ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ

Σχολή Χρηματοοικονομικής και Στατιστικής

Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης

ΜΕΤΑΠΤΥΧΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΣΠΟΥΔΩΝ
ΣΤΗΝ **ΕΦΑΡΜΟΣΜΕΝΗ ΣΤΑΤΙΣΤΙΚΗ**

ΣΥΓΚΡΙΣΗ ΤΩΝ ΤΕΧΝΙΚΩΝ ΚΑΘΟΡΙΣΜΟΥ ΤΟΥ ΠΛΗΘΟΥΣ ΟΜΑΔΩΝ ΣΕ ΣΥΝΟΛΑ ΠΟΛΥΔΙΑΣΤΑΤΩΝ ΔΕΔΟΜΕΝΩΝ

Αναστάσιος Ν. Γεωργίου

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
Σεπτέμβριος 2023

Η παρούσα Διπλωματική Εργασία εγκρίθηκε ομόφωνα από την Τριμελή Εξεταστική Επιτροπή που ορίστηκε 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:

- Καθηγητής Μ. Κούτρας (Επιβλέπων)
- Καθηγήτρια Γ. Βεροπούλου
- Αναπληρωτής καθηγητής Χ. Ευαγγελάρας

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

School of Finance and Statistics

Department of Statistics and Insurance Science

POSTGRADUATE PROGRAM IN
APPLIED STATISTICS

COMPARISON OF TECHNIQUES IDENTIFYING THE NUMBER OF CLUSTERS PRESENT IN MULTIVARIATE DATASETS

By

Anastasios N. Georgiou

MSc Dissertation

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

Piraeus, Greece
September 2023

Στους γονείς μου

Νίκο και Ράνια

Περίληψη

Η ανάλυση ομαδοποίησης είναι μια θεμελιώδης τεχνική στην επιστήμη των δεδομένων, με στόχο την αποκάλυψη εγγενών μοτίβων και σχέσεων μέσα σε πολύπλοκα σύνολα δεδομένων. Η παρούσα Διπλωματική Εργασία διερευνά και συγκρίνει διάφορα κριτήρια αξιολόγησης των τεχνικών ομαδοποίησης σε πολυδιάστατα σύνολα δεδομένων για τον εντοπισμό του βέλτιστου αριθμού συστάδων. Για την αξιολόγηση της σταθερότητας και της αποτελεσματικότητας των κριτηρίων αξιολόγησης χρησιμοποιούνται προσομοιωμένα δεδομένα με γνωστές δομές συστάδων. Διάφορα κριτήρια της βιβλιογραφίας όπως το μέτρο Silhouette και το κριτήριο Calinski Harabaz χρησιμοποιούνται για τη σύγκριση και πρόταση του βέλτιστου αριθμού συστάδων. Τα ευρήματα υπογραμμίζουν την ευαισθησία των κριτηρίων αξιολόγησης στο αρχικό επίπεδο διαχωρισμού των ομάδων μέσω της απόστασης αλλά και την ανάγκη για σωστή κάθε φορά επιλογή μεθόδου συσταδοποίησης τονίζοντας τη σημασία της επιλογής των κατάλληλων τεχνικών με βάση τα χαρακτηριστικά των δεδομένων.

Η παρούσα εργασία συνεισφέρει πολύτιμες γνώσεις για την επιλογή του βέλτιστου αριθμού συστάδων και την ανάδειξη των διαφορετικών χαρακτηριστικών των κριτηρίων, προσφέροντας μέσω παραδειγμάτων κάποια συμπεράσματα για το θέμα αυτό. Προτείνονται μελλοντικές κατευθύνσεις έρευνας για τη διερεύνηση υβριδικών προσεγγίσεων και την αντιμετώπιση προκλήσεων στη ομαδοποίηση δεδομένων μεγάλης κλίμακας.

Abstract

Clustering analysis is a fundamental technique in data science, aiming to uncover inherent patterns and relationships within complex datasets. This MSc Thesis investigates and compares various evaluation criteria of clustering techniques on multidimensional datasets to identify the optimal number of clusters. Simulated data with known cluster structures are exploited to evaluate the stability and effectiveness of each method. Criteria such as the Silhouette measure and Calinski Harabasz are used to compare and suggest the optimal number of clusters. The findings of our numerical experimentation highlight the sensitivity of clustering outcomes to the choice of method, emphasizing the significance of selecting the appropriate techniques based on data characteristics.

The Thesis contributes valuable insights into suggesting and selecting the optimal number of clusters and highlighting the different characteristics of criterias, offering through examples some conclusions on this subject. Finally, we offer guidance for method selection and validation. Future research directions are suggested to explore hybrid approaches, and address challenges in large-scale data clustering.

Περιεχόμενα

Κατάλογος Πινάκων

Κατάλογος Σχημάτων

ΚΕΦΑΛΑΙΟ 1

ΕΙΣΑΓΩΓΗ

Η μελέτη πολυδιάστατων δεδομένων καθιστά τη διάκριση ομοιοτήτων και διαφορών μεταξύ των παρατηρήσεων αρκετά δύσκολη και τις περισσότερες φορές αδύνατη ακόμα και για ειδικούς πάνω στο αντικείμενο ενδιαφέροντος.

Η ανάλυση συστάδων αποδεικνύεται ιδιαίτερα χρήσιμη σε τέτοιες περιπτώσεις, καθώς μέσω της χρήσης αλγορίθμων ομαδοποίησης μπορεί κανείς να αποκαλύψει μεγάλο μέρος της ομοιότητας μεταξύ των παρατηρήσεων οι οποίες είναι διαθέσιμες.

Η ανάλυση κατά συστάδες βρίσκει εφαρμογή σε ποικίλους τομείς και διάφορα επιστημονικά πεδία. Σημαντικά παραδείγματα αποτελούν η γεωπονία, η ιατρική και η ψυχολογία, αλλά εξίσου σημαντική είναι επίσης η συνεισφορά της στην ανάπτυξη της αναγνώρισης προτύπων και συνακολούθως στη μηχανική μάθηση και την εξόρυξη δεδομένων.

Ο προσδιορισμός του βέλτιστου αριθμού ομάδων είναι ένα δύσκολο πρόβλημα. Οι αλγόριθμοι ομαδοποίησης δεδομένων δε γνωρίζουν τον πραγματικό αριθμό των ομάδων για ένα σύνολο δεδομένων εκ των προτέρων. Συνεπώς, είναι απαραίτητο να χρησιμοποιηθεί κάποιο κριτήριο για τη διακοπή της λειτουργίας τους, το οποίο να είναι σε θέση να διακρίνει τον όσο το δυνατόν πιο αποδοτικό διαχωρισμό των ομάδων.

Τέτοιου είδους κριτήρια είναι συνήθως δύσκολο να αξιολογηθούν σε πραγματικά δεδομένα. Εξαιτίας αυτού, είθισται να γίνεται χρήση προσομοιωμένων δεδομένων, τα οποία έχουν παραχθεί από γνωστές και ευρέως κατανοητές κατανομές. Κατ' αυτό το τρόπο, ο αριθμός των ομάδων και ο τρόπος με τον οποίο έχουν κατανεμηθεί σε ένα πολυδιάστατο χώρο ενδιαφέροντος είναι πλήρως ελεγχόμενοι και κατανοήσιμοι.

1.1 Είδη δεδομένων

Είναι σημαντικό να ορίσουμε τους τύπους δεδομένων που μπορεί κανείς να συναντήσει προχωρώντας σε διαδικασίες ανάλυσης συστάδων. Πιο συγκεκριμένα, διακρίνουμε τις ακόλουθες περιπτώσεις:

- **Αριθμητικά (Numeric):** τα αριθμητικά δεδομένα χωρίζονται σε δύο υποκατηγορίες
    - **Διακριτά**: πεπερασμένο πλήθος τιμών (π.χ. ο αριθμός των ατόμων που είναι συνδρομητές σε κάποια υπηρεσία).

- **Συνεχή:** μη αριθμήσιμο πλήθος τιμών (π.χ. ο χρόνος διεκπεραίωσης μιας ιατρικής εξέτασης).

- **Ποιοτικά (Ordinal):** μη αριθμητικές τιμές που ακολουθούν κάποια ιεραρχία (π.χ. αξιώματα στο στράτευμα).
- **Κατηγορικά (Categorical):** μη αριθμητικές τιμές, χωρίς κάποια ιεραρχία. (π.χ. ασθενής ή όχι ασθενής). Όταν οι τιμές τέτοιων παρατηρήσεων είναι είτε 0 είτε 1, τότε τα δεδομένα ονομάζονται διχοτομικά (binary).

1.2 Μέτρα Απόστασης

Τα μέτρα απόστασης αποτελούν μία από τις βασικότερες έννοιες στην ανάλυση συστάδων. Εκφράζουν ειδικά κατασκευασμένες μετρικές οι οποίες μας δείχνουν πόσο όμοιες ή ανόμοιες είναι δύο παρατηρήσεις μεταξύ τους. Παρατηρήσεις που έχουν πολλά κοινά τείνουν να έχουν σημαντικά μικρότερη απόσταση από αυτές που είναι πολύ διαφορετικές. Βρίσκουν εφαρμογή ανάλογα με το τύπο δεδομένων που μπορούν να διαχειριστούν. Συνεπώς, είναι πολύ σημαντικό να γίνεται εκτενής μελέτη πριν εφαρμοστεί κάποιο μέτρο, καθώς αν η επιλογή είναι εσφαλμένη ενδέχεται να οδηγήσει σε λανθασμένα ή παραπλανητικά αποτελέσματα.

Παρακάτω δίνονται κάποια από τα πιο σημαντικά και ευρέως χρησιμοποιούμενα μέτρα απόστασης.

- **Ευκλείδεια Απόσταση**

  Τα διανύσματα $\boldsymbol{x}=(x_1, x_2, \ldots, x_\nu)$ και $\boldsymbol{y}=(y_1, y_2, \ldots, y_\nu)$ αντιπροσωπεύουν σύνολα τιμών ή παρατηρήσεων σε ένα ν-διάστατο χώρο. Κάθε διάνυσμα αποτελείται από ν συνιστώσες, όπου $x_1, x_2, \ldots, x_\nu$ είναι οι συνιστώσες του διανύσματος $\boldsymbol{x}$, και $y_1, y_2, \ldots, y_\nu$ είναι οι συνιστώσες του διανύσματος $\boldsymbol{y}$.

  Αυτά τα διανύσματα μπορούν να χρησιμοποιηθούν για την αναπαράσταση διαφόρων τύπων δεδομένων, ανάλογα με το περιβάλλον. Για παράδειγμα, στη μηχανική μάθηση, τα στοιχεία των διανυσμάτων θα μπορούσαν να αντιπροσωπεύουν χαρακτηριστικά ή χαρακτηριστικά ενός συνόλου δεδομένων και κάθε διάνυσμα θα αντιστοιχεί σε ένα συγκεκριμένο σημείο δεδομένων ή παράδειγμα.

  Οι τιμές των στοιχείων $x_1, x_2, \ldots, x_\nu$ και $y_1, y_2, \ldots, y_\nu$ μπορεί να διαφέρουν ανάλογα με τη συγκεκριμένη εφαρμογή ή τομέα. Θα μπορούσαν να αντιπροσωπεύουν αριθμητικές μετρήσεις, κατηγορικές μεταβλητές ή ακόμα και δυαδικές τιμές. Η ερμηνεία αυτών των στοιχείων εξαρτάται από το πλαίσιο στο οποίο χρησιμοποιούνται.

  Ίσως το πιο σημαντικό και συχνά χρησιμοποιούμενο μέτρο απόστασης μεταξύ δύο παρατηρήσεων είναι η ευκλείδεια απόσταση και δίνεται από τον ακόλουθο τύπο.

$$d(x, y) = \sqrt{\sum_{i=1}^{n} (y_i - x_i)^2} \qquad (1.1)$$

Η Ευκλείδεια απόσταση έχει πλεονέκτημα όταν εφαρμοστεί σε δεδομένα χαμηλών διαστάσεων, όμως χωλαίνει όταν χρησιμοποιείται σε πολυδιάστατα δεδομένα εξαιτίας της **κατάρας της διάστασης** (curse of dimensionality).

Επίσης, δεν διαχειρίζεται καλά παρατηρήσεις οι οποίες βρίσκονται σε διαφορετικές κλίμακες. Απαιτείται λοιπόν συχνά να εφαρμόζεται κανονικοποίηση των δεδομένων προτού εφαρμοστεί σε αυτά.

- **Απόσταση Manhattan**

Η απόσταση του Μανχάταν, επίσης γνωστή ως απόσταση city block ή απόσταση taxicab, είναι μια μέτρηση απόστασης που χρησιμοποιείται συνήθως στα μαθηματικά, την επιστήμη των υπολογιστών και την ανάλυση δεδομένων. Είναι ένα μέτρο της απόστασης μεταξύ δύο σημείων σε ένα χώρο, που υπολογίζεται αθροίζοντας τις απόλυτες διαφορές των συντεταγμένων τους.

Σε αντίθεση με την Ευκλείδεια απόσταση, η οποία μετρά την ευθεία απόσταση μεταξύ δύο σημείων, η απόσταση του Μανχάταν λαμβάνει υπόψη μόνο τις οριζόντιες και κάθετες κινήσεις. Πήρε το όνομά της από τη διάταξη των δρόμων στο Μανχάταν, όπου μπορεί κανείς να ταξιδέψει μόνο κατά μήκος των τετραγώνων της πόλης.

Τυπικά, η απόσταση του Μανχάταν μεταξύ δύο σημείων A(x_1 ,y_1) και B(x_2 ,y_2) σε ένα δισδιάστατο χώρο δίνεται από τον ακόλουθο τύπο:

$$D(x, y) = \sum_{i=1}^{n} |x_i - y_i| \qquad (1.2)$$

Ένα αξιοσημείωτο χαρακτηριστικό της απόστασης του Μανχάταν είναι ότι δεν είναι ευαίσθητη στην κλίμακα των επιμέρους συντεταγμένων. Αντιμετωπίζει κάθε διάσταση ανεξάρτητα και εξίσου, καθιστώντας την κατάλληλη για σύγκριση διανυσμάτων που περιέχουν πραγματικές τιμές που θα μπορούσαν να μετρηθούν σε διαφορετικές κλίμακες.

Η απόσταση του Μανχάταν έχει διάφορες εφαρμογές, όπως:

1. Αλγόριθμοι ομαδοποίησης: Χρησιμοποιείται συνήθως σε αλγόριθμους ομαδοποίησης όπως το *k*-means, όπου η απόσταση μεταξύ των σημείων υπολογίζεται για να τα εκχωρήσει σε συστάδες με βάση την εγγύτητά τους.
2. Επεξεργασία εικόνας: Μπορεί να χρησιμοποιηθεί για τη μέτρηση της ομοιότητας μεταξύ των εικόνων συγκρίνοντας τις εντάσεις των pixel τους.
3. Σχεδιασμός διαδρομής: Στα συστήματα πλοήγησης, η απόσταση του Μανχάταν μπορεί να βοηθήσει στον προσδιορισμό της συντομότερης διαδρομής μεταξύ δύο σημείων σε ένα πλέγμα πόλης, όπου επιτρέπονται μόνο κάθετες και οριζόντιες κινήσεις.

- **Απόσταση Chebyshev**

Η απόσταση Chebyshev, επίσης γνωστή ως η απόσταση σκακιέρας ή ο κανόνας του απείρου, είναι μια μέτρηση απόστασης που χρησιμοποιείται για τη μέτρηση της ανομοιότητας ή της απόστασης μεταξύ δύο σημείων σε ένα χώρο. Πήρε το όνομά του από τον Ρώσο μαθηματικό Pafnuty Chebyshev.

Η απόσταση Chebyshev ορίζεται ως η μέγιστη απόλυτη διαφορά μεταξύ των συντεταγμένων των δύο σημείων σε όλες τις διαστάσεις. Με άλλα λόγια, αντιπροσωπεύει τη μεγαλύτερη κάθετη ή οριζόντια απόσταση μεταξύ των σημείων με πλέγμα.

Τυπικά, για δύο σημεία $A(x_1, y_1)$ και $B(x_2, y_2)$σε ένα δισδιάστατο χώρο, η απόσταση Chebyshev δίνεται από τον ακόλουθο τύπο:

$$d(A,B) = \max(|x_2 - x_1|, |y_2 - y_1|) \quad (1.3)$$

Ο τύπος υπολογίζει τις απόλυτες διαφορές μεταξύ των συντεταγμένων x και των συντεταγμένων y των σημείων και επιλέγει τη μεγαλύτερη από τις δύο διαφορές ως απόσταση. Μπορεί να ερμηνευθεί ως ο αριθμός των κινήσεων που θα έκανε ένα βασιλικό κομμάτι σκακιού για να μετακινηθεί από το σημείο A στο σημείο B σε μια σκακιέρα, όπου επιτρέπονται οι διαγώνιες κινήσεις.

Η απόσταση Chebyshev μπορεί επίσης να επεκταθεί σε χώρους υψηλότερων διαστάσεων. Σε έναν n-διάστατο χώρο, η απόσταση Chebyshev μεταξύ δύο σημείων A και B με n συντεταγμένες μπορεί να υπολογιστεί ως:

$$d(A,B) = \max(|x_1 - x_2|, |y_1 - y_2|, \dots, |x_n - y_n|) \quad (1.4)$$

Κάθε απόλυτη διαφορά μεταξύ των αντίστοιχων συντεταγμένων υπολογίζεται και η μέγιστη διαφορά σε όλες τις διαστάσεις επιλέγεται ως απόσταση.

Η απόσταση Chebyshev έχει διάφορες εφαρμογές, όπως:

- Αλγόριθμοι εύρεσης μονοπατιών: Σε αλγόριθμους εύρεσης μονοπατιών που βασίζονται σε πλέγμα, όπως αυτοί που χρησιμοποιούνται στη ρομποτική ή στην ανάπτυξη παιχνιδιών, η απόσταση Chebyshev μπορεί να χρησιμοποιηθεί για την εκτίμηση της συντομότερης διαδρομής μεταξύ δύο σημείων όταν επιτρέπεται η διαγώνια κίνηση.
- Επεξεργασία εικόνας: Μπορεί να χρησιμοποιηθεί για τη μέτρηση της ανομοιότητας μεταξύ δύο εικόνων συγκρίνοντας τη μέγιστη διαφορά στις τιμές των pixel σε όλες τις αντίστοιχες θέσεις.

- Ανίχνευση ακραίων τιμών: Η απόσταση Chebyshev μπορεί να βοηθήσει στον εντοπισμό ακραίων τιμών σε ένα σύνολο δεδομένων μετρώντας τη μέγιστη απόκλιση ενός σημείου δεδομένων από το κέντρο ή τη μέση τιμή.
- Συνοπτικά, η απόσταση Chebyshev είναι μια μέτρηση απόστασης που υπολογίζει τη μέγιστη απόλυτη διαφορά μεταξύ των συντεταγμένων σε ένα διάστημα. Είναι ιδιαίτερα χρήσιμο σε σενάρια όπου επιτρέπεται η διαγώνια κίνηση ή όταν συγκρίνει ο χρήστος την ανομοιότητα μεταξύ σημείων δεδομένων σε διάταξη σαν πλέγμα ή σαν σκακιέρα.

- **Απόσταση Minkowski**

Η απόσταση Minkowski είναι μια γενική μέτρηση απόστασης που περιλαμβάνει τόσο την Ευκλείδεια απόσταση όσο και την απόσταση του Μανχάταν ως ειδικές περιπτώσεις. Είναι ένα μέτρο της ανομοιότητας ή της απόστασης μεταξύ δύο σημείων σε ένα διάστημα και πήρε το όνομά του από τον Ρώσο μαθηματικό Χέρμαν Μινκόφσκι.

Η απόσταση Minkowski ορίζεται ως εξής:

$$D(x, y) = (\sum_{i=1}^{n} |x_i - y_i|^p)^{\frac{1}{p}} \quad (1.5)$$

Για διαφορετικές τιμές του p προκύπτουν οι εξής αποστάσεις

- $p = 1$ – $Manhattan$
- $p = 2$ – $Ευκλείδεια$
- $p = \infty$ – $Chebyshev$

Η απόσταση Minkowski είναι μια ευέλικτη μέτρηση απόστασης που χρησιμοποιείται σε διάφορους τομείς, όπως:

- Ομαδοποίηση: Χρησιμοποιείται σε αλγόριθμους ομαδοποίησης, όπως ο αλγόριθμος K-means, για τη μέτρηση της ανομοιότητας μεταξύ σημείων δεδομένων και την ανάθεση τους σε συστάδες με βάση την εγγύτητά τους.
- Επιλογή χαρακτηριστικών: Στις τεχνικές επιλογής χαρακτηριστικών, η απόσταση Minkowski μπορεί να χρησιμοποιηθεί για την αξιολόγηση της συνάφειας ή της σημασίας διαφορετικών χαρακτηριστικών συγκρίνοντας τις αποστάσεις τους με τη μεταβλητή στόχο.
- Μηχανική μάθηση: Η απόσταση Minkowski μπορεί να χρησιμοποιηθεί ως μέτρο απόστασης σε αλγόριθμους μηχανικής μάθησης, συμπεριλαμβανομένων των k-πλησιέστερων γειτόνων (k-NN), όπου βοηθά στον προσδιορισμό των πλησιέστερων γειτόνων ενός δεδομένου σημείου δεδομένων.
- Η επιλογή της παραμέτρου p στην απόσταση Minkowski εξαρτάται από το συγκεκριμένο πρόβλημα και τα χαρακτηριστικά των δεδομένων.

Προσαρμόζοντας την τιμή του p, μπορεί κανείς να προσαρμόσει την απόσταση Minkowski για να ταιριάζει στις απαιτήσεις της εργασίας ανάλυσης ή μοντελοποίησης.

- **Ομοίοτητα συνημιτόνου (Cosine Similarity)**

Η απόσταση συνημιτόνου είναι ένα μέτρο ομοιότητας που χρησιμοποιείται συνήθως για να ποσοτικοποιήσει την ανομοιότητα μεταξύ δύο διανυσμάτων σε ένα χώρο υψηλών διαστάσεων. Προέρχεται από την έννοια της ομοιότητας συνημιτόνου, η οποία μετρά το συνημίτονο της γωνίας μεταξύ δύο διανυσμάτων.

Η απόσταση συνημιτόνου μεταξύ δύο διανυσμάτων, που συμβολίζονται ως A και B, υπολογίζεται ως:

$$d(A,B) = 1 - \left(cosine_{similarity(A,B)}\right) \tag{1.6}$$

Η ομοιότητα συνημιτόνου δύο διανυσμάτων υπολογίζεται παίρνοντας το γινόμενο των τελειών των διανυσμάτων και διαιρώντας το με το γινόμενο των μεγεθών τους:

$$cosine_{similarity(A,B)} = \frac{(A \cdot B)}{(||A|| * ||B||)} \tag{1.7}$$

Εδώ, το A · B αντιπροσωπεύει το εσωτερικό γινόμενο των διανυσμάτων ***A*** και ***B***, και ||A|| και ||B|| αντιπροσωπεύουν τα μεγέθη (ή τους Ευκλείδειους κανόνες) των διανυσμάτων ***A*** και ***B***, αντίστοιχα.

Η προκύπτουσα τιμή ομοιότητας συνημιτόνου κυμαίνεται μεταξύ -1 και 1. Η τιμή 1 υποδηλώνει ότι τα διανύσματα έχουν την ίδια κατεύθυνση (είναι πανομοιότυπα), ενώ η τιμή -1 υποδεικνύει αντίθετες κατευθύνσεις. Η τιμή 0 δείχνει ότι τα διανύσματα είναι ορθογώνια (κάθετα) μεταξύ τους.

Για να ληφθεί υπόψη η ομοιότητα του συνημιτόνου αφαιρείται από το 1. Αυτή η μετατροπή επιτρέπει στο μέτρο της απόστασης να κυμαίνεται από 0 έως 2, όπου το 0 αντιπροσωπεύει πανομοιότυπα διανύσματα και το 2 αντιπροσωπεύει διανύσματα που είναι στο μέγιστο ανόμοια ή αντίθετα.

Η ομοιότητα συνημιτόνου χρησιμοποιείται συχνά σε εργασίες εξόρυξης κειμένου, ομαδοποίησης εγγράφων και ανάκτησης πληροφοριών. Στην επεξεργασία φυσικής γλώσσας, για παράδειγμα, τα έγγραφα τυπικά αναπαρίστανται ως διανύσματα υψηλών διαστάσεων χρησιμοποιώντας τεχνικές όπως το TF-IDF ή ενσωματώσεις λέξεων. Η ομοιότητα συνημιτόνου μπορεί στη συνέχεια να χρησιμοποιηθεί για τη μέτρηση της ανομοιότητας μεταξύ των εγγράφων με βάση τις διανυσματικές αναπαραστάσεις τους.

Ένα πλεονέκτημα της ομοιότητας συνημιτόνου είναι η μη ευαισθησία της στο μέγεθος ή την κλίμακα των διανυσμάτων. Εστιάζει στην κατεύθυνση ή τον προσανατολισμό των διανυσμάτων παρά στα μεγέθη των μεμονωμένων συστάδων τους. Αυτή η ιδιότητα την καθιστά ιδιαίτερα χρήσιμη κατά τη σύγκριση εγγράφων ή αποσπασμάτων κειμένου που μπορεί να έχουν διαφορετικό μήκος ή συχνότητα.

- **Απόσταση Mahalanobis**

Η απόσταση Mahalanobis είναι ένα μέτρο της ανομοιότητας ή της απόστασης μεταξύ δύο σημείων σε έναν πολυμεταβλητό χώρο. Λαμβάνει υπόψη τη δομή συσχέτισης των δεδομένων και είναι ιδιαίτερα χρήσιμο όταν ασχολείται με σύνολα δεδομένων που έχουν συσχετισμένες μεταβλητές.

Η απόσταση Mahalanobis θεωρεί τον πίνακα συνδιακύμανσης των δεδομένων για τον υπολογισμό της απόστασης μεταξύ δύο σημείων. Μπορεί να θεωρηθεί ως ένα μέτρο του πόσες τυπικές αποκλίσεις απέχει ένα σημείο από ένα άλλο, λαμβάνοντας υπόψη τη δομή συνδιακύμανσης των μεταβλητών.

Δεδομένων δύο σημείων δεδομένων ***A*** και ***B*** με διανύσματα χαρακτηριστικών διαστάσεων **p**, η απόσταση Mahalanobis μεταξύ τους ορίζεται ως:

$$d(A,B) = \sqrt{(A - B)' * C^{-1} * (A - B)} \qquad (1.8)$$

Σε αυτόν τον τύπο, το $A - B$ αντιπροσωπεύει το διάνυσμα διαφοράς μεταξύ των σημείων ***A*** και ***B***, το C^{-1} είναι ο αντίστροφος του πίνακα συνδιακύμανσης C και το *(A - B)'* υποδηλώνει τον ανάστροφο του διανύσματος διαφοράς.

Η απόσταση Mahalanobis εξετάζει το σχήμα και τον προσανατολισμό της κατανομής δεδομένων ενσωματώνοντας τον πίνακα αντίστροφης συνδιακύμανσης. Αντιπροσωπεύει τις συσχετίσεις μεταξύ των μεταβλητών, οι οποίες μπορούν να οδηγήσουν σε ακριβέστερους υπολογισμούς απόστασης σε χώρους υψηλών διαστάσεων σε σύγκριση με τη χρήση της Ευκλείδειας απόστασης.

Χρησιμοποιώντας την απόσταση Mahalanobis, μπορεί κανείς να μετρήσει την ανομοιότητα μεταξύ των σημείων, ενώ υπολογίζει τις ποικίλες κλίμακες και τις συσχετίσεις των μεταβλητών. Αυτό είναι ιδιαίτερα χρήσιμο σε εφαρμογές όπως η ανίχνευση ακραίων τιμών, η ομαδοποίηση, η ταξινόμηση και η αναγνώριση προτύπων.

Για τον υπολογισμό της απόστασης Mahalanobis, χρειάζεται να εκτιμηθεί ο πίνακας συνδιακύμανσης από το σύνολο δεδομένων. Σε περιπτώσεις όπου το σύνολο δεδομένων έχει περιορισμένα δείγματα ή μεταβλητές, μπορούν να χρησιμοποιηθούν τεχνικές τακτοποίησης ή άλλες μέθοδοι για να εξασφαλιστεί μια καλά ρυθμισμένη και αξιόπιστη εκτίμηση του πίνακα συνδιακύμανσης.

1.3 Κατηγορίες μεθόδων ομαδοποίησης

Οι τεχνικές ομαδοποίησης διακρίνονται στις εξής κύριες κατηγορίες: **ιεραρχικές** (hierarchical), **μη ιεραρχικές** (centroid based), **πυκνότητας** (density based) και **βασισμένες σε κατανομές** (distribution based).

Οι ιεραρχικές αποτελούνται από τις **συσσωρευτικές** (agglomerative) και τις **διαιρετικές** (divisive). Οι πρώτες χρησιμοποιούν μια "**bottom up"** προσέγγιση, εκκινώντας από κάθε παρατήρηση ως μία ξεχωριστή ομάδα και σταδιακά ενώνοντας ζεύγη παρατηρήσεων. Αντίθετα, οι διαιρετικές μέθοδοι ακολουθούν μια **"top down"** λογική, εκκινώντας με όλες τις παρατηρήσεις ως μία ομάδα και αναδρομικά χωρίζοντάς τις σε μικρότερες.

Οι αλγόριθμοι ιεραρχικής ομαδοποίησης δε χρησιμοποιούν κάποιο προκαθορισμένο αριθμό ομάδων, αλλά κάνουν χρήση ενός εσωτερικού πίνακα ομοιότητας. Οι αλγόριθμοι ιεραρχικής ομαδοποίησης είναι μια κατηγορία αλγορίθμων ομαδοποίησης που δεν απαιτούν έναν προκαθορισμένο αριθμό ομάδων ή συστάδων ως είσοδο. Αντίθετα, χτίζουν μια ιεραρχία συστάδων συγχωνεύοντας ή διαχωρίζοντας επαναληπτικά συστάδες με βάση την ομοιότητά τους.

Οι αλγόριθμοι ξεκινούν αντιμετωπίζοντας κάθε σημείο δεδομένων ως μεμονωμένη συστάδα. Στη συνέχεια, υπολογίζουν ένα μέτρο ομοιότητας ή ανομοιότητας μεταξύ όλων των ζευγών των συστάδων. Αυτό γίνεται συνήθως χρησιμοποιώντας μια μέτρηση απόστασης όπως η Ευκλείδεια απόσταση ή η ομοιότητα συνημιτόνου.

Με βάση το μέτρο ομοιότητας, ο αλγόριθμος προσδιορίζει τις δύο ομοιότερες συστάδες και τις συγχωνεύει σε μια ενιαία συστάδα. Αυτή η διαδικασία συνεχίζεται επαναληπτικά, σχηματίζοντας σταδιακά μια ιεραρχία ή μια δομή που μοιάζει με δέντρο από συστάδες. Το βήμα συγχώνευσης επαναλαμβάνεται έως ότου όλα τα σημεία δεδομένων ομαδοποιηθούν σε μία ενιαία συστάδα ή μέχρι να ικανοποιηθεί ένα κριτήριο διακοπής.

Η ιεραρχία των συστάδων μπορεί να αναπαρασταθεί χρησιμοποιώντας ένα δενδρόγραμμα, το οποίο είναι μια οπτική αναπαράσταση της διαδικασίας συγχώνευσης. Το δενδρόγραμμα δείχνει τη σειρά με την οποία συγχωνεύτηκαν οι συστάδες και την ομοιότητα με την οποία συνέβη κάθε συγχώνευση.

Για να καθοριστεί ποιες συστάδες θα συγχωνευθούν σε κάθε βήμα, οι αλγόριθμοι ιεραρχικής ομαδοποίησης βασίζονται στον πίνακα εσωτερικής ομοιότητας, γνωστό και ως πίνακας εγγύτητας ή πίνακας απόστασης. Αυτός ο πίνακας περιέχει τις ομοιότητες ή ανομοιότητες ανά ζεύγη μεταξύ όλων των σημείων δεδομένων ή των συστάδων.

Ο πίνακας εσωτερικής ομοιότητας επιτρέπει στον αλγόριθμο να αξιολογήσει την ομοιότητα μεταξύ διαφορετικών συστάδων σε κάθε στάδιο. Η επιλογή του μέτρου ομοιότητας και η μέθοδος που χρησιμοποιείται για τον υπολογισμό του πίνακα μπορεί να ποικίλλει ανάλογα με τον συγκεκριμένο αλγόριθμο ιεραρχικής ομαδοποίησης που χρησιμοποιείται.

Χρησιμοποιώντας τον πίνακα εσωτερικής ομοιότητας, οι αλγόριθμοι ιεραρχικής ομαδοποίησης μπορούν να συλλάβουν την υποκείμενη δομή των δεδομένων με συγκεντρωτικό (από κάτω προς τα πάνω) ή διαιρετικό (από πάνω προς τα κάτω). Η προκύπτουσα ιεραρχία παρέχει μια ευέλικτη αναπαράσταση που επιτρέπει διαφορετικά επίπεδα ευαισθησίας στη λύση ομαδοποίησης. Επιτρέπει την εξερεύνηση διαφορετικών διαμορφώσεων συμπλέγματος χωρίς να απαιτείται εκ των προτέρων καθορισμός του αριθμού των συστάδων.

Η επιλογή του αριθμού των συστάδων καθορίζεται ερμηνεύοντας το δενδρόγραμμα ή χρησιμοποιώντας ένα σημείο αποκοπής στην κλίμακα ομοιότητας. Αυτό επιτρέπει στον αλγόριθμο να προσαρμόζεται στα δεδομένα και να αναγνωρίζει συστάδες διαφορετικών μεγεθών και σχημάτων. Ως απόρροια όλων αυτών, προκύπτει ότι οι αλγόριθμοι αυτοί απαιτούν σημαντικούς υπολογιστικούς πόρους για μεγάλα σύνολα δεδομένων.

Τα τελευταία χρόνια έχουν κατασκευαστεί πακέτα λογισμικού σε γλώσσες χαμηλού επιπέδου, οι οποίες κάνουν χρήση πολυπύρηνων επεξεργαστών και αποδοτικών υλοποιήσεων. Συνέπεια αυτού είναι η σημαντική μείωση του χρόνου εκτέλεσης αυτής της οικογένειας αλγορίθμων.

Οι μη ιεραρχικές μέθοδοι αντίστοιχα έχουν ικανοποιητική απόδοση σε μεγάλα σύνολα δεδομένων, δημιουργούν ομάδες παραπλήσιου μεγέθους και είναι συνήθως πιο απλοί. Εξαρτώνται όμως σε μεγάλο βαθμό από την επιλογή του αριθμού των ομάδων πριν εκτελεστεί ο αλγόριθμος συσταδοποίησης. Είναι επίσης ευαίσθητοι σε ακραίες τιμές (outliers) και δεν είναι πάντα σε θέση να εντοπίσουν μοτίβα που δεν είναι γραμμικά διαχωρίσιμα.

Οι μέθοδοι πυκνότητας ανιχνεύουν περιοχές στις οποίες η συγκέντρωση παρατηρήσεων είναι υψηλή. Οι περιοχές αυτές ενώνονται μεταξύ τους εφόσον αυτό είναι εφικτό εντοπίζοντας μοτίβα αυθαίρετου σχήματος. Αντιμετωπίζουν προβλήματα σε δεδομένα που παρουσιάζουν ποικίλα επίπεδα πυκνότητας και σε υψηλές διαστάσεις.

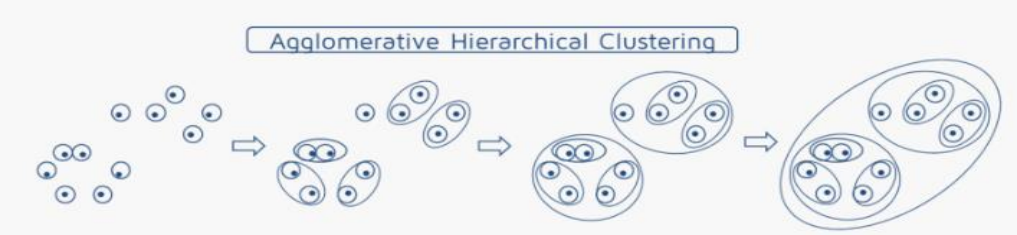
Τέλος, οι αλγόριθμοι που βασίζονται σε κατανομές κάνουν την υπόθεση ότι τα δεδομένα ακολουθούν κάποια γνωστή κατανομή. Όσο η απόσταση από το κέντρο μιας κατανομής αυξάνεται, τόσο η πιθανότητα μια παρατήρηση να ανήκει σε αυτή μειώνεται. Τέτοιου είδους αλγόριθμοι είναι καλό να αποφεύγονται όταν δεν υπάρχει πληροφόρηση σχετικά με την κατανομή που ακολουθούν τα δεδομένα.

1.3.1 Ιεραρχικές Μέθοδοι

1.3.1.1 Συσσωρευτικές Μέθοδοι

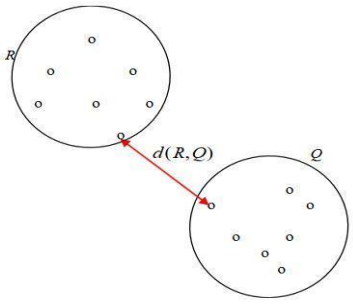
Οι συσσωρευτικές μέθοδοι ξεκινούν με *n* ομάδες και καταλήγουν σε μία μεγάλη ομάδα με διαδοχικές συγχωνεύσεις. Χρησιμοποιούν αποστάσεις ομοιότητας ανά ζεύγος παρατηρήσεων δημιουργώντας έτσι ένα πίνακα αποστάσεων. Βάσει αυτού του πίνακα, δημιουργείται ένα δενδρόγραμμα που αναπαριστά τις συγχωνεύσεις των αντικειμένων έως ότου σχηματιστεί μία

τελική ομάδα. Τα κριτήρια με τα οποία καθορίζεται η ομοιότητα μεταξύ ομάδων είναι και η ειδοποιός διαφορά των μεθόδων αυτής της οικογένειας.



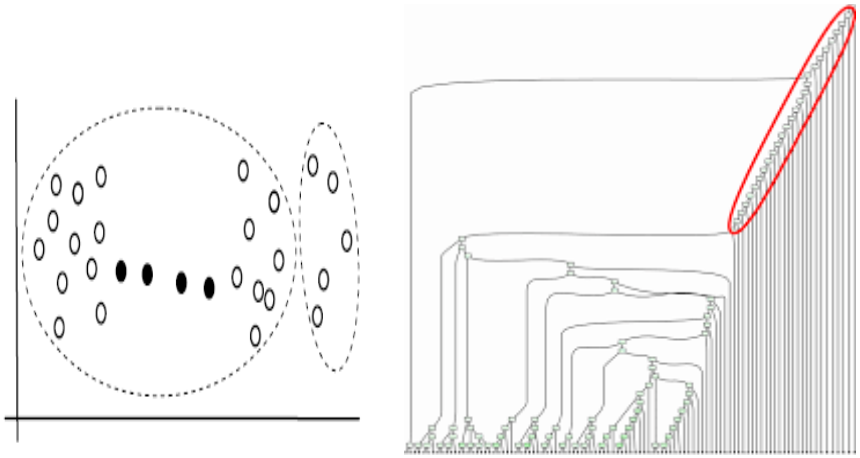
Σχήμα 1.1 Γραφική απεικόνιση συσσωρευτικών μεθόδων

Μέθοδος απλής συνένωσης (Single Linkage): Σε αυτή τη μέθοδο επιλέγονται οι παρατηρήσεις με τη μικρότερη απόσταση. Κάθε συνένωση οδηγεί στη μείωση των ομάδων κατά μία, και η απόσταση μεταξύ των ομάδων μετριέται ως η μικρότερη απόσταση μεταξύ μιας παρατήρησης της μίας από κάποια παρατήρηση της άλλης ομάδας. Βάσει αυτής της μέτρησης κρίνεται και η ομοιότητα μεταξύ δύο ομάδων.



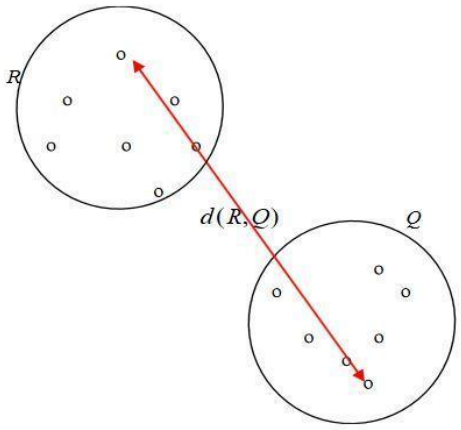
Σχήμα 2.2 Γραφική απεικόνιση μεθόδου απλής συνένωσης

Η μέθοδος αυτή υστερεί συγκριτικά με άλλες μεθόδους, καθώς έχει τη τάση να μη δημιουργεί νέες ομάδες όταν συναντά απομακρυσμένα σημεία, αλλά να τα ενώνει με ήδη υπάρχουσες. Αυτό το φαινόμενο ονομάζεται *chaining*, και οδηγεί σε κακώς διαχωρισμένες ομάδες οι οποίες περιέχουν σημεία που διαφέρουν σημαντικά (βλ. εικόνα 3).



Σχήμα 3.4 Δημιουργία κακώς διαχωρισμένων ομάδων με τη μέθοδο απλής συνένωσης

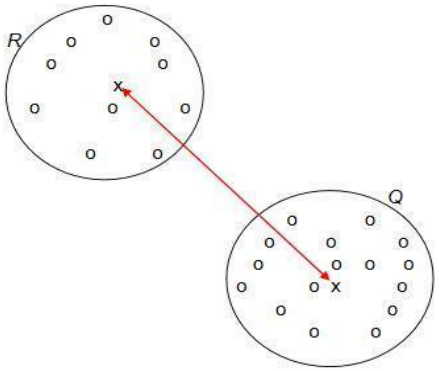
Μέθοδος πλήρους συνένωσης (Complete Linkage Method): Η απόσταση που επιλέγεται σε αυτή τη περίπτωση είναι αυτή των δύο πιο απομακρυσμένων παρατηρήσεων κάθε ομάδας. Η λογική αυτού του αλγορίθμου είναι η ακριβώς αντίθετη από αυτή του **Single Linkage**, δημιουργώντας μεγάλες ομάδες που αποτελούνται από πολλές παρατηρήσεις. Το μειονέκτημα που προκύπτει από αυτή τη στρατηγική είναι η αδυναμία της μεθόδου να εντοπίσει μικρότερες ομάδες.



Σχήμα 1.4 Γραφική απεικόνιση μεθόδου πλήρους συνένωσης

Μέθοδος των μέσων (Average of all pairs): Η απόσταση μεταξύ των ομάδων ορίζεται ως ο μέσος των αποστάσεων όλων των στοιχείων της μιας ομάδας με τα στοιχεία της άλλης.

Μέθοδος των κέντρων (Centroid method): Η μέθοδος αυτή υπολογίζει την απόσταση μεταξύ των κέντρων των ομάδων. Η ελάχιστη απόσταση θεωρείται κριτήριο συνένωσης των ομάδων. Ένας περιορισμός της κεντροειδούς μεθόδου είναι ότι είναι ευαίσθητη στην αρχική τοποθέτηση των κεντροειδών. Διαφορετικές αρχικοποιήσεις μπορεί να οδηγήσουν σε διαφορετικές τελικές ομαδοποιήσεις. Για να μετριαστεί αυτό, μπορούν να χρησιμοποιηθούν πολλαπλές εκτελέσεις του αλγορίθμου με διαφορετικές αρχικοποιήσεις ή προηγμένες τεχνικές αρχικοποίησης, όπως k-means++. Η μέθοδος των κέντρων παράγει συνήθως συμπαγείς και ελλειπτικές ομάδες.



Σχήμα 1.5 Γραφική απεικόνιση μεθόδου του κέντρου

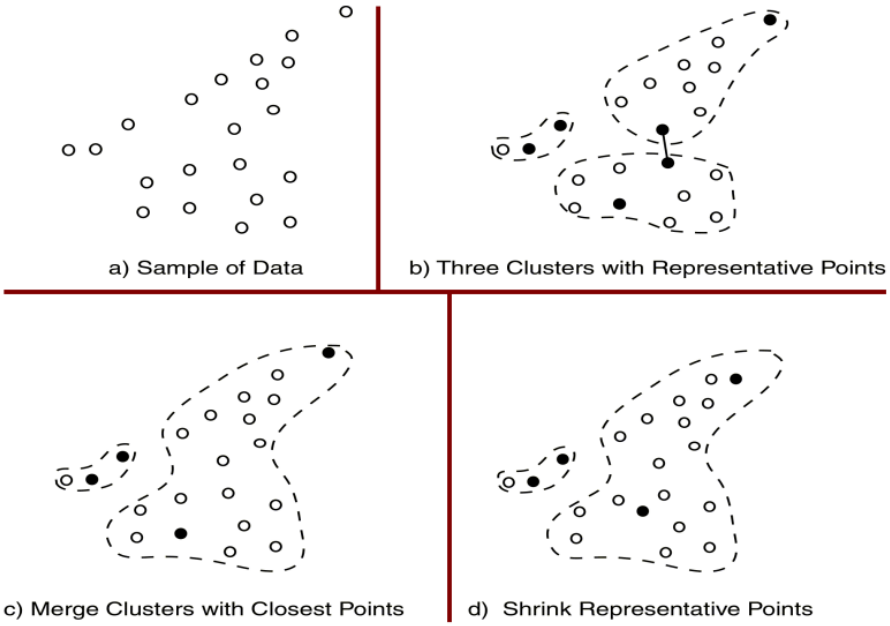
Μέθοδος του Ward: Η μέθοδος Ward είναι ένας ιεραρχικός αθροιστικός αλγόριθμος ομαδοποίησης που στοχεύει στην ελαχιστοποίηση της διακύμανσης εντός του συμπλέγματος κατά τη συγχώνευση συστάδων. Χρησιμοποιείται συνήθως στην ιεραρχική ομαδοποίηση για την κατασκευή ενός δενδρογράμματος, το οποίο αντιπροσωπεύει τη δομή ομαδοποίησης των δεδομένων.

Η μέθοδος Ward ξεκινά θεωρώντας κάθε σημείο δεδομένων ως μεμονωμένη συστάδα. Στη συνέχεια, σε κάθε επανάληψη, συγχωνεύει τις δύο συστάδες που οδηγούν στην ελάχιστη αύξηση της συνολικής διακύμανσης εντός του συμπλέγματος.

Η διακύμανση εντός συστάδας, γνωστή και ως σφάλμα αθροίσματος τετραγώνων (SSE), μετρά τη μεταβλητότητα ή την ανομοιότητα των σημείων δεδομένων σε κάθε συστάδα. Προσδιορίζει πόσο διαφέρουν τα σημεία δεδομένων σε ένα συστάδα από το κέντρο αυτού του συμπλέγματος. Η μέθοδος Ward επιδιώκει να ελαχιστοποιήσει την αύξηση της SSE κατά τη συγχώνευση συστάδων.

Μέθοδος CURE: Η μέθοδος CURE (Clustering Using REpresentatives) είναι ένας αποδοτικός αλγόριθμος που ανήκει στις συσσωρευτικές μεθόδους. Σε αντίθεση με τους περισσότερους αλγόριθμους αυτής της κατηγορίας, η τακτική του CURE είναι να διατηρεί ένα σταθερό αριθμό *c* από καλά διαχωρισμένες αντιπροσωπευτικές παρατηρήσεις, αντί ενός κέντρου. Οι αντιπρόσωποι επιλέγονται ώστε να είναι μακρινά σημεία από το κέντρο βάρους, αλλά και απομακρυσμένοι μεταξύ τους. Ο αλγόριθμος λειτουργεί κάνοντας χρήση της Ευκλείδειας απόστασης και επιτρέπει την εύρεση ομάδων οι οποίες έχουν αυθαίρετα σχήματα.

Τα βήματα που ακολουθεί είναι αρχικά να επιλέξει ένα τυχαίο δείγμα παρατηρήσεων που χωρούν στην κύρια μνήμη, και στη συνέχεια να τα ομαδοποιήσει μερικώς. Οι ομαδοποιήσεις γίνονται εξετάζοντας τη μικρότερη απόσταση μεταξύ ζευγών αντιπροσώπων. Από το πρώτο αποτέλεσμα της ομαδοποίησης αφαιρούνται οι έκτοπες τιμές και τα εναπομείναντα σημεία ομαδοποιούνται ξανά. Το δεύτερο αποτέλεσμα αποθηκεύεται σε κάποια συσκευή non-volatile μνήμης (π.χ. σκληρός δίσκος).



Σχήμα 1.6 Βήματα μεθόδου CURE

Οι παρατηρήσεις που δεν ανήκουν στο δείγμα που επιλέχθηκε, ομαδοποιούνται κατευθείαν στη συσκευή αποθήκευσης, κάνοντας χρήση των αντιπροσωπευτικών παρατηρήσεων. Αυτή η τεχνική καθιστά τον αλγόριθμο CURE κατάλληλο για χρήση σε μεγάλες βάσεις δεδομένων όπου είναι ανέφικτο να χρησιμοποιηθούν όλα τα δεδομένα μαζί για την επίτευξη της ομαδοποίησης.

Μέθοδος ROCK: Η μέθοδος Robust Clustering using Links (ROCK) είναι ένας ισχυρός αλγόριθμος ομαδοποίησης που έχει σχεδιαστεί για να χειρίζεται δεδομένα με ακραίες τιμές και θόρυβο. Παρουσιάστηκε από τους Guha, Rastogi και Shim το 1999.

Το ROCK συνδυάζει ιδέες από την ιεραρχική ομαδοποίηση και την ομαδοποίηση με βάση την πυκνότητα για να επιτύχει ανθεκτικότητα έναντι των ακραίων τιμών. Προσδιορίζει συστάδες λαμβάνοντας υπόψη τόσο τη συνδεσιμότητα με βάση την πυκνότητα όσο και τη σύνδεση μεταξύ σημείων δεδομένων.

Ο αλγόριθμος προχωρά ως εξής:

1. Υπολογισμός Πυκνότητας: Το ROCK υπολογίζει την πυκνότητα κάθε σημείου δεδομένων με βάση την τοπική του γειτονιά. Η πυκνότητα συνήθως μετριέται χρησιμοποιώντας τον αριθμό των γειτονικών σημείων σε μια καθορισμένη ακτίνα.
2. Επιλογή βασικών σημείων: Τα βασικά σημεία προσδιορίζονται ως σημεία δεδομένων με πυκνότητα μεγαλύτερη από ένα προκαθορισμένο όριο. Αυτά τα σημεία είναι πιθανό να βρίσκονται στις πυκνές περιοχές των συστάδων.
3. Απόσταση αμοιβαίας προσβασιμότητας: Για κάθε κεντρικό σημείο, υπολογίζεται η απόσταση αμοιβαίας προσβασιμότητας σε άλλα σημεία πυρήνα. Αυτή η απόσταση μετρά

τη συνδεσιμότητα μεταξύ των βασικών σημείων με βάση τους κοινόχρηστους γείτονές τους.

4. Ιεραρχική ομαδοποίηση: Ένας αλγόριθμος ιεραρχικής ομαδοποίησης, όπως Single Linkage ή Complete Linkage, εφαρμόζεται στα βασικά σημεία με βάση τις αποστάσεις αμοιβαίας προσβασιμότητας. Αυτό το βήμα δημιουργεί ένα δενδρόγραμμα που καταγράφει τη σύνδεση μεταξύ των συστάδων.
5. Εξαγωγή συστάδων: Το δενδρόγραμμα κόβεται σε ένα ορισμένο επίπεδο για να εξαχθούν μεμονωμένες συστάδες. Το σημείο κοπής μπορεί να προσδιοριστεί χρησιμοποιώντας ένα κατώφλι απόστασης ή λαμβάνοντας υπόψη τον αριθμό των επιθυμητών συστάδων.

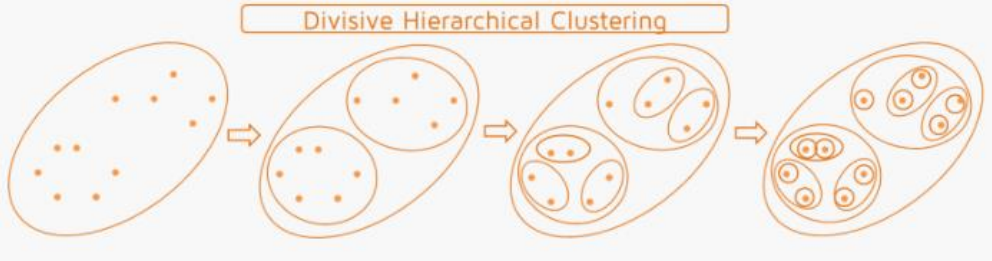
Το ROCK είναι ιδιαίτερα χρήσιμο σε σενάρια όπου τα δεδομένα μπορεί να περιέχουν ακραίες τιμές ή θόρυβο. Ενσωματώνοντας συνδεσιμότητα με βάση την πυκνότητα και λαμβάνοντας υπόψη τη σύνδεση μεταξύ των βασικών σημείων, μπορεί να χειριστεί καταστάσεις όπου οι παραδοσιακοί αλγόριθμοι ομαδοποίησης μπορεί να δυσκολεύονται να προσδιορίσουν με ακρίβεια τις συστάδες λόγω της παρουσίας ακραίων τιμών.

Η ευρωστία του ROCK του επιτρέπει να είναι αποτελεσματικό σε διάφορους τομείς, συμπεριλαμβανομένης της εξόρυξης δεδομένων, της αναγνώρισης προτύπων και της ανίχνευσης ανωμαλιών. Έχει εφαρμοστεί σε τομείς όπως η κατάτμηση εικόνων, η ανάλυση δικτύου και η βιοπληροφορική.

Αξίζει να σημειωθεί ότι το ROCK είναι μόνο μία από τις πολλές διαθέσιμες μεθόδους ομαδοποίησης και η καταλληλότητά του εξαρτάται από τα χαρακτηριστικά και τις απαιτήσεις του συγκεκριμένου συνόλου δεδομένων ή προβλήματος.

1.3.1.2 Διαιρετικές Μέθοδοι

Οι διαιρετικές μέθοδοι ξεκινούν με μια μόνο ομάδα που περιέχει *n* παρατηρήσεις και διαιρείται σε όλο μικρότερες ομάδες έως ότου φτάσουν στο σημείο που κάθε στοιχείο αποτελεί μία ομάδα. Προσπαθούν να βρουν υποομάδες των ήδη διαμορφωμένων ομάδων που είναι περισσότερο απομακρυσμένες και τις διαχωρίζουν. Οι αλγόριθμοι που ανήκουν σε αυτή την οικογένεια είναι υπολογιστικά πιο απαιτητικοί από ότι οι συσσωρευτικές μέθοδοι, καθώς για τη διαμέριση *n* δεδομένων σε 2 ομάδες απαιτείται η εξέταση $2^{n-1} - 1$ περιπτώσεων.



Σχήμα 1.7 Γραφική Απεικόνιση Διαιρετικών μεθόδων

Η παλαιότερη διαιρετική μέθοδος είναι αυτή των Edwards and Cavalli-Sforza (1965). Σε κάθε βήμα, επιλέγεται από όλες τις δυνατές διαμερίσεις εκείνη η οποία ελαχιστοποιεί το

άθροισμα των τετραγωνικών αποκλίσεων. Για κάθε δύο ομάδες που δημιουργούνται από τη διαίρεση μιας προηγούμενης, προκύπτει ότι τα σημεία που τις απαρτίζουν χαρακτηρίζονται από το μέγιστο άθροισμα των τετραγώνων μεταξύ των δύο ομάδων.

1.3.1.3 BIRCH

Ο αλγόριθμος BIRCH (Balanced Iterative Reducing and Clustering Hierarchies) ανήκει στη κατηγορία των ιεραρχικών αλγορίθμων, όμως χειρίζεται τα δεδομένα με διαφορετικό τρόπο από ότι οι συσσωρευτικές και διαιρετικές μέθοδοι.

Η μέθοδος αυτή είναι πολύ αποτελεσματική σε μεγάλα σύνολα δεδομένων, και ο τρόπος με τον οποίο λειτουργεί είναι να δημιουργεί μικρές περιοχές (summaries) των αρχικών δεδομένων οι οποίες διατηρούν όση περισσότερη πληροφορία είναι δυνατόν για τα δεδομένα αυτά. Οι περιοχές αυτές χρησιμοποιούνται στη διαδικασία της ομαδοποίησης αντί των αρχικών δεδομένων. Για τη διαδικασία της ομαδοποίησης είναι δυνατόν να χρησιμοποιηθεί κάποιος άλλος αλγόριθμος (π.χ K-Means) και για αυτό το λόγο ο BIRCH συχνά χρησιμοποιείται ως πρώτο βήμα πριν εφαρμοστεί κάποιος άλλος αλγόριθμος.

Το μειονέκτημα του BIRCH είναι ότι δεν χειρίζεται κατηγορικά δεδομένα και συνεπώς τα δεδομένα που μπορεί να δεχτεί περιορίζονται σε αυτά που μπορούν να αναπαρασταθούν στον Ευκλείδειο χώρο.

Το BIRCH στοχεύει στην αποτελεσματική και αποτελεσματική ομαδοποίηση μεγάλων συνόλων δεδομένων κατασκευάζοντας μια δομή που μοιάζει με δέντρο στη μνήμη που ονομάζεται Δέντρο Χαρακτηριστικών Συστάδων (CFT). Το CFT επιτρέπει την ταχύτερη ομαδοποίηση και διευκολύνει τη συμπαγή αναπαράσταση του συνόλου δεδομένων.

Ο αλγόριθμος προχωρά στα εξής βήματα:

1. Αρχικοποίηση: Το BIRCH αρχικοποιεί το CFT ορίζοντας παραμέτρους ενός ο μέγιστος αριθμός καταχωρήσεων ανά κόμβο (συντελεστής διακλάδωσης), η μέγιστη ακτίνα ενός συμπλέγματος (απόσταση κατωφλίου) και ο μέγιστος αριθμός συστάδων.
2. Φάση σάρωσης: Το BIRCH διαβάζει τα σημεία δεδομένων ένα προς ένα και τα εισάγει στους κατάλληλους κόμβους φύλλων του CFT. Κάθε κόμβος φύλλου αντιπροσωπεύει ένα συστάδα και διατηρεί στατιστικά στοιχεία όπως ο αριθμός των σημείων, το άθροισμα των τιμών των χαρακτηριστικών και το τετράγωνο άθροισμα των τιμών των χαρακτηριστικών.
3. Εξαγωγή χαρακτηριστικών ομαδοποίησης: Καθώς εισάγονται νέα σημεία δεδομένων στους κόμβους φύλλων, το CFT ενημερώνει δυναμικά τα στατιστικά του συμπλέγματος. Αυτό το βήμα διασφαλίζει τη συμπαγή αναπαράσταση των δεδομένων συνοψίζοντας την κατανομή τους εντός του CFT.
4. Φάση συγχώνευσης: Το BIRCH εφαρμόζει μια διαδικασία συγχώνευσης από κάτω τους τα πάνω για να μειώσει τον αριθμό των συστάδων και να δημιουργήσει μια ιεραρχική δομή.

Συγχωνεύει παρόμοια συστάδες με βάση την απόσταση μεταξύ των κεντροειδών τους και την καθορισμένη απόσταση κατωφλίου.

5. Επανεκχώρηση συμπλέγματος: Μετά τη συγχώνευση, το BIRCH εκχωρεί σημεία δεδομένων στις συγχωνευμένες συστάδες με βάση τις αποστάσεις τους από τα νέα κεντροειδή.

Ο αλγόριθμος BIRCH παρέχει πολλά πλεονεκτήματα για σύνολα δεδομένων μεγάλης κλίμακας. Συμπιέζει αποτελεσματικά τα δεδομένα μέσω του CFT, μειώνοντας τις απαιτήσεις μνήμης και βελτιώνοντας την απόδοση επεξεργασίας. Η ιεραρχική δομή επιτρέπει διαφορετικά επίπεδα ευαισθησίας στα αποτελέσματα ομαδοποίησης. Επιπλέον, μπορεί να χειριστεί τόσο αριθμητικά όσο και κατηγορικά δεδομένα.

Ωστόσο, είναι σημαντικό να σημειωθεί ότι το BIRCH είναι ευαίσθητο στην επιλογή των παραμέτρων του, όπως ο παράγοντας διακλάδωσης και η απόσταση κατωφλίου. Αυτές οι παράμετροι θα πρέπει να επιλέγονται προσεκτικά για να διασφαλίζονται τα κατάλληλα αποτελέσματα ομαδοποίησης.

Το BIRCH έχει εφαρμοστεί σε διάφορους τομείς, συμπεριλαμβανομένης της εξόρυξης δεδομένων, της μηχανικής μάθησης και της αναγνώρισης προτύπων, όπου μεγάλα σύνολα δεδομένων πρέπει να ομαδοποιούνται αποτελεσματικά. Η αποτελεσματικότητά του και η ικανότητά του να χειρίζεται σύνολα δεδομένων μεγάλης κλίμακας το καθιστούν δημοφιλή επιλογή για διερευνητική ανάλυση δεδομένων και εργασίες προεπεξεργασίας.

1.3.2 Μη-ιεραρχικές Μέθοδοι

Οι μη ιεραρχικές μέθοδοι προσπαθούν να χωρίσουν τα δεδομένα βάσει κάποιου προκαθορισμένου κριτηρίου και αριθμού ομάδων. Χρησιμοποιούν επαναληπτικές διαδικασίες με σκοπό των ανάθεση παρατηρήσεων στις κατάλληλες ομάδες.

Τα K-means και K-medoids είναι δύο δημοφιλείς μη ιεραρχικοί αλγόριθμοι ομαδοποίησης που χρησιμοποιούνται για τη διαίρεση ενός συνόλου δεδομένων σε K διακριτές συστάδες. Και οι δύο μέθοδοι στοχεύουν στην ελαχιστοποίηση της διακύμανσης εντός του συμπλέγματος, αλλά διαφέρουν στον τρόπο με τον οποίο ορίζουν τα κεντροειδή ή τους αντιπροσώπους των συστάδων.

K-means Ομαδοποίηση:

Η ομαδοποίηση K-means είναι ένας επαναληπτικός αλγόριθμος που εκχωρεί σημεία δεδομένων σε συστάδες με βάση την εγγύτητά τους με τον μέσο όρο ή το κέντρο του συμπλέγματος. Ο αλγόριθμος προχωρά ως εξής:

1. Αρχικοποίηση: Επιλέξτε τυχαία K αρχικά κεντροειδή από το σύνολο δεδομένων.
2. Εκχώρηση: Αντιστοιχίστε κάθε σημείο δεδομένων στο πλησιέστερο κέντρο με βάση μια μέτρηση απόστασης, συνήθως την Ευκλείδεια απόσταση.

3. Ενημέρωση: Υπολογίστε ξανά τα κεντροειδή λαμβάνοντας τον μέσο όρο των σημείων δεδομένων που έχουν εκχωρηθεί σε κάθε συστάδα.
4. Επανάληψη: Επαναλάβετε τα βήματα ανάθεσης και ενημέρωσης μέχρι τη σύγκλιση, όπου τα κεντροειδή δεν αλλάζουν πλέον σημαντικά ή πληρείται ένα προκαθορισμένο κριτήριο διακοπής.

Το K-means στοχεύει να ελαχιστοποιήσει το άθροισμα των τετραγωνικών σφαλμάτων εντός του συμπλέγματος (SSE) βελτιστοποιώντας επαναληπτικά την εκχώρηση σημείων δεδομένων σε συστάδες και ενημερώνοντας ανάλογα τα κεντροειδή.

Ομαδοποίηση K-medoids:

Το K-medoids clustering, γνωστό και ως PAM (Partitioning Around Medoids), είναι μια παραλλαγή του K-means που χρησιμοποιεί αντιπροσωπευτικά αντικείμενα, γνωστά ως medoids, αντί για κεντροειδή. Τα medoids είναι πραγματικά σημεία δεδομένων από το σύνολο δεδομένων.

1. Αρχικοποίηση: Επιλέξτε τυχαία K αρχικά medoids από το σύνολο δεδομένων.
2. Εκχώρηση: Αντιστοιχίστε κάθε σημείο δεδομένων στο πλησιέστερο medoid με βάση μια μέτρηση απόστασης, όπως η Ευκλείδεια απόσταση ή η απόσταση του Μανχάταν.
3. Ενημέρωση: Για κάθε συστάδα, αξιολογήστε τη συνολική απόσταση κάθε σημείου δεδομένων από όλα τα άλλα σημεία του συμπλέγματος. Επιλέξτε το σημείο δεδομένων με τη χαμηλότερη συνολική απόσταση ως το νέο medoid.
4. Επανάληψη: Επαναλάβετε τα βήματα ανάθεσης και ενημερώστε μέχρι τη σύγκλιση, όπου τα medoids δεν αλλάζουν πλέον ή πληρείται ένα κριτήριο διακοπής.

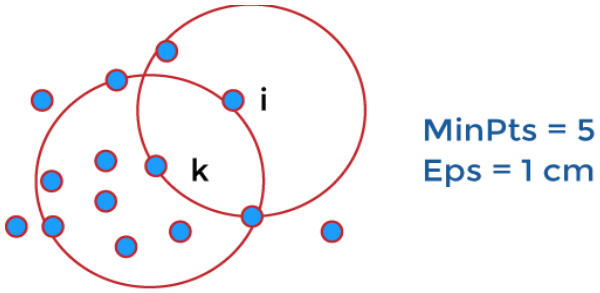
Το K-medoids στοχεύει στην ελαχιστοποίηση της συνολικής ανομοιότητας ή της απόστασης εντός των συστάδων βελτιστοποιώντας επαναληπτικά την εκχώρηση σημείων δεδομένων σε medoids και ενημερώνοντας ανάλογα τα medoids.

Τόσο τα K-means όσο και τα K-medoids έχουν τα δυνατά και τα αδύνατα σημεία τους. Το K-means είναι υπολογιστικά αποδοτικό και λειτουργεί καλά με μεγάλα σύνολα δεδομένων, αλλά μπορεί να είναι ευαίσθητο στην αρχική επιλογή του κέντρου και μπορεί να επηρεαστεί από ακραίες τιμές. Το K-medoids, από την άλλη πλευρά, είναι πιο ανθεκτικό σε ακραίες τιμές και μπορεί να χειριστεί μη ευκλείδειες μετρήσεις απόστασης, αλλά μπορεί να είναι λιγότερο αποτελεσματικό και κατάλληλο για μεγαλύτερα σύνολα δεδομένων.

Η επιλογή μεταξύ K-means και K-medoids εξαρτάται από τα χαρακτηριστικά των δεδομένων, την παρουσία ακραίων τιμών, την επιθυμητή ερμηνευσιμότητα των αποτελεσμάτων και τους διαθέσιμους υπολογιστικούς πόρους.

1.4 Μέθοδοι πυκνότητας

Οι μέθοδοι πυκνότητας (Density Based) ψάχνουν για περιοχές υψηλής συγκέντρωσης παρατηρήσεων. Οι αλγόριθμοι αυτής της κατηγορίας χρησιμοποιούν τη μέθοδο του κοντινότερου γείτονα για να εντοπίσουν κοντινές παρατηρήσεις.



Σχήμα 1.5 Γραφική απεικόνιση μεθόδων πυκνότητας. Το i είναι άμεσα προσβάσιμο στην πυκνότητα από το k

1.4.1 DBSCAN

Η πιο γνωστή τέτοια μέθοδος είναι ο αλγόριθμος **DBSCAN** (Ester et al., 1996) ο οποίος χρησιμοποιεί μια ακτίνα *ε* γύρω από ένα σημείο και ένα ελάχιστο αριθμό γειτόνων (minPts) που πρέπει να συμπεριλάβει σε αυτή τη περιοχή. Με τη χρήση αυτών των δύο παραμέτρων χωρίζει τα σημεία που εντοπίζει σε **κύρια** (core), **συνοριακά** (border) και **έκτοπα** (outlier).

Ένα κύριο σημείο θεωρείται αυτό στο οποίο η περιοχή που ανήκει περιέχει τουλάχιστον minPts γείτονες. Ένα συνοριακό σημείο θεωρείται αυτό που η γειτονιά στην οποία ανήκει περιέχει λιγότερα από minPts σημεία αλλά είναι προσβάσιμο από ένα κύριο σημείο. Σημειώνεται ότι η έννοια «προσβάσιμο» αναφέρεται στην ικανότητα προσέγγισης ενός σημείου δεδομένων από ένα άλλο μέσω μιας διαδρομής γειτονικών σημείων. Ένα έκτοπο σημείο είναι αυτό που ανήκει σε άλλη περιοχή και συνεπώς θεωρείται ότι ανήκει σε άλλη κλάση. Συγκεκριμένα, ένα σημείο δεδομένων θεωρείται προσβάσιμο από έναν πυρήνα ή κύριο σημείο εάν υπάρχει μια αλυσίδα ή μια ακολουθία γειτονικών σημείων που τα συνδέει. Αυτά τα γειτονικά σημεία μπορεί να μην είναι άμεσοι γείτονες, αλλά συνδέονται μέσω μιας σειράς βημάτων εντός της καθορισμένης ακτίνας γειτονιάς ή κατωφλίου απόστασης.

Για παράδειγμα, ας εξετάσουμε ένα σενάριο όπου έχουμε τρία σημεία δεδομένων Α, Β και Γ. Εάν το σημείο Α είναι ένα κεντρικό σημείο και το σημείο Β είναι ένα οριακό σημείο, το σημείο Β θεωρείται προσβάσιμο από το σημείο Α εάν υπάρχει μονοπάτι γειτονικού σημεία (τα οποία μπορεί να περιλαμβάνουν άλλα οριακά σημεία ή σημεία πυρήνα) που τα συνδέουν. Αυτή η διαδρομή μπορεί να περιλαμβάνει τη μετάβαση από το σημείο Α σε ένα γειτονικό σημείο Χ, μετά από το σημείο Χ σε ένα άλλο γειτονικό σημείο Υ και, τέλος, στο σημείο Β.

Η έννοια της προσβασιμότητας είναι σημαντική στους αλγόριθμους ομαδοποίησης που βασίζονται στην πυκνότητα, επειδή βοηθά στον προσδιορισμό της συμμετοχής των οριακών σημείων. Ακόμη και αν η τοπική πυκνότητα γύρω από ένα οριακό σημείο είναι χαμηλότερη

από το καθορισμένο όριο, εφόσον είναι προσβάσιμη από ένα κεντρικό σημείο, θεωρείται μέρος του ίδιου συμπλέγματος.

Λαμβάνοντας υπόψη την προσβασιμότητα, οι αλγόριθμοι ομαδοποίησης που βασίζονται στην πυκνότητα μπορούν να αναγνωρίσουν συστάδες που έχουν ποικίλες πυκνότητες και σχήματα, προσαρμόζοντας διαφορετικά μοτίβα στη διανομή δεδομένων.

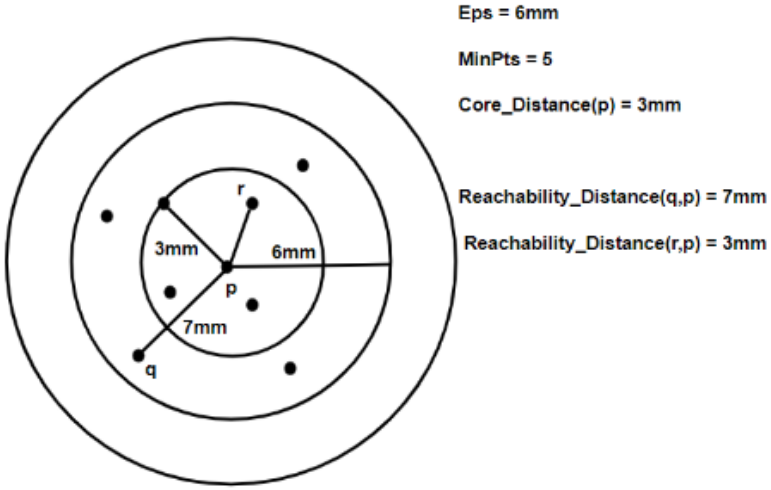
Τα βήματα που ακολουθεί ο αλγόριθμος είναι πρώτα να διαλέγει ένα σημείο που δεν έχει ανατεθεί σε κάποια ομάδα και να ελέγχει αν είναι κύριο σημείο. Αν αποδειχθεί ότι είναι, τότε δημιουργεί μια ομάδα γύρω από αυτό το σημείο, ειδάλλως το θεωρεί ως έκτοπο. Στη πρώτη περίπτωση, εντάσσει όλα τα σημεία πρόσβασης άμεσα στην ομάδα αυτή. Αν κάποιο προηγούμενα έκτοπο σημείο προστεθεί στην ομάδα, τότε αυτό θεωρείται ως συνοριακό. Ο αλγόριθμος τερματίζει όταν όλα τα σημεία έχουν ανατεθεί σε κάποια ομάδα ή έχουν θεωρηθεί έκτοπα.

1.4.2 OPTICS

Ο αλγόριθμος OPTICS βασίζεται στον DBSCAN εισάγοντας όμως και δύο καινούργιες έννοιες οι οποίες δίνονται παρακάτω.

Core Distance: Είναι η ελάχιστη τιμή της ακτίνας που απαιτείται ώστε ένα σημείο να θεωρηθεί ως κύριο. Εάν ένα σημείο δε θεωρείται κύριο, τότε η core distance θεωρείται απροσδιόριστη.

Reachability Distance: Ορίζεται ως προς ένα άλλο σημείο από αυτό που εξετάζεται και είναι η μέγιστη τιμή μεταξύ του core distance και μια άλλης μετρικής (π.χ. Ευκλείδειας) που ορίζεται για τη τρέχουσα ανάλυση. Η συγκεκριμένη απόσταση δεν ορίζεται αν το εξεταζόμενο σημείο δεν αποτελεί κύριο σημείο.



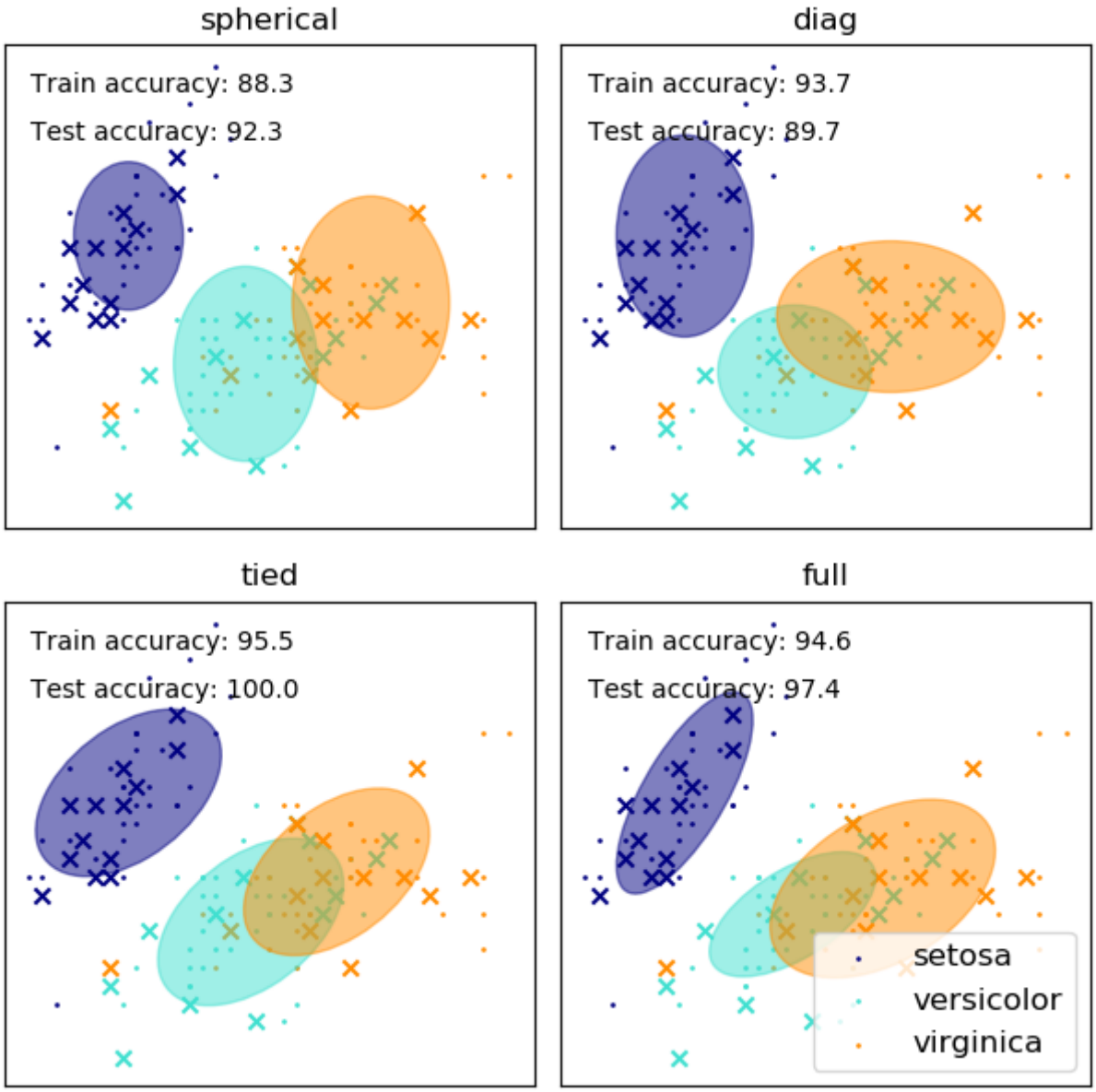
Σχήμα 1.9 Γραφική αναπαράσταση του αλγορίθμου OPTICS

Η τεχνική αυτή διαφέρει από άλλες τεχνικές ομαδοποίησης διότι δεν χωρίζει απευθείας τα δεδομένα σε ομάδες, αλλά χρησιμοποιεί μια οπτικοποίηση των reachability distances για να ομαδοποιήσει τα δεδομένα.

Ο αλγόριθμος OPTICS απαιτεί περισσότερη μνήμη και υπολογιστική ισχύ από τον DBSCAN, καθώς διατηρεί μια ουρά προτεραιότητας για να προσδιορίσει ποιο σημείο είναι πλησιέστερο στο εξεταζόμενο. Στον αντίποδα αυτού, δεν απαιτεί την παράμετρο ε όπως ο DBSCAN, και συνεπώς χρειάζεται λιγότερο χρόνο βελτιστοποίησης των παραμέτρων του.

1.5 Μέθοδοι βασιζόμενες σε κατανομές

Οι αλγόριθμοι που ανήκουν σε αυτή τη κατηγορία κάνουν την εικασία ότι τα δεδομένα ακολουθούν γνωστές κατανομές. Το συνηθέστερο παράδειγμα κατανομών είναι οι Γκαουσιανές κατανομές και το μοντέλο που χρησιμοποιείται είναι το **Gaussian Mixture Model** (GMM).



Σχήμα 1.6 Γραφική απεικόνιση Gaussian Mixture Models

Ένα μοντέλο Gaussian Mixture (GMM) είναι ένα πιθανολογικό μοντέλο που αντιπροσωπεύει ένα σύνολο δεδομένων ως συνδυασμό πολλαπλών κατανομών Gauss. Είναι ένα παραμετρικό μοντέλο που υποθέτει ότι τα σημεία δεδομένων δημιουργούνται από ένα μείγμα κατανομών Gauss.

Σε ένα Gaussian Mixture Model, κάθε στοιχείο αντιπροσωπεύει μια Gaussian κατανομή, που χαρακτηρίζεται από τον μέσο όρο, τη συνδιακύμανση και το βάρος του. Το βάρος κάθε στοιχείου αντιπροσωπεύει τη συμβολή του ή την πιθανότητα ενός σημείου δεδομένων να ανήκει σε αυτό το στοιχείο.

Το GMM υποθέτει ότι τα σημεία δεδομένων δημιουργούνται επιλέγοντας ένα από τα στοιχεία Gaussian με συγκεκριμένη πιθανότητα και στη συνέχεια δημιουργώντας το σημείο δεδομένων από αυτό το επιλεγμένο στοιχείο. Τα βασικά βήματα που εμπλέκονται στην προσαρμογή ενός μοντέλου Gaussian Mixture σε ένα σύνολο δεδομένων είναι τα εξής:

1. Αρχικοποίηση: Αρχικοποίηση των παραμέτρων του μοντέλου, συμπεριλαμβανομένου του αριθμού των συστάδων (K), των μέσων, των συνδιακυμάνσεων και των βαρών κάθε στοιχείου.
2. Αλγόριθμος Προσδοκίας-Μεγιστοποίησης (EM): Χρησιμοποιήστε τον αλγόριθμο Προσδοκίας-Μεγιστοποίησης για να υπολογίσετε τις παραμέτρους του GMM. Ο αλγόριθμος EM περιλαμβάνει μια επαναληπτική διαδικασία με βήματα:
3. Βήμα προσδοκίας (E-βήμα): Υπολογίστε την πιθανότητα ή την πιθανότητα κάθε σημείου δεδομένων να ανήκει σε κάθε στοιχείο χρησιμοποιώντας τις τρέχουσες εκτιμήσεις παραμέτρων. Αυτό γίνεται χρησιμοποιώντας την έννοια των ευθυνών ή των μεταγενέστερων πιθανοτήτων.
4. Βήμα μεγιστοποίησης (M-βήμα): Ενημερώστε τις παραμέτρους των συνιστωσών Gaussian μεγιστοποιώντας την πιθανότητα των δεδομένων με βάση τις τρέχουσες ευθύνες. Αυτό περιλαμβάνει τον υπολογισμό νέων εκτιμήσεων για τα μέσα, τις συνδιακυμάνσεις και τα βάρη με βάση τις τρέχουσες ευθύνες.
5. Το E-step και το M-step επαναλαμβάνονται συνεχώς μέχρι τη σύγκλιση, όπου οι παράμετροι σταθεροποιούνται ή πληρείται ένα κριτήριο διακοπής.
6. Επιλογή μοντέλου: Προσδιορίστε τον κατάλληλο αριθμό εξαρτημάτων (K) για το GMM. Αυτό μπορεί να γίνει χρησιμοποιώντας διάφορα κριτήρια επιλογής μοντέλου, όπως το Akaike Information Criterion (AIC) ή το Bayesian Information Criterion (BIC), για να εξισορροπηθεί η πολυπλοκότητα και η καλή προσαρμογή του μοντέλου.

Μόλις το GMM προσαρμοστεί στα δεδομένα, μπορεί να χρησιμοποιηθεί για διάφορους σκοπούς, όπως εκτίμηση πυκνότητας, ομαδοποίηση και δημιουργία νέων δειγμάτων συνθετικών δεδομένων. Τα Gaussian Mixture Models είναι ευέλικτα και ισχυρά μοντέλα για την αναπαράσταση πολύπλοκων διανομών δεδομένων. Έχουν εφαρμογές σε διάφορους τομείς, συμπεριλαμβανομένης της αναγνώρισης προτύπων, της ομαδοποίησης δεδομένων, της τμηματοποίησης εικόνων και της ανίχνευσης ανωμαλιών.

ΚΕΦΑΛΑΙΟ 2

ΕΣΩΤΕΡΙΚΑ ΚΑΙ ΕΞΩΤΕΡΙΚΑ ΚΡΙΤΗΡΙΑ

Οι αλγόριθμοι ομαδοποίησης εντοπίζουν διαχωρισμούς μεταξύ των παρατηρήσεων ενός σύνολο δεδομένων. Το αποτέλεσμα ενός αλγορίθμου δεν είναι πάντοτε το βέλτιστο, καθώς ο αλγόριθμος ενδέχεται να τερματίσει τη λειτουργία του νωρίτερα ή αργότερα από το βέλτιστο διαχωρισμό. Για το λόγο αυτό αναπτύχθηκαν κριτήρια για την εύρεση του βέλτιστου αριθμού των ομάδων και την καλύτερη ομαδοποίηση των παρατηρήσεων εντός των ομάδων αυτών.

Η πρώτη σχετική μελέτη πραγματοποιήθηκε από τον Thorndike (1953), στην οποία προτάθηκε να υλοποιηθεί μια γραφική απεικόνιση σε ένα άξονα του μέσου όρου των αποστάσεων εντός των ομάδων και σε ένα δεύτερο του αριθμού των ομάδων. Σε κάθε βήμα αύξησης του αριθμού των ομάδων, παρατηρείται μείωση του μέσου όρου των αποστάσεων. μέσα στις ομάδες. Στη πλειοψηφία των περιπτώσεων εμφανίζεται κάποια θέση στο γράφημα όπου παρατηρείται απότομη μείωση του μέσου όρου και στη συνέχεια οριζοντιοποίηση του γραφήματος. Η εικασία του Thorndike ήταν ότι το σημείο αυτό δείχνει το βέλτιστο πλήθος ομάδων. Στη πράξη, ο Thorndike εξετάζοντας το παραπάνω με χρήση προσομοιωμένων δεδομένων ανακάλυψε ότι δεν ήταν πάντα αληθής η εικασία που είχε προτείνει. Ωστόσο, η προσέγγιση που πρότεινε υπήρξε το έναυσμα για τη διερεύνηση του χώρου γύρω από αντίστοιχες τεχνικές εύρεσης του βέλτιστου πλήθους ομάδων.

Οι αλγόριθμοι ομαδοποίησης σε συνδυασμό με τα κριτήρια διακοπής μπορούν να κατηγοριοποιήσουν τα εκάστοτε δεδομένα με μεγαλύτερη ή μικρότερη επιτυχία ανάλογα με τη φύση τους. Τα κριτήρια τερματισμού διακόπτουν την εκτέλεση ενός αλγορίθμου σε κάποιο βήμα του, και για να επιτευχθεί ο καλύτερος δυνατός διαχωρισμός σε ομάδες, είναι απαραίτητη η βελτιστοποίηση τους.

Στο πλαίσιο της αξιολόγησης ομαδοποίησης, χρησιμοποιούνται εσωτερικά και εξωτερικά κριτήρια για την αξιολόγηση της ποιότητας και της απόδοσης των αλγορίθμων ομαδοποίησης.

Εσωτερικά κριτήρια:

Τα εσωτερικά κριτήρια αξιολογούν την ποιότητα των συστάδων με βάση αποκλειστικά τα χαρακτηριστικά των δεδομένων και τα ίδια τα αποτελέσματα της ομαδοποίησης, χωρίς καμία εξωτερική αναφορά ή βασική αλήθεια. Αυτά τα κριτήρια παρέχουν πληροφορίες για την εγγενή δομή ή τη συνοχή των δεδομένων.

Παραδείγματα εσωτερικών κριτηρίων περιλαμβάνουν:

- Άθροισμα τετράγωνων σφαλμάτων (SSE) ή Διακύμανση εντός συμπλέγματος: Μετρά τη συμπάγεια συστάδες υπολογίζοντας το άθροισμα των τετραγωνικών αποστάσεων μεταξύ των σημείων δεδομένων και των κεντροειδών συστάδων τους. Το χαμηλότερο SSE υποδηλώνει καλύτερη ομαδοποίηση.
- Συντελεστής Silhouette: Αξιολογεί τη συμπάγεια και τον διαχωρισμό των συστάδων λαμβάνοντας υπόψη τη μέση απόσταση των σημείων δεδομένων από το δικό τους συστάδα σε σύγκριση με την απόσταση από το πλησιέστερο γειτονικό συστάδα. Ένας υψηλότερος συντελεστής σιλουέτας υποδηλώνει καλύτερη ομαδοποίηση.
- Δείκτης Davies-Bouldin: Μετρά την ποιότητα της ομαδοποίησης με βάση τις αποστάσεις μεταξύ των συστάδων. Λαμβάνει υπόψη τόσο τη διασπορά εντός του συμπλέγματος όσο και τον διαχωρισμό μεταξύ των συστάδων, με χαμηλότερες τιμές να υποδεικνύουν καλύτερη ομαδοποίηση.

Τα εσωτερικά κριτήρια είναι χρήσιμα όταν η βασική αλήθεια ή οι ετικέτες αληθούς συμπλέγματος δεν είναι διαθέσιμες ή όταν αξιολογείται η δομή ομαδοποίησης ανεξάρτητα από εξωτερικές πληροφορίες. Ωστόσο, ενδέχεται να μην καταγράφουν την πραγματική ορθότητα της ομαδοποίησης.

Εξωτερικά κριτήρια:

Τα εξωτερικά κριτήρια αξιολογούν την απόδοση των αλγορίθμων ομαδοποίησης συγκρίνοντας τα αποτελέσματα της ομαδοποίησης με μια αλήθεια αναφοράς ή βάσης. Αυτά τα κριτήρια απαιτούν προηγούμενη γνώση ή εξωτερικά παρεχόμενες ετικέτες κλάσης για τα σημεία δεδομένων.

Παραδείγματα εξωτερικών κριτηρίων περιλαμβάνουν:

- Δείκτης Rand (RI): Μετρά την ομοιότητα μεταξύ των αποτελεσμάτων ομαδοποίησης και της βασικής αλήθειας συγκρίνοντας συμφωνίες κατά ζεύγη ή διαφωνίες μεταξύ των εκχωρημένων συστάδων και των αληθών κλάσεων.
- Προσαρμοσμένος δείκτης Rand (ARI): Παρόμοιος με τον δείκτη Rand, αλλά προσαρμοσμένος για τυχαία συμφωνία. Ένα υψηλότερο ARI υποδηλώνει καλύτερη συμφωνία μεταξύ της ομαδοποίησης και της βασικής αλήθειας.
- Δείκτης Fowlkes-Mallows (FMI): Υπολογίζει τη γεωμετρική μέση ακρίβεια και ανάκληση μεταξύ της ομαδοποίησης και του διαμερίσματος αναφοράς. Μετρά την ομοιότητα μεταξύ των δύο συνόλων κατατμήσεων.
- Τα εξωτερικά κριτήρια παρέχουν ένα μέτρο για το πόσο καλά ευθυγραμμίζεται η ομαδοποίηση με τη γνωστή βασική αλήθεια, καθιστώντας τα χρήσιμα για την αξιολόγηση αλγορίθμων ομαδοποίησης σε εποπτευόμενες ή επισημασμένες ρυθμίσεις.

Τόσο τα εσωτερικά όσο και τα εξωτερικά κριτήρια έχουν τα πλεονεκτήματα και τους περιορισμούς τους. Τα εσωτερικά κριτήρια επικεντρώνονται στην εγγενή δομή των δεδομένων, ενώ τα εξωτερικά κριτήρια αξιολογούν τη συμφωνία με μια γνωστή αναφορά. Η επιλογή των κριτηρίων που θα χρησιμοποιηθούν εξαρτάται από τη διαθεσιμότητα της βασικής αλήθειας, τους στόχους της ανάλυσης και τη φύση των δεδομένων.

Η εύρεση του σωστού αριθμού των ομάδων, η κατάλληλη ανάθεση παρατηρήσεων σε αυτές, καθώς και η ορθή εφαρμογή σε νέα μη προηγουμένως γνωστά αλλά παρεμφερή δεδομένα χαρακτηρίζουν το βαθμό επιτυχίας ενός κριτηρίου.

Για την αντικειμενική αξιολόγηση των παραπάνω, οι ερευνητές του χώρου χρησιμοποιούν προσομοιωμένα δεδομένα αντί δεδομένων τα οποία έχουν συλλεχθεί από τον πραγματικό κόσμο. Κατ' αυτό τον τρόπο, ο ερευνητής έχει πλήρη έλεγχο της διαδικασίας παραγωγής των δεδομένων καθώς αυτά προκύπτουν από γνωστές κατανομές με χρήση μεθόδων προσομοίωσης. Ακόμη, είναι σε θέση να καθορίσει τον αριθμό των ομάδων εκ των προτέρων, και συνεπώς είναι εφικτό να συγκριθούν τα αποτελέσματα με την πραγματικότητα.

Η χρήση κριτηρίων τερματισμού δεν εγγυάται απαραίτητα την λήψη μιας σωστής απόφασης. Συγκεκριμένα, ενδέχεται ένα κριτήριο να οδηγήσει έναν αλγόριθμο σε τερματισμό έχοντας καταλήξει σε μεγαλύτερο *(Type I Error)* ή μικρότερο αριθμό ομάδων *(Type II Error)* από ότι υπάρχουν πραγματικά. Ενδέχεται ακόμη και αν ο αριθμός των ομάδων είναι σωστός, η ανάθεση των παρατηρήσεων να είναι ανακριβής.

Η συνήθης τακτική για να προσδιοριστεί ο ακριβής αριθμός των ομάδων είναι η εκτέλεση του αλγορίθμου πολλαπλές φορές, ώστε το πιο κοινό αποτέλεσμα μεταξύ των εκτελέσεων να είναι το πιο αντιπροσωπευτικό της πραγματικότητας. Για να αντιστοιχίσει ο χρήστης παρατηρήσεις σε συγκεκριμένες ομάδες κατά τη ομαδοποίηση, μια κοινή προσέγγιση είναι η δημιουργία τυχαίων αριθμών από διαφορετικές κατανομές. Κάθε κατανομή αντιπροσωπεύει μία συστάδα και έχει μια συγκεκριμένη διασπορά ή εξάπλωση των παρατηρήσεων γύρω από το κέντρο του.

Δημιουργώντας τυχαίους αριθμούς, προσομοιώνουμε τη διαδικασία εκχώρησης παρατηρήσεων σε συστάδες. Η κατανομή αυτών των τυχαίων αριθμών αντιπροσωπεύει τα υποκείμενα χαρακτηριστικά κάθε συστάδας, όπως ο μέσος όρος και η διακύμανσή του.

Η διασπορά ή η εξάπλωση των παρατηρήσεων γύρω από τα κέντρα των συστάδων είναι μια σημαντική παράμετρος. Καθορίζει πόσο στενά ή ευρέως είναι κατανεμημένες οι παρατηρήσεις μέσα σε ένα συστάδα γύρω από το κέντρο του συμπλέγματος. Μια μικρότερη διασπορά υποδηλώνει ότι οι παρατηρήσεις μέσα σε ένα συστάδα είναι πιο σφιχτά συσκευασμένες ή συγκεντρωμένες, ενώ μια μεγαλύτερη διασπορά υποδηλώνει ότι οι παρατηρήσεις είναι πιο διάσπαρτες ή απλωμένες.

Ελέγχοντας τη διασπορά, μπορούμε να προσαρμόσουμε τη συνοχή ή το διαχωρισμό των συστάδων. Μια μικρότερη διασπορά ενθαρρύνει τις παρατηρήσεις μέσα σε ένα συστάδα να είναι πιο όμοιες μεταξύ τους, οδηγώντας σε πιο διακριτά και στενά ομαδοποιημένα συστάδες. Από την άλλη πλευρά, μια μεγαλύτερη διασπορά επιτρέπει μεγαλύτερη μεταβλητότητα εντός

των συστάδων, με δυνητικά αποτέλεσμα συστάδες που είναι λιγότερο συμπαγείς ή πιο επικαλυπτόμενες.

Η διαδικασία δημιουργίας τυχαίων αριθμών από αυτές τις κατανομές με συγκεκριμένες διασπορές μας επιτρέπει να εκχωρήσουμε παρατηρήσεις σε ομάδες με τρόπο που να αντικατοπτρίζει τα επιθυμητά χαρακτηριστικά ομαδοποίησης. Με το χειρισμό της διασποράς, μπορούμε να ελέγξουμε το επίπεδο διαχωρισμού και συνοχής μεταξύ των παρατηρήσεων σε κάθε συστάδα.

Συνολικά, αυτή η προσέγγιση βοηθά στην επίτευξη ακριβούς τοποθέτησης των παρατηρήσεων σε ομάδες, προσομοιώνοντας τα χαρακτηριστικά κατανομής και διασποράς των συστάδων μέσω της δημιουργίας τυχαίων αριθμών.

Συνεπώς, είναι απαραίτητο να χρησιμοποιούνται μέθοδοι σύγκρισης των αποτελεσμάτων κάθε κριτηρίου με αυτά των υπολοίπων που χρησιμοποιούνται. Η σύγκριση αυτή απαιτείται να γίνεται ως προς τον αριθμό των ομάδων και ως προς την επιλογή ανάθεσης των παρατηρήσεων. Οι μέθοδοι σύγκρισης αυτές ονομάζονται εξωτερικοί δείκτες ή εξωτερικά κριτήρια.

2.1 Εσωτερικά κριτήρια

Ένας από τους πιο διαδεδομένους τρόπους εύρεσης του βέλτιστου πλήθους των ομάδων είναι η εξέταση του δενδρογράμματος που προκύπτει από μια ιεραρχική συσσωρευτική μέθοδο. Το σημείο στο οποίο 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. Ο αριθμός των τομών μας υποδεικνύει μια λογική τιμή για το πλήθος των ομάδων.

Τέτοιου είδους μέθοδοι υπόκεινται στη κρίση του κάθε ερευνητή και συνεπώς υπήρξε η ανάγκη να οριστούν κριτήρια που με τυποποιημένο τρόπο και με χρήση λογικών βημάτων να οδηγούν στην εύρεση του βέλτιστου αριθμού των ομάδων. Η προσέγγιση των ερευνητών έχει κατά κύριο λόγο βασιστεί σε τεχνικές προσομοίωσης. Οι Milligan και Cooper (1985) ανέλυσαν 30 κριτήρια διακοπής τα οποία προέκυψαν χρησιμοποιώντας Monte Carlo τεχνικές προσομοίωσης. Στις επόμενες υπο-ενότητες παρουσιάζονται κάποιες από τις πιο σημαντικές κατηγορίες κριτηρίων διακοπής.

2.1.1 Μέθοδοι Απόστασης

Τα μέτρα απόστασης είναι δυνατόν να περιγράψουν τη δομή των δεδομένων που ανήκουν σε μια ομάδα, και την ομοιότητα αυτών με σημεία που ανήκουν σε κάποια άλλη. Αυτός είναι και ο λόγος που πολλοί ερευνητές έχουν χρησιμοποιήσει μέτρα απόστασης για τον προσδιορισμό του βέλτιστου αριθμού των ομάδων.

1. **Κριτήριο του Glasbey:** Το κριτήριο του Glasbey είναι ένας συνδυασμός των μεθόδων του πλησιέστερου γείτονα και της πλήρους συνένωσης που περιγράφηκαν σε προηγούμενη ενότητα. Η πρώτη μέθοδος μεγιστοποιεί την ελάχιστη απόσταση μεταξύ των ομάδων και η δεύτερη ελαχιστοποιεί τη μέγιστη απόσταση εντός των ομάδων.

Ο αλγόριθμος του πλησιέστερου γείτονα εφαρμόζεται υπολογίζοντας την ελάχιστη απόσταση μεταξύ των ομάδων σε κάθε βήμα. Στη συνέχεια, ο αλγόριθμος της πλήρους συνένωσης υπολογίζει την απόσταση των πιο απομακρυσμένων σημείων της ομάδας, καθώς και τη μέγιστη απόσταση μεταξύ των ομάδων, αλλά και τη μέγιστη απόσταση των σημείων εντός των ομάδων. Όταν η τελευταία είναι ίση ή μεγαλύτερη από την ελάχιστη απόσταση μεταξύ των ομάδων, ο αλγόριθμος τερματίζει και οι σχηματισθέντες ομάδες είναι αυτές του τελευταίου βήματος πριν από το τρέχον.

2. **Το κριτήριο Mountford:** Το κριτήριο που αναπτύχθηκε από τον Mountford (1970)είναι ένα τεστ που χρησιμοποιείται για να καθοριστεί εάν δύο ομάδες πρέπει να συγχωνευθούν στο πλαίσιο των αλγορίθμων ομαδοποίησης. Αυτό το κριτήριο έχει σχεδιαστεί ειδικά για την αξιολόγηση της καταλληλότητας της συγχώνευσης δύο ομάδων με βάση ένα γνωστό όριο ή επίπεδο.

Όταν εξετάζεται η συγχώνευση δύο ομάδων, το κριτήριο αξιολογεί τα σημεία ή τις παρατηρήσεις που εμπλέκονται στη συγχώνευση. Εάν η ομάδα που πρόκειται να συγχωνευτεί περιέχει λιγότερα από τέσσερα σημεία, το κριτήριο εκχωρεί μια τιμή μηδέν. Αυτό σημαίνει ότι εάν μια ομάδα έχει λιγότερους από τέσσερις βαθμούς, δεν θεωρείται κατάλληλη για συγχώνευση με βάση αυτό το κριτήριο.

Το πλεονέκτημα αυτού του κριτηρίου είναι η ικανότητά του να δίνει ικανοποιητικές λύσεις όταν ασχολείται με μεγάλο αριθμό ομάδων. Με άλλα λόγια, αυτό το κριτήριο είναι αποτελεσματικό στην καθοδήγηση της διαδικασίας συγχώνευσης και στον καθορισμό των ομάδων που πρέπει να συνδυαστούν σε καταστάσεις όπου υπάρχουν πολλές ομάδες που πρέπει να ληφθούν υπόψη.

Χρησιμοποιώντας το κριτήριο Mountford, οι αλγόριθμοι ομαδοποίησης μπορούν να λάβουν τεκμηριωμένες αποφάσεις σχετικά με το εάν θα συγχωνευτούν δύο ομάδες με βάση τον αριθμό των σημείων που εμπλέκονται σε κάθε ομάδα. Αυτό το κριτήριο βοηθά στον έλεγχο της διαδικασίας ομαδοποίησης και διασφαλίζει ότι η συγχώνευση πραγματοποιείται με τρόπο που να συνάδει με το επιθυμητό επίπεδο μεγέθους και δομής ομάδας.

3. **Το κριτήριο Cubic Clustering:** Το κριτήριο κυβικής ομαδοποίησης (CCC) είναι ένα μέτρο αξιολόγησης που χρησιμοποιείται για την αξιολόγηση της ποιότητας των λύσεων ομαδοποίησης. Προτάθηκε από τους Salvador και Chan το 2004 ως δείκτης επικύρωσης για αλγόριθμους ιεραρχικής ομαδοποίησης.

Το CCC αξιολογεί τη συνοχή, το διαχωρισμό και την απομόνωση των συστάδων με βάση τις αποστάσεις μεταξύ σημείων δεδομένων και κεντροειδών συστάδων. Λαμβάνει υπόψη τόσο την παραλλαγή εντός συστάδας όσο και την παραλλαγή μεταξύ συστάδων.

Ο υπολογισμός του CCC περιλαμβάνει τρία στοιχεία:

Συμπυκνότητα: Αυτό το στοιχείο μετρά τη στεγανότητα ή την εγγύτητα των σημείων δεδομένων σε κάθε συστάδα. Υπολογίζεται ως το άθροισμα των τετραγωνικών αποστάσεων μεταξύ των σημείων δεδομένων και των κεντροειδών συστάδων τους.

Διαχωρισμός: Αυτό το στοιχείο αξιολογεί την ανομοιότητα ή την απόσταση μεταξύ διαφορετικών συστάδων. Υπολογίζεται ως το άθροισμα των τετραγωνικών αποστάσεων μεταξύ των κεντροειδών συστάδων.

Απομόνωση: Αυτό το στοιχείο καταγράφει την ανομοιότητα μεταξύ μεμονωμένων σημείων δεδομένων και των πλησιέστερων γειτονικών συστάδων τους. Υπολογίζεται ως το άθροισμα των τετραγωνικών αποστάσεων μεταξύ των σημείων δεδομένων και των κεντροειδών των πλησιέστερων γειτονικών συστάδων τους.

Το CCC συνδυάζει αυτά τα τρία στοιχεία χρησιμοποιώντας μια κυβική συνάρτηση για να υπολογίσει μια συνολική μέτρηση της ποιότητας του συμπλέγματος. Ο στόχος είναι να βρεθεί μια λύση ομαδοποίησης που μεγιστοποιεί τη συμπαγή, μεγιστοποιεί τον διαχωρισμό μεταξύ συστάδων και ελαχιστοποιεί την απομόνωση μεταξύ σημείων δεδομένων και γειτονικών συστάδων.

Συγκρίνοντας τις τιμές CCC μεταξύ διαφορετικών λύσεων ομαδοποίησης ή διαφορετικών αριθμών συστάδων, είναι δυνατό να προσδιοριστεί η λύση ομαδοποίησης που βελτιστοποιεί την αντιστάθμιση μεταξύ συμπαγούς, διαχωρισμού και απομόνωσης.

Το πλεονέκτημα του κριτηρίου Cubic Clustering είναι η ικανότητά του να αποτυπώνει πολλαπλές πτυχές της ποιότητας της ομαδοποίησης, συμπεριλαμβανομένων των παραλλαγών εντός και μεταξύ συστάδων. Παρέχει ένα ολοκληρωμένο μέτρο αξιολόγησης που λαμβάνει υπόψη τόσο τα τοπικά όσο και τα παγκόσμια χαρακτηριστικά των clusters.

Το CCC έχει εφαρμοστεί σε διάφορους τομείς, όπως η κατάτμηση εικόνας, η αναγνώριση προτύπων και η ανάλυση δεδομένων, για την αξιολόγηση της αποτελεσματικότητας των αλγορίθμων ιεραρχικής ομαδοποίησης και την καθοδήγηση της επιλογής βέλτιστων λύσεων ομαδοποίησης.

4. **Δείκτης *C*-Index:** Ο δείκτης *C*-Index εισήχθη από τους Hubert και Levin (1976) και δίνεται από τον παρακάτω τύπο:

$$d_w - \frac{\min(d_w)}{\max(d_w) - \min(d_w)} \tag{2.1}$$

όπου d_w είναι το άθροισμα των αποστάσεων μέσα στην ομάδα και $\min(d_w)$, $\max(d_w)$

είναι η ελάχιστη και μέγιστη απόσταση μέσα στην ομάδα. Ο δείκτης *C* μετρά τη συνοχή ενός συμπλέγματος λαμβάνοντας υπόψη το εύρος των αποστάσεων μέσα στη συστάδα. Ένας χαμηλότερος δείκτης *C* υποδηλώνει ένα πιο συμπαγές και καλά καθορισμένη συστάδα.

Για να προσδιοριστεί ο βέλτιστος αριθμός ομάδων ή συστάδων, ο δείκτης *C* υπολογίζεται για διαφορετικά επίπεδα ιεραρχίας στη λύση ιεραρχικής ομαδοποίησης. Ο βέλτιστος αριθμός συστάδων λαμβάνεται με τον προσδιορισμό του ιεραρχικού επιπέδου στο οποίο ο δείκτης *C* φτάνει την ελάχιστη τιμή του. Αυτό το επίπεδο αντιστοιχεί στο σημείο όπου οι συστάδες είναι πιο ευδιάκριτα και καλά διαχωρισμένα.

Βρίσκοντας την ελάχιστη τιμή του *C*-Index μεταξύ διαφορετικών επιπέδων, ο *C*-Index παρέχει ένα κριτήριο για τον προσδιορισμό του βέλτιστου αριθμού ομάδων στην ιεραρχική ομαδοποίηση. Βοηθά στην επιλογή της κατάλληλης λύσης ομαδοποίησης που επιτυγχάνει μια ισορροπία μεταξύ συμπαγούς συμπλέγματος και διαχωρισμού.

Ο δείκτης *C* είναι ένα από τα πολλά διαθέσιμα κριτήρια για τον προσδιορισμό του βέλτιστου αριθμού συστάδων στην ιεραρχική ομαδοποίηση. Παρέχει ένα ποσοτικό μέτρο για την καθοδήγηση της επιλογής λύσεων ομαδοποίησης με βάση τη συνοχή και τον διαχωρισμό των συστάδων σε διαφορετικά επίπεδα της ιεραρχίας.

5. **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.

6. **Κριτήριο Ball and Hall:** Το κριτήριο Ball and Hall, που εισήχθη από τους Ball and Hall το 1965, είναι ένα κριτήριο που χρησιμοποιείται για τον προσδιορισμό του αριθμού των συστάδων σε ένα σύνολο δεδομένων με βάση τη μέση απόσταση των στοιχείων από τα κέντρα συμπλέγματος.

Η βασική ιδέα πίσω από το κριτήριο Ball and Hall είναι να αξιολογηθεί η στεγανότητα των συστάδων εξετάζοντας τη μέση απόσταση των σημείων δεδομένων από τα αντίστοιχα κέντρα συστάδων. Η υπόθεση είναι ότι τα καλά διαχωρισμένα και διακριτά σμήνη θα έχουν μικρότερες μέσες αποστάσεις σε σύγκριση με τις συστάδες που είναι λιγότερο συμπαγή ή πιο διάσπαρτα.

Η διαδικασία εφαρμογής του κριτηρίου Ball and Hall περιλαμβάνει τα ακόλουθα βήματα:

Εκτέλεση ομαδοποίησης: Χρησιμοποιήστε έναν αλγόριθμο ομαδοποίησης (π.χ. k-means, ιεραρχική ομαδοποίηση) για να ομαδοποιήσετε τα δεδομένα σε διαφορετικές ομάδες ή συστάδες.

Υπολογισμός κέντρων συμπλέγματος: Προσδιορίστε το κέντρο ή αντιπροσωπευτικό σημείο για κάθε συστάδα. Το κεντροειδές συνήθως υπολογίζεται ως ο μέσος όρος ή η διάμεσος των σημείων δεδομένων μέσα στη συστάδα.

Υπολογίστε τις μέσες αποστάσεις: Υπολογίστε τη μέση απόσταση κάθε σημείου δεδομένων από το κέντρο συστάδας του. Αυτό μπορεί να γίνει χρησιμοποιώντας μια μέτρηση απόστασης όπως η Ευκλείδεια απόσταση ή η απόσταση του Μανχάταν.

Αξιολογήστε τη συμπαγή συστάδα: Εκτιμήστε το πόσο συμπαγείς είναι οι συστάδες εξετάζοντας τις μέσες αποστάσεις. Οι μικρότερες μέσες αποστάσεις υποδεικνύουν πιο συμπαγείς συστάδες, ενώ οι μεγαλύτερες μέσες αποστάσεις υποδηλώνουν λιγότερη συνοχή ή περισσότερα διασκορπισμένα σμήνη.

Προσδιορίστε τον αριθμό των συστάδων: Χρησιμοποιήστε τις μέσες αποστάσεις ως μετρικό για να προσδιορίσετε τον βέλτιστο αριθμό συστάδων. Ο αριθμός των συστάδων επιλέγεται συνήθως έτσι ώστε οι μέσες αποστάσεις να ελαχιστοποιούνται και οι συστάδες να είναι καλά διαχωρισμένες.

Αναλύοντας τις μέσες αποστάσεις των σημείων δεδομένων από τα κέντρα συστάδων τους, το κριτήριο Ball and Hall παρέχει ένα μέτρο της συμπαγούς συστάδας και βοηθά στην καθοδήγηση της επιλογής του κατάλληλου αριθμού συστάδων. Προσφέρει μια προσέγγιση βάσει δεδομένων για τον προσδιορισμό της βέλτιστης λύσης ομαδοποίησης με βάση τη συνοχή και τον διαχωρισμό των συστάδων.

Αξίζει να σημειωθεί ότι το κριτήριο Ball and Hall είναι μόνο μία από τις πολλές προσεγγίσεις για τον προσδιορισμό του αριθμού των συστάδων και η αποτελεσματικότητά του μπορεί να ποικίλλει ανάλογα με τα συγκεκριμένα χαρακτηριστικά και την κατανομή των δεδομένων.

7. **Silhouette Score:** Η βαθμολογία σιλουέτας είναι ένα μέτρο του πόσο καλά ταιριάζει κάθε σημείο δεδομένων στη συστάδα που έχει εκχωρηθεί, υποδεικνύοντας την ποιότητα μιας λύσης ομαδοποίησης. Προσδιορίζει ποσοτικά τη συμπάγεια σημεία δεδομένων εντός των συστάδων τους και τον διαχωρισμό μεταξύ διαφορετικών συστάδων. Όσο υψηλότερη είναι η βαθμολογία της σιλουέτας, τόσο καλύτερη είναι η ποιότητα της ομαδοποίησης.

Για τον υπολογισμό της βαθμολογίας σιλουέτας για ένα σημείο δεδομένων, εκτελούνται τα ακόλουθα βήματα:

- Υπολογισμός της μέσης απόστασης από άλλα σημεία δεδομένων εντός του ίδιου συμπλέγματος (a_i): Υπολογίστε τη μέση απόσταση μεταξύ του σημείου δεδομένων ενδιαφέροντος και όλων των άλλων σημείων δεδομένων στην ίδια συστάδα.
- Υπολογίστε τη μέση απόσταση από τα σημεία δεδομένων στο πλησιέστερο γειτονικό συστάδα (b_i): Προσδιορίστε τη μέση απόσταση μεταξύ του σημείου

δεδομένων ενδιαφέροντος και όλων των σημείων δεδομένων στο πλησιέστερο γειτονικό συστάδα (τη συστάδα με το οποίο η απόσταση είναι μικρότερη).

- Υπολογίστε το σκορ της σιλουέτας (s_i) για το σημείο δεδομένων: Αφαιρέστε b_i
- από το a_i και διαιρέστε το αποτέλεσμα με το μέγιστο των a_i και b_i. Ο τύπος είναι: $s_i = \frac{b_i - a_i}{\max(a_i, b_i)}$.
- Η βαθμολογία σιλουέτας για μια λύση ομαδοποίησης λαμβάνεται με τον μέσο όρο των βαθμολογιών σιλουέτας όλων των σημείων δεδομένων στο σύνολο δεδομένων. Ο τύπος είναι:

$$Silhouette\ Score = \left(\frac{1}{n}\right) * \sum_{i=1}^{n} s_i \qquad (2.2)$$

όπου *n* είναι ο συνολικός αριθμός των σημείων δεδομένων.

Η βαθμολογία σιλουέτας κυμαίνεται από -1 έως 1. Μια βαθμολογία κοντά στο 1 υποδηλώνει ότι το σημείο δεδομένων είναι καλά ομαδοποιημένο, με καλό διαχωρισμό από άλλα συστάδες. Μια βαθμολογία κοντά στο 0 υποδηλώνει ότι το σημείο δεδομένων βρίσκεται πάνω ή πολύ κοντά στο όριο απόφασης μεταξύ δύο συστάδων. Μια αρνητική βαθμολογία υποδηλώνει ότι το σημείο δεδομένων μπορεί να εκχωρηθεί σε λάθος συστάδα.

Η συνολική βαθμολογία σιλουέτας μιας λύσης ομαδοποίησης παρέχει μια αξιολόγηση της ποιότητας της ομαδοποίησης, λαμβάνοντας υπόψη τόσο τη συμπάγεια των συστάδων όσο και τον διαχωρισμό μεταξύ τους. Συχνά χρησιμοποιείται ως μέτρηση επικύρωσης για τη σύγκριση διαφορετικών αλγορίθμων ομαδοποίησης ή για τον προσδιορισμό του βέλτιστου αριθμού συστάδων επιλέγοντας τη λύση με την υψηλότερη βαθμολογία σιλουέτας.

2.1.2 Μέθοδοι μέγιστης πιθανοφάνειας

Ο γνωστότερος έλεγχος πιθανοφάνειας είναι αυτός του λόγου πιθανοφάνειας που πρότεινε ο Wolfe (1970) και συναντάται με τον όρο likelihood ratio test. Για να αποφασίσει κανείς ότι ένα σύνολο δεδομένων αποτελείται από ένα αριθμό ομάδων k_1 έναντι ενός άλλου αριθμού k_2, μπορεί να γίνει χρήση της στατιστικής συνάρτησης $-2 * \log \lambda$, όπου λ είναι ο λόγος των πιθανοφανειών $\frac{L_{k_1}}{L_{k_2}}$. Η πρόταση του Wolfe ήταν ο παρακάτω τύπος:

$$\lambda = -\frac{2}{n}\left(n - k - \frac{k_2}{2}\right) \log\left(\frac{L_{k_1}}{L_{k_2}}\right) \qquad (2.3)$$

Εδώ, το n αντιπροσωπεύει τον συνολικό αριθμό των σημείων δεδομένων, k είναι ο αριθμός των παραμέτρων στο μοντέλο, k_2 είναι ο αριθμός των παραμέτρων στο εναλλακτικό μοντέλο (με διαφορετικό αριθμό συστάδων), L_{k_1} είναι η πιθανότητα του μοντέλου με k_1 συστάδες, και L_{k_2} είναι η πιθανότητα του μοντέλου με k_2 συστάδες.

Η δοκιμή αναλογίας πιθανότητας αξιολογεί τη διαφορά στις λογαριθμικές πιθανότητες των δύο μοντέλων, προσαρμοσμένη για τον αριθμό των παραμέτρων. Μετρά τον βαθμό βελτίωσης ή προσαρμογής που επιτυγχάνεται από το μοντέλο με k_1 συστάδες σε σύγκριση με το μοντέλο με k_2 clusters.

Συγκρίνοντας την αναλογία πιθανότητας (-2 * log λ) με μια κρίσιμη τιμή από μια στατιστική κατανομή, όπως η κατανομή x_i-τετράγωνο, είναι δυνατό να προσδιοριστεί εάν η βελτίωση της προσαρμογής που επιτεύχθηκε από το μοντέλο με k_1 συστάδες είναι στατιστικά σημαντική σε σύγκριση στο μοντέλο με k_2 συστάδες. Αυτό μπορεί να βοηθήσει στον καθορισμό του κατάλληλου αριθμού συστάδων για ένα δεδομένο σύνολο δεδομένων.

Η δοκιμή αναλογίας πιθανότητας παρέχει ένα επίσημο στατιστικό πλαίσιο για σύγκριση και επιλογή μοντέλων με βάση τις πιθανότητες διαφορετικών υποθέσεων. Χρησιμοποιείται ευρέως σε διάφορους τομείς, συμπεριλαμβανομένης της ανάλυσης ομαδοποίησης, για τη λήψη αποφάσεων βάσει δεδομένων σχετικά με τον βέλτιστο αριθμό συστάδων.

Μέσω ανάλυσης Monte Carlo που διεξήχθη από τον Everitt (1981) διαπιστώθηκε ότι όταν το μέγεθος του δείγματος είναι 10 φορές μεγαλύτερο από τον αριθμό διαστάσεων, ο παραπάνω τύπος φαίνεται να αποδίδει.

Ο Day (1969) πρότεινε μια εναλλακτική μέθοδο που υποθέτει ότι τα δεδομένα θα δημιουργηθούν από ένα μείγμα δύο πολυμεταβλητών κατανομών Gauss. Αυτή η μέθοδος αναφέρεται συνήθως ως μέθοδος ημέρας ή ομαδοποίηση ημέρας. Η προσέγγιση υποθέτει ότι το σύνολο δεδομένων αποτελείται από παρατηρήσεις που είναι ένα μείγμα δύο υποκείμενων πολυμεταβλητών κατανομών Gauss. Κάθε κατανομή Gauss αντιπροσωπεύει ένα ξεχωριστό συστάδα στα δεδομένα.

Η μέθοδος Day's στοχεύει να εκτιμήσει τις παραμέτρους των κατανομών Gauss και να εκχωρήσει σημεία δεδομένων στις αντίστοιχες συστάδες τους. Οι παράμετροι τυπικά περιλαμβάνουν το μέσο διάνυσμα και τον πίνακα συνδιακύμανσης για κάθε Gaussian στοιχείο. Η διαδικασία ομαδοποίησης στη μέθοδο Day's περιλαμβάνει τα ακόλουθα βήματα:

Εκτίμηση παραμέτρων: Υπολογίστε τις παραμέτρους των δύο πολυμεταβλητών κατανομών Gauss. Αυτό γίνεται συνήθως χρησιμοποιώντας τεχνικές όπως ο αλγόριθμος Προσδοκίας-Μεγιστοποίησης (EM) ή η εκτίμηση μέγιστης πιθανότητας.

Υπολογισμός Πιθανοτήτων: Υπολογίστε την πιθανότητα κάθε σημείο δεδομένων να ανήκει σε καθεμία από τις δύο συνιστώσες Gauss. Αυτό επιτυγχάνεται με την αξιολόγηση της συνάρτησης πυκνότητας πιθανότητας των πολυμεταβλητών κατανομών Gauss για κάθε σημείο δεδομένων.

Εκχώρηση συστάδων: Αντιστοιχίστε κάθε σημείο δεδομένων στη συστάδα με την υψηλότερη πιθανότητα. Εάν η πιθανότητα να ανήκει στην πρώτη συνιστώσα Gauss είναι μεγαλύτερη από την πιθανότητα της δεύτερης συνιστώσας, το σημείο δεδομένων εκχωρείται στην πρώτη συστάδα και αντίστροφα.

Η μέθοδος Day's υποθέτει ότι τα σημεία δεδομένων σε κάθε συστάδα ακολουθούν μια πολυμεταβλητή κατανομή Gauss και η ομαδοποίηση πραγματοποιείται με εκτίμηση των παραμέτρων αυτών των κατανομών. Παρέχει ένα πιθανό πλαίσιο για ομαδοποίηση και υποθέτει ότι τα δεδομένα μπορούν να χαρακτηριστούν καλά από ένα μείγμα δύο κατανομών Gauss.Αυτή η μέθοδος έχει χρησιμοποιηθεί σε διάφορες εφαρμογές και μπορεί να επεκταθεί για να χειριστεί σύνολα δεδομένων με περισσότερες από δύο συστάδες λαμβάνοντας υπόψη μείγματα πολλαπλών κατανομών Gauss. Αξίζει να σημειωθεί ότι εφόσον η μέθοδος Day προϋποθέτει μια συγκεκριμένη υποκείμενη κατανομή (μείγμα Gaussian), η αποτελεσματικότητά της μπορεί να εξαρτάται από την πραγματική κατανομή των δεδομένων. Άλλοι αλγόριθμοι ομαδοποίησης, όπως μέθοδοι που βασίζονται σε πυκνότητα ή κεντροειδείς, είναι συχνά πιο ευέλικτοι και μπορούν να χειριστούν διάφορους τύπους διανομών δεδομένων.

2.1.3 Μέθοδοι ανάλυσης διακύμανσης

Πολλά σύνολα δεδομένων αποτελούνται από πολυμεταβλητές συνεχείς παρατηρήσεις. Μετά το διαχωρισμό των δεδομένων σε ομάδες, το αποτέλεσμα μοιάζει με αυτό της πολυμεταβλητής στατιστικής ανάλυσης διακύμανσης (MANOVA).

Αναλύοντας τις συνολικές τετραγωνικές αποκλίσεις σε δύο επιμέρους τμήματα μπορούμε να κατασκευάσουμε το ένα τμήμα ώστε να δείχνει τις αποκλίσεις μέσα στην ομάδα

$$W = \sum_{j=1}^{k} \sum_{i=1}^{n_j} (\boldsymbol{x}_{ij} - \boldsymbol{x}_j)(\boldsymbol{x}_{ij} - \bar{\boldsymbol{x}_j})' \quad (2.4)$$

Εδώ, το k αντιπροσωπεύει τον συνολικό αριθμό των ομάδων, το n_j είναι ο αριθμός των παρατηρήσεων στην ομάδα j, το x_{ij} είναι η ίη παρατήρηση στην ομάδα j, το x_j είναι ο μέσος όρος της ομάδας j και το $(x_j)'$ υποδηλώνει τη μετάθεση του μέσου διανύσματος για ομάδα j.

Οι αποκλίσεις εντός της ομάδας καταγράφουν τη μεταβλητότητα των παρατηρήσεων σε κάθε ομάδα και παρέχουν μια εικόνα για τη διασπορά των σημείων δεδομένων εντός των συστάδων.

Το άλλο τμήμα να δείχνει τις αποκλίσεις μεταξύ των ομάδων

$$B = \sum_{j=1}^{k} n_j (\bar{\boldsymbol{x}_j} - \bar{\boldsymbol{x}})(\bar{\boldsymbol{x}_j} - \bar{\boldsymbol{x}})' \quad (2.5)$$

Εδώ, το n_j αντιπροσωπεύει τον αριθμό των παρατηρήσεων στην ομάδα j, $\boldsymbol{x}_j$ είναι ο μέσος όρος της ομάδας j, $\boldsymbol{x}$ είναι ο συνολικός μέσος όρος σε όλες τις ομάδες και $(\bar{\boldsymbol{x}}_j)'$ και $(\bar{\boldsymbol{x}})'$ υποδηλώνουν τη μεταφορά του μέσου διανύσματος της ομάδας και του συνολικού μέσου όρου διάνυσμα, αντίστοιχα.

Οι διαφορές μεταξύ των ομάδων καταγράφουν τις διακυμάνσεις μεταξύ των μέσων της ομάδας και παρέχουν πληροφορίες σχετικά με το πόσο διαφορετικές είναι οι ομάδες μεταξύ τους. Με την αποσύνθεση των συνολικών τετραγωνικών αποκλίσεων σε αποκλίσεις εντός της ομάδας (*W*) και διαφορές μεταξύ ομάδων (*B*), η δομή τύπου MANOVA βοηθά στην κατανόηση των πηγών διακύμανσης στα δεδομένα πολλαπλών μεταβλητών μεταξύ διαφορετικών ομάδων ή συστάδων. Επιτρέπει την αξιολόγηση της μεταβλητότητας τόσο εντός της ομάδας όσο και μεταξύ της ομάδας, η οποία είναι πολύτιμη για στατιστικά συμπεράσματα και ερμηνεία των αποτελεσμάτων ομαδοποίησης.

Οι πίνακες *B, W* που προκύπτουν μπορούν να εκφράσουν την επιτυχία ή μη της ομαδοποίησης. Όταν ο πίνακας *B* είναι όσο το δυνατόν πιο μεγάλος και ο *W* πιο μικρός, βάσει κάποιου προκαθορισμένου κριτηρίου σύγκρισης πινάκων, τόσο πιο επιτυχημένη θεωρείται μια ομαδοποίηση. Στην ομαδοποίηση, ο στόχος είναι να μεγιστοποιηθούν οι διαφορές μεταξύ των συστάδων (όπως αντικατοπτρίζονται από το *B*) ενώ ελαχιστοποιούνται οι διακυμάνσεις σε κάθε συστάδα (όπως αντικατοπτρίζονται από το *W*). Αυτό βοηθά να διασφαλιστεί ότι τα συμπλέγματα είναι ευδιάκριτα και καλά διαχωρισμένα, καταγράφοντας σημαντικά μοτίβα ή δομή στα δεδομένα.

Συγκρίνοντας τα μεγέθη των πινάκων *B* και *W* με βάση κάποιο προκαθορισμένο κριτήριο, μπορεί κανείς να αξιολογήσει την ποιότητα της ομαδοποίησης. Μπορούν να χρησιμοποιηθούν διαφορετικά κριτήρια για σύγκριση, ανάλογα με τους συγκεκριμένους στόχους και τις απαιτήσεις της εργασίας ομαδοποίησης.

Για παράδειγμα, ορισμένα κοινά κριτήρια σύγκρισης περιλαμβάνουν:

Λόγος διακύμανσης: Ο λόγος διακύμανσης υπολογίζεται ως ο λόγος της διακύμανσης μεταξύ της ομάδας (*B*) προς τη διακύμανση εντός της ομάδας (*W*). Ένας μεγαλύτερος λόγος διακύμανσης υποδηλώνει καλύτερο διαχωρισμό μεταξύ των συστάδων και προτείνει μια πιο επιτυχημένη ομαδοποίηση.

Στατιστική F: Η στατιστική F υπολογίζεται διαιρώντας το άθροισμα των τετραγώνων μεταξύ της ομάδας (B) με το άθροισμα τετραγώνων εντός της ομάδας (*W*) και προσαρμόζοντας τους βαθμούς ελευθερίας. Μια υψηλότερη στατιστική F υποδηλώνει καλύτερο διαχωρισμό μεταξύ των ομάδων και υποδηλώνει μια πιο επιτυχημένη ομαδοποίηση.

Βαθμολογία σιλουέτας: Η βαθμολογία σιλουέτας, όπως συζητήθηκε προηγουμένως, παρέχει ένα μέτρο της συμπάγειας εντός των συστάδων (*W*) και του διαχωρισμού μεταξύ των συστάδων (*B*). Μια υψηλότερη βαθμολογία σιλουέτας υποδηλώνει καλύτερα αποτελέσματα ομαδοποίησης.

Αυτά τα κριτήρια σύγκρισης μπορούν να χρησιμοποιηθούν για την αξιολόγηση της επιτυχίας μιας λύσης ομαδοποίησης ποσοτικοποιώντας την αντιστάθμιση μεταξύ των διαφορών μεταξύ της ομάδας (*B*) και των παραλλαγών εντός της ομάδας (*W*). Όσο μεγαλύτερος είναι ο πίνακας *B* και όσο μικρότερος ο πίνακας *W*, όπως καθορίζεται από το επιλεγμένο κριτήριο, τόσο πιο επιτυχημένη θεωρείται η ομαδοποίηση.

Είναι σημαντικό να σημειωθεί ότι η επιλογή του κριτηρίου για την αξιολόγηση της επιτυχίας της ομαδοποίησης εξαρτάται από το συγκεκριμένο πλαίσιο, τους στόχους και τα χαρακτηριστικά των δεδομένων. Διαφορετικά κριτήρια μπορεί να είναι κατάλληλα για διαφορετικά σενάρια και είναι σημαντικό να επιλέξετε το καταλληλότερο κριτήριο με βάση τις συγκεκριμένες απαιτήσεις της εργασίας ομαδοποίησης.

Χρησιμοποιώντας ιδιότητες όπως οι προαναφερθείσες μπορούμε μέσω του ίχνους ή της ορίζουσας των πινάκων να κατασκευάσουμε κριτήρια για την εύρεση της βέλτιστης ομαδοποίησης. Παράλληλα όμως, πρέπει να ληφθεί υπόψη ο αριθμός των ομάδων στις οποίες οδηγεί η ελαχιστοποίηση του μεγέθους του πίνακα, ειδάλλως μπορεί να οδηγηθούμε σε μεγάλο αριθμό προκυπτουσών ομάδων.

Κριτήριο Calinski & Harabasz

Το κριτήριο των Calinski-Harabasz (1974) είναι ένα από τα πιο γνωστά κριτήρια διακοπής. Η ομαδοποίηση για την οποία ο ακόλουθος τύπος μεγιστοποιείται μας δίνει το βέλτιστο αριθμό ομάδων που ισούται με *k.*

$$c = \frac{\frac{trace(B)}{k-1}}{\frac{trace(W)}{n-k}} \tag{2.6}$$

Ο υπολογισμός του μέγιστου λόγου μας δίνει ομάδες με όμοιες παρατηρήσεις μέσα στην ομάδα, και παράλληλα μεγάλες διαφορές ανάμεσα στις ομάδες. Σε αυτό το σενάριο, οι τιμές του πίνακα *B* είναι μεγάλες και οι τιμές του *W* είναι μικρές.

Παραδοσιακά, ο αλγόριθμος Calinski-Harabasz δεν εφαρμόστηκε εκτενώς, ειδικά σε μεγάλα σύνολα δεδομένων, λόγω των υπολογιστικών του απαιτήσεων. Ο αλγόριθμος περιλαμβάνει τον υπολογισμό διαφόρων αποστάσεων, τον υπολογισμό πινάκων διασποράς συστάδων και την εκτέλεση υπολογισμών ιδιοτιμών, οι οποίοι μπορεί να είναι υπολογιστικά εντατικοί για μεγαλύτερα σύνολα δεδομένων. Ωστόσο, με τις προόδους στην τεχνολογία υλικού (hardware technology) και την αυξημένη οικονομική προσιτότητα των ισχυρών υπολογιστικών πόρων, οι υπολογιστικές απαιτήσεις του αλγορίθμου Calinski-Harabasz έχουν γίνει πιο διαχειρίσιμες. Αυτό έχει κάνει τον αλγόριθμο πιο εφαρμόσιμο και εφικτό σε ένα ευρύτερο φάσμα περιπτώσεων, συμπεριλαμβανομένων μεγαλύτερων συνόλων δεδομένων. Η ταχεία βελτίωση του hardware, όπως οι ταχύτεροι επεξεργαστές, οι μεγαλύτερες

χωρητικότητες μνήμης και οι παράλληλες υπολογιστικές δυνατότητες, έχει μειώσει σημαντικά τον υπολογιστικό χρόνο που απαιτείται για την εκτέλεση του αλγορίθμου.

Επιπλέον, το κόστος απόκτησης τέτοιου υλικού έχει μειωθεί με την πάροδο του χρόνου, καθιστώντας το πιο προσιτό σε ερευνητές και επαγγελματίες. Ως αποτέλεσμα, ο αλγόριθμος Calinski-Harabasz έχει αποκτήσει δημοτικότητα και πλέον εφαρμόζεται ευρύτερα, συμπεριλαμβανομένων περιπτώσεων που αφορούν μεγαλύτερα σύνολα δεδομένων. Παρέχει ένα αξιόπιστο και ακριβές μέτρο της ποιότητας της ομαδοποίησης, λαμβάνοντας υπόψη τόσο τη συμπάγεια των συστάδων όσο και τον διαχωρισμό μεταξύ των συστάδων. Η δυνατότητα εφαρμογής του αλγορίθμου έχει επεκταθεί με τη διαθεσιμότητα πιο αποτελεσματικών πόρων υλικού, επιτρέποντας στους ερευνητές και τους επαγγελματίες να αξιολογούν αποτελεσματικά τα αποτελέσματα της ομαδοποίησης.

Αξίζει να σημειωθεί ότι ενώ ο αλγόριθμος Calinski-Harabasz είναι υπολογιστικά απαιτητικός, παραμένει ένα πολύτιμο εργαλείο για την αξιολόγηση της ποιότητας της ομαδοποίησης. Επιπλέον, περαιτέρω πρόοδοι στην τεχνολογία υλικού και στις υπολογιστικές τεχνικές είναι πιθανό να συνεχίσουν να διευκολύνουν την εφαρμογή του αλγορίθμου σε ακόμη μεγαλύτερα και πιο σύνθετα σύνολα δεδομένων.

Κριτήριο Trace W

Οι Edwards και Sforza (1965) εισήγαγαν το κριτήριο Trace *W* για την εύρεση ενός σημείου διακοπής. Το κριτήριο προτείνει την εύρεση της ελάχιστης τιμής του ίχνους του πίνακα *W*, με σκοπό να βρεθεί ο βέλτιστος αριθμός των ομάδων.

Για κάθε ομαδοποίηση ισχύει η ισότητα $T = W + B$, και καθώς ο πίνακας *T* είναι σταθερός για όλες τις ομάδες, η ελαχιστοποίηση του ίχνους του *W* ισοδυναμεί με τη μεγιστοποίηση του ίχνους του B σύμφωνα με την ιδιότητα του ίχνους των πινάκων $\text{Trace}(T) = \text{Trace}(W) + \text{Trace}(B)$

Κριτήρια Trace $W^{-1}B$ και $\frac{|T|}{|W|}$

Τα κριτήρια Trace $W^{-1}B$ και $\frac{|T|}{|W|}$ προτάθηκαν από τους Friedman και Rubin (1967), και η μεγιστοποίησή τους οδηγεί στην εύρεση του βέλτιστου αριθμού των ομάδων. Το πλεονέκτημα των κριτηρίων αυτών είναι ότι παραμένουν αμετάβλητα για κάθε γραμμικό μετασχηματισμό των αρχικών δεδομένων. Μια βελτίωση του δεύτερου τύπου προτάθηκε από τον Symons (1971) με τον ακόλουθο τύπο $nlog\left(\frac{|T|}{|W|}\right)$, όπου *n* ο αριθμός των παρατηρήσεων του συνόλου δεδομένων. Αυτή η παραλλαγή βρέθηκε ότι οδηγεί σε καλύτερα αποτελέσματα σύμφωνα με μια μελέτη που εκπονήθηκε από τον Arnold (1979). Υπάρχουν μερικοί λόγοι για τους οποίους αυτή η παραλλαγή της φόρμουλας μπορεί να οδηγήσει σε καλύτερα αποτελέσματα:

Ποινικοποίηση μικρών μεγεθών δείγματος: Με την ενσωμάτωση του αριθμού των παρατηρήσεων (n) στον τύπο, η βελτίωση του Symons τιμωρεί λύσεις με μικρότερα μεγέθη δείγματος. Όταν το μέγεθος του δείγματος είναι μικρό, υπάρχει μεγαλύτερη πιθανότητα υπερπροσαρμογής και λήψης λιγότερο αξιόπιστων αποτελεσμάτων ομαδοποίησης. Η συμπερίληψη του *n* στον τύπο βοηθά στην αποφυγή υπερβολικής πολυπλοκότητας στη λύση ομαδοποίησης και προωθεί την πιο ισχυρή και ουσιαστική ομαδοποίηση.

Προκατάληψη προς πιο φειδωλές λύσεις: Η συμπερίληψη του $nlog\left(\frac{|T|}{|W|}\right)$, στον τύπο εισάγει μια ποινή για υπερβολικά πολύπλοκες λύσεις. Η μεγιστοποίηση αυτής της εκλεπτυσμένης φόρμουλας ενθαρρύνει την εύρεση λύσεων ομαδοποίησης που είναι πιο φειδωλές και ερμηνεύσιμες. Η παρρησία αναφέρεται στην απλότητα ή στη χρήση λιγότερων παραμέτρων ή συστάδων για την εξήγηση των δεδομένων. Με την προώθηση της φειδωλότητας, η εκλεπτυσμένη φόρμουλα βοηθά στην αποφυγή υπερβολικής προσαρμογής και παρέχει πιο αξιόπιστα και ουσιαστικά αποτελέσματα ομαδοποίησης.

Συνέπεια μεταξύ των γραμμικών μετασχηματισμών: Ένα πλεονέκτημα τόσο των αρχικών κριτηρίων ανίχνευσης $W^{(-1)B}$ όσο και της βελτίωσης του Symons είναι ότι παραμένουν αμετάβλητα για κάθε γραμμικό μετασχηματισμό των αρχικών δεδομένων. Οι γραμμικοί μετασχηματισμοί περιλαμβάνουν λειτουργίες όπως κλιμάκωση, περιστροφή και μετάφραση των δεδομένων. Αυτή η ιδιότητα διασφαλίζει ότι τα κριτήρια αξιολόγησης είναι αμετάβλητα στους γραμμικούς μετασχηματισμούς, καθιστώντας τα πιο ισχυρά και εφαρμόσιμα σε διαφορετικές αναπαραστάσεις δεδομένων ή συστήματα συντεταγμένων.

Συνολικά, η βελτίωση που προτείνει ο Symons, συμπεριλαμβανομένου του $nlog\left(\frac{|T|}{|W|}\right)$ στον τύπο, μπορεί να οδηγήσει σε καλύτερα αποτελέσματα ομαδοποίησης τιμωρώντας μικρά μεγέθη δειγμάτων, ευνοώντας λιτές λύσεις και παρέχοντας συνέπεια μεταξύ των γραμμικών μετασχηματισμών. Αυτοί οι παράγοντες συμβάλλουν σε βελτιωμένες λύσεις ομαδοποίησης που είναι πιο αξιόπιστες, ερμηνεύσιμες και γενικεύσιμες σε διαφορετικά σύνολα δεδομένων και μετασχηματισμούς.

Δείκτης Duda & Hart

Οι Duda και Hart (1973) πρότειναν τον ακόλουθο λόγο για την εύρεση του βέλτιστου αριθμού ομάδων: $\frac{SSE(2)}{SSE(1)} < Cr$. Ο αριθμητής είναι το άθροισμα των τετραγωνικών σφαλμάτων μέσα στην ομάδα όταν τα δεδομένα είναι χωρισμένα σε δύο ομάδες. Ο παρονομαστής είναι το αντίστοιχο άθροισμα μόνο για μία ομάδα. Όταν ο λόγος είναι μικρότερος από την κρίσιμη τιμή *Cr*, η υπόθεση ότι υπάρχει μία ομάδα αντί για δύο απορρίπτεται.

Η εφαρμογή για τη διερεύνηση του βέλτιστου αριθμού συστάδων είναι ότι αυτή η αναλογία βοηθά στον προσδιορισμό του αριθμού των συστάδων που παρέχουν την πιο ουσιαστική βελτίωση όσον αφορά τη μείωση της διακύμανσης εντός της ομάδας. Όταν ο λόγος είναι κάτω

από την κρίσιμη τιμή, υποδηλώνει ότι τα δεδομένα μπορούν να επεξηγηθούν ή να αναπαρασταθούν καλύτερα χωρίζοντάς τα σε δύο διακριτές συστάδες.

Για τη διερεύνηση του βέλτιστου αριθμού συστάδων, μπορεί κανείς να εφαρμόσει αυτή την αναλογία για διαφορετικούς αριθμούς συστάδων, όπως η σύγκριση SSE(3)/SSE(1), SSE(4)/SSE(1) και ούτω καθεξής. Εξετάζοντας πώς αλλάζει ο λόγος με τον αριθμό των συστάδων και συγκρίνοντάς τον με την κρίσιμη τιμή, μπορεί κανείς να προσδιορίσει τον αριθμό των συστάδων που οδηγούν στην πιο σημαντική μείωση της διακύμανσης εντός της ομάδας και, επομένως, στην πιο βέλτιστη λύση ομαδοποίησης.

Είναι σημαντικό να σημειωθεί ότι η επιλογή της κρίσιμης τιμής (*Cr*) είναι κρίσιμη και εξαρτάται από το συγκεκριμένο πλαίσιο, τους στόχους και το επιθυμητό επίπεδο σημασίας. Επιπλέον, αυτός ο λόγος είναι μόνο ένα από τα πολλά διαθέσιμα κριτήρια για τον προσδιορισμό του βέλτιστου αριθμού συμπλεγμάτων και η αποτελεσματικότητά του μπορεί να εξαρτάται από τα χαρακτηριστικά του συνόλου δεδομένων και τον αλγόριθμο ομαδοποίησης που χρησιμοποιείται.

Κριτήριο των Frey and Van Groenewoud

Ο 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:

$$K = \frac{\overline{du}_{j+1} - \overline{du}_j}{\overline{ds}_{j+1} - \overline{ds}_j} \tag{2.7}$$

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.

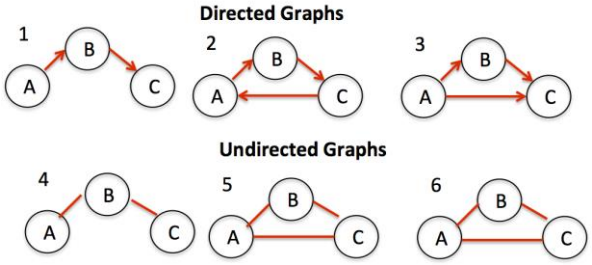
Κριτήριο Marriot

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.

2.1.4 Γραφικές Μέθοδοι

Προτού αναφερθούμε σε μεθόδους γράφων θα χρησιμοποιήσουμε τον ορισμό ενός γράφου από τα διακριτά μαθηματικά. Ένας **γράφος (G)** είναι μια αφηρημένη αναπαράσταση ενός συνόλου στοιχείων συνδεδεμένων μεταξύ τους. Τα στοιχεία του ονομάζονται **κορυφές (K)** και οι δεσμοί **ακμές (A).** Οι ακμές ενός γράφου μπορούν να είναι κατευθυνόμενες ή μη κατευθυνόμενες.



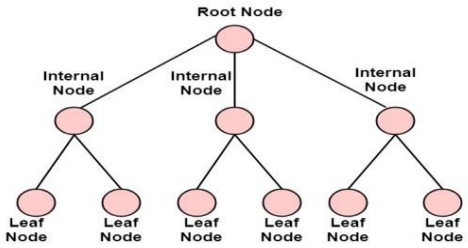
Σχήμα 2.1 Κατευθυνόμενες και μη κατευθυνόμενες ακμές γράφου

Ένας πλήρης γράφος μπορεί να αναπαρασταθεί ως ένας πίνακας γειτνίασης (n * n) στοιχείων, όπου τα σημεία ($\mathbf{x_1}$,$\mathbf{x_2}$,..,$\mathbf{x_n}$) αντιστοιχούν στις γραμμές και στήλες του και ο αριθμός δεσμών που περιέχει ισούται με

$$\binom{n}{2} = \frac{n(n-1)}{2}. \qquad (2.8)$$

Σε ένα μη κατευθυνόμενο γράφημα, κάθε ακμή μπορεί να αναπαρασταθεί από ένα ζεύγος διακριτών σημείων (x_i, x_j). Λαμβάνοντας υπόψη όλα τα πιθανά ζεύγη σημείων, ο συνολικός αριθμός των ακμών ή των συνδέσμων σε ένα πλήρες γράφημα δίνεται από το " $\binom{n}{2}$". Αυτό σημαίνει ότι ένα πλήρες γράφημα με 4 σημεία θα έχει 6 άκρες ή συνδέσμους που συνδέουν όλα τα πιθανά ζεύγη σημείων.

Αντίστοιχα ένα **δέντρο** είναι ένας γράφος ο οποίος δεν περιέχει κύκλους, δηλαδή δεν υπάρχει δεσμός που να συνδέει δύο μη διαδοχικά σημεία. Ένα δέντρο αποτελείται από μια **ρίζα**, η οποία αποτελεί τον αρχικό κόμβο. Συνεχίζοντας προς τα κάτω οι δεσμοί μεταξύ των κόμβων ονομάζονται **κλαδιά** και καταλήγουν σε **φύλλα**, τα οποία αποτελούν τα άκρα. Μια διαδρομή από ένα κόμβο του δέντρου προς ένα άλλο ονομάζεται η διαδοχή των δεσμών από την εναρκτήρια κορυφή προς μια άλλη. Ένα δέντρο ***n*** κόμβων αποτελείται από *n* - 1 δεσμούς.



Σχήμα 7 Απεικόνιση δέντρου n κόμβων

Μία πολύ σημαντική έννοια αποτελεί το δέντρο ελαχίστων αποστάσεων (Δ.Ε.Α) ή Minimum Spanning Tree. Ένα δέντρο ελαχίστων αποστάσεων προκύπτει από ένα πλήρες γράφημα λαμβάνοντας υπόψη τις ελάχιστες αποστάσεις που συνδέουν τους κόμβους. Διατάσσοντας $n - 2$ συνδέσμους του γραφήματος κατά αύξουσα σειρά και επιλέγοντας τους $n - 1$ που δεν σχηματίζουν κύκλους δημιουργείται το Δ.Ε.Α. Αντίστοιχα, μπορεί να σχηματιστεί και το δέντρο μεγίστων αποστάσεων (Δ.Μ.Α) εάν η παραπάνω διάταξη γίνει σε φθίνουσα σειρά.

Το Minimum Spanning Tree (MST) είναι μια θεμελιώδης έννοια στη θεωρία γραφημάτων και στη βελτιστοποίηση δικτύου. Είναι ένα υποσύνολο ακμών από ένα πλήρες γράφημα που συνδέει όλους τους κόμβους μαζί με το ελάχιστο συνολικό βάρος ή απόσταση ακμών.

Για να αποκτήσουμε ένα ελάχιστο εκτεινόμενο δέντρο, ξεκινάμε με ένα πλήρες γράφημα όπου κάθε κόμβος συνδέεται με κάθε άλλο κόμβο. Η διαδικασία περιλαμβάνει την επιλογή ακμών με τα μικρότερα βάρη, διασφαλίζοντας παράλληλα ότι το δέντρο που προκύπτει παραμένει συνδεδεμένο και δεν σχηματίζει κύκλους.

Ένας συνηθισμένος αλγόριθμος που χρησιμοποιείται για την εύρεση του ελάχιστου εκτεινόμενου δέντρου είναι ο αλγόριθμος του Kruskal. Ακολουθεί αυτά τα βήματα:

- Ταξινόμηση των άκρων του πλήρους γραφήματος σε αύξουσα σειρά με βάση το βάρος τους.
- Ξεκινήστε με ένα κενό γράφημα (δέντρο) που δεν περιέχει άκρες.
- Εξετάστε τις άκρες με τη σειρά ταξινόμησης και προσθέστε τις στο δέντρο μία προς μία, ξεκινώντας από τη μικρότερη σταθμισμένη άκρη.
- Για κάθε εξεταζόμενη άκρη, ελέγξτε εάν η προσθήκη της θα δημιουργήσει έναν κύκλο στο τρέχον δέντρο. Αν όχι, προσθέστε την άκρη στο δέντρο.
- Επαναλάβετε το βήμα 4 έως ότου το δέντρο έχει $(n - 1)$ άκρες, όπου n είναι ο αριθμός των κόμβων στο αρχικό γράφημα.

Το προκύπτον δέντρο που λαμβάνεται μέσω αυτής της διαδικασίας είναι το ελάχιστο εκτεινόμενο δέντρο, όπου ελαχιστοποιείται το συνολικό βάρος ή η απόσταση των άκρων.

Ομοίως, μπορούμε να δημιουργήσουμε ένα μέγιστο εκτεινόμενο δέντρο (MST) αντιστρέφοντας τη σειρά ταξινόμησης των άκρων. Αντί να επιλέγουμε τις μικρότερες

σταθμισμένες ακμές, επιλέγουμε τις μεγαλύτερες σταθμισμένες άκρες, ενώ παράλληλα διασφαλίζουμε ότι το δέντρο που προκύπτει είναι συνδεδεμένο και δεν περιέχει κύκλους.

Ακολουθεί ένα παράδειγμα για την απεικόνιση της έννοιας:

Θεωρήστε ένα πλήρες γράφημα με τέσσερις κόμβους (A, B, C, D) και οι αποστάσεις μεταξύ τους είναι οι εξής:

AB: 2 AC: 4 AD: 3 BC: 5 BD: 1 CD: 6

Για να βρούμε το ελάχιστο εκτεινόμενο δέντρο χρησιμοποιώντας τον αλγόριθμο του Kruskal, ταξινομούμε τις άκρες σε αύξουσα σειρά:

BD: 1 AB: 2 AD: 3 AC: 4 BC: 5 CD: 6

Ξεκινώντας με ένα άδειο δέντρο, προσθέτουμε τις άκρες μία-μία χωρίς να δημιουργούμε κύκλους μέχρι να έχουμε (n-1) άκρες. Σε αυτήν την περίπτωση, το ελάχιστο εκτεινόμενο δέντρο θα ήταν:

BD: 1 AB: 2 AD: 3

Αυτό το ελάχιστο εκτεινόμενο δέντρο συνδέει όλους τους κόμβους (A, B, D) με την ελάχιστη συνολική απόσταση 6.
Παρομοίως, μπορούμε να δημιουργήσουμε ένα μέγιστο εκτεινόμενο δέντρο αντιστρέφοντας τη σειρά ταξινόμησης. Για αυτό το παράδειγμα, το μέγιστο εκτεινόμενο δέντρο θα ήταν:

CD: 6 BC: 5 AC: 4

Το μέγιστο εκτεινόμενο δέντρο συνδέει όλους τους κόμβους (C, B, A) με μέγιστη συνολική απόσταση 15. Τόσο το δέντρο ελάχιστης έκτασης όσο και το δέντρο μέγιστης έκτασης παρέχουν χρήσιμες πληροφορίες για την υποκείμενη δομή ενός γραφήματος, επισημαίνοντας τις πιο αποτελεσματικές ή τις πιο εκτεταμένες συνδέσεις μεταξύ κόμβων.

Παρακάτω δίνονται κάποιοι αλγόριθμοι που χρησιμοποιούν τέτοιες τεχνικές.

Κριτήριο GRAPH

Ο αλγόριθμος GRAPH λειτουργεί χρησιμοποιώντας τις δύο παραπάνω μεθόδους Δ.Ε.Α και Δ.Μ.Α. Μέσω του Δ.Ε.Α. αναπαρίσταται ο αλγόριθμος Single Linkage ο οποίος για να ενώσει ***n*** ομάδες αξιολογεί τις ελάχιστες αποστάσεις μεταξύ τους, χρησιμοποιώντας τις κορυφές και τους δεσμούς του πρώτου.

Ακόμη μέσω του Δ.Μ.Α. ελέγχει αν υπάρχει σύνδεση μεταξύ δύο σημείων που ανήκουν σε κάθε μια από τις ομάδες που πρόκειται να ενωθούν, χρησιμοποιώντας αυτή τη συνθήκη ως κριτήριο διακοπής.

Μπορούμε να περιγράψουμε τη διαδικασία που ακολουθεί ο αλγόριθμος με τα εξής βήματα:

1. Αρχικοποίηση Δ.Ε.Α και Δ.Μ.Α
2. Ένωση των κοντινότερων σημείων
3. Εξαγωγή του δεσμού με τη μικρότερη τιμή από το Δ.Ε.Α.
4. Σε περίπτωση που υπάρχει τουλάχιστον μία ομάδα που συνδέεται με συνδέσμους του Δ.Μ.Α., θεωρείται ότι βρέθηκε λύση. Αντίθετα, τα βήματα 3-4 επαναλαμβάνονται μέχρι να ικανοποιηθεί η προηγούμενη συνθήκη.

Για την περαιτέρω αποσαφήνιση της διαδικασίας του αλγορίθμου GRAPH χρησιμοποιώντας το Δ.Ε.Α. και Δ.Μ.Α. μεθόδους, ας δούμε ένα απλό παράδειγμα.

Εξετάζουμε ένα σύνολο δεδομένων με έξι σημεία (A, B, C, D, E, F) και θέλουμε να εφαρμόσουμε τον αλγόριθμο Single Linkage για να ομαδοποιήσουμε αυτά τα σημεία. Θα χρησιμοποιήσουμε το Δ.Ε.Α. και Δ.Μ.Α. μεθόδους εντός του αλγόριθμου GRAPH για την εκτέλεση της ομαδοποίησης.

Αρχικοποίηση:

Ξεκινήστε με κάθε σημείο ως μεμονωμένη ομάδα: [A], [B], [C], [D], [E], [F].

Αρχικοποιήστε το Δ.Ε.Α. μητρώο, το οποίο αντιπροσωπεύει τις αποστάσεις μεταξύ των ομάδων, με βάση τις αποστάσεις μεταξύ μεμονωμένων σημείων.

Αρχικοποιήστε το Δ.Μ.Α. μητρώο, που αντιπροσωπεύει τις συνδέσεις μεταξύ σημείων διαφορετικών ομάδων, με βάση τις αρχικές αναθέσεις ομάδων.

Ένωση των πλησιέστερων σημείων:

Βρείτε την ελάχιστη απόσταση στη Δ.Ε.Α. μήτρα, που αντιπροσωπεύει τη μικρότερη απόσταση μεταξύ οποιωνδήποτε δύο ομάδων.

Συγχωνεύστε τις δύο ομάδες που αντιστοιχούν σε αυτή την ελάχιστη απόσταση.

Ενημέρωση του μητρώου Δ.Ε.Α και Δ.Μ.Α. με βάση τις νέες ομαδικές αναθέσεις.

Εξαγωγή του συνδέσμου με τη μικρότερη τιμή από Δ.Ε.Α.:

Επαναλάβετε το βήμα 2 μέχρι να συγχωνεύσετε όλες τις ομάδες σε ένα ενιαίο σύμπλεγμα ή μέχρι να ικανοποιηθεί ένα κριτήριο διακοπής.

Κριτήριο διακοπής (συνθήκη D.M.A.):

Σε κάθε βήμα, ελέγξτε εάν υπάρχει σύνδεση μεταξύ δύο σημείων που ανήκουν σε διαφορετικές συστάδες στο Δ.Μ.Α μητρώο.

Αν υπάρχει τέτοια σύνδεση, θεωρήστε την ως λύση.

Διαφορετικά, συνεχίστε τη συγχώνευση ομάδων και την ενημέρωση των πινάκων μέχρι να ικανοποιηθεί το κριτήριο διακοπής.

Ας επεξηγήσουμε αυτά τα βήματα με ένα απλό παράδειγμα χρησιμοποιώντας Ευκλείδεια απόσταση:

Δεδομένων των αποστάσεων κατά ζεύγη:

AB: 2	BC: 5	CE: 3
AC: 4	BD: 1	CF: 8
AD: 3	BE: 5	ΔE: 2
AE: 6	BF: 6	DF: 4
AF: 7	CD: 3	EF: 9

Πίνακας 1. Παράδειγμα με αποστάσεις ζευγών

Αρχικοποίηση:

Ξεκινήστε με κάθε σημείο ως μεμονωμένη ομάδα: [A], [B], [C], [D], [E], [F].

Αρχικοποιήστε το μητρώο Δ.Ε.Α. και Δ.Μ.Α. με βάση τις ζευγαρωμένες αποστάσεις.

Ένωση των πλησιέστερων σημείων:

Βρείτε την ελάχιστη απόσταση στο μητρώο Δ.Ε.Α.: BD (απόσταση = 1).

Συγχώνευση ομάδων B και D: [B, D], [A], [C], [E], [F].

Ενημέρωση των μητρώων Δ.Ε.Α. και Δ.Μ.Α.

Εξαγωγή του συνδέσμου με τη μικρότερη τιμή από Δ.Ε.Α.:

Βρείτε την ελάχιστη απόσταση στην ενημερωμένη Δ.Ε.Α. μητρώο: DE (απόσταση = 2).

Συγχώνευση ομάδων D και E: [B, D, E], [A], [C], [F].

Ενημέρωση των μητρώων Δ.Ε.Α. και Δ.Μ.Α.

Κριτήριο διακοπής (συνθήκη Δ.Μ.Α.):

Ελέγξτε το μητρώο Δ.Μ.Α.: Δεν υπάρχει σύνδεση μεταξύ σημείων διαφορετικών ομάδων.

Συνέχιση της συγχώνευσης με βάση το μητρώο Δ.Ε.Α.

Εξαγωγή του συνδέσμου με τη μικρότερη τιμή από Δ.Ε.Α.:

Βρείτε την ελάχιστη απόσταση στο ενημερωμένο Δ.Ε.Α. μητρώο: AC (απόσταση = 4).

Συγχώνευση ομάδων A και C: [B, D, E], [A, C], [F].

Ενημέρωση των μητρώων Δ.Ε.Α. και Δ.Μ.Α.

Κριτήριο διακοπής (συνθήκη Δ.Μ.Α.):

Ελέγξτε το Δ.Μ.Α. μητρώο: Υπάρχει σύνδεση μεταξύ των σημείων A και C.

Θεωρήστε το ως λύση: [B, D, E], [A, C], [F].

Σε αυτό το παράδειγμα, ο αλγόριθμος σταματά όταν υπάρχει σύνδεση μεταξύ των σημείων A και C, υποδεικνύοντας ότι έχει βρεθεί μια λύση. Οι συστάδες που προκύπτουν είναι [B, D, E], [A, C] και [F].

Πρέπει να λάβουμε υπόψη ότι οι αποστάσεις και τα συμπλέγματα που προκύπτουν ενδέχεται να διαφέρουν ανάλογα με το συγκεκριμένο σύνολο δεδομένων και τη μέτρηση απόστασης που χρησιμοποιείται.

2.2 Εξωτερικά κριτήρια

Τα εξωτερικά κριτήρια χρησιμοποιούνται για την αξιολόγηση του αποτελέσματος της διαδικασίας της ομαδοποίησης. Αξιοποιούν πληροφορία η οποία προέρχεται εκτός της διαδικασίας αυτής, και αναφέρεται στην πραγματική κατανομή των παρατηρήσεων σε ομάδες. Ουσιαστικά, τα εξωτερικά κριτήρια συγκρίνουν το αποτέλεσμα των μεθόδων ομαδοποίησης με την πραγματικότητα. Συνεπώς, έχουν εφαρμογή μόνο σε περιπτώσεις όπου κανείς γνωρίζει την κατανομή των παρατηρήσεων σε ομάδες εκ των προτέρων (ground truth).

Μετά την εύρεση αποτελέσματος από ένα αλγόριθμο ομαδοποίησης και έχοντας γνώση του ground truth, γίνεται χρήση ενός πίνακα σύγχυσης (confusion matrix) ο οποίος μας δείχνει κατά πόσο οι παρατηρήσεις έχουν ομαδοποιηθεί σωστά, όπως φαίνεται στο ακόλουθο σχήμα.

Clustering 1 / Clustering 2	Same	Different
Same	a	b
Different	c	d

Σχήμα 8 Απεικόνιση Confusion Matrix

Η τιμή *a* δείχνει πόσα ζεύγη παρατηρήσεων ομαδοποιήθηκαν σωστά **(true positive)** και η τιμή *b* δείχνει πόσα ζεύγη παρατηρήσεων κατατάχθηκαν στην ίδια ομάδα χωρίς αυτό να είναι η σωστή επιλογή **(false negative)**. Αντιθέτως, η τιμή *c* δείχνει τα ζεύγη παρατηρήσεων που τοποθετήθηκαν σε διαφορετικές ομάδες, όμως στην πραγματικότητα ανήκουν στην ίδια **(false positive)**. Τέλος, η τιμή *d* δείχνει τα παραδείγματα που τοποθετήθηκαν σε διαφορετικές ομάδες και όντως ανήκουν σε αυτές **(true negative)**.

Ο παραπάνω πίνακας σύγχυσης χρησιμοποιείται από πολλά εξωτερικά κριτήρια τα οποία χρησιμοποιώντας διάφορους μαθηματικούς τύπους δείχνουν την απόκλιση της ομαδοποίησης από την πραγματικότητα.

Παρακάτω δίνονται κάποιοι από τους γνωστότερους:

1	Rand	$(a+d)\ /\ (a+b+c+d)$
2	Corrected Rand	$(a+d-n_c)\ /\ (a+b+c+d\ -n_c)$

3	Jacard Coefficient	$a\ /\ [(a+b)(a+c)]^{1/2}$
4	Folkes and Mallow index	$a\ /\ (a+b+c)$

Πίνακας 2. Τύποι εξωτερικών κριτηρίων

Όπου

$$n_c = \sum\sum \frac{{N_i}^2 {N_j}^2}{N^2} + \frac{N(N-1)}{2} - \sum \frac{{N_i}^2}{2} - \sum \frac{{N_j}^2}{2} \quad (2.9)$$

και N_i, N_j, N είναι τα περιθώρια και τα ολικά αθροίσματα.

Συγκεκριμένα

N_i: Το άθροισμα των μετρήσεων στη σειρά i του πίνακα σύγχυσης, που αντιπροσωπεύει τον συνολικό αριθμό των παρατηρήσεων που ανήκουν στη συστάδα i στη βασική αλήθεια.

N_j: Το άθροισμα των μετρήσεων στη στήλη j του πίνακα σύγχυσης, που αντιπροσωπεύει τον συνολικό αριθμό των παρατηρήσεων που έχουν εκχωρηθεί στη συστάδα j στο αποτέλεσμα της ομαδοποίησης.

N: Ο συνολικός αριθμός των παρατηρήσεων στο σύνολο δεδομένων.

Ο τύπος για το n_c υπολογίζει τον αναμενόμενο αριθμό συμβατών ζευγών με την υπόθεση της ανεξαρτησίας μεταξύ του αποτελέσματος ομαδοποίησης και των ετικετών αληθείας βάσης. Περιλαμβάνει αθροίσματα και υπολογισμούς με βάση τα οριακά και τα σύνολα που προέρχονται από τον πίνακα σύγχυσης.

Χρησιμοποιώντας τον πίνακα σύγχυσης και τα στοιχεία του (a, b, c, d), μαζί με τις οριακές μετρήσεις (N_i, N_j) και το συνολικό πλήθος (N), αυτές οι μετρήσεις αξιολόγησης παρέχουν μέτρα ομοιότητας ή συμφωνίας μεταξύ του αποτελέσματος ομαδοποίησης και της βασικής αλήθειας, τονίζοντας την ποιότητα της ομαδοποίησης σε σύγκριση με τις γνωστές ετικέτες.

ΚΕΦΑΛΑΙΟ 3

ΣΥΓΚΡΙΣΗ ΚΡΙΤΗΡΙΩΝ – ΕΥΡΕΣΗ ΒΕΛΤΙΣΤΟΥ ΑΡΙΘΜΟΥ ΟΜΑΔΩΝ

3.1.Σκοπός πρακτικού μέρους και μεθοδολογία έρευνας

Η ανάλυση συστάδων είναι μια δημοφιλής τεχνική που χρησιμοποιείται σε πολλά πεδία για την ομαδοποίηση παρόμοιων παρατηρήσεων με βάση τα χαρακτηριστικά τους. Χρησιμοποιείται συνήθως σε πολυδιάστατα σύνολα δεδομένων για την εύρεση μοτίβων που μπορεί να μην ανιχνευθούν εύκολα από ανθρώπινους παρατηρητές. Υπάρχουν αρκετοί διαθέσιμοι αλγόριθμοι ομαδοποίησης, ο καθένας με τα δικά του πλεονεκτήματα και αδυναμίες. Στη βιβλιογραφία υπάρχουν αρκετά κριτήρια τα οποία είναι διαθέσιμα ώστε να αξιολογήσουνε την ποιότητα της ομαδοποίησης. Σκοπός της παρούσας διπλωματικής εργασίας είναι να συγκρίνει και να αξιολογήσει διαφορετικά κριτήρια αξιολόγησης που υπάρχουν στη βιβλιογραφία. Η μελέτη χρησιμοποιώντας προσομοιωμένα δεδομένα με εκ των προτέρων γνωστό αριθμό ομάδων θα προχωρήσει στην αξιολόγηση των κριτηρίων αυτών ως προς την ακρίβειά τους κάτω από διαφορετικές συνθήκες διαχωρισμού των ομάδων.

Ο σχεδιασμός της έρευνας για αυτή τη μελέτη είναι συγκριτικός, καθώς στοχεύει στη σύγκριση και αξιολόγηση διαφορετικών κριτηρίων σε πολυδιάστατα σύνολα δεδομένων. Τα προσομοιωμένα δεδομένα θα χρησιμοποιηθούν για λόγους επίδειξης. Επικέντρωση θα γίνει στην αξιολόγηση του βέλτιστου (σωστού) αριθμού συστάδων στα σύνολα δεδομένων χρησιμοποιώντας διάφορα κριτήρια αξιολόγησης, τόσο εσωτερικά όσο και εξωτερικά.

3.2.Συλλογή και επεξεργασία δεδομένων

Τα δεδομένα που χρησιμοποιούνται σε αυτή τη μελέτη θα αποτελούνται από πολυδιάστατα σύνολα δεδομένων, τα οποία είναι προσομοιωμένα σύνολα δεδομένων. Πιο συγκεκριμένα, τα δεδομένα θα δημιουργηθούν χρησιμοποιώντας συγκεκριμένες κατανομές και ο μέσος όρος και η τυπική απόκλιση θα τροποποιηθούν για να δημιουργηθούν σύνολα δεδομένων με διαφορετικά επίπεδα διαχωρισμού μεταξύ συστάδων.

Πριν από την ομαδοποίηση των δεδομένων, θα χρησιμοποιηθούν τεχνικές παραγωγής των δεδομένων μέσω της γλώσσας R. Τα βήματα παραγωγής θα περιλαμβάνουν τη δημιουργία πολυδιάστατων κανονικών κατανομών με προκαθορισμένες παραμέτρους $\boldsymbol{\mu}$, Σ.

Θα χρησιμοποιηθεί η εντολή mvnorm για την παραγωγή και δημιουργία των απαραίτητων για την παρουσίαση δεδομένων μέσω τυχαίου τρόπου.

3.3.Τεχνικές ομαδοποίησης και κριτήρια αξιολόγησης και μελέτη προσομοίωσης

Η μελέτη θα χρησιμοποιήσει την μέθοδο K-means για τον διαχωρισμό των ήδη προκαθορισμένων ομάδων. Η τεχνική αυτή επιλέχθηκε επειδή είναι δημοφιλής και χρησιμοποιείται ευρέως στην ανάλυση ομαδοποίησης .

Για την αξιολόγηση της απόδοσης της ομαδοποίησης, χρησιμοποιήθηκαν διαφορετικά κριτήρια αξιολόγησης κάποια εκ των οποίων αναλύθηκαν περαιτέρω ώστε να εξαχθούν χρήσιμα συμπεράσματα.

Διεξάχθηκε μελέτη προσομοίωσης για την αξιολόγηση και σύγκριση της απόδοσης των κριτηρίων αξιολόγησης σε πολυδιάστατα σύνολα δεδομένων. Τα σύνολα δεδομένων δημιουργήθηκαν μέσω πολυδιάστατων κανονικών κατανομών που θεωρήθηκαν κατάλληλες για να αναδείξουν το αντικείμενο της συγκεκριμένης εργασίας και μέσω των παραμέτρων τους να οδηγηθούμε στα αντίστοιχα πρίσματα.

3.4. Στατιστική ανάλυση

Θα διεξαχθεί στατιστική ανάλυση για να συγκριθούν οι τεχνικές ομαδοποίησης χρησιμοποιώντας τα κριτήρια αξιολόγησης. Τα αποτελέσματα αναλύθηκαν χρησιμοποιώντας περιγραφικές στατιστικές, όπως η μέση και τυπική απόκλιση, και στατιστικές συμπερασμάτων

Δεν χρησιμοποιήθηκαν προσωπικά δεδομένα στη μελέτη και όλα τα δεδομένα θα είναι ανώνυμα για την προστασία του απορρήτου των ατόμων.

Οι περιορισμοί της μελέτης περιλαμβάνουν τη χρήση συγκεκριμένων τεχνικών ομαδοποίησης και κριτηρίων αξιολόγησης. Άλλες τεχνικές ομαδοποίησης και κριτήρια αξιολόγησης θα μπορούσαν να παράγουν διαφορετικά αποτελέσματα. Τα αποτελέσματα της μελέτης ενδέχεται επίσης να περιορίζονται στα σύνολα δεδομένων που χρησιμοποιούνται στη μελέτη και τα αποτελέσματα ενδέχεται να μην είναι γενικά σε άλλα σύνολα δεδομένων.

3.5.Περιγραφή και δημιουργία των προσομοιωμένων δεδομένων στην R

Όπως αναφέρθηκε παραπάνω διάφορες πολυδιάστατες κανονικές κατανομές χρησιμοποιήθηκαν για την παραγωγή των δεδομένων. Παρακάτω ορισμένα στοιχεία για τη συγκεκριμένη κατανομή και τις παραμέτρους της.

Η πολυδιάστατη κανονική κατανομή αποτελεί τη βάση για τις περισσότερες απλές στατιστικές εφαρμογές. Η πολυδιάστατη που ορίζεται για το τυχαίο διάνυσμα x διαστάσεων p x 1 έχει από κοινού συνάρτηση πυκνότητας πιθανότητας

$$f(\boldsymbol{x}) \quad = \frac{1}{(2\pi)^{\frac{p}{2}}\,|\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})'\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right) \tag{3.1}$$

όπου $\boldsymbol{x}'=(x_1, x_2, \ldots\ldots, x_p)$, $\boldsymbol{\mu}$ ένα διάνυσμα p x 1 και Σ ένας πίνακας p x p.

Η ερμηνεία των παραμέτρων δεν είναι σύνθετη. Το διάνυσμα $\boldsymbol{\mu}$ περιλαμβάνει τις αναμενόμενες τιμές κάθε μιας μεταβλητής, δηλαδή πρόκειται για το διάνυσμα των μέσων, ενώ ο πίνακας Σ είναι ο πίνακας με τις συνδιακυμάνσεις των μεταβλητών του τυχαίου διανύσματος, δηλαδή:

$$\boldsymbol{\mu} = E(\boldsymbol{X}) \tag{3.2}$$

$$\Sigma = Cov(\boldsymbol{X}) \tag{3.3}$$

Μια σύντομη απεικόνιση των παραπάνω:

$$E(\boldsymbol{X}) = \begin{pmatrix} E(X_1) \\ \ldots \\ E(X_P) \end{pmatrix} \tag{3.4}$$

$$Cov(\boldsymbol{X}) = \begin{pmatrix} Var(X_1) & Cov(X_1, X_2) & \ldots & Cov(X_1, X_p) \\ Cov(X_1, X_2) & Var(X_2) & \ldots & \ldots \\ \ldots & \ldots & \ldots & \ldots \\ Cov(X_2, X_p) & Cov(X_2, X_p) & & Var(X_p) \end{pmatrix} \tag{3.5}$$

3.6.Δεδομένα και περιγραφικά στατιστικά

Όπως αναφέρθηκε παραπάνω τα δεδομένα μας θα αποτελούνται από διαφορετικές πολυδιάστατες κανονικές κατανομές. Θα γίνουν διάφορες δοκιμές μεταξύ των παραμέτρων των κατανομών.

Αρχικά επιλέχθηκε να δημιουργηθούν 3 πολυδιάστατες κανονικές κατανομές (3 διαστάσεων) που καθεμία από αυτές θα αποτελεί ένα cluster.

Πιο συγκεκριμένα, μέσω της εντολής mvnorm στην R προχωράμε στη δημιουργία των δεδομένων με τα ακόλουθα διανύσματα μέσων για την καθεμία κατανομή.

```
# Number of data points in each cluster
n <- 500
mu1 <- c(0, 1, 0)
mu2 <- c(12, 10, 9)
mu3 <- c(4, 6, 5)
```

Επιλέξαμε αρχικά να καθορίσουμε ανεξάρτητους μεταξύ τους πληθυσμούς και έτσι ως πίνακα διακύμανσης συνδιακύμανσης ορίστηκε ο διαγώνιος πίνακας με την παρακάτω εντολή.

```
sigma <- diag(3)
```

Παρακάτω η μορφή του διαγώνιου πίνακα:

```
> sigma
     [,1] [,2] [,3]
[1,]    1    0    0
[2,]    0    1    0
[3,]    0    0    1
```

Σχήμα 3.1 Απεικόνιση του αρχικού πίνακα Σ των προσομοιωμένων αρχικών δεδομένων

Η εντολή mvnorm παίρνει ως βασικά ορίσματα τα παρακάτω:

n-> ο αριθμός των δειγμάτων μας

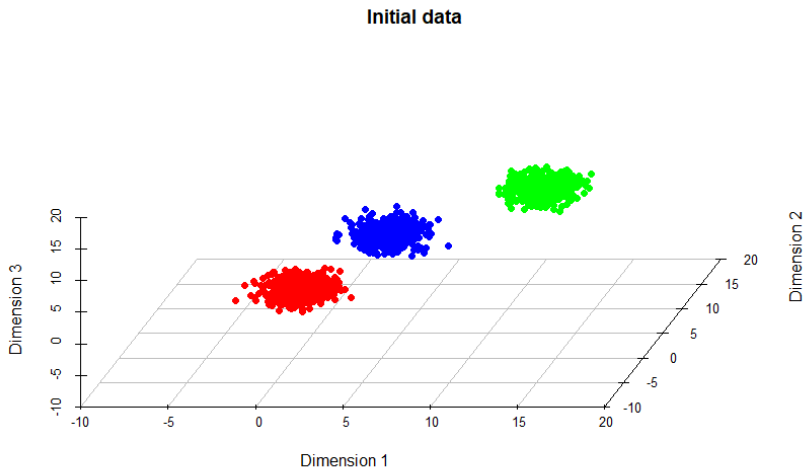
mu->ένα διάνυσμα που δίνει τους μέσους όρους των μεταβλητών

sigma->ένας θετικά ορισμένος και συμμετρικός πίνακας που καθορίζει τον πίνακα διακύμανσης συνδιακύμανσης των μεταβλητών

Στη συνέχεια δημιουργούμε τα δεδομένα για τα 3 διαφορετικά clusters και τα τοποθετούμε σε ένα dataframe ώστε να προχωρήσουμε περαιτέρω στην ανάλυσή μας.Παρακάτω οι εντολές στην R.

```
# Generate data for each cluster
cluster1 <- mvrnorm(n, mu1, sigma)
cluster2 <- mvrnorm(n, mu2, sigma)
cluster3 <- mvrnorm(n, mu3, sigma)

# Combine the clusters
data <- rbind(cluster1, cluster2, cluster3)
```

Η αρχική μορφή των δεδομένων στον τρισδιάστατο χώρο είναι ως εξής:



Σχήμα 3.2 Απεικόνιση των αρχικών δεδομένων χωρισμένα σε 3 διακριτά clusters

Παρακάτω δίνεται ένας σύντομος πίνακας σχετικά με τα περιγραφικά μέτρα για κάθε cluster. Από τις τιμές min,max οι οποίες δεν είναι αλληλοκαλυπτόμενες ανάμεσα στα clusters επιβεβαιώνουμε αυτό που και οπτικά είδαμε ότι οι αρχικές 3 ομάδες είναι τελείως διακριτές.

```
> summary(cluster1)
       V1                  V2                  V3
 Min.   :-3.30665   Min.   :-1.7800   Min.   :-3.75742
 1st Qu.:-0.59328   1st Qu.: 0.3462   1st Qu.:-0.68572
 Median : 0.07301   Median : 1.1005   Median : 0.04232
 Mean   : 0.05185   Mean   : 1.0676   Mean   : 0.03775
 3rd Qu.: 0.62066   3rd Qu.: 1.7530   3rd Qu.: 0.77726
 Max.   : 3.30857   Max.   : 4.3110   Max.   : 2.98568
> summary(cluster2)
       V1               V2               V3
 Min.   : 9.426   Min.   : 6.996   Min.   : 6.044
 1st Qu.:11.354   1st Qu.: 9.388   1st Qu.: 8.280
 Median :12.019   Median :10.032   Median : 9.062
 Mean   :12.023   Mean   :10.021   Mean   : 9.014
 3rd Qu.:12.710   3rd Qu.:10.686   3rd Qu.: 9.738
 Max.   :14.640   Max.   :12.729   Max.   :11.656
> summary(cluster3)
       V1              V2              V3
 Min.   :1.087   Min.   :3.362   Min.   :1.835
 1st Qu.:3.351   1st Qu.:5.351   1st Qu.:4.219
 Median :4.077   Median :6.044   Median :4.937
 Mean   :4.034   Mean   :6.035   Mean   :4.955
 3rd Qu.:4.709   3rd Qu.:6.714   3rd Qu.:5.626
 Max.   :7.698   Max.   :9.019   Max.   :8.103
```

Σχήμα 3.3. Απεικόνιση των περιγραφικών στατιστικών για κάθε αρχικό cluster

Στη συνέχεια θα χρησιμοποιήσουμε την τεχνική clustering K-means και ακολούθως θα προχωρήσουμε στην αξιολόγηση των κριτηρίων έχοντας ως πρότερη γνώση ότι αναμένουμε 3 ομάδες.

Για την διεκπεραίωση των παραπάνω θα χρησιμοποιήσουμε τη βιβλιοθήκη nbclust. Η συγκεκριμένη βιβλιοθήκη χρησιμοποιείται για να βοηθήσει τον ερευνητή να αξιολογήσει τον βέλτιστο αριθμό clusters για τα εκάστοτε δεδομένα.

Πιο συγκεκριμένα η μορφή της εντολής μαζί με τα βασικά ορίσματα που δέχεται είναι η ακόλουθη:

NbClust(data = NULL, diss = NULL, distance = "euclidean", min.nc = 2, max.nc = 15, method = NULL, index = "all", alphaBeale = 0.1)

Τα βασικά ορίσματα που θα χρησιμοποιήσουμε είναι τα **min.nc**, **max.nc**, index, distance.Ως min.nc ορίζεται ο ελάχιστος αριθμός συστάδων προς εξέταση με ελάχιστη τιμή το 2 και αντίστοιχα το όρισμα **max.nc** αφορά το μέγιστο εξεταζόμενο αριθμό συστάδων και προφανώς θα πρέπει να ισχύει max.nc>min.nc.

Το όρισμα distance αναφέρεται στον τύπο απόστασης που θα χρησιμοποιηθεί για τον υπολογισμό του πίνακα ανομοιότητας(dissimilarity matrix). Υπάρχουν διάφοροι τύποι αποστάσεων όπως η ευκλείδια, η Canberra distance, Manhattan distance και άλλες. Για τις

ανάγκες της έρευνας χρησιμοποιήθηκε η πιο ευρέως χρησιμοποιούμενη απόσταση, η ευκλείδια.

Το όρισμα diss αναφέρεται στον πίνακα ανομοιότητας τον οποίο θα αφήσουμε να υπολογιστεί με βάση το όρισμα distance.

Το όρισμα method αφορά τη μέθοδο συνάθροισης των δεδομένων. Μερικές από τις μεθόδους που περιέχονται είναι το Complete Linkage Method,Average of all pairs,Centroid,Ward αλλά και η μέθοδος *K*-means που έχει επιλεχθεί για την έρευνά μας. Ουσιαστικά παρέχονται διάφοροι τύποι αποστάσεων που χρησιμοποιούνται από τους ιεραρχικούς αλγορίθμους και επιπλέον από τους μη ιεραρχικούς είναι διαθέσιμη η μέθοδος K-means.

Το όρισμα index αφορά τα κριτήρια που περιέχονται στο πακέτο και είναι συνολικά 30.Παρέχεται η δυνατότητα επιλογής των κριτήριων που θέλουμε να απεικονιστούν. Επίσης υπάρχουν 2 επιλογές ,η επιλογή all και η επιλογή all long.Με τη δεύτερη προκύπτει η παροχή όλων των κριτηρίων ενώ με την πρώτη που και θα χρησιμοποιηθεί παρουσιάζονται 25 εκ των 30 και παραλείπονται ορισμένα τα οποία απαιτούν μεγάλη υπολογιστική ισχύ και χρόνο.

Τέλος το όρισμα alphaBeale αφορά το επίπεδο σημαντικότητας για το κριτήριο Beale's Index και βιβλιογραφικά συνήθως χρησιμοποιείται το 0.1.

Μία γρήγορη αναφορά στο τι περιμένουμε σαν output και ουσιαστικά θα αξιολογήσουμε από την παραπάνω βιβλιοθήκη. Παράγονται 4 πίνακες ως κάτωθι:

All.index: Αφορά της τιμές των δεικτών για κάθε partition του συνόλου δεδομένων που λήφθηκαν για κάθε αριθμό clusters που εξετάζουμε. Στην περίπτωσή μας και θέτοντας τις κατάλληλες τιμές στα ορίσματα θα εξετάσουμε τα κριτήρια για αριθμό ομάδων από 2 εώς 8.

All Critical values: Περιέχει τις κρίσιμες τιμές ορισμένων δεικτών για κάθε partition των δεδομένων που λαμβάνονται με κάθε αριθμό συστάδων ανάμεσα στο min.nc και max.nc.

Best Partition: Ο διαμερισμός των δεδομένων που αντιστοιχεί στον καλύτερο αριθμό συστάδων.

Best.nc: Είναι ο πίνακας που έχει όλη την ουσία για την ερευνάς μας και σε αυτόν θα επικεντρωθούμε. Περιέχει τον καλύτερο αριθμό συστάδων που προτείνεται από κάθε δείκτη με την αντίστοιχη τιμή του δείκτη.

Μια υποσημείωση σχετικά με τον όρο τιμές των κριτηρίων. Κάθε κριτήριο προτείνει τον βέλτιστο αριθμό ομάδων με βάση μία τιμή η οποία μπορεί να είναι η μεγαλύτερη μεταξύ του εξεταζόμενου αριθμού ομάδων ή η μικρότερη. Τέλος υπάρχουν κριτήρια τα οποία προτείνουν τον βέλτιστο αριθμό μέσω γραφήματος. Παρακάτω συγκεντρωτικά όλα τα κριτήρια και το πώς αποφασίζουν σχετικά με το βέλτιστο αριθμό.

Index in NbClust	Optimal number of clusters
1. "kl" or "all" or "alllong" (Krzanowski and Lai 1988)	Maximum value of the index
2. "ch" or "all" or "alllong" (Calinski and Harabasz 1974)	Maximum value of the index
3. "hartigan" or "all" or "alllong" (Hartigan 1975)	Maximum difference between hierarchy levels of the index
4. "ccc" or "all" or "alllong" (Sarle 1983)	Maximum value of the index
5. "scott" or "all" or "alllong" (Scott and Symons 1971)	Maximum difference between hierarchy levels of the index
6. "marriot" or "all" or "alllong" (Marriot 1971)	Max. value of second differences between levels of the index
7. "trcovw" or "all" or "alllong" (Milligan and Cooper 1985)	Maximum difference between hierarchy levels of the index
8. "tracew" or "all" or "alllong" (Milligan and Cooper 1985)	Maximum value of absolute second differences between levels of the index
9. "friedman" or "all" or "alllong" (Friedman and Rubin 1967)	Maximum difference between hierarchy levels of the index
10. "rubin" or "all" or "alllong" (Friedman and Rubin 1967)	Minimum value of second differences between levels of the index
11. "cindex" or "all" or "alllong" (Hubert and Levin 1976)	Minimum value of the index
12. "db" or "all" or "alllong" (Davies and Bouldin 1979)	Minimum value of the index
13. "silhouette" or "all" or "alllong" (Rousseeuw 1987)	Maximum value of the index
14. "duda" or "all" or "alllong" (Duda and Hart 1973)	Smallest *n_{c}* such that index > criticalValue
15. "pseudot2" or "all" or "alllong"	Smallest *n_{c}* such that index < criticalValue

16. "beale" or "all" or "alllong" (Beale 1969)	*n_{c}* such that critical value of the index >= alpha
17. "ratkowsky" or "all" or "alllong" (Ratkowsky and Lance 1978)	Maximum value of the index
18. "ball" or "all" or "alllong" (Ball and Hall 1965)	Maximum difference between hierarchy levels of the index
19. "ptbiserial" or "all" or "alllong" (Milligan 1980, 1981)	Maximum value of the index
20. "gap" or "alllong" (Tibshirani et al. 2001)	Smallest *n_{c}* such that criticalValue >= 0
21. "frey" or "all" or "alllong" (Frey and Van Groenewoud 1972)	the cluster level before that index value < 1.00
22. "mcclain" or "all" or "alllong" (McClain and Rao 1975)	Minimum value of the index
23. "gamma" or "alllong" (Baker and Hubert 1975)	Maximum value of the index
24. "gplus" or "alllong" (Rohlf 1974) (Milligan 1981)	Minimum value of the index
25. "tau" or "alllong" (Rohlf 1974) (Milligan 1981)	Maximum value of the index
26. "dunn" or "all" or "alllong" (Dunn 1974)	Maximum value of the index
27. "hubert" or "all" or "alllong" (Hubert and Arabie 1985)	Graphical method
28. "sdindex" or "all" or "alllong" (Halkidi et al. 2000)	Minimum value of the index
29. "dindex" or "all" or "alllong" (Lebart et al. 2000)	Graphical method
30. "sdbw" or "all" or "alllong" (Halkidi and Vazirgiannis 2001)	Minimum value of the index

Σχήμα 3.4 Απεικόνιση των κριτηρίων και του τρόπου απόφασης του βέλτιστου αριθμού ομάδων

3.7. Εφαρμογή *K*-means και κριτήρια αξιολόγησης

Στο συγκεκριμένη ενότητα θα προχωρήσουμε σε διάφορες δοκιμές πέρνα των αρχικών(Initial) δεδομένων με διαφορικές παραμέτρους $\boldsymbol{\mu}$, Σ ώστε να έχουμε διαφορετικό επίπεδο διαχωρισμού μεταξύ των ομάδων.

- **Initial Data**

Η μορφή των αρχικών δεδομένων καθώς και οι παράμετροι που αρχικά χρησιμοποιήθηκαν παρουσιάστηκαν παραπάνω. Παρακάτω προχωρούμε στην ομαδοποίηση των δεδομένων μας εξετάζοντας για τελικά πιθανά clusters από 2 εώς 8 γνωρίζοντας εκ των προτέρων ότι για βασική αλήθεια θα χρησιμοποιηθούν οι 3 αρχικά ορισμένες ομάδες.

Παρακάτω οι εντολές που χρησιμοποιήθηκαν:

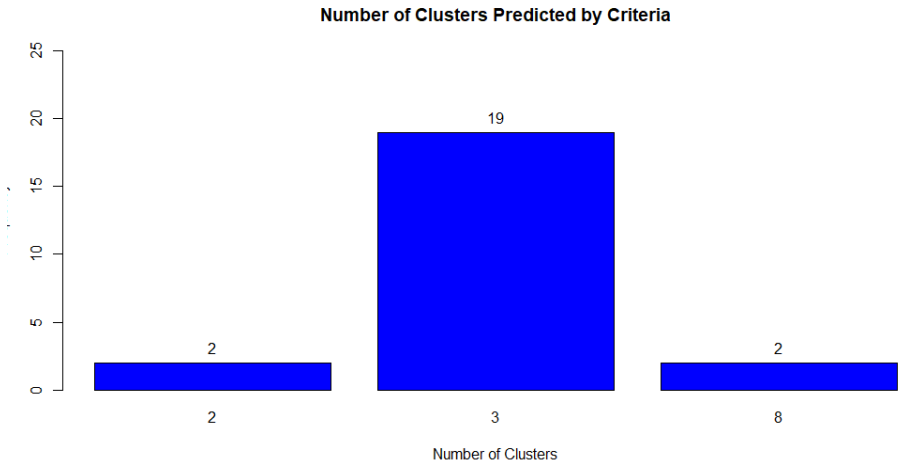
nbclust_result <- NbClust(data, distance = "euclidean", min.nc = 2, max.nc = 8, method = "kmeans")
print(nbclust_result)

Οι πίνακες που προαναφέραμε βρίσκονται παρακάτω (all Index ,Critical Values και Best.nc) μαζί με τα διαθέσιμα κριτήρια.

```
$All.index
       KL        CH  Hartigan     CCC     Scott      Marriot  TrCovW    TraceW Friedman   Rubin Cindex     DB Silhouette
2  2.0982  4422.382 5270.4874 39.3392  5149.832 168427898884 3943404 20769.292  85.3837  9.9617 0.3257 0.4537     0.6644
3 19.7262 12617.750  116.2692 58.4384  8846.262  32238155412 2397385  4596.655 133.4238 45.0106 0.3173 0.4098     0.7350
4  1.7665  9097.840  137.8401 43.8312  9226.496  44479391097 2023634  4265.371 147.6400 48.5065 0.3097 1.1640     0.5472
5  0.9104  7481.536   90.7639 36.5909  9743.363  49241411068 1630819  3905.520 170.1812 52.9758 0.2969 1.5592     0.3818
6  1.1764  6362.496   80.4783 44.2359  9947.662  61878845237 1651139  3681.981 174.7626 56.1921 0.2937 1.4409     0.3728
7  1.9098  5597.355  182.0285 40.7545 10150.901  73551601750 1344966  3493.779 182.8560 59.2190 0.2924 1.3241     0.3781
8 27.5377  5405.060  123.4238 40.7571 10654.763  68658290800 1092138  3114.103 200.9413 66.4391 0.2640 1.5449     0.2297
    Duda  Pseudot2   Beale Ratkowsky       Ball Ptbiserial    Frey McClain   Dunn Hubert SDindex Dindex   SDbw
2 0.1603 5226.4895  8.9065    0.5977 10384.6458     0.7844  0.7574  0.2732 0.0536      0  0.4419 3.3872 0.2519
3 0.7816  139.1423  0.4747    0.5595  1532.2182     0.7916 13.3543  0.3830 0.3886      0  0.3660 1.6159 0.0538
4 0.7654  152.6268  0.5207    0.4856  1066.3428     0.7000  6.7978  0.5175 0.0166      0  1.4809 1.5491 0.2564
5 1.0689  -22.2521 -0.1094    0.4355   781.1040     0.6119  8.6034  0.7067 0.0166      0  1.3458 1.4761 0.3067
6 1.0838  -18.2538 -0.1309    0.3983   613.6635     0.5807  9.6575  0.7960 0.0166      0  1.4255 1.4288 0.4577
7 0.7544  162.1581  0.5532    0.3691   499.1113     0.5648  2.4760  0.8476 0.0166      0  1.4960 1.3886 0.6273
8 1.0639  -19.9268 -0.1018    0.3462   389.2629     0.4756  2.0091  1.2065 0.0166      0  1.5359 1.3112 0.4694

$All.CriticalValues
  CritValue_Duda CritValue_PseudoT2 Fvalue_Beale
2         0.7172           393.5738       0.0000
3         0.6880           225.8668       0.6999
4         0.6880           225.8668       0.6681
5         0.6507           185.2162       1.0000
6         0.6191           145.2206       1.0000
7         0.6880           225.8668       0.6460
8         0.6443           183.2799       1.0000

$Best.nc
                     KL       CH Hartigan     CCC    Scott     Marriot  TrCovW   TraceW Friedman   Rubin Cindex     DB
Number_clusters  8.0000     3.00    3.000  3.0000    3.000 3.00000e+00       3     3.00   3.0000   3.000  8.000 3.0000
Value_Index     27.5377 12617.75 5154.218 58.4384 3696.431 1.48431e+11 1546019 15841.35  48.0401 -31.553  0.264 0.4098
                Silhouette   Duda PseudoT2  Beale Ratkowsky     Ball PtBiserial Frey McClain   Dunn Hubert SDindex Dindex
Number_clusters      3.000 3.0000   3.0000 3.0000    2.0000    3.000     3.0000    1  2.0000 3.0000      0   3.000      0
Value_Index          0.735 0.7816 139.1423 0.4747    0.5977 8852.428     0.7916   NA  0.2732 0.3886      0   0.366      0
                  SDbw
Number_clusters 3.0000
Value_Index     0.0538
```

Σχήμα 3.5 Απεικόνιση των αποτελεσμάτων της Δοκιμής στα αρχικά δεδομένα

Επικεντρώνοντας το ενδιαφέρον μας στον πίνακα Best.nc και τη γραμμή Number_clusters μπορούμε να δούμε τις προτάσεις κάθε κριτηρίου σχετικά με το βέλτιστο αριθμό clusters.Πιο συγκεκριμένα από τα 23 κριτήρια που προκύπτουν προτάσεις μέσω τιμών και όχι μέσω γραφημάτων προκύπτει πως τα 19 προτείνουν 3 clusters όπως φαίνεται και στο παρακάτω γράφημα.



Σχήμα 3.6 Απεικόνιση barplot σχετικά με το πλήθος των κριτηρίων που πρότειναν κάθε αριθμό clusters

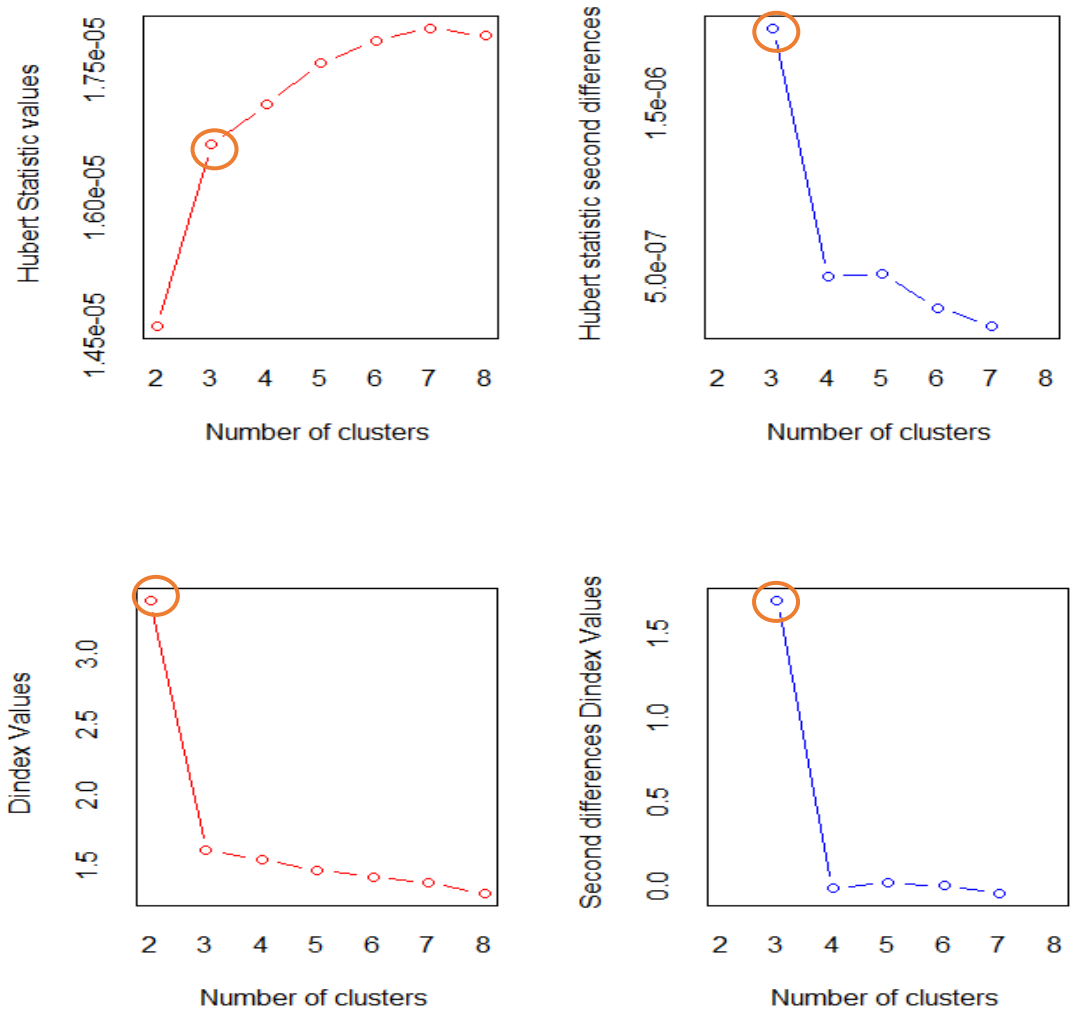
Λαμβάνοντας υπόψη τη δομή των αρχικών μας δεδομένων στο πολυδιάστατο χώρο όπου οι 3 ομάδες είναι εμφανώς διαχωρισμένες παρατηρούμε πως κάποια κριτήρια δεν προτείνουν το «σωστό» αριθμό clusters.Τα κριτήρια αυτά είναι το KL, C Index, McClain και Ratkowsky.

Ας δούμε και τα 2 κριτήρια που προτείνουν τον βέλτιστο αριθμό clusters μέσω γραφικής αναπαράστασης.

Το output της R μας δίνει σχετικές οδηγίες για το πώς θα αξιολογήσουμε τα γραφήματα.

```
*** : The Hubert index is a graphical method of determining the number of clusters.
                In the plot of Hubert index, we seek a significant knee that corresponds to a
                significant increase of the value of the measure i.e the significant peak in Hubert
                index second differences plot.

*** : The D index is a graphical method of determining the number of clusters.
                In the plot of D index, we seek a significant knee (the significant peak in Dindex
                second differences plot) that corresponds to a significant increase of the value of
                the measure.
```

Σχήμα 3.7 Οδηγίες αξιολόγησης των κριτηρίων μέσω γραφικών μεθόδων



Σχήμα 3.8 Γραφικές μεθόδοι κριτηρίων

Όπως παρατηρούμε και στις 2 περιπτώσεις τα κριτήρια προτείνουν 3 clusters. Συνολικά λοιπόν έχουμε 21 στα 25 διαθέσιμα κριτήρια να προτείνουν το «σωστό» αριθμό. Φαίνεται όμως ότι υπάρχει μία ομάδα κριτηρίων η οποία ακόμη και σε ευκόλως διακριτές και καλά διαχωρισμένες ομάδες δε λειτουργεί όπως θα έπρεπε καθώς προτείνει διαφορετικό αριθμό. Χαρακτηριστικό παράδειγμα είναι τα κριτήρια KL και C Index που προτείνουν 8 ομάδες. Το πρόβλημα με αυτά είναι εμφανές καθώς δε δίνουν λάθος αριθμό χάνοντας μια ομάδα από ότι τον πραγματικό αριθμό αλλά δίνουν τελείως λάθος νούμερο. Στις επόμενες δοκιμές θα ελέγξουμε και πάλι τα κριτήρια ώστε να αντιληφθούμε καλύτερα τη συμπεριφορά τους. Σε υποθετικό επίπεδο, μέσα σε κάποιο από τα 3 clusters υπάρχουν εστίες υψηλής συγκέντρωσης σημείων και το κριτήριο τα αντιλαμβάνεται ως μία ομάδα.

Κλείνοντας τη συγκεκριμένη πρώτη δοκιμή στα αρχικά δεδομένα αξίζει να αναφερθεί πως ο πιο συνηθισμένος τρόπος επιλογής του βέλτιστου αριθμού ομάδων είναι μέσω του κανόνα της πλειοψηφίας, έτσι και εδώ αν έπρεπε να επιλέξουμε τον βέλτιστο αριθμό ομάδων αυτός θα ήταν οι 3 .

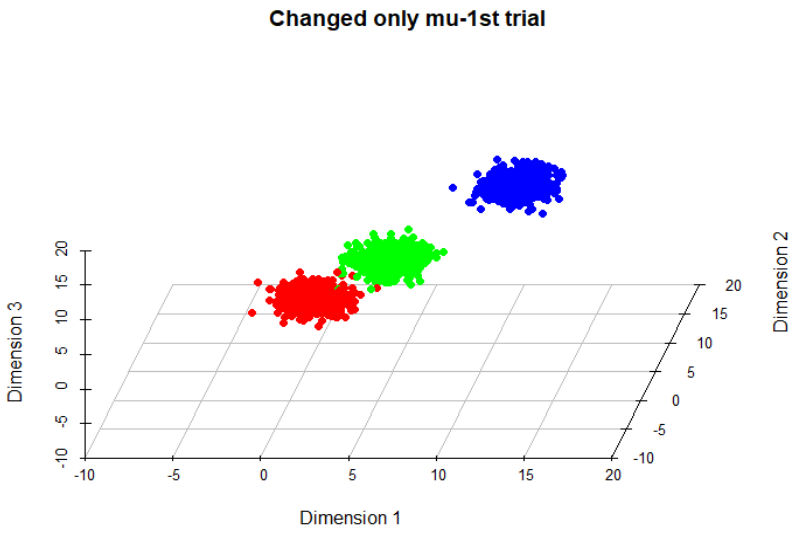
- **Δοκιμή 1 (Διαφορετικοί μέσοι ,ίδιο Σ)**

Στη δεύτερη δοκιμή μας, αποφασίσθηκε αρχικά να «πειράξουμε» μόνο τα διανύσματα των μέσων κρατώντας σταθερό τον ίδιο πίνακα διακύμανσης συνδιακύμανσης. Θα έχουμε έτσι μία πρώτη εικόνα αναφορικά με το πόσο ευαίσθητα μπορεί να είναι τα κριτήρια σε αλλαγές των μέσων, τέτοιες μάλιστα που όπως βλέπουμε παρακάτω τα clusters έρχονται πιο κοντά μεταξύ τους ιδιαίτερα 2 από αυτά.

Παραθέτουμε τα νέα διανύσματα μέσων:

```
# 1st trial bring closer the 2 clusters

mu1v2 <- c(1, 2, 3)
mu2v2 <- c(5, 4, 7)
mu3v2 <- c(12, 6, 16)
```

Τα υπόλοιπα κομμάτια του κώδικα ουσιαστικά παραμένουν στην ίδια λογική.



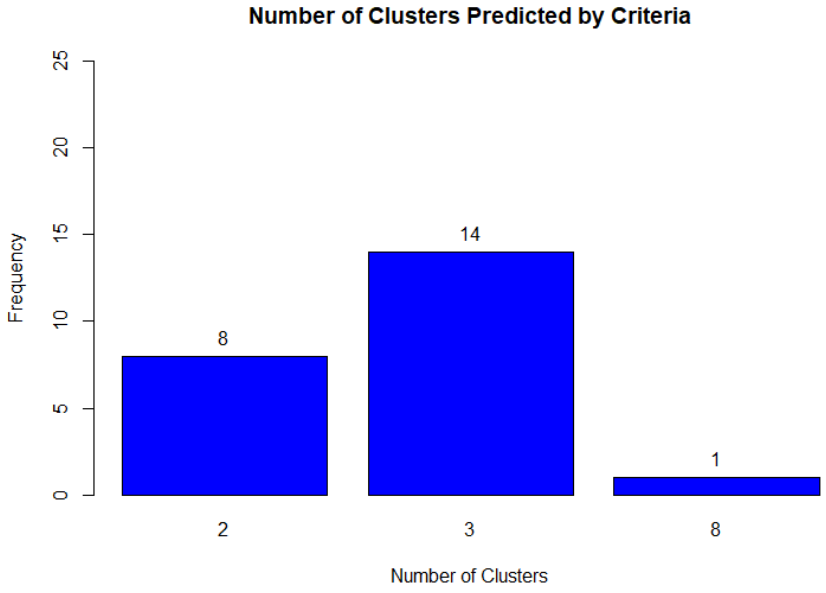
Σχήμα 3.9 Απεικόνιση των δεδομένων στο στάδιο της Δοκιμής 1

Στη συνέχεια χρησιμοποιώντας τις ίδιες εντολές και προσαρμόζοντάς τες αναλόγως παραθέτουμε και πάλι το output με τα πινακάκια των κριτηρίων και το προτεινόμενο βέλτιστο αριθμό συστάδων προς αξιολόγηση και σύγκριση.

```
$All.index
       KL        CH  Hartigan     CCC     Scott      Marriot  TrCovW    TraceW Friedman   Rubin Cindex     DB Silhouette   Duda  Pseudot2
2  5.2270  7705.095 2855.4996 65.5096  6037.178 108167005317 6565779 13776.918 108.6257 20.0134 0.3209 0.3608     0.7408 0.2605 2832.3892
3 10.2285 12615.628  123.5378 64.2344  8932.262  35323584142 2602099  4740.514 171.0035 58.1630 0.3095 0.4979     0.6849 0.7847  137.1580
4  1.6872  9139.546  136.9766 48.3754  9328.820  48208874905 2331252  4379.132 178.2069 62.9628 0.3300 1.1602     0.5414 0.7676  150.4430
5  0.9519  7511.505   94.1872 39.9542  9756.087  56655295500 1842711  4011.804 187.4441 68.7278 0.3170 1.5246     0.4136 0.9976    0.7931
6  0.9862  6402.345   68.7440 34.2628 10059.758  66631734111 1582408  3774.034 202.7479 73.0578 0.3132 1.4069     0.4155 1.3414  -64.6511
7  2.2561  5588.496  180.6519 41.3556 10220.901  81455231916 1516072  3608.017 213.5393 76.4194 0.3121 1.3205     0.4168 0.7498  166.1803
8  1.5854  5391.936   96.3103 41.3178 10706.729  76955784871 1232030  3218.572 251.4838 85.6661 0.2983 1.5502     0.2276 1.1862  -51.3251
    Beale Ratkowsky      Ball Ptbiserial    Frey McClain   Dunn Hubert SDindex Dindex   SDbw
2  4.8267    0.6062 6888.4592     0.8886  1.7177  0.2142 0.4791      0  0.3215 2.7827 0.1215
3  0.4661    0.5393 1580.1712     0.7517 10.3015  0.3889 0.0295      0  0.4621 1.6361 0.0480
4  0.5143    0.4729 1094.7830     0.6660  6.9682  0.5224 0.0175      0  1.7257 1.5634 0.2056
5  0.0041    0.4270  802.3607     0.5825  7.5505  0.7127 0.0175      0  1.7143 1.4914 0.2471
6 -0.4306    0.3906  629.0057     0.5529 13.6561  0.8024 0.0175      0  1.7783 1.4396 0.2792
7  0.5669    0.3620  515.4310     0.5371  2.5475  0.8573 0.0137      0  1.9180 1.4058 0.6398
8 -0.2660    0.3394  402.3215     0.4523  2.6498  1.2225 0.0144      0  1.5656 1.3335 0.5319

$All.CriticalValues
  CritValue_Duda CritValue_PseudoT2 Fvalue_Beale
2         0.7172           393.5053       0.0024
3         0.6881           226.6686       0.7060
4         0.6878           225.6236       0.6725
5         0.6502           175.4169       0.9996
6         0.6119           161.1134       1.0000
7         0.6880           225.8668       0.6368
8         0.6403           183.7032       1.0000

$Best.nc
                     KL       CH Hartigan     CCC    Scott     Marriot  TrCovW   TraceW Friedman    Rubin Cindex     DB Silhouette   Duda
Number_clusters  3.0000     3.00    3.000  2.0000    3.000           3       3    3.000   3.0000   3.0000 8.0000 2.0000     2.0000 3.0000
Value_Index     10.2285 12615.63 2731.962 65.5096 2895.084 85728711937 3963680 8675.023  62.3778 -33.3498 0.2983 0.3608     0.7408 0.7847
                PseudoT2  Beale Ratkowsky     Ball PtBiserial Frey McClain   Dunn Hubert SDindex Dindex  SDbw
Number_clusters    3.000 3.0000    2.0000    3.000     2.0000   NA  2.0000 2.0000      0  2.0000      0 3.000
Value_Index      137.158 0.4661    0.6062 5308.288     0.8886   NA  0.2142 0.4791      0  0.3215      0 0.048
```

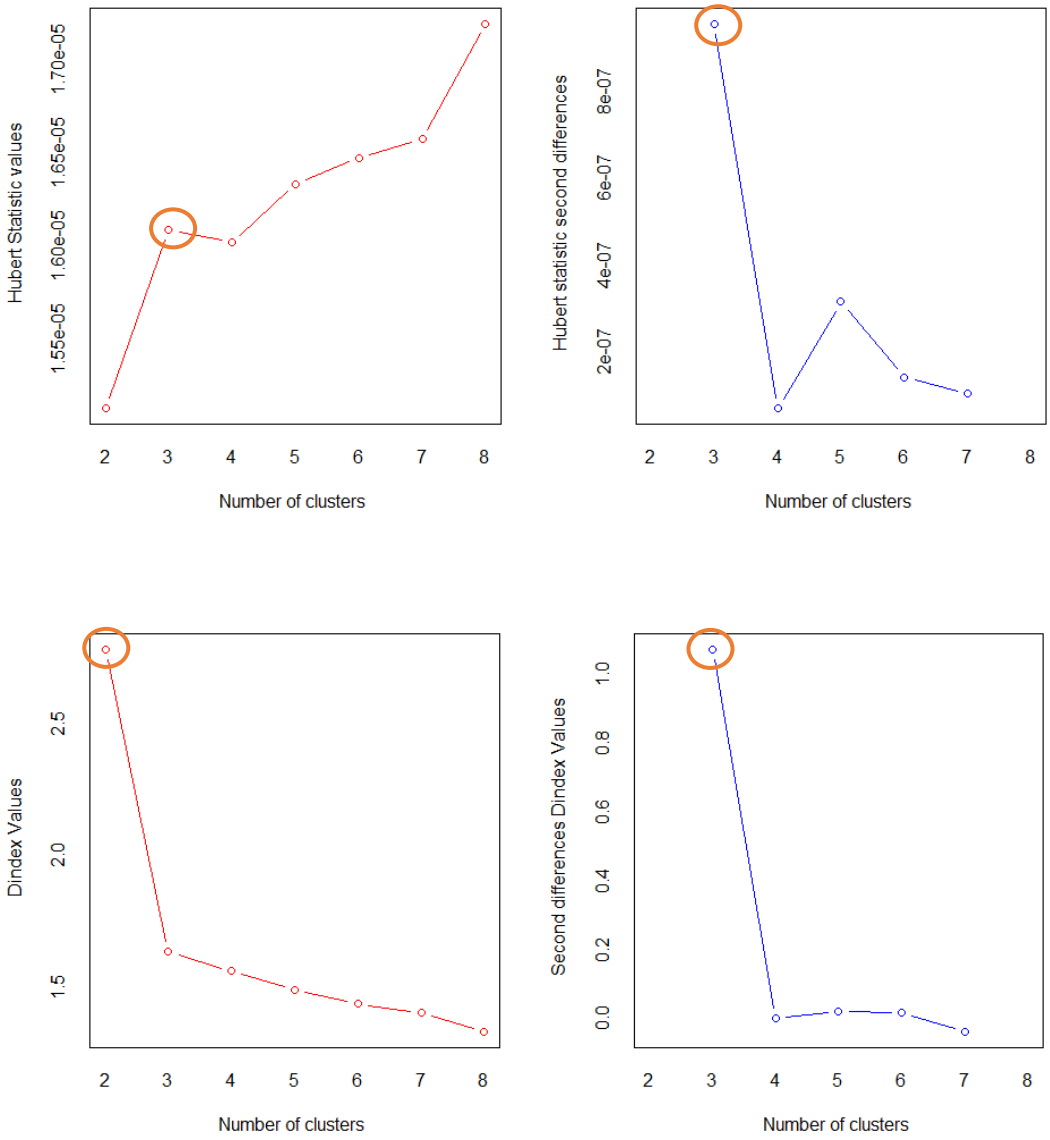
Σχήμα 3.10 Απεικόνιση των αποτελεσμάτων της Δοκιμής 1

Από το παρακάτω συγκεντρωτικό γράφημα μπορούμε να διαπιστώσουμε ότι 14 κριτήρια προτείνουν 3 clusters, 8 κριτήρια προτείνουν 2 clusters και 1 κριτήριο προτείνει 8 clusters.



Σχήμα 3.11 Απεικόνιση barplot σχετικά με το πλήθος των κριτηρίων που πρότειναν κάθε αριθμό clusters στη Δοκιμή 1

Παρακάτω και η παρουσίαση των κριτηρίων μέσω γραφικών μεθόδων.



Σχήμα 3.12 Γραφικές μεθόδοι κριτηρίων-Δοκιμή 1

Τα δύο κριτήρια όπως παρατηρούμε προτείνουν 3 clusters.

Σε σχέση με την πρώτη δοκιμή, τώρα που οι ομάδες είναι εγγύτερα παρατηρούμε ότι περισσότερα κριτήρια διαφοροποιούνται από τον αρχικά ορισμένο αριθμό των 3 ομάδων με τις περισσότερες εξ αυτών να προτείνουν 2. Πιο συγκεκριμένα, τα κριτήρια που προτείνουν δύο ομάδες είναι τα CCC, Ratkowsky, McClain, PtBiserial, Silhouettte, SD index, Dunn,DB ενώ 8 clusters προτείνει το Cindex, η υπόθεση που κάνουμε και σε αυτή την περίπτωση φαίνεται να ενισχύεται καθώς είναι προς το παρόν το μοναδικό κριτήριο που δείχνει τελείως διαφορετικά αποτελέσματα από τα υπόλοιπα.

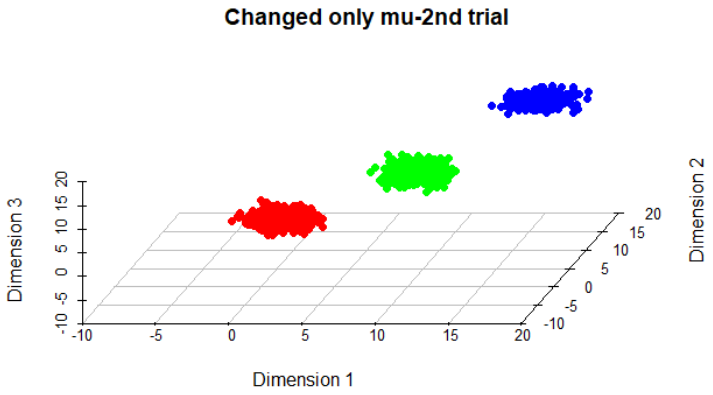
Συγκρίνοντας τις 2 πρώτες δοκιμές παρατηρούμε κάποια κοινά κριτήρια τα οποία σταθερά προτείνουν λανθασμένο αριθμό συστάδων, ενώ άλλα φαίνεται πως επηρεάστηκαν από την αλλαγή των μέσων και τις αποστάσεις των ομάδων, περαιτέρω συμπεράσματα θα αναπτυχθούν στο τέλος των δοκιμών.

- **Δοκιμή 2 (Διαφορετικοί μέσοι ,ίδιο Σ)**

Στη δεύτερη δοκιμή και εφόσον έχει ήδη εντοπιστεί από τις παραπάνω δοκιμές μειωμένη απόδοση αναφορικά με την ακρίβεια των ομάδων ακόμη και όταν έχουμε ισχυρά διακριτά clusters δοκιμάσθηκε να απομακρυνθούν περαιτέρω οι μέσοι των ομάδων σε σχέση με τα initial data και να διατηρηθεί και πάλι το ίδιο Σ. Παρακάτω οι νέοι μέσοι για αυτή τη δοκιμή.

```
mu1v1 <- c(1, 2, 3)
mu2v1 <- c(9, 8, 8)
mu3v1 <- c(16, 14, 19)
```

Παρακάτω η δομή των δεδομένων όπως προκύπτει από τη ρύθμιση των παραμέτρων.



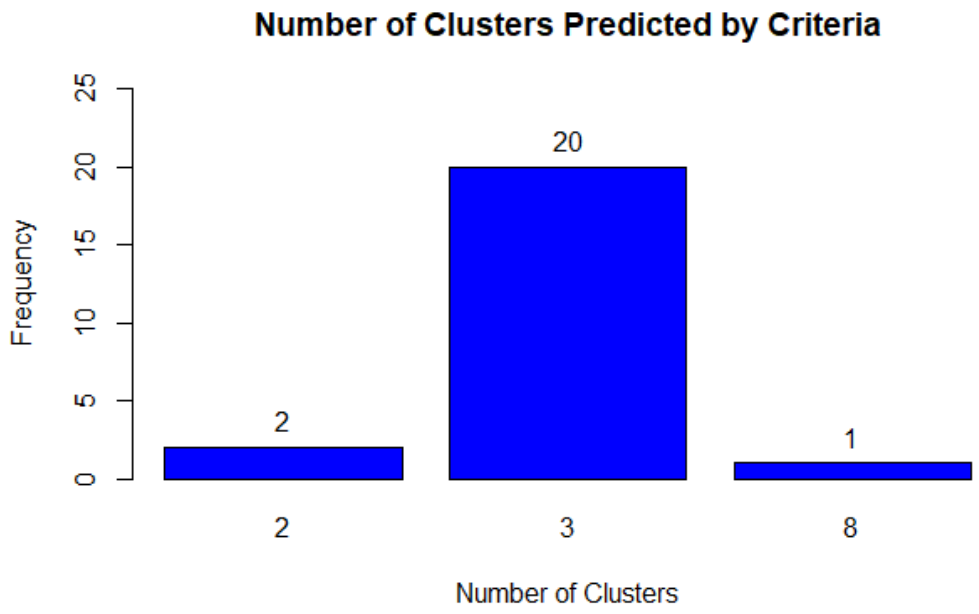
Σχήμα 3.13 Απεικόνιση των δεδομένων στο στάδιο της Δοκιμής 2

Παρατηρούμε ότι οι αρχικές πλέον ομάδες είναι ξεκάθαρα οριοθετημένες και θα περιμέναμε πως, αν όχι το σύνολο των κριτηρίων, η συντριπτική τους πλειοψηφία να προβλέψει 3 clusters. Παρουσιάζουμε παρακάτω τα αποτελέσματα όπως προέκυψαν τρέχοντας τις αντίστοιχες εντολές.

```
$All.index
       KL        CH  Hartigan     CCC     Scott      Marriot   TrCovW    TraceW Friedman    Rubin Cindex     DB Silhouette   Duda  Pseudot2   Beale
2  2.2595  5324.861 9866.9945 51.5639  5731.166 346559315776 23476228 36172.126 209.9671  14.4759 0.3362 0.3894     0.7110 0.0928 9759.8480 16.6319
3 39.9224 25115.992  122.2166 82.4846 10346.693  35944129497  2612381  4767.784 326.6538 109.8251 0.3102 0.2958     0.8089 0.7796  140.8248  0.4804
4  1.7014 18139.605  137.1264 64.5437 10739.752  49170363681  2337439  4407.917 353.1622 118.7913 0.3015 1.0776     0.6133 0.7656  152.4925  0.5203
5  0.9543 14876.069   94.5491 54.6670 11170.360  57656681395  1848493  4037.804 385.2123 129.6800 0.3176 1.5201     0.4256 0.9770    7.6600  0.0400
6  0.9669 12663.941   67.3272 47.7673 11479.814  67548528314  1597525  3797.629 416.1932 137.8814 0.3138 1.4044     0.4277 1.3693  -68.7698 -0.4562
7  2.3705 11032.700  179.2125 42.6202 11642.900  82469091471  1556763  3633.868 431.2802 144.0950 0.3129 1.3154     0.4289 0.7498  166.1803  0.5669
8  1.4825 10610.215   95.4938 40.9244 12111.582  78809371844  1260782  3244.423 483.6259 161.3915 0.2908 1.5523     0.2276 1.1862  -51.3251 -0.2660
  Ratkowsky       Ball Ptbiserial    Frey McClain    Dunn Hubert SDindex Dindex   SDbw
2    0.6177 18086.0627     0.8089  0.7730  0.2407 0.4570      0  0.3567 4.4042 0.1839
3    0.5684  1589.2615     0.8128 16.4173  0.2736 0.5474      0  0.2461 1.6389 0.0282
4    0.4930  1101.9794     0.7180 10.4213  0.3690 0.0159      0  1.6970 1.5664 0.1926
5    0.4416   807.5608     0.6247 11.3709  0.5066 0.0175      0  1.5808 1.4942 0.2256
6    0.4034   632.9381     0.5924 25.6875  0.5707 0.0175      0  1.6422 1.4423 0.2963
7    0.3737   519.1240     0.5759  3.9956  0.6089 0.0137      0  1.7777 1.4090 0.5747
8    0.3499   405.5529     0.4804  4.1426  0.8733 0.0140      0  1.4686 1.3367 0.4662

$All.CriticalValues
  CritValue_Duda CritValue_PseudoT2 Fvalue_Beale
2         0.7172           393.5053       0.0000
3         0.6880           225.8668       0.6959
4         0.6880           225.8668       0.6684
5         0.6504           174.6783       0.9893
6         0.6113           162.1218       1.0000
7         0.6880           225.8668       0.6368
8         0.6403           183.7032       1.0000

$Best.nc
                     KL       CH Hartigan     CCC    Scott      Marriot   TrCovW   TraceW Friedman   Rubin Cindex     DB Silhouette   Duda PseudoT2
Number_clusters  3.0000     3.00    3.000  3.0000    3.000            3        3     3.00   3.0000   3.000 8.0000 3.0000     3.0000 3.0000   3.0000
Value_Index     39.9224 25115.99 9744.778 82.4846 4615.528 323841420463 20863847 31044.47 116.6867 -86.383 0.2908 0.2958     0.8089 0.7796 140.8248
                  Beale Ratkowsky    Ball PtBiserial Frey McClain   Dunn Hubert SDindex Dindex   SDbw
Number_clusters 3.0000    2.0000     3.0     3.0000    1  2.0000 3.0000      0  3.0000      0 3.0000
Value_Index     0.4804    0.6177 16496.8     0.8128   NA  0.2407 0.5474      0  0.2461      0 0.0282
```

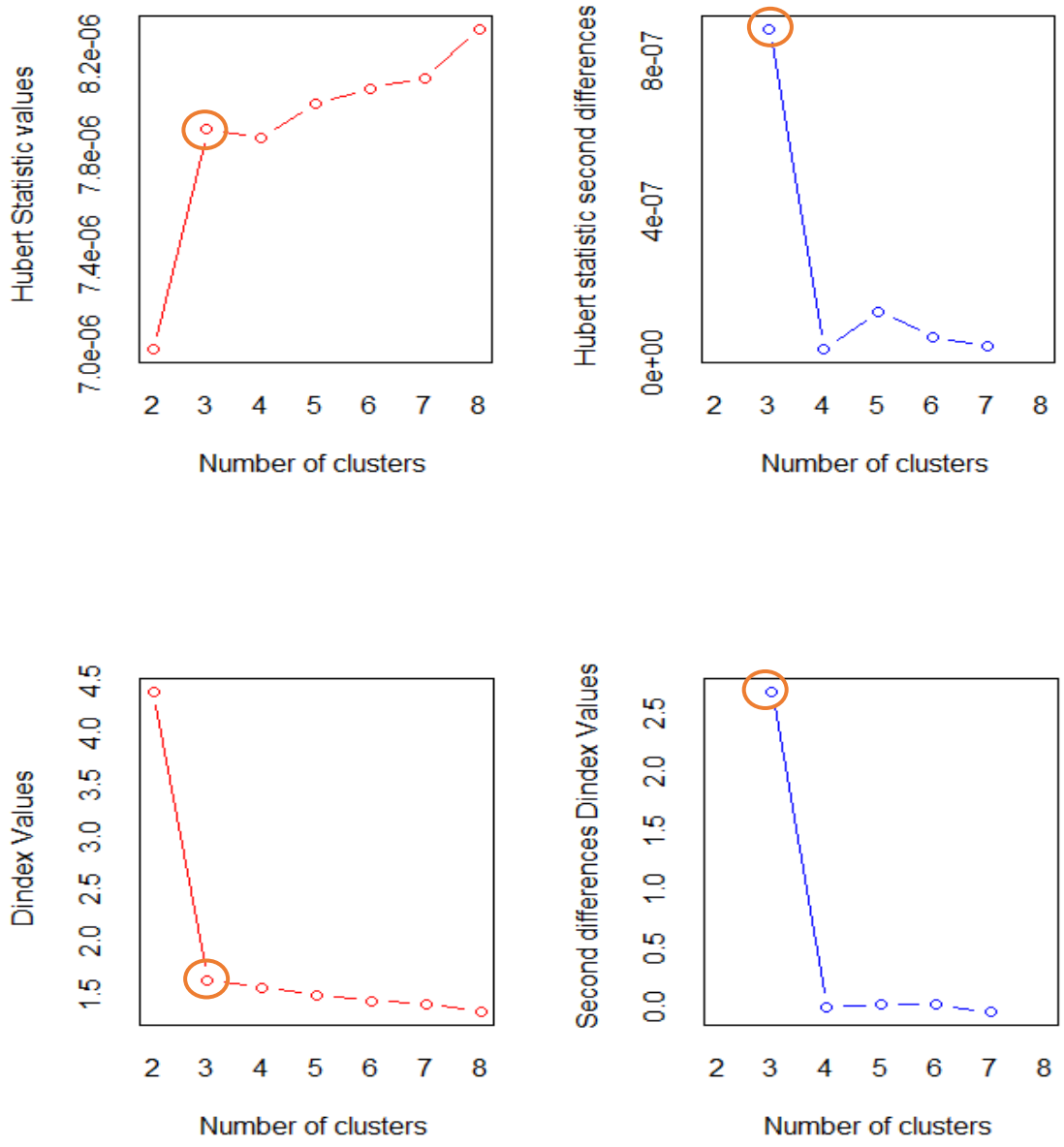
Σχήμα 3.14 Απεικόνιση των αποτελεσμάτων της Δοκιμής 2

Όπως φαίνεται και από το παρακάτω γράφημα στη συγκεκριμένη δοκιμή έχουμε τον υψηλότερο αριθμό κριτηρίων να πετυχαίνει και να προτείνει 3 clusters.



Σχήμα 3.15 Απεικόνιση barplot σχετικά με το πλήθος των κριτηρίων που πρότειναν κάθε αριθμό clusters στη Δοκιμή 2

Σε συνέχεια των παραπάνω αποτελεσμάτων, παρατηρούμε ότι και τα κριτήρια που προτείνουν μέσω γραφικών μεθόδων καταλήγουν στα 3 clusters.



Σχήμα 3.16 Γραφικές μεθόδοι κριτηρίων-Δοκιμή 2

Σε αυτή τη δοκιμή τα 3 κριτήρια που χάνουν το «σωστό» αριθμό ομάδων είναι το κριτήριο McClain, Ratkowsky και το κριτήριο Cindex.

Παρατηρείται από τις 3 πρώτες δοκιμές ότι τα υπάρχουν συγκεκριμένα κριτήρια που αποτυγχάνουν να προβλέψουν επανειλημμένως είτε διαχωρίζουμε πλήρως τις ομάδες είτε μειώνουμε την μεταξύ τους απόσταση.

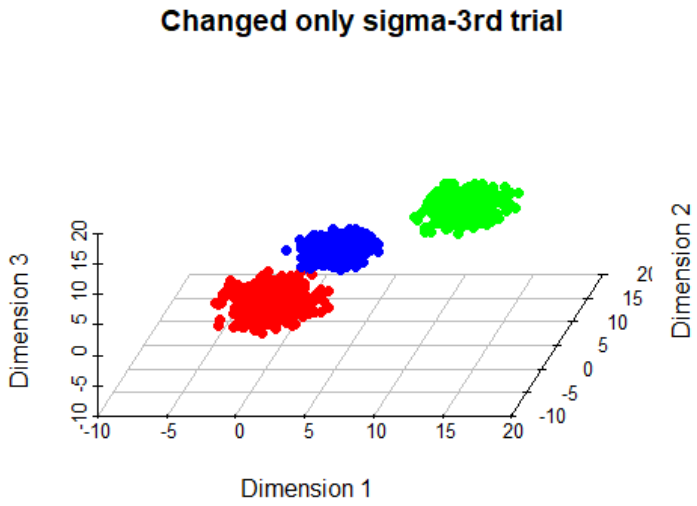
- **Δοκιμή 3 (Μέσοι από initial data, διαφορετικό Σ)**

Αφού προχωρήσαμε σε δοκιμές με διαφορετικές παραλλαγές των διανυσμάτων των μέσων, τώρα θα εξετάσουμε την περίπτωση που διαφοροποιείται το Σ.
Πιο συγκεκριμένα επιλέχθηκε ο αρχικός ταυτοτικός πίνακας να παραλλαχθεί ως προς τα στοιχεία της κύριας διαγωνίου του, τις θέσεις δηλαδή που αφορούν τη διασπορά κάθε πληθυσμού. Παρακάτω η διαμόρφωση των νέων πινάκων Σ για κάθε πληθυσμό. Σε αυτή τη φάση τα στοιχεία εκτός διαγωνίου επιλέχθηκε να παραμείνουν μηδενικά. Τέλος, θα χρησιμοποιήσουμε τα αρχικά διανύσματα μέσων για κάθε ομάδα.

```
> sigma1
     [,1] [,2] [,3]
[1,]    2    0    0
[2,]    0    2    0
[3,]    0    0    2
> sigma2
     [,1] [,2] [,3]
[1,]  1.6  0.0  0.0
[2,]  0.0  1.6  0.0
[3,]  0.0  0.0  1.4
> sigma3
     [,1] [,2] [,3]
[1,]  1.1  0.0  0.0
[2,]  0.0  1.1  0.0
[3,]  0.0  0.0  1.1
```

Σχήμα 3.17 Απεικόνιση του πίνακα Σ των προσομοιωμένων δεδομένων στη Δοκιμή 3

Ας δούμε τη δομή των δεδομένων μας όπως προέκυψε από τη ρύθμιση των παραμέτρων μας.



Σχήμα 3.18 Απεικόνιση των δεδομένων στο στάδιο της Δοκιμής 3

Από το παραπάνω γράφημα έχουμε τις 2 ομάδες να είναι αρκετά κοντά μεταξύ τους και η 3η να απέχει αρκετά. Ο βαθμός διαχωρισμού των 3 ομάδων είναι εμφανής σε αυτή την περίπτωση και ουσιαστικά θα δούμε κατά πόσο τα κριτήρια θα αλλάξουν την πρόταση τους σχετικά με το βέλτιστο αριθμό ομάδων σε σχέση με τα αποτελέσματα των initial data που τα διανύσματα των μέσων είναι ίδια, θέλουμε λοιπόν σε αυτή την περίπτωση κατά πόσο μία αλλαγή αποκλειστικά στον πίνακα Σ μπορεί να μεταβάλει τα αποτελέσματα των κριτηρίων.

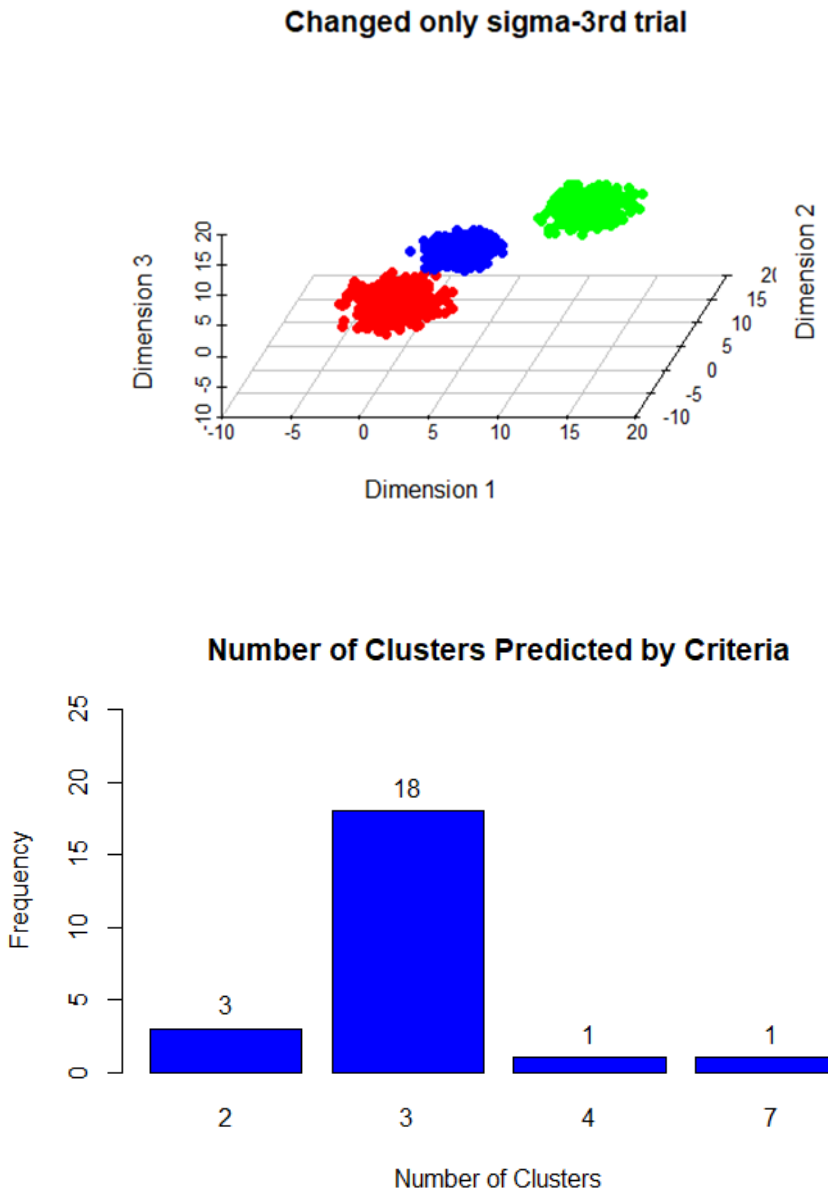
Παρουσιάζουμε παρακάτω τα αποτελέσματα όπως προέκυψαν τρέχοντας τις αντίστοιχες εντολές.

```
> print(nbclust_result2)
$All.index
       KL       CH  Hartigan     CCC    Scott      Marriot  TrCovW    TraceW Friedman   Rubin Cindex     DB Silhouette   Duda  Pseudot2   Beale Ratkowsky       Ball
2  2.0723 3829.543 3391.6918 35.1984 4538.494 476920313984 8163339 24034.824  53.0687  8.6878 0.3280 0.5067     0.6323 0.2268 3405.7058  5.7979    0.5860 12017.4119
3 12.6468 7940.637  125.2320 45.6976 7676.979 132417073054 6325331  7363.279  82.0221 28.3582 0.2707 0.4974     0.6730 1.5604 -275.1051 -0.6102    0.5502  2454.4264
4  3.0693 5774.493  182.2870 32.9532 8061.890 182128698166 5560859  6794.854  90.1529 30.7305 0.2623 1.1885     0.5114 1.6093 -278.2762 -0.6433    0.4786  1698.7134
5  0.8238 4900.878  127.7843 37.7600 8666.472 190175228456 4155245  6056.831 106.7602 34.4750 0.2756 1.5515     0.3799 0.9650   11.7549  0.0615    0.4302  1211.3662
6  0.8768 4278.516   91.6091 34.4740 8998.069 219537661687 3313489  5579.893 114.1559 37.4217 0.2933 1.4236     0.3801 1.2009 -103.7354 -0.2843    0.3939   929.9821
7 13.1385 3796.780  133.7620 31.5676 9252.138 252256392875 3202043  5257.512 119.6542 39.7163 0.2943 1.3395     0.3744 0.7495  166.7847  0.5679    0.3656   751.0732
8  0.1110 3562.673   76.5106 30.6592 9614.274 258807650200 2728091  4825.209 127.8045 43.2746 0.2789 1.5404     0.2291 1.2969  -84.0144 -0.3880    0.3431   603.1511
  Ptbiserial   Frey McClain   Dunn Hubert SDindex Dindex   SDbw
2     0.7636 0.7744  0.2972 0.1583      0  0.4444 3.6811 0.3202
3     0.7669 7.7418  0.4739 0.0916      0  0.3905 2.0230 0.0949
4     0.6824 2.4416  0.6323 0.0141      0  1.1790 1.9307 0.2352
5     0.6102 2.1558  0.8191 0.0161      0  1.0684 1.8289 0.2689
6     0.5846 2.3201  0.9003 0.0177      0  1.1030 1.7557 0.2992
7     0.5708 4.0408  0.9482 0.0160      0  1.1780 1.7090 0.4588
8     0.4757 6.0047  1.4121 0.0160      0  1.3822 1.6329 0.4100

$All.CriticalValues
  CritValue_Duda CritValue_PseudoT2 Fvalue_Beale
2         0.7172           393.8996       0.0006
3         0.6884           346.7756       1.0000
4         0.6883           332.8949       1.0000
5         0.6502           174.3407       0.9800
6         0.6882           280.9390       1.0000
7         0.6881           226.2153       0.6362
8         0.6409           205.6023       1.0000

$Best.nc
                     KL       CH Hartigan     CCC    Scott      Marriot  TrCovW    TraceW Friedman    Rubin Cindex     DB Silhouette   Duda  PseudoT2   Beale Ratkowsky
Number_clusters  7.0000    3.000     3.00  3.0000    3.000            3       3      3.00   3.0000    3.0000 4.0000 3.0000      3.000 3.0000    3.0000  3.0000     2.000
Value_Index     13.1385 7940.637  3266.46 45.6976 3138.485 394214866042 1838008 16103.12  28.9534 -17.2981 0.2623 0.4974      0.673 1.5604 -275.1051 -0.6102     0.586
                   Ball PtBiserial Frey McClain   Dunn Hubert SDindex Dindex   SDbw
Number_clusters    3.000     3.0000    1  2.0000 2.0000      0  3.0000      0 3.0000
Value_Index     9562.986     0.7669   NA  0.2972 0.1583      0  0.3905      0 0.0949
```

Σχήμα 3.19 Απεικόνιση των αποτελεσμάτων της Δοκιμής 3

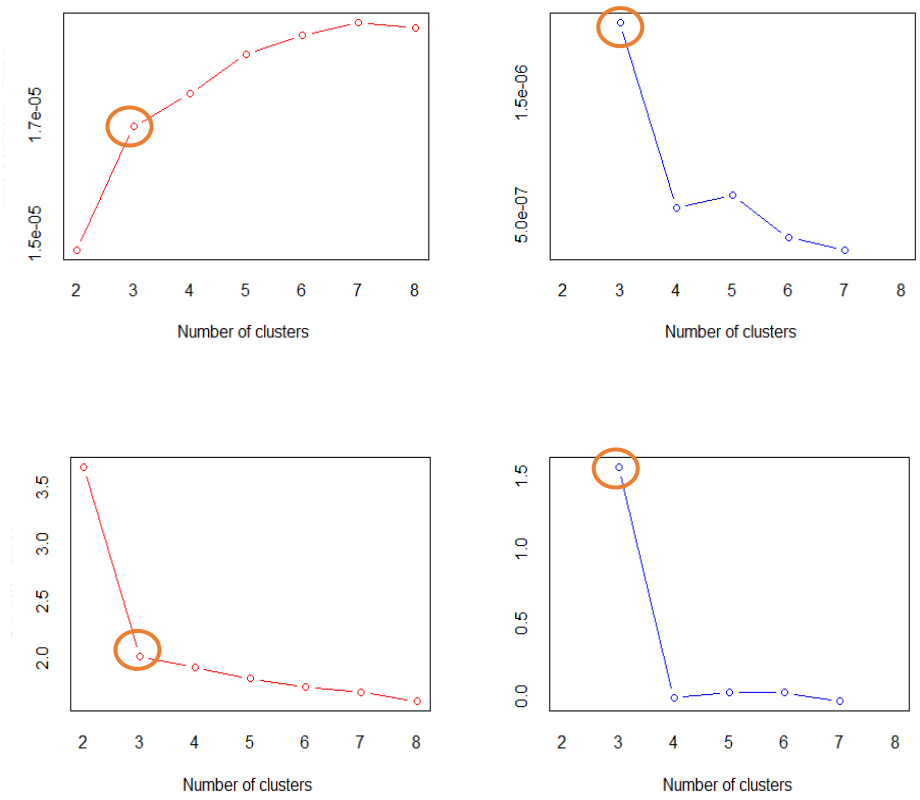
Όπως φαίνεται και από το παρακάτω γράφημα στη συγκεκριμένη δοκιμή έχουμε τον υψηλότερο αριθμό κριτηρίων να πετυχαίνει και να προτείνει 3 clusters.

Πιο συγκεκριμένα 18 κριτήρια προτείνουν 3 clusters σε σχέση με την δοκιμή στα Initial data όπου είχαμε 19 κριτήρια.



Σχήμα 3.20 Απεικόνιση barplot σχετικά με το πλήθος των κριτηρίων που πρότειναν κάθε αριθμό clusters στη Δοκιμή 3

Ας δούμε και τα κριτήρια μέσω γραφικής αναπαράστασης



Σχήμα 3.21 Γραφικές μεθόδοι κριτηρίων-Δοκιμή 3

Τα κριτήρια που αποτυγχάνουν να προτείνουν τον αναμενόμενο αριθμό ομάδων είναι τα εξής: KL, McClain, Ratkowsky, Dunn και Cindex.

Αυτό που παρατηρείται είναι πως πέραν του κριτηρίου του Dunn, τα υπόλοιπα κριτήρια είχαν αστοχία και κατά τη δοκιμή των Initial data.

Επιπλέον παρατηρούμε ότι το κριτήριο KL σε αμφότερες τις περιπτώσεις προβλέπει τελείως λανθασμένο αριθμό ομάδων, προφανώς σε κάποιο cluster έχουμε σημεία με αυξημένο πυρήνα σημείων.

Το κριτήριο Cindex παρόλο που δε φαίνεται να είναι αξιόπιστο κριτήριο για τις δοκιμές φαίνεται πως επηρεάζεται αρκετά σε αυτή τη δοκιμή καθώς προτείνει 4 ομάδες αντί 8 στα Initial data.

- **Δοκιμή 4 (Διαφορετικοί μέσοι, διαφορετικό Σ)**

Στην συγκεκριμένη δοκιμή αποφασίσθηκε να αλλαχθούν και οι δυο παράμετροι ενδιαφέροντος με τέτοιο τρόπο ώστε να φέρουμε κοντά 2 ομάδες και μία μακρύτερα και από τις 2 .

Πιο συγκεκριμένα, παίρνοντας τα initial διανύσματα των μέσων φέραμε πιο κοντά 2 από αυτά πολλαπλασιάζοντας τα με 0.7 αμφότερα. Αναφορικά με το πίνακα Σ αυτός

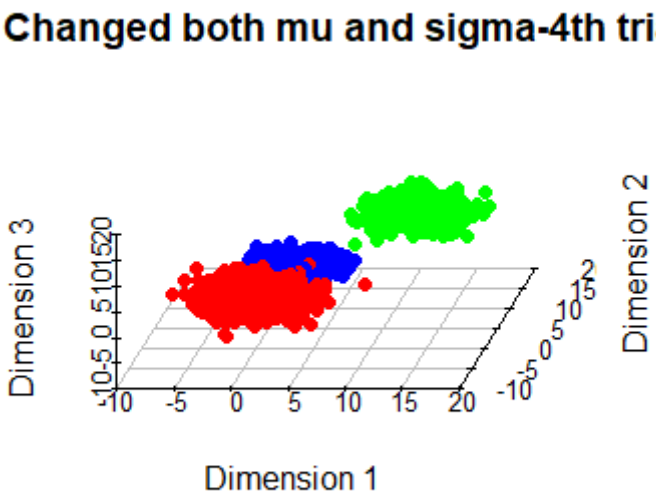
δημιουργήθηκε τυχαία με για κάθε ομάδα. Ουσιαστικά οι παράμετροι ρυθμίστηκαν κάθε φορά ώστε να αναδεικνύουν το θέμα που εξετάζουμε, στην προκειμένη θα θέλαμε να δούμε με την αλλαγή και των 2 παραμέτρων με τυχαίο τρόπο και δημιουργώντας ουσιαστικά 2 ομάδες αρκετά κοντά και μια πιο απομακρυσμένη τι θα προτείνουν τα κριτήρια. Παραθέτουμε τους τελικούς πίνακες Σ και τα νέα διανύσματα μέσων.

```
> sigma11
     [,1] [,2] [,3]
[1,] 5.45 0.00 0.00
[2,] 0.00 3.15 0.00
[3,] 0.00 0.00 4.35
> sigma22
     [,1] [,2] [,3]
[1,] 3.92 0.00 0.00
[2,] 0.00 1.62 0.00
[3,] 0.00 0.00 2.82
> sigma33
     [,1] [,2] [,3]
[1,] 2.66 0.00 0.00
[2,] 0.00 0.36 0.00
[3,] 0.00 0.00 1.56
```

```
> mu1*0.7
[1] 0.0 0.7 0.0
> mu2
[1] 12 10  9
> mu3*0.7
[1] 2.8 4.2 3.5
```

Σχήμα 3.22 Παράμετροι μ, Σ των προσομοιωμένων δεδομένων στη Δοκιμή 4

Παρακάτω μπορούμε να δούμε τη μορφή των δεδομένων μας μετά τη ρύθμιση των παραμέτρων και ακριβώς παρακάτω παρουσιάζονται τα αποτελέσματα των κριτηρίων όπως προέκυψαν.



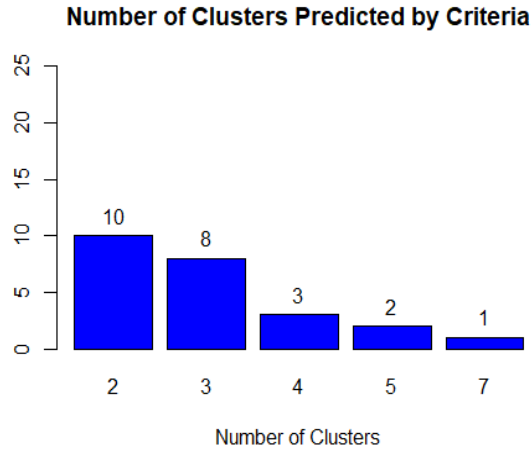
Σχήμα 3.23 Απεικόνιση των δεδομένων στο στάδιο της Δοκιμής 4

```
$All.index
       KL       CH Hartigan     CCC    Scott      Marriot   TrCovW    TraceW Friedman   Rubin Cindex     DB Silhouette   Duda
2  8.3907 4968.361 970.8166 38.5777 4315.457 1.053507e+12 24670116 22155.031  30.3281  8.8171 0.3016 0.4866     0.6723 0.5062
3  3.1009 4576.467 177.3329 27.8607 6181.742 6.830856e+11 24766456 13442.974  44.4660 14.5313 0.2626 0.8193     0.5250 0.6878
4 21.0421 3469.186 246.3078 18.7715 6567.989 9.386906e+11 18399458 12019.194  48.2943 16.2526 0.2512 1.2447     0.3883 1.5406
5  0.1518 3089.785 153.4607 23.1042 7241.712 9.360083e+11 11506233 10320.056  56.2389 18.9285 0.2483 1.4037     0.3540 1.2830
6  1.3446 2754.408 191.2170 20.9581 7585.080 1.072079e+12 10004579  9359.327  60.2511 20.8715 0.2561 1.3537     0.3341 0.9034
7  0.6310 2619.236  94.1615 21.0272 8178.957 9.821454e+11  7379041  8297.350  71.1419 23.5428 0.2691 1.3201     0.2797 1.6493
8  0.7976 2398.496  68.1324 19.4620 8443.589 1.075327e+12  6525522  7805.093  75.7661 25.0277 0.2620 1.3488     0.2722 1.2466
   Pseudot2   Beale Ratkowsky       Ball Ptbiserial     Frey McClain   Dunn Hubert SDindex Dindex   SDbw
2  974.4601  1.6589    0.6156 11077.5156     0.8504   2.1223  0.2655 0.1221      0  0.3125 3.5351 0.2443
3  225.5628  0.7711    0.5350  4480.9912     0.6933   3.7857  0.5974 0.0295      0  0.5188 2.6870 0.1862
4 -272.6571 -0.5963    0.4663  3004.7985     0.6255   1.1754  0.7802 0.0134      0  0.7571 2.5321 0.3013
5 -142.7310 -0.3748    0.4212  2064.0112     0.5794   1.2282  0.9559 0.0149      0  0.7176 2.3681 0.2809
6   49.0651  0.1816    0.3864  1559.8879     0.5540   3.2211  1.0632 0.0162      0  0.7759 2.2684 0.5768
7 -211.4033 -0.6681    0.3598  1185.3357     0.4770   1.0317  1.4881 0.0101      0  0.8758 2.1300 0.3786
8  -77.9393 -0.3354    0.3375   975.6367     0.4639 -21.8645  1.5790 0.0078      0  0.9311 2.0709 0.3183

$All.CriticalValues
  CritValue_Duda CritValue_PseudoT2 Fvalue_Beale
2         0.7172            393.8996       0.1737
3         0.6879            225.5183       0.5102
4         0.6925            344.9720       1.0000
5         0.6902            290.4210       1.0000
6         0.6880            208.1785       0.9089
7         0.6622            273.8975       1.0000
8         0.6477            214.2684       1.0000

$Best.nc
                     KL       CH Hartigan     CCC    Scott      Marriot  TrCovW   TraceW Friedman   Rubin Cindex     DB Silhouette
Number_clusters  4.0000    2.000   3.0000  2.0000   3.000            3       5    3.000    3.000  3.0000 5.0000 2.0000     2.0000
Value_Index     21.0421 4968.361 793.4838 38.5777 1866.285 626026587411 6893225 7288.278   14.138 -3.9928 0.2483 0.4866     0.6723
                   Duda  PseudoT2  Beale Ratkowsky     Ball PtBiserial   Frey McClain   Dunn Hubert SDindex Dindex   SDbw
Number_clusters 4.0000    4.0000 2.0000    2.0000    3.000     2.0000 7.0000  2.0000 2.0000      0  2.0000      0 3.0000
Value_Index     1.5406 -272.6571 1.6589    0.6156 6596.524     0.8504 1.0317  0.2655 0.1221      0  0.3125      0 0.1862
```

Σχήμα 3.22 Απεικόνιση των αποτελεσμάτων της Δοκιμής 4

Από τα παραπάνω outputs παρατηρούμε ότι παρουσιάζεται μία οριακή κατάσταση μεταξύ των κριτηρίων αναφορικά με τον προτεινόμενο αριθμό ομάδων. Και αυτό διότι όπως θα δούμε για πρώτη φορά δεν μπορούμε να έχουμε ξεκάθαρη πρόταση μέσω αυτών.

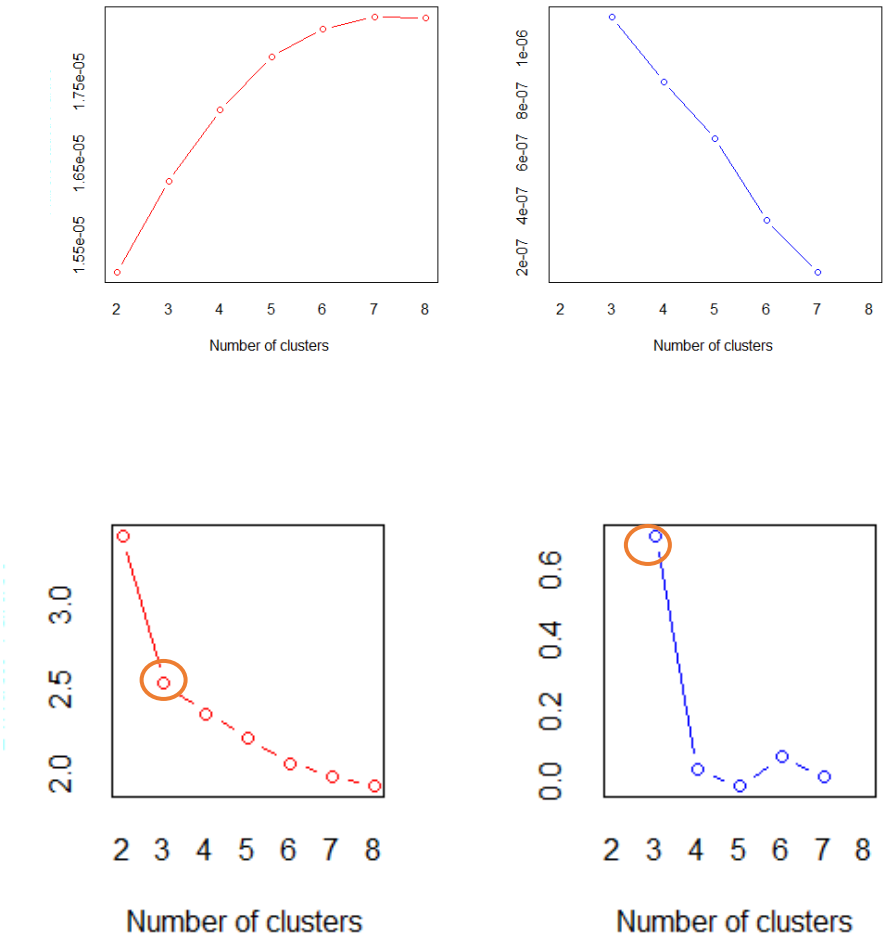
Όπως φαίνεται στο παρακάτω γράφημα 10 κριτήρια προτείνουν ως βέλτιστο αριθμό ομάδων τις 2 , ενώ 8 κριτήρια προτείνουν 3 ομάδες.



Σχήμα 3.24 Απεικόνιση barplot σχετικά με το πλήθος των κριτηρίων που πρότειναν κάθε αριθμό clusters στη Δοκιμή 4

Παρατηρούμε την οριακή κατάσταση στην οποία είμαστε και από τα παρακάτω γραφήματα όπου στο πρώτο τουλάχιστον σετ γραφημάτων(κριτήριο Hubert) δεν είναι και γραφικά

ξεκάθαρο το αν η πρόταση είναι 2 ή 3 ομάδες. Αναφορικά με το δεύτερο σετ γραφημάτων που αφορούν το κριτήριο Dindex έχουμε πρόταση 3 ομάδων.



Σχήμα 3.25 Γραφικές μεθόδοι κριτηρίων-Δοκιμή 4

Για την παραπάνω δοκιμή δεν μπορούμε να αποφανθούμε με βεβαιότητα μέσω των κριτηρίων για το βέλτιστο αριθμό ομάδων. Στη συγκεκριμένη περίπτωση λοιπόν, μία σκέψη θα ήταν η περαιτέρω και σε μεγαλύτερο βάθος κατανόηση της δομής των δεδομένων μας και η περαιτέρω αξιολόγηση των καταλληλότερων κριτηρίων για να καταλήξουμε σε ένα πόρισμα.

Από τα παραπάνω γίνεται αντιληπτό πως τα κριτήρια είναι εδώ για να μας βοηθήσουν αλλά δεν μπορούν σε όλες τις περιπτώσεις να μας οδηγήσουν σε συμπεράσματα, καθώς προφανώς κάθε κριτήριο επηρεάζεται διαφορετικά και σε διαφορετικό βαθμό από την απόσταση και την πυκνότητα των ομάδων με βάση τον τρόπο υπολογισμού του.

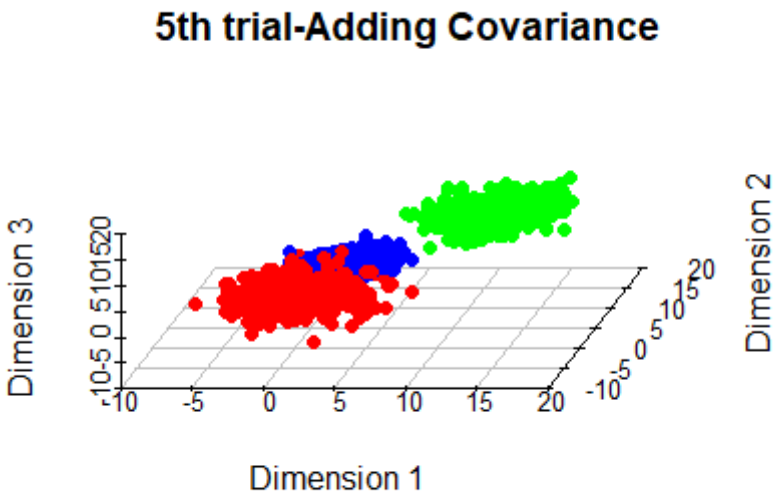
- **Δοκιμή 5 (Παράμετροι από δοκιμή 4, αντικατάσταση των μη διαγώνιων στοιχείων του πίνακα Σ με μη μηδενικές τιμές)**

Στην παρακάτω δοκιμή θα κρατήσουμε τις παραμέτρους από τη δοκιμή 4 ίδιες , όμως θα τροποποιήσουμε τα μη διαγώνια στοιχεία του πίνακα Σ και μόνο αυτά. Στόχος μας είναι να υπάρχει πλέον εξάρτηση μεταξύ των ομάδων και να δούμε αν και κατά πόσο μπορεί να επηρεάσει αυτή την πρόταση βέλτιστου αριθμού ομάδων από τα κριτήρια. Τα στοιχεία στην κύρια διαγώνιο παραμένουν σταθερά.
Παρακάτω παρουσιάζονται οι τελικοί πίνακες Σ για κάθε ομάδα.

```
> sigma111
     [,1] [,2] [,3]
[1,] 5.45 0.50 0.30
[2,] 0.50 3.15 0.20
[3,] 0.30 0.20 4.35
> sigma222
     [,1] [,2] [,3]
[1,] 3.92 0.40 0.60
[2,] 0.40 1.62 0.70
[3,] 0.60 0.70 2.82
> sigma333
     [,1] [,2] [,3]
[1,] 2.66 0.30 0.80
[2,] 0.30 0.36 0.40
[3,] 0.80 0.40 1.56
```

Σχήμα 3.26 Παράμετροι Σ των προσομοιωμένων δεδομένων στη Δοκιμή 5

Η δομή των δεδομένων παίρνει την κάτωθι μορφή και ουσιαστικά μας δίνει πιο μακρόστενες ομάδες.



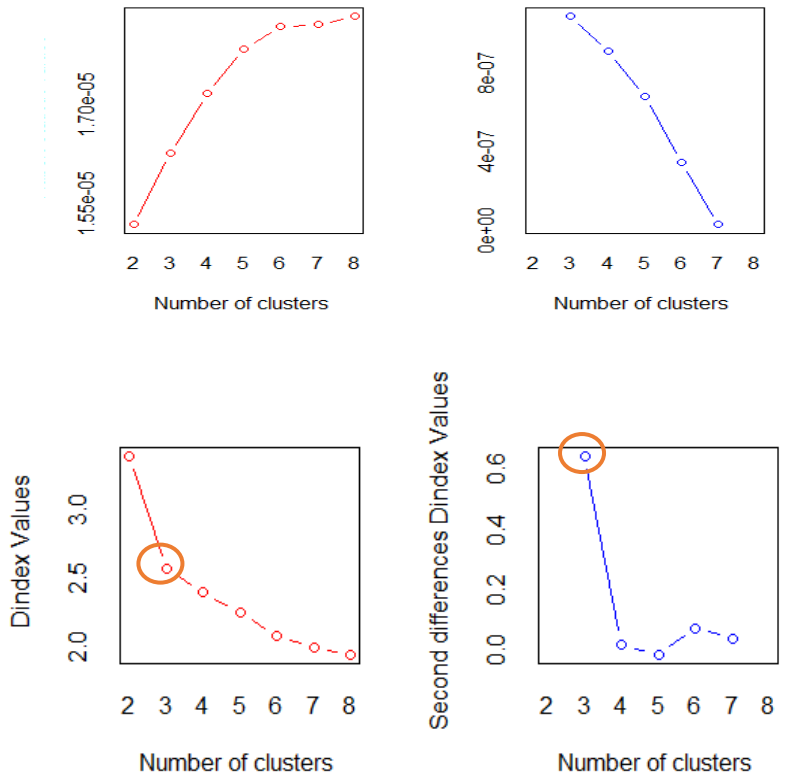
Σχήμα 3.27 Απεικόνιση των δεδομένων στο στάδιο της Δοκιμής 5.

Όπως παρατηρούμε από τους παρακάτω πίνακες προκύπτει και πάλι μία σχετικά οριακή εικόνα των κριτηρίων που αμφιταλαντεύονται ανάμεσα στις 2 και 3 ομάδες ως βέλτιστο αριθμό.

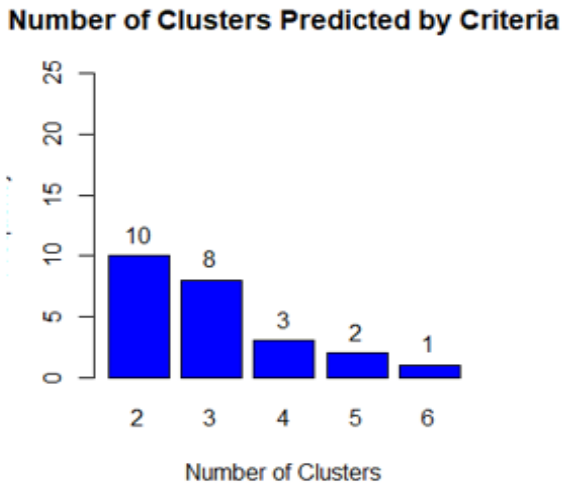
```
$All.index
        KL       CH  Hartigan     CCC    Scott      Marriot   TrCovW    TraceW Friedman   Rubin Cindex     DB Silhouette   Duda  Pseudot2   Beale Ratkowsky
2  6.8557 4819.256 1090.3984 36.7633 4222.214 788890359895 17212666 22464.367  36.1411  8.6830 0.2954 0.4944     0.6660 0.4708 1121.9741  1.9120    0.6121
3  4.9919 4705.659  207.6760 27.1903 5853.700 598185344544 16950973 13000.944  46.9143 15.0034 0.2539 0.7752     0.5317 0.6579  258.9407  0.8834    0.5360
4 31.1036 3639.104  236.6119 18.0801 6333.867 772130175754 15744773 11417.074  51.3751 17.0847 0.2407 1.1620     0.4026 1.5574 -279.5350 -0.6082    0.4684
5  0.1158 3218.007  221.7708 20.9525 6932.168 809626340756  9280145  9857.916  58.6956 19.7869 0.2282 1.3555     0.3637 1.0602  -29.8172 -0.0965    0.4223
6  0.7183 2998.644  123.6124 20.3067 7470.103 814515079976  8835779  8584.480  66.5666 22.7221 0.2145 1.3430     0.2972 1.7097 -190.1129 -0.7045    0.3880
7  0.6085 2724.402   79.5535 18.4215 7870.998 848636152171  7655071  7928.483  72.5423 24.6022 0.2689 1.3010     0.2954 1.5999 -201.3538 -0.6364    0.3611
8  1.6767 2469.341   87.5609 16.2760 8152.759 918602016625  6906420  7527.391  78.3980 25.9131 0.2544 1.3803     0.2826 1.2027  -67.0659 -0.2858    0.3387
        Ball PtBiserial   Frey McClain   Dunn Hubert SDindex Dindex   SDbw
2 11232.1837     0.8296 1.9640  0.2683 0.0562      0  0.3137 3.5188 0.3288
3  4333.6479     0.6863 3.1651  0.5900 0.0166      0  0.4959 2.6398 0.2033
4  2854.2685     0.6218 1.2746  0.7640 0.0121      0  0.7453 2.4709 0.2961
5  1971.5832     0.5734 2.7911  0.9426 0.0118      0  0.7157 2.3147 0.2652
6  1430.7467     0.4942 0.8975  1.3249 0.0038      0  0.8229 2.1487 0.2613
7  1132.6405     0.4769 1.3665  1.4306 0.0155      0  0.8432 2.0742 0.3432
8   940.9239     0.4598 4.3633  1.5462 0.0140      0  0.9135 2.0218 0.4343

$All.CriticalValues
  CritValue_Duda CritValue_PseudoT2 Fvalue_Beale
2         0.7172            393.5738       0.1254
3         0.6880            225.8668       0.4490
4         0.6926            346.6068       1.0000
5         0.6899            235.9685       1.0000
6         0.6651            230.6107       1.0000
7         0.6636            272.2257       1.0000
8         0.6517            212.7093       1.0000

$Best.nc
                    KL       CH Hartigan     CCC    Scott      Marriot  TrCovW   TraceW Friedman  Rubin Cindex     DB Silhouette   Duda PseudoT2 Beale
Number_clusters  4.0000    2.000   3.0000  2.0000    3.000            3       5    3.000   3.0000  3.000 6.0000 2.0000       2.000 4.0000    4.000 2.000
Value_Index     31.1036 4819.256 882.7224 36.7633 1631.485 364649846561 6464628 7879.554  10.7732 -4.239 0.2145 0.4944       0.666 1.5574 -279.535 1.912
                Ratkowsky     Ball PtBiserial   Frey McClain   Dunn Hubert SDindex Dindex   SDbw
Number_clusters    2.0000    3.000     2.0000 5.0000  2.0000 2.0000      0  2.0000      0 3.0000
Value_Index        0.6121 6898.536     0.8296 2.7911  0.2683 0.0562      0  0.3137      0 0.2033
```

Σχήμα 3.28 Απεικόνιση των αποτελεσμάτων της 5ης Δοκιμής

Παρατηρούμε ότι παρόμοια είναι η εικόνα και για τα κριτήρια που προτείνουν μέσω γραφικών μεθόδων.



Σχήμα 3.29 Γραφικές μεθόδοι κριτηρίων-Δοκιμή5



Σχήμα 3.30 Απεικόνιση barplot σχετικά με το πλήθος των κριτηρίων που πρότειναν κάθε αριθμό clusters στη Δοκιμή 5

Συγκεντρωτικά παρατηρούμε ότι τα αποτελέσματα μεταξύ των 2 δοκιμών είναι ίδια πέρα του κριτηρίου C index που έτσι και αλλιώς δε φαίνεται να είναι αξιόπιστο στη ερευνά μας και το κριτήριο Frey με το οποίο δεν καταπιαστήκαμε καθώς δεν είχαμε διαθέσιμα αποτελέσματα σε όλες τις δοκιμές.

Θα μπορούσαμε να ισχυριστούμε ότι, η τοποθέτηση μη μηδενικών τιμών στα στοιχεία που δεν ανήκουν στην κύρια διαγώνιο των πινάκων διακύμανσης συνδιακύμανσης είχε μηδαμινή επιρροή στα αποτελέσματα του βέλτιστου αριθμού ομάδων που πρότειναν τα κριτήρια.

- **Δοκιμή 6 (Εφαρμογή ιεραρχικού clustering στα δεδομένα της Δοκιμής 4)**

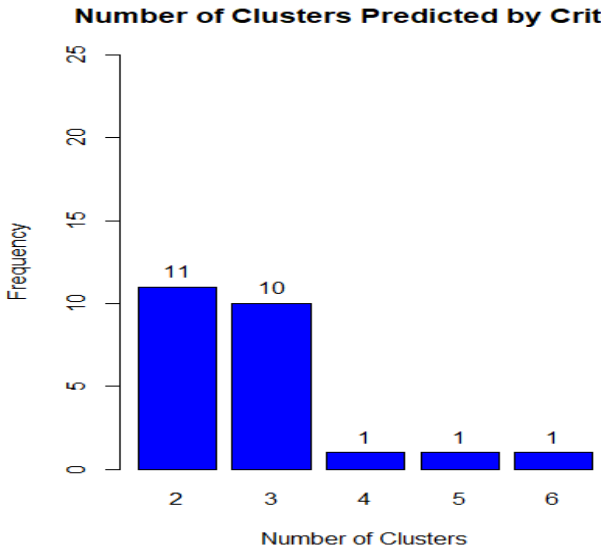
Στην τελευταία δοκιμή θα προσπαθήσουμε να παρατηρήσουμε εάν και κατά πόσο έχοντας ίδια δεδομένα παίζει ρόλο η μέθοδος με την οποία γίνεται το clustering. Πιο συγκεκριμένα, θα χρησιμοποιήσουμε τα δεδομένα της δοκιμής 4 όπως ήταν και θα εφαρμόσουμε ιεραρχική ομαδοποίηση μέσω της μεθόδου του Ward.

Ένας επιπλέον έλεγχος που θα μπορούσε να γίνει στην προκειμένη περίπτωση είναι να ελέγξουμε αν τα κριτήρια προτείνουν τον ίδιο αριθμό ομάδων με αυτόν που θα επιλέγαμε μέσω του δενδρογράμματος.

Η δομή των δεδομένων παραμένει ίδια, οπότε παρακάτω θα δούμε τον πίνακα με τα αποτελέσματα των κριτηρίων σχετικά με το βέλτιστο αριθμό ομάδων.

```
> print(nbclust_result4hier)
$All.index
      KL       CH Hartigan     CCC    Scott      Marriot   TrCovW    TraceW Friedman   Rubin Cindex     DB Silhouette   Duda
2 9.4047 5276.459 945.2624 40.0729 4372.389 8.283170e+11 20204160 20932.362   33.3323  9.2913 0.3080 0.4664     0.6845 0.5225
3 4.1506 4772.439 215.5238 28.3191 6168.348 5.628538e+11 20679188 12833.938   48.1513 15.1542 0.2335 0.8277     0.5239 0.7324
4 0.9592 3709.048 144.7260 19.5760 6722.052 6.917665e+11 16014649 11218.767   54.8376 17.3360 0.2232 1.1857     0.4840 0.6499
5 1.7239 3085.019 141.1539 21.1452 7116.118 8.311618e+11 14725315 10229.176   60.0489 19.0131 0.2211 1.2700     0.3886 0.7778
6 3.2868 2727.444 146.3095 18.5431 7413.140 9.818645e+11 10846441  9346.687   63.3420 20.8082 0.2053 1.5924     0.2304 0.6854
7 0.3590 2518.154  97.8111 17.3392 7773.527 1.050999e+12  8081310  8512.997   67.1813 22.8460 0.2455 1.4954     0.2286 0.7045
8 1.3416 2312.245  92.0747 15.7550 8041.568 1.148102e+12  6518367  7989.575   70.8439 24.3427 0.2611 1.4827     0.2246 0.5718
  Pseudot2  Beale Ratkowsky       Ball Ptbiserial     Frey McClain   Dunn Hubert SDindex Dindex   SDbw
2 911.8892 1.5540    0.6192 10466.1812     0.8557  2.2042  0.2582 0.2220      0  0.3101 3.4133 0.2135
3 163.6657 0.6205    0.5368  4277.9794     0.6932  1.5343  0.5814 0.0316      0  0.5384 2.6263 0.1962
4 295.1722 0.9153    0.4686  2804.6917     0.6562 23.5853  0.6811 0.0327      0  0.7879 2.4770 0.3102
5 142.3056 0.4855    0.4211  2045.8352     0.5550  1.9414  1.0140 0.0198      0  1.1611 2.3364 0.3986
6 121.1820 0.7785    0.3863  1557.7811     0.4878  0.5596  1.3712 0.0198      0  0.9633 2.2274 0.3213
7  76.3416 0.7102    0.3597  1216.1424     0.4772  0.6281  1.4396 0.0249      0  0.9573 2.1441 0.3308
8 145.2508 1.2681    0.3376   998.6969     0.4703  1.2823  1.4840 0.0272      0  0.9922 2.0849 0.3049

$All.CriticalValues
  CritValue_Duda CritValue_PseudoT2 Fvalue_Beale
2         0.7172            393.5053       0.1985
3         0.6826            208.3408       0.6017
4         0.6926            243.2017       0.4327
5         0.6880            225.8668       0.6924
6         0.6509            141.5699       0.5061
7         0.6232            110.0212       0.5462
8         0.6284            114.7406       0.2845

$Best.nc
                     KL       CH Hartigan     CCC    Scott      Marriot  TrCovW   TraceW Friedman   Rubin Cindex     DB
Number_clusters 2.0000    2.000   3.0000  2.0000    3.000            3       4    3.000    3.000  3.0000 6.0000 2.0000
Value_Index     9.4047 5276.459 729.7386 40.0729 1795.959 394375850222 4664538 6483.253   14.819 -3.6812 0.2053 0.4664
                Silhouette    Duda PseudoT2 Beale Ratkowsky     Ball PtBiserial   Frey McClain  Dunn Hubert SDindex Dindex   SDbw
Number_clusters     2.0000 3.0000   3.0000 2.000    2.0000    3.000     2.0000 5.0000  2.0000 2.000      0  2.0000      0 3.0000
Value_Index         0.6845 0.7324 163.6657 1.554    0.6192 6188.202     0.8557 1.9414  0.2582 0.222      0  0.3101      0 0.1962
```

Σχήμα 3.31 Απεικόνιση των αποτελεσμάτων στη Δοκιμή 6



Σχήμα 3.32 Απεικόνιση barplot σχετικά με το πλήθος των κριτηρίων που πρότειναν κάθε αριθμό clusters στη Δοκιμή 6

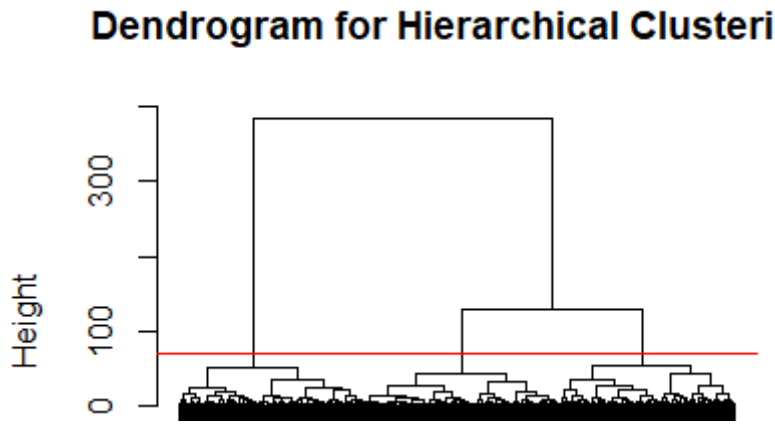
Υπάρχει μικρή διαφοροποίηση αφού πλέον 11 κριτήρια προτείνουν 2 ομάδες έναντι 10 στη διατέλεση της δοκιμής 4. Επιπλέον στην παρούσα δοκιμή έχουμε 10 κριτήρια να προτείνουν 3 ομάδες ενώ στη δοκιμή 4 είναι 8.

Εκ πρώτης όψεως θα μπορούσαμε να πούμε πως το πείραμα παρατηρούμε ότι και βασιζόμενοι στο κριτήριο της πλειοψηφίας δεν αλλάζει η τελική πρόταση, όμως αν κοιτάξουμε κάθε κριτήριο ξεχωριστά και παρατηρήσουμε τον προτεινόμενο αριθμό ομάδων μεμονωμένα θα δούμε πως υπάρχουν 3 κριτήρια που διαφοροποιούν τον αριθμό ομάδων που προτείνουν.

Με βάση το γεγονός πως αρκετοί ερευνητές στρέφονται στον κανόνα της πλειοψηφίας ώστε να εξάγουν όσο το δυνατόν πιο έγκυρο πόρισμα αναφορικά με το βέλτιστο αριθμό ομάδων 3 κριτήρια είναι αρκετά για να αλλάξουν την απόφαση αυτή.

Στη συνέχεια και λαμβάνοντας υπόψη τον κανόνα της πλειοψηφίας δηλαδή ότι τα κριτήρια αποτύπωσαν ως βέλτιστο αριθμό ομάδων κατά πλειοψηφία τις 3 ομάδες, κατασκευάσαμε το δενδρόγραμμα που προκύπτει σαν αποτέλεσμα της ιεραρχικής μεθόδου clustering.

Η αξιολόγηση μέσω δενδρογράμματος όπως έχει προηγουμένως αναλυθεί υπόκεινται στην υποκειμενικότητα κάθε ερευνητή. Παρακάτω παρουσιάζεται το γράφημα στην περίπτωσή μας.



Σχήμα 3.33 Απεικόνιση δενδρογράμματος με τα αποτελέσματα clustering δοκιμής 6

Παρατηρούμε ότι η απόσταση των κόμβων μεγαλώνει αρκετά στο σημείο που βρίσκεται η κόκκινη οριζόντια γραμμή η οποία τέμνεται σε 3 σημεία. Μπορούμε να πούμε λοιπόν πως με βάση το δενδρόγραμμα ο βέλτιστος αριθμός ομάδων που προτείνεται είναι οι 3.

Συμπεραίνουμε λοιπόν πως με βάση το δενδρόγραμμα και τον κανόνα πλειοψηφίας από τα κριτήρια αξιολόγησης δεν έχουμε την ίδια πρόταση ομάδων. Παρόλα αυτά τα πράγματα είναι πολύ οριακά μέσω των κριτηρίων και επιπλέον η απόφαση με το δενδρόγραμμα είναι υποκειμενική, κάποιος άλλος ερευνητής θα μπορούσε μέσω δενδρογράμματος να αποφασίσει τις 2 ομάδες.

Συμπερασματικά, παρατηρούμε κοντινά αλλά διαφοροποιημένα αποτελέσματα εφαρμόζοντας 2 διαφορετικές τεχνικές ομαδοποίησης στα ίδια δεδομένα κάτι το οποίο είναι αναμενόμενο. Παρατηρούνται για τα συγκεκριμένα δεδομένα στις συγκεκριμένες συνθήκες που πραγματοποιήθηκαν οι δοκιμές ότι υπάρχουν διαφορές μεταξύ των κριτηρίων και αυτό μπορεί να μεταβάλλεται αναλόγως της δομής, πυκνότητας των δεδομένων.

Οι διαφοροποιήσεις ως ένα βαθμό είναι λογικές καθώς αλλάζει ο τρόπος ομαδοποίησης των δεδομένων μέσα από διαφορετικές μεθόδους και ο ερευνητής θα πρέπει να προσφύγει σε πιο διεξοδικό έλεγχο ώστε να καταλήξει στην πιο ταιριαστή μέθοδο και κριτήρια πριν λάβει την τελική απόφαση για το βέλτιστο αριθμό ομάδων.

3.8. Πίνακας αποτελεσμάτων και γενικά συμπεράσματα

Στον παρακάτω πίνακα συνοψίζονται τα αποτελέσματα των δοκιμών μας αναφορικά με το βέλτιστο αριθμό ομάδων που προτείνουν τα κριτήρια κατά τις διαφορετικές δοκιμές που πραγματοποιήθηκαν. Περιέχονται όλα τα κριτήρια πέραν του κριτηρίου Frey που δεν είχαμε διαθέσιμη σε όλες τις δοκιμές την τιμή του και τα κριτήρια μέσω γραφικών μεθόδων.

Κριτήρια/Δοκιμές	Initial Data	Δοκιμή 1	Δοκιμή 2	Δοκιμή 3	Δοκιμή 4	Δοκιμή 5	Δοκιμή 6
KL	8	3	3	7	4	4	2
CH	3	3	3	3	2	2	2
Hartigan	3	3	3	3	3	3	3
CCC	3	2	3	3	2	2	2
Scott	3	3	3	3	3	3	3
Marriot	3	3	3	3	3	3	3
TrCovW	3	3	3	3	5	5	4
TraceW	3	3	3	3	3	3	3
Friedman	3	3	3	3	3	3	3
Rubin	3	3	3	3	3	3	3
Cindex	8	3	8	4	5	6	6
DB	3	2	3	3	2	2	2
Silhouette	3	2	3	3	2	2	2
Duda	3	3	3	3	4	4	3
Pseudot2	3	3	3	3	4	4	3
Beale	3	3	3	3	2	2	2
Ratkowsky	2	2	2	2	2	2	2
Ball	3	3	3	3	3	3	3
Ptbiserial	3	2	3	3	2	2	2
McClain	2	2	2	2	2	2	2
Dunn	3	2	3	2	2	2	2
SDindex	3	2	3	3	2	2	2
SDbw	3	3	3	3	3	3	3

Πίνακας 3. Τελικός συγκεντρωτικός πίνακας αποτελεσμάτων

Όπως παρατηρούμε τα κριτήρια πήραν διάφορες τιμές με τις περισσότερες από αυτές να κυμαίνονται στο διάστημα 2 έως 4. Με δεδομένο ότι τα προσομοιωμένα δεδομένα αρχικώς δημιουργήθηκαν με σκοπό να αποτελούν 3 ξεχωριστές διακριτές ομάδες, θα μπορούσαμε να πούμε ότι στην πλειοψηφία τους τα κριτήρια ακόμη και αν δεν προβλέψουν το σωστό αριθμό ομάδων είναι κοντά σε αυτόν.

Στις δοκιμές ασχοληθήκαμε με την απόδοση των κριτηρίων σε περιπτώσεις που αλλάζει ο βαθμός διαχωρισμού των ομάδων. Υπάρχουν κριτήρια που μοιάζουν να είναι ανθεκτικά σε

αυτές τις μεταβολές αποστάσεων λόγω αλλαγής των παραμέτρων $\boldsymbol{\mu}$, Σ και πιο συγκεκριμένα τα κριτήρια αυτά είναι του Hartigan, Scott, Marriot, TraceW, Friedman, Rubin, Ball, SDbw. Θα μπορούσαμε να ισχυριστούμε ότι τα συγκεκριμένα κριτήρια παρουσιάζουν κάποια ανθεκτικότητα στις μεταβολές στις οποίες υποβλήθηκαν κάτι που ίσως να μην είναι ιδιαίτερα βοηθητικό σε πιο περίπλοκες δομές δεδομένων.

Παρατηρήσαμε επίσης πως υπήρχαν κριτήρια που συστηματικά τα αποτελέσματά τους απέκλιναν των υπολοίπων και ήταν αρκετά μακριά από αυτά. Φαίνεται πως δυσκολεύτηκαν με τη συγκεκριμένη δομή των δεδομένων μας, κάτι το οποίο τα καθιστά μη βοηθητικά για την έρευνα μας στις συνθήκες που αυτή διεξάχθηκε. Παρόλα αυτά, θα μπορούσαν να φανούν πολύ χρήσιμα σε άλλες δομές δεδομένων. Τα κριτήρια αυτά είναι το KL,TRCovW και to Cindex.

Υπήρχαν επίσης περιπτώσεις που κριτήρια μετέβαλαν την πρόταση του βέλτιστου αριθμού ομάδων ανάμεσα στο 2 και το 4 και αυτό δείχνει ότι είναι ευαίσθητα στις αλλαγές που πραγματοποιήσαμε, αλλά ταυτόχρονα διατήρησαν μια σχετική ακρίβεια στον αριθμό των ομάδων που πρότειναν. Τα κριτήρια αυτά είναι το CH, CCC, DB,Silhouette, Duda, Pseudot2, Beale , Ptbiserial, Dunn, SDindex.

Επιπλέον αξίζει να αναφέρουμε πως 2 κριτήρια, τα Ratkowsky και McClain, πρότειναν σε όλες τις δοκιμές ως βέλτιστο αριθμό ομάδων τις 2, ακόμη και σε περιπτώσεις που οι ομάδες ήταν ευδιάκριτα διαχωρισμένες σε 3 και μπορούμε να ισχυριστούμε πως για τις συγκεκριμένες δοκιμές στα συγκεκριμένα δεδομένα δε λειτούργησαν καλά και δεν έδειξαν κάποια ευαισθησία στις μεταβολές της δομής των ομάδων ανά τις δοκιμές.

Αναφορικά με κάποια κοινά χαρακτηριστικά των δοκιμών όπως για παράδειγμα τις δοκιμές 4 και 5 που ουσιαστικά εκτελέσθηκαν με τα ίδια διανύσματα μέσων και τις ίδιες διασπορές προσθέτοντας μόνο στα μη διαγώνια στοιχεία τιμές διαφορετικές του μηδενός. Στη συγκεκριμένη λοιπόν δοκιμή μπορούμε να ισχυριστούμε πως η προσθήκη εξάρτησης ανάμεσα στους πληθυσμούς δεν είχε κάποιο αντίκτυπο στα αποτελέσματα μας καθώς για όλα τα κριτήρια πέραν του Cindex έχουμε ακριβώς τα ίδια αποτελέσματα. Λαμβάνοντας λοιπόν υπόψη πως το συγκεκριμένο κριτήριο δε θεωρήθηκε ότι είχε καλή επίδοση στις δοκιμές μας μπορούμε να εξάγουμε το πόρισμα πως δεν υπήρχαν διαφορές μεταξύ των δοκιμών.

Επιπλέον, συγκρίναμε διαφορετικές μεθόδους συσταδοποίησης ώστε να ελέγξουμε αν και κατά πόσο επηρεάζονται τα αποτελέσματα των κριτηρίων που εξετάζουμε.

Στις δοκιμές που εκτελέσθηκαν παρουσιάστηκαν διαφοροποιήσεις σε ορισμένα κριτήρια ενώ στα περισσότερα δεν υπήρχε αλλαγή. Στη μεγάλη εικόνα και σε επίπεδο κανόνα πλειοψηφίας θα μπορούσαμε να πούμε πως δεν άλλαξε κάτι καθώς για 15 από τα 20 κριτήρια δεν άλλαξε ο αριθμός προτεινόμενων ομάδων, παρόλα αυτά με δεδομένο πως οι δοκιμές 5 και 6 εκτελέσθηκαν υπό τις ίδιες ακριβώς συνθήκες απλά εφαρμόζοντας άλλο αλγόριθμο συσταδοποίησης είναι αρκετή η διαφοροποίηση σε 5 κριτήρια. Σε πιο πολύπλοκες δομές δεδομένων όπως για παράδειγμα τα κυκλικά δεδομένα ή τα δεδομένα που περιέχουν ακραίες τιμές οι διαφορές αυτές θα ήταν πιο διακριτές αναλόγως των αλγορίθμων που θα εφαρμόζαμε. Είναι γνωστό πως για δεδομένα με ακραίες τιμές οι αλγόριθμοι πυκνότητας είναι πιο

αποδοτικοί από κλασσικούς αλγορίθμους συσταδοποίησης όπως ο K-means λόγω της διαφορετικής λογικής τους.

Εν κατακλείδι, θα μπορούσαμε να πούμε ότι μέσω των δοκιμών που διενεργήθηκαν μπορέσαμε να εντοπίσουμε κάποια χαρακτηριστικά ανάμεσα στα κριτήρια που εξετάσαμε, όπως για παράδειγμα κατά πόσο ανθεκτικά ή ευαίσθητα μπορεί να είναι στις αλλαγές των παραμέτρων $\boldsymbol{\mu}$, Σ στα προσομοιωμένα δεδομένα που η μελέτη χρησιμοποίησε. Μέσω του αλγορίθμου K-means στον όποιο βασίσαμε την έρευνά μας, είδαμε πως περισσότερες μεταβολές στα αποτελέσματα των κριτηρίων παρατηρήθηκαν όταν υπήρχε αλλαγή στον πίνακα διακύμανσης συνδιακύμανσης, δηλαδή περισσότερα κριτήρια τα οποία είχαν σταθερό αριθμό στις πρώτες δοκιμές (δοκιμές initial data, 1,2) παρουσίασαν ευαισθησία είτε όταν μεταβάλαμε αποκλειστικά τον πίνακα Σ είτε όταν ταυτοχρόνως μεταβλήθηκαν και τα διανύσματα των μέσων.

Παρατηρήσαμε τις διαφοροποιήσεις που μπορούν να υπάρξουν στα αποτελέσματα των κριτηρίων με την αλλαγή μεθόδου clustering ακόμη και όταν όλες οι υπόλοιπες συνθήκες (τα δεδομένα που εξετάζουμε) είναι ίδιες και εν γένει έγινε σύγκριση των κριτηρίων μεταξύ τους ώστε να παρέχουμε μία κατεύθυνση ως προς τα χαρακτηριστικά τους που πιθανόν να βοηθήσει μελλοντικές έρευνες.

Τα παρεχόμενα αποτελέσματα μπορεί να διαφοροποιηθούν εφόσον αλλάξει η δομή των δεδομένων μας και κάθε φορά ο ερευνητής θα πρέπει να αξιολογήσει διεξοδικά την καταλληλότητα των κριτηρίων με βάση το τι ερευνά, τι δεδομένα έχει αλλά λαμβάνοντας επίσης υπόψη τις συνθήκες κάτω από τις οποίες θα γίνει η έρευνα.

ΚΕΦΑΛΑΙΟ 4

ΕΠΙΛΟΓΟΣ

4.1.Σκοπός έρευνας και γενικά συμπεράσματα

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 πιο εύκολη και πιο αποδοτική επεξεργασία των δεδομένων πουδιαθέτουμε.

Η παρούσα Διπλωματική διερεύνησε και συνέκρινε διάφορα κριτήρια αξιολόγησης της συσταδοποίησης σε πολυδιάστατα προσομοιωμένα σύνολα δεδομένων με στόχο τον εντοπισμό του βέλτιστου αριθμού συστάδων κάτω από διαφορετικά επίπεδα διαχωρισμού των ομάδων. Τα ευρήματα δίνουν κάποιες κατευθύνσεις και προσπαθούν να βρουν κάποια κοινά χαρακτηριστικά ανάμεσα στα κριτήρια ώστε να παρέχουν πολύτιμα συμπεράσματα για μελλοντικές μελέτες σχετικά με τις προκλήσεις και τις ευκαιρίες στην ανάλυση ομαδοποίησης.

Τα αποτελέσματα όπως προέκυψαν υπογραμμίζουν την ευαισθησία των αποτελεσμάτων που παρέχουν τα κριτήρια στην επιλογή της μεθόδου αλλά και στη δομή των δεδομένων και τονίζει την ανάγκη για τους ερευνητές να εξετάσουν προσεκτικά την καταλληλότητα της προσέγγισης ομαδοποίησης με βάση τα συγκεκριμένα χαρακτηριστικά του συνόλου δεδομένων.

Τα ευρήματα των κριτηρίων έδειξαν ότι ο βέλτιστος αριθμός συστάδων διέφερε σε αρκετές περιπτώσεις, ενισχύοντας την ιδέα ότι ένα μεμονωμένο κριτήριο ομαδοποίησης μπορεί να μην επαρκεί για τον προσδιορισμό του ιδανικού αριθμού συστάδων σε όλες τις περιπτώσεις. Οι ερευνητές πρέπει να συνδυάσουν πολλαπλές μετρήσεις επικύρωσης για να αποκτήσουν πιο ισχυρά και αξιόπιστα αποτελέσματα.

Σημαντικό ερώτημα είναι ο ρόλος για τον οποίο δημιουργήθηκαν τα κριτήρια που αξιολογούν το επίπεδο ομαδοποίησης. Η απάντηση βρίσκεται στην ανάγκη γνώσης και εφαρμογής μίας μεθόδου που θα μπορούσε να ανιχνεύσει το πόσες ομάδες υπάρχουν στα σύνολα δεδομένων. Η πληροφορία αυτή ειδικότερα σε πραγματικά δεδομένα πολλές φορές δεν είναι ακριβής ή δεν υπάρχει. Στόχος των ερευνητών είναι να βρουν χρήσιμες δομές σε ένα σύνολο δεδομένων χωρίς όμως στις περισσότερες περιπτώσεις να αναμένουν κάποιο πολύ συγκεκριμένο αποτέλεσμα. Στην ανάγκη αυτή έπαιξε πρωταγωνιστικό ρόλο η ομαδοποίηση δεδομένων σε αρκετά επιστημονικά πεδία. Μέσω διαφόρων μετρικών λοιπόν προσπαθούν να το πετύχουν και τα κριτήρια είναι εδώ για να αξιολογήσουν τα αποτελέσματα αυτά.

Τέλος αξίζει να αναφέρουμε ότι η ανάλυση κατά συστάδες κατέχει δεσπόζουσα θέση και εφαρμόζεται σε διαφόρους τομείς όπως τις Βιοεπιστήμες, Πολιτικές και Οικονομικές επιστήμες, Ιατρικές επιστήμες και στο Μάρκετινγκ. Τα τελευταία χρόνια ωστόσο φαίνεται πως έχει παίξει πρωταγωνιστικό ρόλο και στον τομέα της Πληροφορικής με πιο χαρακτηριστικά παραδείγματα τη ρομποτική , την αναγνώριση προτύπων. Ιδιαίτερα βοηθητική φαίνεται πως είναι η μέθοδος σε περιπτώσεις όπου οι ερευνητές προσπαθούν να ομαδοποιήσουν τη συμπεριφορά των χρηστών του διαδικτύου, που πλέον έχει γίνει καθημερινότητα. Στην εποχή μας ολοένα και περισσότερα επαγγέλματα προσανατολισμένα στον πελάτη προσπαθούν να τον κατατάξουν σε μία ομάδα συμπεριφοράς ώστε να μεγιστοποιήσουν τα οφέλη τους για παράδειγμα χρήση προσωποποιημένων διαφημίσεων σε συγκεκριμένο κοινό που ανήκει σε μία ομάδα συγκεκριμένης συμπεριφοράς με βάση τη συμπεριφορά τους όταν σερφάρουν σε διαφορετικές σελίδες ή με βάση το τι κλικάρουν σε κάποια σελίδα Καρλής (2005).

4.2.Προτάσεις για μελλοντική έρευνα

Η παρούσα Διπλωματική έχει προσπαθήσει να συνεισφέρει πολύτιμα συμπεράσματα σχετικά με τα κριτήρια αξιολόγησης των τεχνικών ομαδοποίησης καθώς και τις ίδιες τις τεχνικές και την εφαρμογή τους σε πολυδιάστατα σύνολα δεδομένων. Ωστόσο, υπάρχουν αρκετές οδοί για μελλοντική έρευνα που μπορούν να βασιστούν στα ευρήματα και να εμπλουτίσουν περαιτέρω το πεδίο της ανάλυσης ομαδοποίησης. Οι ακόλουθες προτάσεις προσφέρουν οδηγίες για μελλοντικές έρευνες:

1. Ισχυρότητα των μεθόδων ομαδοποίησης: Η μελλοντική έρευνα θα μπορούσε να επικεντρωθεί στην αξιολόγηση της ευρωστίας των μεθόδων ομαδοποίησης σε ένα ευρύτερο φάσμα συνόλων δεδομένων και σενάρια πραγματικού κόσμου. Η διερεύνηση του τρόπου με τον οποίο λειτουργούν διαφορετικές τεχνικές ομαδοποίησης κάτω από διαφορετικές κατανομές δεδομένων, μεγέθη δειγμάτων και επίπεδα θορύβου θα παρέχει μια βαθύτερη κατανόηση της αξιοπιστίας των μεθόδων.

2. Υβριδικές προσεγγίσεις ομαδοποίησης: Η διερεύνηση τεχνικών υβριδικής ομαδοποίησης που συνδυάζουν τα δυνατά σημεία πολλαπλών μεθόδων θα μπορούσε να είναι μια πολλά υποσχόμενη κατεύθυνση. Ενσωματώνοντας τα πλεονεκτήματα διαφορετικών αλγορίθμων ομαδοποίησης, οι ερευνητές μπορούν ενδεχομένως να επιτύχουν πιο ακριβείς και ισχυρές λύσεις ομαδοποίησης για πολύπλοκα σύνολα δεδομένων.

3. Βελτιωμένες τεχνικές επικύρωσης: Η μελέτη νέων ή βελτιωμένων τεχνικών επικύρωσης ειδικών για δεδομένα υψηλών διαστάσεων είναι απαραίτητη. Νέες προσεγγίσεις που αξιολογούν αποτελεσματικά την ποιότητα της ομαδοποίησης, αντιμετωπίζουν την κατάρα της διάστασης και λαμβάνουν υπόψη την πολυδιάστατη φύση των δεδομένων μπορούν να οδηγήσουν σε πιο αξιόπιστα αποτελέσματα ομαδοποίησης.

4. Ομαδοποίηση για συγκεκριμένο τομέα: Η διερεύνηση τεχνικών ομαδοποίησης προσαρμοσμένων σε συγκεκριμένους τομείς, όπως η υγειονομική περίθαλψη, τα οικονομικά ή τα κοινωνικά δίκτυα, μπορεί να δώσει πολύτιμες πληροφορίες για μοτίβα μοναδικά σε αυτούς τους τομείς. Οι μέθοδοι ομαδοποίησης μπορούν να συλλάβουν καλύτερα τα χαρακτηριστικά των δεδομένων για συγκεκριμένο τομέα και να παρέχουν πιο ουσιαστικές λύσεις ομαδοποίησης. Ως συνέπεια, ενδιαφέρον θα είχε η δημιουργία προσαρμοσμένων κριτηρίων για διαφορετικούς τομείς και η σύγκρισή τους με τα ήδη υπάρχοντα κριτήρια.

5. Ανάλυση Δυναμικής Ομαδοποίησης: Η διερεύνηση μεθόδων ομαδοποίησης και κριτηρίων αξιολόγησης που προσαρμόζονται στην αλλαγή δεδομένων με την πάροδο του χρόνου είναι ένας αναδυόμενος τομέας έρευνας. Η ανάλυση δυναμικής ομαδοποίησης θα μπορούσε να εφαρμοστεί σε σύνολα δεδομένων με εξελισσόμενες δομές για την παρακολούθηση των αλλαγών σε συστάδες και τον εντοπισμό χρονικών προτύπων.

6. Ομαδοποίηση δεδομένων μεγάλης κλίμακας: Η ανάλυση συνόλων δεδομένων μεγάλης κλίμακας (big data) θέτει μοναδικές προκλήσεις. Η μελλοντική έρευνα θα μπορούσε να επικεντρωθεί στην ανάπτυξη κλιμακούμενων τεχνικών ομαδοποίησης που χειρίζονται αποτελεσματικά τεράστιους όγκους δεδομένων, διατηρώντας παράλληλα την ακρίβεια και την ερμηνευτικότητα της ομαδοποίησης. Θα είχε ενδιαφέρον η συμπεριφορά των κριτηρίων σε τέτοιου τύπου δεδομένα.

Συμπερασματικά, οι προτάσεις για μελλοντική έρευνα στοχεύουν στην προώθηση της κατανόησης και της εφαρμογής των τεχνικών ομαδοποίησης σε πολυδιάστατα σύνολα δεδομένων και κατ' επέκταση των κριτηρίων αξιολόγησης. Διερευνώντας νέες προσεγγίσεις, μεθόδους επικύρωσης και εφαρμογές για συγκεκριμένους τομείς, οι ερευνητές μπορούν να συμβάλουν στη συνεχή πρόοδο της ανάλυσης ομαδοποίησης και των διαφορετικών εφαρμογών της σε διάφορους τομείς. Αυτές οι κατευθύνσεις προσφέρουν συναρπαστικές ευκαιρίες για περαιτέρω ενίσχυση της χρησιμότητας και του αντίκτυπου των αλγορίθμων ομαδοποίησης στην επιστήμη δεδομένων και όχι μόνο.

ΠΑΡΑΡΤΗΜΑ

```
library(MASS)
library(NbClust)
library("scatterplot3d")

# Set random seed for reproducibility
set.seed(1445)

# Number of data points in each cluster
n <- 500

# Mean vectors for the three clusters
mu1 <- c(0, 1, 0)
mu2 <- c(12, 10, 9)
mu3 <- c(4, 6, 5)


# Covariance matrix for each cluster
sigma <- diag(3)  # Identity matrix for simplicity

# Generate data for each cluster
cluster1 <- mvrnorm(n, mu1, sigma)
cluster2 <- mvrnorm(n, mu2, sigma)
cluster3 <- mvrnorm(n, mu3, sigma)
summary(cluster1)
summary(cluster2)
summary(cluster3)
# Combine the clusters
data <- rbind(cluster1, cluster2, cluster3)

# Plot the data
scatterplot3d(data[,1], data[,2], data[,3], color=c(rep("red", n), rep("green", n), rep("blue", n)),
pch = 16, box = FALSE, angle=70,xlab=" Dimension 1",ylab=" Dimension 2",zlab="
Dimension 3",xlim=c(-10,20),ylim=c(-10,20),zlim=c(-10,20),main="Initial data")

#plot3d(data[, 1], data[, 2], data[, 3], col = c(rep("red", n), rep("green", n), rep("blue", n)))
#CHECKED
#s3d$plane3d(cluster1, cluster2, cluster3, alpha = 0.1)  # Adds planes to visualize the clusters

library(NbClust)
nbclust_result <- NbClust(data, distance = "euclidean", min.nc = 2, max.nc = 8, method =
"kmeans")
print(nbclust_result)

# Assuming you already have the `nbclust_result` from NbClust
```

```
# Extract the number of clusters predicted by each criterion
num_clusters <- nbclust_result$Best.nc

# Calculate the frequency of each number of clusters
cluster_counts <- table(num_clusters)
# Define a vector of valid cluster counts you want to include
valid_counts <- c(2, 3, 4, 5, 6, 7, 8)
# Filter the cluster counts to include only valid counts
filtered_counts <- cluster_counts[names(cluster_counts) %in% valid_counts]
# Convert the names of filtered cluster counts to a factor for proper labeling
cluster_counts_factor <- factor(as.integer(names(filtered_counts)))
# Create a barplot with axis labels and a title
bp<-barplot(filtered_counts,
      xlab = "Number of Clusters",
      ylab = "Frequency",
      col = "blue",
      names.arg = cluster_counts_factor,
      main = "Number of Clusters Predicted by Criteria", ylim = c(0,25),width = 0.2)
text(bp, filtered_counts, labels = filtered_counts, pos = 3)
print(bp)

# 1st trial bring closer the 2 clusters

mu1v2 <- c(1, 2, 3)
mu2v2 <- c(5, 4, 7)
mu3v2 <- c(12, 6, 16)

# Generate data for each cluster
cluster1v2 <- mvrnorm(n, mu1v2, sigma)
cluster2v2 <- mvrnorm(n, mu2v2, sigma)
cluster3v2 <- mvrnorm(n, mu3v2, sigma)

# Combine the clusters
datav2 <- rbind(cluster1v2, cluster2v2, cluster3v2)

# Plot the data
library("scatterplot3d")
scatterplot3d(datav2[,1], datav2[,2], datav2[,3], color=c(rep("red", n), rep("green", n),
rep("blue", n)),pch = 16, box = FALSE,angle=70,xlab=" Dimension 1",ylab=" Dimension
2",zlab=" Dimension 3",xlim=c(-10,20),ylim=c(-10,20),zlim=c(-10,20),main="Changed only
mu-1st trial")

nbclust_resultv2 <- NbClust(datav2, distance = "euclidean", min.nc = 2, max.nc = 8, method
= "kmeans")
print(nbclust_resultv2)

num_clustersv1 <- nbclust_resultv2$Best.nc
```

```
# Calculate the frequency of each number of clusters
cluster_countsv1 <- table(num_clustersv1)

# Define a vector of valid cluster counts you want to include
valid_countsv1 <- c(2, 3, 4, 5, 6, 7, 8)

# Filter the cluster counts to include only valid counts
filtered_countsv1 <- cluster_countsv1[names(cluster_countsv1) %in% valid_countsv1]

# Convert the names of filtered cluster counts to a factor for proper labeling
cluster_counts_factorv1 <- factor(as.integer(names(filtered_countsv1)))

# Create a barplot with axis labels and a title
bpv1<-barplot(filtered_countsv1,
        xlab = "Number of Clusters",
        ylab = "Frequency",
        col = "blue",
        names.arg = cluster_counts_factorv1,
        main = "Number of Clusters Predicted by Criteria", ylim = c(0,25))

text(bpv1, filtered_countsv1, labels = filtered_countsv1, pos = 3)
print(bpv1)

# 2nd trial  more distance
mu1v1 <- c(1, 2, 3)
mu2v1 <- c(9, 8, 8)
mu3v1 <- c(16, 14, 19)
# Generate data for each cluster
cluster1v1 <- mvrnorm(n, mu1v1, sigma)
cluster2v1 <- mvrnorm(n, mu2v1, sigma)
cluster3v1 <- mvrnorm(n, mu3v1, sigma)
# Combine the clusters
datav1 <- rbind(cluster1v1, cluster2v1, cluster3v1)
# Plot the data
library("scatterplot3d")
scatterplot3d(datav1[,1], datav1[,2], datav1[,3], color=c(rep("red", n), rep("green", n),
rep("blue", n)),pch = 16, box = FALSE,angle=70,xlab=" Dimension 1",ylab=" Dimension
2",zlab=" Dimension 3",xlim=c(-10,20),ylim=c(-10,20),zlim=c(-10,20),main="Changed only
mu-2nd trial")

nbclust_resultv1 <- NbClust(datav1, distance = "euclidean", min.nc = 2, max.nc = 8, method
= "kmeans")
print(nbclust_resultv1)
num_clustersv2 <- nbclust_resultv1$Best.nc

# Calculate the frequency of each number of clusters
cluster_countsv2 <- table(num_clustersv2)
```

```
# Define a vector of valid cluster counts you want to include
valid_countsv2 <- c(2, 3, 4, 5, 6, 7, 8)

# Filter the cluster counts to include only valid counts
filtered_countsv2 <- cluster_countsv2[names(cluster_countsv2) %in% valid_countsv2]

# Convert the names of filtered cluster counts to a factor for proper labeling
cluster_counts_factorv2 <- factor(as.integer(names(filtered_countsv2)))

# Create a barplot with axis labels and a title
bpv2<-barplot(filtered_countsv2,
        xlab = "Number of Clusters",
        ylab = "Frequency",
        col = "blue",
        names.arg = cluster_counts_factorv2,
        main = "Number of Clusters Predicted by Criteria", ylim = c(0,25))

text(bpv2, filtered_countsv2, labels = filtered_countsv2, pos = 3)
print(bpv2)
######3rd trial
# Covariance matrices for each cluster
sigma1 <- matrix(c(2, 0, 0, 0, 2, 0, 0, 0, 2), nrow = 3, byrow = TRUE)
sigma2 <- matrix(c(1.6, 0, 0, 0, 1.6, 0, 0, 0, 1.4), nrow = 3, byrow = TRUE)
sigma3 <- matrix(c(1.1, 0, 0, 0, 1.1, 0, 0, 0, 1.1), nrow = 3, byrow = TRUE)

# Generate data for each cluster
cluster11 <- mvrnorm(n, mu1, sigma1)
cluster22 <- mvrnorm(n, mu2, sigma2)
cluster33 <- mvrnorm(n, mu3, sigma3)
# Combine the clusters
data2 <- rbind(cluster11, cluster22, cluster33)
scatterplot3d(data2[,1], data2[,2], data2[,3], color=c(rep("red", n), rep("green", n), rep("blue",
n)),pch = 16, box = FALSE,angle=70,xlab=" Dimension 1",ylab=" Dimension 2",zlab="
Dimension 3",xlim=c(-10,20),ylim=c(-10,20),zlim=c(-10,20),main="Changed only sigma-3rd
trial")
nbclust_result2 <- NbClust(data2, distance = "euclidean", min.nc = 2, max.nc = 8, method =
"kmeans")
# Print the NbClust results
print(nbclust_result2)
num_clustersv3 <- nbclust_result2$Best.nc
# Calculate the frequency of each number of clusters
cluster_countsv3 <- table(num_clustersv3)

# Define a vector of valid cluster counts you want to include
valid_countsv3 <- c(2, 3, 4, 5, 6, 7, 8)

# Filter the cluster counts to include only valid counts
```

```
filtered_countsv3 <- cluster_countsv3[names(cluster_countsv3) %in% valid_countsv3]

# Convert the names of filtered cluster counts to a factor for proper labeling
cluster_counts_factorv3 <- factor(as.integer(names(filtered_countsv3)))

# Create a barplot with axis labels and a title
bpv3<-barplot(filtered_countsv3,
        xlab = "Number of Clusters",
        ylab = "Frequency",
        col = "blue",
        names.arg = cluster_counts_factorv3,
        main = "Number of Clusters Predicted by Criteria", ylim = c(0,25))

text(bpv3, filtered_countsv3, labels = filtered_countsv3, pos = 3)
print(bpv3)

# 4rth trial ,decrease the values at the diagonal of cov matrix (closer clusters only changing
sigma)
#create a cov matrix that would bring clusters closer
common_cov_matrix <- matrix(c(
  3.5, 0, 0,
  0, 1.2, 0,
  0, 0, 2.4), nrow = 3, byrow = TRUE)

sigma11 <- common_cov_matrix + diag(3) * 1.95  # Add a smaller diagonal adjustment
sigma22 <- common_cov_matrix + diag(3) * 0.42  # Add an even smaller diagonal
adjustment
sigma33 <- common_cov_matrix - diag(3) * 0.84  # Subtract a smaller diagonal adjustment
cluster11v2 <- mvrnorm(n, mu1*0.7, sigma11)
cluster22v2 <- mvrnorm(n, mu2, sigma22)
cluster33v2 <- mvrnorm(n, mu3*0.7, sigma33)
# Combine the clusters
data3 <- rbind(cluster11v2, cluster22v2, cluster33v2)
s3d <-scatterplot3d(data3[,1], data3[,2], data3[,3], color=c(rep("red", n), rep("green", n),
rep("blue", n)), pch = 16, box = FALSE, angle=70,xlab=" Dimension 1",ylab=" Dimension
2",zlab=" Dimension 3",xlim=c(-10,20),ylim=c(-10,20),zlim=c(-10,20),main="Changed both
mu and sigma-4th trial")

nbclust_result3 <- NbClust(data3, distance = "euclidean", min.nc = 2, max.nc = 8, method =
"kmeans")

# Print the NbClust results
print(nbclust_result3)
num_clusters3 <- nbclust_result3$Best.nc
# Calculate the frequency of each number of clusters
cluster_counts3 <- table(num_clusters3)
# Define a vector of valid cluster counts you want to include
valid_counts3 <- c(2, 3, 4, 5, 6, 7, 8)
```

```
# Filter the cluster counts to include only valid counts
filtered_counts3 <- cluster_counts3[names(cluster_counts3) %in% valid_counts3]

# Convert the names of filtered cluster counts to a factor for proper labeling
cluster_counts_factor3 <- factor(as.integer(names(filtered_counts3)))

# Create a barplot with axis labels and a title
bp3<-barplot(filtered_counts3,
        xlab = "Number of Clusters",
        ylab = "Frequency",
        col = "blue",
        names.arg = cluster_counts_factor3,
        main = "Number of Clusters Predicted by Criteria", ylim = c(0,25))

text(bp3, filtered_counts3, labels = filtered_counts3, pos = 3)
print(bp3)
# 5th trial means from 1st trial and sigmas from 4th trial
sigma111 <- matrix(c(5.45,0.5, 0.3, 0.5, 3.15, 0.2, 0.3, 0.2, 4.35), nrow = 3, byrow = TRUE)
sigma222 <- matrix(c(3.92, 0.4, 0.6, 0.4, 1.62, 0.7, 0.6, 0.7, 2.82), nrow = 3, byrow = TRUE)
sigma333 <- matrix(c(2.66, 0.3, 0.8, 0.3, 0.36, 0.4, 0.8, 0.4, 1.56), nrow = 3, byrow = TRUE)

cluster11v4 <- mvrnorm(n, mu1*0.7, sigma111)
cluster22v4 <- mvrnorm(n, mu2, sigma222)
cluster33v4 <- mvrnorm(n, mu3*0.7, sigma333)

# Combine the clusters
data5 <- rbind(cluster11v4, cluster22v4, cluster33v4)
s3d <-scatterplot3d(data5[,1], data5[,2], data5[,3], color=c(rep("red", n), rep("green", n),
rep("blue", n)), pch = 16, box = FALSE, angle=70,xlab=" Dimension 1",ylab=" Dimension
2",zlab=" Dimension 3",xlim=c(-10,20),ylim=c(-10,20),zlim=c(-10,20),main="5th trial-
Adding Covariance")

nbclust_result5 <- NbClust(data5, distance = "euclidean", min.nc = 2, max.nc = 8, method =
"kmeans")
print(nbclust_result5)

num_clusters5<- nbclust_result5$Best.nc

# Calculate the frequency of each number of clusters
cluster_counts5 <- table(num_clusters5)

# Define a vector of valid cluster counts you want to include
valid_counts5 <- c(2, 3, 4, 5, 6, 7, 8)

# Filter the cluster counts to include only valid counts
filtered_counts5 <- cluster_counts5[names(cluster_counts5) %in% valid_counts5]
```

```
# Convert the names of filtered cluster counts to a factor for proper labeling
cluster_counts_factor5 <- factor(as.integer(names(filtered_counts5)))

# Create a barplot with axis labels and a title
bp5<-barplot(filtered_counts5,
        xlab = "Number of Clusters",
        ylab = "Frequency",
        col = "blue",
        names.arg = cluster_counts_factor5,
        main = "Number of Clusters Predicted by Criteria", ylim = c(0,25))

text(bp5, filtered_counts5, labels = filtered_counts5, pos = 3)
print(bp5)

# 6th gia hierarchical the above
nbclust_result4hier <- NbClust(data3, distance = "euclidean", min.nc = 2, max.nc = 8, method
= "ward.D2")
print(nbclust_result4hier)
s3d <-scatterplot3d(data3[,1], data3[,2], data3[,3], color=c(rep("red", n), rep("green", n),
rep("blue", n)), pch = 16, box = FALSE, angle=70,xlab=" Dimension 1",ylab=" Dimension
2",zlab=" Dimension 3",xlim=c(-10,20),ylim=c(-10,20),zlim=c(-10,20),main="Hierarchical
Clustering with data of 4th trial-6th trial")

num_clusters4hier<- nbclust_result4hier$Best.nc
# Calculate the frequency of each number of clusters
cluster_counts4hier <- table(num_clusters4hier)

# Define a vector of valid cluster counts you want to include
valid_counts4hier <- c(2, 3, 4, 5, 6, 7, 8)

# Filter the cluster counts to include only valid counts
filtered_counts4hier <- cluster_counts4hier[names(cluster_counts4hier) %in%
valid_counts4hier]

# Convert the names of filtered cluster counts to a factor for proper labeling
cluster_counts_factor4hier <- factor(as.integer(names(filtered_counts4hier)))

# Create a barplot with axis labels and a title
bp4hier<-barplot(filtered_counts4hier,
        xlab = "Number of Clusters",
        ylab = "Frequency",
        col = "blue",
        names.arg = cluster_counts_factor4hier,
        main = "Number of Clusters Predicted by Criteria", ylim = c(0,25))

text(bp4hier, filtered_counts4hier, labels = filtered_counts4hier, pos = 3)
print(bp4hier)
```

```
## initial comparison between k means and hierarchical
nbclust_result <- NbClust(data, distance = "euclidean", min.nc = 2, max.nc = 8, method =
"kmeans")
print(nbclust_result)
############ HIERARCHICAL
nbclust_result_aggl <- NbClust(data, distance = "euclidean", min.nc = 2, max.nc = 8, method
= "ward.D2")
print(nbclust_result_aggl)
# Perform hierarchical clustering
hc_result <- hclust(dist(data3, method = "euclidean"), method = "ward.D2")
# Define the range of k values (from 2 to 8)
min_k <- 2
max_k <- 8

# Create an empty plot to set the x-axis range
plot(1, type = "n", xlim = c(min_k, max_k), ylim = c(0, 5), xlab = "Distance", ylab = "")

# Loop through different values of k
for (k in min_k:max_k) {
  # Cut the dendrogram at the current value of k
  clusters <- cutree(hc_result, k = k)

  # Calculate the position for the current k on the x-axis
  x_position <- k

  # Plot a point at the x-axis position for the current k
  points(x_position, 0, pch = 20, col = "red")

  # Plot the dendrogram with colored clusters
  plot(hc_result, main = paste("Dendrogram for Hierarchical Clustering"), sub = "")
   # Add colored borders to clusters
}
abline(h = 70, col = "red", lty = 1)
```

ΒΙΒΛΙΟΓΡΑΦΙΑ

Ελληνική

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

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).

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

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.

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

Ward, J. H. (1963). Hierarchical grouping to optimize an objective function, *Journal of the American Statistical Association*,**58** , 236-244.

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

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.

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.

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

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

Everitt, B. S. (1979). Unresolved Problems in Cluster Analysis, *Biometrics*, **35**, 169-181.

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

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.

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

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