An information measure for hierarchic classification

D. M. Boulton and C. S. Wallace

Department of Information Science, Monash University, Clayton, Victoria, 3168, Australia

The information measure has been developed as a criterion of merit for intrinsic classifications. The information measure for non-hierarchic classifications has been described previously and a program developed which searches for that classification optimising the information measure. However, hierarchic classifications are often of practical importance and this paper develops the information measure for hierarchic classifications. Two algorithms are outlined for generating hierarchic classifications which minimise the information measure. One of these has been programmed and first tests show a good agreement with conventional taxonomy.

(Received June 1972)

1. Classification by information measure

The problem addressed is the partition of a population of S things into classes, where each thing is characterised by measurement values $x[d, s]$ on D different attributes ($s = 1, 2 \ldots S$, $d = 1, 2 \ldots D$). The aim is to produce classes such that things in the same class are sufficiently similar to be treated as equivalent for some purposes. This is the classic taxonomic problem, and has been attacked by many authors (e.g. Sokal and Sneath, 1963, Lance and Williams, 1967a, 1967b).

In an earlier paper (Wallace and Boulton, 1968), we suggested that progress on this problem could be accelerated if an attempt was made to define a criterion or figure of merit designed to measure the degree to which a particular classification of a population achieved its aims. As a possible criterion, we defined the 'information measure' of a classification.

If the measurement data is used to estimate for each class of a classification the within-class statistical distribution of attributes, the union of the estimated distribution functions can be regarded as a composite frequency distribution function covering the whole population. The function can then be used to estimate the probability, or relative frequency, of occurrence of things having particular attribute value sets. The probabilities thus found can then be used to establish an optimum Shannon-Fano code (Oliver, 1952) for encoding a message which conveys our knowledge of the data $x[d, s]$, in which the segment of message conveying the attributes of a particular thing has a length (in units of $\log_2 e$ bits, or nits) of minus the logarithm of the probability of occurrence of such a thing.

Such a message is unintelligible, as it stands, to anyone without prior knowledge of the classification and the composite distribution function. If we assume that the only prior knowledge available to the receiver is S, D and the nature and range of each attribute, the message must be augmented by a segment which (again using Shannon-Fano encoding) describes the number of classes, their relative abundances, and each within-class distribution function. The overall length of the resultant augmented message is defined as the 'information measure' of the classification.

We propose that the information measure (or rather, its negative) be used as a criterion of goodness of classifications, and have shown (Wallace and Boulton, 1968) that when the classification is chosen to minimise the information measure, the following properties can be asserted of the classification:

1. Each thing is assigned to that class which is most likely to contain such a thing.
2. The estimates of relative abundance and within-class distribution parameters are essentially maximum likelihood estimates.

It may also be asserted that if the known population is a sample of size S from an infinite population having T classes, and the within-class distributions of the infinite population conform to the assumptions made about them in the encoding process (e.g. that if the true within-class distributions are normal for continuous attributes, then normal distribution functions were used in the encoding) then for sufficiently large S the classification which minimises the information measure can be expected to exhibit T classes. This property of minimum information measure classifications follows from the optimal nature of Shannon-Fano codes.

In previous work, we defined the information measure only for non-hierarchic classifications, that is, classifications in which each class is defined and its properties stated independently of other classes. In this paper, we derive the information measures for two kinds of hierarchic classification, and discuss the computer strategies used to generate optimum hierarchic classifications.

In an hierarchic classification, the population is not partitioned into classes in one step. Rather, the population is first partitioned into a few broad classes. Each of these is further partitioned into smaller classes, and each of these further partitioned, and so on, until terminal classes are generated which are not further subdivided. Hierarchic classifications are more widely used in everyday life and scientific work than non-hierarchic classifications. Examples are the Dewey decimal library classification (which has the property that each partition is 10-way) and the biological hierarchy of phylum class, order, family, etc. (which has the property that a definite and fixed hierarchic depth, or number of successive partitions, is used).

Hierarchic classifications may be represented by inverted tree structures, or 'dendrograms' in which the root (highest) node represents the whole population, and the terminal 'twig' nodes represent the terminal classes.

2. Assumptions

In forming within-class attribute distribution functions, and in the encoding of message segments describing these functions, certain assumptions must be made. Following Wallace and Boulton (1968) we treat two kinds of attribute—continuous and multistate.

A continuous attribute d is one measured on a continuum scale, to a specified accuracy $\varepsilon[d]$, Within a terminal class t, a continuous attribute is assumed to have a normal distribution with mean $\mu[d, t]$ and standard deviation $\sigma[d, t]$.

A multistate attribute takes one or another of a finite set of values or 'states'. If attribute d is multistate and can have $M[d]$ states, labelled 1 to M, then its statistical distribution is described by the set $\{p[m, d, t]\}$ ($m = 1, 2 \ldots M[d]$) where

$p[m, d, t]$ is the probability that a thing in terminal class t will have attribute d in state m, i.e. the probability $(x[d, s] = m \mid s \in t)$.

Within a given terminal class, attribute values are assumed to be uncorrelated.

The above assumptions concern only the distribution of attribute values within a single terminal class, and do not concern the distribution of values for the population as a whole. The latter distribution will be expected to exhibit strong inter-attribute correlation, and, if significant class structure exists, will not be unimodel.

Within a single terminal class our assumed distribution for variables is the most general possible marginal distribution. The assumption of normal form for the within-class distribution of continuous attributes is a compromise which we hope to be adequate in most cases. Where good empirical or theoretic reason exists for expecting some other distribution function for a continuous attribute, the information measure can readily be appropriately modified. In some cases, for instance when the within-class distribution of some attribute is expected to be log-normal, a preliminary transformation of the attribute values will lead to transformed values expected to have a normal distribution. We believe the normal form to be the most colourless assumption we can make, having both a position and a scale parameter.

The assumption of no inter-attribute correlation within a single terminal class is unfortunate, in that there are many cases where we would expect to find significant correlations within a group of things which we would not wish to subdivide. For instance, all linear dimensioned attributes, such as length and height, of the members of a single species can be expected to have a strong positive correlation, reflecting their common dependence on an 'overall size' factor. It would be possible, but very time consuming, to incorporate interattribute correlations in the assumed distribution function for each terminal class. However, to do so would require a computation similar to a principal components analysis to be performed for each terminal class for each iteration of the analysis. An alternative expedient, which appears to suffice in the practical cases we have attempted, is to attempt to remove such correlations from the original data, e.g. by reducing a set of linear-dimensioned attributes to a single 'size' attribute and a set of dimensionless 'shape' attributes. Similar expedients are necessary in most classification methods.

It may be felt that by assuming a specific functional form for each within-class density function, we are being more restrictive than other classification techniques which make no such explicit assumption. However, other techniques which we have studied seem to rely for their rationale on tacit assumptions of a similar nature. For instance, the 'nearest neighbour' technique, based on some dissimilarity measure, is rational only if it is assumed that the probability that a member of a given class would be found to have certain attribute values is a decreasing function of the minimum of the dissimilarities between its attribute values and the values of all known members of the class, and that this function is the same for all classes. This is, in fact, quite a specific assumption about the expected form of class density distributions, and is possibly less consistent with the empirical distribution of accepted terminal taxonomic classes than is our assumption.

For the non-hierarchic case, we have shown (Wallace and Boulton, 1968) that the message specifying the class properties, i.e. the μ, σ and p values for a class, should not, for optimum encoding, specify these quantities too accurately. In fact, we show that a standard deviation is best quoted to an accuracy $\sigma\sqrt{6/(N-1)}$ where N is the number of things in the class, and that a mean μ is best quoted to an accuracy $\sigma\sqrt{12/N}$. The relative frequency of a state m of a multistate variable d is best quoted to an accuracy $\sqrt{12Np[m, d, t]}/N$, likewise the relative abundance of a class is best quoted to an accuracy $\sqrt{12N}/S$. The form for the resulting length of the message required to state the measurements $x[d, s]$, allowing for the inaccuracies of the encoding distribution, has been derived. It can be shown that in an hierarchic classification, the same optimum accuracies of quotation obtain.

It thus follows that the length of the part of the message which states the attribute values $x[d, s]$ is unaffected by a change from a non-hierarchic to an hierarchic classification. The hierarchic classification differs from the non-hierarchic in the way in which the terminal class properties are encoded. In a non-hierarchic classification, the properties of one class are described without reference to those of any other class. For instance, for a continuous attribute, each class's class mean is conceptually free to take any value within a range limited only by the known limits of the population as a whole, and must be described by a message segment sufficiently long to specify the class mean to the required accuracy within this range. However, if two terminal classes have similar properties, these properties may, by a suitable coding scheme, be more efficiently described by a message which first describes the union of the classes, and then indicates how they differ. It is this possibility which an hierarchic description seeks to exploit. We restrict all branchings of a tree to two-way branches for computational reasons.

Suppose that two terminal classes are to be described via a description of a non-terminal parent class which is their union. We must now consider what kind of description of the parent should be given.

The description of a terminal class comprises its relative abundance, the class mean and standard deviation of each continuous attribute, and the class probability for each state of each multistate attribute. Since the purpose of the description of the parent is to aid in the description of its two member classes, quantities described for the parent should be restricted to functions of the above properties of its member classes. In principle, any functions of these properties could be included in the description of the parent. However, we believe it to be consistent with the intent of an hierarchic classification to require that the description of a non-terminal class should be a function only of the attribute values of the things in that class, and should not assume or imply the terminal classification of the things.

The only functions of the terminal class properties which can be expressed in terms of the attribute values of things in the parent class, without knowledge of their terminal classification, are:

(*a*) the relative abundance of the parent class,

(*b*) the parent class mean and standard deviation of each continuous attribute

(*c*) the parent class probability of each state of each multistate attribute.

The description of a non-terminal class which is the union of two terminal classes therefore has exactly the same content as the description of a terminal class. The argument above may be extended to imply that the description of any non-terminal class at any level of the hierarchy should likewise have the same content as the description of a terminal class.

The accuracy to which the properties of non-terminal classes should be quoted for highest efficiency can be shown to follow the same rules as for terminal classes.

It should be noted that, although the same properties are described for non-terminal classes as for terminal classes, it is not assumed or implied that non-terminal classes have distribution functions conforming to the terminal class model. In particular, it is not assumed that the marginal distribution of a continuous attribute within a non-terminal class has normal

form. In fact, no use is made in the classification process of any assumed distribution function for non-terminal classes. A calculable density function is required only for terminal classes.

3. Hierarchic encoding

An hierarchic structure is, in our context, a scheme for the economic encoding of information about the terminal classes of a classification. Since each partition of our hierarchies is two-way, the encoding at all levels of the tree follows the pattern:

Given the properties of a class X to certain accuracy, state the relative abundance of two classes A and B which are the subclasses of X, and give their properties to the appropriate accuracy.

We will derive the lengths of the message segments required to achieve such a step in the encoding of the class properties in later sections. It is also necessary to encode at each partition the occurrence of that partition (i.e. the fact that the class being divided is non-terminal) and the relative abundances of the two sub-classes into which it is split. We now proceed to derive the length of this part of the message.

The relevant prior knowledge assumed is the total size S of the population, which we assume to be known exactly. The first partition (if it exists) will split the population into two subclasses, of sizes U and V. ($U + V = S$). The message which describes this partition will quote U and V only to limited accuracy. However, the accuracy is sufficient for the receiver of the message to infer that all members of the first subclass will be labelled as such with an identifiable message segment of length $\ln (S/U)$, and to deduce what this label is. Similarly, the receiver can deduce the form of the label used to denote membership of the second subclass. Thus, the receiver can, after decoding the message quoting U and V approximately, scan the labels attached to each thing in the population to determine the subclass to which each thing belongs. He can thus acquire exact knowledge of U and V. Thus, if one of these subclasses, say the first, is further subdivided into classes of sizes A and B, we again have an encoding problem for that partition where U, the size of the parent class, is known exactly, and A and B, the sizes of the subclasses, need be quoted only to limited accuracy.

We may further economise the message length requirement by adopting the convention that at each partition, the first-quoted subclass will be the one of smaller size. Thus, for each partition, we require for the description of the class sizes, the quotation of subclass size A to limited accuracy within a range 1 to $U/2$, where U is the exactly-known size of the parent class. In Wallace and Boulton (1968) the length, if A has a prior expectation uniformly distributed in the range 0 to $U/2$, is shown to be

$$\tfrac{1}{2}(\ln (U^3/12AB) + 1) - \ln 2 \qquad (3.0)$$

However if this range is limited to 1 to $U/2$ (as no class can have no members) (3.0) becomes

$$\tfrac{1}{2}(\ln (U(U-1)^2/12AB) + 1) - \ln 2 \qquad (3.1)$$

It is readily shown that the hierarchic structure itself, i.e. the tree structure, can be topologically described by the addition of one bit of information for each node save the root, given the prior expectation that an arbitrary class will have probability of one half of being terminal.

The resulting contribution to the message length due to the description of the splitting of one non-terminal class into two sub-classes is

$$\tfrac{1}{2}(\ln (U(U-1)^2/12AB) + 1) + \ln 2 \qquad (3.2)$$

This form covers the statement that the split occurs, and the relative abundance of the subclasses. We must now discuss the information needed to specify the subclass distribution functions, given the distribution functions of the parent class. Because no interattribute correlations are included in the assumed forms for the distribution functions, each attribute may be considered separately.

4. Hierarchic specification of continuous distributions

Suppose that at a particular partition of the hierarchic tree, a class A of N members is divided into two subclasses B and C of L and M members respectively. Suppose that for some continuous attribute, the description of A specified the class mean μ and class standard deviation σ to accuracies $\pm\frac{1}{2}\sigma\sqrt{12/N}$ and $\pm\frac{1}{2}\sigma\sqrt{6/(N-1)}$ respectively. Given this information, we have to describe for classes B and C their means a and b, and their standard deviations r and t, to the accuracies shown below:

$$a \pm \tfrac{1}{2} r\sqrt{12/L} \qquad r \pm \tfrac{1}{2} r\sqrt{6/(L-1)}$$
$$b \pm \tfrac{1}{2} t\sqrt{12/M} \qquad t \pm \tfrac{1}{2} t\sqrt{6/(M-1)}$$

We shall for simplicity in the following drop the -1 in the formulae specifying the standard deviation errors. The derivation below is unaffected in principle but algebraically much complicated by its retention.

Notice that since the accuracy of quotation of the means, and hence the encoding scheme used for means, depends on the standard deviations, the message must begin by stating r and t.

The optimum encoding of a, b, r and t, and hence the contribution to the information measure, depends on the prior expectation distributions assumed for them once μ and σ are known. We wish to make our prior assumptions as nearly colourless as possible, and seek to assume uniform prior distributions where possible.

Define

$$\rho = \sqrt{L}\, r, \quad \tau = \sqrt{M}\, t, \quad d = a - b \qquad (4.1)$$

$$\delta = \sqrt{LM/N}\, d \qquad (4.2)$$

Then we have from the ideal relation

$$Lr^2 + Mt^2 + LMd^2/N = N\sigma^2$$
$$\rho^2 + \tau^2 + \delta^2 = N(\sigma \pm \tfrac{1}{2}\sigma\sqrt{6/N})^2 \qquad (4.3)$$
$$= (R \pm \tfrac{1}{2}\sqrt{6}\,\sigma)^2$$

where

$$R = \sigma\sqrt{N}$$

Thus, our prior knowledge of σ, N, L and M restricts the possible sets (ρ, τ, δ) to lie in a $\frac{1}{4}$-sphere shell of radius R and thickness $\sigma\sqrt{6}$. The shell is a $\frac{1}{4}$-sphere because of the conditions $\rho \geq 0$, $\tau \geq 0$.

If we assume a uniform prior distribution of expectation for (ρ, τ, δ) sets within the shell, it can be shown that the corresponding marginal prior expectation functions for ρ, τ, and δ are each approximately uniform, the first two in the range $0 - R$ and the last in the range $-R$ to R.

We first find the information needed to specify ρ and τ simultaneously.

Projecting the uniform expectation throughout the shell onto the ρ, τ plane, we find a prior expectation density at a pair (ρ, τ) given by

$$P(\sigma, \tau) = 2/\pi R\delta \qquad (4.4)$$

Hence, to quote ρ to accuracy $\pm\frac{1}{2} r\sqrt{6}$ and τ to accuracy $\pm\frac{1}{2} t\sqrt{6}$ requires a message of length

$$\ln (\pi R\delta/12rt) \qquad (4.5)$$

Having specified r and t to the required accuracy, we have by implication specified δ through the equation (4.3), which can be rewritten

$$\delta^2 \doteq R^2 - \rho^2 - \tau^2 \qquad (4.6)$$

However, each term on the right hand side of (4.6) is known only approximately. The resulting uncertainty in δ can be estimated by assuming that the squared error in δ^2 is approximately the sum of the squared errors in R^2, ρ^2 and τ^2. The range of uncertainty in δ is found to be

$$\frac{\sqrt{6}}{\delta}\sqrt{(N\sigma^4 + Lr^4 + Mt^4)} \tag{4.7}$$

The resulting range of uncertainty in d is

$$\frac{1}{\delta}\sqrt{\frac{6N}{LM}(N\sigma^4 + Lr^4 + Mt^4)} \tag{4.8}$$

This gives us some information about the means a and b. We also know $\mu = \frac{1}{N}(La + Mb)$ to an accuracy $\frac{1}{2}\sigma\sqrt{12/N}$.

Since the Jacobian from μ, d space to a, b space has magnitude 1, our approximate knowledge of μ and d locates a and b in a, b space to within an area of uncertainty given by the product of the uncertainty ranges in d and μ, i.e.

$$\frac{\sigma}{\delta}\sqrt{72(N\sigma^4 + Lr^4 + Mt^4)/LM} \tag{4.9}$$

We require to specify a and b to within ranges of size $r\sqrt{12/L}$ and $t\sqrt{12/M}$ respectively, or to within an area of uncertainty

$$12\,rt\sqrt{1/LM} \tag{4.10}$$

This information required to specify a and b to the required accuracies is the logarithm of the ratio of 4.9 to 4.10, i.e.

$$\ln\left(\sqrt{1/2}\,\frac{\sigma}{rt\delta}\sqrt{(N\sigma^4 + Lr^4 + Mt^4)}\right) \tag{4.11}$$

Adding from (4.5) the information needed to specify r and t gives, for the total information needed to specify the subclass distribution parameters for this attribute:

$$\ln\left(\frac{\pi\sigma^2}{12\sqrt{2}\,r^2t^2}\sqrt{N(N\sigma^4 + Lr^4 + Mt^4)}\right) \tag{4.12}$$

However, the derivation, in which a and b were determined from d^2 and μ, leaves an ambiguity in the sign of d, requiring ln 2 nits to resolve. When this is added, the total message length is

$$\ln\left(\frac{\pi\sigma^2}{6\sqrt{2}\,r^2t^2}\sqrt{N(N\sigma^4 + Lr^4 + Mt^4)}\right) \tag{4.13}$$

The only significant assumption made in the above derivation is the assumption that, within the quarter-sphere shell of possible (ρ, τ, δ) values, all parts of the shell have equal probability density. The consequence of this assumption is that the coding scheme developed does not expect any greater similarity between the subclasses than can be strictly inferred from the description of their union, and hence cannot exploit a stronger similarity when one is found. Even if the subclasses happen, for the attribute considered, to have nearly equal means and standard deviations, the message required to specify these parameters is little if at all shorter than if they differed by the greatest amount possible consistent with the description of their union. It could be argued that in many applications of numerical taxonomy, subclasses of a parent class may reasonably be expected to resemble the parent class and each other very closely for most attributes, and that our coding scheme should therefore concentrate its prior expectation in that region of the quarter-sphere where the subclasses are most similar, i.e., near the point

$$\{\delta = 0,\ \rho = \sqrt{L}\,\sigma,\ \tau = \sqrt{\mu}\,\sigma\}\,.$$

Such a choice would have the effect that the advantages of hierarchic description would be very great for similar subclasses, but would disappear completely for strongly-differing subclasses. Our present choice is a more cautious approach, designed to get some modest benefit from even weak similarity between classes, and therefore not able to achieve dramatic reductions in the information measure even from strong similarities.

The best choice of prior expectation can be made only on the basis of empirical evidence. In any field of application, our expectations about the distribution of subclass properties given the properties of their parent should be based on the observation of the distributions exhibited by hierarchies known to be useful in that field. Unfortunately we have little or no quantitative information of this kind, and are forced at this stage to make the weakest possible assumptions.

5. Hierarchic specification of multistate distributions

Suppose that for some multistate attribute having Q possible states, the parent class was described as having n_i members in state i $(i = 1, 2 \ldots Q)$. Given this information, we have to describe the distribution of the multistate attribute in each of the two subclasses. Let the number of members of B having state i of the attribute be l_i and the number of members of C having this state be m_i. We wish to determine the length of message required to specify the set of values $\{l_i, m_i\}$ $(i = 1, 2 \ldots, Q)$ given prior knowledge of L, M, N and the set $\{n_i\}$.

Before developing this result, it will be helpful briefly to present a derivation of the message length required for the non-hierarchic description of the occupancy of a single multistate attribute within a single class. Let there be N things in the class, and let the number having state i of the attribute be n_i. Assume all possible sets of values for the n_i to be equally likely.

Were all n_i values and N specified exactly, we would have

$$\sum_{i=1}^{Q} n_i = N \tag{5.1}$$

The number of sets of non-negative values n_i obeying (5.1) is given by Y, where

$$Y = {}^{N+Q-1}C_{Q-1} \tag{5.2}$$

For $N \gg Q$, (5.3) is sufficiently approximated by

$$Y = N^{(Q-1)}/(Q-1)! \tag{5.3}$$

If all such sets are equally likely, and the n_i values were specified exactly, the information required would be simply ln Y. However, it can be shown that in the optimum encoding, each value n_i should be stated only to an absolute accuracy of order $\sqrt{n_i}$, thus lowering the information required by an amount $\frac{1}{2}\ln n_i$ for each state. However, the errors in quoting the n_i values result in (5.1) not being obeyed exactly. The expected error in (5.1) is of order $\sqrt{N}$. Hence, the number of possible sets of n_i values is increased from Y to approximately $Y\sqrt{N}$, thus increasing the information required to nominate one set by $\ln\sqrt{N}$. The total information required is hence approximately

$$\begin{aligned}&\ln(Y\sqrt{N}) - \sum_i(\tfrac{1}{2}\ln n_i)\\ &= (Q - \tfrac{1}{2})\ln N - \tfrac{1}{2}\sum_i \ln n_i - \ln(Q-1)!\end{aligned} \tag{5.4}$$

A more rigorous derivation, justifying the choice of accuracy in quoting n_i values, is given in Wallace and Boulton (1968).

We now apply the same reasoning to the hierarchic problem of specifying a set of l_i and m_i values, given the constraints

$$l_i + m_i = n_i \text{ (all } i) \tag{5.5}$$

$$\sum_i l_i = L \tag{5.6}$$

$$\sum_i m_i = M \tag{5.7}$$

Let the number of sets satisfying all constraints exactly be Y. Then the information to specify one such set exactly would be ln Y. However, each l_i or m_i value is to be specified to an absolute accuracy of order $\sqrt{l_i}$ or $\sqrt{m_i}$ respectively, thus reducing the information requirement by a total of $\frac{1}{2}\sum_i \ln(l_i m_i)$.

As a result of inaccuracies in stating l_i and m_i values, the constraint equations (5.5), (5.6) and (5.7) will be satisfied only to an absolute accuracy given, in each case, by the square root of the right-hand side. We thus expect the number of possible $\{l_i, m_i\}$ sets to be increased from Y to $Y\sqrt{L}\sqrt{M}(\prod_i \sqrt{n_i})$. However, the errors in equations (5.5), (5.6) and (5.7) are not independent, being constrained by the equation

$$L + M = \sum_i n_i$$

which is itself obeyed to an absolute accuracy of order $\sqrt{N}$. Thus the number of possible $\{l_i, M_i\}$ sets is increased by the errors in the constraints to

$$Y\sqrt{LM/N}\prod_i \sqrt{n_i}$$

To specify one of these sets, each l_i and m_i value being given to an accuracy of the order of its square root, requires a total message length given by

$$\ln\{Y\sqrt{LM/N}\prod_i \sqrt{n_i}\} - \tfrac{1}{2}\sum_i \ln(l_i m_i)$$

$$= \ln\left\{\prod_i \left(\frac{n_i}{l_i m_i}\right)\frac{LM}{N}\right\} + \ln Y \qquad (5.8)$$

A more exact analysis, following Wallace and Boulton (1968), modified (5.8) by a constant, giving

$$\tfrac{1}{2}\ln\left\{\frac{LM}{N}\prod_{i=1}^{Q}\frac{n_i}{l_i m_i}\right\} + \ln Y - \tfrac{1}{2}(Q-1)\ln 12$$

The determination of Y is, in general, difficult. For binary attributes ($Q = 2$), it is easily shown that

$$Y = 1 + \min(L, M, n_1, n_2)$$

For $Q > 2$, we have given exact algorithms for the calculation of Y in Boulton and Wallace (1973).

The only significant assumption in the above is that all Y sets of numbers are equally likely. This assumption is subject to comments essentially the same as those appearing at the end of Section 4. As is the case for continuous variables, the fact that the description of a non-terminal class involves, for a multistate attribute, the same kind of information as appears in the description of a terminal class does not mean that the distribution of things within a non-terminal class is modelled by the same kind of density function as is employed for terminal classes. In fact, no model is proposed or used for the distribution of a multistate attribute within a non-terminal class.

6. Describing the things

The main body of the message consists of the specification of the class membership of each thing and its D attribute values.

After receipt of the class description message the receiver will have knowledge of the composite distribution formed from the union of the T terminal class distributions. The relative frequency of a thing's set of D attribute values which is used for their encoding is estimated from the distribution of the terminal class to which the thing belongs. Hence the need to first specify the terminal class membership of each thing.

6.1 *The class membership*

The ultimate terminal class membership of each thing is specified by a series of subclass membership labels. The total length of the set of labels for things is given by:

$$\ln S/N_1 + \ln N_1/N_2 + \ldots \ln N_{t-1}/N_t = \ln S/N_t \quad (6.1)$$

where N_t is the population of the terminal class and $N_1 \ldots N_{t-1}$ the populations of the intermediate classes to which s belongs. The relative abundance of terminal class t is N_t/S and hence (6.1) is the optimum label length to specify the terminal class membership of thing s non-hierarchically.

Thus the length of message to specify the class membership of each thing depends only on the set of terminal class relative abundances and is independent of the sizes of the non-terminal classes. This fact is made use of in the agglomerative classification algorithm described later.

6.2 *Encoding values of attributes*

It is shown in Wallace and Boulton (1968), that for a terminal class of N members, the information required to specify the values possessed by its members for a continuous attribute is approximately

$$N(\ln(\sqrt{2\pi}\sigma/\varepsilon) + \tfrac{1}{2}) + \tfrac{1}{2} \qquad (6.2)$$

where σ is the usual unbiased estimate of standard deviation,

$$(N-1)\sigma^2 = \sum_i (x[i] - \mu)^2 \qquad (6.3)$$

and ε is the range of error inherent in each given measurement.

It is also shown that the information required to specify the values of the members for a multistate attribute is

$$\sum_{q=1}^{Q} n[q]\ln(N+Q)/(n[q]+1) + \tfrac{1}{2}(Q-1) \qquad (6.4)$$

where Q is the number of states and $n[q]$ is the number of things in the qth state.

To find the total information required for a particular class and its members, to give all the attribute values, we sum (6.2) over continuous attributes, and (6.4) over discrete attributes, and add $N \ln S/N$ for the class membership labels. This result can be summed over all terminal classes to give the total attribute message length, to which must then be added the information needed for the hierarchic encoding of the class properties and dendrogram structure.

7. Minimisation strategies

If the information measure is accepted as a criterion of merit for a classification, it is desirable to devise classifications which minimise the measure. No explicit method has been found for calculating the optimum classification. However, we have devised methods of making successive improvements to a classification by optimising in each step some variables of the classification while holding others constant. The steps are listed below.

7.1. *Distribution adjustment*

With fixed class membership and hierarchic structure, the information measure is minimised with respect to the within-class distribution parameters of terminal classes by setting each of these to values which are essentially maximum likelihood estimates based on the measurements of the class members. Thus, the mean of a continuous attribute distribution is set equal to the mean of the measurements of the class members, and so on. These optima are in fact assumed in formulae (6.1), (6.2) and (6.4).

7.2 *Reclassifying*

With fixed class properties and hierarchic structure, the measure is minimised with respect to the classification of each thing if each of the S things is assigned to that terminal class within which its measurements may be most economically encoded. This choice is equivalent to assigning each thing to the terminal

class most likely to contain such a thing, i.e. to the terminal class whose density distribution function is greatest in the neighbourhood of the thing.

It can be shown that the optimum choices in both steps 7.1 and 7.2 are independent of the hierarchic structure. Thus, a repeated cycle in which reclassifying and distribution adjustment alternate can be used to bring the classification to at least a local optimum, which cannot be improved by either of steps 7.1 or 7.2. Other steps must be employed to gain further improvement, and in particular, to vary the number of terminal classes.

7.3 *Class merging*

Given two terminal classes, we may ask whether or not the classification will be improved, i.e. the information measure reduced, by combining the two classes into one. Combining two classes into one will inevitably (unless the classes have identical properties) increase the message length needed to specify the attribute values of their members. However, it will also reduce the amount of class description information required.

Let us assume for the moment that, when this question is asked, an hierarchic description of the classification has been set up, and that the terminal class description information has been optimised by distribution adjustment. Since it will presumably be advantageous to combine two terminal classes only if they are very similar, we will consider combining two terminal classes only if they are immediate subclasses of the same non-terminal class.

Let the non-terminal class be A, of size N, and let its terminal subclasses be B and C of sizes L and M respectively where $N = L + M$. We can work out I_B and I_C, the message lengths required to give the attributes of the members of classes B and C, and can also work out I_A, the message length that would be required for specifying the attributes of all N things were classes B and C combined into a single class A, which would then be terminal.

We also work out $I(B, C|A)$, the information needed to specify the existence, relative sizes, and properties of classes B and C given the properties of the non-terminal class A.

Then it is advantageous to combine classes B and C into the single terminal class A if

$$I_B + I_C + I(B, C|A) > I_A \qquad (7.1)$$

Terms I_B, I_C, and I_A depend only on the class properties and the given measurements of the things. They are thus independent of the hierarchic structure and of the properties and members of classes other than A.

Moreover, our choice of hierarchic encoding methods for class properties, as summarised in formulae (3.2), (4.13) and (5.7), leads to a form for $I(B, C|A)$ which is also independent of the hierarchic structure and classes other than A. Thus, the choice as to whether two terminal classes which are immediate subclasses of the same non-terminal class should be combined can be made without reference to other classes.

Suppose now that the hierarchic structure had not yet been established, and that we had only a set of terminal classes whose properties had been optimised by distribution adjustment. We may consider all pairs of terminal classes, and for each pair, evaluate the criterion (7.1). If this condition is satisfied, it implies that the information measure will favour combining the pair into a single class, rather than describing them as subclasses of a single non-terminal class. It does not indicate whether some other arrangement is preferable in which the classes are preserved as separate, but not described as immediate subclasses of the same non-terminal class.

However, if nonetheless we adopt (7.1) as a rule for deciding for any pair of terminal classes, whether they should be combined, we find that there is only one case in which we are led astray. This case requires that the optimum choice is separate classes not immediate subclasses of the same class, but that combining the classes is preferable to having them as separate immediate subclasses of the same class. This case represents a quite remote possibility, as it requires that the classes be sufficiently similar so that I_A does not much exceed $I_B + I_C$, yet sufficiently dissimilar that other classes should be associated with classes B and C at the lowest hierarchic level. We therefore adopt (7.1) as a decision rule for combining classes even when the hierarchic structure is unknown.

7.4 *Half-classes*

A rule for combining classes, and so reducing the number of terminal classes, was outlined above. We also need to incorporate in the minimisation strategy a way of increasing the number of terminal classes. For this purpose, we introduce 'half-classes', which are subclasses of the terminal classes, two per terminal class.

Class Description parameters are maintained for half-classes as for terminal classes. Whenever, during a reclassify step, a thing is assigned to a terminal class, it is also assigned to one or the other of that terminal class's half-classes. The parameters of half-classes are optimised during each distribution adjustment step. Normally, we would expect that if criterion (7.1) were applied to the half-classes of a terminal class, it would indicate that their separate existence was not justified. However, if it is found that (7.1) is not satisfied, the half-classes can be taken as proper terminal classes thenceforth. They are then themselves given half-classes, initially by a random partition of their members.

8. The agglomerative strategy

A program has been written in ALGOL to generate minimum information hierarchic classifications using the optimisation techniques described above. It is an iterative process, which seeks to improve upon a given, or initially random, classification. An iteration commences with some number of terminal classes, with known class properties. One or more cycles are then performed which improve these classes by alternate reclassify and distribution adjustment steps. Half-classes are maintained and improved for each terminal class during these cycles.

After these cycles have been performed, the criterion (7.1) is applied to discover advantageous class combining operations. However, it is applied to half-classes rather than to terminal classes. All half-class pairs are considered, whether or not they belong to the same class. Normally, half-classes of the same original terminal class will be recombined, but the opportunity exists for new terminal classes to be created.

The combinations are carried out in decreasing order of advantage. Whenever two classes are combined into one, the new one also becomes a candidate for further combinations. This process stops when no beneficial combinations can be found. The iteration is then complete. Notice that the iteration does not require the erection of the hierarchic dendrogram.

The dendrogram is established by a direct non-iterative agglomerative technique. Each pair of terminal classes is considered and that pair (B and C, say) found having the minimum value of $I(B, C|A)$ where A is defined as the union of B and C. B and C are then expressed as subclasses of a new, non-terminal class A, and eliminated from further treatment, their place being taken by A. The new set of classes, with B and C replaced by A, is again examined for the lowest-cost pair, which is replaced by a non-terminal class, and so on until all classes have been brought into a single dendrogram.

Actually, the program incorporates an additional feature in that if at any stage of the agglomerative process it is found that the classes remaining can be more economically described on a non-hierarchic basis than by further upward growth of the

dendrogram, the former choice is made. Thus, the final result may consist of a number of unlinked dendrograms. The agglomerative process thus includes non-hierarchic classifications as a subset of hierarchic ones, and will generate a non-hierarchic, or partially, or a fully hierarchic classification, according to the dictates of the information measure.

9. A divisive strategy

The agglomerative process above produces a dendrogram in which all classes, terminal and non-terminal, are described in basically the same way. Although this is convenient in many ways, it makes the construction of a simple identification rule difficult.

If terminal classes B and C are subclasses of A, and terminal classes Y and Z are subclasses of X, no simple rule based on the properties of A and X can necessarily determine whether a thing belongs to class A or to class X. The described properties of the non-terminal classes A and X do not define the distribution density function within these classes, since no functional form for the distribution density within a non-terminal class is assumed in the analysis. Hence, it is not possible to compute the boundary between A and X, which is defined by points of equal A and X density. Hence identification of the non-terminal class membership of a new thing cannot be made by using only the non-terminal class properties. Instead, its terminal class must be found, using the terminal class properties and the assumed functional form for distribution density functions within terminal classes.

However, it may be that in practice, few things will be misclassified at a high hierarchic level if the boundaries between non-terminal classes are approximated by assuming a density distribution function within a non-terminal class of the same form as is assumed for terminal classes. Even though this assumption is clearly incorrect in general, the differences between high-level classes are expected to be more pronounced than the differences between some terminal classes, so that even an inaccurate model of the density distribution within non-terminal classes may suffice to identify most of their member things.

We have not attempted to verify the above conjecture, but, if it is close to the truth, a different classification strategy may be employed.

In this strategy, one starts with a single class comprising the whole population. By optimisation of its half-classes, this may be split into subclasses. These are then given half-classes, which, after further optimisation, may be further split, and so on, until a stage is reached when no class has half-classes for which the combine criterion (7.1) does not hold.

Note that in this process, once a class has been created, its properties and membership are not subsequently altered.

This divisive strategy will yield a 'decision tree' type of identification rule, exactly paralleling the structure of the dendrogram. However, it will not in general have the property that each thing is assigned to that terminal class most likely to contain such a thing. Thus, its information measure will in general be greater than that of an agglomerative classification of the same population, and a non-hierarchic identification rule cannot be used.

Moreover, it is open to the theoretical objection that it embodies the clearly inconsistent assumption that classes at all hierarchic levels have density distributions of the same functional form. This objection, incidentally, can be raised against most numerical taxonomy techniques, with the notable exceptions of the nearest-neighbour strategy and the agglomerative method described in this paper.

Apart from its possibly greater convenience, the divisive strategy may be preferable on theoretical grounds in populations such as some biological ones where the dendrogram may be considered a model of an actual generative process of the population, and where differentiation of a class into subclasses proceeds independently of the differentiation of other classes.

Results

The agglomerative algorithm for minimising the information measure is the only one which has, as yet, been programmed. Unfortunately testing is made difficult because, apart from our own efforts, there does not appear to be any objective means for testing a classification. However, the program has been used to classify 99 representatives of nine accepted species of the genus Pediastrum.

The data consisted of 14 attributes: five binary, eight multistate and one continuous measure. The species content of the sample is given in **Table 1.**

The data was collected from prepared slides by second year botany students during a laboratory class at Monash University. Record was kept for each plant by the student who measured it but neither this information nor the species of each plant was input to the program.

Table 1 Species content of sample

SPECIES NUMBER	SPECIES NAME	NUMBERS PRESENT
1	P. Biradiatum	14
2	P. Tetras	8
3	P. Sp. LB114	14
4	P. Duplex	7
5	P. Boryanum v Boryanum	14
6	P. Clathratum	14
7	P. Angulosum	7
8	P. Boryanum v Longicorne	7
9	P. Simplex	14
	Total	99

Table 2 Contingency table showing comparison between the nine species and the seven program produced terminal classes

	Program Classes							
	1	2	3	4	5	6	7	
1	14							14
2		8						8
3			14					14
4	7							7
5					14			14
6						14		14
7							7	7
8							7	7
9				7		7		14
	21	8	14	7	14	21	14	99

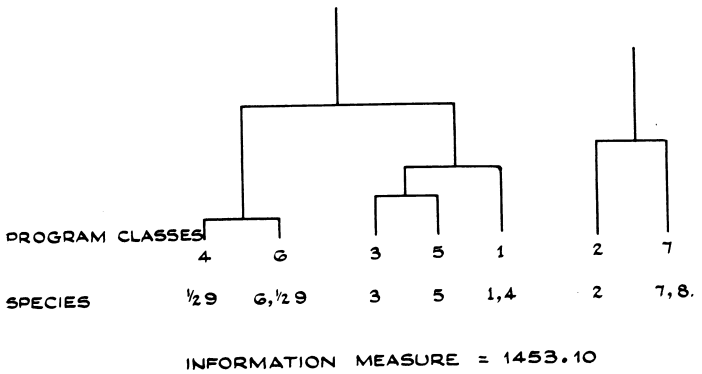

Fig. 1 Dendrogram for best classification.

The best solution obtained by the program comprised seven terminal classes. These are compared with the classification into nine species in **Table 2,** The best dendrogram is shown in **Fig. 1.**

This example has actually yielded two unlinked dendrograms. The two classes (4, 6) and (1, 2, 3, 5, 7) are so different that their parameters are more economically specified as two independent classes rather than by an hierarchical specification of the parameters of their union. An inspection of their properties verifies a marked difference between them.

We have also classified this data using our non-hierarchical program (Boulton and Wallace, 1970). The best classification it found contained six classes, five of them being identical to hierarchical terminal classes 1, 2, 3, 5 and 7, and the other identical with the union of terminal classes 4 and 6. When the non-hierarchical program was forced to limit the number of classes to three the best solution corresponded to the union of hierarchical terminal classes (4 and 6), (2 and 7) and (1, 3 and 5) as occurs in the dendrogram. Thus the hierarchical information measure appears consistent with the non-hierarchical measure.

The hierarchic analysis of the above data was performed by an ALGOL program running on a B5500 computer. It required nine iterations and took $3\frac{1}{2}$ minutes.

Acknowledgements

This work was carried out in the Department of Information Science, Monash University. During this time one of us (D. M. Boulton) was in receipt of a Commonwealth Post Graduate Scholarship. We are grateful to Dr. G. Scott of the Monash University Botany Department and to his students for supplying the Pediastrum data.

References

BOULTON, D. M. and WALLACE, C. S. (1970). A Program for Numerical Classification, *The Computer Journal*, Vol. 13, No. 1, p. 63.

BOULTON, D. M. and WALLACE, C. S. (1973). Occupancy of a Rectangular Array, *The Computer Journal* Vol. 16, No. 1, p. 57

LANCE, G. N. and WILLIAMS, W. T. (1967a). A general Theory of Classificatory Sorting Strategies, I. Hierarchical Systems, *The Computer Journal*, Vol. 9, No. 4, p. 373.

LANCE, G. N. and WILLIAMS, W. T. (1967b). A General Theory of Classificatory Sorting Strategies, II. Clustering Systems, *The Computer Journal*, Vol. 10, No. 3, p. 271.

OLIVER, B. M. (1952). Efficient Coding, *Bell Sys. Techn. J.*, Vol. 31, p. 724.

WALLACE, C. S. and BOULTON, D. M. (1968). An information measure for classification, *The Computer Journal*, Vol. 11, No. 2, p. 185.

Book reviews

Computer Applications of Numerical Methods, by Shan S. Kuo, 1972; xii + 415 pages. (*Addison-Wesley*, £5·75)

This is a revised version of a book first published in 1965. It gives an introduction to FORTRAN programming and numerical methods, with numerous examples. The first five chapters are devoted to computers, flowcharts, floating-point arithmetic, and programming. Some of the information is specific to IBM machines, and much of it is IBM-oriented, so that the reader would feel most at home if he worked in a System 370 installation. The description of the FORTRAN language is informal and is given mainly through examples; the majority of users would need a reference manual as well if they were actually writing programs.

The numerical methods section, which occupies about two-thirds of the book, discusses non-linear equations, initial-value problems in ordinary differential equations, linear equations and eigenvalues, interpolation, curve-fitting and quadrature, the Monte Carlo method, and linear programming. Some of the sections probably date from the first edition, and should have been revised, e.g. the Runge-Kutta-Gill method (page 142), Jacobi's method for eigenvalues (page 217), an interpolation method for Chebyshev curve-fitting (page 267). The presentation is rather uneven; some sections give a good background to the method described, leading up to a useful example, while others are rather sketchy, and plunge the reader into a long and complicated program. It seems to me that much of the space used for programs would have been better given to more practical discussion of the methods, including reasons for choosing one method rather than another, and common pitfalls.

The unevenness of the book makes it rather unsuitable for a class text, and a number of better books are available in this field for reference.

JOAN WALSH (Manchester)

Introduction to Computational Methods for Students of Calculus, by S. S. McNeary, 1972; 196 pages. (*Prentice-Hall International*, £4·25)

The preface to this book tells one that it is neither intended as a text in programming (yet the first sixth is a résumé of FORTRAN), nor as a text in numerical analysis. As an introduction to other books it is far too expensive, so I was left wondering just what this book is to do and for whom.

The material covered, besides FORTRAN, are formula evaluation, convergent sequences, errors, solution of equations, linear equations, polynomial approximation and numerical integration. The treatment throughout is elementary, assuming rather less than 'A-level' mathematics and, indeed, not going beyond a modern mathematics 'A-level' syllabus.

The references are entirely to American text books and among the list of 'Periodicals oriented toward computation' there are none published outside the USA.

P. A. SAMET (London)