Pattern Recognition Pergamon Press 1971. Vol. 3, pp. 269–280. Printed in Great Britain

Behavioral Problems of Deaf Children: Clustering of Variables Using Measures of Association and Similarity

ROBERT M. HARALICK and JOY GOLD HARALICK

Center for Research, Inc., University of Kansas, Lawrence, Kansas, U.S.A.

(*Received* 9 *February* 1970 *and in revised form* 3 *September* 1970)

Abstract—A statistical measure of similarity distinct from a measure of association is introduced. By this measure two things are highly similar, if and only if their respective associations with all other things are the same. In this paper, measures of association and similarity are combined to perform a cluster analysis of variables concerned with behavioral problems of deaf children. The four major clusters which emerge are identified as characteristics of anxiety, hot temperedness, inattentiveness, and social withdrawal. The clusters obtained by the combined measures are compared with the clusters obtained by each measure alone.

INTRODUCTION

THERE are numerous instances of the use of association measures in data analysis. Sometimes these measures are called "association," sometimes "similarity," but in almost all cases the measures indicate a degree of co-occurrence or contiguity between two data items or characteristics. (GOODMAN and KRUSKAL,[1,2] have a comprehensive discussion of association measures). However, the degree of co-occurrence or continguity is not the only type of statistical relatedness that can exist between two things.

Another type of statistical relatedness is concerned with sameness or similarity, and it is of an entirely different character than association. A clear example of this difference occurs in the information retrieval area where one is concerned with the relationship between index terms used to characterize documents.[3,4] For instance, there will be a high pairwise association between terms like electromagnetic and wave, frequency and wavelength, neutrinos and lifetime, acceleration and velocity and high similarities between terms like acoustics and sound, beams and rays, bright and intense, constant and fixed, absolute and relative. Highly associated terms are different aspects of one area of knowledge, while terms with high similarity are terms which are synonyms, near synonyms, or antonyms. Similar words are words which occur in the same or similar context; LIBBEY,[5] LEWIS, BAXENDALE and BENNETT,[6] ROSENFELD, HUANG and SCHNEIDER,[7] and STILES[8] have suggested measures of such similarity.

Both the measures of association and similarity are convenient for clustering variables. In such clustering procedures one attempts to group together in clusters those variables which are measuring the same processes or subsystems in an environment. The grouping is usually done as follows: those highly related variables are put in the same cluster while those which are unrelated are put in different clusters. Hence, clusters are maximal groups

of highly related variables. BALL,[9] SOKAL and SNEATH,[10] BONNER[11] discuss various clustering procedures, and GASKING[12] has an excellent discussion of the cluster concept for those readers who would like an in-depth study of this subject. In this paper we use a measure of similarity together with a measure of association to cluster some behavioral data taken of audially handicapped children.[13]

GENERAL DISCUSSION OF CLUSTERING

The scientist, in attempting an understanding of the environment which he chooses to study, decides on a set of N instruments, or variables $x_1, x_2, \ldots, x_N$ which he considers to be of probable relevance. When the instrument or variable x_i is applied to measure a given unit u or sample member in the environment, it takes on some value $x_i(u)$ in its possible range set of values L_i. The scientist chooses M sample units $u_1, u_2, \ldots, u_M$ for the purpose of gaining information about the environment of which the units are supposed representative. Each of the sample units is measured by each of the N instruments or variables. The data sequence D, thus obtained, is a sequence of M measurements, each measurement being an N-tuple, the N depending on the number of variables used.

$$\begin{aligned} D = \langle & [x_1(u_1), x_2(u_1), \ldots, x_N(u_1)], \\ & [x_1(u_2), x_2(u_2), \ldots, x_N(u_2)], \\ & \vdots \\ & [x_1(u_M), x_2(u_M), \ldots, x_N(u_M)] \rangle \end{aligned}$$

The first type of examination the scientist usually makes of the data sequence D is one which determines the strength of relationships between the instruments or variables $x_1, \ldots, x_N$. The idea behind such an examination is to determine those processes or subsystems within the environment which are seen as distinct by the instruments. The underlying hypothesis is that those instruments or variables which are highly related are measuring characteristics having to do with the same environmental processes or subsystems. To determine subsystems or processes, the variables are usually grouped on the basis of strength of relationships, and each distinct group or cluster of variables is identified with a separate subsystem or process. Ideally, each group is a subset of variables all highly interrelated with one another and minimally related to those variables outside the group. The groups of variables are also minimally overlapping.

One way to begin grouping the variables is to pick out those pairs of variables having the highest association, and put them in the same group. In doing this, we construct a binary relation R on the set of variables X.

At this point it is well to make some basic formal definitions. A set A containing the elements $a_1, a_2, \ldots, a_T$ will be written as $A = \{a_1, a_2, \ldots, a_T\}$. A Cartesian product of two sets is the combination of each element in each set with each element of the other set. It is a cross-product set. The Cartesian product $A \times B$ of a set $A = \{a_1, a_2, \ldots, a_T\}$ and a set $B = \{b_1, b_2, \ldots, b_S\}$ is defined by

$$\begin{aligned} A \times B = \{ & (a_1, b_1), (a_1, b_2), \ldots, (a_1, b_S), (a_2, b_1), (a_2, b_2), \ldots, (a_2, b_S) \\ & \ldots, (a_T, b_1), (a_T, b_2), \ldots, (a_T, b_S)\}. \end{aligned}$$

A set R which is a subset of $A \times B$, $R \subset A \times B$, is called a binary relation from A to B. A set R which is a subset of $A \times A$, $R \subset A \times A$, is called a binary relation on A. If ρ_{ij} is the measure of association between variable x_i and variable x_j, the binary relation R on the set X, which we have constructed by picking out these pairs of variables with high enough associations is defined by

$$R = \{(x_i, x_j) | \rho_{ij} > \theta\}$$

(read as R is the set of all pairs of variables (x_i, x_j) whose association ρ_{ij} is greater than θ). BONNER[11] has suggested an algorithm suitable for the computer which determines "clusters" from the relation R. He considers an ideal core cluster to be a subset C of X such that $C \times C$ is a subset of R, and C is maximal. Because this definition of core is too strict to be a cluster, (it allows many small but ideal highly overlapping subsets each to be called a cluster) his algorithm determines clusters by starting with the cores and enlarges them by adding to them other cores which highly overlap with them.

Because our present concern is with the determination of R, using both association and similarity measures, rather than with an algorithm for the determination of the clusters, we will discuss a direct intuitive approach to the formation of clusters from the relation R. Consider the nature of R: R is a set of ordered pairs and a listing of the many ordered pairs of variables which compose the relation R is a poor way of obtaining the clustering information, because the list may be so long that it cannot be meaningfully absorbed. The problem here is not with the adequacy of the information but with the way in which the information is displayed.

One solution to the display problem is to display R as a single picture of arrows connecting numbered circles or nodes. Such a picture is called a digraph. The nodes represent elements in X, and the arrows represent elements in R. If (a_i, a_j) is in R, then an arrow is drawn connecting node a_i to node a_j. An example of this is provided by Fig. 1. If the relation R is symmetric, that is if $(a_i, a_j) \in R$ implies $(a_j, a_i) \in R$, the two headed arrows in the picture may be replaced by a line. The picture is then called a graph.

Drawing a graph or digraph of the relation R does not necessarily guarantee that its information is more readily absorbed than it would be if it were presented in list form.

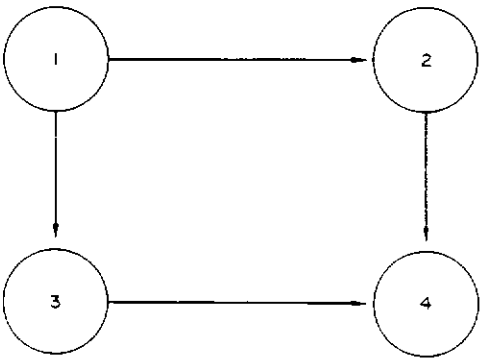

FIG. 1. Illustrates the digraph for relation R.

However, one has better opportunities to organize the picture than to organize the list. For example, if the graph is organized so that those nodes which are all highly interconnected remain close together, then the crossing of lines will be minimized, and the respective clusters of variables will become readily apparent. This kind of organization shows the clusters (highly interconnected nodes) and also shows the bottlenecks, which are nodes or lines lying between clusters. We call such nodes and lines bottlenecks, because if we imagine automobiles traveling from node to node via the lines, traffic bottlenecks will occur on those nodes and lines lying between the clusters.

Once a graph is organized in a cluster-bottleneck fashion, the nodes are interpreted in accordance with the following hypothesis. The nodes in a single cluster represent variables or instruments measuring one environmental subsystem or process. The nodes which lie between clusters represent instruments or variables measuring a common characteristic of two environmental processes or subsystems. Those nodes which are connected to each other, but which lie in two different clusters, represent variables measuring characteristics of a third less important process which overlaps the two processes represented by the two clusters. Within a given cluster, the nodes which connect to the maximal number of other nodes in that cluster, indicate characteristics which are the focus or essence of the process represented by the cluster: they give the cluster the "name" which designates the underlying dimension being measured.

SIMILARITY

Our general discussion would be over if its interpretation were limited only to nodes in a graph constructed from a relation derived from association coefficients. The weakness behind such a limited interpretation is that some variables (nodes) may each be measuring characteristics which are indirectly related to more than one environmental subsystem. As a matter of fact, it is probable that this is the usual case, because one subsystem usually controls or is controlled by another subsystem. Such indirect relationships cause high associations, and subsystems whose respective variables are highly associated will not be separated by the clusters of an association graph. Therefore, another measure of relationship must be used in conjunction with an association measure to enable a distinction to be made between highly associated subsystems. We suggest that one such useful relationship measure is that of similarity.

Similarity is a relationship of commonness. Once the notion of "common with respect to" is defined, the ordinary language notion of similarity works in the following way. The extent to which two things have common parts, elements, or relations is the extent to which they are similar. For us, common will mean common associations. Two variables which have almost identical associations with all the remaining variables will be highly similar. TRYON,[14] in 1939, used this idea of similarity in his correlation profile analyses.

Completely associated variables are completely similar; highly associated variables are generally highly similar. However, variables which are highly similar are not necessarily highly associated. The concept of functional similarity in biology, anthropology and sociology is an example of cases where variables may even be mutually exclusive but be highly associated with the same other variable, and thus *with respect to that variable* highly similar. For example, God-directed prayer and magical incantation are similar in their relationship to personal stress. They tend however, to usually be more or less exclusive of each other in that they arise in different social groups.

Unlike association, similarity is a context dependent concept. In our case, the context is the variable set X: similarity depends on the set of variables chosen to measure the environment.

Association is context independent, because it is only defined on the basis of the pairwise joint distributions of the pair of variables involved. This joint distribution is the same regardless of which other variables one choses to put in the variable set. Similarity, because it involves the comparison of all the inter-variable associations, must certainly depend on which variables are put in the variable set.

There are obvious disadvantages of context dependent measures. For example, suppose that two researchers are examining the same population, using the same sample. One researcher uses variables $x, y, a_1, a_2, \ldots, a_N$; the other researcher uses variables $x, y, b_1, b_2, \ldots, b_N$. The similarity coefficient between x and y calculated by the first researcher is not necessarily the same as that calculated by the second researcher, because x and y may be very similar in terms of both having a high association with a_i variables and may not be very similar at all in the association with any of the b_i variables. However, we hypothesize that if all the variables or instruments are chosen carefully and if N is large, the two different values for similarity will indeed be close.

The formal definition for similarity s_{ij} between the ith and jth variables is as follows:

$$s_{ij} = 1 - \frac{1}{2(N-2)} \sum_{\substack{n=1 \\ n \neq i,j}}^{N} |\rho_{in} - \rho_{jn}|,$$

where ρ_{in} is the association between the ith and nth variables and $-1 \leq \rho_{in} \leq +1$.

If variables x_i and x_j are highly similar s_{ij} will be just less than one: if they are almost completely dissimilar s_{ij} will be close to zero.

SIMILARITY AND ASSOCIATION RELATION

Both the association and similarity measures may be used to construct the relation R. First the association relation R_a is constructed, as usual.

$$R_a = \{(x_i, x_j) | \rho_{ij} > \theta_a\}.$$

Then the similarity relation R_s is constructed by

$$R_s = \{(x_i, x_j) | s_{ij} > \theta_s\}.$$

The total relation R, which includes only those variables which are both highly associated and highly similar, is defined by the intersection of R_a and R_s:

$$R = R_a \cap R_s = \{(x_i, x_j) | \rho_{ij} > \theta_a \text{ and } s_{ij} > \theta_s\}.$$

The pair (x_i, x_j) is a member of relation R if and only if both the association ρ_{ij} and the similarity s_{ij} are great enough.

The total relation R may have the same shortcoming which the association relation R_a has: however, it intuitively seems that two environmental subsystems or processes which are both highly similar and highly associated are one and the same. If a counter-example to this claim can be found, it would be most instructive, for it would probably lead to still further concepts of statistical relatedness.

1. Oddness, bizarre behavior
2. Restlessness, inability to sit still
3. Attention-seeking, "show-off" behavior
4. Stays out late at night
5. Doesn't know how to have fun; behaves like a little adult
6. Self-consciousness, easily embarrassed
7. Fixed expression, lack of emotional reactivity
8. Disruptiveness, tendency to annoy and bother others
9. Feelings of inferiority
10. Steals in company with others
11. Boisterousness, roudiness
12. Crying over minor annoyances and hurts
13. Preoccupation: "in a world of his own"
14. Shyness, bashfulness
15. Social withdrawal, preference for solitary activities
16. Dislike for school
17. Jealousy over attention paid other children
18. Belongs to a gang
19. Repetitive speech
20. Short attention span
21. Lack of self-confidence
22. Inattentiveness to what others say
23. Easily flustered and confused
24. Incoherent speech
25. Fighting
26. Loyal to delinquent friends
27. Temper tantrums
28. Reticence, secretiveness
29. Truancy from school
30. Hypersensitivity; feelings hurt easily
31. Laziness in school and in performance of other tasks
32. Anxiety, chronic general fearfulness
33. Irresponsibility, undependability
34. Excessive daydreaming
35. Masturbation
36. Has bad companions
37. Tension, inability to relax
38. Disobedience, difficulty in disciplinary control
39. Depression, chronic sadness
40. Uncooperativeness in group situations
41. Aloofness, social reserve
42. Passivity, suggestibility; easily led by others
43. Clumsiness, awkwardness, poor muscular coordination
44. Hyperactivity; "always on the go"
45. Distractibility
46. Destructiveness in regard to his own and/or other's property
47. Negativism, tendency to do the opposite of what is requested
48. Impertinence, sauciness
49. Sluggishness, lethargy
50. Drowsiness
51. Profane language, swearing, cursing
52. Nervousness, jitteriness, jumpiness; easily startled
53. Irritability; hot-tempered, easily aroused to anger
54. Enuresis, bed-wetting
55. Often has physical complaints; e.g., headaches, stomach ache

FIG. 2. Behavioral problem checklist developed by Dr. Ronald Peterson and Dr. Herbert Quay at the Childrens Research Center, University of Illinois, Champaign, Illinois, 1967.

RESULTS

To determine the validity of the above discussion, the following experiment was tried for binary data obtained from a behavioral problem checklist on deaf children (see Fig. 2).

Since the data was binary, the ϕ coefficient was used as a measure of association:

$$\phi_{ij} = \frac{P(x_i = 1, x_j = 1)P(x_i = 0, x_j = 0) - P(x_i = 1, x_j, 0)P(x_i = 0, x_j = 1)}{\sqrt{P(x_i = 1)P(x_i = 0)P(x_j = 1)P(x_j = 0)}}.$$

The association relation, illustrated in Fig. 3, was defined by

$$R_a = \{(x_i, x_j) | \phi_{ij} \text{ is the top decile of } \phi \text{ coefficients}\}$$

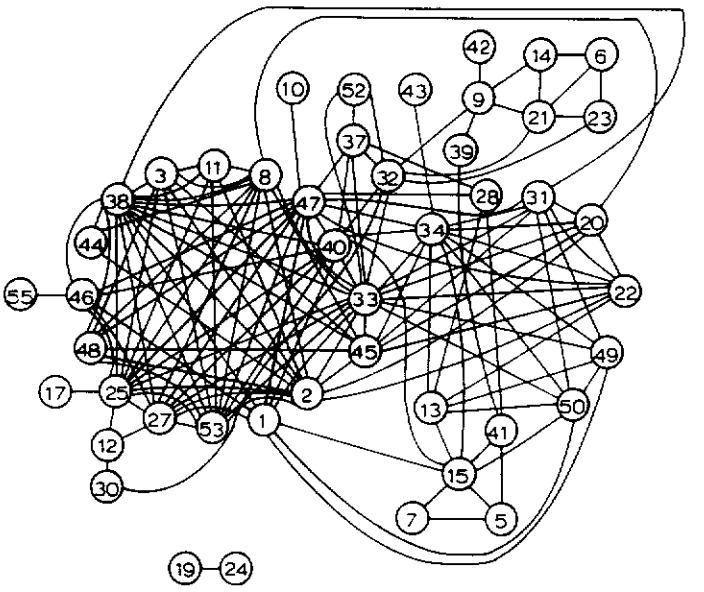

FIG. 3. Ten per cent association graph using ϕ coefficient.

and consists of all those ordered pairs (x_i, x_j) having ϕ coefficients ϕ_{ij} taking values in the top 10 per cent of ϕ coefficients in the set $\{\phi_{nm} | n = 1, 2, \ldots, N, m = 1, 2, \ldots, N\}$. The similarity relation, illustrated in Fig. 4, was defined by

$$R_s = \{(x_i, x_j) | s_{ij} \text{ is in the top decile of similarity coefficients}\}$$

and consists of all those ordered pairs (x_i, x_j) having similarity coefficients taking values in the top 10 per cent of similarity coefficients. The total relation R, illustrated in Fig. 5, was defined by

$$R = R_a \cap R_s.$$

The graphs obtained by using R_a, R_s and R differ markedly from one another. This may be seen by examining the respective graphs illustrated in Figs. 3–5 and the general picture of the clusters of variables shown in Fig. 6.

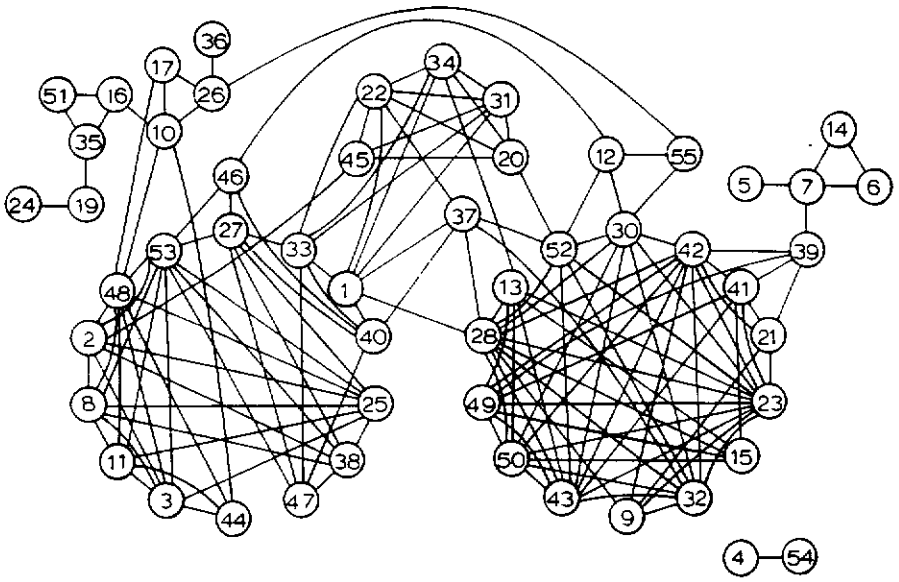

FIG. 4. Ten per cent similarity graph using ϕ coefficient.

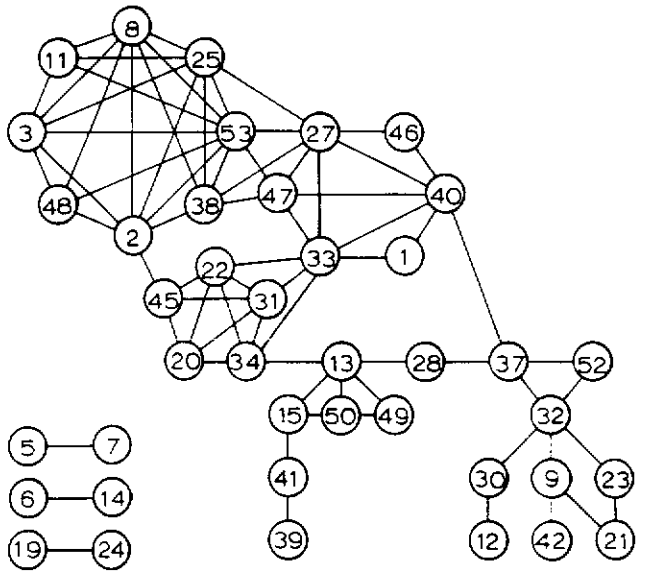

FIG. 5. Ten per cent similarity and association graph using ϕ coefficient.

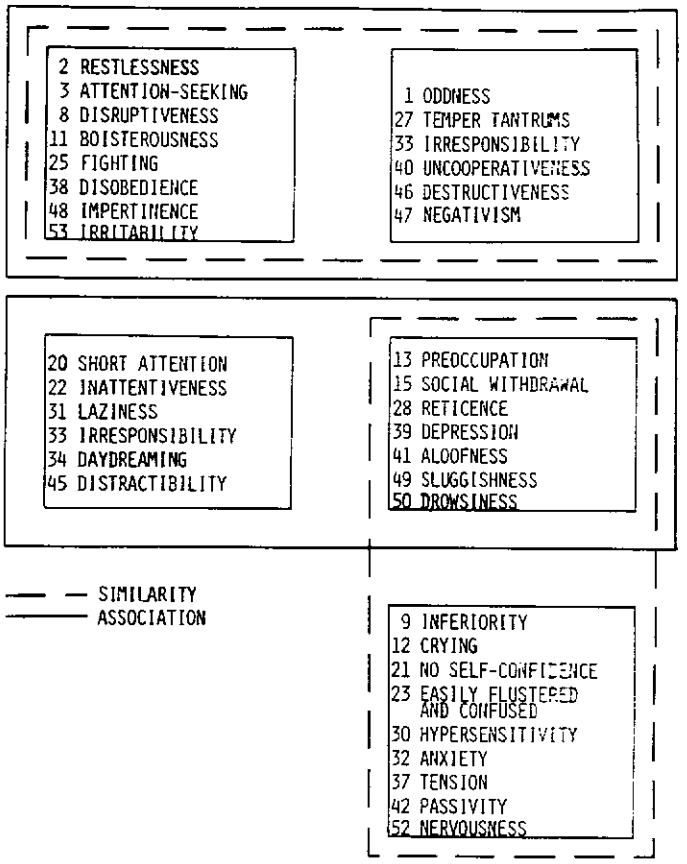

FIG. 6. Illustrates the relationships between the similarity association and combined clusters.

THE ASSOCIATION CLUSTERS

The association measure yields three major clusters of variables. These clusters may be labeled as aggressive disobedience, social withdrawal, and anxiety. The largest cluster is that characterized as aggressive disobedience. This cluster includes a very wide range of variables including: irresponsibility, negativism, disruptiveness, stealing, tantrums, irritability, impertinence and crying. The focal nodes of this cluster are disobedience with links to fifteen other variables within the cluster and fighting with links to thirteen other variables in the cluster. Restlessness and disruptiveness have a large number of intra-cluster links (twelve and eleven respectively), irritability, tantrums, and disruptiveness are linked to eleven other variables. There are twenty-two variables in this cluster.

The next largest cluster obtained from the association measure is the social withdrawal cluster which contains fifteen variables. These variables include social withdrawal, sluggishness, aloofness, reticence and drowsiness. The focal nodes are daydreaming which is connected to nine of the cluster variables and laziness which is connected to eight.

The smallest cluster obtained from the association measure is the anxiety cluster. This cluster has thirteen variables which include: shyness, hypersensitivity, anxiety, nervousness,

tension, and depression. The focal nodes are anxiety, with links to six variables, and lack of self confidence, inferiority feelings and negativism, with links to five other variables in the cluster.

THE SIMILARITY CLUSTERS

The similarity measure yields five major clusters of variables. These clusters may be labeled bad companions, inattentiveness, temper related outbursts, expressive problems and anxiety.

The largest cluster is the anxiety cluster with twenty-one variables. These variables include anxiety, nervousness, crying, aloofness, drowsiness and social withdrawal. The focal nodes of this cluster are easily flustered with links to twelve other variables, passive with links to eleven other variables, reticent with links to ten other variables, and clumsy with links to ten other variables in the cluster.

The next largest cluster is that of temper related outbursts. This cluster contains thirteen variables. These include tantrums, irritability, restlessness, impertinence, disruptiveness, fighting and disobedience. The focal nodes are irritability with links to ten other variables, fighting with links to eight other variables, impertinence with links to eight other variables, and disruptiveness, restlessness and "showing-off" with links to seven other variables.

The third cluster is that of inattentiveness. This cluster has eight variables including laziness, short attention span, daydreaming and inattentiveness. Its focal nodes are inattentiveness with links to seven other variables, and laziness and daydreaming, each with links to six other variables.

The fourth cluster is that containing expressive problems. There are five variables in this cluster including dislike for school, repetitive speech, swearing, incoherent speech and masturbation. All of the nodes are focal.

The fifth cluster obtained by the similarity measure is that of bad companions. This is also a five variable cluster containing delinquent friends, stealing and bad companions. There are three focal nodes in this cluster. These nodes are stealing, linked to three cluster variables, and loyalty to delinquent friends and jealousy, each linked to two.

THE COMBINED CLUSTERS

The combined use of the similarity and association measures yields five clusters, which may be labeled inattentiveness, aggressive-temper, depressed withdrawal, anxiety and uncooperativeness.

The largest cluster is the aggressive cluster with ten variables. These include restlessness, temper tantrums, fighting and disruptiveness. Its focal nodes are irritability with links to nine other variables; disruptiveness and fighting with links to seven other variables.

The next cluster is that of anxiety with nine variables including tension, anxiety, crying and passivity. All nodes are focal.

Inattentiveness has six variables including inattentiveness, daydreaming, laziness and irresponsibility. Its focal nodes are inattentiveness and laziness which each connect to five other variables in the cluster.

The depressed withdrawal cluster also has six variables including drowsiness, withdrawal, depression and preoccupation. Its focal nodes are preoccupation, social withdrawal and drowsiness, all linking to three other cluster variables.

The uncooperative cluster has six variables including irresponsibility, uncooperativeness and tantrums. The focal nodes are uncooperativeness and irresponsibility with links to five and four other cluster variables, respectively.

DISCUSSION OF THE ASSOCIATION, SIMILARITY AND COMBINED RELATIONS

Figure 6 illustrates the inter-relationships between the association, similarity, and combined relations. These inter-relationships show that information can be lost when only one measure of statistical relationship is used. The clusters of inattentiveness and social withdrawal, which are distinct in the combined relation, are inseparable in the association relation. The clusters of social withdrawal and anxiety, which are distinct in the combined relation, are inseparable in the similarity relation. Finally, the clusters of show-off disruptiveness and aggressive tempers, which are just barely distinct in the combined relation, are confused in both the association relation and similarity relation.

The clusters obtained by the combined similarity and association relation seem to edit out those variables which tend not to be dominant ones and provide a neater theoretical picture of the variable interactions. This editing job is a real aid in the first round of data analysis when it is important to focus on the main effects. After main effects have been taken into account, the less dominant variables need to enter the picture and their place may be discovered upon examination of the association or similarity relation. Hence, all three relations are useful.

CONCLUSION

A context dependent measure of similarity distinct from a measure of association has been defined. Two variables are highly similar if and only if their respective associations with the remaining variables are nearly identical. On the basis of the variable associations, an association relation R_a was defined. On the basis of the variable similarities, a similarity relation R_s was defined. The total relation R was defined as the intersection of R_a and R_s. The clusters for R_a, R_s and R were each determined using data concerned with the behavioral problems of deaf children. Both R_a and R_s combined a pair of clusters which were distinct in R thereby implying that better clustering results may be achieved with the integrated use of the similarity and association measures than with either one of them alone.

REFERENCES

1. L. A. Goodman and W. H. Kruskal. Measures of association in cross classification. *J. Am. Stat. Assoc.* **49**, 732–764 (1954).
2. L. A. Goodman and W. H. Kruskal. Measures of association for cross classification II: Further discussion and references. *J. Am. Stat. Assoc.* **54**, 123–163 (1959).
3. G. Salton, E. Keen and M. Lesk. Design experiments in automatic information retrieval. *The Growth of Knowledge*, edited by Manfred Kochen. Wiley, New York (1967).
4. K. Jones and D. Jackson. Current approaches to classification and clump-finding at the Cambridge Language Research Unit. *Comput. J.* **10**, 29–37 (May 1967).
5. M. A. Libbey. The use of second order descriptors for document retrieval. *Am. Docum.* **18** (January 1967).
6. P. A. Lewis, P. B. Baxendale and J. L. Bennett. Statistical discrimination of the synonomy (antonomy relationship between words). *J. ACM*, **14** (January 1967).
7. A. Rosenfeld, H. Huang and V. Schneider. An application of cluster detection to text and picture processing. *IEEE Trans. Information Theory* **IT-15**, No. 6 (November 1969).
8. H. E. Stiles. The association factor in information retrieval. *J. ACM* **8** (April 1961).

9. G. BALL, Data analysis in the social sciences: What about the details?, *Proc. Fall Joint Computer Conf.* Las Vegas, pp. 533–559 (December 1965).
10. R. SOKAL and P. SNEATH, *Principles of Numerical Taxonomy*. Freeman, San Francisco (1963).
11. R. E. BONNER, On some clustering techniques, *IBM J.* **8**, 22–32 (January 1964).
12. D. GASKING, Clusters, *Australas. J. Phil.* **38**, 1–35 (May 1960).
13. R. REIVICH and I. ROTHROCK, Behavioral problems of deaf children and adolescents: A factor-analytic study, to be published.
14. C. TRYON, *Cluster Analysis*. Edwards, Ann Arbor, Michigan (1939).