ISSN:
Print - 2277 - 0593
Online - 2315 - 7461
© FUNAAB 2012

Journal of Natural Science, Engineering and Technology

HIERARCHICAL CLASSIFICATION OF MORPHOLOGICAL FEATURES OF TILAPIA CABREA

***1S. O. N. AGWUEGBO, 2A. P. ADEWOLE AND 3M. G. ISENAH**

1 Department of Statistics, Federal University of Agriculture, Abeokuta, Nigeria.
2 Department of Computer Science, University of Lagos, Nigeria.
3 Department of Statistics, University of Ibadan, Nigeria.
***Corresponding author**: agwuegbo_son@yahoo.com **Tel:** +2348038004887

ABSTRACT

This article proposes an effective data visualization of multidimensional data. These displays are useful to represent the existence or absence of relationships among objects corresponding to hierarchical classifications, bifurcation or evolutionary structure. The display in this article used some morphological features of Tilapia Cabrea, as represented in the *dendrogram* or *cluster tree* which illustrates the successive fusions of objects into groups or divisions made at each successive stage of the analysis. Effectively, this clustering reduces the dimensionality and makes interpretations easier.

Keyword: Multi-dimensional data, Fusion level / height, Hierarchical classification, Dendrogram.

INTRODUCTION

Over the years, a wealth of algorithms and computer programs has been developed in an attempt to display multi-dimensional data effectively. Many graphical methods for displaying multivariate data consist of arrangements of multiple displays of one or two variables using scatter-plot matrices and parallel coordinate plots (Hurley, 2004). Wayner (1983) gives a survey of methods for displaying multivariate data in which each object is represented by an icon such as polygon, made up of parts that vary in size or shape with the measured attributes. Some other commonly used techniques are loop plots (Bertin, 1967), bi-plots (Gabriel, 1971, Chernoff, 1971); faces-(Chernoff, 1973) and boxes-(Hartigan, 1975).

In principle, these methods generalize to arbitrary numbers of variables, but in practice, as the dimensions increase, they become less effective and difficult to interpret for even moderate number of variables. A solution to this apparent shortcoming was given by *Kleiner and Hartigan (1981)*, who applied a hierarchical clustering algorithm to p variables and then represent each object by a tree or a scale. For hierarchical cluster analysis, it has always been natural to treat distance (or similarity measure) and details of the cluster algorithm like the *linkage* methods at par, and most software implementations reflect this fact, offering the user a wide range of different distance measures (Leisch, 2005). In programmatic environments like S, hierarchical clustering often works off a distance matrix, allowing for arbitrary distance measures to be used (Becker *et al*, 1998).

Kaufman *et* al (2005) proposed techniques for solving large problems in the area of hierarchical clustering analysis. These techniques roughly can be divided into two groups; with the first group concerned with the adaptation of existing algorithms in order to reduce either the storage or the number of calculations. Most work on this first group has been done on variants of the single linkage algorithm that involve fewer calculations but yield the same results. The second group techniques consists of new methods that have been specifically designed for clustering large data sets, although they are often based on concepts used in classical methods. The most significant aspect of the procedure is that objects are inserted into hierarchical tree and used effectively for classifying new objects. A significant part of the hierarchical classification is based on approximation by a dendrogram. Kaufman *et al* (2005) introduced two measures of clustering strength, namely, the agglomerative and divisive coefficients, which are appropriate when the groups are unknown.

When data are collected from many units that are somehow similar, the statistical problem is on how to combine the information from the various units, in order to understand better the phenomenon under study. Usually, there is substantial variability among units and a natural way to approach the problem is based on the notion that there exists a configuration of points in a higher dimensional space, and this in turn requires that the multivariate data to be quantitative so that values can be used as coordinates of points.

This study is concerned with data visualization using discrete mathematics and combinatorics to represent, interpret and reveal structures of relationships existing within groups of variables. To carry out a cluster analysis of a set of *n-dimensional* multivariate data, it becomes necessary and useful to impose some structures into metric space by classifying objects as a *Euclidean* high dimensional state space.

We expand this reasoning through the general concept of a data matrix, which consist of *n* rows of objects and *p* columns of variables or features. The data matrix ***X*** may be further specified as a *norm space*, in order to study the geometry of the *vector space*, and to discuss those aspects which depend only upon the notion of distance between two points.

METHODOLOGY AND DATA

Many of the concepts in the study can be easily generalized to an *n-dimensional* linear space $\mathcal{E}_p^n$. In a linear space, besides the operation of addition and multiplication by scalars, it is convenient to introduce a norm (i. e., the length) into the linear space. The norm of the vectors $X = (X_1, X_2, \dots, X_n)$ is defined as:

$$||X||_p = \left[\sum_{k=1}^{n} |X_k|^p \right]^{1/p} \quad for\ 1 < p < \infty \qquad (1)$$

We regard the set of all distribution functions as a topological space in order to introduce a metric, that is, a distance between pairs of elements or points of the space. The metric is defined as:

$$d(i,j) = \left[\sum_{k=1}^{n} |X_{ik} - X_{jk}|^{p} \right]^{1/p} \qquad (2)$$

This is referred to as *Minkowski's* distance or ℓ_p^n metric. The quantity $d(i,j)$ is the distance between two points X_i and X_j and this is equal to the dissimilarity between the corresponding points. The set ℓ_p^n together with the metric defined in (2) above is a metric space.

The distance between two objects i and j is a function of their observed values and can be defined as Euclidean distance:

$$d(i,j) = \left[\sum_{k=1}^{n} |X_{ik} - X_{jk}|^{2} \right]^{1/2} \qquad (3)$$

Trosset (2005) asserts that if we replace Euclidean distance in (3) with some measure of dissimilarity, then there are two natural ways to proceed. The first is to restrict attention to methods (*e.g., complete linkage cluster analysis and (or) nearest neighbor classification*) that operate directly on dissimilarities. The second is on the use of multidimensional scaling to embed the objects to be clustered or classified in the Euclidean space. This paper is restricted to the use of multidimensional scaling techniques.

Multidimensional scaling (*MDS*) is a collection of techniques for constructing configurations of points (*typically in a low dimensional Euclidean space*) from dissimilarity data. The basic idea is to find a configuration for which the inter-point distance approximates the specified dissimilarities. The variation in the dissimilarity measure provides opportunity to observe how the objects may vary with the intended use of the data. *Kaufman and Rousseauw (2005)* pointed out that dissimilarities among objects can be computed using the group average clustering algorithm AGNES (*abbreviation for Agglomerative Nesting*) as available in S-Plus.

Our intension is not to establish a tree structure from a combinatorial view point, but to assign optimal lengths to the edges of a given tree. The technique used to group the data is a hierarchical clustering algorithm. The actual application of the procedure for the Euclidean distance is by a cluster tree or dendrogram which shows how objects are successively amalgamated into groups or clusters at various values of dissimilarity measure. The cluster tree shows clearly the order and value of the measure at which the clusters are formed. At any value on the measure axis, each horizontal line in the tree represents a cluster of one or more objects from which representative objects must be chosen. The number of objects and the object membership for any cluster can be determined from the cluster tree by following the branch of the tree to the extreme left, at which point the objects in the cluster can be identified. *Mojena (1977)* suggests considering the relative sizes of the different fusion levels in the cluster tree to determine the number of clusters to be selected. His proposal is to select the number of clusters corresponding to the first stage in the dendrogram for which

$$\alpha_{j+1} = \bar{\alpha} + kS_{\alpha} \qquad (4)$$

Where $\alpha_0, \alpha_1, \ldots, \alpha_{n-1}$ are the fusion levels corresponding to the stages with $n, n-1, \ldots, 1$ clusters. The terms $\bar{\alpha}$ and S_{α} are respectively, the mean and unbiased standard deviation of the α -values and k is a constant. He recommends that $k \in (2.75, 3.50)$. Milligan and Cooper (1985), however, suggest that a more satisfactory value of k is 1.25.

The data set used for this study was abstracted from a similar study conducted by Uchendu and Nwishi (2007). The data consist of ten morphological features of *Tilapia Cabrea* and since our aim is to determine the similarities between the morphological features, we evaluated the transpose of the standardized data matrix of the morphological features.

RESULTS AND DISCUSSION

Our study examined the differences among the fusion levels in the cluster tree resulting from the application of *group average hierarchical clustering* of the morphological features. The dendrogram is presented in Fig. 1 in the appendix. Classifications corresponding to a particular point in this process can be obtained by first graphing a matrix of scatter plots of the variables as shown in Fig. 2. Applying the *Mojena (1977) scheme*, the estimated mean and unbiased standard deviations of the fusion levels are 6.3545 and 1.7284 respectively. Using a *k-value* of 1.23 gives the two-cluster classification of the morphological features of *Tilapia cabrea* shown in Table 1. The fusion levels and *k-values* for a two-cluster average linkage classification are also shown in Table 2 and the by-plot of clusters of the first-two principal components of the standardized morphological features are shown in Figure 3.

As shown in Table 1 and Fig. 2, it implies that the morphological features of Group 1 are similar and are the major features that determine the size of *Tilapia cabrea* fish. The correlations between these features are positive and are very significant as verified by Uchendu and Nwishi (2010). On the other hand, the features that constitute Group 2 are also similar but are sparingly correlated with those of Group 1. Thus, this group's contribution to the size of *Tilapia cabrea* is negligible.

APPENDIX

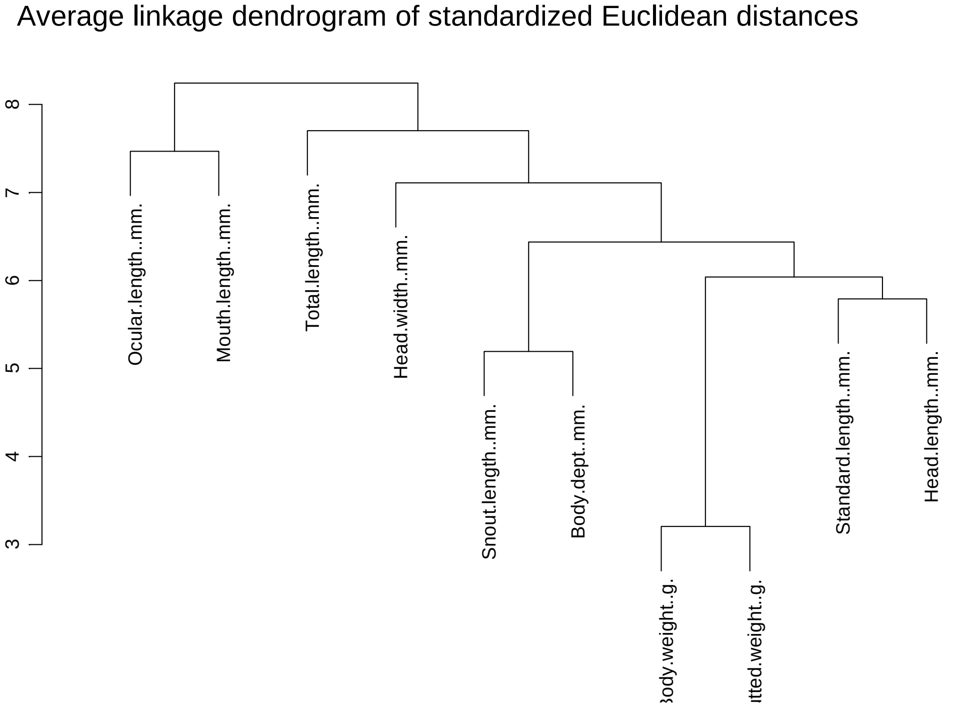

Fig. 1: Average linkage dendrogram of standardized morphological features of *Tilapia cabrea*

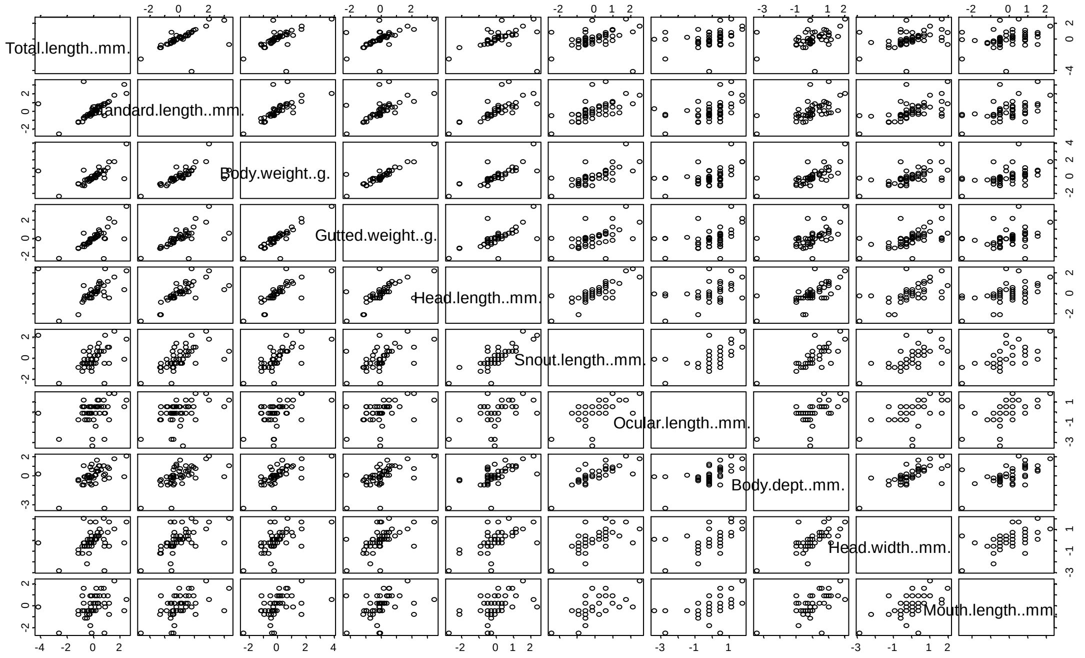

Fig. 2: Pair-wise scatter plot of standardized morphological features of *Tilapia cabrea*

Table 1: Two-cluster hierarchical classification of standardized morphological features of *Tilapia cabrea*

Group 1	Total length, Standard length, Body weight, Gutted weight, Head length, Snout length, Body weight, Head width.
Group 2	Ocular length, Mouth length.

Table 2: Fusion levels and k-values of average linkage clustering of morphological features of *Tilapia cabrea*

Fusion level	3.2045	5.1947	5.7905	6.0404	6.4374	7.1105	7.4689	7.7012	8.2427
k-value	-2.0503	-0.7549	-0.3671	-0.2044	0.0539	0.4921	0.7253	0.8765	1.2290

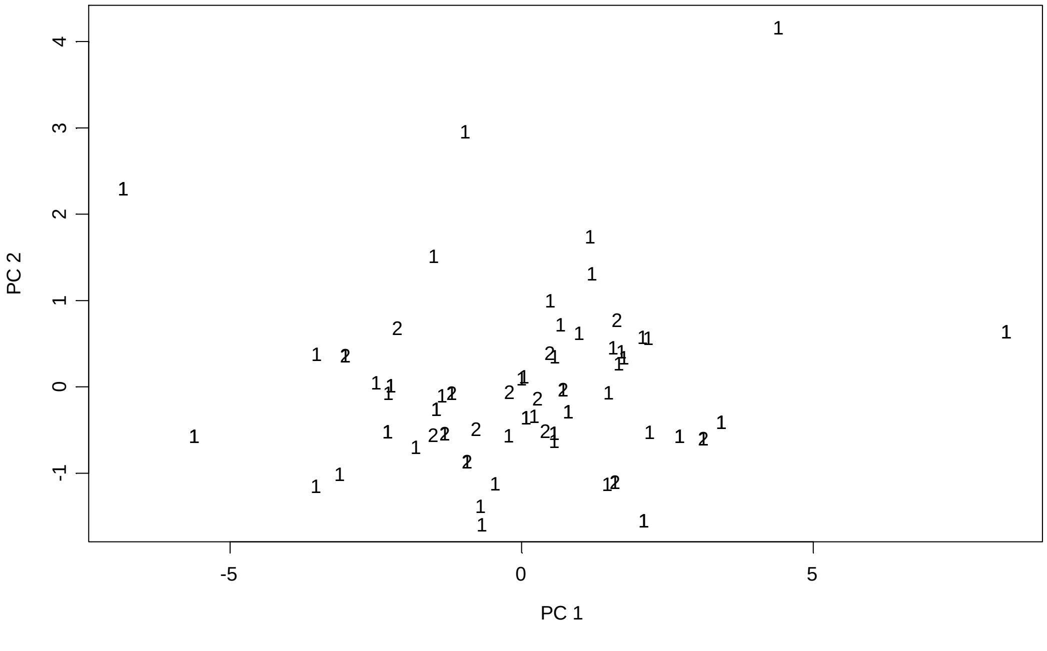

Fig. 3: Plot of the first-two principal components of two-cluster classification of standardized morphological features of *Tilapia cabrea*.

Data on Morphological Features of *Tilapia cabrea*

S/N	Total length (mm)	Standard length (mm)	Body weight (g)	Gutted weight (g)	Head length (mm)	Snout length (mm)	Ocular length (mm)	Body depth (mm)	Head width (mm)	Mouth length (mm)
1	135	85	41.72	38.65	31	11	10	43	16	18
2	116	87	39.2	37.81	30	10	11	50	19	10
3	104	81	30.12	28.69	29	11	11	39	17	18
4	102	74	27	24.93	28	9	10	39	16	16
5	67	53	5.73	4.9	19	6	6	22	9	10
6	165	140	40.79	37.71	34	12	11	45	17	16
7	140	108	62.62	58.09	38	15	11	52	20	20
8	109	84	38.06	28.57	31	11	9	44	11	15
9	96	73	30.12	21.52	22	10	9	42	14	15
10	114	88	37.34	38.47	30	11	10	40	23	18
11	115	90	40.79	37.71	34	12	11	45	18	17
12	102	74	27	24.93	28	11	11	39	14	20
13	127	90	64.13	53.31	37	16	12	53	19	17
14	96	74	31.13	22.52	22	10	9	43	15	16
15	112	86	40.94	33.27	30	14	11	46	16	16
16	168	124	109.88	92.87	43	17	12	60	23	19
17	136	106	56.17	38	44	18	10	47	17	17
18	110	87	40.06	30.96	31	12	10	46	18	16
19	106	84	38.06	28.57	31	11	9	45	16	14
20	113	87	41.47	31.22	32	13	10	44	18	15
21	119	95	42.77	37.49	33	11	9	44	17	16
22	117	97	41.18	38.51	32	12	5	45	18	16
23	122	94	44.87	40.62	32	11	8	46	19	16
24	122	96	42.51	39.68	33	11	10	46	18	18
25	112	86	35.19	31.38	30	10	10	42	17	16
26	142	110	74.1	72.44	30	15	13	51	23	18
27	114	88	37.34	35.47	30	11	10	40	17	12
28	105	80	26.42	24.65	30	10	10	41	16	16
29	116	90	45.43	38.29	32	11	10	54	17	20
30	121	94	47.57	45.17	33	13	11	46	18	20
31	134	103	49	4.47	33	14	11	46	18	20
32	125	100	48.38	42.1	34	13	12	49	20	18
33	134	106	57.68	47.47	36	14	12	53	24	22
34	128	100	57.04	50.34	38	14	10	53	21	20
35	125	100	51.67	45	33	15	11	49	20	18
36	130	103	55.82	47.46	37	14	11	50	20	22
37	124	94	51.42	45.75	35	14	11	51	19	20
38	150	121	74.43	66.11	40	19	13	58	21	24
39	126	102	48.88	43.11	35	13	11	49	19	18
40	123	93	47.73	42.2	33	12	11	49	18	22
41	102	74	27	24.91	28	11	11	39	14	20
42	115	90	40.79	37.71	34	12	11	45	18	17
43	114	84	37.34	35.47	30	11	10	40	23	18
44	96	73	30.12	21.52	22	10	9	42	14	15
45	109	84	38.06	28.57	31	11	9	44	16	15

46	140	108	62.62	58.09	38	15	11	52	20	20
47	115	90	40.79	37.71	34	12	11	45	17	16
48	67	53	5.73	4.9	19	6	6	22	9	10
49	102	74	27	24.93	28	9	10	39	16	16
50	104	81	30.12	28.69	29	11	11	39	17	18
51	116	87	39.2	37.81	30	9	10	39	16	16
52	135	85	41.72	38.65	31	6	6	22	9	10
53	105	80	26.42	24.65	30	12	11	45	17	16
54	114	88	37.34	35.47	30	15	11	52	20	20
55	142	110	74.1	72.44	30	11	9	44	16	15
56	112	86	35.19	31.38	30	10	9	42	14	15
57	122	96	42.51	39.68	33	11	10	40	23	18
58	122	94	44.87	40.62	32	11	10	46	18	18
59	117	87	41.18	38.51	32	12	6	45	18	16
60	119	95	42.77	37.49	33	11	9	44	17	18
61	113	87	41.47	31.22	32	13	10	44	18	15
62	109	84	38.06	28.57	31	11	9	45	16	14
63	110	87	40.06	30.96	31	12	10	46	18	16
64	36	106	56.17	38	44	18	10	47	17	17
65	168	124	109.88	92.87	43	17	12	60	23	19
66	112	86	40.94	33.27	30	14	11	46	16	16
67	96	74	31.13	22.52	22	10	9	43	15	16
68	127	100	64.13	53.31	37	16	12	53	19	17
69	123	93	47.73	42.2	33	12	11	49	18	22
70	126	102	48.88	43.11	35	13	11	49	19	18
71	150	121	74.43	66.11	40	19	13	58	21	24
72	124	94	51.42	45.74	35	14	11	57	19	20
73	130	103	55.82	47.56	37	14	11	50	20	22
74	125	100	51.67	45	33	15	11	49	20	18
75	128	100	57.04	50.34	38	14	10	53	21	20
76	104	145	57.68	47.47	36	14	12	53	24	22
77	125	100	48.38	42.1	34	13	12	49	20	18
78	134	103	49	47.47	33	14	11	46	18	20
79	116	90	45.43	38.29	32	11	10	45	17	20

CONCLUSION

This study demonstrates the usefulness of the cluster tree structure in the visualization of multidimensional data. The tree provides a hierarchical representation of the feature space. The tree structured rules were constructed by repeated splitting of subsets of the feature space. The procedure provides theoretically sound and consistent basis upon which decisions to group objects can be based.

REFERENCES

Becker, R. A., Chambers, J. M., Wilks, A. R. 1988. *The New S Language*. Chapman and Hall, London. U.K.

Bertin, J. 1967. *Semiologie Graphique*. Gauthier-Villars: Paris.

Chernoff, H. 1973. The use of faces to represent points in k-dimensional space graphically. *Journal of American Statistical Association,* 68: 361 - 368.

Gabriel, K. R. 1971. The Bi-plot Graphic Display of Matrices with Application to Principal Component Analysis. *Biometrika*, 58: 453-467.

Hartigan, J. A. 1975. *Clustering Algorithms*, John Wiley & Sons, New York.

Hurley, C. B. 2004. Clustering Visualizations of Multidimentional Data. *Journal of Computational and Graphical Statistics,* 13(4): 788-806.

Kaufman, L., Rousseeuw, P. J. 2005. *Finding Groups in Data*, John Wiley & Sons, New York.

Kleiner, B., Hartigan, J. A. 1981. Representing Points in Many Dimensions by Trees and Castles. *Journal of American Statistical Association*, 76: 260 - 269.

Leisch, F. 2005. A Toolbox for K-Centroids Clusters Analysis. *Computational Statistics and Data Analysis*, 51: 526-544.

Milligan, G. W., Cooper, M. C. 1985. An Examination of Procedures for Determining the Number of Clusters in a Data Set. *Psychometrika*, 50: 159 - 79.

Mojena, R. 1977. Hierarchical Grouping Methods and Stopping Rules: An Evaluation. *Computer Journal* 20: 359 - 63.

Trosset, M. W. 2005. Visualizing Correlation. *Journal of Computational and Statistical Graphics*, 14(1): 1-19.

Uchendu, B. A., Nwishi Chioma, C. H. 2007. Application of Multivariate Statistical Method in the Study of Morphological Features of Tilapia Cabrea. *Proceedings of the Nigerian Statistical Association.*

Wainer, H. 1983. *On Multivariate Display in Recent Advances in Statistics*, Edited by M. H., Rizzi, J. S., Rustagi and D. Sigmund, Academic, New York. Pp. 469-508.

(Manuscript received: 30th January, 2012; accepted: 1st November, 2012).