Article *NIMHANS Journal*

Cluster Formation in Child Psychiatry - Part I: Some Methodological Evaluation

Volume: 15 Issue: 01 January 1997 Page: 7-17

M Venkataswamy Reddy .

Reprints request

, V G Kaliaperumal,

- *Department of Biostatistics, IMHANS*

S M Channabasavanna, - *Director, National Institute of Mental Health & Neuro Sciences, Bangalore 560 029, India*

Abstract

Using real data on 435 child psychiatric patients each measured on 78 binary variables, seven hierarchical agglomerative and two k-means algorithms were compared on their agreement, number of discriminating variables and level of recovery of marker sample. The capacity to produce clear cut clusters and confirmed number of clusters were studied in hierarchical methods with four measures of proximity. While single linkage and centroid methods have not yielded clear cut clusters with any of the four measures of proximity, complete linkage and average linkage (within) yielded clear cut clusters. Ward method failed to produce classifications with confirmed number of clusters. Hierarchical methods with distance measures were more accurate than with similarity measures. The k-means algorithms produced the excellent recovery of clusters structures when used with the centroids of clusters generated by the best hierarchical methods.

Key words -

Child psychiatry,
Cluster analysis,
Hierarchical methods,
K-means algorithms

Clustering techniques can be profitably used to classify psychiatric patients as an aid to the development of more meaningful diagnostic system [1]. The major methodological problem in cluster analysis is the choice of a clustering procedure; the user must identify a suitable method to classify subjects. Those who employed cluster analysis in child psychiatry resorted to the available packages without any rationale in selecting one method or the other [2], [3], [4], [5] and [6]. As the methods employed differed from paper to paper, the results were not comparable. Furthermore, different methods were seldom employed on the same data. Thus, there was no scope for empirically evaluating different cluster procedures. The major objective of the present study is to select a few methods of cluster analysis as well as a few measures of similarity and compare the results. In this way, a suitable cluster analytical procedure with appropriate measure of association/distance could be suggested to tackle classification problems arising in the field of child psychiatry.

Material and Methods

Data set

Detailed data was collected on 701 consecutive cases registered in child guidance clinic, National Institute of Mental Health & Neuro Sciences, Bangalore . Information was available on demographic factors, chief presenting complaints, commonly occurring symptoms, family atmosphere, milestones of developments, school and study pattern, premorbid temperaments, mental status examination and psychiatric diagnosis. Patients aged 5 to 16 years studying between first and tenth standards and whose both the parents were alive were included in the present study. After fulfilling the inclusion criteria, only 435 cases qualified for inclusion in the present analysis. Out of these, 281(65%) were males, and children below 9 years of age comprised of 26%, while 27% were above 12 years.

Data transformation

In order to avoid severely skewed distribution, qualitative variables were included only if their frequency was between 10% and 90%. Altogether, 66 variables were included for the purpose of the analysis, out of which 52 were binary, six were polychotomous and eight were continuos in nature. The categorical variables as well as the quantitative variables were transformed into binary types. Thus the final list of variables considered for the purpose of cluster analysis consisted of 78 binary variables whose frequency ranged from 10% to 85%.

All these variables were coded as one if the given phenomenon was present and zero otherwise.

Out of 435 patients, 58 were not diagnosed. For 121 patients, the frequency of each category of diagnosis was less than 20, and hence they were excluded from marker sample. Thus the marker sample studied in this report consisted of 314 patients with 22 cases of psychosis, 133 cases of hysteria, 62 cases of conduct disorders, 39 cases of hyperactivity syndrome, and 58 cases without any diagnosis. This marker sample was employed to facilitate one of the evaluations of the classifications to be arrived at by employing different methods of cluster analysis.

The present study utilized a multilevel analysis in which alternate classifications schemes were compared with a view to select a 'best' method. In the first phase, seven hierarchical agglomerative methods, viz,. Single Linkage method (SLINK), Complete Linkage method (CLINK), Average Linkage between the merged groups (ALINKB), Average Linkage within the new group (ALINKW), centroid method (CENTROID), Median method (MEDIAN) and the Ward method (WARD) [7] were employed to cluster 435 cases each measured on 78 binary variables. Each method was employed with four measures of proximity (distance or similarity), viz., Euclidian Distance (ED), Standardised Euclidian Distance (SED), Pearson Correlation Coefficient (r) and Jaccords Association Coefficient (JC). After obtaining 28 classifications, each expressed in the form of Dendrograms, the following sequence of steps of analysis were carried out:

Step 1: Confirm the configuration of Dendrograms to determine whether clear cut clusters were present or not by submitting them to three consultant Bio-statisticians.

Step 2: Determine the number of clusters in each of the classifications by using Mojena's Rule [8].

Step 3: Determine the structure of the data by employing C-Index due to Calinski-Harabasz [9].

Step 4: Determine the level of agreement among the cluster schemes using Rand Index [10].

Step 5: Determine the number of discriminating variables in each of the classifications by employing

Chi-square test of goodness of fit.

Step 6: Determine the level of recovery of Maker sample by matching the empirical groups with the most appropriate clinical diagnostic groups.

In the second phase, the K-means algorithms of Forgy and its variant proposed by Jancey [7] were employed to the same set of data with k-1, k and k +1 number of clusters (k is the number of clusters determined by hierarchical methods). Four procedures for initiating clusters were selected, viz.,
(1) seed points at equal intervals,
(2) seed points at random,
(3) initial clusters at equal length intervals and
(4) initial clusters at random length intervals.

In each of the classifications, the Milligan's point biserial correlations and the number of reallocations were recorded. By this procedure, the number of clusters by k-means, algorithm were determined. In the next step, the k-means algorithms were employed using centroids of classifications obtained by hierarchical methods which yielded confirmed number of clusters. As in the case of hierarchical methods, the k-means algorithms were evaluated with respect to their agreement, number of discriminating variables (internal validation) and level of recovery of marker sample (external validation).

In the final phase, the best classification was selected as one with highest recovery of Marker sample from classifications of hierarchical methods and k-means algorithms which yielded confirmed number of clusters. The sub-routines due to Anderberg [7] were used to carry our the seven hierarchical and two k-means algorithms in the present study. The computer program RAND due to Dreger [12] was used to carry our Rand Index of cluster solution agreement.

Results

Hierarchical methods

Analyzing 435 cases each measured on 78 binary variables, 28 dendrograms were obtained. All 28 dendrograms were submitted to three consultant Biostatisticians to confirm their configuration to ascertain if clear cut clusters were present or not.

The final configuration was based on the majority opinion and thus if at least two consultants agreed to the presence of clusters in a dendrogram, that dendrogram was considered to have the clusters. This process resulted in selecting 14 dendrograms with confirmed clear cut clusters as shown in Table I. They were CLINK and ALINKW with all the four measures of distance/similarity, ALINKB with SED and r, MEDIAN and WARD with both ED and SED.

Table I - Confirmation of dendrograms and determination of number of clusters

Table I - Confirmation of dendrograms and determination of number of clusters

The results of application of Mojena's rule [8] were also presented in Table I. It can be noted that there was no trend or significant value of z-score of Mojena's rule in 14 dendrograms confirming that no clear cut cluster structure was present in these dendrograms. Interestingly, those 14 cluster structures which were selected by consultants and those 14 clusters structures selected by Mojena's rule were the same. In the present study, a z-score value of 2.58 to test for the significance of fusion coefficients was found to be suitable to determine the number of clusters. The Mojena's rule indicated the presence of 3 to 8 clusters as shown in Table I.

The value of C-Index of Calinski Harabasz [9], at some preceding and at some succeeding number of clusters suggested by Mojena's rule, are presented in Table II.

Table II - C-Index of Calinski & Harabasz at three different number of clusters

Table II - C-Index of Calinski & Harabasz at three different number of clusters

It can be noted that except ALINKW-ED, the remaining results indicated negative relationship, suggesting that the number of clusters increased with the decrease in C-Index. Thus the application of C-Index method clearly validated the presence of hierarchical structure in the data.

The mean agreement as measured by Mean Rand Index [10], the number of variables which acquired significant discriminating power between the obtained clusters and the percentage recovery of marker sample were shown in Table III.

Table III - Mean Rand Index, number if discriminating variables and percentage of recovery in hierarchical methods

Table III - Mean Rand Index, number if discriminating variables and percentage of recovery in hierarchical methods

**p < 0.01

ALINKW-ED with 6 clusters had the highest agreement (0.917) and MEDIAN-ED with 3 clusters had the lowest agreement (0.689) with other cluster schemes. CLINK-r with 6 clusters and CLINK-JC with 7 clusters had highest number (71) of discriminating variables and MEDIAN-ED with 3 clusters had the lowest number (64) of discriminating variables. Similarly, CLINK-ED with 6 clusters had the highest recovery rate (73.2%) and ALINKW-r with 7 clusters had the lowest recovery rate (50.6).

k-means algorithms

Since Six-cluster seemed to be a good structure as judged by the results of hierarchical methods, two k-means algorithms, viz: Forgy method and Jancey method have been employed with 5, 6 and 7 number of clusters. Seed points were selected at equal intervals records and random records. The initial partitioning was carried out by partitioning at equal distance and by partitioning at random. Thus 24 clustering schemes were obtained in this procedure. The point-biserial correlation [13] and the number of reallocations required in these 24 k-means procedures are given in Table IV.

Table IV - Point-biserial correlation coefficients and number of reallocations (in brackets) in classifications obtained by k-means algorithms

Table IV - Point-biserial correlation coefficients and number of reallocations (in brackets) in classifications obtained by k-means algorithms

The average correlation coefficient was highest for six cluster solutions (0.461) than for five (0.445) and seven (0.395) cluster solutions. The average correlation coefficient was also highest for seed points with equal intervals (0.452) than for the random seed points (0.431). Random initial clusters had highest average correlation coefficient (0.440) than the equal interval initiation of clusters (0.412).

The average correlation was more for Jancey's method (0.443) than for the Forgy method (0.425). Again, out of 8 procedures for initiating clusters, 5 procedures had minimum number of reallocations for 6-cluster structures, 2 for 5 cluster structures and one for 7 structure clusters. Thus, six cluster structure is confirmed for the data under analysis. Hence, the k-means algorithm was repeated with centroids of classifications of hierarchical procedures which yielded 6-cluster solutions. Thus, there were 20 k-means procedures as listed in Table V. As in the case of hierarchical methods, the mean Rand Index, number of discriminating variables and percentages of recovery of the 20 classifications were obtained and projected in Table V.

Table V - Mean Rand Index, number of discriminating variables and percentage of recovery in k-means algorithms

Table V - Mean Rand Index, number of discriminating variables and percentage of recovery in k-means algorithms

**P < 0.01

All 20 structures had mean Rand Index more than 0.800. The Mean Rand Index of structures due to Forgy's methods ranged from 0.876 to 0.922 while 10 structures due to Jancey's method ranged from 0.850 to 0.906. Structures due to Forgy's method with initiating clusters at random had highest mean Rand Index of 0.922 and the minimum of 0.850 was achieved by the structure generated by Jancey's method with the centroids of MEDIAN-SED. It can be noted that the number of significant variables ranged from 66 to 71 for the Forgy's method and 65 to 68 for the Jancey's method. Except the Forgey method with seed points at equal intervals, all others structures generated by the partitioning method had recovered the marker sample significantly. In Forgy's method, the recovery rate ranged from 51.3% to 76.4% and in the case of Jancey method it ranged from 64.9% to 74.5%. The recovery rate was highest for Forgy method with cluster centroids of CLINK-ED. Further this method had 70 discriminating variables and mean Rand Index of 0.914. Hence, it is selected as the best method in the present study for the data that was analyzed.

Discussion and conclusions

Everitt [14] stated that the SLINK yielded less accurate solution when the number of entities was large. The sample size of 435 with 78 variables was certainly large in the present analysis. Probably because of these factors, SLINK did not produce clear cut clusters in the present study with any of the four measures of proximity. This found an agreement with many studies including the studies by Kuiper and Fisher [15], Mojena [8], Blashfield [16], Mezzich [17], Edelbrock [18], Milligan and Bayne et al [19]. It was stated that SLINK could be used to identify the outliers [16]. In the present study, the data has been 'cleaned'; symptoms and signs which occurred less than 10% or more than 90% were excluded from the analysis. Thus, SLINK did not behave fruitfully under 'nice' condition. On the other hand, the situation was favourable to CLINK which yielded clear cut clusters for all four measures of proximity. It may be noted that CLINK is logically opposite of SLINK in methodology.

Jain et al [20] compared six hierarchical methods on univariate random data from uniform and standard normal distributions and found that the clink method generally was best in not discovering false clusters. In a Monte Carlo study, Hubert [21] found that the CLINK usually had greater power. The results of the present study and also of many other studies confirmed that CLINK was better than SLINK.

Logically, ALINK methods are in between SLINK and CLINK. Several Monte Carlo studies [22] established that ALINK methods were suitable when the cluster size varied. The size of final clusters ranged from 50 to 111 in the present study. As noted earlier, all four dendrograms obtained by using ALINKW yielded clear cut clusters. Out of four dendrograms obtained by using ALINKB two dendrograms yielded clear cut clusters; the SED and r were the measure employed to get these dendrograms.

The CENTROID and MEDIAN were originally defined to classify variables rather than subjects [22], [23], [24]. The reversal phenomena in the case of CENTROID is well known [25]. Accordingly, in the present study the CENTROID could not yield any clear cut clusters at all. On the other hand, the MEDIAN yielded clear cut clusters for distance measures only.

The WARD was defined only for distance measures as the variance is the operating factor [26]. Thus, this method failed to produce meaningful clusters with r and JC. The classifications, suggested by WARD was characterized by a good agreement, recovery of marker sample, and number of significant variables between clusters. However, several Monte Carlo studies [27] concluded that the recovery was found to increase as the number of groups decreased. The clusters distinguished by WARD seemed to be largely separated along a dimension related to profile elevation [28]. In the present analysis, this elevation component was larger enough to conclude that the solution of this method was effective. When the WARD was introduced in 1963, research workers believed that it was the best alternative clustering method and this notion was carried out probably until 1978 [29]. However, it was later discovered that the efficiency of this method depended on the nature of data. It seemed that the WARD was the method of choice for equal size clusters. Though the cluster size ranged from 50 to 111 in the present study, the WARD yielded clear cut and meaningful clusters.

Out of 14 clear cut cluster structures obtained, there ware only six dendrograms with 6 cluster solution. They were CLINK with ED and r, ALINKB with SED, ALINKW with ED and SED, and MEDIAN with SED. Thus, the present study demonstrated that Euclidian Distance and more particularly the Standardized Euclidian Distance would be a better measure to be employed in cluster analysis when the data is binary. Further, the statement of Anderberg [7] that ALINKW frequently gives results that are little different from those obtained with CLINK was confirmed. Several Monte Carlo studies have shown that as the number of observed clusters deviated from the number of true clusters, then the recovery rate would decrease [25]. Comparing these six structures, CLINK-ED had the highest recovery (73.2%) while CLINK-r, had the smallest (60.1%). Similarly, ALINKW-ED had the largest Mean Rand Index (0.917) while MEDIAN-SED had the smallest (0.785). CLINK-r had the highest number of significant variables (71) while ALINKW-ED had the smallest number (67). Since the difference in the number of significant variables was not significant, it could be concluded that the approach did not distinguish the clustering methods which yielded reasonable solutions.

The overall performance of k-means algorithms was interesting in this report. The Forgy method was found to be better than the Jancey's variant in the present study, as judged by the range of mean Rand Index as well as the number of variables discriminating the groups. The clusters produced by k-means algorithms with seed points at equal intervals failed to effect significant recovery rate which found an agreement with other research workers [11]. The k-means algorithms with random seed points generated good clusters schemes which had satisfactory recovery rate in the present study. This was contrary to findings reported by majority of the Monte Carlo studies reported in the literature [13]. Only the studies conducted by Spath [30] found an agreement with the present study. The k-means algorithms with seed points at equal intervals and with initiating clusters at random have yielded good clusters which had good recovery of marker sample as well as high mean Rand Index. This was in agreement with Spath [30] and Milligan [31], Everitt [32] and Scheibler and Schneider [33] concluded that k-means algorithms produced recovery values worse than those of the best hierarchical methods when random starting seeds was used. On the other hand, when the centroids of the clusters generated by the best hierarchical methods were used as the starting seeds, k-means algorithms produced the excellent recovery of cluster structure. The present study confirmed this finding and emphasized that the resultant clusters were better if the initial clusters generated by the hierarchical methods were employed in k-means algorithms.

1.Bartko J J, Strauss J S, Carpenter W T, An evaluation of taxometric techniques for psychiatric data
Classifications Society Bulletin Page: 2: 2-28, 1971
2.Jenkins R L, Psychiatric syndromes in children and their relation to family background
American Journal of Orthopsychiatry Page: 36: 450-7, 1966
3.Wolff S, Dimensions and clusters of symptoms in disturbed children
British Journal of Psychiatry Page: 118: 421-7, 1971
4.Wolkind S N, Everitt B S, A cluster analysis of the behavioural items in the preschool child
Psychological Medicine Page: 4: 422-7, 1974
5.Prior M, Boulton D, Gajzago C, Perry D, The classification of childhood psychoses by numberical taxonomy
Journal of Child Psychology & Psychiatry Page: 16: 321-30, 1975
6.Gdowski C L, Lachar D, Kline R B, A PIC profile typology of children & adolescents:1. Empirically derived alternative to traditional diagnosis
Journal of Abnormal Psychology Page: 94: 346-61, 1985
7.Anderberg M R, Cluster analysis for applications. New York: Academic Press
1973
8.Mojena R, Hierarchical grouping methods and stopping rules: an evaluation
Computer Journal Page: 20: 359-63, 1977
9.Calinski T, Harabasz J, A dendrite method for cluster analysis
Communications in Statistics Page: 3: 1-27, 1974
10.Rand W M, Objective criteria for the evaluation of clustering methods
Journal of the American Statistical Association Page: 66: 846-50, 1971
11.Milligan G W, An examination of the effect of six types of error perturbation on fifteeen clustering algorithms
Psychometrika Page: 45: 325-42, 1980
12.Dreger R M, Microcomputer programs for the Rand Index of cluster similarity
Educational and Psychological Measurement Page: 46: 655-61, 1986
13.Milligan G W & Mahajan V, A note on procedures for testing the quality of a clustering of a set of objects
Decision Sciences Page: 11: 655-61, 1980
14.Everitt B S, Cluster analysis. London. Heinman Educational Books
1974
15.Kuiper F K, Fisher L, A Monte Carlo comparison of six clustering procedures
Biometrics Page: 31: 777-83, 1975

16.Blashfield R K,   Mixture model tests of clsuter analysis. Accuracy of four Agglomerative hierarchical methods
Psychological Bulletin   Page: 83: 377-88, 1976
17.Mezzich J E,   Evaluating clustering methods for psychiatric diagnosis
Biological Psychiatry   Page: 13: 265-81, 1978
18.Edelbrock C S,   Mixture model test of hierarchical clustering algorithms. The problem of classifying everybody
Multivariate Behavioral Research   Page: 14: 367-84, 1979
19.Bayne C K, Beauchamp J J, Begovich C L, Kane V E,   Monte Carlo Comparisons of selected clustering procedures
Pattern Recognition   Page: 12: 51-62, 1980
20.Jain N C, Indrayan A, Goel L R,   Monte Carlo Comparison of six hierarchical clustering methods on random data
Pattern Recognition   Page: 19: 95-99, 1986
21.Hubert L J,   Approximate evaluation techniques for the single-link and complete-link hierarchical clustering procedures
Journal of the American Statistical Association   Page: 69: 698-704, 1974
22.Milligan G W, Sood S C, Sokal L M,   The effect of cluster size, dimensionality and the number of clusters on recovery of true clusters structure
IEEE transaction on Pattern Analysis and machine Intelligence   Page: 5: 40-47, 1983
23.Sokal R R, Michener C D,   A statistical method for evaluating systematic relationships
University of Kansas Science Bulletin   Page: 38: 1409-38, 1958
24.King B F,   Step-wise clustering procedures
Journal of the American Statistical Association   Page: 62: 86-101, 1966
25.Romesburg H C,   Cluster analysis for researchers. Belmont .Lifetime Learning Publications
1984
26.Ward J H,   Hierarchical grouping to optimize an objective function
Journal of the American Statistical Association   Page: 58: 236-44, 1963
27.Hands S, Everitt B S,   A Monte Carlo study of the recovery of cluster structure in binary data by hierarchical clustering techniques
Multivariate Behavioral Research   Page: 22: 235-43, 1987
28.Aldenderfer M S, Blashfield R K,   Cluster analysis. Beverely Hills, California Sage Publications
1984
29.Milligan G W,   A review of Monte Carlo test of cluster analysis
Multivariate Behavioral Research   Page: 16: 379-407, 1981
30.Spath H,   Cluster analysis algorithms. Chichester Ellis Horwood
1980
31.Milligan G W,   A two stage clustering algorithm with robust recovery characteristics
Educational & Psychological Measurement   Page: 40: 755-9, 1980
32.Everitt B S,   Unresolved problems in cluster analysis
Biometrics   Page: 35: 169-81, 1979
33.Scheibler D, Schneider W,   Monte Carlo tests of the accuracy of cluster analysis algorithms: a comparison of hierarchical and non-hierarchical methods
Multivariate Behavioral Research   Page: 20: 283-304, 1985