ΑΝΙΟΥΣΑ ΙΕΡΑΡΧΙΚΗ ΤΑΞΙΝΟΜΗΣΗ

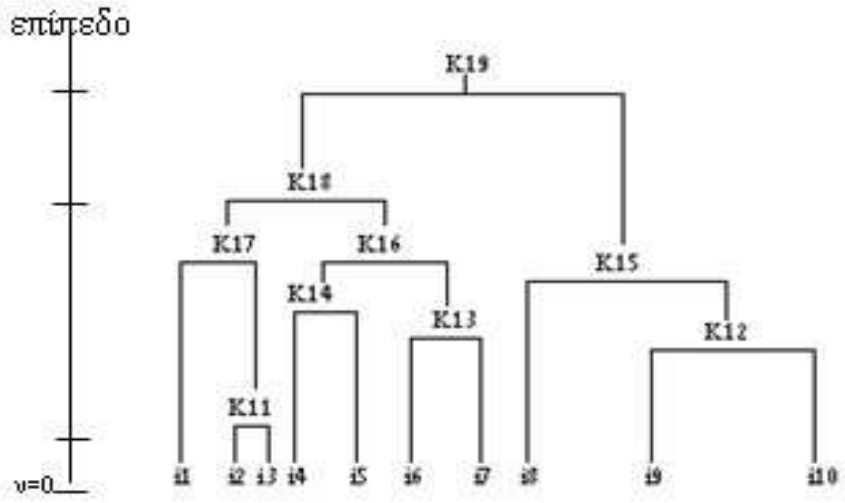
Μία **Ανιούσα Ιεραρχική Ταξινόμηση** (Classification Ascendante Hierarchique-CAH-) των στοιχείων ενός συνόλου I με πληθάριθμο card(I)=n, είναι μία διαδικασία που παράγει μια ακολουθία διαμελισμών του αρχικού συνόλου σε υποσύνολα μη κενά και ξένα ανά δύο μεταξύ τους, τις λεγόμενες **κλάσεις**, τη μία μέσα στην άλλη, συνενώνοντας κάθε φορά δύο μόνο κλάσεις οι οποίες βάσει κάποιας μετρικής παρουσιάζουν σε κάθε βήμα ομαδοποίησης την μικρότερη απόσταση.

Όσο απομακρύνεται κανείς από τον αρχικό διαμελισμό (ο οποίος περιλαμβάνει τόσες κλάσεις όσα είναι τα αντικείμενα που ταξινομούνται), τόσο αυτός γίνεται λιγότερο λεπτομερής.

Η τελευταία κλάση περιλαμβάνει το σύνολο των κλάσεων που δημιουργήθηκαν από τις συνενώσεις στοιχείων και κλάσεων, το πλήθος των οποίων είναι 2n-1, όπου n το πλήθος των στατιστικών μονάδων που ταξινομούνται.

Η ανιούσα ιεραρχική ταξινόμηση είναι μία μέθοδος που χρησιμοποιείται επίσης και ως συμπληρωματική, άλλων μεθόδων της ανάλυσης δεδομένων, όπως λ.χ στην Π.Α.Α για πληρέστερη ερμηνεία του παραγοντικού επιπέδου που προκύπτει μετά από κάθε ανάλυση.

Δενδρόγραμμα της ταξινόμησης



Σχήμα 4.1: Δενδρόγραμμα ταξινόμησης 10 στατιστικών μονάδων κατ' αύξουσα ιεραρχία

Στο σχήμα 4.1 τα i1,i2,…,i10 αντιπροσωπεύουν 10 στατιστικές μονάδες οι οποίες ομαδοποιούνται σε εννέα κλάσεις με πληθάριθμο μεγαλύτερο ή ίσο του 2.

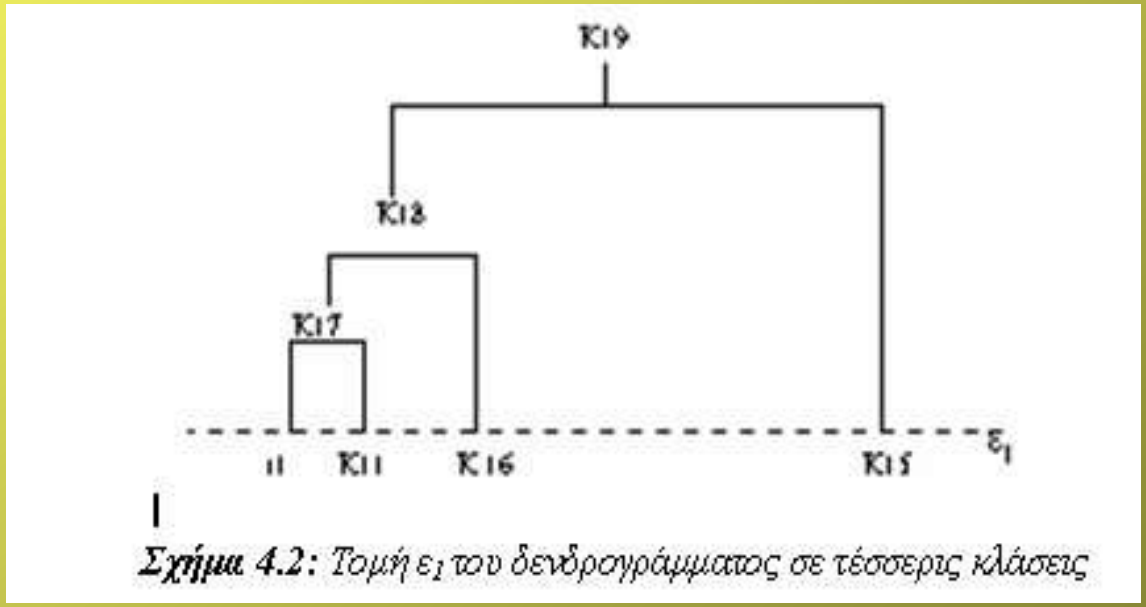
Συγκεκριμένα έχουμε:
την κλάση K11 ={i2,i3} την κλάση K12={i9,i10} την κλάση K13={i6,i7}
την κλάση K14={i4,i5} την κλάση K15={K12,i8} την κλάση K16={K13,K14}
την κλάση K17={i1,K11} την κλάση K18={K16,K17} και την κλάση K19={K15,K18}.

Το σύνολο των εννέα κλάσεων (K11 έως K19} ονομάζεται **τυπολογία** της ιεραρχίας

Τομή του δενδρογράμματος

Αν δεν ενδιαφερόμαστε για την συνολική ιεραρχία των παρατηρήσεων, αλλά μόνο για ένα περιορισμένο αριθμό κλάσεων, δεν έχουμε παρά να πάρουμε μία **«τομή»** του δενδρογράμματος στο επίπεδο ε1, δηλαδή να **"κόψουμε"** το δενδρόγραμμα με μία ευθεία γραμμή, στο σημείο όπου οι κλάδοι που απομένουν να ικανοποιούν τον αριθμό κλάσεων που επιθυμούμε να διατηρήσουμε.

Στο σχήμα 4.2 η ευθεία ε1 παρέχει τον διαμελισμό των 10 παρατηρήσεων σε τέσσερις κλάσεις. Τις ι1,Κ11,Κ16 και Κ15.



Σχήμα 4.2: Τομή ε1 του δενδρογράμματος σε τέσσερις κλάσεις

Απόσταση μεταξύ των στατιστικών μονάδων

Ο πίνακας δεδομένων που υποβάλλεται σε CAH, είναι ένας πίνακας αποστάσεων T(nxn) μεταξύ των n στατιστικών μονάδων. Για τον υπολογισμό των αποστάσεων d(i,i') δύο στατιστικών μονάδων υπάρχουν διάφορα είδη μετρικής, που χρησιμοποιούνται ανάλογα με την φύση των δεδομένων.

A) Για ποσοτικά δεδομένα

- **Η Ευκλείδεια μετρική**

$$d(i,i') = \sqrt{\sum_{j=1}^{p} \left[X(i,j) - X(i',j)\right]^2}$$

- **Η μετρική του City-block (Manhattan)**

$$d(i_i, i_j) = \sum_{k=1}^{n} |x_{ik} - x_{jk}|$$

- **Η μετρική του Mahalanobis**

$$d(i_i, i_j) = \sqrt{(x_i - x_j) \times C^{-1} \times (x_i - x_j)'}$$

- **Η μετρική του Tchebychev**

$$d(i_i, i_j) = \text{Max} \left| x_{ik} - x_{jk} \right|$$

B) Για ποιοτικά δεδομένα

- **Η μετρική του x^2**

$$d^2(i,i') = \sum_{j=1}^{p} \frac{1}{P_{.j}} (P_J^i - P_J^{i'})^2$$

Γ) Για δεδομένα που περιέχονται σε πίνακα συσχετίσεων

$d(i,i') = \sqrt{1 - r_{ij}}$ **όπου r ο συντελεστής συσχέτισης των μεταβλητών X_i και X_j**

Κριτήρια συνένωσης δύο κλάσεων

α) το κριτήριο της ελάχιστης απόστασης ή κριτήριο MIN

$$d(iUi', j)=\min\{d(i,j),d(i',j)\}$$

β) το κριτήριο της μέγιστης απόστασης ή κριτήριο MAX

$$d(iUi', j)=\max\{d(i,j),d(i',j)\}$$

γ) το κριτήριο MINMAX

$$d(iUi', j)=\min\{\max[d(i,j),d(i',j)]\}$$

δ) το κριτήριο της μέσης απόστασης

$$d(iUi', j) = \frac{p(i)d(i,j)+p(i')d(i',j)}{p(i)+p(i')+p(j)}$$

ε) το κριτήριο της απώλειας της διαταξικής αδράνειας

Όπου κατά την διαδικασία της ταξινόμησης η αδράνεια του νέφους των σημείων των παρατηρήσεων χωρίζεται με βάση το **θεώρημα του Huggens** σε δύο συνιστώσες, την **εσωταξική αδράνεια I_E** και την **διαταξική αδράνεια I_Δ**.

$$I_{ολ} = I_E + I_\Delta$$

Είναι φανερό ότι όσο πιο μεγάλη είναι η διαταξική αδράνεια I_Δ, τόσο πιο καλά διαχωρισμένες είναι μεταξύ τους οι κλάσεις.

Διαδικασία δημιουργίας της ιεραρχίας. Κριτήριο του Ward

Το κριτήριο ομαδοποίησης είναι να συγχωνεύουμε δύο κλάσεις Α και Β, όταν η παρουσιαζόμενη απώλεια της διαταξικής αδράνειας μεταξύ δυο κλάσεων είναι η ελάχιστη.

Ο αλγόριθμος της μεθόδου

Η διαδικασία δημιουργίας ενός δενδρογράμματος ανιούσας ιεραρχικής ταξινόμησης ενός συνόλου Ν «αντικειμένων» παρουσιάζει έξι βήματα.

1ο βήμα: Κάθε στοιχείο ενός συνόλου Ν «αντικειμένων» προς κατάταξη θεωρείται ότι αποτελεί μία κλάση, συνεπώς η αρίθμηση των κόμβων πριν την δημιουργία της πρώτης κλάσης, αρχίζει από το 1 και καταλήγει στο Ν. Στο βήμα αυτό υπολογίζονται τα προφίλ των «αντικειμένων».

2ο βήμα: Υπολογίζονται οι **αποστάσεις** όλων των προφίλ των «αντικειμένων» μεταξύ τους.

3ο βήμα: Μετατρέπονται οι αποστάσεις του 2ου βήματος σε **απώλειες διαταξικής αδράνειας**.

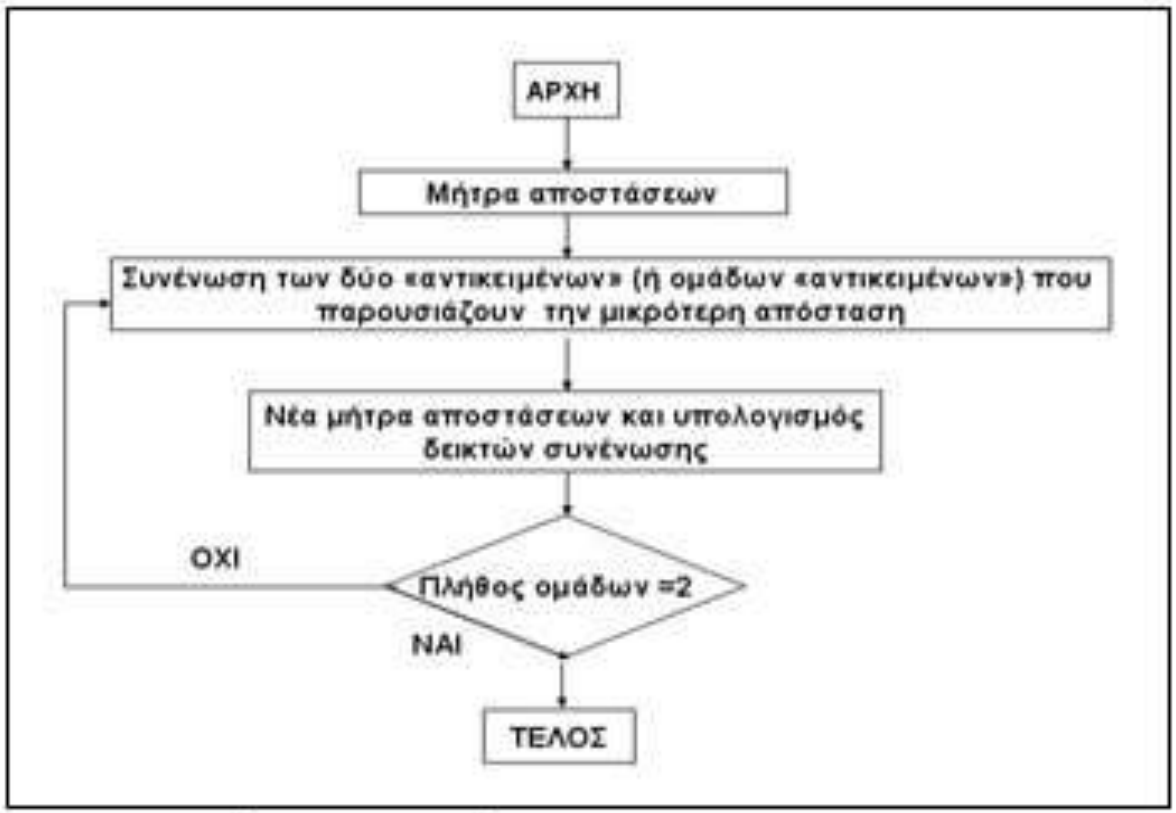
4ο βήμα: Συνενώνονται τα «αντικείμενα» που παρουσιάζουν την ελάχιστη απώλεια αδράνειας. Ο κόμβος που δημιουργείται παίρνει την αρίθμηση K_N+1. Υπολογίζεται ο δείκτης «δ» του κόμβου.

5ο βήμα: Υπολογίζονται οι απώλειες αδράνειας της νέας κλάσης K_N+1 με τα υπόλοιπα Ν-1 «αντικείμενα».

6ο βήμα: Επαναλαμβάνονται τα βήματα 4 και 5 μέχρι να καταλήξουμε σε μία κλάση που περιλαμβάνει και τα Ν «αντικείμενα». Η αρίθμηση του κόμβου που αντιστοιχεί στην τελευταία κλάση της ανιούσας ιεραρχίας είναι 2.Ν-1.

Σχηματικά ο αλγόριθμος της ανιούσας ιεραρχικής ταξινόμησης έχει ως εξής:

Σχηματικά ο αλγόριθμος της ανιούσας ιεραρχικής ταξινόμησης έχει ως εξής:



Σχήμα 4.6: Ο αλγόριθμος της CAH

Κριτήριο διαμελισμού της ιεραρχίας σε r κλάσεις

Το συνηθέστερο κριτήριο που λαμβάνεται υπόψη σ' αυτή την περίπτωση είναι ο παρακάτω λόγος:

$$\lambda_r = \frac{I_{\Delta_r}}{I_{o\lambda}}$$

όπου
$I_{\Delta r}$ είναι η διαταξική αδράνεια I_{Δ} που ερμηνεύεται από το διαμελισμό των σημείων του νέφους N(I) σε r κλάσεις.

Παράδειγμα

Πίνακας 4.10
Περιγραφή των κλάσεων

κλάσεις	A	B	I_E	I_Δ	λ_k
8	1	2	0.000	1.159	1.0000
9	3	4	0.003	1.156	0.9974
10	7	9	0.085	1.074	0.9267
11	5	6	0.188	0.971	0.8378
12	11	10	0.351	0.808	0.6972
13	8	12	1.159	0.000	0.0000

ΜΕΘΟΔΟΙ ΤΑΞΙΝΟΜΗΣΗΣ

➢Η μέθοδος VACOR

Η μέθοδος VACOR χρησιμοποιείται, συνήθως ανεξάρτητα από τις μεθόδους της ανάλυσης δεδομένων, για τη μελέτη των ιδιοτήτων των στατιστικών μονάδων και των αιτίων ομαδοποίησής των.

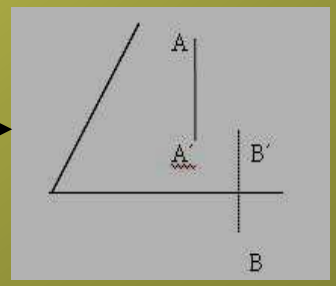
Η μέθοδος VACOR εφαρμόζεται δηλαδή απ' ευθείας στον πίνακα δεδομένων Τ(nxp), παρέχοντας το δενδρόγραμμα της ταξινόμησης, με το οποίο μας δίνεται η δυνατότητα να προσδιορίσουμε το ποσοστό συμβολής της κάθε μεταβλητής στον χαρακτηρισμό της κάθε κλάσης K_i (άρα και των διαδόχων της $A(K_s)$ και $B(K_r)$), καθώς και τις μεταβλητές που συμβάλλουν στη διάσπαση των κόμβων της ταξινόμησης.

➢Η μέθοδος FACOR

Όταν μετά από μία παραγοντική ανάλυση ο ερευνητής επιθυμεί να οριοθετήσει στο παραγοντικό επίπεδο τις διαμορφούμενες ομοιογενείς ομάδες, για καλύτερη ερμηνεία του φαινομένου, καλό είναι να πραγματοποιεί ταξινόμηση των στατιστικών μονάδων, χρησιμοποιώντας τη μέθοδο FACOR.

Με τον τρόπο αυτό η ομαδοποίηση των στατιστικών μονάδων που συντελείται στο 1ο παραγοντικό επίπεδο λαμβάνονται υπόψη όλοι οι παράγοντες (παραγοντικοί άξονες), συνεπώς διαθέτουμε πολύ περισσότερη πληροφορία απ' αυτή που μας παρέχουν οι δύο πρώτοι παράγοντες.

Σε πολλές περιπτώσεις δύο σημεία Α' και Β' του παραγοντικού επιπέδου 1Χ2 που εμφανίζονται το ένα δίπλα στο άλλο, στη πραγματικότητα είναι πολύ πιθανό να απέχουν μεταξύ τους σημαντικά, όπως διαπιστώνει κανείς εύκολα στο διπλανό σχήμα, οπότε η χρησιμότητα της FACOR αποδεικνύεται πολύ χρήσιμη, όταν θελήσουμε να περιγράψουμε στο παραγοντικό επίπεδο τις κλάσεις που διαμορφώνονται.



ΠΑΡΑΤΗΡΗΣΗ 1: Η αντιμετώπιση της συγκεκριμένης περίπτωσης όταν χρησιμοποιούνται τρεις παραγοντικοί άξονες, αντιμετωπίζεται με τα διαγράμματα Καραπιστόλη, (βλέπε ΜΕΘΟΔΟΛΟΓΙΑ ΤΗΣ ΕΡΕΥΝΑΣ)

ΠΑΡΑΤΗΡΗΣΗ 2: Ο εντοπισμός των μεταβλητών που χαρακτηρίζουν μία κλάση της ταξινόμησης αναλύεται διεξοδικά στο κεφάλαιο ΜΕΘΟΔΟΛΟΓΙΑ ΤΗΣ ΕΡΕΥΝΑΣ