Mathematical Geology, Vol. 3, No. 2, 1971

Some Information Theoretic Optimality Criteria for General Classification[1]

George F. Estabrook[2]

INTRODUCTION

Central to the various purposes for which general classification is practiced is the preservation, or summarizing, of information (Michener, 1963; Sneath, 1957; Sokal, 1962; Farris, 1967; Johnson, 1968; Rogers, 1961; Wirth, Estabrook, and Rogers, 1966). This discussion will present one method in which this fundamental concept can be made more precise and, within that context, suggest possible logical formulations for preservation or summarization of information.

FUNDAMENTAL CONCEPTS

The collection of conceptual objects to be classified will be called, S, the study. It is assumed that every effort has been made to render this collection representative. By a classification of the study shall be meant a partition, P_i, of the study. A partition, P_i, of S is a collection, $P_{ij}, j = 1,n_i$, of subsets of S such that $\bigcup_{j=1}^{n_i} P_{ij} = S$ and $P_{ij} \bigcap P_{ik} = \phi$ if $j \neq k$. P_{ij} will be called an equivalence class, or class of P_i.

By a character for S shall be meant a partition, K_i, of S, with classes $K_{ij}, j = 1, m_i$. K_{ij} will be called a character state, or state, of character K_i. In actual practice, K_i corresponds to some basis for making pairwise comparisons of the objects in S, such that for a pair (a,b) of objects in S it can always be determined whether or not a is similar to b with respect to that basis. Thus "is similar to" must be an equivalence relation, and must be the same as "belongs to the same character state of K_i as." It is evident that in this discussion character will mean "qualitative character" or "nominal character" (Andrews and Estabrook, 1971; Estabrook and Rogers, 1966; Hawksworth, Estabrook, and Rogers, 1968).

It is possible (Estabrook, 1967; Lambert and Williams, 1966) to associate with a partition a measure of the information which it may be thought to represent. This

[1] Manuscript received 22 October 1970.

[2] Departments of Botany and Zoology, The University of Michigan (USA).

measure may be intuited as the diversity of the partition's equivalence classes, or as its capacity, as a basis for comparison, to distinguish differences.

For a partition, P_i, of S let

$$H(P_i) \equiv -\sum_{j=1}^{n_i}\left(\frac{|P_{ij}|}{|S|} Ln_2\left[\frac{|P_{ij}|}{|S|}\right]\right)$$

where $|P_{ij}|$ is the number of objects in P_{ij} and $|S|$ is the number of objects in S, be that the measure.

Let P_i and P_k be two partitions of S. The measure of information in P_i restricted to some class P_{kj} of P_k is

$$H(P_i/P_{kj}) \equiv -\sum_{L=1}^{n_i}\left(\frac{|P_{iL}\bigcap P_{kj}|}{|P_{kj}|} Ln_2\left[\frac{|P_{iL}\bigcap P_{kj}|}{|P_{kj}|}\right]\right)$$

This measures the amount of diversity among the objects of P_{kj} which P_i represents; or similarly, this measures the capacity of P_i to distinguish differences in pairs from P_{kj}. Thus,

$$H(P_i) - H(P_i/P_{kj})$$

measures the amount of information about P_i, which is "in" the statement "x belongs to P_{kj}." This difference may be negative in which situation the objects in P_{kj} are more nearly evenly distributed through the classes of P_i than are the objects in S. Thus, if the statement, "x belongs to P_{kj}," is true, P_i can distinguish a greater proportion of pairs of objects; and the information remaining in P_i increases by the amount of confusion, or negative information, introduced by "x belongs to P_{kj}." If this difference is positive, the objects of P_{kj} are even less evenly distributed through the classes of P_i than are the objects of S. Thus, the amount of information remaining in P_i falls by the amount of information about P_i contained in the statement "x belongs to P_{kj}."

We may now define

$$H(P_i/P_k) \equiv \sum_{j=1}^{n_k}\left(\frac{|P_{kj}|}{|S|} H(P_i/P_{kj})\right)$$

This is the expected value for $H(P_i/P_{kj})$ as j varies, which may be intuited as the amount of information in P_i which is not in P_k. Thus, $R(P_i, P_k) \equiv H(P_i) - H(P_i/P_k)$ may be considered a measure of the information common to P_i and P_k. This measure is symmetric, i.e.,

$$R(P_i, P_k) = R(P_k, P_i)$$

which is intuitively pleasing, as we may now conceive of the information associated with two partitions, P_i, P_k (i.e., the information in the cartesian product $P_i \otimes P_k$) as consisting of three distinct types:

that, measured by $H(P_i/P_k)$ which is exclusive to P_i
that, measured by $H(P_k/P_i)$ which is exclusive to P_k, and
that, measured by $R(P_i, P_k)$ which is common to both.

We may write $H(P_i \otimes P_k) = H(P_i/P_k) + H(P_k/P_i) + R(P_i, P_k)$

Because these concepts are applicable to any partition, they apply equally to characters as to classifications.

OPTIMALITY CRITERIA FOR INFORMATION PRESERVATION

Let K_i, $i = 1,n$ be the n characters for a study, S. We have n information containing entities, each containing an amount measured by $H(K_i)$, respectively, of information. By a good classification, P_i, of S, we mean one which preserves or summarizes this character information effectively. Several concepts suggest themselves as appropriate interpretations of "preserve information."

A classification, P_j, of S for which $C_0(P_j) \equiv \sum_{i=1}^{n} R(P_j, K_i)$ assumes its maximum value might be considered desirable as it preserves the most information about the characters. However, it is easy to see that the classification determined by this method is $D = \bigotimes_{i=1}^{n} K_i$ which is typically the discrete partition with each object in a class by itself. In many situations this would be considered undesirable. The number of classes in a partition establishes an upper limit for the amount of information in it; the more classes, the higher proportion of pairs can be distinguished if objects in S are sufficiently equally distributed through those classes. D has many classes; indeed, for any class, K_{ij}, there are classes, D_{ijL}, $L = 1, n_{ij}$ of D such that $K_{ij} = \bigcup_{l=1}^{n_{ij}} D_{ijL}$. Thus, $R(D,K_i) = H(K_i)$ and D preserves all the information in all the characters.

We might wish to require a classification to be efficient as well as effective in preserving information. A classification, P_j, of S for which $C_1(P_j) \equiv C_0(P_j)/H(P_j)$ assumes its maximum value would designate a classification with the highest average fraction, $R(P_j, K_i)/H(P_j)$, of its own information, which is effective for predicting the characters, K_i.

In situations where the characters shared little information, i.e., $R(K_i, K_L)/H(K_i \otimes K_L)$ is low for almost every $i \neq L$, P_j would typically be D, for S is finite, $H(D)$ is relatively low, and D is either equal to, or nearly equal to, the discrete partition.

In situations where some characters, K_i, $s \leq i \leq p$ say, shared enough information, i.e., $R(K_i, K_L)/H(K_i \otimes K_L)$ was not virtually zero for $s \leq i \leq L \leq p$, $E \equiv \bigotimes_{i=s}^{p} K_i$ could conceivably be indicated. Even under these conditions, the typical E is likely to have many classes and some concept of simplicity or efficiency for "good" classifications may be violated.

Two approaches suggest themselves at this point.

1. The function family

$$C_\eta(P_j) \equiv \sum_{i=1}^{n} R(P_j, K_i)/(H(P_j))^\eta \qquad \eta \geq 0$$

might contain a range of members corresponding to an interval of values for η which might indicate a pleasing P_j. If η has a high value, however, C_η may indicate degenerate classifications with one class consisting of most of the members of S and perhaps a few small classes comprising the rest.

2. One of the difficulties with the preceding approaches is that classifications with differing numbers of classes are competing for designation. Because the number of classes in a classification establishes an upper bound for the amount of information it can contain ($Ln_2(N)$ for an N class classification), the domain of C_η contains arguments that are not entirely comparable. Thus, a second approach would be to optimize with $C_{\eta\mu}(P_j) \equiv C_\eta(P_j)$ restricted to classification with μ classes, $\mu > 1$. Perhaps the member of this family most in keeping with our concept of efficiency and simplicity would be C_{12} which would designate a 2 class classification which held the highest average fraction of its information in common with the characters.

OPTIMALITY CRITERIA FOR IDENTIFICATION

Logically classification precedes identification. However, it may be interesting to ask, for a given classification, P_j, of S, how easy is it to identify an object in S into its proper class of P_j from a knowledge of one of the character states to which it belongs. Analogous to the development of optimality criteria for information preservation $I_{\eta\mu}(P_j) \equiv \sum_{i=1}^{n}(R(P_j, K_i)/(H(K_i))^\eta)$ with P_j restricted to classifications with μ classes is suggested.

Here, for example, I_{12} would designate a 2 class classification, P_j, for which the fraction of a character's information held in common with P_j is, on the average, maximum, i.e., the average fraction of a character's information useful for identifying a typical object in S into its proper class of P_j would be as large as possible.

One method to grasp intuitively the difference between $D_{\eta\mu}$ and $I_{\eta\mu}$ is to realize that with $D_{\eta\mu}$ the large characters (i.e., those whose measure of information is great) tend to share a larger absolute amount of information with a typical classification than do the smaller characters; thus the larger characters contribute more heavily to the average fraction. In the situation of $I_{\eta\mu}$, the smaller characters tend to contribute more heavily to the average fraction than they do in the situation of $D_{\eta\mu}$. Thus, speaking intuitively, $I_{\eta\mu}$ is more equitable on a per character basis whereas $D_{\eta\mu}$ is more equitable on a total information basis.

CONCLUDING REMARKS

It should be pointed out that conceptual discussions of this type are rarely sufficient for determining what ought to be the situation in the domain of scientific reality. The purpose of this discussion is to provide one, among many, conceptual structure in which to consider the problem. Mathematical techniques for determining, in general, all the classifications designated by these criteria do not presently exist.

Heuristic (guessing type) computer programs for empirically testing these concepts are being developed. Workers interested in participating in the empirical aspects of this study are encouraged to contact the author.

REFERENCES

Andrews, J. T., and Estabrook, G. F., 1971, Applications of information- and graph-theory to multivariate geomorphological analysis: Jour. Geol., in press.

Estabrook, G. F., 1967, An information theory model for character analysis: Taxon., v. 16, no. 4, p. 86–97.

Estabrook, G. F., and Rogers, D. J., 1966, A general method of taxonomic description for computed similarity measure: BioScience, v. 16, no. 11, p. 789–793.

Farris, J. S., 1967, The meaning of relationship and taxonomic procedure: Systematic Zoology, v. 16, no. 1, p. 44–51.

Hawksworth, F. G., Estabrook, G. F., and Rogers, D. J., 1968, Application of an information theory model for character analysis in the genus *Arceothobium* (*Viscaoeae*): Taxon., v. 17, no. 6, p. 605–619.

Johnson, L. A. S., 1968, Rainbow's end: The quest for an optimal taxonomy: Australian Medical Publ. Co. Ltd., Sydney, NSW, 45 p.

Lambert, J. M., and Williams, W. T., 1966, Multivariate methods in plant ecology: Jour. Ecol., v. 54, p. 635–664.

Michener, C. D., 1963, Some further developments in taxonomy: Systematic Zoology, v. 12, no. 4, p. 151–172.

Rogers, D. J., 1961, Recent endeavors with computers in taxonomy: The Garden Jour. (NYBG), Nov.–Dec., 1961.

Sneath, P. H. A., 1957, Some thoughts on bacterial classification: Jour. Gen. Microbiol., v. 17, p. 184–200.

Sokal, R. R., 1962, Typology and empiricism in taxonomy: Jour. Theo. Biol., v. 3, p. 230–267.

Wirth, M., Estabrook, G. F., and Rogers, D. J., 1966, A graph theory model for systematic biology, with an example for the *Oncidiinae* (*Orchidaceae*): Systematic Zoology, v. 15, no. 1, p. 59–69.