01 Sep 1992

Effect of the X^2 Test on Construction of ID3 Decision Trees

Mayank Thakore

Daniel C. St. Clair
Missouri University of Science and Technology

EFFECT OF THE X^2 TEST ON CONSTRUCTION OF ID3 DECISION TREES

Mayank Thakore* and D. C. St. Clair

CSC-92-19

Department of Computer Science

University of Missouri-Rolla

Rolla, Missouri 65401

*This report is substantially the M.S. thesis of the first author, completed Fall 1992.

ABSTRACT

Inductive machine learning algorithms are knowledge-based learning algorithms which take training instances as input and produce knowledge as output. One popular induction algorithm is Quinlan's ID3 [1986]. This algorithm produces knowledge in the form of a decision tree. Each path in the tree can be interpreted as a rule with the leaves representing rule conclusions. Selected attributes which describe the training instances form the interior nodes of the tree.

The ID3 algorithm is extremely sensitive to noisy training data. In an effort to reduce the effects of noise on tree construction, Quinlan used the χ^2 test to identify noisy attribute values and exclude them at certain points in tree construction. This approach has proven to be effective in some cases and not effective in others.

This paper examines ID3 trees produced from noisy training data. To determine the effects of the χ^2 test in various situations, several test domains were used. Various levels of noise were injected into each training set and the corresponding trees were evaluated. It was observed that the effectiveness of the χ^2 test on noisy data is related to both the type of matching criteria used at leaf nodes and the size of the training set.

TABLE OF CONTENTS

Page

PUBLICATION THESIS OPTION iii

ABSTRACT iv

ACKNOWLEDGEMENTS v

LIST OF ILLUSTRATIONS vii

LIST OF TABLES viii

1. INTRODUCTION 1

2. ID3 3

3. DECISION TREE ACCURACY 5

4. ID3 TREE CONSTRUCTION IN THE PRESENCE OF NOISE 7

5. NOISE AND THE χ^2 TEST 8

6. EXPERIMENTAL RESULTS 9

7. CONCLUSIONS 18

REFERENCES 19

VITA 20

LIST OF ILLUSTRATIONS

Figure Page

1. Sample ID3 decision tree showing the rule

   $A_3(a_{33}) \Lambda A_1(a_{12}) \Lambda A_4(a_{42}) \Rightarrow C(N_5)$. 4

2. Number of multiple conclusions at leaf nodes when χ^2 test is not applied 14

3. Number of multiple conclusions when χ^2 test is applied. 16

LIST OF TABLES

Table Page

I. Tree Structure 12

II. Accuracy % 13

EFFECT OF THE χ^2 TEST ON CONSTRUCTION OF ID3 DECISION TREES

Mayank Thakore
Sun Microsystems, Inc.
Mountain View, CA 94043.
e-mail: mayank@Sun.COM

Daniel C. St. Clair
University of MO - Rolla
Engineering Education Center in St. Louis
St. Louis, MO 63121
e-mail: stclair@umrgec.eec.umr.edu

1. INTRODUCTION

Many efficient machine-learning systems have been developed which produce knowledge from a limited set of training examples. One of these, ID3 [Quinlan 86], observes training examples, and through inductive steps, builds a decision tree. A path in the decision tree from the root to one of the leaves represents a rule which is true if conditions within its path are satisfied. Rule conclusions are associated with the leaf node.

This concept learning system requires the description of the objects of a domain for inducing implications. The description of each object includes attribute-value pairs and a corresponding classification value. It is highly probable that descriptions of these objects may contain errors. Hence all the data should be treated as noisy. Thus systems learning concepts through examples should be capable of handling errors occurring in the description of the objects.

In ID3, noisy data can lead to an inaccurate decision tree. Two approaches can be viewed for solving this problem. First, these errors can be controlled externally

where a separately codified system checks the validity of each concept description. Inconsistencies are then corrected. This technique for finding errors can be costly and tedious. It does not guarantee that all errors have been found. The second approach to error correction is through an error-handling mechanism which is integrated into the inductive system. The system checks the errors and through some algorithm, makes adjustments. Since the second approach is easier to automate and operationally less complex, it is more feasible. Different solutions relating to the problem have been presented. Among these, one popular approach is the application of the χ^2 test to training data during induction. However other approaches, such as Fisher's Exact Test [Finney et al 1963], are also known.

The work presented in this paper evaluates the effects of the χ^2 test when applied to the development of ID3 decision trees. The effects of using different types of conclusion matching criteria are examined on trees constructed with and without the χ^2 test . Such criteria are required when using ID3 on *real* domains. Evaluation criteria include tree complexity and predictive accuracy. Data chosen for the experiments represent a wide spectrum of noisy environments.

A brief description of the inductive system, ID3, is given in Section 2. To clarify the concept of accuracy, Section 3 shows the detail of different calculations of accuracy in inductive systems. Sections 4 and 5 review the types of noise and the χ^2 test, respectively. Description of the experiments and their results are provided in Section 6.

2. ID3

The ID3 classifier system [Quinlan 86] builds a decision tree from a set of training instances. The description of the input objects consists of the attributes along with their values, plus one or more associated classifications. Attribute and classification values may be taken from discrete, continuous, or symbolic domains. The algorithm is non-incremental since it assumes that the sets of all the training instances are available at the time the decision tree is constructed. Each path in the decision tree represents a classification rule.

To be more explicit, let A denote the set of attributes being used to describe the classification, namely,

$$A = \{A_1, A_2, \ldots A_n\}$$

where A_i is the name of an attribute whose values are :

$$\{a_{i1}, a_{i2}, \ldots a_{i\kappa(i)}\}.$$

The function $\kappa(i)$ denotes the number of the values which attribute A_i can assume. Each training instance utilized by ID3 is composed of an n+1 tuple of the form;

$$(a_{1i}, a_{2j} \ldots a_{np}, c_p)$$

where conclusion $c_p \in C$, the set of all possible classifications.

The ID3 algorithm selects an attribute $A_i \in A$, to be the root of the tree. The selection of the "best" attribute to serve as the root node is done using an information theoretic measure known as entropy [Lewis 1962, Quinlan 1986] . Each branch from the root corresponds to a value, a_{ik}, of the root attribute A_i that was found in the set of training instances. The leaf formed by each of these branches consists of all the training instances whose root attribute value matches the value of the branch. The ID3 algorithm then moves to these leaf nodes and recursively applies the same procedure.

The process is complete when leaf nodes can no longer be split. A leaf node is not split when all classifications at the node are identical or when no more attributes are available to determine splitting. As noted in the next section, the leaf node in the tree may contain multiple conclusions.

Figure 1 represents a decision tree constructed by ID3. Each interior node denotes an attribute and every branch bears a value of its parent attribute. The indicated path in the tree represents the rule;

$A_3(a_{33}) \Lambda A_1(a_{12}) \Lambda A_4(a_{42}) \Rightarrow C(N_5)$

where, $A_i(a_{ij})$ indicates that the value of the attribute A_i is a_{ij}. The expression $C(N_5)$ denotes that all the conclusions at node N_5 are applicable. In this case, $N_5 = \{c_1\}$.

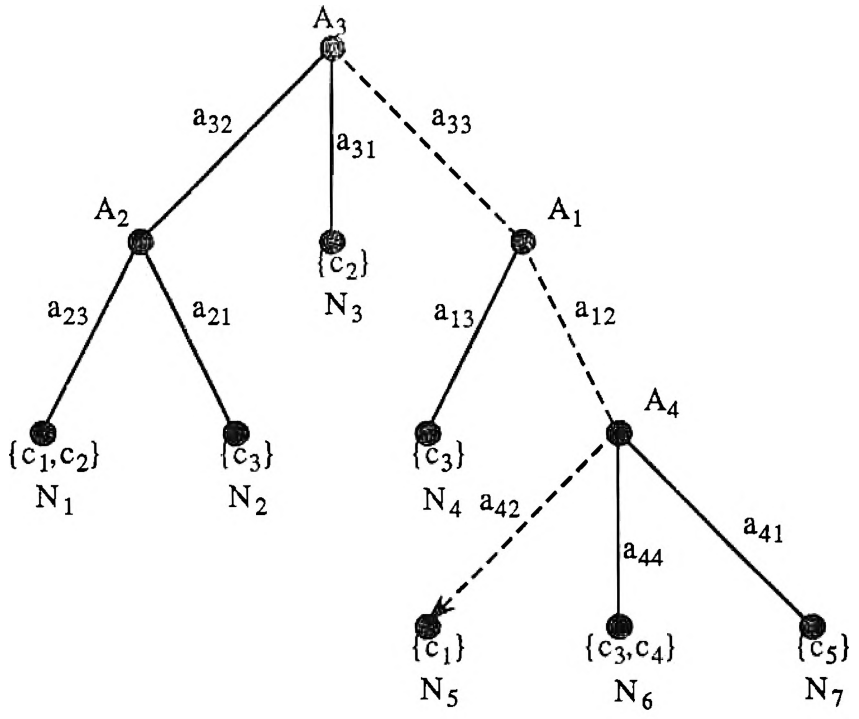

Fig. 1. Sample ID3 decision tree showing the rule

$A_3(a_{33}) \Lambda A_1(a_{12}) \Lambda A_4(a_{42}) \Rightarrow C(N_5)$.

The leaf nodes in the decision tree consisting of only one conclusion are called *single-conclusion leaf nodes*. If there exists more than one conclusion at any leaf node, that leaf node is described as a *multiple-conclusion node*. The decision tree in Fig. 1 has four interior nodes (A_1,A_2,A_3,A_4) inclusive of the root node. The seven leaf nodes are comprised of five single-conclusion leaf nodes, (N_2,N_3,N_4,N_5,N_7), and two multiple conclusion leaf nodes, (N_1,N_6).

3. DECISION TREE ACCURACY

One measurement of the quality of the tree produced by ID3 is the calculation of the accuracy when the decision tree is used to classify instances in test sets. The description of test instances is similar to that of the training instances. A test instance is said to be *correctly classified*, if there exists a path in the decision tree whose attribute value matches the attribute value of the test instance. The test instance (a_{12},a_{23},a_{33},a_{41},c_5) is correctly classified by the tree in Fig. 1. The classification path terminates at the leaf node labeled N_7. Clearly, the attribute value a_{23} is not needed for this classification.

If there does not exist any path in the decision tree for a test instance then this instance is said to be *misclassified*. The node, at which an instance attribute value can not be matched, is the *fail node* of that particular instance. For example, the test instance (a_{12},a_{24},a_{33},a_{43},c_1) is a misclassified instance and the classification path for this instance leads to A_4, but no branch labeled a_{43} leaves A_4. Hence, node A_4 is a fail node. If all the nodes in a path of the decision tree match a test instance, but the classification value of the test instance does not match the classification at the leaf node, then the instance is termed an *unmatched conclusion*. The corresponding leaf node is called the *unmatched node*. In the instance (a_{12},a_{24},a_{33},a_{41},c_1), the class c_1 is the

unmatched conclusion and the leaf node N_7 is the unmatched node. If a fail node or an unmatched conclusion occurs, then correct classification is not possible.

The accuracy of the ID3 tree reflects the ability of the tree to correctly classify training examples. Accuracy is calculated by the formula;

$$A = \frac{C}{T} * 100$$

where,

A - *% Accuracy*,

C - *Number of correctly classified test instances*, and

T - *Total number of test instances*.

As long as there is only one conclusion at any of the leaf nodes, this concept holds. When there is more than one conclusion at any of the leaf nodes, the accuracy depends on the interpretation of the word "match." The present work proposes three different criteria to deal with the issue:

- Multiple Conclusions (MC),
- Single Conclusion (SC), and
- Most Probable Conclusion (MPC).

In the case of Multiple Conclusions (MC), any leaf node consisting of more than one conclusion, is considered as a node with multiple conclusions. Hypothetically, there are multiple leaf nodes having the same path in the decision tree. A match occurs, if the test instance conclusion is in the set of conclusions at the leaf node. This can be a normal situation in a real-life task, since problems may have more

than one solution. This type of conclusion matching should produce higher test accuracy than those produced by Single or Most Probable Conclusion matching.

The Single Conclusion(SC) approach is applicable in cases where the appearance of multiple conclusions at any of the leaf nodes is taken as representing contradictory information. As such, these instances are rejected as incorrect. This assumption rejects the possibility of the existence of more than one conclusion for a single set of input attribute values. This type of match is much more restrictive and much less tolerant of noisy training data than MC. The % accuracy obtained using SC criterion be no larger than the % accuracy obtained using MC criterion. If no multiple conclusions exist, the two accuracies are identical.

In the Most Probable Conclusion (MPC) case, the most frequently occurring conclusion out of all the conclusions at a node is picked as the node conclusion. Noisy training data has a somewhat lesser effect on this type of accuracy. This criterion is more restrictive than MC but less restrictive than SC.

The SC approach is common in the literature [Quinlan 1986], but MC and MPC criteria appear to work well in many real-world tasks. Note also that the three criteria, discussed above, do not exhaust other possibilities. Other solutions have been proposed [Quinlan 1989].

4. ID3 TREE CONSTRUCTION IN THE PRESENCE OF NOISE

Training instances for a real domain are obtained through measurements, observations, and/or normal logical reasoning. As such they are subjected to noise. However machine learning algorithms, such as ID3, treat training instances as both

complete and perfectly correct. Nevertheless, instances being used for training can have incorrect values either in one or more attributes or in the classification information. The classifier system must be able to deal with this kind of noise.

Furthermore, noise is a principal factor affecting accuracy. Noise can be either in the training or in the test instances or in both. The importance of noise-free training instances lies in the fact that the system, if trained from the noisy environment, obviously may not generate correct conclusions in the decision tree, even if the test objects are noise-free. As a result, accuracy may be poor. This results in overall weak conclusions in spite of the noise-free test instances. The generation of incorrect branches in the decision tree can also result from noisy training instances. Consequently, it is important to study the effect of noise in training instances before studying the effect of noise in test instances.

5. NOISE AND THE χ^2 TEST

Training on noisy data can cause ID3 to construct a tree which is more complex than a tree constructed from noise-free data. The resulting tree is likely to perform poorly on the test data. One suggested method for the reduction of the complexity of the tree, is to use the χ^2 statistical test with a high degree of confidence. The following description of the process is based on Quinlan's [1986] development.

Let c_{ij} denote the number of times conclusion c_i is associated with node N_j. Assume Node N_p is the parent node for nodes N_j, $j = 1..\lambda$. The value of λ is determined by the number of values that can be assumed by N_p. Calculate the χ^2 statistic,

$$\chi^2_{calc} = \sum_{j=1}^{\lambda} \left[\sum_{i=1}^{|N_j|} \frac{(c_{ij} - c^*_{ij})^2}{c^*_{ij}} \right]$$

where $|N_j|$ denotes the number of conclusions associated with node N_j and expected value of c_{ij} is,

$$c^*_{ij} = c_{ip} \frac{|N_j|}{|N_p|}$$

The value for χ^2_{calc} is compared to a tabular χ^2 value, $\chi^2_{tab}(\lambda\text{-}1, \alpha)$, for confidence factor α and λ-1 degrees of freedom. At a given node, only those attributes for which $\chi^2_{calc} > \chi^2_{tab}(\lambda\text{-}1, \alpha)$ are considered as potential splitting attributes. Attributes which do not meet this criterion are considered to be too noisy. Because, the χ^2 test is an approximation of an exact distribution, this tree pruning technique works well only if the approximation is good. Hoel [1971] suggests that $|c_{ij}^*| \geq 5$ should be satisfied in order to use this test.

6. EXPERIMENTAL RESULTS

Several experiments were conducted in order to evaluate the effect of the χ^2 test on the construction of ID3 decision trees. The following statistics were recorded:

* % noise in the training data (test data was noise-free),
* Number of splits prevented by χ^2 test,
* Number of leaf nodes,
* Number of interior nodes (the root node is considered as interior),
* Number of multiple conclusion leaf nodes, and
* Accuracy (% correct) for SC, MC, and MPC methods of conclusion matching.

Experiments were run with and without the χ^2 test so the effects of the χ^2 test could be evaluated.

A method for regulating the amount of noise in the training data was essential since the amount of noise in each training set needed to be known. To accomplish this, the Noisy File Generator(NFG) system was developed. The NFG randomly selects a training instance and within that instance, one attribute or classification value is randomly selected. A randomly chosen value from the corresponding domain is then substituted for the current value. All random values were generated from the uniform distribution. Experiments were performed using various percentages of noise: 0%, 20%, 40%, and 60%. For instance, in the case of 40% noise insertion, noise was introduced in 40% of the training instances.

Seven diversified domains were chosen for testing. Several of these domains were real and as such, contained noisy data. Additional noise was added to each of the training sets. Various numbers of training instances were used, while 100% of the original data used as testing instances.

- **Multiplexer Circuit Analysis Domain:** In this domain, there are two address and four data lines for a total of six attributes of a multiplexer circuit. The output of this circuit is the classification value. All attribute and classification values are binary. This domain consists of 64 cases of which 70% were used for training. Note, that this data has a uniform distribution of values. They are composed of correct, but a relatively small number of the instances. This data set is described by Paul Utgoff [1988].

- **Mushroom Classification Domain:** This domain consists of a subset of real data samples. It has twenty-two independent attributes, but only five of them were needed in order to get significant results. Each mushroom is classified into one of two classes: definitely edible or definitely poisonous.

This domain has 4000 cases. Out of these cases, 25% were used for training. This data set was drawn from the The Audubon Society Field Guide to North American Mushrooms [Lincoff 1981] and provided by J. C. Schlimmer of Washington State University.

• **Soybean Disease Diagnosis Domain:** Soybean disease case histories comprise this domain. Each history has thirty-five attributes, twenty-nine of which were used for these experiments. Attribute values were converted to numerical values depending on their characteristics. Every object is classified into one of four classification values. There are 64 cases and all of them are used for both training and testing. This data set is described by Doug Fisher [1987].

• **Thyroid Disease Diagnosis Domain:** This domain describes 150 thyroid case histories. It has nineteen attributes composed of a combination of continuous values, discrete values, and unknown values. The continuous attributes were normalized using value ranges specified by experts in the thyroid field. These examples are classified into one of the three diagnostic classification values: negative, hypothyroid, or sick euthyroid. The data set was drawn from the Garvan Institute in Australia and was obtained from Doug Fisher [1987].

• **Election(1984) Prediction Domain:** This domain contains sixteen attributes, fourteen of which were used for the experiments. The attributes are in the form of queries. The classification values are the responses to these queries given by members of the U. S. House of Representatives. Each object is classified into one of two classification values: Republican or Democrat. The domain consists of 435 examples of which 50% were used for training. This data was drawn from the Congressional Quarterly Almanac, 98th Congress and compiled by J. C. Schlimmer at Washington State University.

• **Artificial Domain:** This artificially created domain models probabilistic classification over disjunction [Quinlan 1987]. It consists of ten Boolean attributes. The class T and F is derived as;

If	$(A_0(a_{11}) \Lambda A_1(a_{12}) \Lambda A_2(a_{13})) V(A_3(a_{21}) \Lambda A_4(a_{22}) \Lambda A_5(a_{23})$ $V (A_6(a_{31}) \Lambda A_7(a_{32}) \Lambda A_8(a_{33}))$
Then	(Class T with 90% probability) (Class F with 10% probability)
Otherwise	(Class T with 10% probability) (Class F with 90% probability)

Note, attribute A_9 is not used for experimentation and does not have any significance. There are 1070 cases and 60% of these were used for training.

• **Election (1986) Prediction Domain:** This domain is similar to the Election (1984) domain. There are seventeen attributes in the form of queries. In this domain there are a total of 95 instances with 80% being used for training. Dr. J. C. Schlimmer at Washington State University compiled these data.

Many of these data sets were obtained from the Machine Learning Database at University of California - Irvine.

The results of these experiments are summarized in the Tables I and II. Table I summarizes the trees produced with and without the χ^2 test for each data set at various levels of training noise. The number of splits prevented by the χ^2 test is the number of times further tree expansion from a leaf node was prevented because the χ^2 test excluded all available attributes. The node then became a leaf node. As expected, the number of splits prevented by the χ^2 test may increase with an increase in training noise. This appears to be more pronounced in data sets such as Mushroom and Artificial which have a large number of training examples.

Tree complexity as measured by counting the number of interior and leaf nodes is noticeably less for χ^2 trees than for the non-χ^2 trees. Within a data set, the number of interior and leaf nodes in non-χ^2 trees usually increases as the amount of

Domain	% Noise in Train. Data	# of Splits Prevented By χ^2	# Leaf Nodes		# Interior Nodes		# Multiple Conclusions at Leaf Nodes	
			no χ^2	χ^2	no χ^2	χ^2	no χ^2	χ^2
(1)Multiplexer	0	2	13	3	12	2	0	2
(45 Training)	20	2	21	2	16	1	4	0
(64 Testing)	40	1	22	2	18	1	3	0
	60	1	29	2	23	1	5	0
(2)Mushroom	0	2	7	3	3	1	2	2
(1000Training)	20	7	32	3	27	1	13	7
(4000 Testing)	40	6	38	2	27	1	18	6
	60	11	43	6	25	2	23	11
(3)Soybean	0	2	15	6	7	3	0	2
(64 Training)	20	2	46	10	34	3	2	2
(64 Testing)	40	8	69	6	48	2	3	8
	60	11	103	12	60	9	5	12
(4)Thyroid	0	7	19	2	10	1	0	7
(102 Training)	20	3	20	2	9	1	0	3
(150 Testing)	40	7	27	2	15	1	0	0
	60	7	32	2	17	1	0	0
(5)Vote-84	0	3	12	5	7	3	0	3
(218 Training)	20	5	89	3	65	1	6	5
(435 Testing)	40	9	126	7	74	4	10	9
	60	7	125	7	102	2	10	7
(6)Artificial	0	11	173	16	167	6	6	11
(642 Training)	20	13	278	12	247	7	24	13
(1070 Testing)	40	7	323	7	265	2	44	7
	60	8	335	7	248	2	60	8
(7)Vote-86	0	2	17	3	13	1	0	2
(76 Training)	20	3	39	5	34	3	2	3
(95 Testing)	40	3	39	2	34	1	1	3
	60	2	34	3	31	1	1	2

Table I. Tree Structure

Domain	% Noise In Training Data	ACCURACY MC		SC		MPC	
		no χ^2	χ^2	no χ^2	χ^2	no χ^2	χ^2
(1)Multiplexer	0	84.4	87.5	84.4	37.5	84.4	62.5
(45 Training)	20	70.3	100.0	64.1	0.0	70.3	59.4
(64 Testing)	40	67.2	100.0	62.5	0.0	67.2	76.6
	60	54.7	100.0	46.9	0.0	54.7	56.2
(2)Mushroom	0	91.2	91.2	69.9	62.5	82.7	82.6
(1000Training)	20	93.4	93.4	48.0	0.0	82.6	82.6
(4000 Testing)	40	93.4	93.4	2.4	0.0	82.6	82.7
	60	93.4	93.4	0.0	0.0	82.6	82.7
(3)Soybean	0	74.0	96.0	74.0	62.7	74.0	83.8
(64 Training)	20	70.7	74.7	64.7	38.0	70.7	80.4
(64 Testing)	40	96.0	96.0	62.0	16.7	69.3	75.4
	60	48.0	91.3	39.3	80.0	48.6	56.2
(4) Thyroid	0	100.0	100.0	100.0	13.0	100.0	100.0
(102 Training)	20	89.1	100.0	89.1	13.0	89.1	92.2
(150 Testing)	40	80.4	97.8	80.4	15.2	80.4	90.5
	60	69.6	93.5	69.6	15.2	69.6	82.4
(5)Vote-84	0	94.7	97.2	94.7	88.7	94.7	94.7
(218 Training)	20	87.1	100.0	81.8	0.3	86.2	91.5
(435 Testing)	40	74.3	97.0	66.0	0.2	74.3	82.3
	60	61.8	99.1	50.6	0.0	60,3	80.5
(6)Artificial	0	72.7	86.0	71.6	83.4	72.7	74.4
(642 Training)	20	66.4	100.0	63.0	0.0	66.4	69.6
(1070 Testing)	40	60.5	97.1	54.9	0.1	60.0	62.4
	60	59.3	100.0	51.9	0.0	58.5	64.5
(7)Vote-86	0	89.5	95.8	89.5	21.1	98.5	89.5
(76 Training)	20	83.2	95.8	78.9	17.9	83.2	83.2
(95 Testing)	40	72.6	97.9	70.5	0.1	72.6	73.7
	60	68.4	100.0	66.3	0.0	68.4	72.6

Table II. Accuracy %

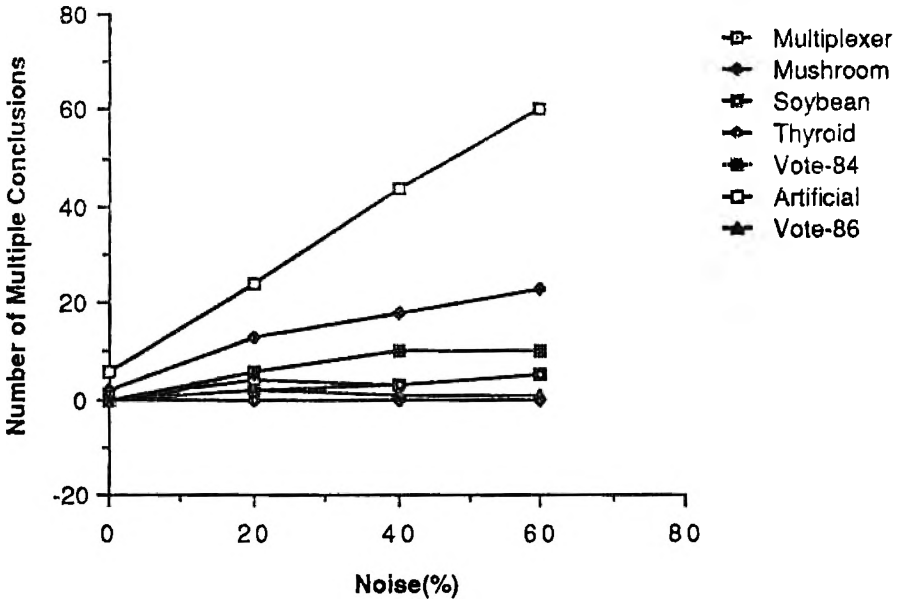

Fig. 2. Number of multiple conclusions at leaf nodes when χ^2 test is not applied

noise in the training data is increased. This relationship does not appear to hold for χ^2 trees.

The number of leaf nodes with more than one conclusion is reported in the right-hand two columns of Table I. Figures 2 and 3 show this information as a function of noise in the training set. As noise increases, the number of multiple conclusions at the leaf nodes increase. In most cases, χ^2 trees contain fewer multiple conclusion leaf nodes.

Examination of the figures for a given data set, indicate that, for the data tested, the number of multiple conclusions tends to change less abruptly as one moves from

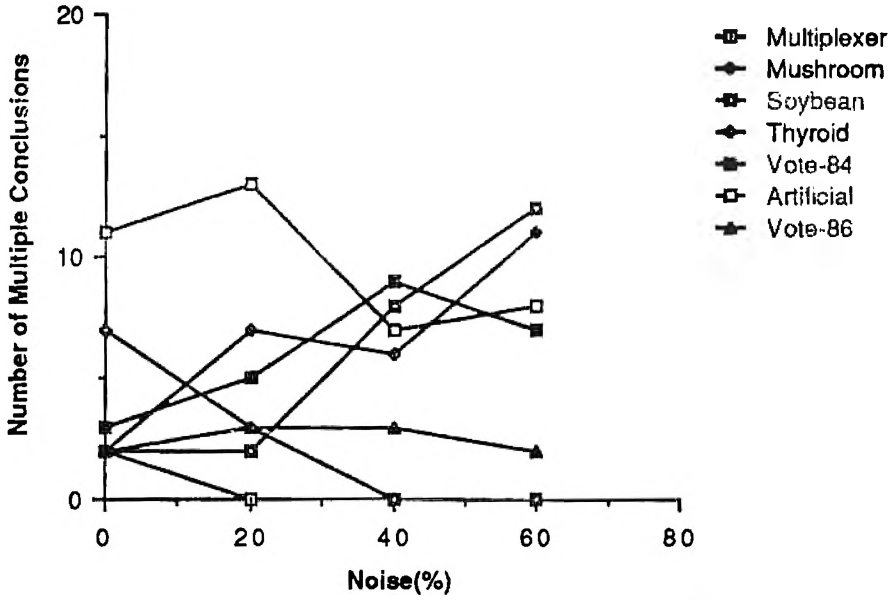

Fig. 3. Number of multiple conclusions when χ^2 test is applied.

one level of noise to the next higher level in cases where the χ^2 test has been applied. For the non χ^2 produced data, the increase appears to be monotone nondecreasing in general. This is not true for the trees produced using χ^2.

Table II shows SC, MC, and MPC % Accuracy obtained from tests shown in Table I. As expected, MC accuracy is higher relative to SC and MPC % Accuracy falls between MC and SC. The difference in MC and SC is 100% in some cases. This suggests that the choice of comparison criteria is critical to performance. As indicated earlier, this choice depends on the domain being learned.

In the results for the MC criterion, increasing the noise level appears to have a smaller effect on the % Accuracy of trees built using the χ^2 test. This % Accuracy remains consistently high throughout all levels of noise. In all cases, the % Accuracy of the trees produced using the χ^2 test is equivalent to or exceeds the % Accuracy of the

trees produced without using the χ^2 test! This fact coupled with the fact that χ^2 produced trees contain fewer nodes than non-χ^2 produced trees, suggest that in situations where the MC criterion is appropriate, ID3 trees should be constructed using the χ^2 test.

Results from the SC criterion are much different from those produced by the MC criterion. For non-χ^2 produced trees, the % Accuracy behaves as expected in that as the percentage of noise increases, the % Accuracy decreases. The strange behavior of % Accuracy for produced trees is easily understood by observing the number of interior and leaf nodes for these trees. For example, the χ^2 produced Mushroom tree derived from 20% noise in the training data shows 0.0 % Accuracy. This tree has one interior node, the root, and three leaf nodes. In addition, it has seven multiple conclusions at the leaf nodes. If each of the three leaf nodes contains more than one conclusion, the SC criterion will refuse to classify any test example as correct! This suggests that only trees produced with few multiple conclusions and a significant number of leaf nodes will be of practical classification value when using the SC criterion.

As expected, MPC produces results which, in general, lie between those of SC and MC. Percent accuracy for MPC non-χ^2 produced trees is similar to that for MC non- χ^2 produced trees. In some cases, accuracy results for MPC χ^2 produced trees is lower than for trees not produced using the χ^2 test. In other cases it is higher. In all cases, an increase in noise in a given data set appears to have less effect on % Accuracy than it does for the other comparison techniques. These results suggest that MPC is more tolerant of noisy training data.

7. CONCLUSIONS

The experimental results suggest that the χ^2 test significantly affects the construction of ID3 decision trees. In general, χ^2 produced trees are less complex in that they contain fewer nodes. In some cases, χ^2 produced trees may be so sparse that their classification performance is poor.

Classification performance of all trees directly depends on both the amount of noise present in the training data and the comparison criteria used when comparing node conclusions. In general, performance achieved using the Most Probable Conclusion criterion falls between that produced by the Single Conclusion and Multiple Conclusion criteria. The choice of criterion depends on the domain being learned. Decision trees produced using the χ^2 test and the Most Probable Conclusion criterion perform consistently well in the presence of increasing noise.

REFERENCES

Finney, D. J., Latscha, R., Bennett, B. M. &Hsu, P. Tables for Testing Significance in a 2 X 2 Contingency Table. Cambridge: Cambridge University Press, 1963.

Fisher, D. H. "Knowledge Acquisition via Incremental Conceptual Clustering." Doctoral dissertation, Department of Information and Computer Science, University of California, Irvine, CA., 1987.

Hoel P. G. Introduction to Mathematical Statistics. John Wiley & Sons, Inc.1971.

Lewis, P. M. "The Characteristic Selection Problem in Recognition Systems." Transactions on Information Theory (1962): 171-178.

Lincoff, G. H. (1981) The Audubon Society Field Guide to North American Mushrooms. New York: Alfred A. Knopf(original not seen).

Quinlan, J. R. "Induction of Decision Trees." Machine Learning 1 (1986): 81-106.

_____. "Simplifying Decision Trees." In International Journal of Man-Machine Studies(1987): 221-234.

_____. "Unknown Attribute Values In Induction." In Proceedings of the Sixth International Workshop On Machine Learning (1989): Morgan Kaufmann Publishers, Inc.:164-168.

Utgoff, P.E. "ID5: An Incremental ID3." In Proceedings of the Fifth International Conference On Machine Learning (1988): 107-120.