ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ

ΤΜΗΜΑ ΣΤΑΤΙΣΤΙΚΗΣ
ΚΑΙ ΑΣΦΑΛΙΣΤΙΚΗΣ ΕΠΙΣΤΗΜΗΣ

ΜΕΤΑΠΤΥΧΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΣΠΟΥΔΩΝ
ΣΤΗΝ ΕΦΑΡΜΟΣΜΕΝΗ ΣΤΑΤΙΣΤΙΚΗ

ΜΕΘΟΔΟΙ ΕΥΡΕΣΗΣ ΒΕΛΤΙΣΤΟΥ ΠΛΗΘΟΥΣ ΟΜΑΔΩΝ ΓΙΑ ΠΟΛΥΔΙΑΣΤΑΤΑ ΔΕΔΟΜΕΝΑ

Φανή Ζαφειροπούλου

Διπλωματική Εργασία

που υποβλήθηκε στο Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς ως μέρος των απαιτήσεων για την απόκτηση του Μεταπτυχιακού Διπλώματος Ειδίκευσης στην Εφαρμοσμένη Στατιστική

Πειραιάς

Ιανουάριος 2007

Η παρούσα Διπλωματική Εργασία εγκρίθηκε ομόφωνα από την Τριμελή Εξεταστική Επιτροπή που ορίσθηκε από τη ΓΣΕΣ του Τμήματος Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς στην υπ' αριθμ. …….. συνεδρίασή του σύμφωνα με τον Εσωτερικό Κανονισμό Λειτουργίας του Προγράμματος Μεταπτυχιακών Σπουδών στην Εφαρμοσμένη Στατιστική

Τα μέλη της Επιτροπής ήταν:

- Μ. Κούτρας (Επιβλέπων)
- Δ. Καφφές
- Γ. Τζαβελάς

Η έγκριση της Διπλωματική Εργασίας από το Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς δεν υποδηλώνει αποδοχή των γνωμών του συγγραφέα.

UNIVERSITY OF PIRAEUS

DEPARTMENT OF STATISTICS AND INSURANCE SCIENCE

POSTGRADUATE PROGRAM IN APPLIED STATISTICS

METHODS OF DETERMINING THE OPTIMAL NUMBER OF CLUSTERS FOR MULTIDIMENSIONAL DATA

By

Fani Zafiropoulou

MSc Dissertation

submitted to the Department of Statistics and Insurance Science of the University of Piraeus in partial fulfillment of the requirements for the degree of Master of Science in Applied Statistics

Piraeus, Greece

January 2007

Περίληψη

Οι μέθοδοι ομαδοποίησης (cluster analysis) είναι τεχνικές της πολυμεταβλητής στατιστικής οι οποίες αποσκοπούν στη δημιουργία ομοιογενών ομάδων έτσι ώστε τα στοιχεία (παρατηρήσεις) που βρίσκονται στην ίδια ομάδα να παρουσιάζουν παρόμοια συμπεριφορά από άποψη κατανομής ενώ τα στοιχεία διαφορετικών ομάδων να αντιστοιχούν σε 'απομακρυσμένες' κατανομές.

Στα πλαίσια της εργασίας αυτής

α. θα παρουσιαστούν οι συνήθεις τεχνικές ομαδοποίησης (μη ιεραρχικές και ιεραρχικές μέθοδοι)

β. θα παρουσιαστούν τεχνικές εύρεσης βέλτιστου πλήθους αριθμών ομάδων

γ. με χρήση από case studies θα γίνει αριθμητική σύγκριση της αποτελεσματικότητας των διαφόρων μεθόδων ομαδοποίησης

δ. θα εξηγηθεί μέσα από (πραγματικά) παραδείγματα η χρησιμότητα των τεχνικών ομαδοποίησης δεδομένων

Abstract

The cluster analysis methods are techniques based on multivariate statistics that aim at the production of homogeneous groups in such a way that the data (units–observations) within the same cluster appear to have similar behavior as far as the aspect of distribution is concerned while the observations that belong to different clusters correspond to 'divergent' distributions.

In the framework of this text

a. The usual grouping techniques are going to be presented (hierarchical and non–hierarchical methods)
b. Techniques of finding the optimum number of clusters are going to be presented
c. Numerical comparison of the efficiency of the various clustering methods will be carried out through the use of case studies
d. The utilities of the data clustering techniques will be explained through the conduction of (real) examples

Περιεχόμενα

Κατάλογος Πινάκων

Κατάλογος Γραφημάτων

ΕΙΣΑΓΩΓΗ

Η ανάλυση κατά συστάδες ή ανάλυση σε ομάδες (Cluster Analysis) είναι μία μέθοδος που έχει ως στόχο να κατατάξει σε ομάδες τις υπάρχουσες παρατηρήσεις χρησιμοποιώντας την πληροφορία που υπάρχει σε κάποιες μεταβλητές. Πιο συγκεκριμένα, η ανάλυση κατά συστάδες εξετάζει πόσο όμοιες είναι κάποιες παρατηρήσεις ως προς κάποιο σύνολο μεταβλητών με σκοπό να δημιουργήσει ομάδες από παρατηρήσεις που μοιάζουν μεταξύ τους.

Η ανάλυση κατά συστάδες χρησιμοποιείται σε διάφορες επιστήμες για να ομαδοποιήσει δεδομένα και έχει παίξει καθοριστικό ρόλο στην ανάπτυξη πολλών επιστημονικών κλάδων, όπως στην αρχαιολογία, την ιατρική, την οικολογία, την ανάλυση προτύπων (Pattern recognition), την ψυχολογία, τις γεωεπιστήμες, την έρευνα αγοράς κλπ.

Ωστόσο, θα πρέπει να τονίσουμε ότι μερικές φορές η ανάλυση σε ομάδες μπορεί να έχει και άλλους σκοπούς από την απλή ομαδοποίηση των δεδομένων. Έτσι η ανάλυση σε ομάδες μπορεί να χρησιμοποιηθεί για

- να αποκτηθεί κάποια γνώση σχετικά με τα δεδομένα, αν για παράδειγμα παρουσιάζουν ομοιότητες κλπ.
- τη διερεύνηση σχέσεων στα δεδομένα. Συνήθως έχοντας ένα σύνολο δεδομένων στα χέρια μας έχουμε μια πολύ απλή ασαφή εικόνα για το τι περιέχουν τα δεδομένα, τι είδους σχέσεις υπάρχουν κλπ.
- τη μείωση των διαστάσεων του προβλήματος. Ειδικά στη σύγχρονη εποχή το πλήθος των δεδομένων που συγκεντρώνεται είναι τεράστιο χωρίς να σημαίνει ότι και η πληροφορία που περιέχεται σε αυτά είναι εξίσου μεγάλη. Υπάρχουν επικαλύψεις, μεταβλητές χωρίς ιδιαίτερο ενδιαφέρον κλπ. Επομένως ομαδοποιώντας τις μεταβλητές αποκτούμε μια εικόνα σχετικά με τις μεταβλητές που παρουσιάζουν ενδιαφέρον και επικεντρωνόμαστε σε αυτές.
- δημιουργία και έλεγχο υποθέσεων σχετικά με τα δεδομένα. Πολλές φορές ο ερευνητής υποψιάζεται την ύπαρξη κάποιων ομάδων με βάση κάποιο θεωρητικό μοντέλο που έχει στο μυαλό του.
- κατάταξη καινούριων παρατηρήσεων. Έχοντας δημιουργήσει ομάδες από παρατηρήσεις, σε πολλές εφαρμογές ενδιαφερόμαστε να κατατάξουμε καινούργιες παρατηρήσεις.

Οι έννοιες της απόστασης και της ομοιότητας είναι δύο βασικές έννοιες για την ανάλυση κατά συστάδες, άλλα όχι μόνο. Οι έννοιες αυτές ουσιαστικά ποσοτικοποιούν αυτό που στην καθημερινή γλώσσα εννοούν. Δηλαδή παρατηρήσεις που μοιάζουν πολύ μεταξύ τους, έχουν με απλά λόγια σχετικά όμοιες τιμές, θα πρέπει να έχουν πολύ μεγάλη τιμή για το μέτρο της ομοιότητας που θα χρησιμοποιήσουμε και πολύ μικρή απόσταση. Οι έννοιες αυτές είναι πολύ χρήσιμες καθώς μας επιτρέπουν να μετρήσουμε πόσο μοιάζουν οι παρατηρήσεις μεταξύ τους και επομένως να τις τοποθετήσουμε στην ίδια ομάδα.

Θα πρέπει να παρατηρήσουμε πως υπάρχουν αρκετές διαφορετικές προσεγγίσεις για το πώς μπορούμε να ομαδοποιήσουμε τα δεδομένα μας. Οι βασικότερες και πιο διαδεδομένες προσεγγίσεις είναι:

- Ιεραρχικές μέθοδοι: Ξεκινάμε με κάθε παρατήρηση να είναι από μόνη της μία ομάδα. Σε κάθε βήμα ενώνουμε τις 2 παρατηρήσεις ή ομάδες που έχουν την πιο μικρή απόσταση.
- Μη ιεραρχικές μέθοδοι: Ο αριθμός των ομάδων είναι γνωστός από πριν. Χρησιμοποιώντας έναν επαναληπτικό αλγόριθμο τοποθετούμε τις παρατηρήσεις στις ομάδες ανάλογα με το ποια ομάδα είναι πιο κοντά στην εκάστοτε παρατήρηση.

Η εύρεση του βέλτιστου αριθμού των ομάδων ενός συνόλου δεδομένων είναι ένα πρόβλημα το οποίο έχει απασχολήσει πολλούς ερευνητές που δραστηριοποιούνται σε αυτόν τον κλάδο. Οι αλγόριθμοι που χρησιμοποιούνται για την ομαδοποίηση των δεδομένων, και τους οποίους θα εξετάσουμε στη συνέχεια, χωρίζουν ένα σύνολο δεδομένων σε ομάδες χωρίς όμως να προσδιορίζουν τον ακριβή αριθμό ομάδων από τις οποίες απαρτίζονται τα δεδομένα. Για να γίνει κάτι τέτοιο πρέπει να εφαρμοσθεί κάποιο κριτήριο διακοπής του αλγορίθμου (Stopping Rule). Η βελτιστοποίηση ενός τέτοιου κριτηρίου μπορεί να τερματίσει έναν αλγόριθμο ομαδοποίησης σε ένα βήμα του ούτως ώστε ο αριθμός των ομάδων που έχει βρεθεί σε αυτό το βήμα να είναι ο καλύτερος δυνατός. Τα κριτήρια αυτά μπορεί να έχουν μεγαλύτερη ή μικρότερη επιτυχία στην εύρεση του αριθμού ομάδων.

Η επιτυχία ενός κριτηρίου διακοπής δεν εξαρτάται μόνο από την εύρεση του σωστού αριθμού των ομάδων, αλλά και από την ορθή τοποθέτηση των δεδομένων στις σωστές ομάδες. Με πραγματικά δεδομένα, είναι δύσκολη η αξιολόγηση ενός κριτηρίου τερματισμού αλγορίθμων ομαδοποιήσεως. Συνήθως για να αξιολογηθεί ένα τέτοιο κριτήριο,

χρησιμοποιούνται μέθοδοι προσομοίωσης, με δεδομένα, που παράγονται από γνωστές κατανομές. Με τον τρόπο αυτό ένας ερευνητής είναι σε θέση να καθορίσει εκ των προτέρων το πλήθος των ομάδων και να έχει πλήρη γνώση της τοποθέτησης των δεδομένων μέσα σε αυτές. Ένα μέτρο αξιολόγησης της επιτυχίας ενός κριτηρίου διακοπής ως προς τη σωστή τοποθέτηση των δεδομένων μέσα στις ομάδες, μπορεί να δοθεί με τη χρήση των εξωτερικών κριτηρίων (external criteria). Τα κριτήρια αυτά συγκρίνουν την πραγματική δομή των ομάδων με τη δομή των ομάδων που έχει προκύψει μέσω μίας διαδικασίας ομαδοποίησης.

Τελειώνοντας αυτήν την εισαγωγή, θα πρέπει να αναφέρουμε πως η εξάπλωση και η ευρεία χρήση των υπολογιστών σε θέματα ανάλυσης σε συστάδες, σε συνδυασμό με την ευρεία διαθεσιμότητα δεδομένων έχει οδηγήσει τα τελευταία χρόνια σε μία καινούργια θεώρηση των προβλημάτων της ανάλυσης που κυρίως βασίζεται στο πως θα αναλυθούν και θα ομαδοποιηθούν τα δεδομένα τεράστιων βάσεων. Αυτό έχει ως συνέπεια να δίνεται ολοένα και μεγαλύτερο βάρος στον υπολογιστικό φόρτο των μεθόδων και ερευνώνται μέθοδοι για να μπορέσει κανείς να διαχειριστεί μεγάλο όγκο δεδομένων. Το τίμημα που συνήθως πληρώνουμε είναι μικρότερη ακρίβεια και ορθότητα των αποτελεσμάτων.

ΚΕΦΑΛΑΙΟ 1

Όπως αναφέραμε και στην εισαγωγή, η ανάλυση κατά συστάδες είναι μία μέθοδος που σκοπό έχει να κατατάξει σε ομάδες τις υπάρχουσες παρατηρήσεις χρησιμοποιώντας την πληροφορία που υπάρχει σε κάποιες μεταβλητές. Μπορεί να πει κανείς πως εξετάζοντας πόσο όμοιες είναι κάποιες παρατηρήσεις ως προς κάποιον αριθμό μεταβλητών, η μέθοδος τείνει να δημιουργεί ομάδες από παρατηρήσεις που μοιάζουν μεταξύ τους.

Μια πετυχημένη ανάλυση θα πρέπει να καταλήξει σε ομάδες στις οποίες οι παρατηρήσεις μέσα σε κάθε ομάδα να είναι όσο γίνεται πιο ομοιογενείς, αλλά οι παρατηρήσεις διαφορετικών ομάδων να διαφέρουν όσο γίνεται περισσότερο.

1.1 Μέτρα Απόστασης και Μέτρα Ομοιότητας

Δύο πολύ βασικές έννοιες για την ανάλυση κατά συστάδες είναι οι έννοιες της απόστασης και της ομοιότητας. Μπορούμε εύκολα να διαπιστώσουμε πως αυτές οι δύο έννοιες είναι αντίθετες, αφού παρατηρήσεις που είναι όμοιες θα έχουν μεγάλη ομοιότητα και μικρή απόσταση. Οι έννοιες αυτές είναι πολύ χρήσιμες καθώς μας επιτρέπουν να μετρήσουμε πόσο μοιάζουν οι παρατηρήσεις μεταξύ τους και επομένως να τις τοποθετήσουμε στην ίδια ομάδα.

Επομένως, ο στόχος της ανάλυσης κατά συστάδες μετατοπίζεται στο να δημιουργήσουμε ομάδες μέσα στις οποίες οι παρατηρήσεις να απέχουν λίγο ενώ παρατηρήσεις διαφορετικών ομάδων να απέχουν μεταξύ τους αρκετά.

Τα δεδομένα που χρησιμοποιούνται για ομαδοποίηση αποτελούνται από n παρατηρήσεις. Κάθε μία από αυτές αναπαρίστανται από ένα διάνυσμα $x_i^{'} = (x_{i1}, x_{i2}, ..., x_{ip})$, όπου p είναι ο αριθμός των μεταβλητών που περιγράφουν μία παρατήρηση. Για το λόγο αυτό ένα σύνολο παρατηρήσεων μπορεί να παρασταθεί ως ένα σύνολο n σημείων σε ένα p-διάστατο χώρο. Οι n × p παρατηρήσεις μπορούν να συγκεντρωθούν σε ένα πίνακα $X = (x_{ij})$ με n γραμμές και p στήλες.

$$\begin{bmatrix} x_{11} & \Lambda & x_{1k} & \Lambda & x_{1p} \\ M & & M & & M \\ x_{i1} & \Lambda & x_{ik} & \Lambda & x_{ip} \\ M & & M & & M \\ x_{n1} & \Lambda & x_{nk} & \Lambda & x_{np} \end{bmatrix}$$

Ένας τέτοιος πίνακας θα λέγεται **πίνακας δεδομένων**.

Για παράδειγμα, στον επόμενο πίνακα δίνονται το βάρος (σε κιλά) και το ύψος (σε εκατοστά) οχτώ ανθρώπων. Στην περίπτωση αυτή, n=8 και p=2.

Όνομα	Βάρος	Ύψος
Πέτρος	16	93
Κατερίνα	50	158
Αγγελική	15	97
Γιώργος	43	163
Λεωνίδας	83	181
Δημήτρης	67	179
Μαρία	14	88
Χριστίνα	11	75

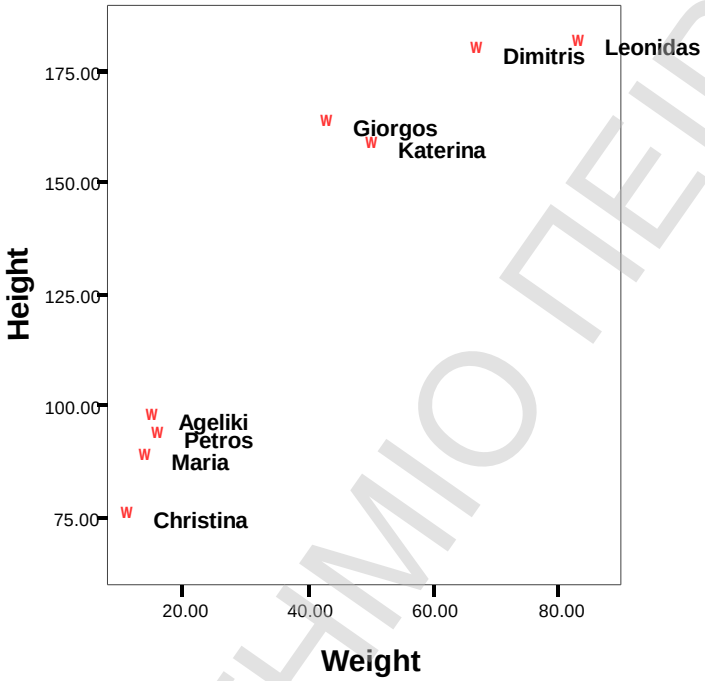
Πίνακας 1.1

Βάρος (σε κιλά) και ύψος (σε εκατοστά) 8 ατόμων

Βασιζόμενοι λοιπόν στα παραπάνω δεδομένα, ο πίνακας δεδομένων μας θα είναι ο εξής:

$$X = \begin{bmatrix} 16 & 93 \\ 50 & 158 \\ 15 & 97 \\ 43 & 163 \\ 83 & 181 \\ 67 & 179 \\ 14 & 88 \\ 11 & 75 \end{bmatrix}$$

Χρησιμοποιώντας τις τιμές των δύο μεταβλητών ως συντεταγμένες, τα παραπάνω δεδομένα απεικονίζονται στο γράφημα που ακολουθεί. Παρατηρούμε ότι στο γράφημα υπάρχουν δύο ομάδες. Η μία ομάδα αποτελείται από άτομα μικρής ηλικίας, ενώ η άλλη ομάδα αποτελείται από ενήλικες.



Γράφημα 1.1

Γραφική απεικόνιση οχτώ ατόμων ως προς τα χαρακτηριστικά βάρους και ύψους

1.1.1 Μέτρα Απόστασης

Τα μέτρα απόστασης είναι κατάλληλες ποσότητες που θα μπορούν να χρησιμοποιηθούν σε μία διαδικασία ομαδοποίησης για να μας δείξουν ότι δύο παρατηρήσεις είναι όμοιες ή ανόμοιες μεταξύ τους. Παρατηρήσεις που μοιάζουν πολύ μεταξύ τους, θα πρέπει να δίνουν πολύ μικρή τιμή στην απόσταση και επομένως να τις τοποθετούμε στην ίδια ομάδα.

Στη συνέχεια παρατίθενται κάποια μέτρα απόστασης τα οποία χρησιμοποιούνται πιο συχνά στην πράξη. Τα μέτρα αυτά έχουν χωριστεί σε ομάδες ανάλογα με το είδος των

δεδομένων στα οποία μπορούν να εφαρμοσθούν. Εμείς θα αναφέρουμε τις αποστάσεις εκείνες οι οποίες χρησιμοποιούνται για συνεχή και για δυαδικά δεδομένα.

(α) Μέτρα απόστασης για συνεχή δεδομένα

Η περίπτωση των συνεχών δεδομένων είναι ίσως η απλούστερη αλλά και η περισσότερο διαδεδομένη. Υπάρχουν πολλές αποστάσεις που έχουν χρησιμοποιηθεί για να μετρήσουν την απόσταση ανάμεσα σε συνεχή δεδομένα.

- **Ευκλείδεια Απόσταση**

Το πιο γνωστό μέτρο απόστασης μεταξύ δύο παρατηρήσεων είναι η ευκλείδεια απόσταση, η οποία στην ουσία είναι η γεωμετρική απόσταση στον πολυδιάστατο χώρο. Ορίζεται από τον τύπο:

$$d_{ij} = d(x_i, x_j) = \sqrt{\sum_{r=1}^{p} (x_{ir} - x_{jr})^2} \qquad (1.1)$$

Η ευκλείδεια απόσταση ικανοποιεί τις επόμενες τρεις ιδιότητες:

I1. $d_{ij} \geq 0$ για κάθε i, j και $d_{ij} = 0 \Leftrightarrow i = j$

I2. $d_{ij} \leq d_{is} + d_{sj}$ (τριγωνική ιδιότητα)

I3. $d_{ij} = d_{ji}$ (συμμετρική ιδιότητα)

Όταν έχουμε n παρατηρήσεις $x_1, x_2, ..., x_n$ τότε τις αποστάσεις τους $d_{ij} = d(x_i, x_j)$, i, j = 1,2,…,n τις τοποθετούμε σε έναν πίνακα $D = [d_{ij}]$ με n γραμμές και n στήλες τον οποίο θα λέμε **πίνακα αποστάσεων** των n σημείων. Λόγω των παραπάνω ιδιοτήτων είναι φανερό ότι όλα τα διαγώνια στοιχεία του πίνακα αποστάσεων θα είναι ίσα με 0 και ο πίνακας είναι συμμετρικός. Για το λόγο αυτό πολλές φορές δεν γράφουμε όλα τα στοιχεία του πίνακα, αλλά μόνο όσα βρίσκονται από την κύρια διαγώνιο και κάτω.

Χρησιμοποιώντας τα δεδομένα του προηγούμενου παραδείγματος προκύπτουν οι επόμενες αποστάσεις:

	Πέτρος	Κατερίνα	Αγγελική	Γιώργος	Λεωνίδας	Δημήτρης	Μαρία	Χριστίνα
Πέτρος	0	73.355	4.123	75.027	110.603	99.985	5.385	18.682
Κατερίνα	73.355	0	70.328	8.602	40.224	27.019	78.715	91.706
Αγγελική	4.123	70.328	0	71.694	108.074	97.098	9.055	22.361
Γιώργος	75.027	8.602	71.694	0	43.863	28.844	80.411	93.638
Λεωνίδας	110.603	40.224	108.074	43.863	0	16.125	115.802	128.141
Δημήτρης	99.985	27.019	97.098	28.844	16.125	0	105.309	118.119
Μαρία	5.385	78.715	9.055	80.411	115.802	105.309	0	13.342
Χριστίνα	18.682	91.706	22.361	93.638	128.141	118.119	13.342	0

Πίνακας 1.2

Οι αποστάσεις για τα δεδομένα του παραδείγματος

οπότε έχουμε τον παρακάτω πίνακα αποστάσεων:

$$D = \begin{bmatrix} 0 & & & & & & & \\ 73.355 & 0 & & & & & & \\ 4.123 & 70.328 & 0 & & & & & \\ 75.027 & 8.602 & 71.694 & 0 & & & & \\ 110.603 & 40.224 & 108.074 & 43.863 & 0 & & & \\ 99.985 & 27.019 & 97.098 & 28.844 & 16.125 & 0 & & \\ 5.385 & 78.715 & 9.055 & 80.411 & 115.802 & 105.309 & 0 & \\ 18.682 & 91.706 & 22.361 & 93.638 & 128.141 & 118.119 & 13.342 & 0 \end{bmatrix}$$

Η ευκλείδεια απόσταση αποτελεί την πιο απλή και την πιο γνωστή απόσταση ανάμεσα σε συνεχή δεδομένα. Παρατηρούμε μάλιστα ότι οι ευκλείδειες αποστάσεις συνήθως υπολογίζονται από τα πρωτογενή δεδομένα και όχι από τα κανονικοποιημένα δεδομένα. Η μέθοδος αυτή έχει συγκεκριμένα πλεονεκτήματα, όπως για παράδειγμα η απόσταση μεταξύ δύο οποιονδήποτε αντικειμένων δεν επηρεάζεται από την προσθήκη νέων αντικειμένων στην ανάλυση, τα οποία μπορεί να είναι outliers. Ωστόσο, οι αποστάσεις μπορούν να επηρεαστούν σε μεγάλο βαθμό από διαφορές στην κλίμακα μέτρησης μεταξύ των διαστάσεων από τις οποίες οι αποστάσεις υπολογίζονται. Επομένως, είναι πιο σωστό να μετασχηματίσουμε τις μεταβλητές ώστε να έχουν συγκρίσιμη κλίμακα. Ένας τρόπος για να φέρουμε όλες τις

μεταβλητές σε συγκρίσιμη κλίμακα είναι να διαιρέσουμε καθεμιά μεταβλητή με την τυπική της απόκλιση. Επομένως, αν συμβολίσουμε με s_r τη διακύμανση της r μεταβλητής

$$s_r = \left[\frac{1}{n-1} \sum_{i=1}^{n} (x_{ir} - \bar{x}_r)^2 \right]^{1/2}$$

όπου $\bar{x}_r = \frac{1}{n} \sum_{i=1}^{n} x_{ir}$ και r = 1,2,…,p, τότε η απόσταση που παίρνουμε έχει τη μορφή:

$$d_{ij} = d(x_i, x_j) = \sqrt{\sum_{r=1}^{p} \frac{(x_{ir} - x_{jr})^2}{s_r^2}} \qquad (1.2)$$

Η απόσταση αυτή είναι γνωστή με την ονομασία **απόσταση του Pearson** και επιτρέπει καλύτερες συγκρίσεις ανάμεσα στις μεταβλητές από ότι η ευκλείδεια απόσταση.

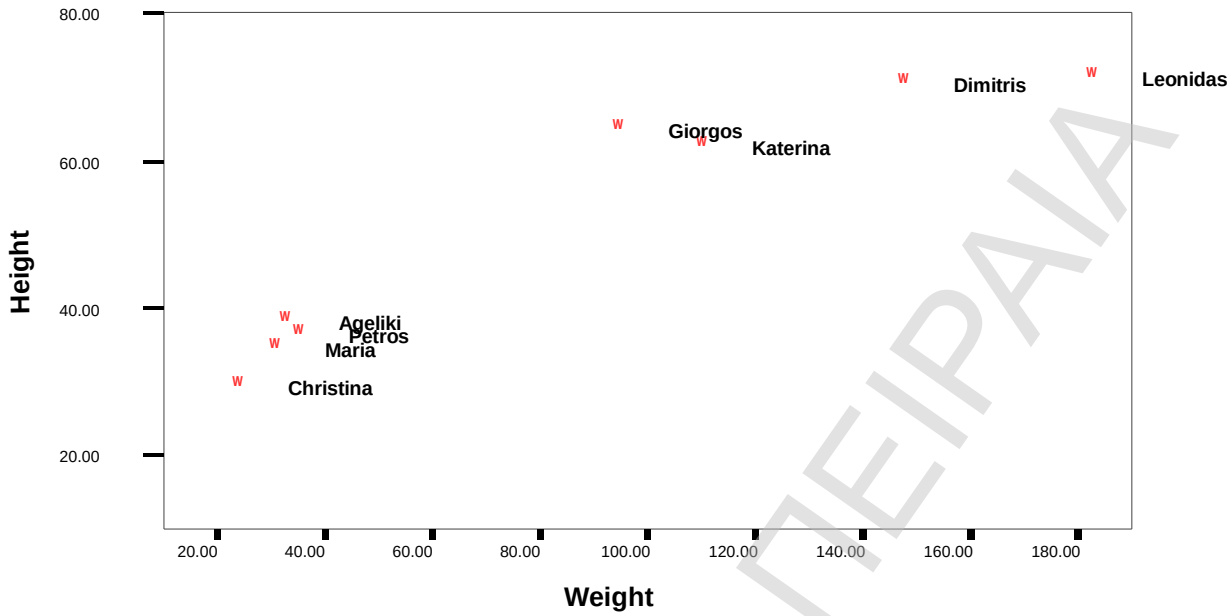
Για να γίνει κατανοητή η αναγκαιότητα αυτή ας επανέλθουμε στο παράδειγμα που παραθέσαμε στην προηγούμενη παράγραφο. Αν το βάρος και το ύψος των ατόμων εκφραστεί σε λίβρες και σε ίντσες αντίστοιχα, τότε τα δεδομένα μας θα είναι:

Όνομα	Βάρος	Ύψος
Πέτρος	35.27	36.61
Κατερίνα	110.23	62.20
Αγγελική	33.07	38.19
Γιώργος	94.80	64.17
Λεωνίδας	182.98	71.26
Δημήτρης	147.71	70.47
Μαρία	30.86	34.65
Χριστίνα	24.25	29.53

Πίνακας 1.3

Το βάρος (σε λίβρες) και το ύψος (σε ίντσες) των ατόμων

Μία λίβρα ισούται με 0.4536 κιλά και μία ίντσα είναι 2.54 εκατοστά. Για το λόγο αυτό, ο πίνακας 1.3 σε σχέση με τον πίνακα 1.1 περιέχει μεγαλύτερους αριθμούς στη στήλη με τα βάρη και μικρότερους αριθμούς στη στήλη με τα ύψη.



Γράφημα 1.2

Το γράφημα που αντιστοιχεί στον πίνακα 1.3

Παρόλο που τα γραφήματα 1.1 και 1.2 απεικονίζουν στην ουσία τα ίδια δεδομένα, παρατηρούμε ότι το γράφημα 1.2 είναι αρκετά πιο φαρδύ. Συγκεκριμένα, στο γράφημα 1.2 η παρουσία της μεταβλητής βάρος είναι πολύ πιο σημαντική από ότι στο γράφημα 1.1. Κατά συνέπεια, στο γράφημα 1.2 οι δύο ομάδες δεν διαχωρίζονται τόσο καλά όσο στο γράφημα 1.1 και αυτό γιατί στο συγκεκριμένο παράδειγμα ο παράγοντας του ύψους ενός ατόμου βοηθά στο να προσδιορίσουμε καλύτερα την ηλικία του ατόμου, από ότι ο παράγοντα του βάρους. Επομένως, για να αποφύγουμε αυτήν την εξάρτηση από την κλίμακα μέτρησης, αρκεί να φέρουμε όλες τις μεταβλητές σε συγκρίσιμη κλίμακα.

- **Η απόσταση του Mahalanobis**

Όπως είδαμε προηγουμένως, ένα μειονέκτημα της ευκλείδειας απόστασης είναι πως δεν λαμβάνει υπόψη τις συνδιακυμάνσεις ανάμεσα στις μεταβλητές. Μία τέτοια απόσταση που λαμβάνει υπόψη τις συνδιακυμάνσεις είναι η απόσταση του Mahalanobis, η οποία δίνεται από τον τύπο:

$$d_{ij}^2 = (x_i - x_j)\Sigma^{-1}(x_i - x_j)' \qquad (1.3)$$

όπου ο Σ είναι ο δειγματικός πίνακας διακύμανσης–συνδιακύμανσης. Στην ειδική περίπτωση όπου ο Σ είναι διαγώνιος με διαγώνια στοιχεία τα s_j^2, η απόσταση αυτή συμπίπτει με την απόσταση Pearson.

Για το παράδειγμα που εξετάζουμε, ο πίνακας αποστάσεων που προκύπτει χρησιμοποιώντας την απόσταση του Mahalanobis θα είναι ο εξής:

$$D = \begin{bmatrix} 0 & & & & & & & \\ 2.335 & 0 & & & & & & \\ 0.166 & 1.853 & 0 & & & & & \\ 5.224 & 1.408 & 3.782 & 0 & & & & \\ 7.193 & 5.651 & 8.469 & 12.433 & 0 & & & \\ 3.671 & 0.529 & 3.550 & 3.151 & 3.114 & 0 & & \\ 0.024 & 2.791 & 0.299 & 5.967 & 7.248 & 4.108 & 0 & \\ 0.597 & 4.741 & 1.361 & 9.282 & 6.940 & 5.715 & 0.386 & 0 \end{bmatrix}$$

- **Απόσταση Manhattan ή City-block metric**:

Η απόσταση Manhattan ορίζεται από τον τύπο:

$$d_{ij} = \sum_{r=1}^{p} \left| x_{ir} - x_{jr} \right| \qquad (1.4)$$

Η απόσταση αυτή είναι απλώς η μέση διαφορά μεταξύ των μεταβλητών. Στις περισσότερες περιπτώσεις, αυτό το μέτρο απόστασης μας δίνει παρόμοια αποτελέσματα με την ευκλείδεια απόσταση, εκτός από την περίπτωση που υπάρχουν έκτροπες παρατηρήσεις (outliers). Στην περίπτωση αυτή τους δίνει μικρότερο βάρος, αφού η διαφορά δεν υψώνεται στο τετράγωνο, οπότε μπορεί να οδηγηθούμε σε καλύτερα αποτελέσματα.

Χρησιμοποιώντας την απόσταση Manhattan, βρίσκουμε τον επόμενο πίνακα αποστάσεων για τα δεδομένα του παραδείγματος.

$$D = \begin{bmatrix} 0 & & & & & & & \\ 99 & 0 & & & & & & \\ 5 & 96 & 0 & & & & & \\ 97 & 12 & 94 & 0 & & & & \\ 155 & 56 & 152 & 58 & 0 & & & \\ 137 & 38 & 134 & 40 & 18 & 0 & & \\ 7 & 106 & 10 & 104 & 162 & 144 & 0 & \\ 23 & 122 & 26 & 120 & 178 & 160 & 16 & 0 \end{bmatrix}$$

- **Απόσταση Minkowski**:

Η απόσταση αυτή δίνεται από τη σχέση:

$$d_{ij} = \left(\sum_{r=1}^{p} \left| x_{ir} - x_{jr} \right|^{\lambda} \right)^{1/\lambda} \qquad (1.5)$$

Παρατηρούμε ότι η απόσταση Minkowski γενικεύει την Ευκλείδεια απόσταση και την απόσταση Manhattan, αφού η πρώτη προκύπτει για λ = 2, ενώ η δεύτερη για λ = 1.

Στο παράδειγμα που εξετάζουμε, αν χρησιμοποιήσουμε την απόσταση Minkowski για λ = 3 προκύπτει ο επόμενος πίνακας αποστάσεων:

$$D = \begin{bmatrix} 0 & & & & & & & \\ 67.964 & 0 & & & & & & \\ 4.021 & 64.622 & 0 & & & & & \\ 71.314 & 7.764 & 67.639 & 0 & & & & \\ 99.404 & 36.369 & 96.803 & 41.180 & 0 & & & \\ 91.605 & 24.201 & 88.450 & 26.169 & 16.010 & 0 & & \\ 5.104 & 73.040 & 9.004 & 76.418 & 104.246 & 96.636 & 0 & \\ 18.128 & 85.776 & 22.044 & 89.388 & 116.083 & 109.153 & 13.053 & 0 \end{bmatrix}$$

ενώ για λ = 4 ο πίνακας αποστάσεων θα είναι ο εξής:

$$D = \begin{bmatrix} 0 & & & & & & & \\ 66.184 & 0 & & & & & & \\ 4.004 & 62.590 & 0 & & & & & \\ 70.384 & 7.417 & 66.528 & 0 & & & & \\ 94.610 & 34.795 & 91.848 & 40.404 & 0 & & & \\ 88.544 & 22.962 & 85.131 & 22.006 & 16.001 & 0 & & \\ 5.032 & 71.193 & 9.000 & 75.416 & 99.362 & 93.512 & 0 & \\ 18.027 & 83.994 & 22.006 & 88.382 & 111.240 & 106.120 & 13.009 & 0 \end{bmatrix}$$

- **Απόσταση max ή Chebyshev**:

Αυτό το μέτρο απόστασης μας δείχνει ότι δύο παρατηρήσεις είναι διαφορετικές αν έχουν μεγάλες διαφορές σε μια τουλάχιστον μεταβλητή. Η απόσταση Chebyshev δίνεται από τον τύπο:

$$d_{ij} = \max_{r=1,2,\ldots,p} \left| x_{ir} - x_{jr} \right| \qquad (1.6)$$

Χρησιμοποιώντας τα δεδομένα του παραδείγματός μας παίρνουμε τον παρακάτω πίνακα αποστάσεων.

$$D = \begin{bmatrix} 0 & & & & & & & \\ 65 & 0 & & & & & & \\ 4 & 61 & 0 & & & & & \\ 70 & 7 & 66 & 0 & & & & \\ 88 & 33 & 84 & 40 & 0 & & & \\ 86 & 21 & 82 & 24 & 16 & 0 & & \\ 5 & 70 & 9 & 75 & 93 & 91 & 0 & \\ 18 & 83 & 22 & 88 & 106 & 104 & 13 & 0 \end{bmatrix}.$$

Αξίζει σε αυτό το σημείο να αναφέρουμε ότι, από τους πίνακες αποστάσεων που προέκυψαν από τη χρήση των διαφόρων τύπων απόστασης, παρατηρούμε ότι παρόλο που οι αποστάσεις μεταξύ ίδιων σημείων δεν παραμένουν ίδιες, εν τούτοις η σχετική διάταξη των αποστάσεων φαίνεται πως δε μεταβάλλεται.

(β) Μέτρα απόστασης για δίτιμες μεταβλητές

Τα μέτρα αυτά προκύπτουν από πίνακες διπλής εισόδου όπου το 1 παριστάνει την ύπαρξη μίας ιδιότητας και το 0 παριστάνει την απουσία αυτής.

Για να ορίσουμε ένα μέτρο απόστασης d_{ij} μεταξύ δυο ατόμων i και j, φτιάχνουμε αρχικά ένα πίνακα συνάφειας 2×2:

		άτομο j		
		1	0	
άτομο i	1	**a**	**b**	a + b
	0	**c**	**d**	c + d
		a + c	b + d	p

Πίνακας 1.4

Πίνακας συνάφειας

όπου a, b, c, d το πλήθος των συνδυασμών (1,1) (1,0) (0,1) (0,0) αντίστοιχα και $p = a + b + c + d$. Με βάση αυτούς τους συμβολισμούς οι πιο συνηθισμένες αποστάσεις που έχουν προταθεί είναι οι παρακάτω:

1. Simple matching distance $$d_{ij} = \frac{b + c}{a + b + c + d}$$
2. Rogers and Tanimoto distance $$d_{ij} = \frac{2(b + c)}{(a + d) + 2(b + c)}$$
3. Sokal and Sneath distance $$d_{ij} = \frac{b + c}{2(a + d) + (b + c)}$$
4. Jaccard distance $$d_{ij} = \frac{b + c}{a + b + c}$$
5. Dice and Sorensen distance $$d_{ij} = \frac{b + c}{2a + b + c}$$

Προτού συνεχίσουμε στα Μέτρα Ομοιότητας, θα ήταν χρήσιμο να παρατηρήσουμε το εξής: Στην ερώτηση «Ποιες μεταβλητές να χρησιμοποιήσουμε;» η απάντηση είναι ότι στην πραγματικότητα δεν υπάρχει κάποιος τρόπος για να μας οδηγήσει στην επιλογή μεταβλητών πριν κάνουμε την ανάλυση. Επομένως διαλέγουμε τις μεταβλητές που πιστεύουμε για κάποιος λόγους ότι έχουν τη δυνατότητα να δημιουργήσουν ομοιογενείς μεταβλητές. Αφού κάνουμε την ανάλυση μπορούμε εκ των υστέρων να δούμε αν κάποιες μεταβλητές τελικά ήταν αδιάφορες, με την έννοια ότι η τιμή τους είναι η ίδια για όλες τις ομάδες που δημιουργήσαμε, και επομένως δε μας προσφέρουν κάποια πληροφορία. Αν μάλιστα θεωρήσουμε ότι δεν μας προσφέρει αυτή η μεταβλητή κάτι σχετικά με την ερμηνεία που αναζητούμε μπορούμε να την αφαιρέσουμε και να χρησιμοποιήσουμε τις υπόλοιπες κάνοντας ξανά τη διαδικασία από την αρχή.

1.1.2 Μέτρα ομοιότητας

Ομοιότητα είναι ένας γενικός όρος, ο οποίος μπορεί να χρησιμοποιηθεί είτε ως συντελεστής ομοιότητας μεταξύ δεδομένων (ποιοτικών, ποσοτικών), ή ως μέτρο απόστασης ποσοτικών δεδομένων.

Τα μέτρα ομοιότητας (similarity measures ή affinity measures) μπορούν να χρησιμοποιηθούν για να μας δείξουν αν δύο παρατηρήσεις είναι όμοιες ή ανόμοιες μεταξύ τους. Συγκεκριμένα, παρατηρήσεις που μοιάζουν πολύ, δίνουν πολύ μεγάλη τιμή στο μέτρο της ομοιότητας, ενώ παρατηρήσεις που είναι ανόμοιες δίνουν πολύ μικρή τιμή.

(α) Δειγματικός συντελεστής συσχέτισης

Το πιο γνωστό μέτρο ομοιότητας για ποσοτικές παρατηρήσεις είναι ο (δειγματικός) συντελεστής συσχέτισης, που δίνεται από τον τύπο

$$s_{ij} = \frac{\sum_{r=1}^{p}(x_{ir} - \bar{x}_{i.})(x_{jr} - \bar{x}_{j.})}{\left(\sum_{r=1}^{p}(x_{ir} - \bar{x}_{i.})^2 \sum_{k=1}^{p}(x_{jr} - \bar{x}_{j.})^2\right)^{1/2}} \qquad (1.7)$$

όπου $\bar{x}_{i.} = \frac{1}{p}\sum_{r=1}^{p} x_{ir}$, $\bar{x}_{j.} = \frac{1}{p}\sum_{r=1}^{p} x_{jr}$

Ο δειγματικός συντελεστής συσχέτισης χρησιμοποιείται συνήθως για την ομαδοποίηση των μεταβλητών οι οποίες έχουν παρατηρηθεί σε έναν πληθυσμό. Επανερχόμενοι στο παράδειγμα της προηγούμενης παραγράφου, θα προσθέσουμε δύο νέες μεταβλητές (τον μήνα και το έτος γέννησης) και θα υπολογίσουμε τις συσχετίσεις μεταξύ των μεταβλητών.

Όνομα	Βάρος	Ύψος	Μήνας	Έτος
Πέτρος	16	93	1	82
Κατερίνα	50	158	5	55
Αγγελική	15	97	11	81
Γιώργος	43	163	7	56
Λεωνίδας	83	181	6	48
Δημήτρης	67	179	6	56
Μαρία	14	88	12	83
Χριστίνα	11	75	1	84

Πίνακας 1.5

Το βάρος, το ύψος, ο μήνας και το έτος γέννησης 8 ατόμων

Στον πίνακα 1.6 δίνονται οι δειγματικοί συντελεστές συσχέτισης μεταξύ των τεσσάρων μεταβλητών. Παρατηρούμε ότι ο δειγματικός συντελεστής συσχέτισης μεταξύ των μεταβλητών βάρος και ύψος είναι 0.953. Συγκεκριμένα, ο συντελεστής παίρνει μία τόσο μεγάλη τιμή γιατί φαίνεται να υπάρχει θετικός συσχετισμός μεταξύ των δύο αυτών μεταβλητών. Όπως άλλωστε φαίνεται και στο γράφημα 1.1, όσο μεγαλύτερο είναι το βάρος ενός ατόμου, τόσο ψηλότερο φαίνεται πως είναι το άτομο αυτό. Αντιθέτως, δεν υπάρχει καμία απολύτως συσχέτιση μεταξύ των μεταβλητών του βάρους και του μήνα γέννησης (ο μήνας της γέννησης ενός ατόμου δεν μπορεί να επηρεάσει και το βάρος του ατόμου), ενώ η συσχέτιση μεταξύ του βάρους και του έτους γέννησης είναι ισχυρά αρνητική (–0.948). Αυτή η αρνητική συσχέτιση έγκειται στο γεγονός ότι τα νεώτερα άτομα φαίνεται να έχουν μικρότερο βάρος. Ομοίως, μεταξύ του ύψους και του έτους γέννησης ενός ατόμου παρατηρείται ισχυρή αρνητική συσχέτιση, ενώ δεν υπάρχει καμία συσχέτιση μεταξύ ύψους και μήνα γέννησης.

	Βάρος	Ύψος	Μήνας	Έτος
Βάρος	1.000	.953	-.026	-.948
Ύψος	.953	1.000	.039	-.984
Μήνας	-.026	.039	1.000	.013
Έτος	-.948	-.984	.013	1.000

Πίνακας 1.6

Οι δειγματικοί συντελεστές συσχέτισης μεταξύ των τεσσάρων μεταβλητών

(β) Συντελεστές ομοιότητας για δυαδικά δεδομένα

Δημιουργούμε και πάλι τον προηγούμενο πίνακα συνάφειας 2×2 όπου a, b, c, d είναι το πλήθος των συνδυασμών (1,1) (1,0) (0,1) (0,0) αντίστοιχα και $p = a + b + c + d$.

Τα πιο γνωστά μέτρα ομοιότητας (similarity measures) που χρησιμοποιούνται για τις ανάγκες προβλημάτων ομαδοποίησης δεδομένων στα οποία παρατηρούνται δίτιμες μεταβλητές είναι τα εξής:

1. Simple matching $$s_{ij} = \frac{a + d}{a + b + c + d}$$

2. Rogers and Tanimoto $$s_{ij} = \frac{a + d}{(a + d) + 2(b + c)}$$

3. Sokal and Sneath $s_{ij} = \frac{2(a+d)}{2(a+d)+(b+c)}$

4. Jaccard coefficient $s_{ij} = \frac{a}{a+b+c}$

5. Dice and Sorensen $s_{ij} = \frac{2a}{2a+b+c}$

6. Rusel and Rao $s_{ij} = \frac{a}{a+b+c+d}$

7. Sokal and Sneath II $s_{ij} = \frac{a}{a+2(b+c)}$

8. Sokal and Sneath III $s_{ij} = \frac{a+d}{b+c}$

9. Kulczynski $s_{ij} = \frac{a}{b+c}$

Τα παραπάνω μέτρα ομοιότητας μπορούν να χωριστούν σε δύο κατηγορίες. Η πρώτη κατηγορία περιλαμβάνει τα μέτρα εκείνα τα οποία είναι χρήσιμα για συμμετρικές δυαδικές μεταβλητές. Στις συμμετρικές μεταβλητές οι τιμές 0 και 1 έχουν την ίδια σημασία και όλα τα κελιά έχουν την ίδια βαρύτητα. Στην κατηγορία αυτή ανήκουν ο simple matching coefficient, ο δείκτης των Rogers and Tanimoto και ο δείκτης των Sokal and Sneath.

Η άλλη κατηγορία περιλαμβάνει τα μέτρα εκείνα τα οποία χρησιμοποιούνται για ασύμμετρες δυαδικές μεταβλητές. Τέτοια μέτρα είναι ο Jaccard coefficient, ο δείκτης των Dice and Sorensen, καθώς και ο δείκτης Sokal and Sneath II. Στις ασύμμετρες δυαδικές μεταβλητές το κύριο βάρος πέφτει στο κελί (1,1), δηλαδή στην κοινή παρουσία κάποιων χαρακτηριστικών.

Ας υποθέσουμε ότι σε τρία άτομα παρατηρούμε την ύπαρξη ή την απουσία δέκα διαφορετικών χαρακτηριστικών και ότι τα αποτελέσματα που πήραμε έχουν ως εξής:

	Χαρακτηριστικά									
Άτομα	1	2	3	4	5	6	7	8	9	10
A	1	1	1	0	0	0	1	0	0	1
B	0	1	1	0	1	1	0	0	0	0
Γ	0	0	0	0	0	1	1	0	0	0

Πίνακας 1.7

Παράδειγμα με δυαδικά δεδομένα

Από τα δεδομένα, για τα άτομα Α και Β προκύπτει ο επόμενος πίνακας συνάφειας:

		Άτομο Β		
		1	0	
Άτομο Α	1	2	3	5
	0	2	3	5
		4	6	10

Πίνακας 1.8

Πίνακας συνάφειας για τα άτομα Α, Β

ενώ για τα άτομα Β και Γ ο πίνακας συνάφειας θα είναι ο εξής:

		Άτομο Γ		
		1	0	
Άτομο Β	1	1	3	4
	0	1	5	6
		2	8	10

Πίνακας 1.9

Πίνακας συνάφειας για τα άτομα Α, Γ

Ας χρησιμοποιήσουμε τον simple matching και τον Jaccard coefficient. Όσον αφορά τον simple matching coefficient παίρνουμε τις παρακάτω τιμές:

$$s_1(\text{A}, \text{B}) = 0.5 \quad \text{και} \quad s_1(\text{B}, \Gamma) = 0.6$$

ενώ με τον Jaccard coefficient έχουμε:

$$s_2(\text{A}, \text{B}) = 0.286 \quad \text{και} \quad s_2(\text{B}, \Gamma) = 0.2$$

Με τη χρήση του simple matching coefficient παρατηρούμε ότι τα άτομα Α και Β είναι λιγότερο όμοια από τα άτομα Β και Γ, ενώ ο Jaccard coefficient υποδεικνύει ακριβώς το αντίθετο, γεγονός το οποίο μπορεί να μας οδηγήσει σε διαφορετική ομαδοποίηση.

1.2 Κατάταξη των μεθόδων ομαδοποίησης

Οι μέθοδοι ομαδοποίησης μπορούν να χωριστούν σε δύο διαφορετικές κατηγορίες ανάλογα με τον τρόπο που προχωρούν στη διαμόρφωση των ομάδων: στις **ιεραρχικές** και στις **μη ιεραρχικές** μεθόδους.

Οι ιεραρχικές μέθοδοι χωρίζονται σε **συσσωρευτικές μεθόδους (agglomerative methods)** οι οποίες ακολουθούν μία σειρά διαδοχικών συγχωνεύσεων n παρατηρήσεων σε ομάδες και σε **διαιρετικές μεθόδους (divisive methods)** οι οποίες χωρίζουν ένα σύνολο n παρατηρήσεων διαδοχικά σε μικρότερες ομάδες. Και οι δύο τύποι των ιεραρχικών μεθόδων, παρόλο που δρουν κατά αντίθετο τρόπο, οι μεν ξεκινώντας από n ομάδες και καταλήγοντας σε μία, οι δε ξεκινώντας από μία ομάδα και καταλήγοντας σε n ομάδες, προσπαθούν να βρουν το βέλτιστο βήμα σε κάθε επίπεδο της διαδοχικής υποδιαίρεσης ή σύνθεσης των ομάδων.

Στην ιεραρχική ομαδοποίηση, ο αριθμός των ομάδων δεν είναι γνωστός από πριν. Επειδή χρησιμοποιούν έναν πίνακα αποστάσεων ή ισοδύναμα έναν πίνακα ομοιότητας χρειάζονται πολύ χρόνο και χώρο στον υπολογιστή και γι' αυτό είναι ασύμφορες για μεγάλα σετ δεδομένων, ενώ υπάρχει η τάση να δημιουργούνται ομάδες με ανομοιογενές μέγεθος.

Στις μη ιεραρχικές μεθόδους θεωρείται ότι ο αριθμός των ομάδων είναι γνωστός από πριν. Με έναν επαναληπτικό αλγόριθμο τοποθετούμε τις παρατηρήσεις στις ομάδες ανάλογα με το ποια ομάδα είναι πιο κοντά στην εκάστοτε παρατήρηση.

Οι μη ιεραρχικές μέθοδοι, ενώ αποφεύγουν τα προβλήματα που δημιουργούνται στις ιεραρχικές μεθόδους (δουλεύουν ικανοποιητικά με μεγάλα δείγματα και δημιουργούν ομάδες παραπλήσιου μεγέθους), εξαρτώνται πολύ από τις αρχικές τιμές που θα χρησιμοποιήσουμε.

Σε οποιαδήποτε μέθοδο θα πρέπει να τονιστεί ότι δυστυχώς υπάρχουν πολλά σημεία στα οποία ο ερευνητής μπορεί να λειτουργήσει υποκειμενικά, με αποτέλεσμα από τα ίδια δεδομένα να διεξαχθούν ακόμα και αντικρουόμενα αποτελέσματα. Από την άλλη, είναι γενική αλήθεια, πως όταν στα δεδομένα υπάρχουν πραγματικά ομοιογενείς ομάδες τότε οποιαδήποτε μέθοδος θα καταφέρει να τις αναγνωρίσει. Επομένως οι αντιφατικές λύσεις είναι μάλλον μία ένδειξη ότι δεν υπάρχει η κατάλληλη δομή στα δεδομένα μας, δηλαδή δεν υπάρχουν ομοιογενείς ομάδες.

1.2.1 Μη ιεραρχικές μέθοδοι ομαδοποίησης

Ο στόχος των μη ιεραρχικών μεθόδων είναι να ομαδοποιήσουν τις n μονάδες σε έναν προκαθορισμένο αριθμό ομάδων.

Η πρώτη γνωστή μέθοδος προτάθηκε από τον Forgy, ενώ η δεύτερη γνωστή τεχνική μη ιεραρχικής ομαδοποίησης είναι η **μέθοδος MacQueen** ή **k-means method**. Η μέθοδος MacQueen δουλεύει επαναληπτικά. Χρησιμοποιεί την έννοια του κέντρου της ομάδας (centroid) και στη συνέχεια κατατάσσει τις παρατηρήσεις ανάλογα με την απόστασή τους από τα κέντρα όλων των ομάδων. Το κέντρο της ομάδας δεν είναι τίποτα άλλο από τη μέση τιμή για κάθε μεταβλητή όλων των παρατηρήσεων της ομάδας. Ο αλγόριθμος αυτός έχει ικανοποιητική απόδοση για μεγάλα σύνολα δεδομένων, επειδή σε αυτήν την περίπτωση δουλεύει πολύ πιο γρήγορα από την ιεραρχική ομαδοποίηση.

Γενικά, η δυναμική του αλγορίθμου είναι πως με τις πρώτες λίγες επαναλήψεις πλησιάζει πολύ κοντά στην τελική λύση και στις υπόλοιπες επαναλήψεις οι διαφοροποιήσεις που προκύπτουν οφείλονται σε μετακίνηση μικρού αριθμού παρατηρήσεων που πιθανώς βρίσκονται στα σύνορα κάποιων ομάδων. Επομένως δεν είναι απαραίτητος ένας μεγάλος αριθμός επαναλήψεων, καθώς η βασική δομή θα σχηματιστεί πολύ γρήγορα.

Συνήθως η τελική ομαδοποίηση που δημιουργεί ο αλγόριθμος περιέχει ομάδες με ίσο περίπου αριθμό παρατηρήσεων.

Το μεγάλο μειονέκτημα του αλγορίθμου είναι ότι εξαρτάται από τα αρχικά κέντρα τα οποία αν δεν είναι σωστά επιλεγμένα μπορεί να οδηγήσουν σε ολότελα διαφορετική ομαδοποίηση από τη φυσική ομαδοποίηση που υπάρχει στα δεδομένα. Για να το ξεπεράσουμε αυτό, μια λύση είναι να τρέχουμε τη μέθοδο με διάφορες επιλογές ώστε να είμαστε σίγουροι πως δεν παγιδεύεται ο αλγόριθμος σε κάποια μη βέλτιστη λύση.

Δύο ακόμη μειονέκτημα της μεθόδου είναι τα εξής: Η ύπαρξη έκτροπων παρατηρήσεων (outliers) μπορεί να οδηγήσει στη δημιουργία ομάδων με πολύ διασπαρμένα στοιχεία. Επίσης, αν είναι γνωστό εκ των προτέρων ότι ο πληθυσμός που μελετάμε αποτελείται από k ομάδες, και συμβεί στο δείγμα μας να μην αντιπροσωπεύεται κάποια από αυτές (συνήθως η σπανιότερη), τότε απαιτώντας το διαχωρισμό σε k ομάδες, θα οδηγηθούμε σε αφύσικες (παραπλανητικές) ομαδοποιήσεις. Για να αποφεύγεται το τελευταίο πρόβλημα, καλό είναι να εφαρμόζουμε τον αλγόριθμο για διάφορες επιλογές του k και να συγκρίνουμε τα

αποτελέσματα χρησιμοποιώντας και τη διαίσθηση μας ώστε να πετύχουμε την καλύτερη ομαδοποίηση.

Τέλος αναφέρουμε τη **μέθοδο των σταθερών ομάδων** (stable clusters) η οποία δημιουργήθηκε με στόχο να μετριαστεί η επιρροή των αρχικών κέντρων στην πορεία που ακολουθεί ο αλγόριθμος του MacQueen (k-means).

1.2.2 Ιεραρχικές μέθοδοι ομαδοποίησης: Συσσωρευτικές μέθοδοι

Ο βασικός αλγόριθμος όλων των συσσωρευτικών μεθόδων (agglomerative methods) είναι περίπου ο ίδιος. Όλες οι μέθοδοι χρησιμοποιούν κάποιο συντελεστή ομοιότητας ή μία απόσταση που υπολογίζεται για όλους τους συνδυασμούς ανά δύο των υπό εξέταση αντικειμένων και έτσι διαμορφώνεται ο πίνακας αποστάσεων. Ο αλγόριθμος επιδρά στον πίνακα αποστάσεων και δημιουργεί ένα δενδρόγραμμα το οποίο δείχνει τις διαδοχικές συγχωνεύσεις των αντικειμένων μέχρι το επίπεδο που όλα τα αντικείμενα σχηματίζουν μία μόνο ομάδα (βλ. παράδειγμα, σελ 26). Σε κάθε επίπεδο της διαδικασίας κάθε ομάδα αποτελείται από τα επιπλέον όμοια αντικείμενα.

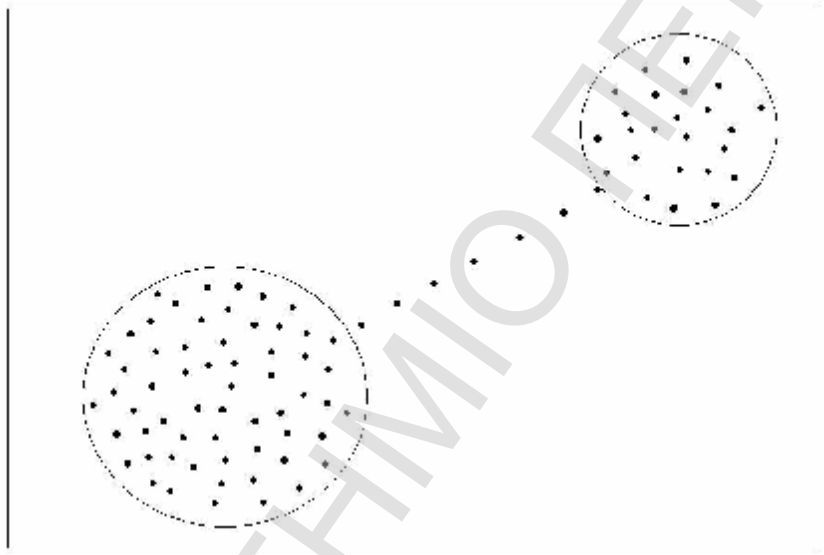
Οι συσσωρευτικές μέθοδοι διαφέρουν μεταξύ τους ως προς τον ορισμό της ομοιότητας μεταξύ των παρατηρήσεων κάθε ομάδας. Μερικοί αλγόριθμοι δίνουν καλύτερα αποτελέσματα για ορισμένους τύπους δεδομένων. Οι συνηθέστερα χρησιμοποιούμενες τεχνικές είναι οι ακόλουθες:

α) Μέθοδος της απλής συνένωσης (Single Linkage Method)

Η μέθοδος της απλής συνένωσης είναι η παλαιότερη και η απλούστερη όλων των ιεραρχικών μεθόδων. Η συνένωση των παρατηρήσεων γίνεται με τη χρήση της απόστασης των πλησιέστερων παρατηρήσεων. Σε κάθε συνένωση ο αριθμός των ομάδων ελαττώνεται κατά ένα. Επομένως, εδώ ορίζουμε ως απόσταση μεταξύ δυο ομάδων τη μικρότερη απόσταση από μια παρατήρηση μέσα στην μια ομάδα με κάποια παρατήρηση στην άλλη ομάδα. Στον ορισμό αυτό οφείλεται η δεύτερη ονομασία της μεθόδου που είναι «**μέθοδος του πλησιέστερου (κοντινότερου) γείτονα**» (nearest neighbor method).

Το βασικότερο μειονέκτημα του αλγορίθμου είναι ότι έχει την τάση να συνδέει απομονωμένα σημεία με ήδη υπάρχουσες ομάδες, αντί να δημιουργεί νέες. Έτσι σε δύο

ομάδες, οι οποίες έχουν έστω και ένα μόνο δεσμό μεταξύ τους (σημεία με μικρή απόσταση), τα μέλη που τις απαρτίζουν δεν μένουν χωριστά. Επομένως οι ομάδες που προκύπτουν από τη Single Linkage μέθοδο μπορεί να είναι κακώς διαμορφωμένες, με δύο μέλη που ανήκουν στην ίδια ομάδα συνδεδεμένα με μια αλυσίδα ενδιάμεσων σημείων. Το φαινόμενο αυτό λέγεται 'chaining' και φαίνεται γραφικά στο επόμενο σχήμα. Είναι φανερό ότι σε μία τέτοια περίπτωση, η ύπαρξη της 'αλυσίδας' που συνδέει τις δύο ομάδες θα οδηγήσει στο σχηματισμό μίας ενιαίας ομάδας χωρίς στην πραγματικότητα να δικαιολογείται αυτό από τη φύση των δεδομένων.



Γράφημα 1.3

Το φαινόμενο της 'αλυσίδας'

β) Μέθοδος της πλήρους συνένωσης (Complete Linkage Method)

Ως απόσταση μεταξύ των ομάδων χρησιμοποιεί την απόσταση των πιο απομακρυσμένων ζευγών σημείων, από τα οποία το ένα ανήκει στη μία ομάδα και το άλλο στην άλλη. Παρατηρούμε λοιπόν, ότι η μέθοδος αυτή σε σχέση με της μεθόδου του πλησιέστερου γείτονα βασίζεται στην ακριβώς αντίθετη έκφραση της απόστασης μεταξύ δυο ομάδων. Ο συγκεκριμένος αλγόριθμος δίνει έμφαση στην εσωτερική συνοχή των ομάδων. Έτσι οι ομάδες που δημιουργούνται με τη μέθοδο του μακρινότερου γείτονα είναι συνήθως μεγάλες και συμπαγείς, όμως η μέθοδος αρκετά συχνά αποτυγχάνει να ξεχωρίσει κάποιες πολύ συμπαγείς μικρές ομάδες.

γ) Μέθοδος των σταθμισμένων μέσων (Weighted Average Linkage Method)

Σε αυτή την περίπτωση η απόσταση είναι ο μέσος των αποστάσεων όλων των στοιχείων της μιας ομάδας με τα στοιχεία της άλλης.

δ) Μέθοδος των κέντρων βάρους (Centroid method)

Η απόσταση τώρα υπολογίζεται ως η απόσταση των κέντρων των ομάδων, δηλαδή ως κριτήριο συνένωσης σε αυτή τη μέθοδο λαμβάνεται η ελάχιστη απόσταση μεταξύ των κέντρων βάρους των ομάδων. Ένα μειονέκτημα της μεθόδου είναι ότι μπορεί να εφαρμοστεί μόνο σε ποσοτικά δεδομένα. Η μέθοδος Centroid έχει μερικές καλές ιδιότητες και παράγει συνήθως ομάδες συμπαγείς και ελλειπτικές.

ε) Μέθοδος του Ward (Ward's method)

Μια πολύ σημαντική μέθοδος της κατηγορίας αυτής είναι η μέθοδος του Ward (1963), υπολογιστικά βελτιωμένη από τον Wishart (1969), η οποία σε κάθε βήμα του αλγορίθμου της ελαχιστοποιεί την απώλεια πληροφορίας, μεταξύ αυτού του βήματος και του προηγούμενου. Ως απώλεια πληροφορίας λαμβάνεται η διαφορά του εντός των ομάδων αθροίσματος τετραγώνων των ομαδοποιήσεων δύο διαδοχικών βημάτων του αλγορίθμου. Ως γνωστόν το άθροισμα τετραγώνων ενός συνόλου δεδομένων, δίνει μία εικόνα του πόσο συγκεντρωμένα ή διεσπαρμένα είναι τα δεδομένα αυτά, γύρω από το μέσο όρο τους. Επομένως η τιμή του μπορεί να χρησιμοποιηθεί ως μέτρο ομοιογένειας αυτού του συνόλου.

Εάν τώρα, σε αυτό το σύνολο δεδομένων, επιδράσει κάποιος αλγόριθμος ομαδοποίησης, που λαμβάνει το άθροισμα τετραγώνων ως μέτρο συνοχής, τα δεδομένα μπορούν να θεωρηθούν, ότι έχουν ομαδοποιηθεί κατά τον καλύτερο τρόπο, όταν το εντός των ομάδων άθροισμα τετραγώνων είναι ελάχιστο. Σε κάθε βήμα του ιεραρχικού αλγορίθμου, πρέπει να υπάρχει μία βέλτιστη κατάσταση. Στην προκειμένη περίπτωση, το άθροισμα τετραγώνων θα πρέπει να είναι ελάχιστο. Επειδή δε, σε κάθε βήμα ενός ιεραρχικού αλγορίθμου το εντός των ομάδων άθροισμα τετραγώνων μεταβάλλεται, η πορεία του αλγορίθμου θα είναι ικανοποιητική, εάν η μεταβολή του αθροίσματος τετραγώνων είναι ελάχιστη.

Επομένως, η μέθοδος του Ward διαφέρει από τις υπόλοιπες και είναι σχεδιασμένη να ελαχιστοποιεί τη διακύμανση μέσα στις ομάδες. Τέλος, αναφέρουμε ότι η συγκεκριμένη μέθοδος έχει μερικές πολύ καλές ιδιότητες και συνήθως δημιουργεί ομάδες με παρόμοιο αριθμό παρατηρήσεων, γι' αυτό και χρησιμοποιείται στην πράξη πολύ συχνά.

στ) Μέθοδος του Gower (Gower's method)

Αυτή η μέθοδος είναι μια παραλλαγή της μεθόδου των κέντρων βαρών (centroid method) και εφαρμόζεται μόνο σε ποσοτικές μεταβλητές με χρήση της ευκλείδειας απόστασης.

ζ) Η Μέθοδος της Διαμέσου (Median Method)

Η μέθοδος αυτή θεωρεί ότι οι ομάδες που πρόκειται να συγχωνευθούν είναι του ίδιου μεγέθους. Ως ομάδες χρησιμοποιούνται τα κέντρα βάρους των δεδομένων στον ευκλείδειο χώρο. Η απόσταση του κέντρου k το οποίο σχηματίζεται από τη συγχώνευση των κέντρων βάρους i, j βρίσκεται πάνω στη διάμεσο του τριγώνου που σχηματίζεται από τα κέντρα βάρους i, j και k.

η) Η Μέθοδος του Μέσου Όρου των Ομάδων (Group Average Method)

Ως απόσταση μεταξύ των ομάδων λαμβάνεται ο μέσος όρος των αποστάσεων ανάμεσα σε όλα τα ζεύγη αντικειμένων των δύο ομάδων.

θ) Η Μέθοδος των Lance & Williams

Είναι ουσιαστικά γενίκευση των προηγουμένων μεθόδων, στην οποία τα μέτρα αποστάσης μεταξύ των ομάδων πληρούν έναν αναδρομικό τύπο. Για την απόσταση μεταξύ μίας ομάδος k και μιας ομάδος (ij) που προέκυψε από τη συγχώνευση των ομάδων i και j ο τύπος είναι:

$$d_{k(ij)} = \alpha_i d_{ki} + \alpha_j d_{kj} + \beta d_{ij} + \gamma \left| d_{ki} - d_{kj} \right| \qquad (1.8)$$

με τους περιορισμούς

$$\alpha_i + \alpha_j + \beta = 1,\ \alpha_i = \alpha_j,\ \beta < 1,\ \gamma = 0$$

Οι περισσότερες από τις μεθόδους ομαδοποίησης που αναφέρθηκαν προηγουμένως μπορούν να προκύψουν ως ειδικές περιπτώσεις της μεθόδου αυτής με κατάλληλη επιλογή των τιμών των συντελεστών α_i, α_j, β, γ. Πιο συγκεκριμένα έχουμε τα εξής:

(α) Για την Single Linkage, $\alpha_i = \alpha_j = 1/2,\ \beta = 0,\ \gamma = 1/2$

(β) Για την Complete Linkage, $\alpha_i = \alpha_j = 1/2,\ \beta = 0,\ \gamma = 1/2$

(γ) Για την Centroid, $\alpha_i = n_i/(n_i + n_j),\ \alpha_j = n_j/(n_i + n_j),\ \beta = -\alpha_i \cdot \alpha_j,\ \gamma = 0$

(δ) Για την Median, $\alpha_i = \alpha_j = 1/2,\ \beta = -1/4,\ \gamma = 0$

(ε) Για την Group Average, $\alpha_i = n_i/(n_i + n_j)$, $\alpha_j = n_j/(n_i + n_j)$, $\beta = \gamma = 0$

(στ) Για τη Μέθοδος του Ward, $\alpha_i = n_k + n_i/n_k + n_i + n_j$, $\beta = -n_k/n_k + n_i + n_j$, $\gamma = 0$

Συγκρίνοντας τις διάφορες μεθόδους μεταξύ τους με δεδομένα προσομοίωσης έχει διαπιστωθεί ότι συνήθως, η καλύτερη ομαδοποίηση επιτυγχάνεται με τη μέθοδο του Ward και τη μέθοδο των σταθμισμένων μέσων (Weighted Average Linkage Method). Η μέθοδος του κοντινότερου γείτονα είναι αυτή με τη χειρότερη επίδοση. Παρόλα αυτά σε πολλά προβλήματα δεν είναι ξεκάθαρο ποια μέθοδος είναι προτιμότερη και η καθεμιά δουλεύει καλύτερα με συγκεκριμένη μορφή δεδομένων. Φυσικά, αν οι ομάδες είναι αρκετά διαφορετικές μεταξύ τους όλες οι μέθοδοι θα οδηγήσουν στη σωστή ομαδοποίηση.

Για να δούμε πως δουλεύουν οι μέθοδοι που αναφέραμε προηγουμένως ας χρησιμοποιήσουμε τα δεδομένα που υπάρχουν στον πίνακα 1.10. Τα δεδομένα αυτά αφορούν 5 χώρες και τα αντίστοιχα ποσοστά γεννήσεων, θανάτων και βρεφικής θνησιμότητας.

Γεννήσεις ανά 1000 κατοίκους	Θάνατοι ανά 1000 κατοίκους	Θάνατοι βρεφών σε 1000 γεννήσεις	Χώρα
24.7	5.7	30.8	Αλβανία
12.5	11.9	14.4	Βουλγαρία
11.6	13.4	14.8	Ουγγαρία
14.3	10.2	16	Πολωνία
13.6	10.7	26.9	Ρουμανία

Πίνακας 1.10

Τα ποσοστά γεννήσεων, θανάτων και βρεφικής θνησιμότητας 5 χωρών

Χρησιμοποιώντας αυτές τις 3 μεταβλητές και την ευκλείδεια απόσταση καταλήγουμε τον επόμενο πίνακα αποστάσεων:

	Αλβανία	Βουλγαρία	Ουγγαρία	Πολωνία	Ρουμανία
Αλβανία	0.000				
Βουλγαρία	21.356	0.000			
Ουγγαρία	22.066	1.794	0.000		
Πολωνία	18.640	2.948	4.355	0.000	
Ρουμανία	12.784	12.607	12.558	10.934	0.000

Πίνακας 1.11

Ο πίνακας αποστάσεων για τις παρατηρήσεις

Βλέπουμε λοιπόν πως οι 2 κοντινές παρατηρήσεις είναι η Βουλγαρία και η Ουγγαρία, άρα αυτές θα ενωθούν και θα αποτελέσουν μία ομάδα. Το ζητούμενο είναι πώς θα υπολογίσουμε την απόσταση κάθε παρατήρησης από τις υπόλοιπες με αυτήν την ομάδα.

Σύμφωνα με τη μέθοδο του κοντινότερου γείτονα, η απόσταση θα είναι η μικρότερη από τις αποστάσεις των στοιχείων της ομάδας με κάθε παρατήρηση. Έτσι για την Αλβανία έχουμε:

d(Αλβανία, Βουλγαρία) = 21.356 και d(Αλβανία, Ουγγαρία) = 22.066

επομένως η απόσταση της Αλβανίας από την ομάδα θα είναι:

d(Αλβανία, {Βουλγαρία,Ουγγαρία}) = 21.356

Με τον ίδιο τρόπο βρίσκουμε και τα υπόλοιπα στοιχεία του πίνακα και έτσι καταλήγουμε στον επόμενο πίνακα αποστάσεων:

	Αλβανία	{Βουλγαρία,Ουγγαρία}	Πολωνία	Ρουμανία
Αλβανία	0.000			
{Βουλγαρία,Ουγγαρία}	21.356	0.000		
Πολωνία	18.640	2.948	0.000	
Ρουμανία	12.784	12.558	10.934	0.000

Πίνακας 1.12

Ο πίνακας αποστάσεων για τις παρατηρήσεις με τη χρήση της μεθόδου του κοντινότερου γείτονα

Στην περίπτωση του μακρινότερου γείτονα θα βρίσκαμε πως:

d(Αλβανία, {Βουλγαρία,Ουγγαρία}) = 22.066

και ο πίνακας αποστάσεων θα έπαιρνε την παρακάτω μορφή:

	Αλβανία	{Βουλγαρία,Ουγγαρία}	Πολωνία	Ρουμανία
Αλβανία	0.000			
{Βουλγαρία,Ουγγαρία}	22.066	0.000		
Πολωνία	18.640	4.355	0.000	
Ρουμανία	12.784	12.6056	10.934	0.000

Πίνακας 1.13

Ο πίνακας αποστάσεων για τις παρατηρήσεις με τη χρήση της μεθόδου του μακρινότερου γείτονα

Παρατηρούμε πως οι πίνακες διαφέρουν κι επομένως είναι πολύ πιθανό να οδηγήσουν σε διαφορετική ομαδοποίηση.

Συνεχίζοντας με την μέθοδο του κοντινότερου γείτονα και βάσει των αποστάσεων του πίνακα 1.12, θα συγχωνεύσουμε στην ομάδα {Βουλγαρία,Ουγγαρία} την Πολωνία και θα πάρουμε τον παρακάτω πίνακα αποστάσεων:

	Αλβανία	{Βουλγαρία,Ουγγαρία,Πολωνία}	Ρουμανία
Αλβανία	0.000		
{Βουλγαρία,Ουγγαρία,Πολωνία}	18.640	0.000	
Ρουμανία	12.784	10.934	0.000

Πίνακας 1.14

Πίνακας αποστάσεων για τις τρεις ομάδες

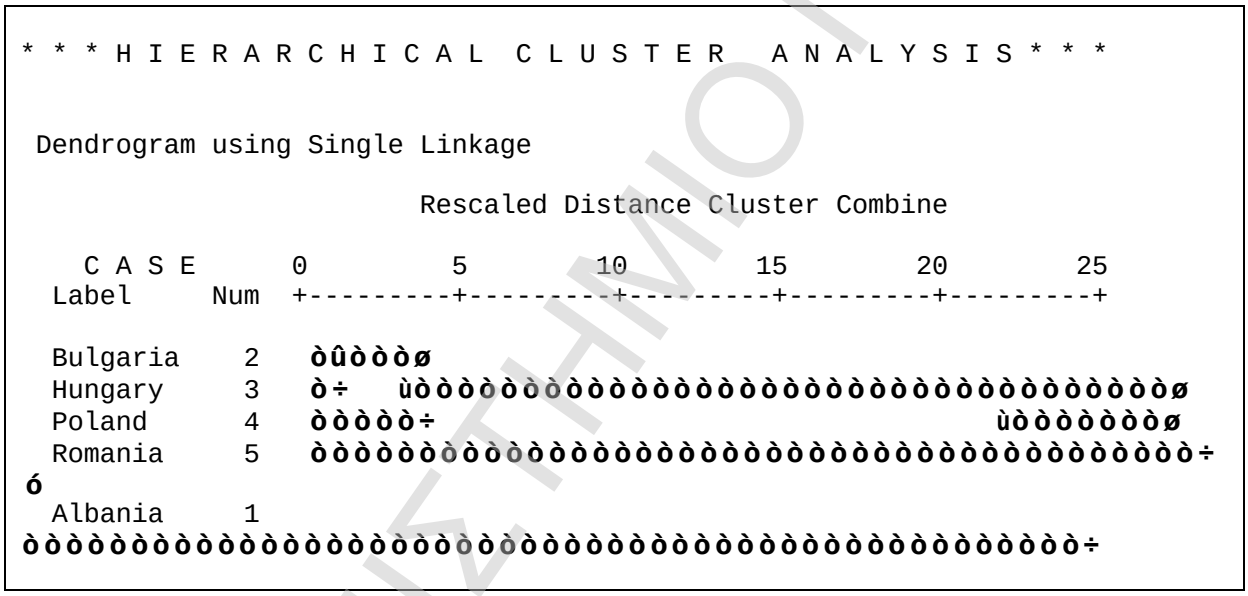
Στο τελευταίο βήμα θα ενώσουμε την υπάρχουσα ομάδα με τη Ρουμανία και θα προκύψει ο επόμενος πίνακας:

	Αλβανία	{Βουλγαρία,Ουγγαρία, Πολωνία,Ρουμανία}
Αλβανία	0.000	
{Βουλγαρία,Ουγγαρία, Πολωνία,Ρουμανία}	12.784	0.000

Πίνακας 1.15

Πίνακας αποστάσεων για τις δύο ομάδες

Μπορούμε να αναπαραστήσουμε όλη τη διαδικασία της ιεραρχικής ομαδοποίησης με ένα δενδρόγραμμα. Για την κατασκευή του γραφήματος χρησιμοποιήθηκε το στατιστικό πακέτο SPSS:



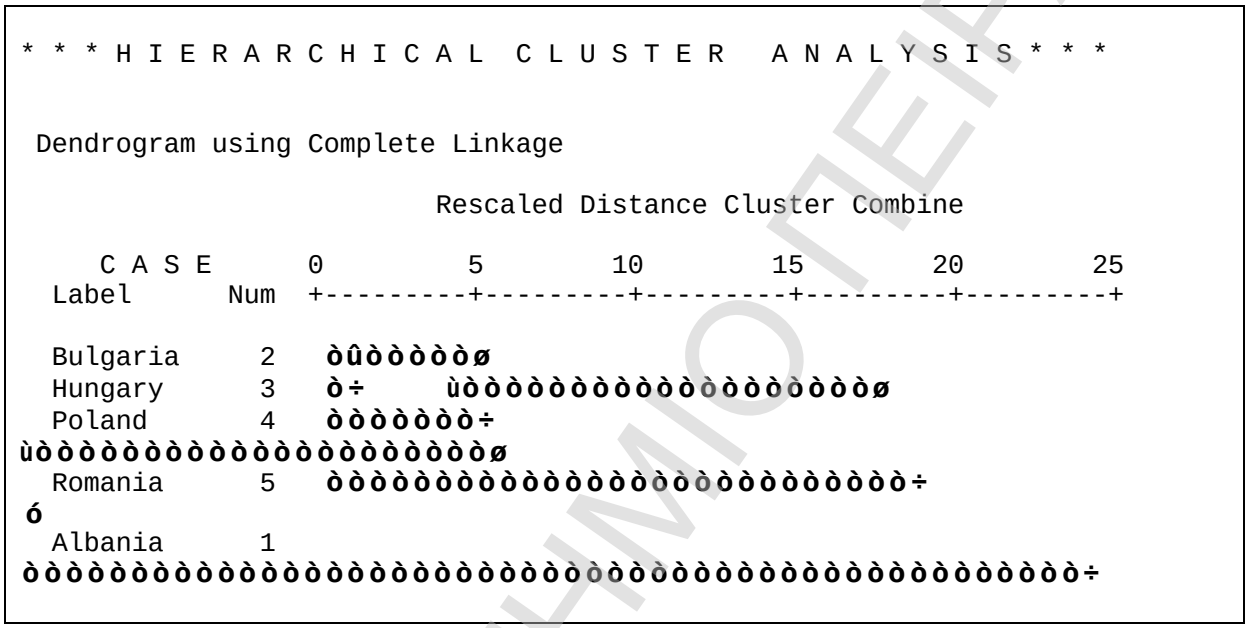
Γράφημα 1.4

Δενδρόγραμμα με τη χρήση της μεθόδου του κοντινότερου γείτονα

Το δενδρόγραμμα ενώνει με μία γραμμή τις παρατηρήσεις ή τις ομάδες που συγχωνεύονται σε κάθε βήμα. Αυτό επαναλαμβάνεται σε κάθε βήμα με αποτέλεσμα στο τέλος όλες οι παρατηρήσεις να είναι ενωμένες με κάποιο μονοπάτι. Συνήθως στον έναν άξονα έχουμε τις παρατηρήσεις και στον άλλον την τιμή της απόστασης με την οποία ενώσαμε τις παρατηρήσεις (ομάδες) ώστε να έχουμε μία ένδειξη πώς προχώρησε η διαδικασία. Όπως θα δούμε αργότερα, αυτό είναι χρήσιμο και για την εύρεση του βέλτιστου αριθμού ομάδων.

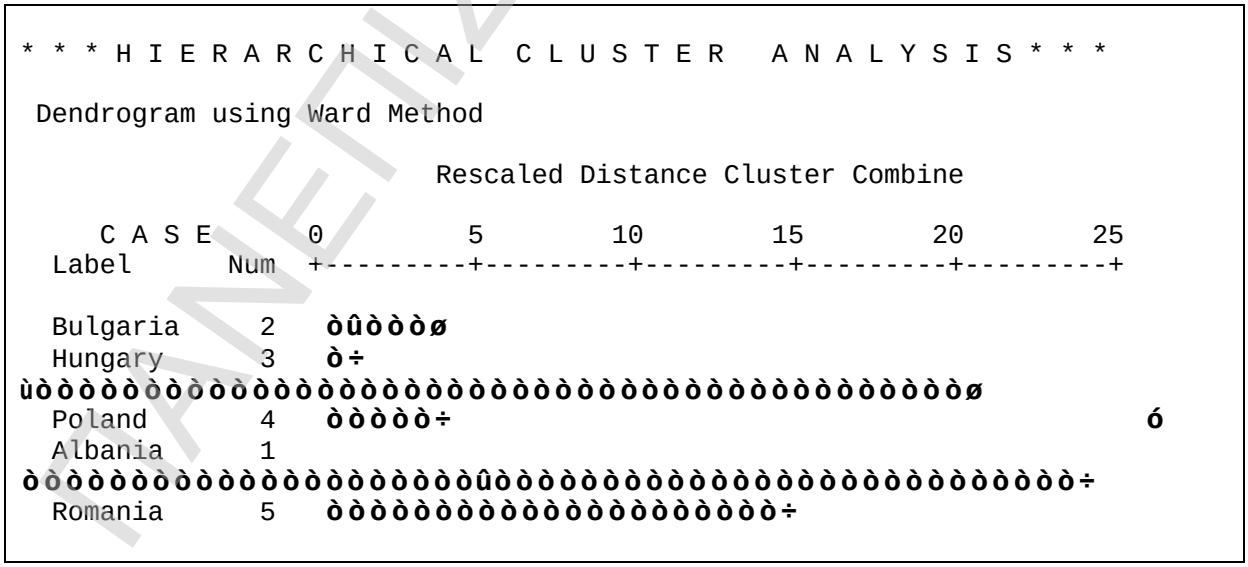
Επιστρέφοντας στο παράδειγμα, στη συνέχεια δίνονται τα δενδρογράμματα που προέκυψαν με την χρήση και άλλων μεθόδων ιεραρχικής ομαδοποίησης. Παρατηρούμε ότι

εκτός από τη μέθοδο του Ward, οι υπόλοιπες μέθοδοι καταλήγουν στο ίδιο δενδρόγραμμα με αυτό της μεθόδου του κοντινότερου γείτονα. Οι μόνες αλλαγές που παρατηρούνται είναι στο επίπεδο της απόστασης όπου γίνονται οι συγχωνεύσεις των ομάδων. Αντίθετα, χρησιμοποιώντας τη μέθοδο του Ward καταλήγουμε σε διαφορετικές ομάδες. Για παράδειγμα, η Αλβανία και η Ρουμανία σχηματίζουν μία ομάδα, γεγονός το οποίο δεν συμβαίνει με τις άλλες μεθόδους.



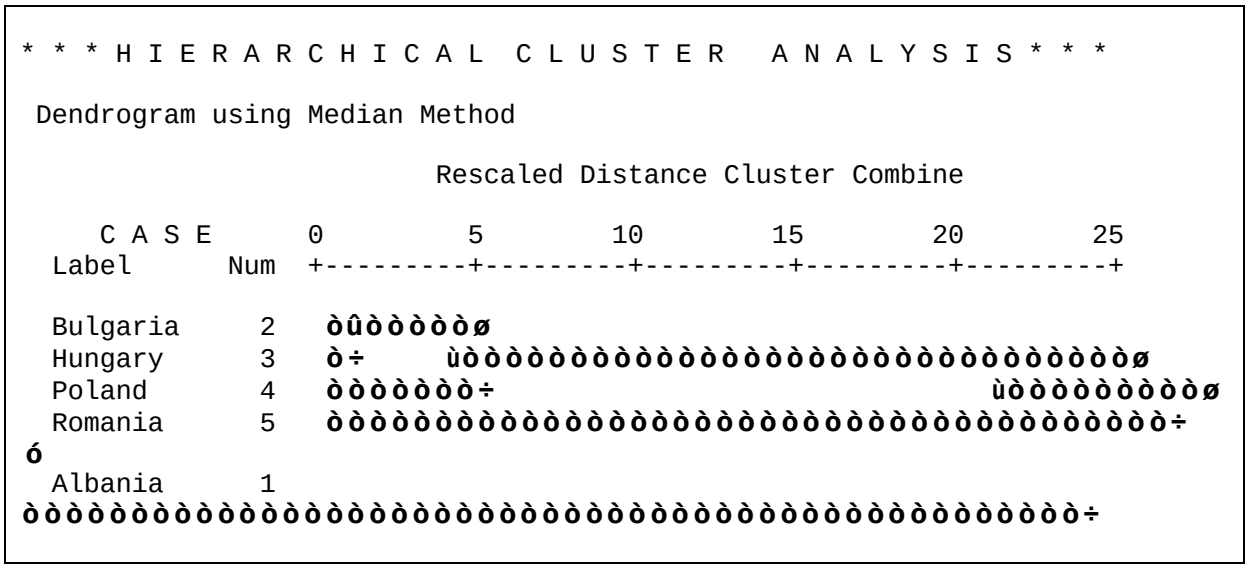
Γράφημα 1.5

Δενδρόγραμμα με τη χρήση της μεθόδου του μακρινότερου γείτονα



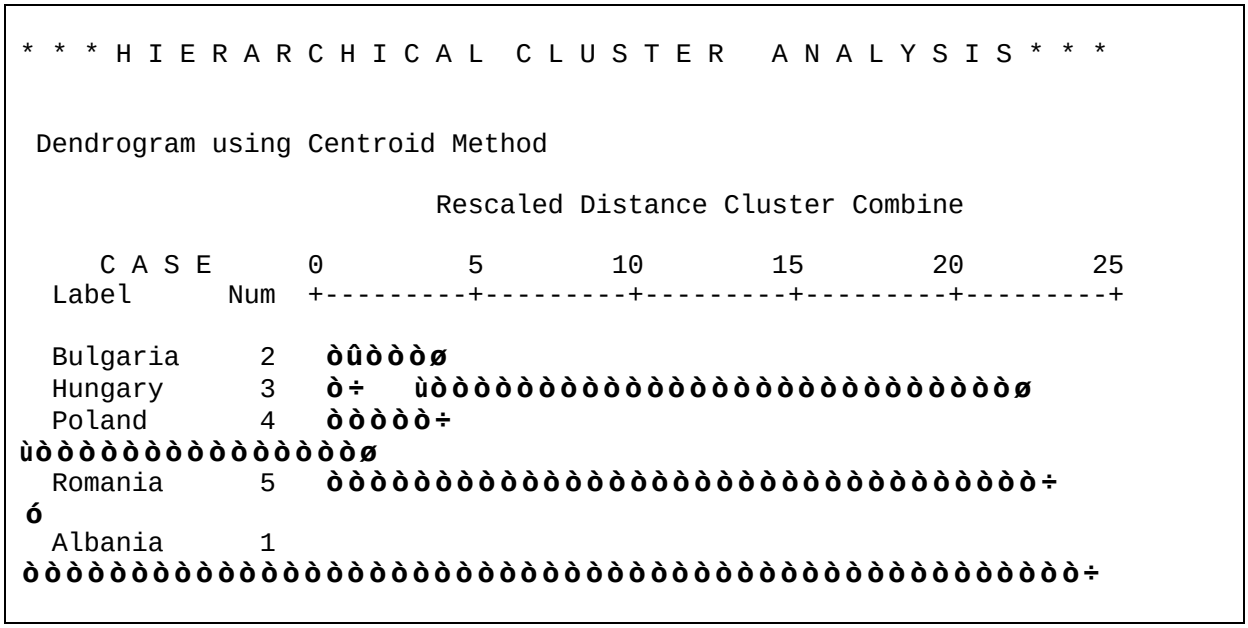
Γράφημα 1.6

Δενδρόγραμμα για τα δεδομένα με τη χρήση της μεθόδου του Ward



Γράφημα 1.7

Δενδρόγραμμα με τη χρήση της μεθόδου της διαμέσου



Γράφημα 1.8

Δενδρόγραμμα με τη χρήση της μεθόδου Centroid

Το δενδρόγραμμα αποτελεί ένα πολύτιμο οπτικό εργαλείο για την ιεραρχική ομαδοποίηση καθώς αφενός περιέχει ολόκληρη τη διαδικασία των διαδοχικών ομαδοποιήσεων και αφετέρου εμφανίζει την δυναμική της μεθόδου που επιλέξαμε. Για παράδειγμα, αυτό που

είδαμε στην εφαρμογή του κοντινότερου γείτονα είναι η χαρακτηριστική περίπτωση της συγκεκριμένης μεθόδου όπου οι παρατηρήσεις συνήθως ενσωματώνονται μία-μία σε μια μεγάλη ομάδα.

1.2.3 Ιεραρχικές μέθοδοι ομαδοποίησης: Διαιρετικές μέθοδοι

Όπως έχουμε ήδη αναφέρει, στις ιεραρχικές μεθόδους ομαδοποίησης ανήκει και η κατηγορία των διαιρετικών μεθόδων (divisive methods). Οι διαιρετικές μέθοδοι ξεκινούν θεωρώντας το σύνολο των δεδομένων ως μία ομάδα και στη συνέχεια το διαιρούν προοδευτικά σε επιμέρους ομάδες, ούτως ώστε χρησιμοποιώντας κάποιο κριτήριο βελτιστοποίησης να προκύψουν τελικά τόσες ομάδες, όσα τα άτομα που περιλαμβάνονται στα δεδομένα.

Η λογική στην οποία βασίζονται οι διαιρετικοί αλγόριθμοι είναι, να βρίσκουν υποομάδες των ήδη διαμορφωμένων ομάδων που είναι περισσότερο απομακρυσμένες και να τις διαχωρίζουν. Έτσι, σε κάθε βήμα διαμερίζουν μια ομάδα σε δυο άλλες μικρότερες έως ότου φτάσουν στο σημείο όπου όλες οι ομάδες περιέχουν ένα μόνο στοιχείο.

Το κυριότερο μειονέκτημα των διαιρετικών μεθόδων είναι ότι είναι φοβερά χρονοβόρες, διότι δοκιμάζουν διάφορους συνδυασμούς δεδομένων. Για τη διαίρεση n δεδομένων σε 2 ομάδες εξετάζονται $2^{n-1}-1$ περιπτώσεις.

Η χαρακτηριστικότερη των μεθόδων αυτών είναι των Edwards και Cavalli-Sforza (1965). Η διαδικασία που ακολουθεί είναι να επιλέγει σε κάθε βήμα από όλες τις δυνατές διαμερίσεις σε δυο ομάδες εκείνη η οποία ελαχιστοποιεί το άθροισμα των τετραγωνικών αποκλίσεων (total error sum of squares) για τις δυο ομάδες. Η λογική της μεθόδου είναι παρόμοια με αυτήν του αλγορίθμου του Ward που γνωρίσαμε στην ενότητα των συσσωρευτικών μεθόδων.

Πιο συγκεκριμένα, η μέθοδος των Edwards και Cavalli-Sforza (1965) χρησιμοποιεί την τεχνική της ανάλυσης διασποράς, σύμφωνα με την οποία όταν ένα σύνολο σημείων χωρίζεται σε n ομάδες, το άθροισμα τετραγώνων των αποστάσεων όλων των σημείων από το μέσο όρο τους (SST) ισούται με το άθροισμα των αθροισμάτων τετραγώνων των αποστάσεων των σημείων κάθε μίας από τις n ομάδες από το μέσο όρο της ($SS_1 + ... + SS_n$ ή SSW), συν το άθροισμα τετραγώνων των αποστάσεων των μέσων όρων των n ομάδων από το συνολικό μέσο όρο (SSB), δηλαδή

$$SST = SSW + SSB. \qquad (1.9)$$

Επακόλουθο της ανάλυσης διασποράς είναι ότι η καλύτερη διαίρεση ενός συνόλου δεδομένων σε k ομάδες θα είναι δυνατή όταν το μεταξύ των ομάδων άθροισμα τετραγώνων γίνεται μέγιστο. Οι Edwards και Cavalli-Sforza χρησιμοποιούν τη μεγιστοποίηση του μεταξύ των ομάδων αθροίσματος τετραγώνων ως κριτήριο ομαδοποίησης ξεκινώντας από τη διαίρεση του συνόλου των δεδομένων σε 2 ομάδες. Σε κάθε βήμα του αλγορίθμου, κάθε ομάδα που προέκυψε στο προηγούμενο βήμα, υποδιαιρείται σε δύο ομάδες έως ότου να προκύψουν ομάδες με ένα σημείο η κάθε μία. Τα σημεία των δύο ομάδων που μόλις έχουν προκύψει από μία ομάδα του προηγούμενου βήματος, χαρακτηρίζονται από το μέγιστο άθροισμα τετραγώνων μεταξύ των ομάδων ή το ελάχιστο άθροισμα τετραγώνων εντός των ομάδων.

1.2.4 Ένας άλλος διαχωρισμός των Αλγόριθμων Ομαδοποιήσεων

Ο Everitt (1981) αναφέρει την επόμενη ταξινόμηση των αλγορίθμων ομαδοποιήσεων σε πέντε κατηγορίες με βάση τους τύπους ταξινόμησης του Cormack (1971) :

1. Οι **Ιεραρχικοί Αλγόριθμοι** με τους οποίους οι ομάδες σχηματίζονται από όμοια αντικείμενα, κατά τέτοιο τρόπο ώστε η κάθε μία εμπεριέχεται στην αμέσως ευρύτερη (nested) και παρίστανται συνήθως με δενδρόγραμμα το οποίο αρχίζει με τόσες ομάδες όσα είναι και τα αντικείμενα και τελειώνει σε μία ομάδα η οποία περιλαμβάνει όλα τα αντικείμενα.

Όπως αναφέραμε και σε προηγούμενη παράγραφο, οι ιεραρχικές τεχνικές ομαδοποίησης χωρίζονται σε συσσωρευτικές και σε διαιρετικές μεθόδους. Και οι δύο τύποι των ιεραρχικών τεχνικών προσπαθούν να βρουν το βέλτιστο βήμα σε κάθε επίπεδο της διαδοχικής υποδιαίρεσης ή σύνθεσης των δεδομένων.

2. Οι **Αλγόριθμοι Βελτιστοποίησης (Optimization) ή Διαχωριστικοί** με τους οποίους τα δεδομένα χωρίζονται σε τελείως ξεχωριστές ομάδες (των οποίων το πλήθος μπορεί να προκαθοριστεί) με τη βελτιστοποίηση κάποιου μαθηματικού κριτηρίου.

Οι μέθοδοι αυτής της κατηγορίας διαφέρουν μεταξύ τους είτε από τον τρόπο του αρχικού διαχωρισμού των δεδομένων, είτε από το κριτήριο που χρησιμοποιούν. Υπάρχουν διάφοροι τύποι κριτηρίων. Τα τρία πρώτα κριτήρια που αναφέρουμε παρακάτω απορρέουν από την

εξίσωση T = W + B, όπου T είναι ο συνολικός πίνακας διασποράς, W είναι ο πίνακας διασποράς εντός των ομάδων και B είναι ο πίνακας διασποράς μεταξύ των ομάδων.

i) Ελαχιστοποίηση του αθροίσματος των στοιχείων της διαγωνίου του πίνακα W (Minimization of Trace(W))

ii) Ελαχιστοποίηση της ορίζουσας του W (Minimization of the Determinant of W)

iii) Μεγιστοποίηση του αθροίσματος των συντελεστών της διαγωνίου του πίνακα $B^{-1}W$ (Maximization of Trace($B^{-1}W$))

iv) Ελαχιστοποίηση ενός μέτρου πληροφορίας (Information measure). Αυτή η μέθοδος (Wallace & Walton (1968)) ελαχιστοποιεί ένα μέτρο πληροφορίας ως συνάρτηση των δεδομένων και των παραμέτρων ορισμένων κατανομών που λαμβάνονται ως μεταβλητές.

3. Οι **Αλγόριθμοι Πυκνότητας (Density or mode-seeking procedures)** οι οποίοι ψάχνουν να βρουν περιοχές με μεγάλη πυκνότητα συγκεντρώσεων αντικειμένων.

Οι μέθοδοι αυτής της κατηγορίας βασίζονται στη μέθοδο του κοντινότερου γείτονα (single linkage) και προσπαθούν να ξεπεράσουν το φαινόμενο της ένταξης απομονωμένων σημείων σε ομάδες (chaining), γεγονός το οποίο είναι το μεγαλύτερο μειονέκτημα της μεθόδου. Πιο συγκεκριμένα υπάρχουν οι παρακάτω μέθοδοι:

i) Η μέθοδος Tampax των Carmichael & Sneath. Αυτή η μέθοδος (Carmichael (1968), Carmichael & Sneath (1969)) προσπαθεί να μιμηθεί τον ανθρώπινο παρατηρητή στο χώρο των δύο και των τριών διαστάσεων συγκρίνοντας σχετικές αποστάσεις σημείων και ψάχνοντας περιοχές υψηλής πυκνότητας, περιβαλλόμενες από σχετικά άδεια διαστήματα. Η μέθοδος βασίζεται στη μέθοδο του κοντινότερου γείτονα, με τη διαφορά ότι ένα καινούριο σημείο δεν γίνεται δεκτό αν η απόστασή του από μία ομάδα είναι μεγαλύτερη από ένα σημείο που είχε γίνει προηγουμένως δεκτό.

ii) Η μέθοδος των Gitman & Levine (Gitman and Levine's method). Οι Gitman & Levine (1970) περιγράφουν έναν αλγόριθμο σχετικό με την ανίχνευση μονοκόρυφων συγκεχυμένων ομάδων (unimodal fuzzy sets), ο οποίος ξεκινά με τη μέθοδο του κοντινότερου γείτονα και χρησιμοποιεί έναν αριθμό n_i ως βαθμό ένταξης στην ομάδα ενός σημείου x_i. Το n_i είναι ο αριθμός των σημείων ενός συνόλου Γ το οποίο ορίζεται ως εξής:

$$\Gamma = \{x \mid d(x_i, x) \leq T\}$$

όπου Τ είναι η μέγιστη τιμή που επιτρέπεται να λάβει η απόσταση και η οποία καθορίζεται από τον αναλυτή.

iii) Η μέθοδος μέτρησης Υπερκύβων (The cartet count method). Αυτή η μέθοδος, η οποία περιγράφεται από τους Cattel & Coulter (1966), χωρίζει έναν χώρο με πολλές μεταβλητές σε υπερκύβους και μετρά τον αριθμό των σημείων σε κάθε ένα από αυτούς.

iv) Ανάλυση κορυφών (Mode analysis). Η ανάλυση κορυφών (Wishart (1969b)) είναι παράγωγο της μεθόδου του κοντινότερου γείτονα. Η μέθοδος αυτή ψάχνει να βρει φυσικές (natural) υποομάδες, θεωρώντας μια σφαίρα ακτίνας R που περιβάλει κάθε σημείο και μετρώντας τον αριθμό των σημείων μέσα στη σφαίρα. Η ακτίνα της σφαίρας ποικίλλει ανάλογα με τον αριθμό των σημείων.

v) Η μέθοδος των μικτών κατανομών (Method of Mixtures). Η κατανομή πιθανότητας ενός πληθυσμού που αποτελείται από περισσότερες από μία ομάδες μπορεί να είναι η συνισταμένη διάφορων κατανομών, κάθε μία από τις οποίες ορίζει μία ομάδα. Ο Wolfe (1965) χρησιμοποίησε τη μέθοδο μεγίστης πιθανοφάνειας για να υπολογίσει τις παραμέτρους των κατανομών που αντιστοιχούν σε κάθε ομάδα. Κάθε αντικείμενο γίνεται μέλος μίας ομάδας όταν η πιθανότητά του να ανήκει στην ομάδα αυτή είναι μέγιστη.

4. **Αλγόριθμοι Συστάδων (Clumping procedures)**, με τους οποίους οι ομάδες που δημιουργούνται μπορούν να έχουν κοινά σημεία (π.χ. νοήματα λέξεων σε γλώσσες). Ένα τέτοιο είδος συστάδας (clump) και το συμπλήρωμά του μπορούν να θεωρηθούν ως διαφορετικοί τύποι ομάδων.

5. **Άλλοι** οι οποίοι δεν εμπίπτουν στις προηγούμενες κατηγορίες. Μερικοί αναλυτές χρησιμοποιούν είτε Παραγοντική ανάλυση (Factor analysis) είτε Ανάλυση κυρίων συνιστωσών (Principal component analysis) για την ανίχνευση της ύπαρξης ομάδων. Κάποιοι άλλοι (Sokal & Rolf (1962), Gower (1970)) προσπαθούν να δώσουν μία γεωμετρική ερμηνεία, ενώ υπάρχουν και κάποιοι αναλυτές οι οποίοι χρησιμοποιούν γραφικές τεχνικές για να προσδιορίσουν τη δομή δεδομένων με πολλές μεταβλητές.

ΚΕΦΑΛΑΙΟ 2

ΕΠΙΛΟΓΗ ΤΟΥ ΠΛΗΘΟΥΣ ΤΩΝ ΟΜΑΔΩΝ

2.1 Ιστορική αναδρομή

Οι αλγόριθμοι ομαδοποίησης (ιεραρχικοί και μη-ιεραρχικοί) χωρίζουν ένα σύνολο δεδομένων σε ομάδες χωρίς όμως να προσδιορίζουν τον ακριβή αριθμό ομάδων από τις οποίες απαρτίζονται τα δεδομένα. Για το λόγο αυτό υπήρξε η ανάγκη ανάπτυξης κάποιων συγκεκριμένων κριτηρίων και διαδικασιών οι οποίες να βρίσκουν τον βέλτιστο αριθμό των ομάδων και τον καλύτερο συνδυασμό των παρατηρήσεων εντός αυτών των ομάδων.

Η ανάπτυξη λογικών κανόνων που οδηγούν στην εύρεση του βέλτιστου αριθμού ομάδων σε ένα σύνολο δεδομένων έχει απασχολήσει αρκετούς ερευνητές. Πρώτος ο Thorndike (1953) πρότεινε μια γραφική προσέγγιση του προβλήματος σύμφωνα με την οποία γίνεται αρχικά μια απεικόνιση σε ένα άξονα του μέσου όρου των αποστάσεων μέσα στις ομάδες (average within-cluster distance) και σε ένα δεύτερο του αριθμού των ομάδων. Σε κάθε αύξηση του αριθμού k των ομάδων υπάρχει αντίστοιχη μείωση στο μέσο όρο των αποστάσεων μέσα στις ομάδες. Στις περισσότερες περιπτώσεις εμφανίζεται κάποια θέση όπου έχουμε απότομη μείωση του μέσου όρου των αποστάσεων μέσα στις ομάδες και στη συνέχεια «οριζοντιοποίηση» του γραφήματος. Η πρόταση του Thorndike είναι να ληφθεί ως βέλτιστο πλήθος ομάδων, η τιμή που αντιστοιχεί στο παραπάνω σημείο.

Στην προσπάθειά του να εξετάσει την ορθότητα αυτού του συλλογισμού με τη χρήση τεχνητών δεδομένων, ο Thorndike διαπίστωσε ότι δεν αντιστοιχεί πάντα στο βέλτιστο πλήθος ομάδων μια τέτοια συμπεριφορά του γραφήματος. Παρότι η εικασία του Thorndike δεν αποδείχτηκε ορθή για όλες τις περιπτώσεις, έδωσε το έναυσμα και σε άλλους ερευνητές να ασχοληθούν με την εξεύρεση παρόμοιων γραφημάτων που μπορούν να δώσουν χρήσιμες πληροφορίες για το πλήθος των ομάδων που υπάρχουν στα δεδομένα. Θα πρέπει πάντως να

σημειωθεί προκαταβολικά ότι τέτοιου είδους προσεγγίσεις έχουν έντονο το στοιχείο της υποκειμενικό (Κούτρας (2005)).

2.2 Μέθοδοι εύρεσης βέλτιστου πλήθους ομάδων

Οι ιεραρχικές τεχνικές που έχουμε περιγράψει στο προηγούμενο κεφάλαιο ξεκινούν με έναν αριθμό ομάδων ίσο με τον αριθμό (n) των ατόμων (στοιχείων) στα οποία έχουν γίνει οι μετρήσεις και μέσω διαδοχικών συγχωνεύσεων καταλήγουν σε μία ομάδα που περιλαμβάνει όλες τις παρατηρήσεις. Μία ιεραρχική δομή μπορεί να απεικονιστεί μέσω ενός δενδρογράμματος το οποίο αποτελείται από (n-1) επίπεδα, στο καθένα από τα οποία πραγματοποιείται μία συγχώνευση. Οι ιεραρχικοί αλγόριθμοι δεν μπορούν να σταματήσουν σε ένα ενδιάμεσο επίπεδο εκτός αν κάποιο κριτήριο ενεργήσει επάνω τους ώστε το δενδρόγραμμα να μπορέσει να κοπεί σε ένα συγκεκριμένο επίπεδο, δείχνοντας με τον τρόπο αυτό έναν βέλτιστο αριθμό ομάδων που είναι μικρότερος ή ίσος από το n.

Ένας πρακτικός τρόπος εύρεσης του πλήθους των ομάδων είναι να εξετάσουμε το δενδρόγραμμα που προκύπτει από μια ιεραρχική συσσωρευτική μέθοδο και από αυτό να καθορίσουμε το βέλτιστο πλήθος. Πιο συγκεκριμένα, σε εκείνο το σημείο του δενδρογράμματος που παρατηρείται η μεγαλύτερη μεταβολή της ποσότητας που καταγράφεται στον οριζόντιο άξονα (απόσταση ή μέτρο ομοιότητας) μπορούμε να φέρουμε μια παράλληλη γραμμή προς τον κατακόρυφο άξονα και να δούμε σε πόσα σημεία τέμνει το δενδρόγραμμα. Το πλήθος k, για το οποίο παρατηρούμε μεγάλες αποστάσεις συνένωσης σε σχέση με το προηγούμενο (k-1 ομάδες) αποτελεί μια λογική τιμή για το βέλτιστο πλήθος των ομάδων.

Ωστόσο, όπως είδαμε και με τη μέθοδο του Thorndike, τα αποτελέσματα που προκύπτουν από μία τέτοια διαδικασία υπόκεινται στην κρίση του κάθε ερευνητή. Για το λόγο αυτό υπήρξε η ανάγκη ανάπτυξης κάποιων συγκεκριμένων κριτηρίων τα οποία να μπορούν να βρίσκουν το βέλτιστο αριθμό των ομάδων και τον καλύτερο συνδυασμό των παρατηρήσεων εντός των ομάδων.

Κατά συνέπεια, μία εναλλακτική μέθοδος για τον προσδιορισμό του βέλτιστου αριθμού των ομάδων είναι η εφαρμογή κάποιου κριτηρίου διακοπής του αλγορίθμου (Stopping Rule). Η βελτιστοποίηση ενός τέτοιου κριτηρίου μπορεί να τερματίσει έναν αλγόριθμο ομαδοποίησης σε ένα βήμα του, ούτως ώστε ο αριθμός των ομάδων που έχει βρεθεί σε αυτό το βήμα να είναι ο καλύτερος.

Τα τελευταία χρόνια έχουν γίνει διάφορες προσπάθειες να λυθεί το πρόβλημα εμπειρικά (Mojena (1977)), αλλά λόγω της έλλειψης επαρκούς θεωρίας στην ανάλυση κατά συστάδες, η σύγκριση των διαφόρων μεθόδων γίνεται με χρήση τεχνικών προσομοίωσης. Οι Milligan και Cooper (1985) περιγράφουν 30 κριτήρια διακοπής τα οποία προήλθαν μέσω προσομοιώσεων Monte Carlo. Παράλληλα γίνεται σύγκριση των 30 αυτών κριτηρίων διακοπής με βάση την απόδοσή τους τόσο στην εύρεση του βέλτιστου πλήθους των ομάδων που περιέχει ένα σύνολο δεδομένων, όσο και στην ορθή σύσταση των ομάδων με την βοήθεια εξωτερικών κριτηρίων που θα εξετάσουμε σε επόμενη παράγραφο.

Τα κριτήρια διακοπής που εξετάζονται μπορεί να έχουν μεγαλύτερη ή μικρότερη επιτυχία στην εύρεση του αριθμού ομάδων. Ωστόσο, η συγκεκριμένη εργασία των Milligan και Cooper θεωρείται ορόσημο στον χώρο των κριτηρίων διακοπής.

2.3 Κριτήρια Διακοπής (Stopping Rules)

Στην παράγραφο αυτή θα εξετάσουμε μία σειρά κριτηρίων διακοπής. Τα κριτήρια αυτά μπορούν να ομαδοποιηθούν σε διάφορες κατηγορίες. Υπάρχουν κριτήρια τα οποία βασίζονται σε μεθόδους ανάλυσης διακύμανσης (ANOVA) (analysis of variance methods), κάποια τα οποία στηρίζονται σε μεθόδους αποστάσεων (distance methods), ενώ διάφορα άλλα κριτήρια διακοπής στηρίζονται σε μεθόδους μεγίστης πιθανοφάνειας (maximum likelihood methods). Ακόμα έχουμε κριτήρια τα οποία χρησιμοποιούν διάφορες κατηγορίες δεικτών που βασίζονται είτε σε μεθόδους t-tests (t-tests methods), είτε σε μεθόδους της θεωρίας γραφημάτων (graph theoretical methods), ενώ υπάρχουν και κάποια κριτήρια που έχουν ως βάση δείκτες συμφωνίας (indices of agreement methods). Τέλος, μία ακόμα κατηγορία αφορά κριτήρια που στηρίζονται σε μη-παραμετρικές μεθόδους (non-parametric methods).

Στη συνέχεια, από κάθε κατηγορία, θα περιγράψουμε τα σημαντικότερα κριτήρια τα οποία συναντάμε πιο συχνά στην πράξη, ενώ θα κάνουμε και μία πιο σύντομη αναφορά στα υπόλοιπα κριτήρια που υπάρχουν στη βιβλιογραφία.

2.3.1 Μέθοδοι ανάλυσης διακύμανσης

Ας υποθέσουμε ότι έχουμε πολυμεταβλητά συνεχή δεδομένα. Έχοντας κατατάξει τις παρατηρήσεις σε ομάδες, ουσιαστικά έχουμε δεδομένα που μοιάζουν με αυτά στην πολυμεταβλητή ανάλυση διακύμανσης (MANOVA). Συνεπώς με την ίδια ακριβώς τεχνική μπορούμε να διαμερίσουμε το συνολικό άθροισμα τετραγώνων σε δύο μέρη. Το ένα μέρος θα δείχνει τις αποκλίσεις μέσα στις ομάδες και το άλλο θα δείχνει τις αποκλίσεις ανάμεσα στις ομάδες.

Συγκεκριμένα, για συνεχή δεδομένα, μπορεί κανείς να αναλύσει τις συνολικές τετραγωνικές αποκλίσεις σε δύο μέρη, όπως φαίνεται στον πίνακα 2.1. Με $\bar{x}_j$ συμβολίζουμε τον μέσο (κέντρο βάρους) της j ομάδας (j = 1,2,…,k) και με $\bar{x}$ συμβολίζουμε τον μέσο (κέντρο βάρους) όλων των παρατηρήσεων.

Μεταξύ των ομάδων (between)	$B = \sum_{j=1}^{k} n_j (\bar{x}_j - \bar{x})(\bar{x}_j - \bar{x})'$ (2.1)
Μέσα στις ομάδες (within)	$W = \sum_{j=1}^{k} \sum_{i=1}^{n_j} (x_{ij} - \bar{x}_j)(x_{ij} - \bar{x}_j)'$ (2.2)
Συνολική	$T = \sum_{j=1}^{k} \sum_{i=1}^{n_j} (x_{ij} - \bar{x}_{\bullet\bullet})(x_{ij} - \bar{x}_{\bullet\bullet})'$ (2.3)

Πίνακας 2.1

Αθροίσματα τετραγωνικών αποκλίσεων

Για μια πετυχημένη ομαδοποίηση θέλουμε ο πίνακας B να είναι όσο γίνεται μεγαλύτερος (με βάση κάποιο κριτήριο διάταξης πινάκων) ενώ ο πίνακας W όσο γίνεται μικρότερος, αφού αντιστοιχεί σε όμοια στοιχεία μέσα στην ομάδα.

Συνεπώς, για να διαλέξουμε την καλύτερη ομαδοποίηση των δεδομένων μας, μπορούμε να φτιάξουμε κριτήρια βασισμένα στις παραπάνω ποσότητες. Αυτά τα κριτήρια μπορεί να βασίζονται στο ίχνος των πινάκων ή στην ορίζουσά τους. Όταν όμως δεν λαμβάνεται καθόλου υπόψη ο αριθμός των ομάδων στις οποίες οδηγεί η ελαχιστοποίηση του μεγέθους του πίνακα, τέτοια κριτήρια οδηγούν συνήθως σε λύσεις με μεγάλο αριθμό ομάδων.

α) Το κριτήριο των Calinski and Harabasz's (1974)

Το κριτήριο των Calinski-Harabasz θεωρείται ένα από τα πιο αξιόπιστα κριτήρια διακοπής. Ένα πλεονέκτημα του συγκεκριμένου κριτηρίου είναι ότι λαμβάνει υπόψη του και τον αριθμό των ομάδων από τις οποίες αποτελούνται τα δεδομένα.

Έστω k ο αριθμός των ομάδων στις οποίες διαχωρίζονται τα δεδομένα σε κάποιο βήμα. Για να βρούμε τη βέλτιστη τιμή του k, επιλέγουμε την ομαδοποίηση για την οποία προκύπτει η μεγαλύτερη τιμή του λόγου

$$c = \frac{\text{trace(B)}}{k-1} \Big/ \frac{\text{trace(W)}}{n-k} \tag{2.4}$$

Ο λόγος για τον οποίο διαλέγουμε τη μεγαλύτερη τιμή είναι γιατί αν φτιάξουμε ομάδες με παρατηρήσεις όμοιες μέσα στην ομάδα, αλλά με μεγάλες διαφορές από ομάδα σε ομάδα, τότε περιμένουμε πως τα στοιχεία του πίνακα B θα έχουν μεγάλες τιμές και τα στοιχεία του πίνακα W θα έχουν μικρές και επομένως το κριτήριο θα πάρει μεγάλη τιμή.

Ένα μειονέκτημα του αλγορίθμου των Calinski-Harabasz είναι ότι είναι φοβερά χρονοβόρος απέναντι σε άλλους αλγόριθμους και μπορεί να χρησιμοποιείται μόνο για μικρά σύνολα δεδομένων, καθόσον είναι πολύ ακριβής. Αξίζει πάντως να αναφέρουμε πως στην μελέτη των Milligan & Cooper (1985), ο αλγόριθμος των Calinski-Harabasz είχε τη μεγαλύτερη επιτυχία.

β) Το Κριτήριο Trace W

Ένα κριτήριο το οποίο έχει προταθεί από τους A.Edwards & L.Cavalli-Sforza(1965) είναι το κριτήριο Trace W. Το κριτήριο αυτό είναι ένας από τους πιο δημοφιλείς δείκτες που έχουν προταθεί για χρήση στο πλαίσιο της ομαδοποίησης. Βάσει του συγκεκριμένου κριτηρίου, για να βρούμε το βέλτιστο αριθμό των ομάδων, επιλέγουμε την ομαδοποίηση εκείνη για την οποία προκύπτει η ελάχιστη τιμή του ίχνους του πίνακα W.

Αξίζει να αναφέρουμε πως στην εργασία τους οι Milligan & Cooper (1985), για να προσδιορίσουν τον αριθμό των ομάδων στα δεδομένα, χρησιμοποίησαν τα μέγιστα σκορ διαφορών, δεδομένου ότι το κριτήριο αυξάνει μονότονα με λύσεις που περιέχουν λιγότερες

ομάδες. Ως σκορ διαφοράς ορίζεται η τιμή που προκύπτει από τη σύγκριση των τιμών ενός κριτηρίου από το ένα επίπεδο της ιεραρχίας στο επόμενο.

Για κάθε διαμέριση των n παρατηρήσεων σε k ομάδες θα ισχύει μεταξύ των πινάκων T, W και B η ισότητα.

$$T = W + B$$

Εφόσον ο πίνακας T είναι σταθερός για όλες τις διαμερίσεις, η ελαχιστοποίηση του ίχνους του W είναι ισοδύναμη με τη μεγιστοποίηση του ίχνους του B, αφού σύμφωνα με ιδιότητα του ίχνους πινάκων έχουμε

$$\text{Trace}(T) = \text{Trace}(W) + \text{Trace}(B)$$

γ) Τα Κριτήρια Trace $W^{-1}B$ και $|T|/|W|$

Οι Friedman και Rubin (1967) πρότειναν δύο εναλλακτικά κριτήρια ομαδοποίησης τα οποία παραμένουν αμετάβλητα κάτω από κάθε γραμμικό μετασχηματισμό του αρχικού συνόλου δεδομένων. Το ένα κριτήριο είναι το κριτήριο Trace[$W^{-1}B$] και άλλο είναι το κριτήριο |T|/|W|. Όσον αφορά το κριτήριο Trace[$W^{-1}B$], ο δείκτης προτάθηκε από τους Friedman και Rubin ως βάση για μία μη ιεραρχική μέθοδο ομαδοποίησης.

Τα δυο μεγέθη που εξετάζουμε μπορούν να εκφραστούν συναρτήσει των ιδιοτιμών του πίνακα $W^{-1}B$ με την παρακάτω μορφή :

$$\frac{|T|}{|W|} = \prod_{i=1}^{t}(1+\lambda_i) \qquad \text{και} \qquad \text{Trace}W^{-1}B = \sum_{i=1}^{t}\lambda_i$$

όπου $t \le p$ και p η διάσταση των δεδομένων. Η μεγιστοποίηση των κριτηρίων θα μας οδηγήσει στην εύρεση του βέλτιστου αριθμού των ομάδων.

Αξίζει να αναφέρουμε ότι ανάμεσα σε διάφορα μέτρα που εξετάστηκαν από τους Scott και Symons (1971), προτάθηκε η χρήση του δείκτη

$$n\log(|T|/|W|)$$

όπου το n είναι ο αριθμός των στοιχείων στο σύνολο των δεδομένων. Εκτός του πολλαπλασιασμού του από μία σταθερά, είναι ο ίδιος δείκτης όπως αυτός εξετάστηκε από τον

Arnold (1979). Βρέθηκε ότι αυτή η μορφή του δείκτη δίνει πολύ καλύτερα αποτελέσματα από τον δείκτη $|T|/|W|$.

δ) Ο Λόγος μεταβολής του αθροίσματος τετραγώνων μέσα στις ομάδες (ΛΜΑΤ)

Το κριτήριο αυτό κατατάσσεται στην κατηγορία των κριτηρίων του αθροίσματος τετραγώνων. Ως εκ τούτου, έχει σχέση με τη μέθοδο του Ward. Όπως έχουμε αναφέρει, η μέθοδος του Ward προσπαθεί να ελαχιστοποιήσει τη διαφορά του συνολικού εντός των ομάδων αθροίσματος τετραγώνων από k + 1 σε k ομάδες, το οποίο παριστάνεται ως:

$$\Delta_{k,k+1} = SS_k - SS_{k+1}$$

όπου

$$SS_k = \sum_{j=1}^{k}\sum_{i=1}^{n_j}\left(x_{ij} - \bar{x}_{.j}\right)^2$$

είναι το συνολικό εντός των ομάδων άθροισμα τετραγώνων για k ομάδες. Το κριτήριο ΛΜΑΤ προσπαθεί να συγκρίνει τη διαφορά $\Delta_{k+1,k+2}$ με τη διαφορά $\Delta_{k,k+1}$, μεταξύ δύο διαδοχικών βημάτων ενός συσσωρευτικού αλγόριθμου. Όταν η διαφορά $\Delta_{k,k+1}$ είναι μεγαλύτερη της $\Delta_{k+1,k+2}$, σημαίνει ότι η ομαδοποίηση από k + 1 σε k ομάδες, είναι ασθενέστερη της ομαδοποίησης από k + 2 σε k + 1 ομάδες. Αυτό μπορεί να εκφραστεί με τη μορφή του λόγου

$$\frac{\Delta_{k,k+1}}{\Delta_{k+1,k+2}} \qquad (2.5)$$

Εάν υπάρχει ένα k, για το οποίο ο λόγος (2.5) γίνεται μέγιστος, στο βήμα αυτό πρέπει να τερματισθεί ο συσσωρευτικός αλγόριθμος, διότι η ομαδοποίηση από k + 1 σε k ομάδες, είναι η ασθενέστερη, επειδή εμφανίζει τη μεγαλύτερη απώλεια πληροφορίας (κατά τη μέθοδο του Ward) μεταξύ δύο διαδοχικών βημάτων του αλγορίθμου. Κάτι τέτοιο μάλιστα δηλώνει κάποια διατάραξη στην πορεία των διαδοχικών ομαδοποιήσεων. Επομένως το κριτήριο ΛΜΑΤ, τερματίζει έναν αλγόριθμο ομαδοποίησης, όταν τα δεδομένα έχουν χωρισθεί σε k + 1 ομάδες.

Το κριτήριο ΛΜΑΤ έχει ιδιαίτερη επιτυχία, όταν εφαρμόζεται στη μέθοδο του Ward, επειδή και οι δύο βασίζονται στο ίδιο θεωρητικό υπόβαθρο.

ε) Ο δείκτης των Duda και Hart

Οι Duda και Hart (1973) πρότειναν ένα κριτήριο λόγου

$$SSE(2)/SSE(1) < C_r$$

όπου SSE(2) είναι το άθροισμα των τετραγωνικών σφαλμάτων μέσα στην ομάδα (sum of squared errors within cluster) όταν τα δεδομένα είναι χωρισμένα σε δύο ομάδες και το SSE(1) δίνει το αντίστοιχο άθροισμα όταν εμφανίζεται μόνο μία ομάδα. Η υπόθεση ότι υπάρχει μία ομάδα (και όχι δύο) απορρίπτεται όταν ο λόγος είναι μικρότερος από μία προκαθορισμένη κρίσιμη τιμή (C_r). Η κρίσιμη αυτή τιμή είναι μία συνάρτηση του μεγέθους του δείγματος και του αριθμού των διαστάσεων.

Θα πρέπει να σημειωθεί πως η συγκεκριμένη μέθοδος των Duda και Hart εφαρμόζεται μόνο σε εκείνο το υποσύνολο των δεδομένων που παίρνουν μέρος στη συγχώνευση των ομάδων.

στ) Το κριτήριο Marriot

Το κριτήριο του Marriot (1971) υποδεικνύει να ληφθεί ως βέλτιστο πλήθος k των ομάδων εκείνο το οποίο ελαχιστοποιεί την ποσότητα $k^2|W|$, όπου W είναι ο πίνακας της διασποράς μέσα στις ομάδες.

Σε κατανομές με ένα μέγιστο (μονοκόρυφες, unimodal) η ελαχιστοποίηση της παραπάνω ποσότητας πιθανόν να μας οδηγεί στην λύση k = 1, ενώ σύμφωνα με την παρατήρηση του Everitt (1979) όταν έχουμε ομοιόμορφη κατανομή η τιμή του κριτηρίου θα παραμένει συνεχώς σταθερή. Η μέγιστη διαφορά μεταξύ των διαδοχικών επιπέδων προσδιορίζει το καλύτερο επίπεδο διαχωρισμού.

ζ) Ο δείκτης του Beale

Ο Beale (1969) πρότεινε τη χρήση ενός λόγου F για να ελέγξει την υπόθεση της ύπαρξης k_2 ομάδων έναντι k_1 ομάδων στα δεδομένα ($k_2 > k_1$). Ο λόγος F συγκρίνει την αύξηση στη

μέση τετραγωνική απόκλιση από τα κέντρα των ομάδων καθώς μεταφερόμαστε από k_2 σε k_1 ομάδες, έναντι της μέσης τετραγωνικής αποκλίσεως όταν είναι παρούσες στα δεδομένα k_2 ομάδες.

Ο λόγος F που πρότεινε ο Beale δίνεται από τον τύπο:

$$F(k_1, k_2) = \frac{R_{k_1} - R_{k_2}}{R_{k_2}} \Big/ \left[\left\{ \frac{N - k_1}{N - k_2} \right\} \left(\frac{k_2}{k_1} \right)^{2/p} - 1 \right]$$

όπου $R_k = (N-k)S^2{}_k$ και $S_k{}^2$ είναι η μέση τετραγωνική απόκλιση από τα κέντρα των ομάδων στο δείγμα. Η παραπάνω στατιστική συνάρτηση συγκρίνεται με το άνω α σημείο της κατανομής F με $p(k_2-k_1)$ και $p(N-k_2)$ βαθμούς ελευθερίας, όπου p είναι η διάσταση των δεδομένων.

Στην πραγματικότητα, ο Beale έδωσε έναν F-έλεγχο με τον οποίο ελέγχεται εάν μια υποδιαίρεση σε k_2 ομάδες είναι στατιστικά καλύτερη από μια υποδιαίρεση σε ένα μικρότερο αριθμό ομάδων k_1. Η παραπάνω μέθοδος εξάγει χρήσιμα συμπεράσματα στις περιπτώσεις που οι ομάδες είναι αρκετά διαχωρισμένες μεταξύ τους και έχουν σφαιρική μορφή.

η) Ο δείκτης των Davies and Bouldin

Οι Davies και Bouldin (1979) πρότειναν ένα γενικό πλαίσιο κριτηρίων που χρησιμεύουν στη διαδικασία διαχωρισμού των ομάδων. Ο συνολικός δείκτης ορίζεται ως ο μέσος των δεικτών που υπολογίστηκαν από την κάθε μία ομάδα χωριστά. Ως δείκτης κάθε μεμονωμένης ομάδας θεωρείται η μέγιστη ανά δύο σύγκριση μεταξύ της παραπάνω ομάδας και των άλλων επιμέρους ομάδων της λύσης. Κάθε ανά δύο σύγκριση υπολογίζεται ως ο λόγος που έχει αριθμητή το άθροισμα των αποστάσεων των εντός της ομάδας στοιχείων από τα στοιχεία της άλλης ομάδας και παρονομαστή ένα μέτρο διαχωρισμού μεταξύ των ομάδων. Η απόσταση μεταξύ των κέντρων των ομάδων χρησιμοποιείται ως μέτρο διαχωρισμού μεταξύ των ομάδων.

Η ελάχιστη τιμή του συνολικού δείκτη μεταξύ των επιπέδων της ιεραρχίας χρησιμοποιείται για να δείξει το βέλτιστο αριθμό των ομάδων στα δεδομένα.

θ) Το κριτήριο των Frey and Van Groenewoud

Οι Frey και Van Groenewoud (1972) πρότειναν τη χρήση ενός δείκτη ο οποίος προκύπτει ως ο λόγος των σκορ των διαφορών μεταξύ δύο διαδοχικών επιπέδων στην ιεραρχία. Ο αριθμητής είναι η διαφορά ανάμεσα στους μέσους των αποστάσεων μεταξύ των ομάδων (between cluster distances average) για καθένα από τα δύο επίπεδα. Ο παρονομαστής δηλώνει τη διαφορά ανάμεσα στους μέσους των αποστάσεων εντός των ομάδων (within cluster distances average) από τα δύο επίπεδα. Πιο συγκεκριμένα ο δείκτης των Frey και Van Groenewoud (1972) δίνεται από τον τύπο:

$$K = \frac{\bar{d}_{u_{j+1}} - \bar{d}_{u_j}}{\bar{d}_{s_{j+1}} - \bar{d}_{s_j}}$$

όπου το σύμβολο d_u δίνει τις αποστάσεις μεταξύ των ομάδων και το σύμβολο d_s δίνει τις αποστάσεις εντός των ομάδων. Οι Milligan & Cooper (1985) πρότειναν, να θεωρήσουμε ότι έχουμε το βέλτιστο αριθμό των ομάδων όταν ο λόγος K παίρνει την τιμή 1. Στην πράξη, ο λόγος K παίρνει τιμές μεγαλύτερες ή μικρότερες του 1.

ι) Ο δείκτης $\bar{c}/\sqrt{k}$

Οι Ratkowsky και Lance (1978) εισήγαγαν ένα κριτήριο για να προσδιορίσουν τον βέλτιστο αριθμό των ομάδων βασιζόμενοι στο λόγο $\bar{c}/\sqrt{k}$, όπου το $\bar{c}$ ισούται με τον μέσο όρο των λόγων $(SSB/SST)^2$ οι οποίοι παρατηρούνται από την κάθε διάσταση στα δεδομένα. Η μεγιστοποίηση του $\bar{c}/\sqrt{k}$ θεωρείται ότι φανερώνει το βέλτιστο αριθμό των ομάδων.

ια) Ο δείκτης log(SSB/SSW)

Ο Hartigan (1975) πρότεινε τον συγκεκριμένο δείκτη για την εύρεση του σωστού αριθμού των ομάδων βασισμένο στο άθροισμα τετραγώνων. Οι τιμές SSB και SSW είναι τα αθροίσματα των τετραγωνικών αποστάσεων μεταξύ και εντός των ομάδων αντίστοιχα.

2.3.2 Μέθοδοι απόστασης

Όπως έχουμε αναφέρει, στην ανάλυση κατά συστάδες, ένα σύνολο παρατηρήσεων μπορεί να παρασταθεί ως ένα σύνολο σημείων σε ένα p-διάστατο χώρο. Για το λόγο αυτό, τα μέτρα απόστασης φαίνεται ότι μπορούν ως ένα καλό σημείο να περιγράψουν τις ομοιότητες καθώς και τη δομή των δεδομένων μεταξύ και εντός των ομάδων. Κατά συνέπεια, θα μπορούσαν να χρησιμοποιηθούν και για τον προσδιορισμό του αριθμού των ομάδων από τις οποίες αποτελούνται τα δεδομένα. Αυτό είναι το βασικό πλεονέκτημα των μεθόδων που βασίζονται στις αποστάσεις και ο λόγος για τον οποίο αρκετοί συγγραφείς χρησιμοποίησαν μέτρα απόστασης ή συναρτήσεις αυτών.

α) Το κριτήριο του Glasbey

Το κριτήριο του Glasbey είναι ένας συνδυασμός των αλγορίθμων Single Linkage και Complete Linkage. Και οι δύο μέθοδοι βασίζονται στην εσωτερική συνοχή των ομάδων και στην εξωτερική απομόνωσή τους (Cormack, (1971)). Η μεν Single Linkage μεγιστοποιεί την ελάχιστη απόσταση μεταξύ των ομάδων. Η δε Complete Linkage ελαχιστοποιεί τη μέγιστη απόσταση εντός των ομάδων.

Η μέθοδος του Glasbey εφαρμόζει τον αλγόριθμο Single Linkage υπολογίζοντας την ελάχιστη απόσταση μεταξύ των ομάδων σε κάθε βήμα του. Επίσης χρησιμοποιώντας τον Complete Linkage αλγόριθμο υπολογίζει αφενός τη διάμετρο της ομάδας που μόλις έχει σχηματισθεί, δηλαδή την απόσταση των πιο απομακρυσμένων σημείων της και αφετέρου υπολογίζει τη μέγιστη διάμετρο μέσα σε όλες τις ομάδες που έχουν δημιουργηθεί μέχρι αυτή τη στιγμή, δηλαδή υπολογίζει τη μέγιστη απόσταση μεταξύ των ομάδων αλλά και τη μέγιστη απόσταση των σημείων εντός των ομάδων. Όταν η μέγιστη απόσταση εντός των ομάδων είναι ίση ή μεγαλύτερη από την ελάχιστη απόσταση μεταξύ των ομάδων, ο αλγόριθμος σταματά και οι ομάδες που έχουν διαχωριστεί είναι αυτές που βρέθηκαν στο προηγούμενο βήμα του αλγορίθμου.

β) Δείκτης C-Index

Ο δείκτης C-Index προτάθηκε από τους Hubert και Levin (1976) και υπολογίζεται από τη σχέση:

$$[d_w - \min(d_w)]/[\max(d_w) - \min(d_w)]$$

όπου d_w είναι το άθροισμα των αποστάσεων μέσα στην ομάδα και $\min(d_w)$ και $\max(d_w)$ είναι αντίστοιχα η ελάχιστη και η μέγιστη απόσταση μέσα στην ομάδα. Η ελάχιστη τιμή μεταξύ των ιεραρχικών επιπέδων χρησιμοποιείται για να αναδειχθεί ο βέλτιστος αριθμός των ομάδων.

γ) Cubic Clustering Criterion

Το κριτήριο του Cubic Clustering προτάθηκε από τους Ray (1982) και Sarle (1983) και είναι ο στατιστικός έλεγχος που παρέχεται στο πακέτο SAS. Ο δείκτης έχει τον ακόλουθο τύπο:

$$\frac{\log[(1 - E(R^2))/(1 - R^2)]}{\sqrt{np/2}/(0.001 + E(R^2))^{1.2}}$$

όπου R^2 είναι η αναλογία της διακύμανσης (proportion of variance) που αποδίδεται σε κάθε ομάδα. Η αναμενόμενη τιμή του R^2 έχει καθοριστεί κάτω από την υπόθεση ότι τα δεδομένα προέρχονται από ένα δείγμα ομοιόμορφης κατανομής βασισμένα σε έναν υπερκύβο. Το p είναι η εκτίμηση της διάστασης της διακύμανσης μεταξύ των ομάδων (an estimate of the dimensionality of the between cluster variation).

Οι σταθεροί όροι έχουν επιλεγεί βάσει εκτεταμένων προσομοιωμένων αποτελεσμάτων. Η μέγιστη τιμή ανάμεσα στα επίπεδα της ιεραρχίας χρησιμοποιείται για να καθοριστεί ο βέλτιστος αριθμός ομάδων στα δεδομένα.

δ) Το κριτήριο Stepsize

Το κριτήριο Stepsize χρονολογείται πριν από τη δουλειά των Sokal και Sneath (1963). Αυτό το μάλλον απλό κριτήριο συνίσταται στην εξέταση των αποστάσεων ανάμεσα στις συγχωνευμένες τιμές μεταξύ των επιπέδων της ιεραρχίας. Μία μεγάλη απόσταση δείχνει ότι τα δεδομένα έχουν υπερομαδοποιηθεί στην τελευταία συγχώνευση. Για το λόγο αυτό, μία μέγιστη διαφορά θεωρείται ότι φανερώνει τον βέλτιστο αριθμό των ομάδων στα δεδομένα.

ε) Το κριτήριο των Ball and Hall

Οι Ball και Hall (1965) πρότειναν ότι η μέση απόσταση των στοιχείων από τα αντίστοιχα κέντρα των ομάδων μπορεί να χρησιμεύσει ως μέτρο για την εύρεση του αριθμού των ομάδων στα δεδομένα.

στ) Το κριτήριο McClain and Rao

Οι McClain και Rao (1975) ασχολήθηκαν με ένα κριτήριο το οποίο αποτελείται από έναν λόγο δύο όρων. Ο πρώτος όρος είναι ο μέσος των αποστάσεων εντός της ομάδας διαιρεμένος με τον αριθμό των αποστάσεων εντός της ομάδας. Ο παρονομαστής είναι ο μέσος των αποστάσεων μεταξύ των ομάδων διαιρεμένος με τον αριθμό των αποστάσεων μεταξύ των ομάδων. Η ελάχιστη τιμή του δείκτη βρέθηκε ότι δίνει το καλύτερο αποτέλεσμα.

ζ) Το κριτήριο Mountford

Ο Mountford (1970) ανέπτυξε μία διαδικασία ελέγχου για να καθορίσει αν δύο ομάδες πρέπει να συγχωνευτούν. Ο αριθμητής του στατιστικού τύπου του ελέγχου υπολογίστηκε ως το άθροισμα των αποστάσεων εντός της ομάδας μείον τη μέση απόσταση μεταξύ των ομάδων. Ο παρονομαστής είναι ένα μέτρο της διακύμανσης εντός της ομάδας. Σε ένα δοθέν επίπεδο, ο δείκτης βασίζεται μόνο σε αυτά τα σημεία που πραγματικά παίρνουν μέρος στη συγχώνευση. Αν η συγχώνευση της ομάδας περιέχει λιγότερο από τέσσερα σημεία, ο δείκτης παίρνει την τιμή 0. Το καλύτερο αποτέλεσμα για τη εύρεση του βέλτιστου αριθμού των ομάδων στα δεδομένα παρατηρείται όταν παίρνουμε τη μέγιστη διαφορά μεταξύ των επιπέδων. Ωστόσο, το κριτήριο του Mountford έχει την τάση να βρίσκει λύσεις με μεγάλο πλήθος ομάδων.

2.3.3 Μέθοδοι μεγίστης πιθανοφάνειας

Ένας άλλος έλεγχος που χρησιμοποιείται είναι αυτός του λόγου πιθανοφάνειας (likelihood ratio test) (Wolfe (1970)). Για να ελέγξουμε την υπόθεση ότι τα δεδομένα μας αποτελούνται από k_1 ομάδες έναντι k_2 ομάδων χρησιμοποιούμε την στατιστική συνάρτηση -2logλ, όπου λ είναι ο λόγος των πιθανοφανειών, δηλ. $\lambda = L_{k_1} / L_{k_2}$. Κάτω από ειδικές συνθήκες αυτή η

στατιστική συνάρτηση προσεγγίζεται από τη χ^2- κατανομή με βαθμούς ελευθερίας ίσους με την διαφορά των παραμέτρων στην υπόθεση.

Ο Wolfe (1971) πρότεινε τη χρήση του εναλλακτικού τύπου:

$$\lambda = -\frac{2}{n}\left(n - k - \frac{k_2}{2}\right)\log\left(\frac{L_{k_1}}{L_{k_2}}\right)$$

και προχώρησε στην αξιολόγηση της νέας μεθόδου μέσω προσομοίωσης Monte Carlo.

Η μέθοδος του Wolfe βασίζεται στην υπόθεση της πολυμεταβλητής κανονικότητας. Ο Everitt (1981) διεξήγαγε μία λεπτομερή Monte Carlo ανάλυση για τη διαδικασία που πρότεινε ο Wolfe και βρήκε ότι ο τύπος του Wolfe φαίνεται να αποδίδει μόνο για περιπτώσεις όπου το μέγεθος του δείγματος είναι περίπου 10 φορές μεγαλύτερο από τον αριθμό των διαστάσεων. Ακόμα ο Binder (1978) έδειξε ότι η στατιστική συνάρτηση του ελέγχου δεν κατανέμεται ασυμπτωτικά ως κατανομή χ^2 και επομένως η θεωρητική προσέγγιση της προτεινόμενης διαδικασίας δε φαίνεται να καλύπτεται ικανοποιητικά.

Ο Day (1969) πρότεινε μία εναλλακτική μέθοδο στην περίπτωση που τα δεδομένα μπορεί να θεωρηθούν ότι προέρχονται από μία μίξη δύο πολυμεταβλητών κανονικών κατανομών (multivariate normal distributions). Τότε, μπορούμε να προχωρήσουμε σε εκτίμηση μέσω της μεθόδου μεγίστης πιθανοφάνειας, της γενικευμένης απόστασης μεταξύ δύο ομάδων. Στην πράξη συνήθως χρησιμοποιείται η μέγιστη αύξηση στην τιμή της απόστασης μεταξύ των ιεραρχικών επιπέδων για να εντοπισθεί το σημείο της υπερ-ομαδοποίησης (overclustering).

2.3.4 Μέθοδοι που βασίζονται σε t-tests

Μία άλλη κατηγορία κριτηρίων διακοπής βασίζονται σε μεθόδους t-tests. Στην κατηγορία αυτή υπάγονται τρία κριτήρια: το κριτήριο $|\log(p)|$, το κριτήριο του Sneath και το κριτήριο του Mojena.

Οι Gnanadesikan, Kettenring και Landwher (1977) πρότειναν έναν κανόνα διακοπής βασισμένο στο $|\log(p)|$, όπου p είναι το p–value που προκύπτει κατά τον έλεγχο T^2 του Hotelling.

Ο Sneath (1977) δημιούργησε μία μέθοδο για να ελέγξει την διακριτότητα των ομάδων βασιζόμενος σε ένα κριτήριο λόγου t_w. Ο αριθμητής του λόγου t_w είναι η απόσταση μεταξύ των κέντρων των δύο ομάδων κάτω από τον παράγοντα της συγχώνευσης. Ο παρονομαστής του δείκτη είναι ένα μέτρο της διασποράς ή της επικάλυψης των δύο ομάδων. Επομένως, ο έλεγχος δεν βασίζεται σε όλα τα δεδομένα στην ανάλυση κατά συστάδες, αλλά μόνο σε εκείνα τα στοιχεία που εμπλέκονται στην συγχώνευση. Ο λόγος αυτός συγκρίνεται με μία κρίσιμη τιμή που παρατηρείται από μία μη κεντρική t κατανομή. Η υπόθεση της μίας ομάδας απορρίπτεται αν το t_w υπερβαίνει την αυτή τιμή.

Τέλος, ένα ακόμα κριτήριο διακοπής που βασίζεται σε t-tests προτάθηκε από τον Mojena (1977). Το κριτήριο διακοπής αυτό είναι όμως ελάχιστα γνωστό και έχει ερευνηθεί λίγο. Ο κανόνας αντιστοιχεί σε ένα μονόπλευρο διάστημα εμπιστοσύνης βασισμένο στις τιμές που έχουν συγχωνευτεί σε κάθε επίπεδο της ιεραρχίας. Υπολογιστικά, παίρνουμε τη μέση συγχωνευμένη τιμή και της προσθέτουμε ένα μέτρο του τυπικού σφάλματος των συγχωνευμένων τιμών από ολόκληρη την ιεραρχία πολλαπλασιασμένο με μία κρίσιμη τιμή.

2.3.5 Γραφικές μέθοδοι

Προτού προχωρήσουμε στην περιγραφή των γραφικών μεθόδων, θα αναφερθούμε σε κάποια βασικά στοιχεία της θεωρίας γραφημάτων, τα οποία είναι απαραίτητα για την κατηγορία των αλγορίθμων που βασίζονται σε αυτή.

Έστω ένα σύνολο N, αποτελούμενο από n σημεία x_1, x_2, …, x_n συνδεδεμένα μεταξύ τους με συνδέσμους (δεσμούς, πλευρές) (x_1, x_2), …, (x_{n-1}, x_n), οι οποίοι είναι στοιχεία ενός συνόλου A. Το σύνολο G, το οποίο περιγράφεται από τα σύνολα N και A, ονομάζεται Γράφημα και συμβολίζεται με:

$$G = [N ; A]$$

Ένα συνδεδεμένο γράφημα (connected graph) είναι ένα γράφημα του οποίου όλοι οι κόμβοι είναι συνδεδεμένοι. Ένα πλήρες γράφημα (complete graph) n σημείων έχει

$$|A| = \binom{n}{2}$$

δεσμούς, οι οποίοι συνδέουν ανά δύο τα n σημεία. Το |A| δηλώνει το πλήθος των στοιχείων του συνόλου A. Σε ένα πλήρες γράφημα αντιστοιχεί ένας $(n \times n)$ πίνακας D, του οποίου οι στήλες και γραμμές αντιστοιχούν στα n σημεία $x_1, x_2, \ldots, x_n$ και τα στοιχεία του αντιστοιχούν στις αποστάσεις $d(x_i,x_j)$ των σημείων $x_1, x_2, \ldots, x_n$ ανά δύο.

Ένα δέντρο είναι ένα συνδεδεμένο γράφημα G = [N ; A], το οποίο δεν περιέχει κύκλους. Δηλαδή, εάν υπάρχει μία διαδοχή n σημείων, τα οποία συνδέονται μεταξύ τους με τους δεσμούς $(x_1, x_2), \ldots, (x_{n-1}, x_n)$ δεν μπορεί να υπάρχει δεσμός (x_i, x_j) που να συνδέει δύο μη διαδοχικά σημεία. Η διαδοχή των δεσμών αυτών λέγεται δρόμος (path). Ένα δέντρο αποτελείται από μία ρίζα, κλαδιά που είναι οι δεσμοί μεταξύ των σημείων και φύλλα τα οποία συνδέονται μόνο από τη μία πλευρά τους με άλλα σημεία. Ένα δέντρο G, n σημείων έχει |A| = n – 1 δεσμούς.

Ένα δέντρο ελαχίστων αποστάσεων (Δ.Ε.Α.) (Minimum Spanning Tree), είναι ένα δέντρο, το οποίο έχει προκύψει από ένα πλήρες γράφημα, όταν ληφθούν υπόψη οι ελάχιστες αποστάσεις που συνδέουν τα n σημεία. Εάν οι $\binom{n}{2}$ σύνδεσμοι ενός πλήρους γραφήματος διαταχθούν κατά αύξουσα τάξη και από αυτούς επιλεχθούν n – 1 σύνδεσμοι που δεν σχηματίζουν κύκλους, τότε αυτοί οι σύνδεσμοι απαρτίζουν το Δ.Ε.Α. Αντίστοιχα, το Δ.Μ.Α. μπορεί να σχηματιστεί εάν οι πλευρές ενός πλήρους γραφήματος διαταχθούν κατά φθίνουσα τάξη και από αυτές επιλεχθούν οι πρώτες n – 1 πλευρές που δεν σχηματίζουν κύκλους.

α) Το κριτήριο GRAPH

Ο Αλγόριθμος GRAPH χρησιμοποιεί ένα συνδυασμό του Δέντρου ελαχίστων αποστάσεων (Δ.Ε.Α.) και του Δέντρου μεγίστων αποστάσεων (Δ.Μ.Α.).

Ο αλγόριθμος GRAPH είναι ουσιαστικά ο Single Linkage αλγόριθμος, ο οποίος απεικονίζεται άμεσα από το Δ.Ε.Α. και έχει ως κριτήριο διακοπής την ύπαρξη συνδέσμου του Δ.Μ.Α., ανάμεσα σε δύο σημεία, που κάθε ένα από αυτά ανήκει σε μία από τις δύο προς συγχώνευση ομάδες, σε κάποιο βήμα του αλγορίθμου.

Ο Single Linkage αλγόριθμος, για τη συγχώνευση παρατηρήσεων n ομάδων, λαμβάνει υπόψη τις ελάχιστες αποστάσεις μεταξύ τους, ακολουθώντας τις κορυφές (που απεικονίζουν τα αντικείμενα) και τους συνδέσμους (που απεικονίζουν τις αποστάσεις μεταξύ των

αντικειμένων) του Δ.Ε.Α., ξεκινώντας από τις πλησιέστερες παρατηρήσεις. Η όλη δομή των ομαδοποιήσεων παριστάνεται με ένα δενδρόγραμμα.

Ο αλγόριθμος GRAPH φτιάχνει τα δέντρα των ελαχίστων αποστάσεων (Δ.Ε.Α.) και μεγίστων αποστάσεων (Δ.Μ.Α.) και αρχίζει να εξετάζει τους συνδέσμους (links) του Δ.Ε.Α. κατά αύξουσα τάξη. Τη στιγμή που εξετάζεται ένας σύνδεσμος γίνεται συνένωση των κορυφών τις οποίες συνδέει και τα σημεία που αντιστοιχούν στις κορυφές είτε σχηματίζουν μια ομάδα από μόνα τους ή προσαρτώνται σε ομάδες που ήδη υπάρχουν, είτε εάν ανήκουν σε δύο διαφορετικές ομάδες αυτές ενώνονται και σχηματίζουν μια νέα ομάδα. Σε κάθε βήμα του αλγορίθμου γίνεται έλεγχος μήπως υπάρχει σύνδεσμος του Δ.Μ.Α. που συνδέει δύο σημεία των υπό συγχώνευση ομάδων. Η ύπαρξη ενός τέτοιου σημείου σημαίνει ότι αν οι δύο ομάδες συγχωνευτούν θα υπάρχουν δύο σημεία μέσα στην καινούργια ομάδα τα οποία θα απέχουν πολύ μεταξύ τους συγχρόνως με δύο σημεία που είναι πολύ κοντά μεταξύ τους. Μια τέτοια ομαδοποίηση δεν θα πρέπει να γίνει εφόσον απαραίτητη προϋπόθεση για ομαδοποίηση σημείων είναι η εσωτερική συνοχή των ομάδων. Εάν δεν έχουν εξαντληθεί οι σύνδεσμοι του Δ.Ε.Α., η ύπαρξη ενός συνδέσμου του Δ.Ε.Α. και ενός του Δ.Μ.Α. είναι απαγορευτική ως προς τη δημιουργία μίας ομάδας από τις προϋπάρχουσες. Εάν οι σύνδεσμοι του Δ.Ε.Α. έχουν εξαντληθεί τότε έχει τελειώσει η ομαδοποίηση των σημείων. Το πλήθος των ομάδων, που θεωρείται σωστό από τον αλγόριθμο είναι αυτό που έχει βρεθεί στο προηγούμενο βήμα της περατώσεως του αλγορίθμου.

Ο αλγόριθμος GRAPH μπορεί να περιγραφεί με τα ακόλουθα υπολογιστικά βήματα, με την προϋπόθεση, ότι υπάρχουν τουλάχιστον 3 αντικείμενα για ομαδοποίηση.

1. Δημιουργούμε τα Δ.Ε.Α. και Δ.Μ.Α.
2. Συγχωνεύουμε τα σημεία, τα οποία είναι πλησιέστερα μεταξύ τους.
3. Εξετάζουμε το σύνδεσμο του Δ.Ε.Α., με την πιο μικρή τιμή από τους υπόλοιπους συνδέσμους. Εάν υπάρχουν μία ή περισσότερες ομάδες, που συνδέονται με συνδέσμους του Δ.Μ.Α., η προηγούμενη ομαδοποίηση θεωρείται ως λύση.
4. Διαφορετικά, το βήμα 3 επαναλαμβάνεται μέχρι να βρεθούν δύο ομάδες, οι οποίες εκτός του συνδέσμου του Δ.Ε.Α., συνδέονται και με ένα σύνδεσμο του Δ.Μ.Α.

Για παράδειγμα, έστω ότι έχουμε τον επόμενο πίνακα των αποστάσεων:

$$D = \begin{array}{c} \\ 2 \\ 3 \\ 4 \\ 5 \end{array} \begin{array}{c} \begin{array}{cccc} 1 & 2 & 3 & 4 \end{array} \\ \begin{bmatrix} 2 & & & \\ 6 & 5 & & \\ 10 & 9 & 4 & \\ 9 & 8 & 5 & 3 \end{bmatrix} \end{array}$$

Τα Δ.Ε.Α. και Δ.Μ.Α είναι τα εξής:

Κορυφές	**Αποστάσεις**
1 – 2	2
4 – 5	3
3 – 4	4
2 – 3	5

Πίνακας 2.2

Δέντρο ελαχίστων αποστάσεων

Κορυφές	**Αποστάσεις**
1 – 4	10
1 – 5	9
2 – 4	9
1 – 3	6

Πίνακας 2.3

Δέντρο μεγίστων αποστάσεων

Έτσι θα πάρουμε τα παρακάτω βήματα του αλγορίθμου:

Βήμα 1 : Τα αντικείμενα 1, 2 ενώνονται. Οι ομάδες είναι: (1,2), (3), (4), (5)

Βήμα 2 : Τα αντικείμενα 4, 5 ενώνονται. Οι ομάδες είναι: (1,2), (3), (4,5)

Βήμα 3 : Οι ομάδες (3) και (4,5) ενώνονται. Οι ομάδες είναι: (1,2), (3,4,5)

Βήμα 4 : Οι ομάδες που πρόκειται να ενωθούν είναι: (1,2), (3,4,5). Αυτές όμως οι ομάδες συνδέονται με τους συνδέσμους 1-4, 1-5, 2-4 και 1-3 του Δ.Μ.Α. Επομένως η συγχώνευση των ομάδων (1,2) και (3,4,5) δεν μπορεί να γίνει.

Επομένως ο αλγόριθμος GRAPH σταματά στο βήμα 3 και το πλήθος των ομάδων που βρέθηκε είναι δύο.

β) Το κριτήριο των Lingoes and Cooper

Οι Lingoes και Cooper (1971) εισήγαγαν ένα κριτήριο βασισμένο στα επιχειρήματα της θεωρίας GRAPH και άμεσα παρέχει ένα p-value για σύγκριση με μία καθορισμένη τιμή σφάλματος τύπου I. Η ομαδοποίηση συνεχίζεται μέχρι η τιμή του p-value να μην είναι πλέον σημαντική.

2.3.6 Δείκτες συμφωνίας

α) Μέθοδος των Goodman and Kruskal's (1954)

Σε αυτόν τον κανόνα γίνονται κάποιες συγκρίσεις μεταξύ όλων των αποστάσεων μέσα στις ομάδες (within-cluster) και των αποστάσεων μεταξύ των ομάδων (between-cluster).

Μια σύγκριση θεωρείται 'αρμονική' αν μια απόσταση μέσα στην ομάδα είναι μικρότερη από μια απόσταση μεταξύ των ομάδων

$$G1(k) \equiv (S_+ - S_-)/(S_+ + S_-)$$

όπου με S_+ συμβολίζουμε τον αριθμό των 'αρμονικών' συγκρίσεων, ενώ με S_- τον αριθμό των μη 'αρμονικών' συγκρίσεων. Δηλαδή έχουμε δυο σύνολα αποτελούμενα από αποστάσεις, το μεν ένα από τις αποστάσεις μέσα στις ομάδες και το δε άλλο από αποστάσεις μεταξύ των ομάδων. Στη συνέχεια συγκρίνουμε κάθε μέλος του συνόλου των αποστάσεων μέσα στις ομάδες με κάθε ένα μέλος από το άλλο σύνολο. Ανάλογα με το αποτέλεσμα της σύγκρισης αυξάνει ένας από τους δυο δείκτες S_+, S_-. Αυτή η διαδικασία επαναλαμβάνεται σε όλα τα

στάδια της ιεραρχικής διαδικασίας (δηλ. για κάθε k) και η τιμή του k η οποία μεγιστοποιεί το δείκτη G1(k) θεωρείται ότι καθορίζει το πλήθος των ομάδων σε ένα σύνολο δεδομένων.

β) Ο δείκτης G(+)

Ο συγκεκριμένος δείκτης εισήχθη από τον Rohlf (1974) και εξετάστηκε από τον Milligan (1981a). Υπολογίζεται από τη σχέση:

$$[2S_{-}]/[n_d(n_d-1)]$$

όπου το S_{-} καθορίζεται όπως στο κριτήριο των Goodman and Kruskal's και n_d είναι ο αριθμός των αποστάσεων μέσα στην ομάδα.

Οι ελάχιστες τιμές χρησιμοποιούνται για να καθοριστεί ο αριθμός των ομάδων στα δεδομένα.

2.3.7 Μη-παραμετρικές μέθοδοι

Στην κατηγορία αυτή υπάγονται δύο κριτήρια: το κριτήριο του Bock (1977) και το κριτήριο που βασίζεται στον απαραμετρικό δείκτη συσχέτισης τ (Tau).

Ο Bock (1977) πρότεινε μία διαδικασία ελέγχου βασιζόμενη σε μία προσέγγιση η οποία εμπλέκει τόσο την εκτίμηση της πυκνότητας όσο και τη μη παραμετρική στατιστική συνάρτηση U.

Ο δείκτης τ ελέγχθηκε από τον Milligan (1981a). Υπολογιστικά, η τυπική μη παραμετρική συσχέτιση τ υπολογίζεται μεταξύ των εισόδων σε δύο μετρήσεις. Ο πρώτος πίνακας περιλαμβάνει τις αποστάσεις μεταξύ των στοιχείων και ο δεύτερος 0 / 1 πίνακας δείχνει αν κάθε ζεύγος σημείων είναι ή δεν είναι εντός της ίδιας ομάδας. Η μέγιστη τιμή στην ιεραρχική ακολουθία θεωρείται ότι μας δείχνει το σωστό αριθμό των ομάδων.

2.4 Επίλογος

Όπως είδαμε, τα κριτήρια διακοπής μπορούν να ταξινομηθούν βάσει των μεθόδων που χρησιμοποιούν. Έτσι η πλειοψηφία των κριτηρίων που εξετάσαμε βασίζεται σε μεθόδους

ανάλυσης διακύμανσης (MANOVA), ενώ ένας σημαντικός αριθμός κριτηρίων είναι βασισμένα σε αποστάσεις.

Κλείνοντας αυτήν την παράγραφο, αξίζει να σημειώσουμε ότι κάποιοι από τους δείκτες που αναφέραμε είχαν προταθεί αρχικά ως μέτρα για χρήση στη μελέτη της κατηγορικής ομαδοποίησης. Παρόλα αυτά, οι περισσότεροι δείκτες συγκρίνουν ένα μέρος των στοιχείων των δεδομένων χρησιμοποιώντας έναν πίνακα ομοιοτήτων μεταξύ των στοιχείων. Επομένως, οι δείκτες αυτοί μπορούν να χρησιμοποιηθούν ως μέτρα εσωτερικών κριτηρίων σε ένα πιο ευρύ πλαίσιο ομαδοποίησης.

ΚΕΦΑΛΑΙΟ 3

ΣΥΓΚΡΙΣΗ ΤΩΝ ΜΕΘΟΔΩΝ ΕΝΤΟΠΙΣΜΟΥ ΠΛΗΘΟΥΣ ΟΜΑΔΩΝ

Για τον προσδιορισμό του βέλτιστου αριθμού των ομάδων από τις οποίες απαρτίζονται τα δεδομένα θα πρέπει να εφαρμοσθεί κάποιο κριτήριο διακοπής του αλγορίθμου (Stopping Rule). Η βελτιστοποίηση αυτού του κριτηρίου είδαμε πως μπορεί να τερματίσει ένα αλγόριθμο ομαδοποίησης σε ένα βήμα του έτσι ώστε ο αριθμός των ομάδων που έχει βρεθεί σε αυτό το βήμα να είναι ο καλύτερος.

Τα κριτήρια που περιγράψαμε μπορεί να έχουν μεγαλύτερη ή μικρότερη επιτυχία στην εύρεση του αριθμού των ομάδων ανάλογα με τη φύση των δεδομένων στα οποία εφαρμόζονται. Η επιτυχία ενός κριτηρίου τερματισμού ενός αλγορίθμου ομαδοποιήσεως μπορεί να χαρακτηριστεί από 3 παράγοντες: (α) από την εύρεση του σωστού αριθμού ομάδων, (β) από τη σωστή τοποθέτηση των δεδομένων σε σωστές ομάδες και (γ) από την ευστάθειά του, ως προς την εύρεση παρόμοιων αποτελεσμάτων, που προκύπτουν από παρεμφερή μεταξύ τους δεδομένα.

Με πραγματικά δεδομένα, είναι δύσκολη η αξιολόγηση ενός κριτηρίου τερματισμού αλγορίθμων ομαδοποιήσεως. Συνήθως για να αξιολογηθεί ένα τέτοιο κριτήριο, χρησιμοποιούνται μέθοδοι προσομοίωσης, με δεδομένα, που παράγονται από γνωστές κατανομές, όπου ένας ερευνητής μπορεί να καθορίσει εκ των προτέρων το πλήθος των ομάδων και να έχει πλήρη γνώση της τοποθέτησης των δεδομένων μέσα σε αυτές. Έτσι όταν

εφαρμοσθεί το κριτήριο, τόσο το πλήθος των ομάδων, όσο και η τοποθέτηση των δεδομένων μέσα σε αυτές τις ομάδες, που προκύπτουν από την εφαρμογή του, μπορούν να συγκριθούν με τα αντίστοιχα χαρακτηριστικά των δεδομένων από τα οποία προέκυψαν και να αποφασισθεί, αν το κριτήριο είναι περισσότερο ή λιγότερο επιτυχές.

Η εφαρμογή ενός κριτηρίου τερματισμού αλγορίθμων ομαδοποιήσεως μπορεί να οδηγήσει σε μια σωστή, ή σε μία λανθασμένη απόφαση. Υπάρχουν δύο τύποι σφαλμάτων αποφάσεων. Ο πρώτος τύπος σφάλματος είναι όταν ένα κριτήριο αποφασίζει για την ύπαρξη περισσότερων ομάδων από όσες στην πραγματικότητα υπάρχουν και ο δεύτερος τύπος συμβαίνει όταν ένα κριτήριο αποφασίζει για την ύπαρξη λιγότερων ομάδων. Ο δεύτερος τύπος σφάλματος θεωρείται σοβαρότερος, γιατί με τη συγχώνευση ομάδων, που κανονικά θα έπρεπε να είναι χωριστές, χάνεται πληροφορία (Milligan, 1985). Επίσης, ακόμα και αν το κριτήριο βρίσκει το σωστό αριθμό των ομάδων, μπορεί η τοποθέτηση των δεδομένων σε αυτές τις ομάδες να είναι λανθασμένη.

Για να κριθεί λοιπόν αν ένα κριτήριο τερματισμού αλγορίθμων ομαδοποιήσης είναι επιτυχές, πρέπει να υπάρχει αφενός ένας τρόπος σύγκρισής του με άλλα κριτήρια ως προς την επιτυχία της εύρεσης του αριθμού των ομάδων και αφετέρου να υπάρχει ένας δείκτης επιτυχίας της τοποθέτησης των δεδομένων στη σχετική ομάδα. Το πρώτο γίνεται με την επανάληψη του αλγορίθμου σε έναν ικανοποιητικό αριθμό φορών και το δεύτερο με την παραγωγή τυχαίων αριθμών από διάφορες κατανομές με δεδομένα κέντρα και δεδομένη ακτίνα συγκεντρώσεως σημείων (truncation) γύρω από αυτά, ανάλογα με το αν θέλουμε οι ομάδες να είναι περισσότερο ή λιγότερες απομονωμένες μεταξύ τους. Μετά τη δημιουργία των δεδομένων και την ενεργοποίηση του αλγορίθμου ομαδοποίησης σε συνδυασμό με το κριτήριο τερματισμού του αλγορίθμου, μπορεί να εφαρμοστεί κάποιος δείκτης που σκοπό έχει τη σύγκριση των ομαδοποιήσεων. Τέτοιοι δείκτες ονομάζονται εξωτερικοί δείκτες (external criteria). Οι επικρατέστεροι στη σύγχρονη βιβλιογραφία είναι ο δείκτης Rand και ο Corrected Rand.

Η ευστάθεια της επιτυχίας ή μη ενός κριτηρίου είναι πολύ σημαντική, διότι μπορεί να δώσει κάποια εγγύηση για τη χρήση του κριτηρίου σε πραγματικά δεδομένα. Ένα μέτρο αξιολόγησής του μπορεί να εκφραστεί με τα ποσοστά επιτυχίας του κριτηρίου, που προκύπτουν από την επανάληψη της παραγωγής δεδομένων, από την ίδια κατανομή.

3.1 Εξωτερικά κριτήρια (External criteria)

Ένα σημαντικό πρόβλημα που προκύπτει με τη χρήση των αλγορίθμων ομαδοποίησης και των κριτηρίων διακοπής είναι η δυσκολία να αξιολογηθεί το αποτέλεσμα του διαχωρισμού των δεδομένων που συντελέστηκε.

Όπως είδαμε, ένα κριτήριο διακοπής, εκτός από τον προσδιορισμό του αριθμού των ομάδων, θα πρέπει να πετυχαίνει και τη σωστή τοποθέτηση των δεδομένων στις ομάδες αυτές. Επειδή τα κριτήρια διακοπής χρησιμοποιούν πληροφορία που προκύπτει εξολοκλήρου μέσα από την διαδικασία ομαδοποίησης, τα κριτήρια αυτά ονομάζονται και εσωτερικά κριτήρια.

Ένας τρόπος αξιολόγησης της ομαδοποίησης που έχει πραγματοποιηθεί στα δεδομένα γίνεται με τη χρήση των εξωτερικών κριτηρίων. Στην ουσία, οι δείκτες των εξωτερικών κριτηρίων χρησιμοποιούν πληροφορία η οποία προέρχεται εκτός της διαδικασίας της ομαδοποίησης. Η πληροφορία αυτή είναι 'εξωτερική' γιατί αναφέρεται στην πραγματική δομή των ομάδων, η οποία είναι εκ των προτέρων γνωστή.

Τα εξωτερικά κριτήρια συγκρίνουν τις ομοιότητες μεταξύ της πραγματικής δομής των ομάδων και της δομής των ομάδων που έχει συντελεστεί μέσω μίας διαδικασίας ομαδοποίησης. Με άλλα λόγια, οι δείκτες των εξωτερικών κριτηρίων χρησιμοποιούνται για την αξιολόγηση της αποκατάστασης της πραγματικής δομής των ομάδων.

Έστω λοιπόν ότι εφαρμόζουμε έναν αλγόριθμο σε ένα σύνολο δεδομένων με γνωστή ομαδοποίηση. Μετά την εφαρμογή του αλγορίθμου υπάρχουν οι εξής τέσσερις περιπτώσεις οι οποίες εμφανίζονται στον πίνακα 3.1. Τα τέσσερα κελιά του πίνακα υποδηλώνουν αν τα ζεύγη των παρατηρήσεων ομαδοποιήθηκαν σωστά μαζί. Για παράδειγμα, το κελί a δείχνει τη συχνότητα του αριθμού των ζευγών στοιχείων τα οποία ομαδοποιήθηκαν σωστά μαζί μέσω της διαδικασίας της ομαδοποίησης. Το κελί b δίνει το πλήθος των ζευγαριών που προκύπτουν από τη διαδικασία της ομαδοποίησης κατά την οποία δύο στοιχεία τοποθετούνται στην ίδια ομάδα, ενώ στην πραγματικότητα προέρχονται από διαφορετικές ομάδες. Το c περιγράφει το ακριβώς αντίθετο, δηλαδή δίνει τη συχνότητα του πλήθους των ζευγών όπου ο αλγόριθμος τοποθέτησε ένα ζευγάρι στοιχείων σε διαφορετικές ομάδες, ενώ τα στοιχεία ανήκουν στην πραγματικότητα στην ίδια ομάδα. Τέλος, το d είναι το πλήθος των ζευγών όπου η διαδικασία τοποθέτησε τα στοιχεία σε διαφορετικές ομάδες, γεγονός που αντιπροσωπεύει και την ορθή λύση. Επομένως, τα κελιά a και d αντιπροσωπεύουν τη συχνότητα εμφάνισης της ορθής

ομαδοποίησης κατά ζεύγη, ενώ τα κελιά b και c αντιπροσωπεύουν τη λανθασμένη κατά ζεύγη ομαδοποίηση. Το πλήθος όλων των δυνατών συνδυασμών των ζευγών θα ισούται με n(n-1)/2, όπου n είναι ο αριθμός των στοιχείων που παίρνουν μέρος στην ομαδοποίηση. Επομένως θα έχουμε: α+b+c+d = n(n-1)/2

ΛΥΣΗ ΑΛΓΟΡΙΘΜΟΥ	**ΟΡΘΗ ΛΥΣΗ**	
	Ζευγάρι στην ίδια ομάδα	Ζευγάρι όχι στην ίδια ομάδα
Ζευγάρι στην ίδια ομάδα	**a**	**b**
Ζευγάρι όχι στην ίδια ομάδα	**c**	**d**

Πίνακας 3.1

Οι περιπτώσεις που προκύπτουν μετά την εφαρμογή του αλγορίθμου

Ο πίνακας 3.2 που ακολουθεί παρέχει τους υπολογιστικούς τύπους για τις ποσότητες που παρουσιάζονται στον πίνακα 3.1. Με N_{ij} καθορίζουμε τον αριθμό των σημείων που βρίσκονται βάσει του αλγορίθμου στην ομάδα i και τα οποία είναι επίσης στη ομάδα j της ορθής λύσης. Ακόμα, τα $N_{i.}$, $N_{.j}$ και $N_{..}$ αντιπροσωπεύουν τα περιθώρια (marginal) και ολικά αθροίσματα αντίστοιχα.

ΛΥΣΗ ΑΛΓΟΡΙΘΜΟΥ	ΟΡΘΗ ΛΥΣΗ	
	Ζευγάρι στην ίδια ομάδα	Ζευγάρι όχι στην ίδια ομάδα
Ζευγάρι στην ίδια ομάδα	$\sum\sum N_{ij}^2 / 2 - N_{..} / 2$	$\sum N^2{}_{i.} / 2 - \sum\sum N_{ij}^2 / 2$
Ζευγάρι όχι στην ίδια ομάδα	$\sum N^2{}_{.j} / 2 - \sum\sum N_{ij}^2 / 2$	$\sum\sum N_{ij}^2 / 2 + N^2{}_{..} / 2 - \sum N^2{}_{i.} / 2 - \sum N^2{}_{.j} / 2$

Πίνακας 3.2

Οι υπολογιστικοί τύποι του πίνακα 3.1

Για παράδειγμα, ας θεωρήσουμε ένα σύνολο οχτώ ατόμων x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8 το οποίο γνωρίζουμε ότι χωρίζεται σε δύο ομάδες. Η πρώτη ομάδα αποτελείται από τα άτομα x_1, x_2, x_3, x_4 και x_5, ενώ η δεύτερη ομάδα αποτελείται από τα άτομα x_6, x_7 και x_8. Επομένως έχουμε τις επόμενες δύο ομάδες

(1,2,3,4,5) και (6,7,8).

Μετά την εφαρμογή ενός αλγορίθμου ομαδοποίησης τα δεδομένα χωρίστηκαν σε δύο ομάδες. Ωστόσο, η δομή των ομάδων που συντελέστηκε παρουσιάζει κάποιες διαφορές σε σύγκριση με την πραγματική δομή των ομάδων. Συγκεκριμένα, μέσω της διαδικασίας της ομαδοποίησης, η πρώτη ομάδα που δημιουργήθηκε αποτελείται από τα άτομα x_1, x_2, x_3 και x_6, ενώ η δεύτερη ομάδα αποτελείται από τα άτομα x_4, x_5, x_7, x_8. Άρα έχουμε τις επόμενες δύο ομάδες

(1,2,3,6) και (4,5,7,8).

Επειδή το σύνολο το δεδομένων μας αποτελείται από 8 άτομα, το πλήθος όλων των δυνατών συνδυασμών των ζευγών θα ισούται με $n(n-1)/2 = 28$. Από την εφαρμογή του αλγορίθμου ομαδοποίησης προκύπτουν τα εξής:

- a = 5. Τα ζεύγη των ατόμων τα οποία ομαδοποιήθηκαν σωστά μαζί είναι τα: (1,2), (1,3), (2,3), (4,5), (7,8)
- b = 7. Συγκεκριμένα, τα ζεύγη των ατόμων όπου τα άτομα κατά τη διαδικασία της ομαδοποίησης τοποθετήθηκαν στην ίδια ομάδα, ενώ στην πραγματικότητα προέρχονται από διαφορετικές ομάδες είναι τα: (1,6), (2,6), (3,6), (4,7), (4,8), (5,7), (5,8)
- c = 8. Τα ζεύγη όπου ο αλγόριθμος τοποθέτησε ένα ζευγάρι ατόμων σε διαφορετικές ομάδες, ενώ τα άτομα ανήκουν στην πραγματικότητα στην ίδια ομάδα είναι τα: (1,4), (1,5), (2,4), (2,5), (3,4), (3,5), (6,7), (6,8)
- d = 8. Τα ζεύγη όπου ορθά η διαδικασία τοποθέτησε τα στοιχεία σε διαφορετικές ομάδες είναι τα: (4,6), (5,6), (1,7), (1,8), (2,7), (2,8), (3,7), (3,8)

Έτσι προκύπτει ο ακόλουθος πίνακας:

ΛΥΣΗ ΑΛΓΟΡΙΘΜΟΥ	**ΟΡΘΗ ΛΥΣΗ**

	Ζευγάρι στην ίδια ομάδα	Ζευγάρι όχι στην ίδια ομάδα
Ζευγάρι στην ίδια ομάδα	**5**	**7**
Ζευγάρι όχι στην ίδια ομάδα	**8**	**8**

Πίνακας 3.3

Τα αποτελέσματα του παραδείγματος

Χρησιμοποιώντας τους τύπους του πίνακα 3.2, για i = 1,2 και j = 1,2, θα καταλήξουμε στα ίδια αποτελέσματα με αυτά του πίνακα 3.3. Όπως αναφέραμε, το N_{ij} είναι ο αριθμός των σημείων που βρίσκονται βάσει του αλγορίθμου στην ομάδα i και τα οποία είναι επίσης στη ομάδα j της ορθής λύσης. Επομένως: $N_{11} = 3$, $N_{12} = 1$, $N_{21} = 2$, $N_{22} = 2$. Ακόμα $N_{..} = 8$, ενώ περιθώρια αθροίσματα θα ισούνται με: $N_{1.} = 4$, $N_{2.} = 4$, $N_{.1} = 5$, $N_{.2} = 3$.

Στη συνέχεια παρατίθενται τέσσερα γνωστά εξωτερικά κριτήρια. Για τον προσδιορισμό των συγκεκριμένων κριτηρίων χρησιμοποιήθηκαν οι τιμές του πίνακα 3.1:

1.	Rand	$\frac{a+d}{a+b+c+d}$
2.	Corrected Rand	$\frac{a+d-n_c}{a+b+c+d-n_c}$
3.	Fowlkes – Mallows	$\frac{a}{[(a+b)(a+c)]^{1/2}}$
4.	Jaccard	$\frac{a}{a+b+c}$

Πίνακας 3.4

Εξωτερικά κριτήρια

Ο όρος n_c που συναντάμε στο κριτήριο Corrected Rand καλείται παράγοντας διόρθωσης και βασίζεται σε περίπλοκους υπολογιστικούς τύπους.

Παρατηρώντας τον πίνακα 3.4 κάποιος μπορεί εύκολα να διαπιστώσει ότι ο δείκτης του Rand είναι όμοιος με τον simple matching coefficient που είδαμε στο κεφάλαιο 1. Από την άλλη, ο δείκτης του Jaccard και ο δείκτης των Fowlkes – Mallows αγνοούν στους τύπους τους την συχνότητα d, τόσο στον αριθμητή όσο και στον παρανομαστή.

Ο δείκτης του Rand είναι ένας από τους πρώτους δείκτες που προτάθηκε και χρησιμοποιήθηκε εκτεταμένα σε Monte Carlo αναλύσεις. Οι υπόλοιποι τρεις δείκτες προτάθηκαν πιο πρόσφατα σε μία προσπάθεια να ξεπεράσουν κάποιες ανεπιθύμητες ιδιότητες του δείκτη Rand. Για παράδειγμα, ο δείκτης του Rand προσεγγίζει το ανώτερο όριό του, που είναι η τιμή 1, καθώς ο αριθμός των ομάδων αυξάνεται απότομα.

Όταν η ομαδοποίηση από τον αλγόριθμο είναι ίδια ακριβώς με την ορθή ομαδοποίηση τότε οι τιμές των b και c είναι ίσες με το μηδέν και όλα τα εξωτερικά κριτήρια που εξετάζουμε παίρνουν την τιμή 1. Τα κριτήρια Rand, Jaccard και Fowlkes-Mallows μπορούν θεωρητικά να πάρουν ως ελάχιστη τιμή το μηδέν. Ωστόσο κάτι τέτοιο σπανίως συμβαίνει σε πραγματικά δεδομένα. Η ελάχιστη τιμή του δείκτη Corrected Rand εξαρτάται από τον ακριβή διαχωρισμό των δεδομένων σε ομάδες.

Με βάση τον πίνακα 3.2, οι εξωτερικοί δείκτες υπολογίζονται με τους ακόλουθους τύπους:

1.	Rand	$\frac{\sum\sum N_{ij}^2 + N_{..}(N_{..}-1)/2 - \sum N_{i.}^2/2 - \sum N_{.j}^2/2}{N_{..}(N_{..}-1)/2}$	(A1)
2.	Corrected Rand	$\frac{\sum\sum N_{ij}^2 - \left(\sum\sum N_{i.}^2 N_{.j}^2\right)/N_{..}^2}{\sum N_{i.}^2/2 + \sum N_{.j}^2/2 - \left(\sum\sum N_{i.}^2 N_{.j}^2\right)/N_{..}^2}$	(A2)
3.	Fowlkes – Mallows	$\frac{\sum\sum N_{ij}^2 - N_{..}}{\left(\sum N_{i.}^2 - N_{..}\right)\left(\sum N_{.j}^2 - N_{..}\right)}$	(A3)
4.	Jaccard	$\frac{\sum\sum N_{ij}^2 - N_{..}}{\sum N_{i.}^2 + \sum N_{.j}^2 - N_{..} - \sum\sum N_{ij}^2}$	(A4)

Πίνακας 3.5

Οι τύποι των εξωτερικών κριτηρίων με βάση τον πίνακα 3.2

Χρησιμοποιώντας τον πίνακα 3.5 για να καθορίσουμε τις σχέσεις (A1)-(A4), ίσως να είναι χρήσιμη κάποια αλγεβρική απλοποίηση. Στον δείκτη Corrected Rand, ο διορθωτικός παράγοντας n_c δίνεται από τη σχέση:

$$n_c = \sum\sum N_{i.}^2 N_{.j}^2 / N_{..}^2 + N_{..}\left(N_{..} - 1\right)/2 - \sum N_{i.}^2 / 2 - \sum N_{.j}^2 / 2$$

Επιστρέφοντας στο παράδειγμα που εξετάζουμε, θα υπολογίσουμε τον δείκτη του Rand και τον δείκτη του Jaccard χρησιμοποιώντας τους τύπους του πίνακα 3.4. Βάσει τον αποτελεσμάτων του πίνακα 3.3, ο δείκτης του Rand ισούται με 0.464 ενώ ο δείκτης του Jaccard ισούται με 0.25. Επομένως, καταλήγουμε στο συμπέρασμα ότι ο αλγόριθμος ομαδοποίησης που εφαρμόστηκε δεν τοποθέτησε σωστά τα δεδομένα στις ομάδες.

3.2 Σύγκριση των κριτηρίων διακοπής με τη χρήση τεχνητών δεδομένων

3.2.1 Παραγωγή των δεδομένων του ελέγχου

Για την μελέτη των κριτηρίων διακοπής χρησιμοποιούνται συνήθως δεδομένα τα οποία προέρχονται από γνωστές κατανομές και η δομή τους είναι εκ των προτέρων γνωστή. Με αυτόν τον τρόπο καθίσταται δυνατή η σύγκριση των αποτελεσμάτων που προκύπτουν με τη χρήση των κριτηρίων διακοπής και ο ερευνητής είναι σε θέση να αποφασίσει ποια από αυτά τα κριτήρια έχουν βρει το σωστό αριθμό των ομάδων ή πόσο απέχουν από αυτόν, καθώς και να ελέγξει το κατά πόσο καταφέρνουν τα κριτήρια να αναγνωρίσουν την πραγματική δομή των δεδομένων.

Οι Milligan και Cooper (1985) προτείνουν έναν αλγόριθμο για την παραγωγή συνόλων τεχνητών δεδομένων. Ο αλγόριθμος αυτός έχει γραφτεί σε γλώσσα προγραμματισμού Fortran και χρησιμεύει στην παραγωγή δεδομένων τα οποία προκύπτουν μέσω προσομοίωσης με υπολογισμούς Monte Carlo. Τα τεχνητά σύνολα δεδομένων περιέχουν 2, 3, 4 ή 5 διακεκριμένες ομάδες. Κάθε σύνολο δεδομένων αποτελείται συνολικά από 50, 100, 150 ή 200 στοιχεία και οι ομάδες ενσωματώνονται εξ ορισμού σε έναν ευκλείδειο χώρο διαστάσεων

4, 6 ή 8. Για ευκολία, στη μελέτη μας θα θεωρήσουμε ότι το κάθε σύνολο δεδομένων θα περιέχει 50 στοιχεία.

Οι ομάδες που δημιουργούνται με τη χρήση του αλγορίθμου είναι κατασκευασμένες κατά τέτοιο τρόπο ώστε να ικανοποιούν τις ιδιότητες της εσωτερικής συνοχής και της εξωτερικής απομόνωσης. Η εσωτερική συνοχή προϋποθέτει όλα τα στοιχεία μίας ομάδας να είναι όμοια, ενώ η εξωτερική απομόνωση απαιτεί οι ομάδες να είναι καλά διαχωρισμένες μεταξύ τους, δηλαδή τα στοιχεία που απαρτίζουν μία ομάδα να απέχουν όσο γίνεται περισσότερο από τα στοιχεία μίας άλλης ομάδας.

Η εσωτερική συνοχή των ομάδων επιτυγχάνεται με τη χρήση μίξεων πολυμεταβλητών κανονικών κατανομών. Στην κάθε διάσταση, τα στοιχεία της κάθε ομάδας απέχουν το πολύ 1,5 τυπική απόκλιση από τον μέσο της ομάδας. Από την άλλη, όσον αφορά την ιδιότητα της εξωτερικής απομόνωσης, χρησιμοποιήθηκε μία γεωμετρική προσέγγιση. Συγκεκριμένα, στα σύνορα των ομάδων δεν επιτρέπεται επικάλυψη στην πρώτη διάσταση στο χώρο των μεταβλητών. Η απόλυτη ελάχιστη απόσταση μεταξύ των ορίων γειτονικών ομάδων στην πρώτη διάσταση ισούται με 0.25 φορές το άθροισμα των τυπικών αποκλίσεων μέσα σε αυτές τις δυο ομάδες.

Αποτέλεσμα της παραπάνω διαδικασίας είναι η προκύπτουσα δομή του συνόλου των δεδομένων να αποτελείται από ‘φυσικές’ ομάδες οι οποίες να είναι σαφώς διαχωρισμένες μεταξύ τους και να έχουν υψηλή συνοχή στο εσωτερικό τους. Ο όρος του τυχαίου θορύβου που περιλαμβάνεται δημιουργεί, στην πραγματικότητα, ένα σύνολο δεδομένων με μία συστάδα. Τα πρώτα πειράματα με τον αλγόριθμο έδειξαν μικρή διαφορά στα αποτελέσματα της ομαδοποίησης για τον όρο του τυχαίου θορύβου όταν χρησιμοποιήθηκαν ομοιόμορφα δεδομένα σε αντιδιαστολή με δεδομένα κανονικά κατανεμημένα.

Ο παράγοντας της πυκνότητας των στοιχείων, μέσα σε κάθε 50-αδα, αποτελείται από τρία επίπεδα. Το πρώτο επίπεδο παράγει ίσο αριθμό στοιχείων σε κάθε ομάδα (ή όσο το δυνατόν πιο κοντά στην ισότητα). Το δεύτερο επίπεδο απαιτεί μία ομάδα να περιέχει πάντα το 10% των δεδομένων, ενώ το τρίτο επίπεδο απαιτεί μία ομάδα να περιέχει το 60% των στοιχείων. Τα υπόλοιπα σημεία κατανέμονται όσο το δυνατόν ισομερώς ανάμεσα στις άλλες ομάδες που παρουσιάζονται στα δεδομένα.

Βάσει λοιπόν του αλγορίθμου των Milligan και Cooper, οι τρεις παράγοντες που αντιστοιχούν στον αριθμό των συστάδων, στον αριθμό των διαστάσεων και στο επίπεδο της πυκνότητας διαμορφώνουν ένα τριπλό παραγοντικό σχέδιο το οποίο αποτελείται από 36

κελιά. Σε κάθε κελί πραγματοποιούνται τρεις επαναλήψεις με αποτέλεσμα τη δημιουργία 108 συνόλων δεδομένων. Κάθε σύνολο δεδομένων χρησιμοποιείται για να παραχθεί ένας πίνακας αποστάσεων (από ευκλείδειες αποστάσεις) πάνω στον οποίο θα επιδράσουν τέσσερις μέθοδοι ομαδοποίησης. Με τον τρόπο αυτό παράγεται ένα σύνολο 432 λύσεων. Οι τέσσερις ιεραρχικοί μέθοδοι ομαδοποίησης που χρησιμοποιήθηκαν για την παραγωγή των λύσεων είναι η μέθοδος single linkage, η μέθοδος complete linkage, η μέθοδος του μέσου όρου των ομάδων (group average) και η μέθοδος του Ward.

3.2.2 Έλεγχος και σύγκριση των μεθόδων

Βασιζόμενοι στα σύνολα των τεχνητών δεδομένων, τα οποία παρήχθησαν από τον αλγόριθμο των Milligan και Cooper, είμαστε σε θέση να συγκρίνουμε μία σειρά κριτηρίων διακοπής και να αξιολογήσουμε τα αποτελέσματά τους.

Παρόλο που η φύση των ομάδων που δημιουργήθηκαν ικανοποιεί τις επιθυμητές ιδιότητες (εσωτερική συνοχή και εξωτερική απομόνωση), χρησιμοποιήθηκαν και δύο μέτρα εξωτερικών κριτηρίων για την αξιολόγηση των κριτηρίων διακοπής. Τα δύο αυτά εξωτερικά κριτήρια είναι ο δείκτης του Rand και ο δείκτης Jaccard. Τα αποτελέσματα και των δύο κριτηρίων έδειξαν ότι στην πλειοψηφία των λύσεων των συνόλων των δεδομένων, η ευδιάκριτη ομαδοποίηση ήταν παρούσα στο σωστό επίπεδο της ιεραρχίας. Συγκεκριμένα, από το σύνολο των 432 λύσεων, οι δείκτες έδειξαν 413 ή 412 φορές βέλτιστη αποκατάσταση στο σωστό επίπεδο. Η πληροφορία αυτή, σε συνδυασμό και με τις λεπτομέρειες που προκύπτουν από τη διαδικασία της κατασκευής των ομάδων, φανερώνει πως η ομαδοποίηση των δεδομένων είναι αρκετά ισχυρή.

Τα αποτελέσματα της μελέτης συνοψίζονται στους πίνακες (3.6) έως (3.8). Κάθε εφαρμογή ενός κριτηρίου διακοπής μπορεί να οδηγήσει είτε στην ορθή απόφαση, όσον αφορά τον προσδιορισμό των ομάδων που υπάρχουν στα δεδομένα μας, είτε σε κάποιο σφάλμα. Το σφάλμα αυτό διακρίνεται σε δυο κατηγορίες. Η πρώτη είναι όταν το κριτήριο διακοπής εντοπίζει k ομάδες ενώ στην πραγματικότητα υπάρχουν λιγότερες από k ομάδες στα δεδομένα. Η δεύτερη κατηγορία είναι όταν συμβαίνει ακριβώς το αντίθετο, δηλαδή από το κριτήριο διακοπής εντοπίζονται λιγότερες ομάδες από ότι υπάρχουν στην δομή των

δεδομένων. Ακόμα, υπάρχει και διάκριση των τεχνητών δεδομένων ανάλογα με το πλήθος των ομάδων που περιέχουν. Επομένως, υπάρχει μέτρηση της απόδοσης κάθε κριτηρίου διακοπής χωριστά για κάθε αριθμό ομάδων και το άθροισμα αυτών, καθώς εξάγεται και ένα ποσοστό επιτυχίας της μεθόδου στο επίπεδο της ορθής ομαδοποίησης.

Για ένα δεδομένο κριτήριο διακοπής, η συχνότητα των λύσεων που αποτελούνται από πολύ λίγες και πάρα πολλές συστάδες παρουσιάζονται μαζί με τον αριθμό των σωστών προσδιορισμών. Στους πίνακες εμφανίζονται οι συνολικές αριθμήσεις της συχνότητας οι οποίες χωρίζονται περαιτέρω για τα σύνολα των στοιχείων που περιέχουν 2, 3, 4 και 5 συστάδες.

Τα αποτελέσματα για τους πρώτους έξι κανόνες διακοπής παρουσιάζονται στον πίνακα (3.6). Οι καταχωρήσεις για τη γραμμή με τίτλο "2 ή λιγότερες ομάδες" δίνουν τον αριθμό των περιστατικών όπου ο κανόνας διακοπής παρήγαγε λύσεις με δύο ή λιγότερες συστάδες από όσες ήταν πραγματικά παρούσες στα δεδομένα. Για παράδειγμα, ας υποθέσουμε ότι υπήρξαν πέντε συστάδες παρούσες στα δεδομένα. Εάν ο κανόνας διακοπής πρότεινε ότι υπάρχουν είτε μια είτε δύο είτε τρεις συστάδες παρούσες, τότε το αποτέλεσμα θα καταγραφόταν σε αυτήν την κατηγορία. Η γραμμή "1 ομάδα λιγότερη" δείχνει τον αριθμό των λύσεων όπου ο κανόνας επέλεξε ένα επίπεδο με μια λιγότερη συστάδα από όσες παρουσιάζονται στα δεδομένα. Ομοίως, οι γραμμές "1 ομάδα περισσότερη" ή "2 ομάδες περισσότερες" διευκρινίζουν ότι ο αριθμός λύσεων που έδωσε ο κανόνας είναι μια ή δύο συστάδες αντίστοιχα περισσότερες απ' ότι ήταν πραγματικά παρούσες στα δεδομένα. Τέλος, η γραμμή "3 ή περισσότερες" δείχνει την αρίθμηση της συχνότητας για τον αριθμό των λύσεων όπου ο κανόνας διακοπής επέλεξε ένα επίπεδο που περιέχει 3 ή περισσότερες συστάδες από ότι είναι παρούσες στα δεδομένα. Για παράδειγμα, εάν τα στοιχεία αποτελούνταν από δύο συστάδες και ο κανόνας πρότεινε ότι υπάρχουν πέντε ή περισσότερες συστάδες, το αποτέλεσμα θα αντιστοιχούσε σε αυτήν την κατηγορία.

Βεβαίως, τα αποτελέσματα που παρουσιάζονται στους πίνακες δείχνουν ότι οι κανόνες διακοπής που υπάρχουν μπορούν να είναι αποτελεσματικοί στον καθορισμό του σωστού αριθμού των συστάδων για δεδομένα στα οποία υπάρχει ευδιάκριτη συγκέντρωση. Τα δύο εξωτερικά κριτήρια, ο δείκτης του Jaccard και ο δείκτης Rand παρέχουν τα ανώτερα όρια στην απόδοση ενός κανόνα διακοπής. Λαμβάνοντας υπόψη το γεγονός ότι αυτά τα εξωτερικά κριτήρια παρήγαγαν 413 και 412 σωστούς προσδιορισμούς, αντίστοιχα, καταλήγουμε στο συμπέρασμα ότι ο δείκτης Calinski και Harabasz και ο κανόνας SSE(2)/SSE(1), που

αναπτύχθηκε από τους Duda και Hart (1973), παρείχαν την άριστη αποκατάσταση. Περαιτέρω, όταν εμφανίστηκαν σφάλματα, έτειναν να είναι κοντά στις αποτυχίες. Ο δείκτης Duda και Hart είχε κάποια δυσκολία στο επίπεδο των δύο συστάδων, όπου βρέθηκαν διάφορες λύσεις με πολύ λίγες συστάδες. Αν και η αποκατάσταση στο επίπεδο 2 συστάδων δεν είναι φτωχή, πρέπει να αναγνωριστεί ότι αυτός ο κανόνας διακοπής επιτρέπει στο χρήστη να εξετάσει αν είναι παρούσα στα δεδομένα μόνο μια συστάδα. Ο δείκτης Calinski και Harabasz, αφ' ετέρου, εκτελείται μάλλον με συνέπεια για κάθε αριθμό συστάδων.

Παρόμοια αποτελέσματα με αυτά του δείκτη των Duda και Hart εμφανίζονται και για τους επόμενους τρεις κανόνες του πίνακα (3.6). Ο C-Index, ο δείκτης Gamma των Goodman και Kruskal's (1954) και ο δείκτης Beale εμφανίζουν μια πτώση στην αποκατάσταση στο επίπεδο των δύο συστάδων. Ειδικότερα, ο C-δείκτης και ο κανόνας Gamma παρήγαγαν έναν ουσιαστικό αριθμό ελλείψεων σε αυτό το επίπεδο. Είναι καθησυχαστικό ότι, ο δείκτης που χρησιμοποιείται στο στατιστικό πρόγραμμα SAS, το cubic clustering criterion, δίνει ορθά αποτελέσματα σε ένα ανταγωνιστικό ποσοστό. Ωστόσο, αυτό το κριτήριο έδωσε το υψηλότερο ποσοστό καθορισμού πάρα πολλών συστάδων που έχουν εμφανιστεί μέχρι τώρα, αλλά παρήγαγε έναν σχετικό χαμηλό αριθμό λύσεων με πολύ λίγες συστάδες. Κατά συνέπεια, εάν ο δείκτης δώσει λάθος απόφαση, είναι πιθανότερο να οδηγήσει σε λύσεις με περισσότερες συστάδες από ότι είναι πραγματικά παρούσες στα δεδομένα.

Είναι χρήσιμο να σημειωθεί ότι από τους δείκτες που έχουμε εξετάσει ως τώρα, οι περισσότεροι εμφάνισαν τα χαμηλότερα ποσοστά αποκατάστασης για τις λύσεις όπου δύο συστάδες ήταν παρούσες. Καταλήγουμε λοιπόν στο συμπέρασμα ότι η περίπτωση δύο συστάδων είναι η δυσκολότερη δομή για ανίχνευση από τους κανόνες διακοπής.

Τα αποτελέσματα για τους επόμενους έξι κανόνες εμφανίζονται στον πίνακα (3.7). Οι δείκτες αυτοί αντιπροσωπεύουν τους λιγότερο αποτελεσματικούς κανόνες, όσον αφορά τα αποτελέσματά τους. Τα κριτήρια stepsize και |log(p)| παρήγαγαν την καλύτερη αποκατάσταση στο επίπεδο των δύο συστάδων, ενώ πρότειναν σχετικά λίγες λύσεις με πάρα πολλές συστάδες. Ο likelihood ratio κανόνας παρήγαγε αρκετά σταθερά ποσοστά αποκατάστασης για κάθε αριθμό συστάδων, κατά τρόπο παρόμοιο με τον κανόνα Calinski και Harabasz. Είναι ενδιαφέρον να σημειωθεί ότι ο κανόνας stepsize αντιστοιχεί στο απλούστερο κριτήριο στο πείραμα, ενώ η διαδικασία likelihood ratio είναι μία από τις υπολογιστικά πιο σύνθετες, έχοντας και μια θεωρητική ανάπτυξη. Ωστόσο, υπάρχει μικρή διαφορά στην απόδοση μεταξύ αυτών των δύο κανόνων.

Συνεχίζοντας με τους δείκτες, ας δούμε τις διαδικασίες που παρήγαγαν περισσότερα λάθη από ότι σωστούς προσδιορισμούς. Σε αντίθεση με το δείκτη τ, όλοι οι υπόλοιποι κανόνες διακοπής έχουν ένα ποσοστό λάθους είτε για πολύ λίγες είτε για πάρα πολλές συστάδες, το οποίο και υπερβαίνει το σωστό ποσοστό αποκατάστασης. Για παράδειγμα, ο κανόνας log(SSB/SSW) του πίνακα (3.7) παρουσίασε μεροληψία προς τη λύση των τριών συστάδων, ανεξαρτήτως του πραγματικού αριθμού των συστάδων στα δεδομένα. Το κριτήριο αυτό φαίνεται να δίνει ένα υψηλό ποσοστό σωστής αποκατάστασης στο επίπεδο τριών συστάδων, ένα υψηλό ποσοστό λάθους για τις λύσεις με πάρα πολλές συστάδες όσον αφορά τα σύνολα δεδομένων με δύο συστάδες και διογκωμένα ποσοστά για την άλλη κατηγορία λάθους στα σύνολα δεδομένων τεσσάρων και πέντε συστάδων. Παρόμοια αποτελέσματα εμφανίζονται και στον κανόνα $k^2|W|$ του Marriot του πίνακα (3.8). Αξίζει να σημειωθεί ότι και ο δείκτης τ αντιμετώπισε μία μη αμελητέα δυσκολία στην εύρεση του σωστού αριθμού ομάδων όταν ήταν παρούσες στα δεδομένα πέντε συστάδες.

Ο πίνακας (3.8) περιλαμβάνει τα αποτελέσματα για ένα από τα πιο ευρέως προτεινόμενα κριτήρια διακοπής. Ο δείκτης Trace W είναι ένας μάλλον δημοφιλής δείκτης για χρήση στο πλαίσιο της ομαδοποίησης, συμπεριλαμβανομένης της χρήσης του στον καθορισμό του αριθμού των συστάδων στα δεδομένα. Τα αποτελέσματα για το δείκτη δείχνουν ότι είναι μία ιδιαίτερα ανεπιτυχής επιλογή, δοθέντων των διαθέσιμων εναλλακτικών λύσεων.

Ολοκληρώνοντας με τον έλεγχο των δεικτών, όσον αφορά τον κανόνα του Mountford (1970), παρατηρούμε ένα διογκωμένο ποσοστό για τις λύσεις με πάρα πολλές συστάδες. Επιπλέον, ο δείκτης $|T|/|W|$ δεν ανίχνευσε ποτέ στις 432 προσπάθειες το σωστό αριθμό συστάδων στα δεδομένα. Είναι σαφές από τον πίνακα ότι εκτός από το κριτήριο $TraceW^{-1}B$, οι δείκτες Mountford και $|T|/|W|$ επέλεξαν λύσεις που πρότειναν, σε πολύ μεγάλο ποσοστό, ότι ήταν παρούσες στα δεδομένα πάρα πολλές συστάδες.

Στη συνέχεια παρατίθενται οι πίνακες με τα αποτελέσματα του πειράματος. Οι μέθοδοι παρουσιάζονται με αύξουσα σειρά ανάλογα με την επιτυχία που είχαν στην σύγκριση. Το πιο επιτυχημένο κριτήριο είναι αυτό του Calinski and Harabasz (1974) το οποίο και δίνει ένα αρκετά μεγάλο ποσοστό επιτυχίας (90,27%). Συγκεκριμένα, από τις 432 συνολικά επιλύσεις, 390 φορές βρίσκει το σωστό αριθμό των ομάδων. Ακολουθούν τα υπόλοιπα κριτήρια και τέλος ο λόγος $|T|/|W|$ που δεν κατάφερε να πετύχει καμία ορθή ομαδοποίηση, καθώς όλες οι

ομαδοποιήσεις που έκρινε ως βέλτιστες είχαν περισσότερες ομάδες απ' ότι η πραγματική ομαδοποίηση.

	Πραγματικός αριθμός ομάδων					
Κριτήριο Διακοπής	**2**	**3**	**4**	**5**	**Σύνολο**	**%Ποσοστό**
Calinski and Harabasz						
2 ή λιγότερες ομάδες	-	-	1	0	1	
1 ομάδα λιγότερη	-	12	6	0	18	
σωστός αριθμός ομάδων	96	95	97	102	390	90,27%
1 ομάδα περισσότερη	3	0	3	6	12	
2 ομάδες περισσότερες	4	0	1	0	5	
3 ομάδες περισσότερες	5	1	0	0	6	
Duda and Hart						
2 ή λιγότερες ομάδες	-	0	0	0	0	
1 ομάδα λιγότερη	16	4	3	0	23	
σωστός αριθμός ομάδων	77	101	103	107	388	89,81%
1 ομάδα περισσότερη	4	2	1	1	8	
2 ομάδες περισσότερες	7	0	0	0	7	
3 ομάδες περισσότερες	4	1	1	0	6	
C-Index						
2 ή λιγότερες ομάδες	-	-	4	0	4	
1 ομάδα λιγότερη	-	5	12	12	29	
σωστός αριθμός ομάδων	71	89	91	96	347	80,32%
1 ομάδα περισσότερη	2	7	1	0	10	
2 ομάδες περισσότερες	2	0	0	0	2	
3 ομάδες περισσότερες	33	7	0	0	40	
Goodman and Kruskal						
2 ή λιγότερες ομάδες	-	-	8	0	8	
1 ομάδα λιγότερη	-	9	16	12	37	
σωστός αριθμός ομάδων	74	86	83	96	339	78,47%
1 ομάδα περισσότερη	3	7	1	0	11	
2 ομάδες περισσότερες	2	0	0	0	2	
3 ομάδες περισσότερες	29	6	0	0	35	
Beale						
2 ή λιγότερες ομάδες	-	3	4	0	7	
1 ομάδα λιγότερη	34	8	5	8	55	
σωστός αριθμός ομάδων	57	87	95	92	331	76,62%
1 ομάδα περισσότερη	0	1	0	0	1	
2 ομάδες περισσότερες	4	0	0	0	4	
3 ομάδες περισσότερες	13	9	4	8	34	
Cubic Clustering						
2 ή λιγότερες ομάδες	-	0	0	0	0	
1 ομάδα λιγότερη	13	5	4	0	22	
σωστός αριθμός ομάδων	67	88	82	84	321	74,30%
1 ομάδα περισσότερη	5	7	10	13	35	
2 ομάδες περισσότερες	6	0	7	9	22	
3 ομάδες περισσότερες	17	8	5	2	32	

Πίνακας 3.6

Ο πρώτος πίνακας αποτελεσμάτων από την εφαρμογή διαφόρων κριτηρίων διακοπής

	Πραγματικός αριθμός ομάδων					
Κριτήριο Διακοπής	**2**	**3**	**4**	**5**	**Σύνολο**	**%Ποσοστό**
G(+)						
2 ή λιγότερες ομάδες	-	-	4	0	4	
1 ομάδα λιγότερη	-	8	16	12	36	
σωστός αριθμός ομάδων	52	70	79	96	297	68,75%
1 ομάδα περισσότερη	1	4	4	0	6	
2 ομάδες περισσότερες	0	0	0	0	0	
3 ομάδες περισσότερες	55	26	8	0	89	
Stepsize						
2 ή λιγότερες ομάδες	-	-	37	29	66	
1 ομάδα λιγότερη	-	51	18	11	80	
σωστός αριθμός ομάδων	96	56	53	68	273	63,20%
1 ομάδα περισσότερη	6	1	0	0	7	
2 ομάδες περισσότερες	1	0	0	0	1	
3 ομάδες περισσότερες	5	0	0	0	5	
Likelihood ratio						
2 ή λιγότερες ομάδες	-	0	0	0	0	
1 ομάδα λιγότερη	12	0	4	0	16	
σωστός αριθμός ομάδων	64	72	64	68	268	62%
1 ομάδα περισσότερη	12	16	17	25	70	
2 ομάδες περισσότερες	9	10	16	9	44	
3 ομάδες περισσότερες	11	10	7	6	34	
|log(p)|						
2 ή λιγότερες ομάδες	-	-	43	52	95	
1 ομάδα λιγότερη	-	35	19	11	65	
σωστός αριθμός ομάδων	78	71	45	43	237	54,86%
1 ομάδα περισσότερη	11	1	1	1	14	
2 ομάδες περισσότερες	10	0	0	1	11	
3 ομάδες περισσότερες	9	1	0	0	10	
Log(SSB/SSW)						
2 ή λιγότερες ομάδες	-	-	0	22	22	
1 ομάδα λιγότερη	-	0	66	20	86	
σωστός αριθμός ομάδων	0	104	42	66	212	49,07%
1 ομάδα περισσότερη	52	3	0	0	55	
2 ομάδες περισσότερες	19	0	0	0	19	
3 ομάδες περισσότερες	37	1	0	0	38	
Tau						
2 ή λιγότερες ομάδες	-	-	25	52	77	
1 ομάδα λιγότερη	-	28	52	46	126	
σωστός αριθμός ομάδων	85	77	30	10	202	46,76%
1 ομάδα περισσότερη	7	2	1	0	10	
2 ομάδες περισσότερες	7	0	0	0	7	
3 ομάδες περισσότερες	9	1	0	0	10	

Πίνακας 3.7

Ο δεύτερος πίνακας αποτελεσμάτων

	Πραγματικός αριθμός ομάδων					
Κριτήριο Διακοπής	**2**	**3**	**4**	**5**	**Σύνολο**	**%Ποσοστό**
Marriot						
2 ή λιγότερες ομάδες	-	-	0	73	73	
1 ομάδα λιγότερη	-	0	92	6	98	
σωστός αριθμός ομάδων	0	104	15	27	146	33,80%
1 ομάδα περισσότερη	95	4	1	2	102	
2 ομάδες περισσότερες	8	0	0	0	8	
3 ομάδες περισσότερες	5	0	0	0	5	
Trace W						
2 ή λιγότερες ομάδες	-	-	0	88	88	
1 ομάδα λιγότερη	-	0	92	20	112	
σωστός αριθμός ομάδων	0	104	16	0	120	27,78%
1 ομάδα περισσότερη	62	3	0	0	65	
2 ομάδες περισσότερες	19	0	0	0	19	
3 ομάδες περισσότερες	27	1	0	0	28	
Trace $W^{-1}B$						
2 ή λιγότερες ομάδες	-	0	0	69	69	
1 ομάδα λιγότερη	0	0	56	19	75	
σωστός αριθμός ομάδων	0	52	23	9	84	19,44%
1 ομάδα περισσότερη	85	48	27	3	163	
2 ομάδες περισσότερες	15	7	0	0	22	
3 ομάδες περισσότερες	8	1	2	8	19	
Mountford						
2 ή λιγότερες ομάδες	0	0	0	0	0	
1 ομάδα λιγότερη	0	0	0	0	0	
σωστός αριθμός ομάδων	1	6	1	2	10	2,30%
1 ομάδα περισσότερη	0	0	7	4	11	
2 ομάδες περισσότερες	3	3	3	3	12	
3 ομάδες περισσότερες	104	99	97	99	399	
|T|/|W|						
2 ή λιγότερες ομάδες	-	0	0	0	0	
1 ομάδα λιγότερη	-	0	0	0	0	
σωστός αριθμός ομάδων	0	0	0	0	0	0%
1 ομάδα περισσότερη	0	0	0	0	0	
2 ομάδες περισσότερες	0	0	0	0	0	
3 ομάδες περισσότερες	108	108	108	108	432	

Πίνακας 3.8

Ο τρίτος πίνακας αποτελεσμάτων

3.2.3 Γενικότερα συμπεράσματα

Τα αποτελέσματα της μελέτης των Milligan και Cooper είναι ενδιαφέροντα λόγω του ευρέως φάσματος απόδοσης που βρέθηκε για τα κριτήρια διακοπής. Φαίνεται ότι έχει προσδιοριστεί ένα σχετικά ακριβές σύνολο κριτηρίων που να μπορεί να βοηθήσει τους ερευνητές στον καθορισμό του αριθμού συστάδων σε ένα σύνολο δεδομένων. Επιπλέον, κάποια κριτήρια διακοπής βρέθηκαν να είναι ιδιαίτερα ατελέσφορα στην παραγωγή των σωστών προσδιορισμών σε σύνολα δεδομένων που κατείχαν ιδιαίτερα ισχυρή και ευδιάκριτη δομή συστάδων.

Αν και τα αποτελέσματα σχετικά με τις καλύτερες μεθόδους στον πίνακα (3.6) είναι ιδιαίτερα ενθαρρυντικά, πρέπει να σημειωθεί ότι τα συμπεράσματα είναι πιθανό να είναι κάπως εξαρτώμενα από τα δεδομένα. Δεν θα μας εξέπληττε αν διαπιστώναμε ότι η σειρά των δεικτών θα άλλαζε εάν χρησιμοποιούνταν διαφορετικές δομές στα δεδομένα. Εντούτοις, αν χρησιμοποιούνταν διαφορετικές δομές δεδομένων, θα φαινόταν μάλλον απίθανο ένας δείκτης που βρίσκεται μεταξύ των καλύτερων δεικτών στην παρούσα μελέτη να τοποθετηθεί σε αυτούς με τις χειρότερες αποδόσεις.

Ως παράδειγμα της εξάρτησης των δεδομένων, θα αναφέρουμε τη διαδικασία των Duda and Hart (1973) η οποία απαιτεί τον υπολογισμό μιας κρίσιμης τιμής για λόγους απόφασης. Δεν θα φαινόταν αδικαιολόγητο να αναμείνουμε ότι ίσως να αλλάξει η βέλτιστη τιμή εάν η δομή των δεδομένων αλλάξει. Ως εκ τούτου, η βέλτιστη αυτή κρίσιμη τιμή είναι πιθανό να εξαρτάται από τα δεδομένα. Για το λόγο αυτό απαιτείται περαιτέρω έρευνα για την επίδραση εναλλακτικών δομών δεδομένων στις παραμέτρους του ελέγχου. Ωστόσο, είναι χρήσιμο να σημειωθεί ότι υπάρχουν περιπτώσεις όπου οι παράμετροι του ελέγχου δεν είναι απαραίτητες για την ανάπτυξη ενός αποτελεσματικού κριτηρίου διακοπής. Για παράδειγμα, η διαδικασία Calinski και Harabasz, το καλύτερο κριτήριο διακοπής που βρέθηκε στο πείραμα, δεν εξαρτάται από τον καθορισμό μίας κρίσιμης τιμής. Αφ' ετέρου, είναι αξιοπρόσεκτο ότι η διαδικασία Duda και Hart υπολογίζεται μόνο από την πληροφορία που παρέχεται από τα στοιχεία που περιλαμβάνονται στην τελευταία συγχώνευση των συστάδων.

Ωστόσο, θα μπορούσε κάποιος να υποστηρίξει, ότι κάποιοι άλλοι κανόνες μπορούν να είναι σε θέση να παρουσιάσουν βελτιωμένη απόδοση κάτω από όρους που διαφέρουν από την παρούσα μελέτη. Για παράδειγμα, ο κανόνας likelihood ratio θα περιμέναμε να αποδώσει καλύτερα για σύνολα στοιχείων που αποτελούνται από ένα πολύ μεγαλύτερο δειγματικό

μέγεθος. Ωστόσο, εάν οι καλύτερες μέθοδοι εκτέλεσης που βρήκαμε στην παρούσα μελέτη αποδίδουν επίσης καλά παρουσία μεγαλύτερων δειγματικών μεγεθών, τότε δεν υπάρχει κανένας ιδιαίτερος λόγος να ακολουθήσουν τις ασυμπτωτικές μεθόδους, δεδομένου ότι πέρα από το χρήσιμο εύρος δειγματικού μεγέθους δεν παρέχουν με συνέπεια ανώτερη απόδοση.

Στη συνέχεια θα κάνουμε μερικά σχόλια σχετικά με τη χρήση των σκορ διαφοράς. Ένα σκορ διαφοράς είναι η σύγκριση των τιμών ενός κριτηρίου από το ένα επίπεδο της ιεραρχίας στο επόμενο. Ουσιαστικά, αρκετά κριτήρια διακοπής ενσωματώνουν στο βασικό καθορισμό τους συγκρίσεις μεταξύ των επιπέδων, όπως για παράδειγμα ο κανόνας SEE(2)/SSE(1) των Duda και Hart, ο κανόνας του Beale, καθώς και οι δείκτες stepsize και likelihood ratio. Τα σκορ διαφοράς υιοθετήθηκαν από μερικά κριτήρια διακοπής στην παρούσα μελέτη, επειδή βελτιστοποίησαν την απόδοση ορισμένων κριτηρίων, όπως οι κανόνες |log(p)|, log(SSB/SSW), ο δείκτης $k^2|W|$ του Marriot και οι δείκτες TraceW, $TraceW^{-1}B$ και |T|/|W|. Αξίζει να σημειωθεί ότι οι περισσότερες από αυτές τις εφαρμογές έδωσαν σχετικά μία φτωχή αποκατάσταση του σωστού αριθμού των συστάδων. Για το λόγο αυτό, η εφαρμογή των σκορ διαφοράς φαίνεται να είναι μικρής αξίας.

Διάφοροι συγγραφείς έχουν υποστηρίξει την ανάπτυξη αυστηρών στατιστικών τεχνικών στη μέθοδο της ανάλυσης των συστάδων (Fleiss & Zubin (1969), Goodall (1966)). Η πρόοδος σε αυτήν την περιοχή είναι αργή, γεγονός το οποίο, χωρίς καμία αμφιβολία, οφείλεται εν μέρει στο πρόβλημα για το ποια κατανομή θα χρησιμοποιηθεί. Οι στατιστικοί τείνουν να ξεκινούν μία ανάλυση υποθέτοντας πολυμεταβλητή κανονικότητα. Αυτή η πρακτική έχει εξεταστεί σοβαρά από τον Gower (1981), ο οποίος υποστήριξε ότι υπάρχουν καλύτερες εναλλακτικές στατιστικές στρατηγικές. Ασφαλώς, η μάλλον κακή απόδοση των τυποποιημένων πολυμεταβλητών κανονικών κριτηρίων στην παρούσα μελέτη, όπως το TraceW, το Trace $W^{-1}B$ και |T|/|W|, προσδίδει αξιοπιστία στα επιχειρήματα του Gower. Για το λόγο αυτό, θα ήταν χρήσιμο οι ερευνητές να εξετάσουν εναλλακτικές στρατηγικές όσον αφορά το είδος της κατανομής που θα χρησιμοποιηθεί.

3.3 Επίλογος

Τις τελευταίες δεκαετίες έχει πραγματοποιηθεί πολλή θεωρητική και εφαρμοσμένη έρευνα ώστε να αντιμετωπισθεί το πρόβλημα της εύρεσης του βέλτιστου αριθμού ομάδων. Τα

αποτελέσματα ωστόσο δεν είναι εντελώς ενθαρρυντικά. Τα τελευταία χρόνια συγκεκριμένοι τομείς όπως η επεξεργασία εικόνας, τα ιατρικά διαγνωστικά συστήματα κ.τ.λ. έχουν αναπτυχθεί γρήγορα λόγω της υποστήριξής τους από την προχωρημένη τεχνολογία των υπολογιστών και ως συνέπεια αυτών υπάρχει μία ισχυρή απαίτηση να λυθεί το πρόβλημα του αριθμού των ομάδων, προκειμένου να δοθούν πιο ακριβείς λύσεις στο πρόβλημα που έχουν να επιλύσουν.

Όπως είδαμε, οι διάφοροι μέθοδοι ομαδοποίησης που χρησιμοποιούνται δεν μπορούν να προσδιορίσουν τον ακριβή αριθμό των ομάδων από τις οποίες απαρτίζονται τα δεδομένα. Με τον συνδυασμό λοιπόν ενός κριτηρίου διακοπής μπορούμε να τερματίσουμε έναν αλγόριθμο ομαδοποίησης ώστε να καταλήξουμε στον βέλτιστο διαχωρισμό ενός συνόλου δεδομένων.

Στην εργασία μας μελετήσαμε μία σειρά κριτηρίων διακοπής, βασιζόμενοι στα αποτελέσματα της εργασίας των Milligan και Cooper (1985). Τα κριτήρια αυτά προήλθαν από μία ποικιλία πηγών μέσω προσομοιώσεων Monte Carlo. Τα δεδομένα που χρησιμοποιήθηκαν είχαν μία δεδομένη δομή και οι ομαδοποιήσεις που πραγματοποιήθηκαν από τις μεθόδους ομαδοποίησης, σε συνδυασμό με τα κριτήρια διακοπής, εξετάστηκαν με τη χρήση εξωτερικών κριτηρίων.

Τα αποτελέσματα έδειξαν ότι αρκετά από τα κριτήρια διακοπής που ελέγξαμε μπορούν να καθορίσουν με επιτυχία τον βέλτιστο αριθμό των συστάδων σε ένα σύνολο δεδομένων, καθώς και να τοποθετήσουν σωστά τα δεδομένα μέσα στις συστάδες. Για το λόγο αυτό, κάποια από τα κριτήρια θεωρούνται από πολλούς ερευνητές ως βάσεις για περαιτέρω μελέτες πάνω στο συγκεκριμένο πεδίο της εύρεσης του βέλτιστου αριθμού των ομάδων στην ανάλυση κατά συστάδες.

ΕΠΙΛΟΓΟΣ - ΕΦΑΡΜΟΓΕΣ

Η μέθοδος της ανάλυσης σε ομάδες είναι μία τεχνική η οποία συγκαταλέγεται στις τεχνικές της πολυμεταβλητής στατιστικής ανάλυσης (multivariate statistical analysis). Στην πράξη τα δεδομένα ενός ερευνητή είναι από τη φύση τους πολυμεταβλητά και σπάνια ο σκοπός του ερευνητή είναι να μελετήσει μια μεταβλητή ανεξάρτητα και απομονωμένα από τις υπόλοιπες.

Σκοπός της ανάλυσης σε ομάδες είναι η δημιουργία ομάδων από παρατηρήσεις για τις οποίες τα δεδομένα δείχνουν πως έχουν παρόμοια χαρακτηριστικά. Με τον τρόπο αυτό επιτυγχάνουμε την ευκολότερη και αποδοτικότερη επεξεργασία των δεδομένων που διαθέτουμε.

Τα αποτελέσματα της ανάλυσης σε ομάδες μπορούν να συμβάλουν στην δημιουργία διαφόρων κατηγοριών ή ταξινομήσεων. Για παράδειγμα, στη βιολογία και στη βοτανολογία, μία από τις κύριες εφαρμογές της ανάλυσης σε ομάδες είναι η κατασκευή συστημάτων ταξινόμησης. Σε κάποιες άλλες περιπτώσεις, η ανάλυση σε ομάδες μπορεί να χρησιμοποιηθεί είτε για τη μείωση των διαστάσεων του προβλήματος, είτε για την κατάταξη καινούργιων παρατηρήσεων. Ένα κλασικό παράδειγμα κατάταξης νέων παρατηρήσεων αφορά το αν θα δοθεί ένα δάνειο ή όχι σε έναν υποψήφιο πελάτη μίας τράπεζας. Από τα δεδομένα που έχει η τράπεζα γνωρίζει την εξέλιξη για ένα μεγάλο πλήθος δανείων (αν το δάνειο αποπληρώθηκε σωστά ή όχι) καθώς και όλα τα στοιχεία των δανειοληπτών. Επομένως μπορεί να κατασκευάσει έναν κανόνα σχετικά με το ποια χαρακτηριστικά του δανειολήπτη είναι εκείνα που επιδρούν στο να αποπληρώσει κανονικά ή όχι το δάνειο. Όταν λοιπόν ένας καινούργιος πελάτης ζητήσει δάνειο μπορεί ο ερευνητής χρησιμοποιώντας τον κανόνα αυτό να κατατάξει τον νέο πελάτη είτε στην κατηγορία των ‘καλών’ πελατών είτε στην κατηγορία των ‘κακών’ πελατών.

Μία σημαντική ερώτηση που πρέπει να απαντηθεί προτού εφαρμοστεί μία μέθοδος ομαδοποίησης είναι πόσες ομάδες υπάρχουν σε ένα σύνολο δεδομένων. Η πληροφορία αυτή μπορεί να μην είναι γνωστή ***a priori*** και στην πραγματικότητα μπορεί να μην υπάρχει ακριβής απάντηση στο συγκεκριμένο ερώτημα. Οι ερευνητές ενδιαφέρονται να βρουν χρήσιμες δομές σε ένα σύνολο δεδομένων χωρίς συνήθως να προσδοκούν κάποιο συγκεκριμένο αποτέλεσμα. Με την εύρεση του βέλτιστου αριθμού των ομάδων σε ένα

σύνολο δεδομένων επιτυγχάνεται ο προσδιορισμός της αντιπροσωπευτικότερης ομαδοποίησης ως προς τα διαφορετικά χαρακτηριστικά των δεδομένων.

Η ομαδοποίηση των δεδομένων έχει παίξει καθοριστικό ρόλο στην ανάπτυξη πολλών επιστημονικών κλάδων. Για παράδειγμα, οι αρχαιολόγοι ενδιαφέρονται να κατατάξουν τα ευρήματα μιας ανασκαφής σε ομάδες που θα μπορούσαν για παράδειγμα να αντανακλούν διαφορετικές χρονικές περιόδους ή κατηγορίες ευρημάτων, όπως ξύλινα εργαλεία, αντικείμενα ταφής κλπ. Για να το πετύχουν αυτό χρησιμοποιήσουν μια σειρά από μετρήσεις σχετικές με τα ευρήματα και με βάση αυτές τις μετρήσεις ομαδοποιούν τα ευρήματά τους.

Στις Βιο-επιστήμες (Βιολογία, Βοτανολογία, Ζωολογία, Εντομολογία, Μικροβιολογία, Οικολογία, Παλαιοντολογία κλπ.) εξετάζουμε ζωντανούς οργανισμούς όπως φυτά, ζώα, έντομα, μικροοργανισμούς, απολιθώματα. Η ομαδοποίηση των οργανισμών χρησιμεύει στον καθορισμό διαφόρων ειδών/ ποικιλιών. Για παράδειγμα, οι βιολόγοι ενδιαφέρονται να κατατάξουν διαφορετικά είδη ζώων σε ομάδες με βάση κάποια χαρακτηριστικά τους. Η βέλτιστη ομαδοποίηση έγκειται στη δημιουργία αντιπροσωπευτικών ταξινομήσεων ούτως ώστε σε ένα ευδιάκριτο άλλα ποικίλο είδος να περιοριστεί η δημιουργία υποομάδων.

Άμεση σχέση με τις Βιο-επιστήμες έχουν και οι Ιατρικές επιστήμες (Ψυχιατρική, Παθολογία και άλλες ειδικότητες που εστιάζουν σε κλινικές διαγνώσεις). Τα άτομα που εξετάζουμε μπορεί να είναι ασθενείς, ασθένειες, συμπτώματα ή κλινικές διαγνώσεις. Στις περιπτώσεις αυτές είναι απαραίτητη η χρήση της ανάλυσης κατά συστάδες για την αποτελεσματικότερη διάγνωση και θεραπεία των διαφόρων ασθενειών. Έτσι, για παράδειγμα, στον τομέα της Ψυχιατρικής, η βέλτιστη ομαδοποίηση των συμπτωμάτων των διαφόρων ασθενειών, όπως παράνοια, σχιζοφρένεια κλπ, μπορεί να χρησιμεύσει στην ταχύτερη και εγκυρότερη διάγνωση της ασθένειας και κατά επέκταση σε μία επιτυχή θεραπεία.

Η μέθοδος της ανάλυσης σε συστάδες εφαρμόζεται ευρέως και στις Πολιτικές και Οικονομικές επιστήμες στους κλάδους πληροφοριών, πολιτικής αποφάσεων, έρευνας αγοράς και επιχειρησιακής έρευνας. Τα άτομα τα οποία ομαδοποιούνται μπορεί να είναι κράτη ή πόλεις, ψήφοι, βιομηχανίες, καταναλωτές, προϊόντα, επενδύσεις κλπ. Έστω για παράδειγμα ότι έχουμε ένα σύνολο δεδομένων το οποίο αφορά διαφορετικές χώρες και τα αντίστοιχα ποσοστά εργασιακής απασχόλησης που συναντώνται σε διάφορους κλάδους της βιομηχανίας. Από τη φύση τους τα δεδομένα είναι πολυμεταβλητά και σκοπός της ανάλυσης κατά συστάδες είναι να ομαδοποιήσει τις χώρες βάσει των εργασιακών τους προτύπων. Η εύρεση

του βέλτιστου αριθμού των ομάδων μπορεί να βοηθήσει στο να αντιληφθούμε άλλες τυχόν σχέσεις που υπάρχουν μεταξύ των χωρών που ομαδοποιήθηκαν (οικονομική κατάσταση, πολιτική κατάσταση, γεωγραφική θέση κλπ.).

Στο μάρκετινγκ, οι ερευνητές αγοράς ενδιαφέρονται να γνωρίζουν ποια είναι τα χαρακτηριστικά των αγοραστών. Επομένως, μελετώντας δεδομένα που αφορούν τα δημογραφικά χαρακτηριστικά των πελατών και την καταναλωτική τους συνήθεια ως προς κάποιο προϊόν, δημιουργούν ομάδες αγοραστών και με βάση την ομαδοποίηση αυτή μπορούν να κατευθύνουν τις μελλοντικές τους κινήσεις, όπως διαφήμιση, προσφορές κλπ. Κατά επέκταση, η βελτιστοποίηση ενός αλγορίθμου ομαδοποίησης μπορεί να φανεί ιδιαίτερα χρήσιμη στον σωστό προσδιορισμό του αριθμού των τμημάτων της αγοράς, καθώς και στον προσδιορισμό του αριθμού των ευδιάκριτων προτύπων εξόδων σε μία μελέτη που αφορά τη συμπεριφορά των καταναλωτών.

Διάφορα παραδείγματα εφαρμογών της ανάλυσης σε ομάδες συναντάμε και στις Κοινωνικές επιστήμες, όπως Ψυχολογία, Κοινωνιολογία, Εγκληματολογία, Εκπαίδευση κλπ., όπου οι ερευνητές εξετάζουν τα χαρακτηριστικά ενός συνόλου ατόμων (κουλτούρα, μέθοδοι εκπαίδευσης, κοινωνικές συμπεριφορές, εγκλήματα κλπ.) και ενδιαφέρονται να αναγνωρίσουν μέσα από το σύνολο άτομα με παρόμοιους στόχους και συμπεριφορές.

Τέλος, είναι σημαντικό να αναφέρουμε ότι τα τελευταία χρόνια η ανάλυση κατά συστάδες έχει παίξει σημαντικό ρόλο στις σύγχρονες τεχνολογίες της πληροφορικής (αναγνώριση προτύπων, τεχνητή νοημοσύνη, ρομποτική, ηλεκτροδιαγράμματα, σήματα από ραντάρ κλπ.). Ιδιαίτερο ενδιαφέρον παρουσιάζει η ομαδοποίηση δεδομένων στο Διαδίκτυο (Internet) (Καρλής (2005)). Οι σχεδιαστές ηλεκτρονικών σελίδων ενδιαφέρονται να βρουν και να ομαδοποιήσουν τη συμπεριφορά των χρηστών του Internet ανάλογα με τον τρόπο με τον οποίο σερφάρουν σε διαφορετικές σελίδες. Επομένως, η συμπεριφορά τους, όπως καταγράφεται με τη διαδοχική εναλλαγή σελίδων προσφέρει δεδομένα με σκοπό την ομαδοποίηση των χρηστών.

Εκτός από τα παραδείγματα που αναφέραμε, η ανάλυση κατά συστάδες βρίσκει πληθώρα εφαρμογών σχεδόν σε κάθε επιστήμη και επομένως αποτελεί ένα πολυτιμότατο εργαλείο στα χέρια όλων των επιστημονικών κλάδων. Ωστόσο, η ανάλυση κατά συστάδες θα πρέπει να χρησιμοποιείται με μεγάλη προσοχή και αυτό γιατί μπορεί να είναι εύκολο μέσω του

υπολογιστή να πάρουμε κάποια ομαδοποίηση, όμως θα πρέπει να είμαστε σε θέση να διακρίνουμε αν η συγκεκριμένη ομαδοποίηση που συντελέστηκε έχει νόημα.

Από μία πιο θεωρητική σκοπιά, η ανάλυση κατά συστάδες μπορεί να χρησιμοποιηθεί για την ανάπτυξη επαγωγικών γενικεύσεων. Συνήθως όμως ένα σύνολο αποτελεσμάτων βρίσκει εφαρμογή μόνο στο δείγμα στο οποίο βασίστηκε. Με κατάλληλους ωστόσο μετασχηματισμούς τα αποτελέσματα αυτά μπορούν να περιγράψουν επαρκώς τις ιδιότητες και άλλων δειγμάτων και κατά επέκταση τον πληθυσμό από τον οποίο προέρχονται τα δεδομένα.

ΒΙΒΛΙΟΓΡΑΦΙΑ

Ελληνική

Καρλής, Δ. (2005). *Πολυμεταβλητή στατιστική ανάλυση*, Εκδόσεις Αθ. Σταμούλης, Αθήνα.

Κούτρας, Μ. (2005). *Εφαρμοσμένη Πολυμεταβλητή Ανάλυση: Ανάλυση κατά συστάδες*, Πανεπιστημιακές παραδόσεις, Πανεπιστήμιο Πειραιώς.

Ξένη

Arnold, S. J. (1979). A test for clusters, *Journal of Marketing Research*, **19**, 545-551.

Ball, G. H. and Hall, D. J. (1965). *ISODATA, A novel method of data analysis and pattern classification*, Menlo Park: Stanford Research Institute, (NTIS No. AD 699616).

Beale, E. M. L. (1969). *Cluster Analysis*, London: Scientific Control Systems.

Binder, D. A. (1978). Bayesian cluster analysis, *Biometrika*, **65**, 31-38.

Bock, H. H. (1977). On tests concerning the existence of a classification, *In First international symposium on data analysis and informatics*, **2**, 449-464, Rocquencourt, France: IRIA.

Calinski, R. B. and Harabasz, J. (1974). A dendrite method for cluster analysis, *Communications in Statistics*, **3**, 1-27.

Carmichael, J. W. and Sneath, P. H. A. (1969). Taxometric Maps, *Syst. Zool.*, **18**, 402-415.

Carmichael, J. W., Georges, J. A. and Julius, R. S. (1968). Finding Natural Clusters, *Syst. Zool.*, **17**, 144-150.

Cattel, R. B. and Coulter, M. A. (1966). Principles of behavioral taxonomy and the mathematical basis of the taxonomy computer program, *Br. J. Math. Statist. Phychol.*, **19**, 237-269.

Cormack, R. M. (1971). A review of classification, *J. R. Statist. Soc.*, Series A, **134**, No. 3, 321-367.

Davies, D. L., Bouldin, D. W. (1979). A cluster separation measure, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **1**, 224-227.

Day, N. E. (1969). Estimating the components of a mixture of normal distributions, *Biometrika*, **56**, 463-474.

Duda, R. O. and Hart, P. E. (1973). *Pattern classification and scene analysis*, New York: Wiley.

Edwards, A. W. F. and Cavalli-Sforza, L. (1965). A method for cluster analysis, *Biometrics*, **21**, 362-375.

Everitt, B. S. (1981). *Cluster Analysis*, Heinemann Educational Books, London

Everitt, B. S. (1979). Unresolved Problems in Cluster Analysis, *Biometrics*, 3**5**, 169-181.

Fleiss, J. L. and Zubin, J. (1969). On the methods and theory of clustering, *Multivariate Behavioral Research*, **4**, 235-250.

Friedman, H. P. and Rubin, J. (1967). On some invariant criteria for grouping data, *Journal of the American Statistical Association*, **62**, 1159-1178.

Frey, T. and Van Groenewoud, H. (1972). A cluster analysis of the D-squared matrix of white spruce stands in Saskatchewan based on the maximum-minimum principle, *Journal of Ecology*, **60**, 873-886

Gitman, I. and Levine, M. D. (1970). An algorithm for detecting unimodal fuzzy sets and its application as a clustering technique, *IEEE, Trans. Comp*., **19**, 583-593.

Gnanadesikan, R., Kettenring, J. R. and Landwehr, J. M. (1977). Interpreting and assessing the results of cluster analyses, *Bulletin of the International Statistical Institute*, **47**, 451-463.

Goodall, D. W. (1966). Hypothesis testing in classification, *Nature*, **221**, 329-330.

Goodman, L. A. and Kruskal, W. H. (1954). Measures of association for cross-classifications, *Journal of the American Statistical Association*, **49**, 732-64.

Gower, J. C. (1981, June*). Is classification statistical?*, Paper presenting at the meeting of the Classification Society, Toronto.

Gower, J. C. (1970). Classification and Geology, *Rev. I.S.I.*, **38**, 35-41.

Hartigan, J. A. (1975). *Clustering algorithms*, New York: Wiley.

Hubert, L. J. and Levin, J. R. (1976). A general statistical framework of assessing categorical clustering in free recall, *Psychological Bulletin*, **83**, 1072-1080.

Lingoes, J. C. and Cooper, T. (1971). PEP-I: A FORTRAN IV (G) program for Guttman-Lingoes nonmetric probability clustering, *Behavioral Science*, **16**, 259-261.

Marriot, F. H. C. (1971). Practical problems in a method of cluster analysis, *Biometrics*, **27**, 501-514.

McClain, J. O. and Rao, V. R. (1975), CLUSTISZ: A program to test for the quality of clustering of a set of objects, *Journal of Marketing Research*, **12**, 456-460.

Milligan, G. W. (1981a). A Monte Carlo study of thirty internal criterion measures for cluster analysis, *Psychometrika*, **46**, 187-199.

Milligan, G. W. and Cooper, M. C. (1985). An examination of procedures for determining the number of clusters in a data set, *Psychometrika*, **50**, 159-179.

Mojena, R. (1977). Hierarchical grouping methods and stopping rules: An evaluation, *The Computer Journal*, **20**, 359-363.

Mountford , M. D. (1970). *A test for the difference between clusters*, In G. P. Patil, E. C. Pielou and W. E. Waters (Eds.), Statistical Ecology, **3**, 237-257, University Park, Pa.: Pennsylvania State University Press.

Ratkowsky, D. A. and Lance, G. N. (1978). A criterion for determining he number of groups in a classification, *Australian Computer Journal*, **10**, 115- 117.

Ray, A. A. (Ed.). (1982). *SAS use's guide: Statistics*, Cary, North Carolina: SAS Institute.

Rohlf, F. J. (1974). Methods of comparing classifications, *Annual Review of Ecology and Systematics*, **5**, 101-113.

Sarle, W. S. (1983). *Cubic clustering criterion (Tech. Rep. A-108)*, Cary, North Carolina: SAS Institute.

Scott, A. J. and Symons, M. J. (1971). Clustering methods based on likelihood ratio criteria, *Biometrics*, **27**, 387-397.

Sneath, P. H. A. (1977). A method for testing the distinctness of clusters: A test of the disjunction of two clusters in Euclidean space as measured by their overlap, *Mathematical Geology*, **9**, 123-143.

Sokal, R. R. and Rohlf, F. J. (1962). The comparison of dendrograms by objective methods, *Taxon*, **11**, 33-40.

Sokal, R. R. and Sneath, P. H. A. (1963). *Principles of numerical taxonomy*, San Francisco: Freeman.

Thorndike, R. L. (1953). Who belongs in a family?, *Psychometrika*, 18, 267-276.

Wallace, C. S. and Boulton, D. M. (1968). An information measure for classification, *Comp. J.*, **11**, 185-194.

Ward, J. H. (1963). Hierarchical grouping to optimize an objective function, *J. Am. Statist. Ass.*, **58**, 236-244.

Wishart, D. (1969a). Numerical classification method for deriving natural classes, *Nature*, **221**, 97-98.

Wishart, D. (1969b). *Mode Analysis*, Numerical Taxonomy (A.J.Cole,ed.), 282-308, New York: Academic Press.

Wolfe, J. H. (1971). *A Monte Carlo study of the sampling distribution*, Naval Personel and Training Research Laboratory Technical Bulletin STB, 72-2, San Diego, California, USA.

Wolfe, J. H. (1970). Pattern Clustering by multivariate mixture analysis, *Multivariate Behavioral Research*, **5**, 329-350.

Wolfe, J. H. (1965). *A Computer Program for the maximum likelihood analysis of types*, Technical Bulletin, 65-15, U.S. Naval Personel Research Activity, San Diego.