An information measure for single link classification

D. M. Boulton and C. S. Wallace

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

The information measure is an objective measure of the quality of a classification and results from an information transmission view of the classification problem. So far information measures have only been derived for the case where an explicit assumption is made about the form of the distribution of attribute values within a class. One important method which involves no such explicit assumption is single link. In this paper we derive a new information measure which is optimised by classifications produced by the single link method. By investigating the properties of this information measure we are able to gain more insight into the single link method and also determine the type of problem to which it best applies.

(Received October 1973)

1. Introduction

Many methods of automatically classifying things on the basis of a set of measurements (attributes) have been developed over the last decade or so. Unfortunately, 'comparative studies have shown that when different methods are applied to the same data set there are often major discrepancies between the results obtained', (Jardine and Sibson, 1971a). This state of affairs has led a number of authors to consider ways of comparing the quality of different methods, (see for example: Jardine and Sibson, 1971a; Williams, Clifford and Lance, 1971).

Many classification methods are of the type that proceed via a matrix of dissimilarities to a numerically stratified hierarchic classification which is usually represented by a dendrogram. Jardine and Sibson (1971b) have considered such methods and set up a number of adequacy conditions which, they believe, methods of this type should satisfy. The method called single link (or nearest neighbour) is the only one which satisfies all conditions.

We have proposed a classification method which does not involve any dissimilarity measure and hence is not a method of the above type. It is based on a measure of classification goodness which is called the information measure as it results from an information transmission view of classification. The classification structure is considered as providing a framework for encoding a message which conveys the attribute value sets of all the things being classified. The information measure is the length of this message and we consider that the best classification is that which minimises the information measure.

We have derived two information measures, one for non-hierarchic classifications (Wallace and Boulton, 1968) and the other for hierarchic classifications (Boulton and Wallace, 1973). The basis of both is that the classification provides a simple piecewise model of the distribution of things in attribute measurement space. Within each class a separate multivariate distribution, is used, whose form is determined by that expected to be found within a class, e.g. normal distributions for continuous attributes. The composite distribution which results when the class distributions are combined, is used to obtain the probabilities of different sets of attribute values for obtaining the optimum message segment lengths for conveying all the things' attribute values.

As classification methods of the same type as single link are not based on any explicit assumption of the form of the distribution of attribute values within a class, neither of the information measures we have so far derived is really applicable. In this paper we will derive a new information measure which covers this other class of methods. With this new information measure the attribute value sets of all things, and hence their positions in measurement space, are encoded as a series of vectors which each locate one thing relative to another.

We will show that if the message length used to convey a vector is considered as the dissimilarity of the two things it connects, then the information measure is always minimised by the single link classification. We will then go on to discuss the applicability of the method of single link by considering the properties of the corresponding information measure.

2. Derivation of the new information measure

The problem is to construct a message which will convey the attribute values of an ordered set of N things, or more conveniently, when the things are represented by points in attribute measurement space, the positions of an ordered set of N points. The message is constructed to have three parts as follows.

1. The position of the first thing is specified relative to the origin of measurement space.

For each remaining thing, say thing number i, the following two pieces of information are specified:

2. the length and direction of a vector which defines the position of this ith thing relative to the position of another thing, say thing number j
3. the identity of thing j, i.e. the value of j.

If the above message is to convey unique positions for all the things and contain no redundancy, the vectors must form a spanning tree. That is, a tree where:

1. no closed loops occur
2. each point (thing) is visited by at least one vector
3. the tree is connected.

These rules imply that, if we consider the positive direction of a vector describing thing i relative to thing j to be directed from j to i, then the unique route through the tree of vectors from thing 1 to thing i is along the positive direction of each vector encountered, for all $i > 1$.

An example of a spanning tree for six things and two continuous attributes is shown in **Fig. 1.** The corresponding message to convey the attribute values (positions) of the things is

$$V_{1,0};\ V_{2,6},\ 6;\ V_{3,1},\ 1;\ V_{4,5},\ 5;\ V_{5,3},\ 3;\ V_{6,3},\ 3\ .$$

where

$V_{i,j}$ is the vector which positions thing i relative to thing j,

$V_{1,0}$ is the vector from the origin of measurement space to the first thing.

Notice that in this message the identity of the thing pointed to by a vector is implied by the ordering of the vectors in the message. For example, the third vector in the message specifies the position of the third thing relative to the specified thing (number 1) at the origin of the vector.

To determine how the three parts of the message are encoded

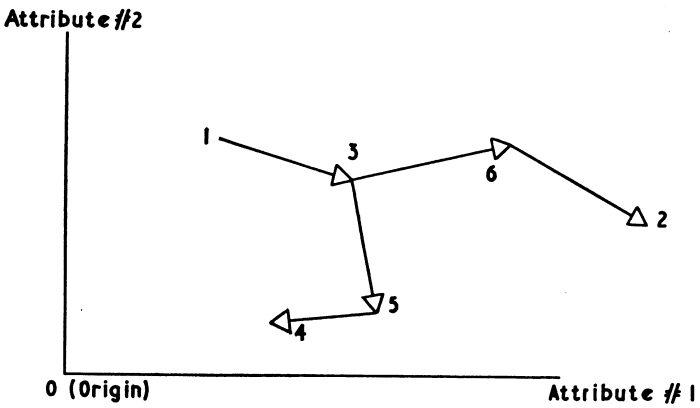

Fig. 1 A spanning tree connecting six things

we must consider the principles of efficient encoding. These were first set down by Shannon (1948), who showed that a message can be considered as signalling one of a set of possible events and that to minimise the expected message length the optimum message for event i should have a length given by

$$-\log_b p_i \qquad (2.1)$$

The base of logarithms is arbitrary and is equal to the size of the code alphabet. For convenience we use natural logarithms.

It is when applying the principle of efficient encoding to obtain the optimum message lengths for each part of the total message that we are forced to make assumptions about the probabilities of different events. There are three sets of different events, one for each of the three parts of the message:

1. all possible positions of the first thing relative to the origin of measurement space
2. all possible vectors which can be used to position one thing relative to another
3. all things relative to which another thing can be located.

Consider the specification of the position of the first thing. The most colourless assumption is that all positions within a predetermined fixed region of measurement space are equally likely. Then the resulting message length is independent of the attribute value set the first thing happens to have. However, because the attribute value set of the first thing is fixed by the data, this first part of the message, regardless of its length, cannot alter the classification which minimises the complete information measure. Thus the distribution assumed for the position of the first thing is irrelevant when different classifications of a single data set are being compared, so this part of the message can be omitted from the information measure.

Consider next how the third part of the message is encoded. If there are N things being classified then the position of a thing can be specified relative to any of the remaining $(N-1)$ things. The most colourless assumption is that with equal probability any one of the remaining $(N-1)$ things could be chosen. Thus the optimum message length is

$$\ln (N-1) \; . \qquad (2.2)$$

This message segment will appear $(N-1)$ times in the whole message, so contributing a total length of

$$(N-1) \ln (N-1) \; . \qquad (2.3)$$

As this length depends only on N, it is independent of the classification and can, together with part (1), be ignored when comparisons are being made.

Finally we come to that part of the message wherein the $(N-1)$ vectors, which position one thing relative to another, are conveyed. To simplify the following discussion we will assume that the distribution of different lengths and directions a vector can have, is independent of the position of the vector in measurement space. We can thus consider just one distribution, and hence code, which will be used for all vectors. It is important to note, however, that this assumption does not invalidate the discussion which follows.

Let the probability density of vectors of different length l and direction θ be

$$f(l, \theta) \; . \qquad (2.4)$$

If each vector points to somewhere in a small region* δu and $f(l, \theta)$ is reasonably constant over this region then the probability of a vector (l, θ) is given approximately by

$$f(l, \theta)\, \delta u \; . \qquad (2.5)$$

Thus the optimum message length to convey the vector is

$$-\ln f(l, \theta)\, \delta u \; . \qquad (2.6)$$

The total message length to specify all the $(N-1)$ vectors is given by

$$\sum_{i=1}^{N-1} -\ln f(l_i, \theta_i)\, \delta u \; . \qquad (2.7)$$

The function $f(l, \theta)$ can, in principle have any form. It should be chosen to conform with our expectation of different vectors, but for the time being let us not commit outselves to any choice of $f(l, \theta)$. In fact we will now show that if the message length to convey a vector is treated as a dissimilarity measure on the two things at the vector's ends, then regardless of the form of $f(l, \theta)$ the classification which minimises the information measure is the same classification which would be yielded by single link with this dissimilarity measure.

Gower and Ross (1969) have shown that a spanning tree with branches whose lengths are equal to the dissimilarity of the pair of things at their ends, represents a single link classification when the sum of the lengths of the branches (the length of the tree) is minimised. The only term in the information measure which depends on the structure of the spanning tree used in the encoding process is part (2) which defines the lengths and directions of all the vectors in the tree. Thus, if the message length to define a vector is a dissimilarity then minimising the information measure is equivalent to minimising the sum of the dissimilarities of things directly linked in the spanning tree. Therefore, the information measure will be minimised by the minimum spanning tree, that is by the single link classification. In other words, if the contours of constant $f(l, \theta)$, i.e. of equal probability, about a thing coincide with contours of equal dissimilarity then the information measure is minimised by the single link classification; this is so because $\ln x$ is a monotonic increasing function of x.

Let us now consider the form we would expect for $f(l, \theta)$ on the basis of the structure we expect a classification to possess. We consider that classification is profitable when the distribution of things in measurement space is nonuniform. The things should form groups each having members concentrated in a small region of measurement space. This implies that we expect short interconnecting vectors to be more probable than longer ones. Thus $f(l, \theta)$ should be some monotonic decreasing function of increasing l, i.e.

$$f(l_1, \theta) < f(l_2, \theta); \; l_1 > l_2 \; .$$

Under such an assumption of $f(l, \theta)$, the implied dissimilarity of two things is also a monotonic increasing function of increasing distance between them. This property is, in fact, possessed by many dissimilarity measures. Ball (1970), for example, has plotted iso-dissimilarity contours for a number of popular measures and all but measures based on the correlation coefficient exhibit the above monotonic property.

*When $f(l, \theta)$ is a density (i.e. when continuous attributes are involved) the message length to specify a vector will be infinite unless each vector ends on a region δu of finite size. This implies that each continuous measurement is made to a finite accuracy.

Discussion

We have shown that an information measure designed for numerically stratified hierarchic classifications favours the method of single link providing that the probability of finding one thing close to another in measurement space is a monotonic decreasing function of increasing dissimilarity of the two things. We have also shown that a reasonable form for the above probability is a function which is monotonic decreasing with increasing distance from a thing. This form of probability density function $f(l, \theta)$ implies that dissimilarity is also a monotonic decreasing function of increasing distance between two things. Many dissimilarity measures in use actually have this property.

It is interesting to consider the form of the probability density $p(x)$ of finding a new thing, with coordinates (attribute values) x, amongst or near a group (i.e. a class) of things. Although no explicit assumption has been made about this density, it is unavoidably implied by the length of the optimum message used to encode the attribute values of such a new thing. This message length must have the form

$$-\ln p(x)\, \delta u \ ,$$

where $p(x)\, \delta u$ is the probability of finding a thing within a small region δu of measurement space with coordinates x. As the position of a thing is optimally encoded by specifying the length and direction of the vector linking it to its nearest neighbour, the distribution $p(x)$ assumed in the region of a class of things is a piecewise combination of sections of the function $f(l, \theta)$ each centred on a thing. A single section of $f(l, \theta)$ centred on a particular thing surrounds the thing as far away as it yields a higher probability than sections of $f(l, \theta)$ centred on other things.

The above composite model distribution for a class can be visualised as a number of 'mounds' which are butted together, not summed. Each mound is centred on a thing and has a shape given by $f(l, \theta)$. Thus we see that the probability of finding a thing in a densely populated region of measurement space can be no greater than finding a thing in a sparsely populated region. All 'mounds' have the same height and the probability of finding a new thing only depends on the distance from the nearest thing. In fact the only way the class distribution responds to dense clusterings of things is by raising the levels of the valleys as they are filled by more and more 'mounds'.

The above shape of class model distribution is extremely uncoloured. It says: we don't know exactly what shape of distribution to expect within a class so all we will assume is that if a thing is observed at some point then it is likely that another thing will occur nearby, that is with similar attribute values. This seems a reasonable assumption when the set of things being classified contains all known things of this type, as when all species in a single family are being classified. In such case the set of things is not a sample from a large population and so the classes cannot be samples from a number of large sub-populations. Therefore we are not really justified in setting up a model probability distribution, such as a normal distribution, within each class.

However, it is probably more often the case that the set of things being classified is a sample from a large population. In this case the form of the class distribution implied by single link is not suitable because it is very limited in the way it can respond to variations in the density of things in measurement space. The probability of finding one thing at a small distance ε away from another thing is assumed to depend only on ε and not on the density of the surrounding region of measurement space.

The main manifestation of this problem is the well known effect called 'chaining'. This has been described by Jardine and Sibson (1971b) as 'one substantial defect' of the single link method. Other undesirable effects can also be demonstrated. For example, the hierarchic level of dissimilarity at which two classes are combined will tend to decrease if the size of the sample being classified is increased. The resulting increased density of things in measurement space will tend to decrease the distance between neighbouring things so causing them to combine at lower levels in the hierarchy.

Conclusion

By showing how the single link method of classification can be expressed in terms of an information measure we have been able to gain more insight into the method. We find that, although single link makes no explicit assumption about the expected distribution of things within a class in measurement space, such a distribution is actually implied in the method. The form of this distribution turns out to be based on a most uncoloured assumption. However, on further consideration we find that the distribution is only reasonable when the things being classified constitute all such things known. It is not suitable when the things are a sample from a large population.

Acknowledgement

We are grateful to C. J. Van Rijsbergen for his helpful comments and for pointing out a number of relevant references.

References

BALL, G. H. (1970). *Classification Analysis*, Stanford Research Institute Project 5533, Menlo Park, California.

BOULTON, D. M., and WALLACE, C. S. (1973). Information Measure for Hierarchic Classifications, *The Computer Journal*, Vol. 16, No. 3, August, pp. 254-261.

GOWER, J. C., and ROSS, G. J. S. (1969). Minimum Spanning Trees and Single Link Cluster Analysis, *Applied Statistics*, Vol. 10, pp. 54-64.

JARDINE, N., and SIBSON, R. (1971a). Choice of methods for automatic Classification, *The Computer Journal*, Vol. 15, No. 4, pp. 404-406.

JARDINE, N., and SIBSON, R. (1971b). *Mathematical Taxonomy*, Wiley, London.

SHANNON, C. E. (1948). A Mathematical Theory of Communication, *Bell Sys. Tech. J.*, Vol. 27, pp. 379-423.

WALLACE, C. S., and BOULTON, D. M. (1968). An Information Measure for Classification, *The Computer Journal*, Vol. 11, p. 185.

WILLIAMS, W. T., CLIFFORD, H. T., and LANCE, G. N. (1971). Group-size dependence: a rationale for choice between numerical classifications, *The Computer Journal*, Vol. 14, No. 2, pp. 157-162.