	0.02	100	194	16
220	*pattern*	0.02	100	194	16

to the patient's final diagnosis (the present 'gold standard'). Stricter criteria, using various thresholds for counting a classification as correct, significantly reduced performance.

All other things being equal, setting *lgrain* equal to *epoch* produced better performance than *lgrain* set to *pattern* and slow learning rates were better than high rates (although these required many epochs to reach convergence). The optimal number of hidden units was somewhat less than the number of input units, but not much less (125). To a great extent, learning depended on the initial starting weights. 'Good' initial weights allowed for faster and more accurate learning because the network's starting position in the error-weight space was in the neighbourhood of a global optimum (that is, the network was not prevented from eventually moving into a global optimum by an intervening local optimum). Nevertheless, after 2000 epochs, all networks managed to classify 99% of the training set and at least 16 test cases correctly. A conservative conclusion, therefore, is that back-propagation networks will always get at least 16 test cases correct.

Although it might be expected that at least some of the signs were redundant, reducing the number of inputs always resulted in degraded performance. From a practical viewpoint, this meant that back-propagation was not able to suggest any way in which the amount of data collected for each case could be reduced. Table 20.2 summarizes the performance of the most successful network (125 hidden units, *lgrain* = *epoch* and *lrate* = 0.01). As may be seen, after 300 epochs, this network correctly classified 232 of the 246 training cases and 19 of the 30 test cases. After 3000 epochs, the network had reduced its error considerably and was able to classify correctly more than 99% of the training cases. So, there is no doubt about the network's ability to learn. Nevertheless, its best performance on the test cases was 63%. This suggests that there were insufficient training cases for the network to learn all the possible variations. Note, also, that test case accuracy was not well correlated with training case accuracy; the network that performed best on the test cases missed 14 of the training cases. Thus, it appears that there was some tradeoff between learning the specifics of the training set and generalization to the testing set. Given the small numbers of cases in the training sets, the network may have overfitted the training data, thus reducing its ability to generalize to the test set.

It should be noted that the optimal network learned fairly slowly, taking 72.5 hours (real time) to converge, running continuously on the Sun 3/50.

Table 20.2 Performance on the training and test sets at selected stages of learning. (The net had 159 input units, 125 hidden units, *lgrain* = *epoch* and *lrate* = 0.01)

		Number of correctly classified cases	
Epochs	Error (sums of squares)	Training set ($N = 246$)	Testing set ($N = 30$)
100	71.3	204	16
300	37.8	232	19
500	22.0	237	18
700	15.7	240	17
1000	9.8	242	17
1500	5.9	244	17
2000	4.8	244	18
3000	4.4	244	18

RULE INDUCTION

Table 20.3 summarizes the effects of changing various parameters on ID3's classification performance. The values in the table are the number of correctly classified test cases out of 30 averaged over 10 trees and rounded to the nearest whole number. Each cell in Table 20.3 represents ID3's performance using a specific combination of parameters. For example, the first row of the table shows that with a CF of 10% and windowing, ID3 was correct on 16 cases using the gain criterion but only classified 15 cases correctly using the gain ratio criterion. As noted earlier, 80 trees were constructed using different combinations of parameters. Using the gain criterion, and subsetting, most of these trees were able to get 16 of 30 test cases correct.

Looking at Table 20.3 as a whole, it is apparent that, unlike back-propagation, ID3's performance was not greatly affected by changes to its parameters. The only exception was the gain criterion, which was consistently better than the gain ratio criterion—a rather unusual finding (see Quinlan, 1986, for example).

Why should the gain criterion outperform the gain ratio criterion? The answer undoubtedly lies in the specific characteristics of the abdominal pain domain. The gain criterion resulted in smaller and 'shallower' trees than the gain ratio criterion. In addition, signs with many values, such as 'pain onset site' (13 values), and age (8 values),

Table 20.3 Classification performance of ID3 under varying conditions (using all inputs)

	Gain criterion			Gain ratio criterion		
		No window (single tree)			No window (single tree)	
Pruning confidence factor (%)	Windows	No subset	Subset	Windows	No subset	Subset
10	16	16	16	15	15	15
20	15	15	15	14	14	11
30	16	16	15	13	13	11
40	16	16	16	13	11	11
50	15	15	16	12	11	11
60	15	15	16	12	11	11
70	14	15	16	12	11	11
80	14	15	16	13	11	11
90	14	15	16	12	11	11
100	14	15	16	11	11	11

tended to be used first (they were closer to the tree root) when trees were constructed using the gain criterion, whereas signs with relatively few values were closer to the root in trees constructed using the gain ratio criterion. In the present domain, signs with only a few values tend to be less diagnostic than those with many values. For example, 'severity of pain' could take on only two values: 'moderate' or 'severe'. This highly subjective distinction does not differentiate well among classifications. Because the gain ratio criterion favours such signs it may have focused excessively on less diagnostic information.

An important exception to this argument is the two-valued clinical sign, 'rebound tenderness' versus 'no rebound tenderness', which was regarded as the most important sign and chosen as the root of the tree under both the gain and the gain ratio criteria. This ability to provide new insight into the data—in this case, the identification of the most important sign—is one of the benefits ID3 has over back-propagation.

Neither the gain nor the gain ratio are inherently superior (both need to be tested to determine which is best for a particular domain). The same is true of the remaining parameters. For example, windowing (using various initial window sizes) made little difference because the number of training cases from which trees were constructed with windowing (an average of 220) was almost the same as without windowing (246). Also, because tree generation required up to ten cycles, it was actually slower than building a single tree from the total number of items. The finding that windowing is slower than no windowing is not typical (see Quinlan, 1986); it emphasizes the differences among problem domains.

Although the data used here were undoubtedly noisy, there is little evidence that ID3 was seriously overfitting the data. Pruning the tree using the CF and subsetting produced only small improvements in performance. Nevertheless, it is possible that some of the signs were redundant to ID3's diagnostic process. To find out, we examined ID'3 performance with reduced numbers of inputs. In contrast to back-propagation, ID3's performance actually improved when the number of inputs was reduced. ID3's best performance—18 cases correctly classified out of 30—was recorded using only 11 inputs, the gain criterion, subsetting and a CF of 20%. Cross-validation, conducted by repartitioning the set into a new training and a new test set, produced the same results. It would appear that ID3 was able to extract general rules from the training cases, thereby eliminating the need for the specifics of each case.

ID3, at its best, was slightly less accurate than back-propagation (18 versus 19 correct), but the two procedures performed similarly, correctly classifying 16 cases or so. ID3 was rather more robust than back-propagation, getting 16 of 30 correct regardless of the parameters used. It was also of more practical value because it was able to identify the most important signs. (Reducing the amount of data necessary for a diagnosis can save time, money and sometimes lives.) Finally, ID3 had a great advantage in speed. In single-window mode, a complete run required only a few minutes.

CONDITIONAL PROBABILITIES

Conditional probabilities were calculated from the training set for each sign given each classification. Using Bayes's formula, these probabilities were used to calculate the posterior probability of each of the three classifications given the particular pattern of signs presented by a case. Using a lenient criterion, each case was assigned to the category with the highest posterior probability. This method classified correctly 17 of the 30 test cases. The MEDICL program (which uses the conditional probabilities derived from the OMGE survey of 6000 cases) classified 19 of the 30 cases correctly but its results are not strictly comparable to those obtained in the present study because it uses a slightly different set of categories.

The sign with the highest conditional probability for appendicitis was 'pain migrating to the right lower quadrant'. This is related to but not the same as the 'rebound tenderness' sign designated most important by ID3. There was only modest overlap between ID3's most important 11 signs and those signs with high conditional probabilities.

PHYSICIANS

The doctors' initial diagnoses correctly classified 21 of the 30 cases in the test set. Thus, their performance was better than any of the other procedures. This is to be expected given their far greater experience with the domain and the strong possibility that some of the information they gained from their examination of the patients was not coded in the 159 signs.

The doctors' 'hit' rate of 70% of the test set compares well with their average hit rate of 76% of training cases. As noted earlier, this outcome was achieved by design to ensure the test set cases were no more difficult than the training set.

SENSITIVITY AND SPECIFICITY

The raw performance of the various classification algorithms is summarized in Table 20.4. As may be seen, no technique reached the level of the doctors. But, given their greater experience and their probable access to additional information, this is not surprising. What is remarkable is how well the various techniques performed given the relatively small training set and their remarkable similarity; the differences among the various classification procedures were too small to justify statistical analysis. This does not mean, however, that the techniques are interchangeable. We used the contingency coefficient (Siegel, 1956) to determine whether the various techniques were all getting the same test cases right (or wrong). These coefficients were remarkably low. For example, the average correlation between the classifications reached by the Bayesian and ID3 procedures was only 0.35, and no correlation exceeded 0.55. It seems safe to conclude that the three procedures are attacking the classification problem in rather different ways.

Of course, accuracy is only part of the story. As noted earlier, different types of error have different costs. For example, it is less costly to misdiagnose a non-specific pain patient as having appendicitis than to misdiagnose an appendicitis patient as non-specific pain. In the first instance, the patient will have an unnecessary operation. In the second case, the patient may die. Looked at this way, pure accuracy becomes less important than **sensitivity** (true-positive rate) and **specificity** (true-negative rate).

Table 20.4 Summary of the best performances

Diagnostician	Appendicitis	Other serious illness	Non-specific pain	Total correct cases*
Back-propagation	14	2	3	19
ID3	15	0	3	18
Bayesian	12	1	4	17
Doctors†	14	2	5	21
Maximum possible cases	16	5	9	30

*Best performance
†Initial diagnosis only

Table 20.5 Sensitivity, specificity and predictive value of diagnosticians for diagnosing appendicitis

Diagnostician	Sensitivity (true positive rate or $1-\alpha$)	Specificity (true negative rate or $1-\beta$)
Back-propagation	88	36
Bayesian	69	36
ID3	94	21
Doctors	88	50

These are summarized in Table 20.5 for appendicitis versus the other two categories.

As may be seen, all three techniques have a relatively high sensitivity for appendicitis, with ID3 performing best. However, ID3's specificity is low. This means that its success in diagnosing appendicitis is achieved by over-using the diagnosis. Because of their large number of false positives, the predictive values (true positives/all positives) of ID3, back-propagation and Bayesian probability revision are all relatively low. Overall, the doctors probably perform best. They miss only two cases of appendicitis and have the lowest number of false positives. However, this is not a fair comparison because only doctors explicitly take into account the cost of mistakes.

DISCUSSION

The present results are limited by the small number of training and test cases. It has been suggested that a training set should contain at least five cases per item of input data per classification (Wasson *et al.*, 1985). In the present case, this would mean a set of 2400 cases. Few medical data bases are anywhere near this size. We should also note that the Bayesian analysis might have performed better if signs were not considered conditionally independent (Seroussi *et al.*, 1986).

Given the relatively small training set, all of the techniques—but especially back-propagation and ID3—performed remarkably well. Still, there were various differences among the techniques worth noting. First, ID3 was generally more robust than back-propagation. For most runs, it achieved 15 or 16 correct classifications. Its worst performance was 11 out of 30. Back-propagation, on the other hand, was highly susceptible to the learning rate parameter. If not set low enough, the algorithm would not converge at all. Back-propagation was also a much slower process than either ID3 or Bayesian probability revision (it took days as compared with minutes running on the same machine). Finally, the workings of ID3 are more accessible than back-propagation; its trees can easily be stated as rules. It was able to give new insights into the data, emphasizing the importance of a particular sign and identifying the 11 most important signs. Back-propagation, in contrast, appeared to be learning patterns specific to the training set and provided no new insights into the data.

Because the techniques made errors on different cases, they cannot be considered mere substitutes for one another. In the clinic, the best technique would be the one that minimized costly errors. In this regard, none of the techniques was as good as the doctors. But this comparison is unfair because none of the techniques was designed to attend to errors. It is possible that they would have performed as well as the doctors had they been biased away from certain categories. Such biases could easily be added to the network and to the Bayesian analysis (by requiring higher output thresholds for certain diagnoses). ID3 could also be trained to be cautious about certain diagnoses. If they had acceptable accuracy rates, techniques which take into consideration the costs and benefits of various outcomes have the potential to become realistic clinical decision aids. The design of such techniques constitutes an important path for future research.

ACKNOWLEDGEMENTS

Preparation of this chapter was assisted by grants from the Australian Research Council and the NHMRC Public Health and Development Committee to the first named author. The authors are also grateful to ICL for access to the MEDICL program.

NOTE

1. An alternative technique is to use the symmetrical sigmoid function as an activation function. It has a range from -1 to $+1$, so males could be coded -1 and females $+1$.

REFERENCES

Adams, I.D., Chan, M., Clifford, P.C., Cooke, W.M., Dallos, V., de Dombal, F.T., Edwards, M.H., Hancock, D.M., Hewett, D.J., McIntyre, N., Somerville, P.G., Spiegelhalter, D.J., Wellwood,

J. and Wilson, D.H. (1986) Computer aided diagnosis of acute abdominal pain: a multicentre study. *British Medical Journal*, **298**, 800–804.

Bounds, D.G., *et. al.* (1988) A multilayer perceptron network for the diagnosis of low back pain, in *Proceedings of the San Diego Conference on Neural Networks*, Vol. 2, pp. 481–489.

de Dombal, F.T. (1984) Computer based assistance for medical decision making. *Gastroenterology and Clinical Biology*, **8**, 135–137.

de Groot, A.D. (1965) *Thought and Choice in Chess*, The Hague: Mouton.

Dietterich, T.G., Hild, H. and Bakiri, G. (1990) A comparative study of ID3 and backpropagation for English text-to-speech mapping, in *Proceedings of the International Workshop on Machine Learning*, Austin, TX.

Gallant, S.I. (1988) Connectionist expert systems. Communications of the *ACM*, **31**, 152–169.

Gough, I. (1988) A study of diagnostic accuracy in suspected acute appendicitis. *Australia and New Zealand Journal of Surgery*, **58**, 555–589.

Hart, A. and Wyatt, J. (1989) Connectionist models in medicine: an investigation of their potential, in *AIME89: Proceedings of the Second European Conference on Artificial Intelligence in Medicine*, Hunter, J., Cookson, J. and Wyatt, J. (eds), Lecture Notes in Medical Informatics, Heidelberg: Springer-Verlag, 115–124.

Hillard, A., Myles-Worsley, M., Johnston, W. and Baxter, B. (1985) The development of radiological schemata through training and experience: a preliminary communication. *Investigative Radiology*, **18**, 422–425.

Hunt, E.B. (1989) Connectionist and rule-based representations of expert knowledge. *Behavior Research Methods, Instruments and Computers*, **21**, 88–95.

Komaroff, A.L. (1982) Algorithms and the 'art' of medicine. *American Journal of Public Health*, **72**, 10–12.

McLelland, J.L. and Rumelhart, D. (1988) *Explorations in Parallel Distributed Processing*, Cambridge, MA: MIT Press.

Quinlan, J.R. (1983) Learning efficient classification procedures and their application to chess endgames, in *Machine Learning: An Artificial Intelligence Approach*, Vol. 1, Michalski, R.S., Carbonell, J.G. and Mitchell, T.M. (eds), Palo Alto, CA: Morgan Kaufmann, pp. 149–166.

Quinlan, J.R. (1986) Induction of decision trees. *Machine Learning*, **1**, 81–106.

Quinlan, J.R. (1987) Simplifying decision trees. *International Journal of Man–Machine Studies*, **27**, 221–234.

Quinlan, J.R. (1988) An empirical comparison of genetic and decision-tree classifiers, in *Proceedings of the Fifth International Conference on Machine Learning*.

Rumelhart, D.E. and McClelland, J.L. (1986) *Parallel Distributed Processing: Explorations in the microstructure of cognition: Vol. 1: Foundations*, Cambridge, MA: MIT Press.

Schwartz, S. (1989) Computer consultants in the clinic, in *Proceedings of the XXIV International Congress of Psychology: Vol. 9. Clinical and Abnormal Psychology*, Lovibond, P. and Wilson, P. (eds), Amsterdam: North-Holland.

Schwartz, S. and Griffin, T. (1986) *Medical Thinking: The psychology of medical judgment and decision making*, New York: Springer-Verlag.

Schwartz, S., *et al.* (1989) Clinical expert systems versus linear models: Do we really have to choose? *Behavioral Science*, 34, 305–311.

Seroussi, B. and the ARC & AURC Cooperative Group (1986) Computer-aided diagnosis of acute abdominal pain when taking into account interactions. *Methods of Information in Medicine*, **25**, 194–198.

Siegel, S. (1956) *Non-Parametric Statistics*, New York: McGraw-Hill.

Sox, H.C., Blatt, M.A., Higgins, M.C. and Marton, K.I. (1988) *Medical Decision Making*, Boston: Butterworths.

Wasson, J.H., Sox, H.C., Neff, R.K. and Goldman, L. (1985) Clinical prediction rules: applications and methodological standards. *New England Journal of Medicine*, **313**, 793–799.

Weiss, S.M. and Kapouleas, I. (1989) An empirical comparison of pattern recognition, neural nets and machine learning classification methods, in *Proceedings of the 11th International Joint Conference on Artificial Intelligence*, Detroit, Vol. 1, pp. 781–787.