AN ELEMENTARY MATHEMATICAL THEORY OF CLASSIFICATION AND PREDICTION *

by

T. T. Tanimoto
International Business Machines Corporation
New York, New York

November 17, 1958

AN ELEMENTARY MATHEMATICAL THEORY OF CLASSIFICATION AND PREDICTION*

T. T. Tanimoto

International Business Machines Corporation
New York, New York

Introduction. --The analysis of qualitative data is one of the areas which, to a great extent, has defied mathematical treatment. With the use of the modern large-scale electronic computers in mind, we shall propose a simple procedure which, when used as a tool, will assist us in solving problems in many fields where most of the important data are qualitative; e.g., taxonomy, organization theory, medical and psychiatric diagnosis, etc. The method does not preclude any quantitative data, since significant specified quantitative intervals can be considered as attributes. A simple enumeration process may be called a system of classification and may be very useful at times, but it certainly is not a scientific system of classification. It is generally felt that a good scientific classification system should be based on over-all similarity of those attributes of objects which are felt to be pertinent to the particular purpose of the classification.[1] Although scientific classification is a system by which we can compress much information about individual objects or ideas into smaller systems,

we certainly would not attempt to classify all the objects in the universe. We must localize our classification to some field or subfields of endeavor. The more local a scientific classification is, the more specific is the information yielded by the classification about the individuals. The decision as to how local or how global a classification system should be is determined by the degree of specificity of the information desired in the classification. Obviously, objects or ideas may be classified in many different ways, depending upon what we wish to accomplish, and thus lead us to consider different sets of attributes; e.g., airplanes may be classified as flying objects whose class would include such things as birds, flying saucers, bats, etc.; or airplanes may be classified as means of transportation, which would take in automobiles, ships, etc. In short, the attributes of the objects or ideas which we wish to classify must be chosen in light of what we wish to accomplish by the classification. One of the most useful results of a good scientific classification system, besides that of gaining new over-all information, is its value in the prediction of the existence or non-existence of an object or an attribute. Our method will be such that the prediction will be of a probabilistic nature.

It is clear that any procedure or theory based on attributes can not be completely objective, in so far as an expert in a particular field must make the decisions as to

(1) which objects are to be considered

(2) what attributes are pertinent

(3) whether a particular object does or does not possess a specific attribute of the set of pertinent attributes.

This degree of subjectivity is not only necessary, but is probably a good thing, in that an expert's personal experience and insight are incorporated into the procedure. We shall propose as our fundamental assumptions:

(i) All the objects with which we are concerned must, to the best of our knowledge, be distinct kinds of objects.

(ii) All the attributes considered must be distinct. This does not preclude any attributes which we, from experience, feel are comprehended by other considered attributes. If a comprehension exists, our procedure will bear this out.

We shall develop our theory upon the subjective aspects (1), (2) and (3) and the assumptions (i) and (ii).

Suppose that B is a finite set of n objects, and let a be a particular attribute possessed by some elements of B, then the classical definition of the probability p that an element of B chosen at random, assuming equal likeliness, will possess the attribute a, is given by

$$p = \frac{N(a,\ B)}{N(B)}$$

where N(a, B) is the number of elements of B which possess the attribute a, and N(B) is the number of elements in B. We now extend this definition to include the concept of similarity of a pair of attributes.

Let $A = \{a_i\}$, $i = 1, 2, \ldots, m$ be a finite set of m attributes (of which some may be the absence of particular attributes) associated with the finite set $B = \{b_j\}$ $j = 1, 2, \ldots, n$ of n objects. Define the $m \times n$ matrix $R = (r_{ij})$ so that $r_{ij} = 1$ if b_j possesses the attribute a_i and $r_{ij} = 0$ if b_j does not possess the attribute a_i. (Note that not considering any particular attribute in the system is completely distinct from considering its absence in the system.) Let us denote by A_i the i^{th} row vector of R (and B_j the j^{th} column vector of R) and define $A_i \cup A_k$ as the n-dimensional row vector consisting of ones and zeros in such a way that if α_i^u and α_k^u are the u^{th} component of A_i and A_k respectively, the u^{th} component of $A_i \cup A_k$ is given by $\alpha_i^u + \alpha_k^u + \alpha_i^u \cdot \alpha_k^u \pmod 2$; i.e., if the the u^{th} component of A_i or A_k (or both) is one, then the u^{th} component of $A_i \cup A_k$ is one, otherwise zero. Also define $A_i \cap A_k$ as an n-dimensional row vector so that its u^{th} component is given by $\alpha_i^u \cdot \alpha_k^u$ if α_i^u and α_k^u are the u^{th} component of A_i and A_k respectively; i.e., the u^{th} component of $A_i \cap A_k$ is one if and only if the u^{th} components of A_i and A_k are both one, otherwise zero. We now define the similarity coefficient σ_{ik} of a pair of attributes a_i and a_k with respect to the given set B of objects by

$$\sigma_{ik} = \frac{N(A_i \cap A_k)}{N(A_i \cup A_k)} ,$$

where the numerator and the denominator of σ_{ik} are the number of ones in the vectors $A_i \cap A_k$ and $A_i \cup A_k$ respectively. Note that $N(A_i \cup A_k) \neq 0$, since a_i and a_k are considered to be attributes which are pertinent to B and hence must be possessed by at least one element of B. If $B^* \subseteq B$ is the subset consisting of elements possessing either the attribute a_i or a_k, and since $N(A_i \cap A_k) \leq N(A_i \cup A_k)$ so that $0 \leq \sigma_{ij} \leq 1$, σ_{ij} is exactly the probability of choosing at random, assuming equal likeliness, an element of the set B^* which has both the attributes a_i and a_k simultaneously. In a similar way we define the dual similarity coefficient s_{jh} of a pair of objects b_j and b_h with respect to the set of attributes A by

$$s_{jh} = \frac{N(B_j \cap B_h)}{N(B_j \cup B_h)}$$

Note that $s_{jj} = 1$; i.e., any object is perfectly similar to itself. The $n \times n$ matrix $S = (s_{jh})$ will be called the matrix of the similarity coefficients of the objects B with respect to the set of attributes A and dually $\Sigma = (\sigma_{ik})$ the $m \times m$ matrix of the similarity coefficients of the attributes A with respect to the set of objects B. Both S and Σ are symmetric matrices with ones along the principal diagonal. Note that mathematically, the problem of classifying objects with respect to attributes is exactly the same as the problem of classifying the attributes with respect to the set of objects. One problem will be called the dual of the other.

A simple geometrical interpretation of the significance of the matrix S is easily established if we consider the objects b_j, $j = 1, 2, \ldots, n$ as points in a semi-metric space H with the distance $d_{ij} \geq 0$ between the points b_i and b_j defined by

$$d_{ij} = -\log_2 s_{ij}.$$

Thus, if two objects b_i and b_j are very similar, i.e., s_{ij} is nearly one, then b_i and b_j, considered as points in H, are very close to each other in the sense that the distance between them is small so that our usual notion of closeness of two objects, attribute-wise, is carried over in a geometrical sense in H. If $d_{ij} < \infty$, we shall say that the point b_i is connected to the point b_j and if $d_{ij} = \infty$, then the point b_i is not connected to the point b_j. Thus the matrix (d_{ij}) defines a graph D in H.② If (g_{ij}) is the point-to-point incidence matrix determined by S, i.e., $g_{ij} = 1$ if $s_{ij} \neq 0$ and $g_{ij} = 0$ if $s_{ij} = 0$, then the incidence matrix determines a graph G which is homeomorphic to D. $0(b_i) = \sum_j g_{ij}$ will be called the ramification order or simply the order of the point b_i; i.e., $0(b_i)$ is the number of arcs emanating from the point b_i in G or D. Let us assume for the time being that there is at least one point of D whose order is n-1. We define the hierarchical power $H(b_i)$ of the point b_i by

$$H(b_i) = \sum_j d_{ij}.$$

Thus have we introduced an order in H so that the set B is a lattice L. In terms of information theory, $H(b_i)$ is the entropy of the system

associated with b_i.③ The element b_{i_o} determined by

$$H(b_{i_o}) = \min_i \sum_j d_{ij}$$

will be called the apex of the lattice L. b_{i_o}, in general, is not necessarily unique.

Theorem. The object b_{i_o} corresponding to the apex of L is the object which is probability-wise most similar to all of the other objects b_j, $j = 1, 2, \ldots, i_o - 1, i_o + 1, \ldots, n$.

Proof. Since all the s_{ij}'s are independent probabilities, and $H(b_i) = \sum_j d_{ij} = -\sum_j \log s_{ij} = -\log \prod_j s_{ij}$, we have $H(b_{i_o}) = \min_i \sum_j d_{ij} = -\max_i \log \prod_j s_{ij}$. Thus i_o is exactly that index determined by $\max_i \prod_j s_{ij}$, the maximal probability of the simultaneous occurrence of choosing at random attributes which are possessed by the object pairs b_{i_o} and b_j, $j = 1, 2, \ldots, n$, $j \neq i_o$, among those possessed by b_{i_o} or b_j, $j = 1, 2, \ldots, n$, $j \neq i_o$.

If all $0(b_j) < n - 1$ for $j = 1, 2, \ldots, n$, then the point (points) of maximal order is (are) considered as possible candidates as the apex (apeces) of L. If the maximum order of the points of D is $n - 2$, then the finite hierarchical power of those points is given by

$$H_{n-2}(b_i) = \sum\nolimits' {}_j d_{ij},$$

where $\sum'$ extends over all the indices j except those for which d_{ij} is infinite. The apex (or apeces) of H in this case is found by

$$\min_i H_{n-2}(b_i).$$

In general if max $O(b_i) = \nu$, then the apex (apeces) is found by

$$\min H_{\nu}(b_i)$$

where the b_i's range over the set of points whose orders are ν. Thus the classification is essentially complete, since all the objects are ordered by their hierarchical powers. The clusterings of the points b_j in the graph D in H determine the classification, and the radius of each of the clusters considered as a classification group is left entirely to the subjective judgment of the expert in the particular field.

Suppose now that the number of objects with which we are concerned is fixed, and we ask the following question: what are the $k<m$ most important attributes of the set of objects? This question is easily answered by considering the problem dual to the one above, thus giving the hierarchical powers of the attribute points a_i in the dual graph D* in H*, the dual semi-metric space. The attribute point with the maximal hierarchical power is eliminated from the matrix R, since that attribute is the one that is least significant. (If $H(a_i) = \infty$, then the point a_i for which $O(a_i)$ is a minimum is eliminated first.) Now the new matrices R* and $\sum$* of order one less are formed, and the process is repeated. This procedure is performed again and again until we have exactly k attributes remaining. In a similar fashion, if we are given a fixed set of attributes, we can easily find the $h<n$ objects which are best

represented as possessing these attributes. One can also perform a reduction of both attributes and objects simultaneously if desired. In practice one should have a sufficiently large matrix R so that attributes and objects may be eliminated by the above process to reduce R to the desired dimensions, thus completely avoiding the question of weighting of attributes or objects by their apparent subjective importance.

The problem in prediction or in diagnosis is the following: given a set of attributes, what object, among the considered objects, is this set of attributes most likely to represent? (e.g., in medical diagnosis: given a set of symptoms, what disease is most likely to be associated with this set of symptoms?) The solution to this problem is accomplished by augmenting the matrix R (R may or may not be found by the above-mentioned elimination process) so that the last column corresponds to the unknown object x with the absence or presence of its attributes a_i. Upon performing the previously mentioned calculations, the point x in the graph D' will lie in some cluster of points, thus identifying it with that particular group. The similarity coefficients of x with all the other objects are the probabilities that x is each of the objects. If the sum of these probabilities is very small when compared to one, this would indicate the lack of other objects in setting up the original classification in order to justify classifying x in the system.

*The problem was originally proposed to the author by Dr. David J. Rogers of the New York Botanical Gardens.

(1) P.H.A. Sneath, J. Gen. Microbiology, 17, 201, 1957.

(2) D. Konig, Theorie der Endlichen und Unendlichen Graphen, Chelsea, New York, 1950.

(3) C.E. Shannon, A Mathematical Theory of Communication, Bell System Tech. Journ., 27, 379, 623, 1948.

November 17, 1958