AN APPROACH TO THE FORMALIZATION
OF CLASSIFICATION

A. A. Dorodnitsyn, M. F. Kaspshitskaya, and I. V. Sergienko

UDC 550.93.551.241(470.477)

A method is proposed for the classification of objects defined by a large ensemble of attributes, in application to problems of the taxonomy of biological objects.

1. INTRODUCTION

A characteristic feature of classification methods applied in natural sciences, for example, in botany and zoology, is the use of a small number of attributes to partition the set of investigated objects into classification groups. This position is imposed, and it is explained by the limited information capacity of the human brain, which is not able to grasp directly a large number of possible groups, arising in the use of a large number of attributes. If a small number of attributes is used, then all of the cells in the attribute space are filled with elements of the set of objects to be classified and, since all of these cells are “equiprobable,” the partitioning of the set of objects into classification groups is totally dependent on the choice of a small number of defining attributes. The choice of the defining attributes is to a substantial degree subjective. This explains the existence of many classification systems, proposed by different authors — it is difficult to find two books, say on zoology, in which the same classification system would be used.

The only correct classification system would be one based on genetic couplings. Unfortunately, biologists are not supposed to construct a classification system from a known genetic tree. On the contrary, their problem is

establish the genetic tree, and to this end the construction of an objective classification system, based on a complete (or at least sufficiently great) ensemble of attributes, would be of great assistance.

Below, in order to make clearer the sense of the approach proposed here, we shall consider a case where each attribute has two values, i.e., may be represented by a binary code − 0 or 1. Then, if n attributes are used for classification, the attribute space will consist of $2^n$ cells. And here two essentially different situations can arise. Let the set $\mathcal{M}$ of objects consist of m distinct elements. If $2^n << m$, then the elements of the set $\mathcal{M}$ fill all the cells of the attribute space. Assume that all attributes have equal weight. Then all the cells of the attribute space have equal weight, there are no formal grounds for grouping them, and the only classification system will be isomorphic to the attribute space (this position is obviously not changed if any of the cells are empty). We shall call such a classification system trivial.

A different situation arises when $2^n >> m$. In this case the elements of the set of objects $\mathcal{M}$ fill only a part of the attribute space and their position in this space can give a basis for grouping. The exposition of a grouping procedure is the subject of this article.

The first situation is characteristic of all classification systems existing in zoology and botany. The number of attributes (not always binary, to be sure), as a rule, is not large and hardly ever exceeds five.

It should be noted that biologists do not adhere dogmatically to classification according to selected “defining” attributes. Thus, for example, mollusks of the class Gastropoda are divided by the type of respiratory organs into three subclasses: Prosobranchia, Opisthobranchia, and Pulmonata. In this case the classification is conducted according to only a single attribute, having three values. But, for example, the ground mollusks of the genus Pomatias, which do not have gills, but have a respiratory organ fulfilling the role of the lung, are assigned by zoologists to the subclass Prosobranchia, since in a number of other anatomic attributes they are very different from other Pulmonata, but very close to the marine family Littorinidae in the subclass of Prosobranchia. Here the classification is conducted by a small number of attributes, to be sure, but with regard to other attributes. Unfortunately, the process of such “regard” is not at all formalized. Therefore, our remark about the subjectivity of existing classification systems relates all the more to classification methodology. As far as concerns its actual realization, here (consciously or unconsciously) there holds a certain “objectivization,” and it can be considered that when higher classificational groups (types, classes, and, apparently, subclasses) are concerned, then no doubt arises in the objectivity of the classification. The lower the degree of classification, the greater the subjectivity, the less stable (in the sense of “general acceptance”) is the classification system, and the greater the number of different variants. In partitioning into groups, considerations of convenience begin to have a strong influence – it is more convenient to manipulate a relatively small set of elements in each classificational group. It is therefore often possible to observe that in the classification of a large number of groups of animals the representatives of the various families have much greater common features than certain representatives of the same genus in a less numerous group.

2. FORMAL CLASSIFICATION PROCEDURE

Let us proceed to the description of the process of formal classification. Formal classification begins after a system of attributes of equal weights has been established and the set of objects to be classified is encoded. The person making the formal classification (let us call him, by convention, a mathematician) is concerned only with the codes of the objects. The concept of “equal weight of attributes” is assumed to be intuitive; this concept is established by a specialist, studying the given ensemble of objects (let us call him, by convention, a biologist). From the viewpoint of the mathematician the attributes need only be logically independent. It is impermissible, for example, to have an attribute $i$ whose values signify the following: $0 =$ “does not breathe with gills,” $1 =$ “breathes with gills,” and a second attribute $k$ with values as follows: $0 =$ “does not breathe with lungs,” $1 =$ “breathes with lungs.”

It is not permissible, however, that the system contain physically dependent attributes. But the establishment of the physical dependency or independence is already the affair of the biologist. What such a physical independence is can be explained by the example of the classification of the limnetic bivalvular mollusks, given at the end of this article. One of the attributes taken by us is the presence of the byssus: $0 =$ byssus absent, $1 =$ byssus present. If, further, the attribute of the mode of life is introduced − $0 =$ freely moving, $1 =$ fixed to the substrate − formally these features will be independent, physically the byssus is the organ of fixation to the substrate, and its presence is equivalent to the fixed mode of life.

Thus, the biologist composes the system of attributes, their semantic description (for himself), and encodes the list of objects, assuring at the same time the condition $2^n >> m$, as well as the logical and physical independence of the attributes.

System of attributes:

attribute 1: $0 =$ (description of the meaning), $1 =$ (description of the meaning),
attribute 2: $0 =$ (description of the meaning), $1 =$ (description of the meaning),

945