AN INTRODUCTION TO ASSOCIATION ANALYSIS

J. M. RANDAL

Applied Mathematics Division, D.S.I.R.,
Wellington

SUMMARY: This is a brief, non-mathematical survey of some of the most common methods of numerically classifying large numbers of individuals according to their attributes. The article is intended as an introduction to the subject, and is not intended to be in any way definitive. An appendix briefly discusses the results of a simple classification exercise.

INTRODUCTION

Given a set of individuals (be they people, paddocks, or pumpkins) each possessing a number of attributes, one can arbitrarily classify them into a number of related but distinct non-overlapping groups. However, there are two major difficulties: effecting the classification and justifying it before one's colleagues. In ecological circles one has no trouble in finding sets of more than a thousand individuals to classify or at least to put into some kind of rational order, and if each of these individuals has 50 or so attributes this task is not trivial even if a computer is to be used to do it.

Association analysis is one of the techniques which, coming to flower at the beginning of this decade, have enabled the taxonomist to handle large arrays of taxonomic data, and the name has to some extent become a generic term for a whole series of variations on the theme of this aspect of numerical taxonomy. In this paper I hope to describe them briefly, without the aid of mathematics, and to give a computer scientist's view of their uses and advantages.

METHODS

Generally speaking, data are presented for analysis in the form of a rectangular array of numbers. In some instances the numbers may be represented by symbols such as +, —, * etc. For the purposes of this paper, each row represents an individual, each column an attribute. Sometimes it is preferable to transpose the array (that is represent the individuals as columns of attributes) but since transposition is merely a mechanical procedure the orientation of the array is immaterial for all present purposes. For any individual, the numerical representation for each attribute may take a number of forms:

Presence-absence

The presence or absence of a particular phenomenon is recorded as a 1 or 0 respectively. Sometimes if the result of a particular test is inconclusive this may be recorded as −1.

Multi-state

This is a generalisation of presence-absence. Given a number of possible outcomes for a particular test, say four, then the score for that particular attribute for the particular individual may be noted as a number between 1 and 4. These states do not have to be mutually exclusive, although it is easier if they are. They usually represent qualitative judgments, e.g. blue eyes . . . 1, green eyes . . . 2, brown eyes . . . 3, pink eyes . . . 4.

They do not have to be ordered in the sense that small responses are represented by small numbers, although this is often convenient.

Quantitative

The score is an actual measurement; that is, the length of a stem, the result of an examination and so on.

Very often, individuals are described by an assortment of these attribute types and, if computing facilities are not capable of accepting this mixture, quantitative data are often reduced to presence-absence by coding a variable as 1 if it is above the mean for that attribute and as 0 otherwise; or, if non-zero scores are infrequent, as 1 for these and 0 otherwise. Often qualitative data must also be reduced in a similar manner but this is clearly a more difficult and intuitive task. However, the main point is perhaps, that the methods to be described are remarkable insofar as they can be made to work on any kind of data regardless of distribution type or scaling; and that, providing that care is taken not to throw away too much information, the data can often be simplified without unduly prejudicing the results.

In order to classify the data numerically, one numerical and two procedural rules should be specified:

Similarity statistic

Given any two individuals in the data array, some measure of their similarity must be calculated. One popular statistic is the distance between the individuals, assuming they represent points in a many-dimensional space; another is the sum of the absolute difference between corresponding attributes. If the data are all quantitative, the correlation coefficient between individual rows is often taken; whereas exclusive multistate and presence-absence data lend themselves to a statistic which estimates the amount of information gained if the two individuals in question were to be fused. Different taxonomists tend to have their pet statistic, or subject their data to a succession of tests, each with a different statistic.

Grouping rule

After having applied the similarity statistic, some action must be taken on the basis of the information it supplies. This depends on the method of classification employed. Some methods use the grouping rule to control the fusing of individuals into groups, other to split groups into subgroups. Generally there is more than one grouping rule available for use with each method, and once again, the method chosen is one of personal preference.

Stopping rule

When using automatic computers, some rule must be laid down about when to stop the classification procedure, either because the information on which the classification is based has been exhausted or because it would be uneconomic or undesirable to proceed further.

Although the methods using these rules fall into three main classes, a host of variants exist. Many of them are direct applications of hand methods to automatic computers, others are variants that have had to be developed because of the shortcomings of particular computing installations.

The three major streams are as follows:

Agglomerative method

Initially, each individual is regarded as a group of one. The similarity measure between each individual and every other is calculated, and the most similar individuals are fused to form a group of two. This new group is regarded as an individual and the two individuals that compose it are forgotten. There is now one less group than there used to be. The process is repeated again and again until there is only one group comprising all the individuals left. The classification process is then complete, and is interpreted by reading the classification history backwards, (see Appendix). The method is simple and well defined; although if it is to be programmed for a computer, presenting the results of the process requires some effort on the programmer's part. Its main disadvantage is that the classification is not necessarily unique. Two pairs of groups or individuals may be equally similar, and the choice of one pair rather than another to begin the classification process may significantly alter the resulting grouping. Usually, the occurrence of this phenomenon may be detected by presenting one's data a second time with the individuals arranged in reverse order. The classification also depends on the measure of similarity and the method of fusing groups used.

Those statistics which are a derivation of the concept of the distance between two individuals (in terms of their attribute scores) have the advantage of being intuitively understandable as well as being applicable to most kinds of data. "Statistical" measures, such as the "information statistic" which tries to extract the most information from the data presented have also been very successful, but are not easy to apply to anything but presence-absence or, at a stretch, multi-state data.

When groups or individuals are fused with others to form bigger groups, the resulting group or its similarity to all the other groups is usually represented as if it were an individual whose attributes reflect the attributes of those individuals that comprise it. The centroid strategy gives the resulting group attributes which are each the "centre of gravity" of the corresponding attributes of the member individual. The "nearest neighbour" strategy provides attributes that will give a similarity measure between groups equal to the measure of similarity between the nearest neighbouring individuals in each group, and so on.

Often quite different groupings may be obtained by using different combinations of statistic and strategy.

Divisive method

All the individuals are supposed at first to belong to the one group. The attributes are correlated each with every other in some way. The attribute with the least correlation with all the others is used as a key to split the group. One subgroup possesses a certain measure of the attribute, the other does not. A similar process is then applied to each subgroup and so on until all the attributes are used up in all the subgroups, or — more usually — until all the groups shrink to a required size, or until a required number of divisions has taken place.

Once again, it is likely that two attributes will be equally uncorrelated, and the grouping produced is thus not necessarily unique.

This method is usually used with presence-absence data by calculating chi-square for each pair of attributes and then dividing the group according to the presence or absence of the attribute yielding the highest chi-square value. However, the method may be applied to quantitative data if one is prepared to accept the additional responsibility of defining the level of the attribute, the possession of which will decide the fate of any individual.

Apart from the difficulties in keeping track of where each individual is when using a computer to do the job, this approach suffers from the choice of dividing on the basis of one key attribute at a time. Methods for dividing on more than one attribute are, however, being developed. There is also the difficulty of deciding when to economically halt the classification process: to carry it too far results in a waste of effort in obtaining meaningless divisions of small groups, to stop it too soon courts the annoyance of missing out on that really important division that was sure to come next.

Gradient methods

These accept a given (and often arbitrary) grouping of individuals and try to improve it by altering the grouping (often by transferring one individual from one group to another), and accepting the new grouping if it is better than the old. A similarity measure is used in conjunction with some criterion of group compactness to find the "loosest" group. The loneliest individual in the group is assigned to another group. If the grouping structure is more compact than before, this scheme is accepted and the process is repeated. The measure of compactness is represented as a number, and the object of the game is to reduce this number as much as possible. This downhill motion gives rise to the word "gradient". The main disadvantages with this method are: That there is no guarantee how long the process will take (indeed under some schemes a cyclic set of alterations might arise which will cause the process to continue indefinitely); the process may never find the most compact grouping but may get hung up on some local minimum of the measure of compactness; lastly, gradient methods seldom leave behind the useful dendrogram (or hierarchy) that the former methods often produce. On the other hand they may well be useful for refining groupings that have been found by other methods.

There are all manner of variations and combinations of these three major approaches, as well as numerical classification methods that do not fit into any of the three most common mentioned above. The discussion of these methods however should give some idea of the general approach to numerical classification procedures in general.

CONCLUSION

The overall advantages that these methods offer are that trial classifications of large amounts of data may be made cheaply, quickly, and without too much initial analysis. The methods work well without any assumptions of the distributions of the attribute scores and are capable of producing results in an easily interpreted form. However, these strengths are also weaknesses, insofar as any classification produced is not statistically "respectable" because no adequate statistical tests have yet been devised to give a thorough measure of the confidence that one may put in a given classification similar to the ones available in, say, factor analysis. The best approach in the face of this shortcoming is still to justify one's classification in traditional terms of the original data. For, although several different classification methods often give a fair degree of agreement, some individuals often attach themselves to radically different groups for different classifications.

Because one may choose both the statistic and the strategy with which to effect a classification, it is clear that these methods should be regarded as consistent rather than objective. This is especially true when one remembers that one must choose the attributes which are to provide the information for the classification. Nevertheless, numerical methods such as those described above are useful for suggesting plausible classifications of large numbers of individuals; and since the subject of numerical taxonomy is very lively, we may hope to see some considerable improvements in all its areas in the near future.

REFERENCES

KERSHAW, K. A. 1967. Ecological methods and computers. *Sci. Prog. Oxford* 55: 437–451.

LANCE, G. N. and WILLIAMS, W. T. 1967. A general theory of classificatory sorting strategies: 1, Hierarchical systems. *Computer. J.* 9: 373.

LANCE, G. N. and WILLIAMS, W. T. 1967. A general theory of classificatory sorting strategies: 2, Clustering systems. *Computer. J.* 10: 271.

LANCE, G. N. and WILLIAMS, W. T. 1967. Mixed data classificatory programs: 1, Agglomerative systems. *Aust. Computer J.* 1: 15.

LANCE, G. N. and WILLIAMS, W. T. 1968. Mixed data classificatory programs: 2, Divisive systems. *Aust. Computer J.* 1: 82.

LANCE, G. N. and WILLIAMS, W. T. 1968. The choice of strategy in the analysis of complex data. *The Statistician* 18: 31.

MACNAUGHTON-SMITH, P. 1965. Some statistical and other techniques for classifying individuals. *Studies in the causes of delinquency and the treatment of offenders* (6). H.M.S.O., London.

SOCAL, R. R. and SNEATH, P. H. A. 1963. *Principles of numerical taxonomy.* W. H. Freeman and Co., San Francisco.

APPENDIX

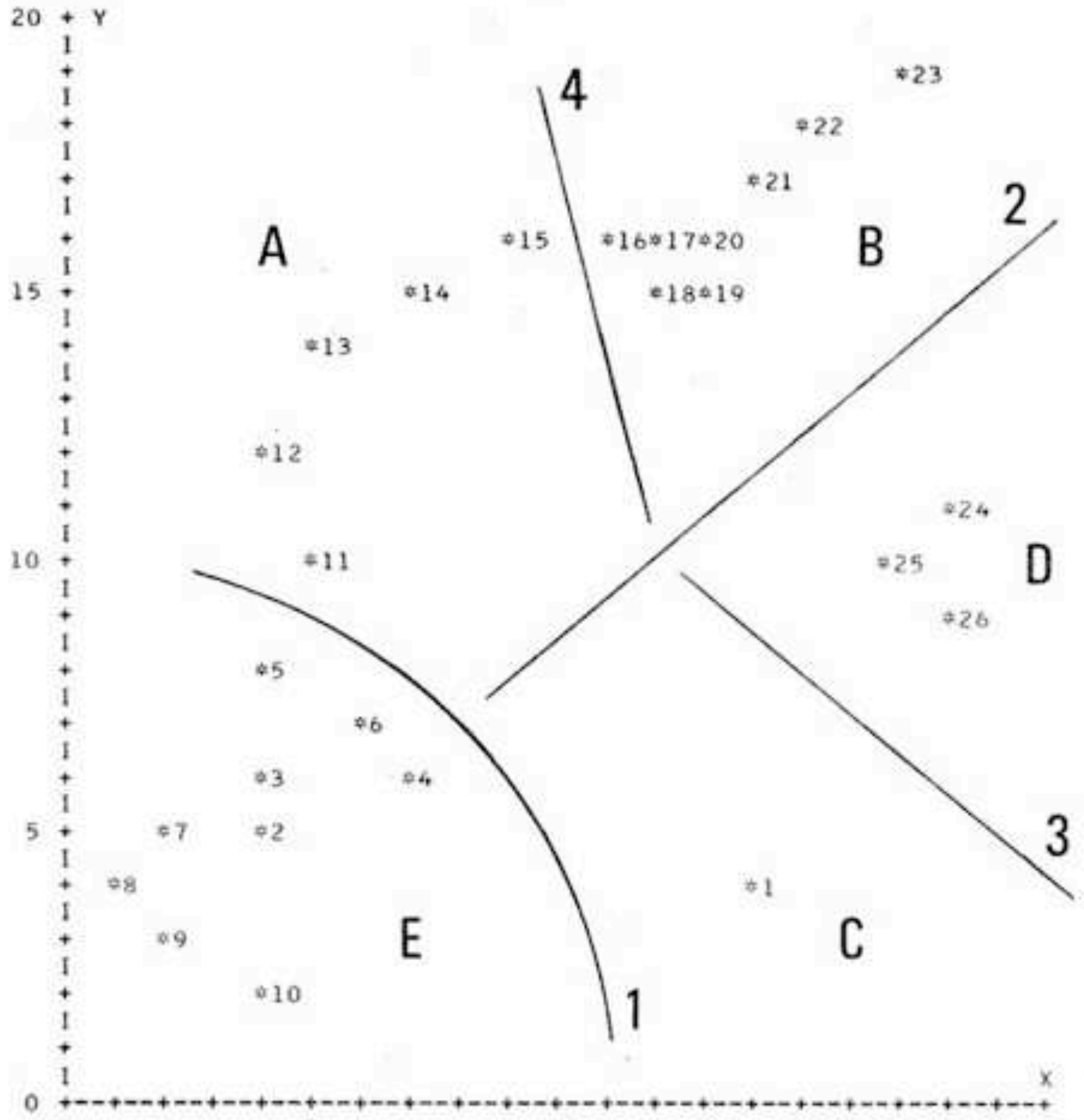

FIGURE 1. *A hypothetical distribution in two dimensions according to centroid strategy.*

An example

Suppose, on some hypothetical featureless swamp, there exists a scattering of 26 nesting sites. Since these have been identified from the air, only their positions have been noted, and these are shown, plotted relative to an arbitrary origin, in Figure 1. Do the nests have a grouping pattern? If so, it would be desirable to organise a trip to each group, rather than visit every site. Each site has two attributes, a distance east, x, and north, y, of the datum. We may classify the sites in terms of these attributes. Let us say that the sites are similar if they are close together; dissimilar if they are far apart, and let us use the square of the distance as a measure of this similarity. We will classify the sites according to the agglomerative method using the 'centroid' statistic. That is, if sites 17, 18, 19, 20 were to form a group, they would be represented as a site in the middle of the square which they form with four times the importance of the component sites.

The computer programme used to effect the classification gives (among other things) two forms of results:

A hierarchy (or dendrogram) as shown in Figure 2. It is formed from the right left; that is, the sites cluster into small groups (24, 25, 26) which, in turn, group together to form larger groups. The square of the distance between the centres of gravity of the groups when they do merge is proportional to the distance between the vertical line which represents the merging and the right-hand margin of the dendrogram. If we interpret the dendrogram from the left right we see that the sites may be split firstly as two big groups divided by the line marked 1. The larger group is split by the line 2 and so on. The first four divisions are marked on both the graph and the dendrogram. There seems little point in considering any further sub-division. Although few would argue about the composition of the groups marked C and D one may appreciate that different opinions may exist about the proper boundaries between groups A and B and A and E. However, the classification by and large is helpful.

A similarity matrix, as shown in Figure 3. This is a symbolic indication of the squared distance between each

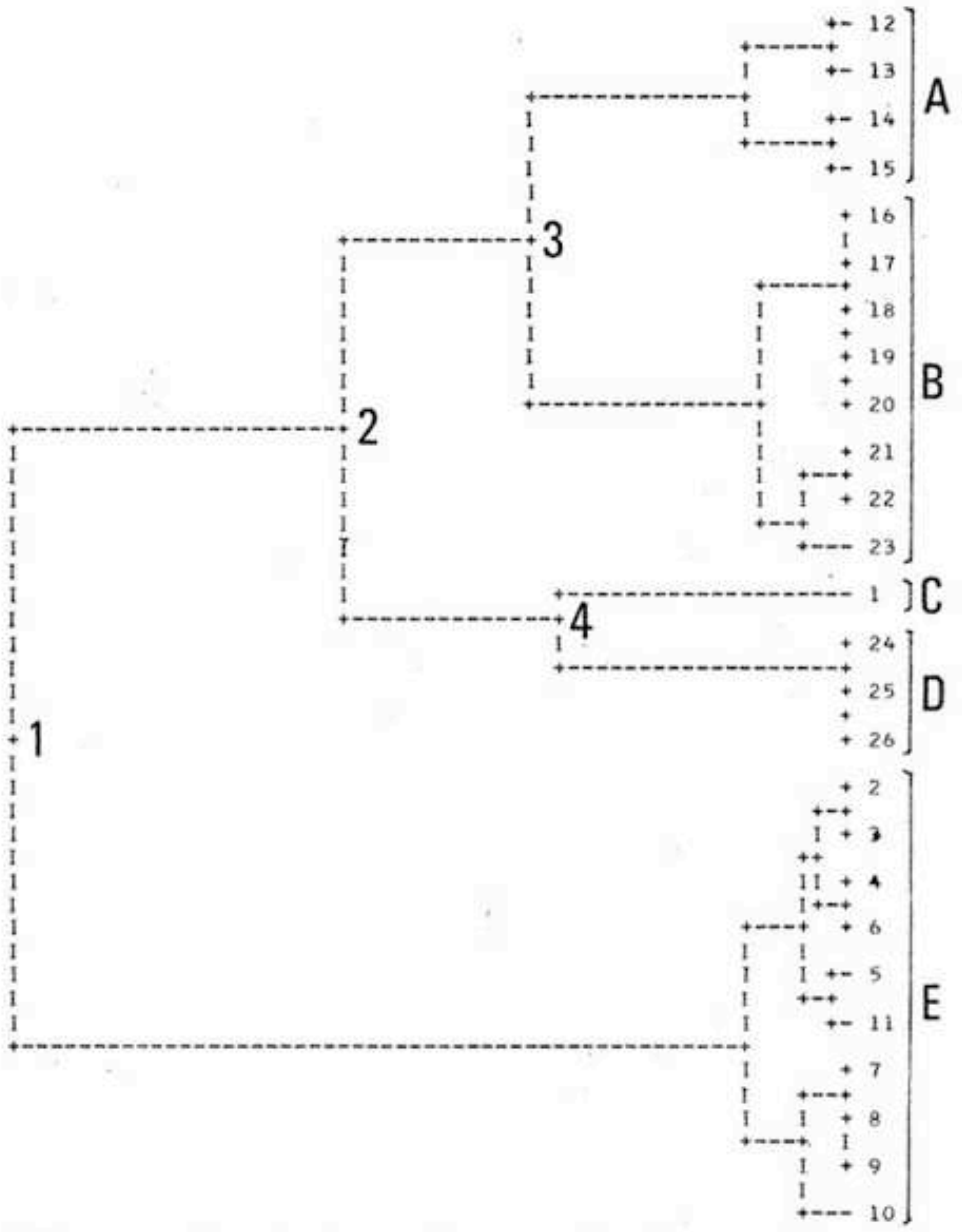

FIGURE 2. *A hierachy or dendrogram of the centroid classification in Figure 1.*

site and every other, after they have been classified and sorted so that the closest sites lie closest together as along the top of the dendrogram. Thus, groups tend to be displayed as triangular areas near the diagonal of the matrix. The groups are marked along the diagonal and the site numbers down the left and along the bottom of the matrix. The similarity matrix gives more information about the relationship between the groups than the dendrogram does. The scatter of high scores in the bottom left-hand corner of the matrix indicates that groups A and E are more closely related than groups B and E; an indication not always shown on the dendrogram. The greater number of high scores in group B indicates there is greater compactness compared with group E. This conclusion is supported by the dendrogram but it would be hard to visualise if each point had more than two attributes.

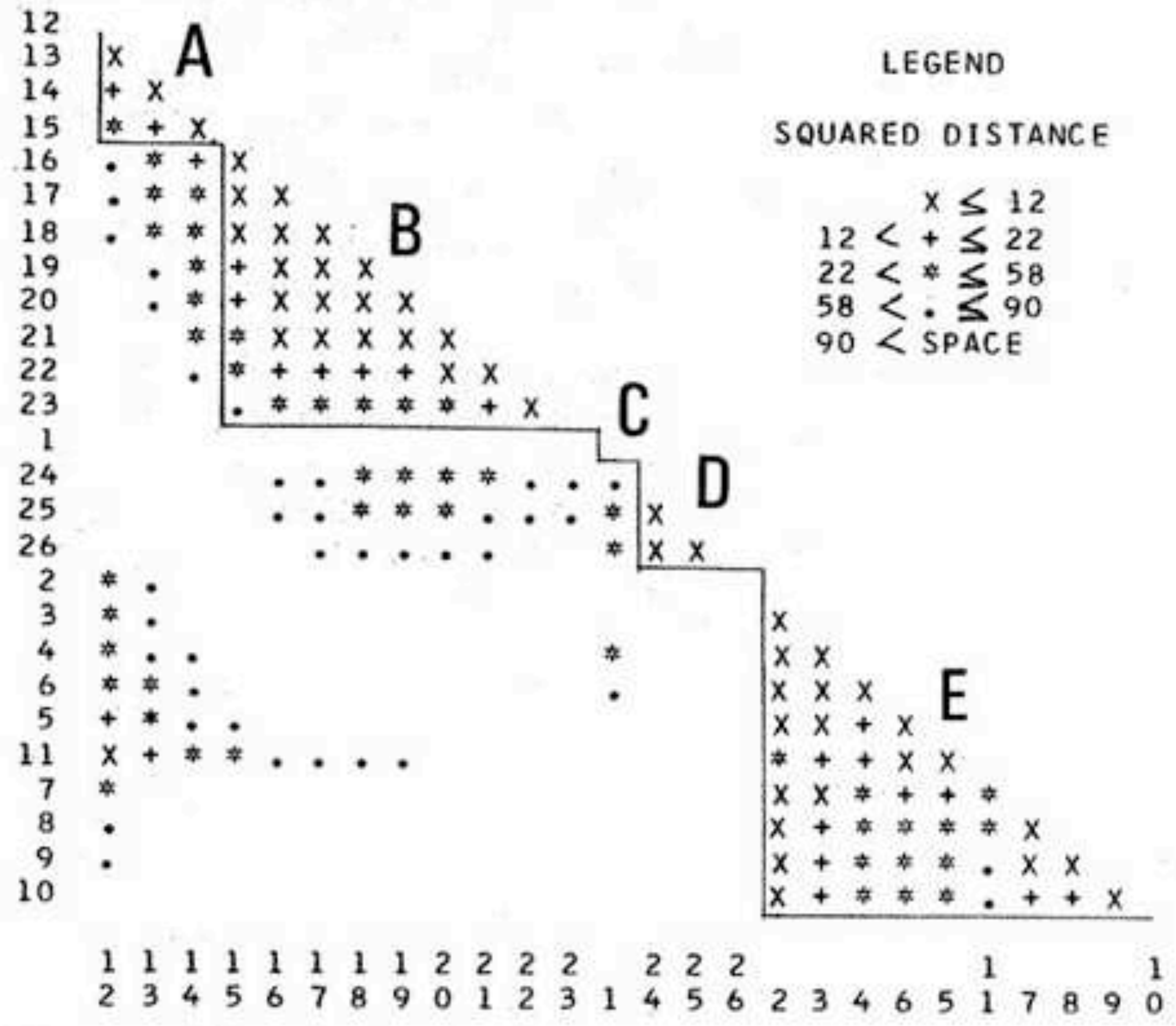

FIGURE 3. *A similarity matrix of the centroid classification in Figure 1.*

Discussion

Figures 4 and 5 show the results of classifying the sites with the same statistic but with the intergroup distance taken as being the distance between the nearest neighbour of each group. Note that not only do none of the groups correspond exactly with the centroid strategy, but that the four major dividing lines are made in a very different order. If one is to choose between these two classifications one must invoke more information than was given by just the co-ordinates of the sites. Such information would be the size of the site or any difference of flora associated with it. The addition of this would make classification 'by eye' more difficult, but one could apply the same methods as used above to the extended data. If the same site information is subjected to a further classification wherein the distance between two groups is taken as being the distance between the furthest neighbours of each group, yet another pattern emerges in which groups C, D and the rest are all given equal status by the first two dividing lines, and the next two divisions divide group E into three. One cannot deny that plenty of ideas for possible groupings may arise, but the production of a numerical classification is not evidence for accepting that classification. Note, however, that these divergent schemes would be less likely to occur if each site possessed more attributes.

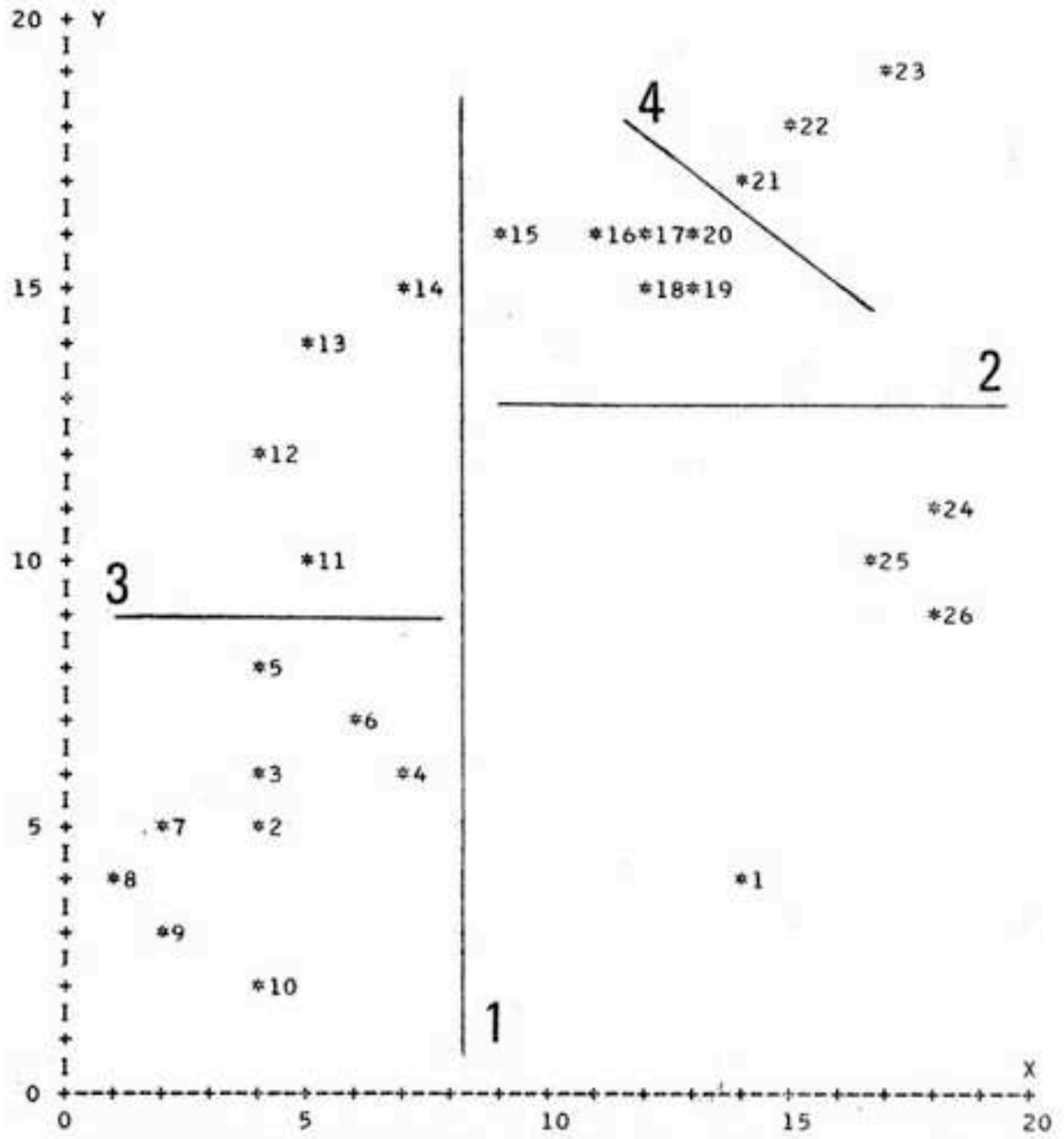

FIGURE 4. *Data in Figure 1, classified according to "nearest neighbour" strategy.*

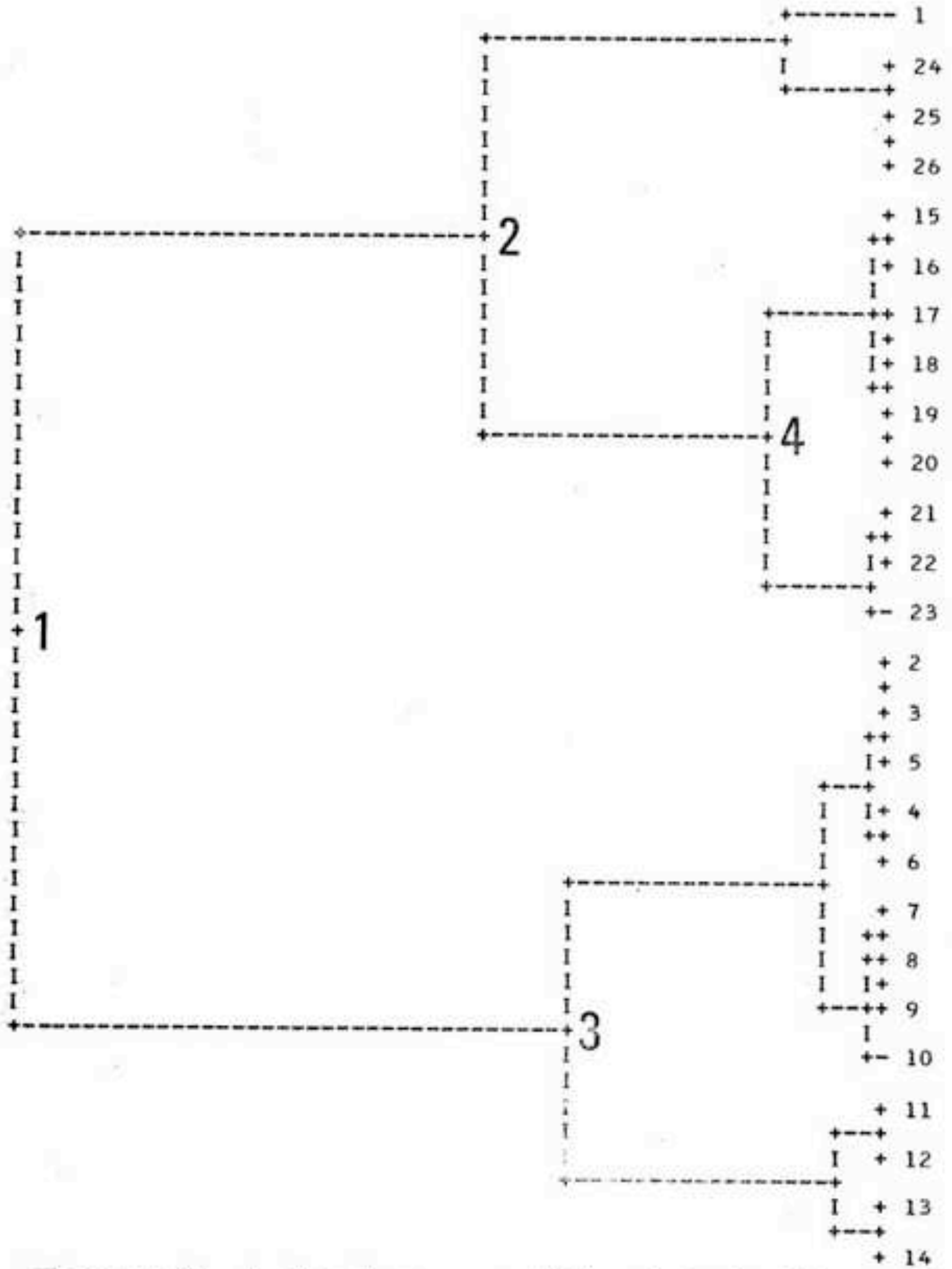

FIGURE 5. *A dendrogram of the data in Figure 1, classified according to the "nearest neighbour" strategy.*