An information measure for classification

By C. S. Wallace* and D. M. Boulton*

This paper derives a measure of the goodness of a classification based on information theory. A classification is regarded as a method of economical statistical encoding of the available attribute information.

The measure may be used to compare the relative goodness of classifications produced by different methods or as the basis of a classification procedure.

A classification program, 'SNOB', has been written for the University of Sydney KDF 9 computer, and first tests show good agreement with conventional taxonomy.

(First received December 1967)

1. Introduction

In all fields of discourse, the basic objects of concern are classified, and names given to the classes, to enable us to make general statements whose meaning applies to many objects rather than to a single object. For such a classification to be useful, the objects within a single class must essentially be equivalent at some level of discourse. The problem of generating a useful classification, exemplified by taxonomy, may be stated as follows:

Given a set of S things and for each a set of D measurements (attributes), to form a partition of the set of things, or, equivalently, a partition of the D-dimensioned measurement space within which each thing may be represented by a point, such that the things within each subset, or region of measurement space, may usefully be treated as equivalent in some discussion.

Many classification processes have been devised in answer to this problem (Sokal and Sneath, 1963; Williams and Dale, 1965). These methods have usually been directed towards producing classes such that members of the same class are as 'similar' as possible and/or members of different classes are as 'dissimilar' as possible. Such aims, while not necessarily equivalent to the general aim described above, can obviously be expected in practice to produce classifications which well serve the general aim. Unfortunately, the different measures of similarity between things and between classes of things which have been used in these processes result in significantly different classifications, and it is usually left to the user to choose that method which produces the most useful result. Moreover, it is difficult in many of these processes to separate a measure of the success of a classification from the process used to generate it. There is no readily applicable objective criterion firmly based on the original aim of the classification which can be used to compare the relative success of different processes.

The aim in this paper is to propose a measure of the goodness of a classification, based on information theory, which is completely independent of the process used to generate the classification.

2. The information measure

A classification may be regarded as a method of representing more briefly the information contained in the $S \times D$ attribute measurements.

These measurements contain a certain amount of information which without classification can be recorded directly as S lists of the D attribute values. If the things are now classified then the measurements can be recorded by listing the following:

1. The class to which each thing belongs.
2. The average properties of each class.
3. The deviations of each thing from the average properties of its parent class.

If the things are found to be concentrated in a small area of the region of each class in the measurement space then the deviations will be small, and with reference to the average class properties most of the information about a thing is given by naming the class to which it belongs. In this case the information may be recorded much more briefly than if a classification had not been used. We suggest that the best classification is that which results in the briefest recording of all the attribute information.

In this context, we will regard the measurements of each thing as being a message about that thing. Shannon (1948) showed that where messages may be regarded as each nominating the occurrence of a particular event among a universe of possible events, the information needed to record a series of such messages is minimised if the messages are encoded so that the length of each message is proportional to minus the logarithm of the relative frequency of occurrence of the event which it nominates. The information required is greatest when all frequencies are equal.

The messages here nominate the positions in measurement space of the S points representing the attributes of the things. If the expected density of points in the measurement space is everywhere uniform, the positions of the points cannot be encoded more briefly than by a simple list of the measured values. However, if the expected density is markedly non-uniform, application

* *Basser Computing Department, School of Physics, University of Sydney, Sydney, Australia.*

of Shannon's theorem will allow a reduction in the total message length by using a brief encoding for positions in those regions of high expected density.

If we consider the universe of all possible sets of D experimental measurements, we can form no reasonable expectation of the expected density in measurement space. However, if the observed distribution of the S given points is markedly non-uniform, the economics of Shannon's theorem can be realised in part by adding to the S messages a further message which approximately describes the average density distribution of the S points, and using this approximate non-uniform distribution as the basis for encoding the attribute messages. The message length needed to describe the average density distribution must be considered as adding to the lengths of the messages now used to encode the attributes of the things, because the receiver of the composite message has no *a priori* knowledge of the distribution.

The greater the detail given about the distribution, the more efficiently may the attributes of individual things be encoded. The 'best' classification will result from an optimum compromise between the efficiency of attribute encoding and the length of message needed to specify the distribution.

We regard a classification as one means, among many, of stating and exploiting a non-uniform density distribution in measurement space. For instance, factor analysis may perhaps be regarded as a technique for describing and exploiting a concentration of the points near a flat subspace. A classification sets up a model of the distribution in which the space is divided, not necessarily by planes, into regions (classes), the distribution within each region being described by a separate, simple, distribution function centred on an area of high density.

A thing within a particular region has its attributes encoded on the basis of the density, in its neighbourhood, of the distribution function for that region. That is, it is regarded as a member of the 'class' corresponding to the region, and the message describing it is encoded on the basis of the expected distribution of attribute values for members of that class, and contains the name of that class.

The $S \times D$ measurements are thus encoded into a message having five component parts:

1. the number of classes;
2. a dictionary of class names;
3. a description of the distribution function for each class;
4. for each thing, the name of the class to which it belongs;
5. for each thing, its attribute values in the code set up for its class.

We assume already known, and not included in the message, the following:

(*a*) The number of things S.
(*b*) The number of attributes D.
(*c*) The nature (discrete, continuous, etc.) of each attribute.

We now proceed to discuss the form of the component messages 1 to 5 above.

3. Class identification format

3.1 Number of classes

We prefer to make no assumption about the expected frequency of different numbers of classes, and will therefore encode the number of classes on the basis that all values with a range of say 1 to 100 are equally likely. The message length is therefore constant and this component will henceforth be ignored.

3.2 Dictionary of class names

In the message about each thing is included a class name or label to identify the density distribution used to encode its attributes. To keep the total length of these labels as short as possible, the label for each class will be chosen with a length dependent on the relative abundance of that class. The labels are therefore not known *a priori* and must be described in a dictionary.

If there are $n[t]$ members of class t ($t = 1, 2, \ldots, T$) then the label used in the description of a thing to say that it belongs to class t will occur $n[t]$ times in the total message. If we estimate the relative frequency of class t as $p[t]\left(\sum_{t=1}^{T} p[t] = 1\right)$ then the information needed to quote the class membership of all things is

$$-\sum_{t=1}^{T} n[t] \ln p[t]. \qquad (1)$$

The choice of $p[t]$ values which minimises (1) is

$$p[t] = n[t]/S.$$

However, the $p[t]$ are not known *a priori* and hence neither are the labels. It is shown in the appendix that when the information needed to describe the labels (class name dictionary) is taken into account, the overall information cost is minimised by tolerating an error of order $1/\sqrt{S}$ in a choice of $p[t]$, thus reducing the information needed for the class name dictionary, but causing the value of (1) to be on average slightly above its optimum. With the optimum choice of accuracy in selecting the $p[t]$, the information needed by the class name dictionary is found in the appendix to be approximately

$$\left(\frac{T-1}{2}\right) \ln \frac{S}{12} - \ln (T-1)! - \tfrac{1}{2} \sum_{t=1}^{T} \ln p[t]. \qquad (1a)$$

The expected additional cost due to the non-optimum selection of $p[t]$ is shown in equations (22) and (23) of the appendix to be about

$$(T-1)/2. \qquad (1b)$$

The total length of class identification labels and class name dictionary is given from (1), (1*a*) and (1*b*) as

$$\frac{T-1}{2}(\ln (S/12) + 1) - \ln (T-1)! - \sum_{t=1}^{T} (n[t] + \tfrac{1}{2}) \ln (p[t]) \qquad (2)$$

We estimate $p[t]$ by $n[t]/S$.

4. The class distribution model

We here treat only two kinds of attribute, continuous measurements (e.g. length) and unordered multistate (e.g. colour or sex). The treatment of ordered multistate data will be deferred to another communication, as it appears to warrant rather more discussion than is appropriate here.

In common with most workers in the field, we presume that within a single class, attributes are uncorrelated. The measure can, however, be extended to accommodate correlations, and we hope in another paper to show that considerable advantages might follow from such an extension.

Having assumed no correlation, the distribution within a class can be separated into independent distributions for each attribute. In calculating the information required for recording attribute values, we assume for simplicity that missing values are noted by a message element of fixed length, independent of the classification. Their contribution to the total message length can then be ignored.

4.1 Multistate attributes

For each class, the distribution of a multistate attribute is specified by the relative probability of occurrence of each state within the class. A label is created to nominate each state of the attribute, and the value of this attribute is specified for members of this class by including in their attribute messages the label nominating the appropriate state. Different labels will in general be used in different classes to nominate the same state of the same attribute.

In order that the labels used within a class may be recognised, the description of the class itself must include a description of the labels used for each state.

The considerations applying to the choice of state labels and their description are exactly analogous to those applying to the choice of class labels and the class name dictionary. However, if the number of things in class t having value m for multistate attribute d is $n[m, d, t]$, the relative frequency of occurrence of state m of attribute d in class t will be estimated by

$$p[m, d, t] = (n[m, d, t] + 1)/(n[d, t] + M[d]) \qquad (3)$$

where $M[d]$ is the number of states possible for attribute d, and $n[d, t]$ is the number of members of class t having known values for attribute d, i.e.

$$\sum_{m=1}^{M[d]} n[m, d, t].$$

If some values are missing, $n[d, t] < n[t]$.

The slight bias towards $(1/M[d])$ introduced by (3) into the estimate of $p[m, d, t]$ has the useful effect that newly-measured things not in the original sample S are not absolutely prevented from being assigned to class t by having a value for attribute d not possessed by any existing member of the class. Also, our approximate estimate of total message length (eqn. 4 below) will diverge if any $p[m, d, t]$ is esimated as zero. Biased estimates of the form (3) are discussed by Good (1965).

The total information needed to encode values of attribute d in class t, including the description of the labels used within the class, is given by the expression, analogous to (2):

$$\left(\frac{M[d]-1}{2}\right)(\ln(n[d, t]/12) + 1) - \ln(M[d]-1)! - \sum_{m=1}^{M[d]} (n[m, d, t] + \tfrac{1}{2}) \ln p[m, d, t]. \qquad (4)$$

4.2 Continuous attributes

We assume that within a class the values of a continuous attribute d are normally distributed with mean $\mu[d, t]$ and standard deviation $\sigma[d, t]$. Other forms of distribution could be used without any difficulty, in principle. In fact, the classification could allow a selection of forms, in which case further class description messages would be needed to specify the form employed.

To simplify the following discussion, d and t subscripts will be dropped, as we will be referring for the rest of Section 4.2 to the distribution of a single continuous attribute within a single class.

It is assumed that the range of measurement for the population as a whole is known *a priori* as is the accuracy of measurement. If a measurement $x[s]$ is quoted to a least count ϵ, i.e. to an accuracy of $\pm\frac{1}{2}\epsilon$, then the probability of getting such a measurement from the distribution (μ, σ) is approximately

$$\frac{K\epsilon}{\sigma} \exp(-(x[s] - \mu)^2/2\sigma^2) \qquad (5)$$

where $K = 1/\sqrt{(2\pi)}$.

The information used to encode the $n = n[t]$ such measurements is therefore

$$\sum_{s \text{ in } t} [\ln(\sigma/K\epsilon) + (x[s] - \mu)^2/2\sigma^2]. \qquad (6)$$

Here, the values μ and σ are the values quoted in the description of the class and used as the basis for the encoding of the measurements, and will not in general be fully accurate, because quotation to full accuracy would require an excessive amount of information.

Suppose that the nominal mean μ is quoted to a least count of b, and hence does not exactly equal the true mean $m = \frac{1}{n}\sum_{s=1}^{n} (x[s])$. Then if the mean is assumed to be in a range of size a, but no *a priori* assumption of its probability distribution within this range is made, the information required to encode μ is

$$\ln \frac{a}{b}. \qquad (7)$$

Writing $f = \mu - m$, f may be assumed to be uniformly distributed over a range $-b/2$ to $+b/2$, and

$$\text{Expectation } f^2 = \frac{1}{3}\left(\frac{b}{2}\right)^2 = \frac{1}{12} b^2. \qquad (8)$$

Likewise, suppose that the nominal standard deviation σ is quoted to a least count of c within a range 0 to p. The information required for its encoding is

$$\ln \frac{p}{c}. \tag{9}$$

The total information requirement for dimension d of type t is found by summing (6), (7) and (9).

$$\begin{aligned} I &= \ln \frac{p}{c} + \ln \frac{a}{b} + \sum_{s=1}^{n} \left(\ln \frac{\sigma}{K\epsilon} + (x_s - \mu)^2/2\sigma^2 \right) \\ &= \ln \frac{pa}{cb} + n \ln \frac{\sigma}{K\epsilon} + \sum_{s=1}^{n} (xt - m - f)^2/2\sigma^2 \\ &= \ln \frac{pa}{cb} + n \ln \frac{\sigma}{K\epsilon} + (nz^2 + nf^2)/2\sigma^2 \end{aligned} \tag{10}$$

where z is the true sample standard deviation, and $nz^2 = \sum_{s=1}^{n} (x[s] - m)^2$.

We now find the optimum value of b.

The expectation value $E(I)$ of I may be found by replacing f^2 by its expectation $b^2/12$. The parts of $E(I)$ dependent on b are then found from (10) to be

$$- \ln b + nb^2/24\sigma^2. \tag{11}$$

Differentiating (11) with respect to b gives for the best b

$$b = \sigma\sqrt{(12/n)}. \tag{12}$$

Substituting (12) in (10) and putting $f^2 = b^2/12 = \sigma^2/n$ gives

$$E(I) = \ln \frac{p}{c} + \ln \frac{a}{\sigma}\sqrt{\left(\frac{n}{12}\right)} + n \ln \frac{\sigma}{K\epsilon} + nz^2/2\sigma^2 + \tfrac{1}{2}. \tag{13}$$

We now determine the optimum value of σ.

The parts of (13) dependent on σ are

$$- \ln \sigma + n \ln \sigma + nz^2/2\sigma^2 = (n - 1) \ln \sigma + nz^2/2\sigma^2. \tag{14}$$

Differentiating (14) with respect to σ gives for the best σ

$$\sigma = z\sqrt{[n/(n - 1)]} \tag{15}$$

Now σ is quoted to a least count c, and hence will in general differ from the value given by (15). Write

$$\sigma = z\sqrt{[n/(n - 1)]} + g \tag{16}$$

where $|g| < c/2$, and the expectation of g^2 is $c^2/12$.

We now determine the optimum value of c. The parts of (13) depending on c or σ are

$$- \ln c + (n - 1) \ln \sigma + nz^2/2\sigma^2. \tag{17}$$

Before substituting (16) in (17) it will be simpler to rewrite (16) as

$$\sigma = w(1 + h) \tag{18}$$

where
$$w = z\sqrt{[n/(n - 1)]} \tag{19}$$

and
$$h = g/w. \tag{20}$$

Substituting (18) in (17) gives

$$- \ln c + (n - 1) \ln w + (n - 1) \ln (1 + h) + nz^2/2w^2(1 + h)^2. \tag{21}$$

The parts of (21) dependent on the choice of c are

$$- \ln c + (n - 1) \ln (1 + h) + nz^2(1 + h)^{-2}/2w^2. \tag{22}$$

Since h, the fractional error in quoting σ, is expected to be small, we expand (22) in powers of h up to h^2, to give

$$\begin{aligned} &- \ln c + (n - 1)(h - h^2/2) + nz^2(1 - 2h + 3h^2)/2w^2 \ldots \\ &= - \ln c + h[(n - 1) - nz^2/w^2] + h^2 [- (n - 1)/2 + 3nz^2/2w^2]. \end{aligned} \tag{23}$$

From (19), we have $nz^2/w^2 = (n - 1)$, and so (23) becomes

$$\begin{aligned} &- \ln c + h^2[- (n - 1)/2 + 3(n - 1)/2] \\ &= - \ln c + h^2(n - 1). \end{aligned} \tag{24}$$

Now, $h = g/w$ and the expectation of g^2 is $c^2/12$. The expectation of (24) is therefore

$$- \ln c + (n - 1)c^2/12w^2. \tag{25}$$

Differentiating (25) with respect to c gives for the best c

$$c = w\sqrt{[6/(n - 1)]}. \tag{26}$$

With this choice of c, the expected value of h^2 is

$$c^2/12w^2 = \tfrac{1}{2}(n - 1). \tag{27}$$

The expected value for I with the optimum choice of b, σ and c is found by substitution in (13)

$$\begin{aligned} E(I) &= \ln \frac{p}{w\sqrt{[6/(n - 1)]}} + \ln \frac{a}{w(1 + h)}\sqrt{\left(\frac{n}{12}\right)} + n \ln \frac{w(1 + h)}{K\epsilon} + nz^2/2w^2(1 + h)^2 + \tfrac{1}{2} \\ &= \ln \frac{ap}{w^2}\sqrt{[n(n - 1)/72]} + n \ln \frac{w}{K\epsilon} + (n - 1) \ln (1 + h) + (n - 1)/2(1 + h)^2 + \tfrac{1}{2} \end{aligned} \tag{28}$$

Expanding (28) in powers of h up to h^2 gives

$$E(I) = \ln \frac{ap}{w^2}\sqrt{[n(n - 1)/72]} + n \ln \frac{w}{K\epsilon} + (n - 1)h^2 + n/2 + Oh^3.$$

Substituting the expectation of h^2 from (27) gives

$$\begin{aligned} E(I) &= \ln \frac{ap}{w^2}\sqrt{[n(n - 1)/72]} + n\left(\ln \frac{w}{K\epsilon} + \tfrac{1}{2}\right) + \tfrac{1}{2} + Oh^3 \\ &= \ln \frac{a}{w}\sqrt{(n/12)} + \ln \frac{p}{w}\sqrt{[(n - 1)/6]} + \left[n\left(\ln \frac{w}{K\epsilon} + \tfrac{1}{2}\right)\right] + Oh^3. \end{aligned} \tag{29}$$

In the latter form, the three major terms of (29) show respectively the information required to encode the mean, standard deviation and individual variations about the mean.

In practice we take the ranges within which the mean and standard deviation are quoted to be $4\sigma_p$ and σ_p respectively, where σ_p is the standard deviation of the population as a whole.

The form (29) is an approximation based on the assumption of a large class population, and requires some modification for classes having very few members.

4.3 Notes on class descriptions

Forms (4) and (29) include the lengths of messages needed to specify the distributions of attributes within each class. As written, these forms assume that each class is described independently of all others. That is, the classes are not embedded in an hierarchic structure. Further economies can in some cases be achieved by describing the class distributions in terms of classes of classes, etc., but as the derivation of the resulting message lengths is rather tedious, it will be deferred to a later communication.

It might be thought an unnecessary extravagance to quote a different standard deviation for the same attribute in different classes. However, we have found, e.g. in the classification of six species of fur seals, that significant differences of standard deviation can occur between species. Further, there may be instances in which a difference in standard deviation is a significant aid in distinguishing two classes. Where no significant difference in standard deviation exists between two rather similar classes, an hierarchic description of their attributes could well allow the same value to be used for both classes.

5. Probabilistic interpretation

Combining the results (2), (4) and (29), we see that the total message can be separated into two major parts, one concerned with the dictionary of class names and the description of class properties, and of length at most logarithmically dependent on S, the other comprising a message for each thing giving its class and attributes.

The latter, which will dominate the total length for large S, can be written, for a member s of class t, as

$$-\ln(n[t]/S)$$
$$-\sum_{d\ \text{multistate}} \ln[(n[x[d,s],d,t]+1)/(n[d,t]+M[d])]$$
$$+\sum_{d\ \text{continuous}} [\ln(\sigma[d,t]/K\epsilon)+(x[d,s]-\mu[d,t])^2/2(\sigma[d,t])^2] \quad (30)$$

where $x[d, s]$ is the value (discrete or continuous) of attribute d of the thing.

If the set of S things may legitimately be treated as a sample from a much larger population, then the first term above can be regarded as minus the logarithm of the probability that an arbitrary thing belongs to class t, where a slightly biased estimate of this probability has been used. The second and third terms can be regarded as minus the logarithm of the probability that an arbitrary member of class t would have the attribute values found for thing s, where again slightly biased estimates have been used.

The value of (30) is therefore minus the logarithm of the probability that an arbitrary member of the whole population would be found to be in class t and to have the measurements $x[d, s]$. That is, (30) is minus the logarithm of the estimated population density in measurement space of class t in the neighbourhood of s. A classification which minimises the information measure will therefore assign each thing to the class with the highest expected density in the neighbourhood of the thing. The classes will not overlap in measurement space, and so (30) will represent minus the logarithm of the probability that an arbitrary member of the population would have measurements $x[d, s]$.

The total length of all attribute messages, found by summing (30) over all s, represents the probability of obtaining all the observed measurements given the following complex hypothesis about the set of S things:

(*a*) The set of S things is a sample drawn from a population which is the union of T subpopulations.
(*b*) Each subpopulation t occurs with a certain relative abundance, estimated by $n[t]/S$.
(*c*) Within each subpopulation t, measurements $x[d, s]$ are distributed normally or multistate with parameters as estimated in Sections 4.2 and 4.1.
(*d*) Within each subpopulation, different measurements are uncorrelated.
(*e*) Each thing is a member of a particular subpopulation.

Minimisation of (30) by choice of the class name labels, class distribution parameters and assignment of things to classes is therefore equivalent to a maximum-likelihood estimation of the relative abundance, distribution parameters, and membership of the subpopulations. However, the number of classes T could not be so estimated, as the value of (30) decreases monotonically with increasing number of classes.

The length of the class name dictionary and class description parts of the total message, which is not included in (30), can now be seen to play the role of minus the logarithm of the *a priori* probability of the hypothesis elements (*a*) to (*c*). Minimisation of the total message length is equivalent to choosing the hypothesis to maximise the posterior probability including these *a priori* factors.

The inclusion of any form of prior probability function may be considered objectionable on the grounds that its form is unknowable. However, it should be noted that, if the prior probability that some parameter lies in the neighbourhood of a value x is written as

$$f(x)dx,$$

the form employed by us corresponds to the assumption

that $f(x)$ is uniform over a range based on the spread of values observed in the whole set. The important parts of our form arise from the factor dx, not from the shape of $f(x)$.

The effect of including *a priori* factors on the estimates of subpopulation distribution parameters is to introduce a small bias into the estimates of continuous standard deviations and multistate probabilities. The normal maximum likelihood estimate of standard deviation

$$\Sigma(x_i - \bar{x})^2/n$$

becomes from (15)

$$\Sigma(x_i - \bar{x})^2/(n - 1)$$

which is actually a less biased estimate. Likewise, the encoding system derived in the appendix for the class description of multistate attributes effectively yields estimates of the relative probabilities of occupation of different states slightly different from their observed relative frequency of occupation. Although we have not investigated this question in detail, it appears that the estimates are on average biased slightly towards equality of probability in all states. The bias appears to be similar in form to, although rather less than, the bias we deliberately introduce for other reasons in (3).

The relative effects of the above biases decrease with increasing S, but the inclusion of the class description message length has the important result that the total message length exhibits a minimum with respect to T, the number of classes. Thus, minimisation of the information measure produces a classification having a number of classes determined by the measure itself, and not by the introduction of an arbitrary stopping rule such as is used in association analysis (Lance and Williams, 1965) or by the arbitrary selection of a certain similarity level on a dendrogram.

As S increases, the decreasing relative importance of the class description information makes it more likely that weakly separated classes will be distinguished in the classification which minimises the information measure. In effect, the increasing information content of the data makes the weak difference between the classes become significant.

6. Discussion

There are obvious similarities between the information measure and methods employed by others. For instance, the information statistic of Lance and Williams (1966) gives a measure between two groups of things which is the information needed to encode their (discrete) attributes on a code based on the union of the groups minus the information needed using different codes for each group. However, it does not include the information needed to specify the class membership of each thing, nor that needed to specify the codes.

The method of cluster analysis based on centroid sorting with a squared Euclidean distance similarity coefficient, in so far as it leads to a classification which minimises the total over classes of squared distances from class centroids, may be regarded as a maximum likelihood estimation of class membership and centroid positions of the class distributions, using the simplifying assumptions of:

(*a*) uncorrelated normal distribution of each attribute within each class;
(*b*) distribution standard deviations equal over all attributes and classes;
(*c*) all classes of equal abundance.

A classification which minimises the information measure will assign each thing to the class having the highest estimated population density in the neighbourhood of the thing. This assignment is that suggested by Sebestyen (1962) in his Bayesian approach to pattern recognition.

Most previous numerical classification systems have been based on one sort or another of inter-individual similarity measure. The wealth of measures proposed in the literature suggests that the notion of similarity has proved difficult to quantify, and difficult to relate to rigorous statistical concepts. It is interesting to note that in a recent comparison of several different similarity coefficients and clustering strategies applied to species of the genus *Stylosanthes*, L. 't Mannetje (1967) found the best correspondence with accepted taxonomy to be achieved by a coefficient and strategy which are purely empirical, and have no obvious basis in statistical or probability theory.

The information measure does not involve any attempt to measure similarity, either between individuals or between classes. Rather, it is based on comparisons between an individual, on the one hand, and a class on the other, and uses the well-defined concept of the probability that a member of a class of known distribution would be found to have certain measured attributes. The question of how in a given context one can best measure the similarity between two individuals is therefore replaced by the question of how in a given context, the distribution of a class can best be described. We have in this paper tried to use the most colourless possible class distribution descriptions. However, in a particular context, such as the classification of species, one might well be able, by appeal to the theory of evolution which is supposed to account for interspecific variation, to find a form of class distribution which more nearly models the observed distributions than do our assumed forms. The form of distribution in measurement space to be expected of a class in a particular context appears to us to be more amenable to systematic investigation than does the form of similarity measure, if any, appropriate to that context.

7. Results

We have not as yet applied the information measure to classifications arrived at by other numerical methods. We have, however, written a program for the University of Sydney KDF 9 computer which attempts to minimise

the information measure. The only data body so far submitted to the program for which there existed an established classification is a set of some 30 measurements of 217 seal skulls, the sample comprising six species. (See **Table 1.**)

Table 1

Description of seal skull sample

	MALE	FEMALE
A. tropicalis tropicalis	15	9
A. australis	31	3
A. tropicalis gazella	60	18
A. pusillus	13	6
A. doriferus	34	10
A. forsteri	16	2

The program did not have access to either the sex of the skulls or the location of capture.

The program distinguished seven classes which corresponded fairly closely to those male and female groups of the six species which were represented by a sample of reasonable size. (See **Table 2.**)

There are three main exceptions: first, the two species *A. pusillus* and *A. doriferus* were indistinguishable to the program. The only division was on the basis of sex. This indistinguishability is apparent from an inspection of the skulls themselves. It is difficult to note any difference for either sex between skulls from the two species. Further, an inspection of the means and standard deviations of the measurements shows that the differences between pairs of means of the same measurement for a particular sex are seldom greater than about 0·3 standard deviations.

Secondly, half of the group of 15 *A. tropicalis tropicalis* males were grouped with the females. These seven males are the younger members of the group and tend to look more like females. This same fact of age could explain a number of other small misclassifications. The four *A. pusillus* and six *A. doriferus* males grouped with the *A. australis* males are both collections of the younger members of their respective groups, and likewise for the three *A. forsteri* males misclassified.

Thirdly, the grouping of the *A. tropicalis tropicalis* and *A. forsteri* males is probably due to a combination of too small sample size and notable similarity between the two groups. Their means for the same measurement seldom differ by more than about 0·5 standard deviations.

Acknowledgements

This work was carried out in the Basser Computing Department, School of Physics, University of Sydney. We are grateful for the use of the KDF 9 computer and other facilities of the School of Physics. One of the authors (D. M. Boulton) was engaged on this work whilst in receipt of a University of Sydney Post-Graduate Studentship. We are indebted to Miss Judith King from the British Museum who supplied us with all the seal skull measurements, and also to Mrs. E. Vincent for her time and patience in typing this paper.

Appendix

The optimum encoding of multistate data

The statement of the class of each thing and the statement of values of multistate attributes present the same problem of how to optimise the labels used to encode this information and how to optimise the description of the label set employed.

Given a set of N things, we wish to produce a message saying in which of M states each thing lies. This message may be regarded as the concatenation of N 'labels', one per thing, each 'label' being a code name

Table 2

Program versus zoological classification for six species of seals

PROGRAM CLASSES	MALES						FEMALES					
	A trop. tropicalis	*A. australis*	*A. trop. gazella*	*A. pusillus*	*A. doriferus*	*A. forsteri*	*A. trop. tropicalis*	*A. australis*	*A. trop. gazella*	*A. pusillus*	*A. doriferus*	*A. forsteri*
4	7	7				12						
3		23		4	6	3						
6	1	1	59									
1				9	28							
7	7						8					1
2							1		17			
5								2	1	5	10	1

for the state of the corresponding thing. For the M names of the M states to be distinguishable from one another, the length (in units of $\log_2 e$ bits) of the name for the mth state must be $-\ln p[m]$, where

$$\sum_{m=1}^{M} p[m] \leqslant 1. \tag{1}$$

Maximum economy is achieved by choosing the $p[m]$ to satisfy the equality sign in (1).

If a number $r[m]$ N things have state m, the total length of the message giving all N states is

$$-N \sum_{m=1}^{M} r[m] \ln p[m]. \tag{2}$$

Expression (2) is minimised by choosing $p[m] = r[m]$. However, the receiver of the message has no *a priori* knowledge of $r[m]$, and cannot deduce from the message the labels employed. Thus he is unable to decode the message as it stands.

Assume that the receiver has exact knowledge of N and M, and that there has been prior agreement between sender and receiver that for this N and M the labels used will be based on one of an agreed finite set of M-tuples $\{p[m]\}$. (A similar agreement, or agreement on a general rule, is needed for any other values of M and N. The complete collection of agreements, or the general rule, constitutes the 'code book' for all such communications.)

The message of length (2) must be preceded by an additional message (the 'class name dictionary') which tells the receiver which one of this finite set of M-tuples has been selected by the sender. The sender will choose that M-tuple of $p[m]$ values which most nearly matches the relative abundances $r[m]$ in order most nearly to minimise (2). If the agreed set of M-tuples contained one M-tuple $\{p[m]\}$ for each possible M-tuple $\{r[m]\}$, then the minimum of (2) could always be achieved. However, there is a large number of possible M-tuples $\{r[m]\}$, and the length of the class name dictionary which nominates the particular M-tuple $\{p[m]\}$ being used would therefore become considerable.

A better compromise is to agree on a relatively small set of M-tuples $\{p[m]\}$, thus reducing the average length of the class name dictionary. In general, the values $p[m]$ adopted will no longer exactly equal the values $r[m]$, and so on average the length of the main part of message, as given by (2), will be slightly greater than the minimum possible.

We would like to determine how many M-tuples $\{p[m]\}$ should be employed, over what range of values of $\{r[m]\}$ each should be used, how much information will be needed to nominate a particular M-tuple, and by how much the expression (2) will on average exceed its minimum.

For some range of values of $\{r[m]\}$, we employ a fixed M-tuple of names whose lengths are minus the logarithms of $\{p[m]\}$, where $\sum_m p[m] = 1$.

Write

$$r[m] = p[m] + a[m] \quad (m = 1, 2, \ldots M) \tag{3}$$

where

$$\sum_m a[m] = 0. \tag{4}$$

The additional cost D incurred in the length of the main part of the message due to the use of the M-tuple $\{p[m]\}$ instead of the optimum M-tuple $\{r[m]\}$ is, from (2), given by

$$\begin{aligned} D &= -N\{\sum_m (p[m] + a[m]) \ln p[m] \\ &\qquad - \sum_m (p[m] + a[m]) \ln (p[m] + a[m])\} \\ &= N \sum_m (p[m] + a[m]) \ln (1 + a[m]/p[m]). \end{aligned} \tag{5}$$

Since $a[m]/p[m]$ is expected to be small, we expand (5) to the second power of $a[m]/p[m]$, and, using (4), obtain

$$D = \frac{N}{2} \sum_m a[m]^2/p[m]. \tag{6}$$

Now the possible M-tuples $\{r[m]\}$ range over the $(M-1)$ dimensional simplex G defined by

$$\sum_m r[m] = 1 \tag{7}$$

$$r[m] \geqslant 0 \quad (\text{all } m) \tag{8}$$

The density of M-tuples is uniform throughout G.

If we assume that all M-tuples within the simplex are equally likely to be encountered, the information needed to nominate the use of a particular label set $\{p[m]\}$ will, in the best encoding, be the minus the logarithm of the fraction of the volume of G over which the labels $\{p[m]\}$ are employed. We therefore wish to choose the region of G within which $\{p[m]\}$ is used, to have the largest volume for a given average value of D over the region. The regions of application of different M-tuples $\{p[m]\}$ must not overlap, and must together fill G. We choose regions of hyper-parallelepiped shape, since they can be packed together. We now attempt to find the optimum dimensions of the region of application of a particular M-tuple $\{p[m]\}$.

The contour in G of points yielding a given additional cost D is the intersection of the M-dimensional ellipsoid F defined by (6) with the hyperplane (7) containing G. The contour is an $(M-1)$ dimensional ellipsoid E.

The linear transformation T defined by

$$b[m] = a[m]/\sqrt{(p[m])} \tag{9}$$

and correspondingly

$$q[m] = r[m]/\sqrt{(p[m])} \tag{10}$$

maps F into the sphere F' defined by

$$\frac{N}{2} \sum_m b[m]^2 = D \tag{11}$$

and G into the simplex G' defined by

$$\sum_m \{\sqrt{(p[m]\, q[m])}\} - 1 = 0 \tag{12}$$

$$q[m] > 0 \quad (\text{all } m). \tag{13}$$

Since the centre of F lies in G, the centre of F' lies in G'. Thus the contour in G' of points with additional cost D, namely the intersection of F' with the hyperplane (12), is an $(M-1)$ dimensional sphere of radius R equal to the radius of F', hence

$$D = \frac{N}{2} R^2. \tag{14}$$

By virtue of the spherical symmetry in G' of the cost D about the M-tuple $\{b[m] = 0\}$, the optimum parallelepiped region of application of this M-tuple must, under T, map into an $(M-1)$ dimensional hypercube C in G' centred on the point $\{b[m] = 0\}$. Let the length of the sides of C be x. Since T is linear, all points within C are equally likely to be encountered. The average squared distance of all points in C from the centre of C is

$$\frac{2^{M-1}}{x^{M-1}} \int_0^{x/2} dy_1 \int_0^{x/2} dy_2 \ldots \int_0^{x/2} dy_{M-1} (y_1^2 + y_2^2 + \ldots + y_{M-1}^2)$$
$$= \frac{M-1}{12} x^2 \tag{15}$$

where the y_i are a set of $(M-1)$ rectangular coordinates aligned with the edges of C.

Hence, from (14), the average additional cost D over the region is

$$D\,av = N(M-1)x^2/24. \tag{16}$$

Since T is linear, the fraction of G' occupied by C equals the fraction of G occupied by the region of application of $\{p[m]\}$. To find the message length needed to nominate the choice of $\{p[m]\}$, we must find the volumes of C and G'.

The volume of C is

$$x^{M-1}. \tag{17}$$

To find the volume of G', consider the M-dimensional simplex S having G' as one face and the origin $\{q[m] = 0\}$ as the opposite vertex. The other vertices of S are the vertices of G', viz. the points (in q-coordinates)

$$\left(\frac{1}{\sqrt{(p[1])}}, 0, 0 \ldots 0\right), \left(0, \frac{1}{\sqrt{(p[2])}}, 0 \ldots 0\right),$$
$$\left(0, 0 \ldots \frac{1}{\sqrt{(p[M])}}\right).$$

The volume of S is therefore

$$\frac{1}{M' \cdot \prod_m \sqrt{(p[m])}} \tag{18}$$

The volume of S is also given by Vh/M, where V is the volume of G', and h is the perpendicular distance of the origin from the plane (12). Since the sum of the squares of the coefficients of $q[m]$ in (12) is $\sum_m p[m] = 1$, h is found by substituting the coordinates $\{p[m] = 0\}$ of the origin into the left hand side of (12). Hence

$$h = 1$$

and from (18)

$$V = \frac{1}{(M-1)! \prod_m \sqrt{(p[m])}}. \tag{19}$$

Combining (17) and (19), we get that the length of message needed to nominate $\{p[m]\}$ is

$$-\ln \{\prod_m \sqrt{(p[m])} x^{M-1} (M-1)!\} \tag{20}$$

The total expected additional message length due to the need to nominate $\{p[m]\}$ and to the non-optimum match between $\{p[m]\}$ and $\{r[m]\}$ is found by adding (20) and (16) to be

$$N(M-1)x^2/24 - \tfrac{1}{2} \sum_m \ln p[m]$$
$$- (M-1) \ln x - \ln (M-1)!. \tag{21}$$

The value of (21) is minimised by choosing

$$x^2 = 12/N \tag{22}$$

giving (21) a minimum value of

$$(M-1)/2 - \tfrac{1}{2} \sum_m \ln p[m] + \frac{M-1}{2} \ln (N/12) - \ln (M-1)!$$
$$= \left(\frac{M-1}{2}\right) (\ln N/12 + 1)$$
$$- \tfrac{1}{2} \sum_m \ln p[m] - \ln (M-1)!. \tag{23}$$

The total additional cost incurred with any particular set of abundances $\{r[m]\}$ will depend on the difference between $\{r[m]\}$ and the nearest $\{p[m]\}$, and hence on the details of how the parallelepiped regions are packed together. We do not wish to consider these details for all N and M, and in any case there are likely to be many different packing patterns of near-optimum efficiency. It would be appropriate to obtain the additional cost we expect to incur with a particular $\{r[m]\}$ by averaging the additional cost over all M-tuples $\{p[m]\}$ which include $\{r[m]\}$ in their ranges of application.

The additional cost incurred by using a particular $\{p[m]\}$, given a particular $\{r[m]\}$, comprises two parts. The first part, D, due to the non-optimum match between the label set used and the relative frequencies of the states, is given by (6). The second, due to the need to nominate the label set used, is given by (20) substituting the optimum value of x from (22), that is, by

$$-\tfrac{1}{2} \sum_m \ln p[m] + \frac{M-1}{2} \ln (N/12) - \ln (M-1)! \tag{24}$$

It is easily seen that the region in G including all the $\{p[m]\}$ having $\{r[m]\}$ in their ranges of application is approximately centred on $\{r[m]\}$, and is similar in position, size and shape to the region of application of an M-tuple $\{p[m]\}$ having $p[m] = r[m]$. Because this region, over which we wish to average, is approximately centred on $\{r[m]\}$, we may approximate the average value of (24) in the region by substituting $r[m]$ for $p[m]$ in (24). The value of (6) depends on the differences $a[m]$

between $p[m]$ and $r[m]$, with a weighting dependent on $p[m]$. If the variation in the weighting over the region of averaging is neglected, (6) may be replaced by

$$D = \frac{N}{2} \sum_m a[m]^2/r[m]. \quad (25)$$

If now the region of averaging is taken as being identical to the range of application of $p[m]$ where $p[m] = r[m]$, the average value of (25) is given by (16), substituting the optimum value of x from (22), as

$$(M - 1)/2 \quad (26)$$

Combining (24) and (26) gives the expected additional cost incurred with relative frequencies of states $r[m]$ as

$$\left(\frac{M-1}{2}\right)(\ln N/12 + 1) - \tfrac{1}{2}\sum_m \ln r[m] - \ln (M-1)! \quad (27)$$

This expression represents the expected total cost additional to the message length which would be required to give the states of the N things using optimum label lengths of $-\ln r[m]$. The expected total message length is therefore

$$\left(\frac{M-1}{2}\right)(\ln N/12 + 1) - \tfrac{1}{2}\sum_m \ln r[m] - \ln(M-1)! - N\sum_m r[m] \ln r[m]$$
$$= \left(\frac{M-1}{2}\right)(\ln N/12 + 1) - \ln (M-1)! - \sum_m (n[m] + \tfrac{1}{2}) \ln r[m] \quad (28)$$

where $n[m]$ is the number of things in state M. The above form is essentially that used in the body of the paper for the information needed to record the class of each thing, and to record the values of multistate variables.

It is worth noting that in size and shape, the region over which an M-tuple $p[m]$ is applied is essentially similar to the region of expected error in the estimation of the probabilities of a multinomial distribution based on a sample of size N. Thus the message transmitted to nominate the M-tuple is sufficient to convey essentially all the available information about the probability of occurrence of each state.

References

GOOD, I. J. (1965). The Estimation of Probabilities: An Essay on Modern Bayesian Methods, *Research Monograph* No. 30, The M.I.T. Press, Cambridge, Mass.

LANCE, G. N., and WILLIAMS, W. T. (1965). Computer programs for monothetic classification (Association Analysis), *Computer Journal*, Vol. 8, pp. 246–249.

LANCE, G. N., and WILLIAMS, W. T. (1966). Computer programs for hierarchical polythetic classification (Similarity Analysis), *Computer Journal*, Vol. 9, pp. 60–64.

'T MANNETJE, L. (1967). *A Comparison of Eight Numerical Procedures in a Taxonomic Study of the Genus* Stylosanthes *Sw.*, C.S.I.R.O. Division of Tropical Pastures, Brisbane, Qld.

SEBESTYEN, G. S. (1962). *Decision Making Processes in Pattern Recognition*, Macmillan, New York.

SHANNON, C. E. (1948). A Mathematical Theory of Communication, *Bell System Tech. J.*, Vol. 27, p. 379 and p. 623.

SOKAL, R. R., and SNEATH, P. H. A. (1963). *Numerical Taxonomy*, W. H. Freemen and Co., San Francisco and London.

WILLIAMS, W. T., and DALE, M. B. (1965). Fundamental Problems in Numerical Taxonomy (in *Advances in Botanical Research 2*).

Correspondence

To the Editor
The Computer Journal

A new method for solving polynomial equations

Sir,

Reading the article about solving polynomial equations by Garside, Jarratt and Mack (this *Journal*, Vol. 11, p. 87), I wonder if the method can be further improved by successive long divisions of $F(Z)$ into a continued fraction with the repeated roots consequently removed. In general, for an nth order polynomial their

$$F(Z) = \cfrac{a_1}{Z + b_1 + \cfrac{a_2}{Z + b_2 \cfrac{a_n}{Z + b_n}}}$$

For a repeated root a_n should $=0$ and $-b_n$ be the repeated root. Should a_n be very small it could cause unnecessary trouble were it purely a round-off error. Consequently, if small values of the coefficients are not to be ignored, they should all be calculated to double length with only the single length used later for evaluation of the continued fraction. This would seem a small price to pay for separating out the repeated roots.

It may happen that part of the way through the repeated division process a leading coefficient (or coefficients) of the numerator becomes zero. This will result in the next division giving a second (or higher) order polynomial with correspondingly less undone divisions. This must be catered for in the program. Also if the leading coefficient is very small, it could cause very large coefficients in the next stage. Should this happen, all the coefficients should be sealed and this may result in the leading one becoming zero.

In calculating the continued fraction, it may happen that one denominator vanishes. This must make the next fraction zero.

Yours sincerely,

J. P. O'BRIEN

English Electric Diesels Limited,
Newton-le-Willows
7 June 1968

(*Further correspondence appears on pp.* 172 *and* 240)