An Appraisal of a Decision Tree approach to Image Classification.*

B. A. Shepherd.

Machine Intelligence Research Unit,
University of Edinburgh, U.K.

ABSTRACT

This paper investigates the applicability to a shape-recognition problem of a concept learning algorithm which generates decision rules from examples. A comprehensive analysis of this algorithm applied to an industrial vision problem is described. This problem has no obvious 'best' solution and much effort has been devoted to performing a realistic appraisal of the algorithm by making a detailed set of comparisons with the performances of appropriate alternative classifiers. Results presented show the algorithm to be comparable in performance with the alternative classifiers but superior in terms of both the cost of making a classification and also the intelligibility of the solution.

L. Introduction,

A nonparametric (or "distribution-free") solution to a pattern recognition problem is one designed only from information contained In representative examples of the pattern classes. This approach is popular In practical problems where detailed statistical information is rarely available. A promising nonparametric solution to classification problems is to produce decision rules which can be expressed in the form of decision trees. Some algorithms for generating such decision trees from examples *have* been suggested (eg Henrichon & Fu 69, Selthi & Sarvaraydu 82) but have concentrated on classification using only numerical measurements of the patterns. A similar limitation arises with 'statistical' classifiers, these are well suited to patterns containing noise (random variation) but poor at dealing with structural information. The algorithm described here has the ability to use both numerical and non-numerical variables. Its origins can be summarised.

An investigation by Hunt (66) into inductive reasoning produced an algorithm called CLS (Concept Learning System) for generating decision trees from examples defined in terms of non-numeric attributes. Each attribute has a predefined set of outcomes which need not represent values along a physical dimension or have any ordering. 103 (Iterative DIchotomiser 3.Quinlan 79) an algorithm based on CLS has been extensively tested on the problem of classifying positions as won or lost in chess endgames (Shapiro & Niblett,82). ID3 has displayed many advantages over alternative solutions. It oan often be used In such a way as to produce solutions with a high degree of human intelligibility and has been extended at MIRU (Blake. 82) to handle both numerical and non-numerical attributes. This extended form is called ACLS (Analogue Concept Learning System). The work described here seeks to evaluate the use of ACLS-generated decision trees in Image classification problems. An industrial vision problem was selected as test domain. This problem has no obvious 'best' solution and much effort has been devoted to performing a realistic appraisal of the ACLS solution by making a comprehensive set of comparisons with the performances of alternative classifiers.

"This research is supported by a grant from the 8ERC for co-operative research with GEC Hirst Research Centre. Wembley. Principal Investigator: Professor Donald Michie.

2 The. vision problem domain-

A sufficiently demanding vision problem has been posed by Rowntree of York concerning the identification of Black Magic chocolates at their factory. In this problem the 12 different types of chocolate have only one stable position but can take any orientation in the horizontal plane. Classification is to be done from a plan view image of individual chocolates digitised to an accuracy of 96x96 pixels. The problem is further defined by making the classification based on shape information (ie. silhouettes) only.

3_ *The* ACLS algorithm,

ACLS inputs examples of the pattern classes. Each example consists of an attribute vector paired with a class value. A decision tree is generated from a collection of examples (training set) by recursively sub-dividing this collection into smaller subsets according to the values of the attributes. The ID3 entropy-based cost measure is applied to each subset to determine which attribute will create the 'best' further sub-division and a node representing the attribute used is added to the tree. The form of the node will depend on whether the attribute is numerical (called Integer) or non-numerical (called logical). Nodes for Integer attributes are binary and test the attribute against a single threshold value. In contrast nodes testing logical attributes have an output branch associated with each value the attribute may take. Sub-dividing the training set continues until each subset contains examples of one class only, and the corresponding terminal node (leaf) is labeled with that class value. If the tree generated is binary then all leaves will be labeled but only with class values for which training examples exist *and* hence It will discriminate only between these particular classes. The multi-branching trees produced using logical attributes may contain many unlabeled leaves (called null). These leaves may correspond to attribute vectors not belonging to any of the pattern classes or they may have arisen only because of the absence of *an* appropriate training example.

4. Objectives of the investigation.

The advantages of a decision rule in tree form can be summarised:

(a) Generating the tree performs an 'attribute selection' function allowing each class to be described by the most appropriate subset of attributes. Thus the set of attributes supplied to the learning phase can be selected without regard to minimising its size: reduction of size can be left to the algorithm itself.

(b) When making a classification only attributes occurring on the decision path need be computed.

(c) Decision trees lend themselves to humanly understandable solutions.

(d) ACLS trees in particular can handle both integer and logical attributes (see notel).

in order to measure these advantages a detailed experimental comparison was made with two alternative classifiers: the K Nearest Neighbour decision rule (K-NN) (Fix & Hodges. 1951.52) and the minimum distance classifier (MDC) (see for eg. Devijver & Kittler. 82). A third alternative means of classification was also undertaken and consisted of showing the chocolate images to a series of human subjects. it was hoped that these latter results would provide a useful measurement to help judge the machine performances.

Notel: A full investigation into the advantages and scope of logical attributes is unrealistic with this domain. However a class of logical attributes which can be investigated is that generated by discretising the integer attributes. If these discrete values can be given conceptual names eg. small, large then the intelligibility of the tree may benefit.

5. Experimental prgcflflurg,

Over 300 boolean images were created, with approximately equal numbers from each chocolate class. For each image the nine numeric shape descriptors summarised in fig. 1 were computed. These attributes were chosen mainly because of their ease to compute. They were not tailored to the problem neither were they expected to totally solve it. The resulting attribute vectors were paired with the correct class values and used as examples. Testing the classifiers was achieved by dividing the total set of examples into various subsets and using each in turn as training and test sets. In order to pursue the use of logical attributes the integer attributes were divided into labeled subranges, and the experiments using the ACLS classifier repeated.

Compactness	area+circle area with same perim
Circularity	area+circle area with same diam
Rectangularity1	area+area of best-fit rectangle
Rectangularity2	area+area of computed rectangle
Aspect ratio 1	length+width
Aspect ratio 2	max. length+min. length
Elongatedness	length+'straightened out' width
norm. area	area+normalising area
norm. perim.	perim. + normalising perim.

Fig. 1 Numerical shape attributes.

6 Results and discussion,

6.1 Classification success rates.

The results obtained using the integer attributes are summarised in fig. 2. In order to both train and test the classifiers the maximum training set size was limited to 216 examples (leaving 108 test examples). All performances improve as the training set size increases to this value. It seems likely that these performances will Improve further (to some steady value) given larger training sets. No major differences among the best performances of the three algorithms are evident from these results which are substantially better than those obtained from human subjects trained on similar material (see later). Results for the smaller training set sizes show the K-NN algorithm to have a higher performance. Note that the K-NN method (K equal to one) and the MDC method are algorithmically Identical when the training set contains only one example of each class.

Av. training set size	12	36	60	108	216
K Nearest Neighbour	56	70	71	75	78
Minimum dist. class.	56	60	68	79	82
ACLS (integer att's)	23	55	60	71	78

Fig. 2: Average success rates

(Note: Each figure is the average of 3 individual estimators. Each of these consists of the ratio of successful to total classifications made for a given training-test example set combination)

Logical attributes were investigated in two ways:

(a) Expressing each integer attribute as a multivalued logical attribute.

This gave inferior results which worsened as N the number of values in each attribute was increased. A larger value of N produced an increase in the number of valid combinations of the attributes which consequently required a greater *number* of training examples in order to be labeled correctly. This is an illustration of a problem often encountered in pattern recognition whenever the dimensionality of the measurement' space Is Increased. In this problem. Increasing N caused the fixed size training sets to become less representative resulting in *an* increasing number of unlabelied (null) leaves appearing In *the* trees generated. The actual decrease in success rate arose because many of these null decisions occurred during the testing phase. The potential advantage of the multi-branching trees to distinguish non-classes from classes is not exercised here since ail test classifications were made using only class examples. Hence all the null classifications which occurred were erroneous.

(b) Expressing each integer attribute as a series of boolean logical attributes:

The idea *here* was to use logical attributes and also generate binary trees. Each outcome of the multi-valued logical attributes was used as a separate boolean attribute taking the values true or false. For example If the attribute 'area' was given the outcomes 'small' and 'large' then the boolean attributes derived In this way would be: 'Is-area-small?' and 'is-area-large?'. Trees generated in this way were were found to *perform* nearly as well as those using integer attributes.

6.2 Cost of classification.

A comparison of costs is shown in fig 3. The memory requirements of both the MDC and ACLS seem very economical. With regard to minimising the computation required ACLS seems far superior to the two alternatives. Only half of the total attribute set is computed for any particular classification and the extra processing required after this is insignificant.

Classif.	memory reqd	att's used	extra processing after attr. calc.
K-NN	tr. set size xvector size =200 x 10 =2000 words +minimal. code	9	dist. calc. for all eg. vectors +finding nearest K vectors =200 sum-squares +200 comparisons
MDC	num. classes xvector size =12 x 10 =120 words +minimal. code	9	dist. calc. for each class prototype +finding nearest =12 sum-squares +12 comparisons
ACLS integer att's	size of rule (in pascal) = 54 instr.	4.5 (av.)	average of 5.5 comparisons

Fig. 3 Comparison of classification costs.

6.3 Intelligibility of the solutions.

Two factors which effect the intelligibility of a decison tree are its size and the intelligibility of the attributes used. The smallest binary tree which can solve this problem must have 11 branch nodes + 12 leaves = 23 nodes, but this would require ideal attributes. Those used in this investigation possess a high degree of variation (see next section) and so the measured average tree size (best performance) of 54 nodes is not excessive. However inspection of the trees shows them to be too large and scattered to be easily comprehended. Replacing each Integer attribute by one many-valued logical attribute can help to conceptually simplify the tree. At each node all possible outcomes of the tested attribute *are* explained and each attribute only occurs once on any decision path. This compares favourably with the binary trees where the same attribute can be scattered intermittently along any path. This option has Its limitations: an excessive number of outcomes in each logical attribute may also result in large and complicated trees. The lower observed performance of the multi-branching trees would seem to suggest a trade-off' between performance and any gain in intelligibility they may provide. However the size of the training set must also be taken Into account here since a larger training set may well raise the performance to an acceptable level (see e.l.cp.

6.4 SUCCESS with the chocolate domain,

The major cause of classification error arose from a large wlthin-class variation In the images which can be split Into manufacturing and image oapture variations An analysis of the attribute values shows a large component of the latter arises from lens distortion effects. Experiments were performed CMowforth.82) to test human subjects on the same Images as used In the above experiments. The average suooess rate measured was only 62% but Increased to 90% when using Images created using a planar lens (no lens distortion). This confirms the lens as a souroe of error and also reflects well on the machine success rates. A performance of 100% is probably unattainable using silhouette images alone. The experimental confusion matrices show a strong confusion between two of the shapes (both squares) and this is confirmed by inspection of the chocolates themselves. A grey level measurement may be necessary to discriminate these particular chocolates.

7. Conclusions and further wprK.

In this domain the decision trees generated by ACLS have performed comparably with those of practical alternative classifiers and all compare favourably with the performance of humans. In addition ACLS trees have displayed advantages regarding both the cost of making a classification and the intelligibility of the solution. Comparisons amongst the various trees have indicated that binary trees may perform better than the logical multi-branching trees, although the latter may have a certain advantage regarding intelligibility. Current work includes increasing the conceptual power of the trees. One approach (Shapiro & Niblett 82) makes large trees more understandable by structuring them into a hierarchy of sub-trees each representing an identifiable sub-problem. The experiments are also being re-run using images with no lens distortion and this together with a small extension to the attribute set is expected to achieve a practically acceptable success rate In this problem domain.

Future work will apply this algorithm to a more structurally complex problem domain.

REFERENCES

[1] Henrichon.E. J & Fu.S.K. . "A Nonparametric Partitioning Procedure for Pattern Classification". IEEE Trns. Comp. vol. C-18 pp 614-624. Jul. 1969

[2] Selthi I.K.& Sarvarayudu Q.P.R. "Hierarchical Classifier Design Using Mutual Information". IEE Trns. Pat. Anal&Mach. Int. . vol PAMI-4, pp 441-445. Jul. 1982

[31 Hunt. E. B. . Marin. J. & Stone. P. "Experiments in Induction". New York. Acad. Press 1966

[41 Quinlan.J.R. "Discovering rules by Induction from large collections of examples". Expert Systems in the Micro Electronic Age.pp 168-201. Ed. D. Michie. Edinburgh University press. 1979

[51 Shapiro. A & Niblett. T. "Automatic Induction of classification rules for a chess endgame". Advances In Computer Chess 3.pp 73-91 Pergamon Press. 1982

[61 Flx.E. & Hodges. J. L. . "Discriminatory analysis. nonparametric classifications". USAF Sch.Aviat. Med. .Rep 4. Feb. 1951.

[71 Devljver.P.A.A Kittler.J. . "Pattern Recognition: A Statistical Approach". Prentice/Hall Int. ,1982

[81 Mowforth P. "A comparison between the abilities of human subjeots and an inductive learning algorithm to categorise chocolates from their silhouettes'. Research Memorandum. MIRU. Univ. Edinburgh. 1982

[9] Peterson. A. & Niblett.T. . "ACLS User Manual". MIRU. ITL. Univ. Edinburgh. 1982.