6 Decision Tree Induction: Using Frequency Tables for Attribute Selection

6.1 Calculating Entropy in Practice

The detailed calculations needed to choose an attribute to split on at a node in the evolving decision tree were illustrated in Section 5.3.3. At each node a table of values such as Figure 5.10(a), reproduced here as Figure 6.1, needs to be calculated for every possible value of every categorical attribute.

Value of attribute				Class
age	specRx	astig	tears	
1	1	1	1	3
1	1	1	2	2
1	1	2	1	3
1	1	2	2	1
1	2	1	1	3
1	2	1	2	2
1	2	2	1	3
1	2	2	2	1

Figure 6.1 Training Set 1 (age = 1) for *lens24* Example

For practical use a more efficient method is available which requires only a

M. Bramer, *Principles of Data Mining*, Undergraduate Topics in Computer Science, DOI 10.1007/978-1-4471-7307-6_6

single table to be constructed for each categorical attribute at each node. This method, which can be shown to be equivalent to the one given previously (see Section 6.1.1), uses a *frequency table.* The cells of this table show the number of occurrences of each combination of class and attribute value in the training set. For the *lens24* dataset the frequency table corresponding to splitting on attribute *age* is shown in Figure 6.2.

	age = 1	age = 2	age = 3
Class 1	2	1	1
Class 2	2	2	1
Class 3	4	5	6
Column sum	8	8	8

Figure 6.2 Frequency Table for Attribute age for *lens24* Example

We will denote the total number of instances by N, so $N = 24$.

The value of E_{new}, the average entropy of the training sets resulting from splitting on a specified attribute, can now be calculated by forming a sum as follows.

(a) For every non-zero value V in the main body of the table (i.e. the part above the 'column sum' row), subtract $V \times \log_2 V$.

(b) For every non-zero value S in the column sum row, add $S \times \log_2 S$.

Finally, divide the total by N.

Figure 6.3 gives the value of $\log_2 x$ for small integer values of x for reference.

Using the frequency table given as Figure 6.2, splitting on attribute *age* gives an E_{new} value of

$$-2\log_2 2 - 1\log_2 1 - 1\log_2 1 - 2\log_2 2 - 2\log_2 2 - 1\log_2 1$$
$$-4\log_2 4 - 5\log_2 5 - 6\log_2 6 + 8\log_2 8 + 8\log_2 8 + 8\log_2 8$$

divided by 24. This can be rearranged as

$$(-3 \times 2\log_2 2 - 3\log_2 1 - 4\log_2 4 - 5\log_2 5 - 6\log_2 6 + 3 \times 8\log_2 8)/24$$

$= 1.2867$ bits (to 4 decimal places), which agrees with the value calculated previously.

6.1.1 Proof of Equivalence

It remains to be proved that this method always gives the same value of E_{new} as the basic method described in Chapter 5.

x	$\log_2 x$
1	0
2	1
3	1.5850
4	2
5	2.3219
6	2.5850
7	2.8074
8	3
9	3.1699
10	3.3219
11	3.4594
12	3.5850

Figure 6.3 Some values of $\log_2 x$ (to 4 decimal places)

Assume that there are N instances, each relating the value of a number of categorical attributes to one of K possible classifications. (For the *lens24* dataset used previously, $N = 24$ and $K = 3$.)

Splitting on a categorical attribute with V possible values produces V subsets of the training set. The jth subset contains all the instances for which the attribute takes its jth value. Let N_j denote the number of instances in that subset. Then

$$\sum_{j=1}^{V} N_j = N$$

(For the frequency table shown in Figure 6.2, for attribute *age*, there are three values of the attribute, so $V = 3$. The three column sums are N_1, N_2 and N_3, which all have the same value (8). The value of N is $N_1 + N_2 + N_3 = 24$.)

Let f_{ij} denote the number of instances for which the classification is the ith one and the attribute takes its jth value (e.g. for Figure 6.2, $f_{32} = 5$). Then

$$\sum_{i=1}^{K} f_{ij} = N_j$$

The frequency table method of forming the sum for E_{new} given above amounts to using the formula

$$E_{new} = -\sum_{j=1}^{V}\sum_{i=1}^{K}(f_{ij}/N).\log_2 f_{ij} + \sum_{j=1}^{V}(N_j/N).\log_2 N_j$$

The basic method of calculating E_{new} using the entropies of the j subsets resulting from splitting on the specified attribute was described in Chapter 5.

The entropy of the jth subset is E_j where

$$E_j = -\sum_{i=1}^{K}(f_{ij}/N_j).\log_2(f_{ij}/N_j)$$

The value of E_{new} is the weighted sum of the entropies of these V subsets. The weighting is the proportion of the original N instances that the subset contains, i.e. N_j/N for the jth subset. So

$$\begin{aligned}
E_{new} &= \sum_{j=1}^{V} N_j E_j/N \\
&= -\sum_{j=1}^{V}\sum_{i=1}^{K}(N_j/N).(f_{ij}/N_j).\log_2(f_{ij}/N_j) \\
&= -\sum_{j=1}^{V}\sum_{i=1}^{K}(f_{ij}/N).\log_2(f_{ij}/N_j) \\
&= -\sum_{j=1}^{V}\sum_{i=1}^{K}(f_{ij}/N).\log_2 f_{ij} + \sum_{j=1}^{V}\sum_{i=1}^{K}(f_{ij}/N).\log_2 N_j \\
&= -\sum_{j=1}^{V}\sum_{i=1}^{K}(f_{ij}/N).\log_2 f_{ij} + \sum_{j=1}^{V}(N_j/N).\log_2 N_j \quad [\text{as } \sum_{i=1}^{K} f_{ij} = N_j]
\end{aligned}$$

This proves the result.

6.1.2 A Note on Zeros

The formula for entropy given in Section 5.3.2 excludes empty classes from the summation. They correspond to zero entries in the body of the frequency table, which are also excluded from the calculation.

If a complete column of the frequency table is zero it means that the categorical attribute never takes one of its possible values at the node under consideration. Any such columns are ignored. (This corresponds to ignoring empty subsets whilst generating a decision tree, as described in Section 4.2, Figure 4.5.)

6.2 Other Attribute Selection Criteria: Gini Index of Diversity

As well as entropy (or information gain) many other methods have been proposed for selecting the attribute to split on at each stage of the TDIDT algorithm. There is a useful review of several methods by Mingers [1].

One measure that is commonly used is the *Gini Index of Diversity*. If there are K classes, with the probability of the ith class being p_i, the Gini Index is defined as $1 - \sum_{i=1}^{K} p_i^2$.

This is a measure of the 'impurity' of a dataset. Its smallest value is zero, which it takes when all the classifications are the same. It takes its largest value $1 - 1/K$ when the classes are evenly distributed between the instances, i.e. the frequency of each class is $1/K$.

Splitting on a chosen attribute gives a reduction in the average Gini Index of the resulting subsets (as it does for entropy). The new average value Gini_{new} can be calculated using the same frequency table used to calculate the new entropy value in Section 6.1.

Using the notation introduced in that section, the value of the Gini Index for the jth subset resulting from splitting on a specified attribute is G_j, where

$$G_j = 1 - \sum_{i=1}^{K} (f_{ij}/N_j)^2$$

The weighted average value of the Gini Index for the subsets resulting from splitting on the attribute is

$$\begin{aligned}
\text{Gini}_{new} &= \sum_{j=1}^{V} N_{j\cdot} G_j / N \\
&= \sum_{j=1}^{V} (N_j/N) - \sum_{j=1}^{V} \sum_{i=1}^{K} (N_j/N).(f_{ij}/N_j)^2 \\
&= 1 - \sum_{j=1}^{V} \sum_{i=1}^{K} f_{ij}^2/(N.N_j) \\
&= 1 - (1/N) \sum_{j=1}^{V} (1/N_j) \sum_{i=1}^{K} f_{ij}^2
\end{aligned}$$

At each stage of the attribute selection process the attribute is selected which maximises the reduction in the value of the Gini Index, i.e. $\text{Gini}_{start} - \text{Gini}_{new}$.

Again taking the example of the *lens24* dataset, the initial probabilities of the three classes as given in Chapter 5 are $p_1 = 4/24$, $p_2 = 5/24$ and $p_3 = 15/24$. Hence the initial value of the Gini Index is $G_{start} = 0.5382$.

For splitting on attribute *age* the frequency table, as before, is shown in Figure 6.4.

We can now calculate the new value of the Gini Index as follows.

1. For each non-empty column, form the sum of the squares of the values in the body of the table and divide by the column sum.
2. Add the values obtained for all the columns and divide by N (the number of instances).

	age = 1	age = 2	age = 3
Class 1	2	1	1
Class 2	2	2	1
Class 3	4	5	6
Column sum	8	8	8

Figure 6.4 Frequency Table for Attribute age for *lens24* Example

3. Subtract the total from 1.

For Figure 6.4 we have

$$\textbf{age = 1:}\ (2^2 + 2^2 + 4^2)/8 = 3$$
$$\textbf{age = 2:}\ (1^2 + 2^2 + 5^2)/8 = 3.75$$
$$\textbf{age = 3:}\ (1^2 + 1^2 + 6^2)/8 = 4.75$$

$G_{new} = 1 - (3 + 3.75 + 4.75)/24 = 0.5208$.

Thus the reduction in the value of the Gini Index corresponding to splitting on attribute *age* is $0.5382 - 0.5208 = 0.0174$.

For the other three attributes, the corresponding values are

specRx: $G_{new} = 0.5278$, so the reduction is $0.5382 - 0.5278 = 0.0104$
astig: $G_{new} = 0.4653$, so the reduction is $0.5382 - 0.4653 = 0.0729$
tears: $G_{new} = 0.3264$, so the reduction is $0.5382 - 0.3264 = 0.2118$

The attribute selected would be the one which gives the largest reduction in the value of the Gini Index, i.e. *tears*. This is the same attribute that was selected using entropy.

6.3 The χ^2 Attribute Selection Criterion

Another useful attribute selection measure that can be calculated using a frequency table is the χ^2 value. χ is the Greek letter often rendered in the Roman alphabet as chi (pronounced 'sky' without the initial 's'). The term χ^2 is pronounced 'chi-square' or 'chi-squared'. It is commonly used in statistics. Its relevance to attribute selection will soon become apparent.

The method will be described in more detail and in a fuller form in a later chapter on discretisation of continuous attributes, so only a fairly brief description will be given here.

Suppose that for some dataset with three possible classifications $c1$, $c2$ and $c3$ we have an attribute A with four values $a1$, $a2$, $a3$ and $a4$, and the frequency table given in Figure 6.5.

	a1	a2	a3	a4	**Total**
c1	27	64	93	124	308
c2	31	54	82	105	272
c3	42	82	125	171	420
Total	100	200	300	400	1000

Figure 6.5 Frequency Table for Attribute A

We start by making the assumption that the value of A has no effect whatsoever on the classification and look for evidence that this assumption (which statisticians call the *null hypothesis*) is false.

It is quite easy to imagine four-valued attributes that are certain or almost certain to be irrelevant to a classification. For example the values in each row might correspond to the number of patients achieving a large benefit, a little benefit or no benefit (classifications $c1$, $c2$ and $c3$) from a certain medical treatment, with attribute values $a1$ to $a4$ denoting a division of patients into four groups depending on the number of siblings they have (say zero, one, two, three or more). Such a division would appear (to this layman) highly unlikely to be relevant. Other four-valued attributes far more likely to be relevant include age and weight, each converted into four ranges in this case.

The example may be made more controversial by saying that $c1$, $c2$ and $c3$ are levels achieved in some kind of intelligence test and $a1$, $a2$, $a3$ and $a4$ denote people who are married and male, married and female, unmarried and male or unmarried and female, not necessarily in that order. Does the test score obtained depend on which category you are in? Please note that we are not trying to settle such sensitive questions in this book, especially not with invented data, just (as far as this chapter is concerned) deciding which attribute should be selected when constructing a decision tree.

From now on we will treat the data as test results but to avoid controversy will not say anything about the kind of people who fall into the four categories $a1$ to $a4$.

The first point to note is that from examining the *Total* row we can see that the people who took the test had attribute values $a1$ to $a4$ in the ratio 1:2:3:4. This is simply a fact about the data we happen to have obtained and in itself implies nothing about the null hypothesis, that the division of test subjects into four groups is irrelevant.

Next consider the $c1$ row. We can see that a total of 308 people obtained classification $c1$. If the value of attribute A were irrelevant we would expect the 308 values in the cells to split in the ratio 1:2:3:4.

In cell $c1/a1$ we would expect a value of $308 * 100/1000 = 30.8$.

In $c1/a2$ we would expect twice this, i.e. $308 * 200/1000 = 61.6$.

In $c1/a3$ we would expect $308 * 300/1000 = 92.4$.

In $c1/a4$ we would expect $308 * 400/1000 = 123.2$.

(Note that the total of the four values comes to 308, as it must.)

We call the four calculated values above the *expected* values for each class/attribute value combination. The actual values in the $c1$ row: 27, 64, 93 and 124 are not far away from these. Do they and the expected values for the $c2$ and $c3$ rows support or undermine the null hypothesis, that attribute A is irrelevant?

Although the 'ideal' situation is that all the expected values are identical to the corresponding actual values, known as the *observed* values, this needs a strong caveat. If you ever read a published research paper, newspaper article etc. where for some data the expected values all turn out to be exact integers that are exactly the same as the observed values for all classification/attribute value combinations, by far the most likely explanation is that the published data is an exceptionally incompetent fraud. In the real world, such perfect accuracy is never achieved. In this example, as with most real data it is in any case impossible for the expected values to be entirely identical to the observed ones, as the former are not usually integers and the latter must be.

Figure 6.6 is an updated version of the frequency table given previously, with the observed value in each of the cells from $c1/a1$ to $c3/a4$ followed by its expected value in parentheses.

	a1	a2	a3	a4	**Total**
c1	27 (30.8)	64 (61.6)	93 (92.4)	124 (123.2)	308
c2	31 (27.2)	54 (54.4)	82 (81.6)	105 (108.8)	272
c3	42 (42.0)	82 (84.0)	125 (126.0)	171 (168.0)	420
Total	100	200	300	400	1000

Figure 6.6 Frequency Table for Attribute A Augmented by Expected Values

The notation normally used is to represent the observed value for each cell by O and the expected value by E. The value of E for each cell is just the product of the corresponding column sum and row sum divided by the grand total number of instances given in the bottom right-hand corner of the table. For example the E value for cell $c3/a2$ is $200 * 420/1000 = 84.0$.

We can use the values of O and E for each cell to calculate a measure of how far the frequency table varies from what we would expect if the null hypothesis (that attribute A is irrelevant) were correct. We would like the measure to be zero in the case that the E values in every cell are always identical to the corresponding O values.

The measure generally used is the χ^2 value, which is defined as the sum of the values of $(O - E)^2/E$ over all the cells.

Calculating the χ^2 value for the updated frequency table above, we have $\chi^2 = (27 - 30.8)^2/30.8 + \ldots + (171 - 168.0)^2/168.0 = 1.35$ (to two decimal places).

Is this χ^2 value small enough to give support for the null hypothesis that attribute A is irrelevant to the classification? Or is it large enough to suggest that the null hypothesis is false?

This question will be important when the same method is used later in connection with the discretisation of continuous attributes, but as far as this chapter is concerned we will ignore the question of the validity of the null hypothesis and simply record the value of χ^2. We then repeat the process with all the attributes under consideration as the attribute to split on in our decision tree and choose the one with the largest χ^2 value as the one likely to have the greatest power of discrimination amongst the three classifications.

6.4 Inductive Bias

Before going on to describe a further method of attribute selection we will introduce the idea of *inductive bias*, which will help to explain why other methods are needed.

Consider the following question, which is typical of those that used to be (and probably still are) set for school children to answer as part of a so-called 'intelligence test'.

> Find the next term in the sequence
> 1, 4, 9, 16, ...

Pause and decide on your answer before going on.

Most readers will probably have chosen the answer 25, but this is misguided. The correct answer is 20. As should be obvious, the nth term of the series is calculated from the formula:

nth term $= (-5n^4 + 50n^3 - 151n^2 + 250n - 120)/24$

By choosing 25, you display a most regrettable bias towards perfect squares.

This is not serious of course, but it *is* trying to make a serious point. Mathematically it is possible to find some formula that will justify any further development of the sequence, for example

$$1, 4, 9, 16, 20, 187, -63, 947$$

It is not even necessary for a term in a sequence to be a number. The sequence

$$1, 4, 9, 16, \text{dog}, 36, 49$$

is perfectly valid mathematically. (A restriction to numerical values shows a bias towards numbers rather than the names of types of animal.)

Despite this, there is little doubt that anyone answering the original question with 20 will be marked as wrong. (Answering with 'dog' is definitely not to be recommended.)

In practice we have a strong preference for hypothesising certain kinds of solution rather than others. A sequence such as

	$1, 4, 9, 16, 25$	(perfect squares)
or	$1, 8, 27, 64, 125, 216$	(perfect cubes)
or	$5, 8, 11, 14, 17, 20, 23, 26$	(values differ by 3)

seems reasonable, whereas one such as

$$1, 4, 9, 16, 20, 187, -63, 947$$

does not.

Whether this is right or wrong is impossible to say absolutely — it depends on the situation. It illustrates an *inductive bias*, i.e. a preference for one choice rather than another, which is not determined by the data itself (in this case, previous values in the sequence) but by external factors, such as our preferences for simplicity or familiarity with perfect squares. In school we rapidly learn that the question-setter has a strong bias in favour of sequences such as perfect squares and we give our answers to match this bias if we can.

Turning back to the task of attribute selection, any formula we use for it, however principled we believe it to be, introduces an inductive bias that is not justified purely by the data. Such bias can be helpful or harmful, depending on the dataset. We can choose a method that has a bias that we favour, but we cannot eliminate inductive bias altogether. There is no neutral, unbiased method.

Clearly it is important to be able to say what bias is introduced by any particular method of selecting attributes. For many methods this is not easy to do, but for one of the best-known methods we can. Using entropy can be shown to have a bias towards selecting attributes with a large number of values.

For many datasets this does no harm, but for some it can be undesirable. For example we may have a dataset about people that includes an attribute 'place of birth' and classifies them as responding to some medical treatment 'well', 'badly' or 'not at all'. Although the place of birth may have some effect on the classification it is probably only a minor one. Unfortunately, the information gain selection method will almost certainly choose it as the first attribute to split on in the decision tree, generating one branch for each possible place of birth. The decision tree will be very large, with many branches (rules) with very low value for classification.

6.5 Using Gain Ratio for Attribute Selection

In order to reduce the effect of the bias resulting from the use of information gain, a variant known as Gain Ratio was introduced by the Australian academic Ross Quinlan in his influential system C4.5 [2]. Gain Ratio adjusts the information gain for each attribute to allow for the breadth and uniformity of the attribute values.

The method will be illustrated using the frequency table given in Section 6.1. The value of E_{new}, the average entropy of the training sets resulting from splitting on attribute *age*, has previously been shown to be 1.2867 and the entropy of the original training set E_{start} has been shown to be 1.3261. It follows that

$$\text{Information Gain} = E_{start} - E_{new} = 1.3261 - 1.2867 = 0.0394.$$

Gain Ratio is defined by the formula

$$\textbf{Gain Ratio} = \textbf{Information Gain}/\textbf{Split Information}$$

where Split Information is a value based on the column sums.

Each non-zero column sum s contributes $-(s/N)\log_2(s/N)$ to the Split Information. Thus for Figure 6.2 the value of Split Information is

$$-(8/24)\log_2(8/24) - (8/24)\log_2(8/24) - (8/24)\log_2(8/24) = 1.5850$$

Hence Gain Ratio = 0.0394/1.5850 = 0.0249 for splitting on attribute *age*.

For the other three attributes, the value of *Split Information* is 1.0 in each case. Hence the values of Gain Ratio for splitting on attributes *specRx*, *astig* and *tears* are 0.0395, 0.3770 and 0.5488 respectively.

The largest value of Gain Ratio is for attribute *tears*, so in this case Gain Ratio selects the same attribute as entropy.

6.5.1 Properties of Split Information

Split Information forms the denominator in the Gain Ratio formula. Hence the higher the value of Split Information, the lower the Gain Ratio.

The value of Split Information depends on the number of values a categorical attribute has and how uniformly those values are distributed (hence the name 'Split Information').

To illustrate this we will examine the case where there are 32 instances and we are considering splitting on an attribute a, which has values 1, 2, 3 and 4.

The 'Frequency' row in the tables below is the same as the column sum row in the frequency tables used previously in this chapter.

The following examples illustrate a number of possibilities.

1. Single Attribute Value

	$a=1$	$a=2$	$a=3$	$a=4$
Frequency	32	0	0	0

Split Information $= -(32/32) \times \log_2(32/32) = -\log_2 1 = 0$

2. Different Distributions of a Given Total Frequency

	$a=1$	$a=2$	$a=3$	$a=4$
Frequency	16	16	0	0

Split Information $= -(16/32) \times \log_2(16/32) - (16/32) \times \log_2(16/32) = -\log_2(1/2) = 1$

	$a=1$	$a=2$	$a=3$	$a=4$
Frequency	16	8	8	0

Split Information $= -(16/32) \times \log_2(16/32) - 2 \times (8/32) \times \log_2(8/32) = -(1/2)\log_2(1/2) - (1/2)\log_2(1/4) = 0.5 + 1 = 1.5$

	$a=1$	$a=2$	$a=3$	$a=4$
Frequency	16	8	4	4

Split Information $= -(16/32) \times \log_2(16/32) - (8/32) \times \log_2(8/32) - 2 \times (4/32) \times \log_2(4/32) = 0.5 + 0.5 + 0.75 = 1.75$

3. Uniform Distribution of Attribute Frequencies

	$a=1$	$a=2$	$a=3$	$a=4$
Frequency	8	8	8	8

Split Information $= -4 \times (8/32) \times \log_2(8/32) = -\log_2(1/4) = \log_2 4 = 2$

In general, if there are M attribute values, each occurring equally frequently, the Split Information is $\log_2 M$ (irrespective of the frequency value).

6.5.2 Summary

Split Information is zero when there is a single attribute value.

For a given number of attribute values, the largest value of Split Information occurs when there is a uniform distribution of attribute frequencies.

For a given number of instances that are uniformly distributed, Split Information increases when the number of different attribute values increases.

The largest values of Split Information occur when there are many possible attribute values, all equally frequent.

Information Gain is generally largest when there are many possible attribute values. Dividing this value by Split Information to give Gain Ratio substantially reduces the bias towards selecting attributes with a large number of values.

6.6 Number of Rules Generated by Different Attribute Selection Criteria

Figure 6.7 repeats the results given in Figure 5.8, augmented by the results for Gain Ratio. The largest value for each dataset is given in bold and underlined.

Dataset	Excluding Entropy and Gain Ratio		Entropy	Gain Ratio
	most	least		
contact_lenses	42	26	**16**	17
lens24	21	**9**	**9**	**9**
chess	155	52	**20**	**20**
vote	116	40	34	**33**
monk1	89	53	**52**	**52**
monk2	142	109	**95**	96
monk3	77	43	28	**25**

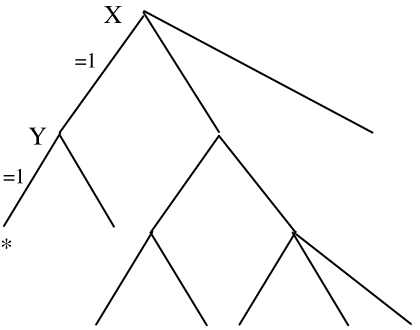
Figure 6.7 TDIDT with Various Attribute Selection Methods

For many datasets Information Gain (i.e. entropy reduction) and Gain Ratio give the same results. For others using Gain Ratio can give a significantly smaller decision tree. However, Figure 6.7 shows that neither Information Gain nor Gain Ratio invariably gives the smallest decision tree. This is in accord with the general result that no method of attribute selection is best for all possible datasets. In practice Information Gain is probably the most commonly used method, although the popularity of C4.5 makes Gain Ratio a strong contender.

6.7 Missing Branches

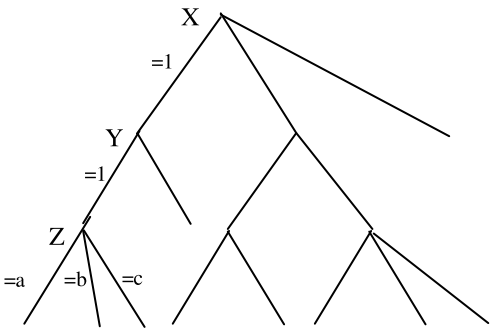
The phenomenon of missing branches can occur at any stage of decision tree generation but is more likely to occur lower down in the tree where the number of instances under consideration is smaller.

As an example, suppose that tree construction has reached the following stage (only some of the nodes and branches are labelled).



The left-most node (marked as $*$) corresponds to an incomplete rule
IF X =1 AND Y = 1 ...

Suppose that at $*$ it is decided to split on categorical attribute Z, which has four possible values a, b, c and d. Normally this would lead to four branches being created at that node, one for each of the possible categorical values. However it may be that for the instances being considered there (which may be only a small subset of the original training set) there are no cases where attribute Z has the value d. In that case only three branches would be generated, giving the following.



There is no branch for $Z = d$. This corresponds to an empty subset of instances where Z has that value. (The TDIDT algorithm states 'divide the instances into non-empty subsets'.)

This *missing branch* phenomenon occurs quite frequently but generally has little impact. Its drawback (if it is one) occurs when the tree is used to classify an unseen instance for which attributes X, Y and Z have the values 1, 1 and d respectively. In this case there will be no branches of the tree corresponding to the unseen instance and so none of the corresponding rules will fire and the instance will remain unclassified. This is not usually a significant problem as it may well be considered preferable to leave an unseen instance unclassified rather than to classify it wrongly. However it would be easy for a practical rule induction system to provide a facility for any unclassified instances to be given a default classification, say the largest class.

6.8 Chapter Summary

This chapter describes an alternative method of calculating the average entropy of the training (sub)sets resulting from splitting on an attribute, which uses frequency tables. It is shown to be equivalent to the method used in Chapter 5 but requires less computation. Two alternative attribute selection criteria, the Gini Index of Diversity and the χ^2 statistic, are illustrated and it is shown how they can also be calculated using a frequency table.

The important issue of inductive bias is introduced. This leads to a description of a further attribute selection criterion, Gain Ratio, which was introduced as a way of overcoming the bias of the entropy minimisation method, which is undesirable for some datasets.

6.9 Self-assessment Exercises for Chapter 6

1. Repeat Exercise 1 from Chapter 5 using the frequency table method of calculating entropy. Verify that the two methods give the same results.
2. When using the TDIDT algorithm, with the *degrees* dataset, find the attribute that will be chosen for the first split on the data using the Gain Ratio and Gini Index attribute selection strategies.
3. Suggest two datasets for which the Gain Ratio attribute selection strategy may be preferable to using entropy minimisation.

References

[1] Mingers, J. (1989). An empirical comparison of pruning methods for decision tree induction. *Machine Learning*, *4*, 227–243.

[2] Quinlan, J. R. (1993). *C4.5: programs for machine learning*. San Mateo: Morgan Kaufmann.