INSTITUT DE RECHERCHE
EN INFORMATIQUE ET EN AUTOMATIQUE

DATA ANALYSIS, LEARNING SYMBOLIC AND NUMERIC KNOWLEDGE

(Proceedings of the Conference on Data Analysis, Learning Symbolic and Numeric Knowledge)

Antibes
September 11 - 14, 1989

Edited by E. Diday

Nova Science Publishers, Inc.
New York•Budapest

Nova Science Publishers, Inc.
283 Commack Road
Suite 300
Commack, New York 11725

Library of Congress Cataloging-in-Publication Data
available upon request

ISBN 0-941743-64-0

Graphic Design by Elenor Kallberg and Peggy Harvey

Printed in the United States of America

SPACE-DISTORTING PROPERTIES IN AGGLOMERATIVE HIERARCHICAL CLUSTERING ALGORITHMS AND A SIMPLIFIED METHOD FOR COMBINATORIAL METHOD

Noboru Ohsumi
Department of Statistical Methodology
The Institute of Statistical Mathematics
4-6-7, Minami-Azabu, Minato-ku
Tokyo (Japan)
Nagatomo Nakamura
Department of Construction
Nihon University Junior College
7-24-1, Narashinodai, Funabashi-shi
Chiba (Japan)

1. Introduction

The objective of this paper is to examine the relationship between the properties of parameters which express agglomerative hierarchical clustering algorithms (AHC algorithms) and "space distortion" resulting from updated distances, and to generalize the Lance and Williams(1967) formula (combinatorial method). The AHC algorithms are based on the following formula proposed by Lance and Williams :

$$d(C_i \cup C_j, C_k) = \alpha_i\, d(C_i, C_k) + \alpha_j\, d(C_j, C_k) + \beta\, d(C_i, C_j) + \gamma\, |d(C_i, C_k) - d(C_j, C_k)| . \quad (1)$$

Using this formula, the dissimilarities, or distances, between a newly-merged cluster $C_i \cup C_j$ and the remaining other clusters C_k are updated and a set of parameters, $\theta \equiv \{\alpha_i, \alpha_j, \beta, \gamma\}$, are used to characterize the clustering methods and these parameters determine the linkage process.

With this formula, Lance and Williams introduced the concepts of "space distortion" (which may be space-conserving, space-contracting, or space-dilating) and the "monotonicity" of updated distances. One of the conditions for monotonicity are given by

$$\alpha_i + \alpha_j + \beta \geq 1. \qquad (2)$$

However, the AHC algorithms may produce reversal of the resulting tree structure. Thus, Milligan (1979) and Batagelj (1981) have presented following necessary and sufficient conditions for suppressing such reversals.

$$\gamma \geq -\min\{\alpha_i , \alpha_j\},$$
$$\alpha_i + \alpha_j \geq 0. \qquad (3)$$

At the same time, the concept of space distortion, as discussed by Lance and Williams, is intuitive. In addition, DuBien and Warde (1979) have derived, under some assumptions, a more sophisticated, theoretical concept of space distortion among distances obtained at different cluster merging levels. Unfortunately, these studies have concentrated on characterization of only a sub-family of the AHC algorithms, namely, the sub-family characterized by the (β,γ) space defined in formula (1).

2. Conditions for Space Distortion

Thus, this paper proposes several extensions to the concept of space distortion which will increase the sophistication of the AHC algorithms. Let us first assume the following condition concerning the distances among three clusters (C_i, C_j, and C_k):

$$d(C_i, C_j) < d(C_i, C_k) < d(C_j, C_k). \qquad (4)$$

In addition, let

$$\Delta^m = \{ d(C_i \cup C_j , C_k) \text{ at } \theta \mid d(C_i, C_j) < d(C_i, C_k) < d(C_j, C_k)\} \qquad (5)$$

be a set of all distances obtained from the result of an agglomeration at the m-th step in a clustering process. In this case, the conditions causing space distortion are defined as follows:

1) Space conservation:

$$d(C_i, C_k) < d(C_i \cup C_j , C_k) < d(C_j, C_k) ; \quad d(C_i \cup C_j , C_k) \in \Delta^m,$$

2) Space dilation:

$$d(C_j, C_k) \leq d(C_i \cup C_j, C_k)\ ;\ d(C_i \cup C_j, C_k) \in \Delta^m,$$

3) Space contraction:

$$d(C_i, C_k) \geq d(C_i \cup C_j, C_k)\ ;\ d(C_i \cup C_j, C_k) \in \Delta^m.$$

The paper examines the properties of space distortion occurring in various clustering methods and the results are summarized (see Table 2). Most clustering methods are based on general agglomerative algorithms using the $(\alpha_i, \alpha_j, \beta, \gamma)$ parameter space (i.e., formula (1)). However, examination of these strategies under the assumptions above clarifies the relationship between space distortion and the parameter space occupied by the particular strategy. Figure 1 shows the region occupied by various methods in the parameter space $(\alpha_i, \alpha_j, \beta)$, as defined below.

1) Region in which space conservation occurs:

$$\{ (\alpha_i, \alpha_j, \beta) \mid \alpha_i + \alpha_j + \beta = 1,\ 0 < \alpha_i, \alpha_j < 1,\ \beta = 0\}.$$

2) Region in which space conservation or space dilation occurs:

$$\{ (\alpha_i, \alpha_j, \beta) \mid \alpha_i + \alpha_j + \beta > 1,\ 0 < \alpha_i, \alpha_j < 1,\ -1 < \beta < 1\}$$
$$\cup \{ (\alpha_i, \alpha_j, \beta) \mid \alpha_i + \alpha_j + \beta = 1,\ 0 < \alpha_i, \alpha_j < 1,\ -1 < \beta < 0\}.$$

3) Region in which space conservation or space contraction occurs:

$$\{ (\alpha_i, \alpha_j, \beta) \mid \alpha_i + \alpha_j + \beta < 1,\ 0 < \alpha_i, \alpha_j < 1,\ -1 < \beta < 1\}$$
$$\cup \{ (\alpha_i, \alpha_j, \beta) \mid \alpha_i + \alpha_j + \beta = 1,\ 0 < \alpha_i, \alpha_j < 1,\ 0 < \beta < 1\}.$$

3. Simplification of the Lance and Williams formula

This paper proves that the single linkage and complete linkage methods are characterized as special cases of the flexible method by simplifying the Lance and Williams formula (1). First, the following two parameters are defined for updating the distances to usage two clusters:

$$\delta = d(C_j, C_k) - d(C_i, C_k) > 0,\ \ \varepsilon = d(C_i, C_k) - d(C_i, C_j) > 0. \qquad (6)$$

The proof involves substituting the values $\delta/(\delta+2\varepsilon)$ for parameter β to derive the single linkage method from the flexible method in formula (1).

$$d(C_i \cup C_j, C_k) = \{ d(C_i, C_k) + d(C_j, C_k) \}\ (1-\beta)/2 + \beta\, d(C_i, C_j)$$
$$= \{ d(C_i, C_k) + d(C_j, C_k) \}\ \varepsilon / (\delta + 2\varepsilon) + \delta\, d(C_i, C_j) / (\delta+2\varepsilon).$$

By substituting $d(C_j,C_k)= \delta + d(C_i,C_k)$, and $d(C_i,C_j)= -\varepsilon + d(C_i,C_k)$ for the above expression, we can obtain the following relation.

$$\begin{aligned} d(C_i \cup C_j, C_k) &= \varepsilon \{ 2d(C_i, C_k) + \delta \} /(\delta+2\varepsilon) + \delta \{ d(C_i,C_k) - \varepsilon \} /(\delta+2\varepsilon) \\ &= \{ 2\varepsilon d(C_i, C_k) + \varepsilon\delta + \delta d(C_i,C_k) - \varepsilon\delta \} /(\delta+2\varepsilon) \\ &= d(C_i, C_k). \end{aligned}$$

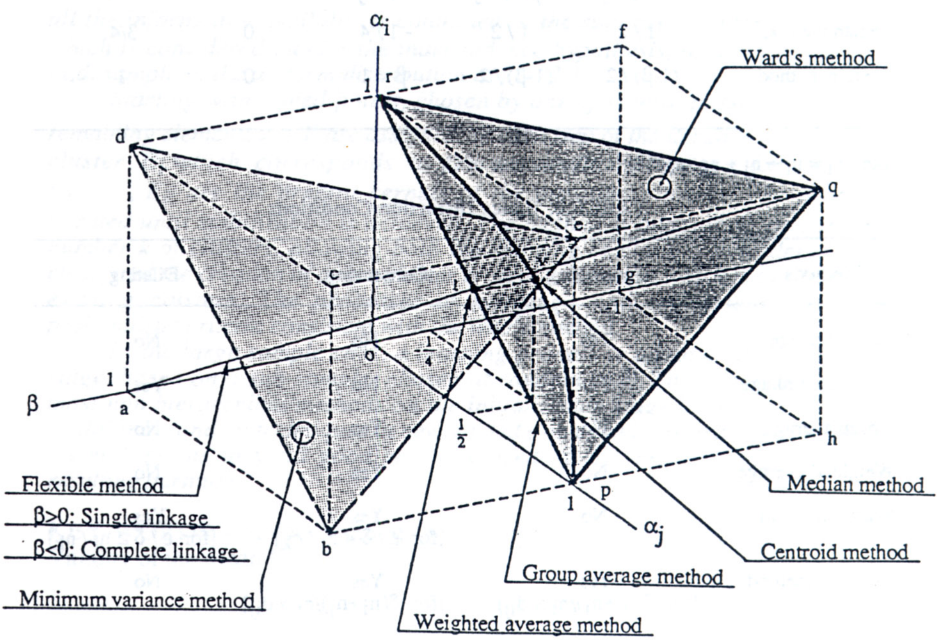
Similarly, substituting the value $-\delta/(\delta+2\varepsilon)$ for parameter β, we can obtain the complete linkage method. Thus, specifying parameter γ can be eliminated, and the formula (1) is simplified as following expression:

$$d(C_i \cup C_j, C_k) = \alpha_i\, d(C_i, C_k) + \alpha_j\, d(C_j, C_k) + \beta\, d(C_i, C_j).$$

Moreover, the condition of monotonicity which contains parameter γ is unnecessary to be considered. Thus, simplification of the Lance and Williams formula is completed.

REFERENCES

[1]Batagelj, V. (1981): Note on ultrametric hierarchical clustering algorithms. *Psychometrika*, **46**, 351-352.
[2]DuBien, J.L. and Warde, W.D. (1979): A mathematical comparison of the members of an infinite family of agglomerative clustering algorithms. *The Canadian Journal of Statistics*, **7**, 1, 29-38.
[3]Lance, G.N. and Williams, W.T. (1967): A general theory of classificatory sorting strategies, I. Hierarchical systems. *The Computer Journal*, **9**, 373-380.
[4]Milligan, G.W. (1979): Ultrametric hierarchical clustering algorithms. *Psychometrika*, **44**, 343-346.
[5]Nakamura, N. and Ohsumi, N. (1988): Space-distorting properties in agglomerative hierarchical clustering and a new simplified algorithm. *The Japan and Korea 5th conference.*
[6]Ohsumi, N. and Nakamura, N. (1988): Some relations among space-distorting properties in agglomerative hierarchical clustering algorithms (in Japanese). *Proceedings of the Annual Meeting of the Japanese Applied Statistics Society 1988*, 16-20.
[7]Ohsumi, N. and Nakamura, N. (1988): On a hierarchical classification method with the adjustable parameters (in Japanese). *Proceedings of the 2nd Annual Meeting of the Japanese Society of Computational Statistics*, 29-32.



α_i
f
Ward's method
1
r
d
q
e
c
g
-1
o
$-\frac{1}{4}$
1
β
a
$\frac{1}{2}$
h
Median method
Flexible method
β>0: Single linkage
β<0: Complete linkage
Minimum variance method
b
1
p
α_j
Centroid method
Group average method
Weighted average method

FIGURE 1

TABLE 1. Hierarchical Clustering Algorithms

Methods	α_i	α_j	β	γ	$\alpha_i+\alpha_j+\beta$
Single linkage	1 / 2	1 / 2	0	- 1 / 2	1
Complete linkage	1 / 2	1 / 2	0	1 / 2	1
Group average	$n_i/(n_i+n_j)$	$n_j/(n_i+n_j)$	0	0	1
Weighted average	1 / 2	1 / 2	0	0	1
Ward's method	$(n_i+n_k)/n_t$	$(n_j+n_k)/n_t$	$-n_k/n_t$	0	1
Centroid method	$n_i/(n_i+n_j)$	$n_j/(n_i+n_j)$	$-\alpha_i*\alpha_j$	0	$1+\beta$
Median method	1 / 2	1 / 2	- 1 / 4	0	3/4
Flexible method	$(1-\beta) / 2$	$(1-\beta) / 2$	$\beta < 1$	0	1

Note: $n_t = n_i + n_j + n_k$ and n_i is Cluster size of C_i.

TABLE 2. Space Distortion Conditions

Methods	Contracting	Conserving	Dilating
Single linkage	Yes	No	No
Complete linkage	No	No	Yes
Group average	No	Yes	No
Weighted average	No	Yes	No
Ward's method	No	Yes [for $\varepsilon / \delta < n_i / n_k$]	Yes [for $\varepsilon / \delta \geq n_i / n_k$]
Centroid method	Yes [for $\delta(n_i+n_j)/n_i \leq d_{ij}$]	Yes [for $\delta(n_i+n_j)/n_i > d_{ij}$]	No
Median method	Yes [for $2\delta \leq d_{ij}$]	Yes [for $2\delta > d_{ij}$]	No
Flexible method	Yes [for $\delta / (\delta+2\varepsilon) \leq \beta$]	Yes [for $-\delta / (\delta+2\varepsilon) < \beta < \delta / (\delta+2\varepsilon)$]	Yes [for $-\delta / (\delta+2\varepsilon) \geq \beta$]
Flexible method (Lance-Williams)	Yes [for $\beta > 0$]	Yes [for $\beta = 0$]	Yes [for $\beta < 0$]

Note: n_i is cluster size of C_i, $d_{ij} = d(C_i,C_j)$, $d(C_i,C_j) < d(C_i,C_k) < d(C_j,C_k)$,

$\delta = d(C_j,C_k) - d(C_i,C_k)$, $\varepsilon = d(C_i,C_k) - d(C_i,C_j)$.